# Essays in Empirical Financial Research

Inauguraldissertation

zur

Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

vorgelegt

von

M.Sc. Frederik André Heinrich Simon

aus

Boston

Referent: Prof. Dr. Dieter Hess
Korreferent: Prof. Dr. Alexander Kempf
Tag der Promotion: 28.10.2024

# Acknowledgements

This thesis marks the end of my doctoral studies at the Department of Finance at the University of Cologne.

First and foremost, I want to express my deepest gratitude towards my supervisor, Prof. Dr. Dieter Hess, for his continuous support, guidance and advice throughout my doctoral journey. His expertise has been instrumental not only in my research endeavors but also in my personal development. I am also indebted to my co-supervisor, Prof. Dr. Alexander Kempf, for his mentorship, encouragement, and the opportunity to be part of the Centre for Financial Research (CFR), Prof. Dr. Ingmar Nolte for reviewing my thesis, and Prof. Dr. Thomas Hartmann-Wendels for chairing the dissertation committee.

Further, I am deeply grateful to my co-authors, Sebastian Weibels and Prof. Dr. Tom Zimmermann, for their collaboration, dedication, and invaluable contributions to our research. I benefited a lot from your input and expertise.

I also thank the participants and discussants at several seminars and conferences, as well as anonymous referees for their helpful comments and suggestions.

Moreover, I am grateful to my colleagues at the Department of Finance at the University of Cologne for their support and all the stimulating discussions we had. I particularly thank Luca Brunke, Dr. Tim Vater, Simon Wolf, Nathalie Zährl, Luca Conrads, Dr. Mario Hendriock, Florian Neitzert, Hendrik Kußmaul and Dr. Alexander Pütz for the time we spent together.

Lastly, I would like to thank my family for their support and the opportunities they have provided me with. It goes without saying that this would not have been possible without you.

Frederik André Heinrich Simon

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The advent of machine learning continues to significantly impact financial research. In his seminal book, Mitchell (1997) broadly defines machine learning as the study of "*computer programs that automatically improve with experience*" (Mitchell, 1997, XV). In the context of empirical financial research, Gu et al. (2020) provide a more granular definition by defining machine learning methods as high-dimensional statistical prediction models, coupled with regularization methods for model selection and mitigation of overfit as well as efficient algorithms for finding optimal model specifications.

Machine learning methods are ideally suited for financial research in a lot of ways. Primarily this is due to two characteristics which many issues in finance share: (1) information sets are typically large in terms of variables and (2) functional forms are oftentimes ambiguous and potentially complex (Kelly and Xiu, 2023). Revisiting the aforementioned definition by Gu et al. (2020), finance hence resembles a fertile soil for machine learning. Yet, financial machine learning also faces some major challenges such as datasets containing only a small number of observations, low signal-to-noise ratios and constantly evolving markets (Israel et al., 2020; Kelly and Xiu, 2023). Nonetheless, a plethora of studies continues to explore the extent to which machine learning methods prove to be successful across various topics in financial research, such as return prediction, analysing risk-return trade-offs, or portfolio optimization (Kelly and Xiu, 2023).

I took this thesis as an opportunity to contribute to that discussion. The first essay of this thesis provides a general machine learning solution to a central issue in finance, i.e., portfolio optimization. Grounded on the seminal work by Markowitz (1952), portfolio optimization has traditionally involved two separate steps: estimation of moments of the return distribution and optimization of some utility function involving these moments.

Much effort has been put into deriving the moments of the return distribution, for which machine learning naturally serves as a powerful tool (e.g., Gu et al., 2020). Nonetheless, even only considering the variances and covariances of stocks in a utility function implies an excessive amount of moments to be estimated. Moreover, return moments are generally difficult to estimate (e.g, Merton, 1980) and thus prone to errors, which leads to sub-optimal portfolio weights when optimizing the respective utility function in the second step (e.g., Michaud, 1989; Best and Grauer, 1991).

Brandt et al. (2009) introduce a framework which allows for direct optimization of portfolio weights conditional on an information set, thus circumventing the aforementioned issues. Their approach originally utilizes a simple linear model mapping a small set of conditioning variables to portfolio weights. DeMiguel et al. (2020) extend their approach by introducing a penalized linear model utilizing a larger set of conditioning variables. My co-authors and I further extend the framework by mapping the conditioning information set to portfolio weights via artificial neural networks. This approach, which we name *Deep Parametric Portfolio Policies* (DPPP), allows the relation between conditioning information and portfolio weights to be of arbitrary, possibly non-linear, functional forms.

We find that the DPPP model outperforms a linear analogue in the spirit of DeMiguel et al. (2020) by a significant margin and for all types of investors considered. Yet, the degree to which the DPPP outperforms a simpler linear specification decreases with increasing risk aversion and realistic portfolio constraints such as e.g., constraints on leverage. Moving beyond performance evaluation per se, we interpret our approach via several interpretation techniques. In line with the performance of the DPPP and its linear analogue converging with increasing risk aversion, we find that the higher the risk aversion, the less prevalent are non-linearities in our approach. Put differently, the DPPP and its linear analogue become more similar in terms of predicted portfolio weights. In line with the literature, we further find that past return-based stock characteristics are more important than accounting-based stock characteristics, and that importance becomes more equally distributed among characteristics with increasing risk aversion (e.g., DeMiguel et al., 2020; Kelly et al., 2024).

The second essay explores the potential of machine learning for forecasting earnings. Traditionally, earnings forecasts have been either derived from analysts or via simple linear statistical models. More recently, several studies have introduced various

more complex machine learning techniques (e.g., Cao and You, 2021; Hendriock, 2022; Van Binsbergen et al., 2023; Jones et al., 2023). However, these studies typically do not fully exploit the possibility of high-dimensional datasets and Chen et al. (2022) even conclude that machine learning techniques perform quite poorly in terms of predicting earnings.[1] Moreover, studies in this context provide limited model interpretation, one of the key issues pertaining to the application of machine learning methods (e.g. Israel et al., 2020). Our research addresses these gaps by introducing a machine learning framework for earnings prediction that leverages high-dimensional financial statement data. Furthermore, we offer comprehensive model interpretation, enhancing our understanding of how precisely future earnings are linked to current fundamental data.

We find that our machine learning framework, an ensemble of several commonly used methods, outperforms common traditional linear methods by a significant margin and for all forecast horizons considered. For example, for a 1-year forecast horizon, our model outperforms the best-performing traditional model by around 12% in terms of forecast accuracy. Our model also demonstrates superior explanatory power for out-of-sample variation, surpassing the out-of-sample $R^2$ values of the best-performing traditional model by approximately 14-19%, depending on the forecast horizon. This superiority in terms of predictive performance translates into more profitable gross investment returns, conditional on these earnings forecasts.[2]

Apart from providing an accurate prediction model, the key contribution of this study is our extensive model interpretation. By utilizing state-of-the-art interpretation techniques we derive the importance of variables and groups of variables for earnings prediction. We find that current income statement data, in particular current earnings, is the most important group of financial statement predictors. As the forecast horizon increases, variable importance becomes more equally distributed. More precisely, balance sheet information gains importance whereas income statement information loses importance.

In addition to providing an in-depth assessment of predictor importance, we disentangle the effects of non-linearity via surrogate modeling. We find that around 80-90% of the variation of our models' prediction can be explained by a linear surrogate model. The remaining variation in predictions cannot be explained by simple interactions among

[1] They use this as an argument for predicting the direction of earnings *changes* using machine learning, rather than predicting levels.

[2] I explore this relation more thoroughly in Chapter 4.

predictions and is hence attributable to non-linearity of the functional form of our model. Lastly, we assess how the non-linearity is expressed at a variable level. We do so by examining partial dependencies for a selection of the most important predictor variables. Our results suggest that the relationship between the variables is approximately linear for profit firms and approximately linear for loss firms. Yet, the strength of the respective input-future earnings relationship differs between profit and loss firms.

The third essay digs deeper into the relation between model-based earnings forecast accuracy and investment returns conditional on forecasted earnings. One of the primary applications of earnings forecast models is their use in implied cost of capital (ICC) estimations. In fact, the seminal study of Hou et al. (2012), which may be considered as marking the beginning of the significant research efforts that have been put into the development of earnings prediction models, explicitly motivates the development of their prediction model by its application in ICC computations. Conceptually, ICC resemble the constant discount rate which links future expected payoffs to current stock prices. The aforementioned link is provided by some equity valuation model which one needs to assume. ICC thus denote return expectations implied by the respective model, making them naturally suitable for use in an investment context. Despite this, there is limited literature that evaluates ICC in practical investment scenarios.

Literature in this context typically constructs long-short portfolios and evaluates the resulting gross returns (e.g., Hou et al., 2012; Li and Mohanram, 2014). However, two key issues pertain to the extant literature on ICC investment performance: first, transaction costs are typically not explicitly considered, despite possibly significantly altering findings. A prominent example for the effect of transaction costs is short-term reversal. The variable predicts returns comparably well, but net of transaction costs, a short-term reversal investment strategy is not profitable (e.g., Novy-Marx and Velikov, 2016; Chen and Velikov, 2023). This is due to the fact that trading conditional on short-term reversal induces excessive turnover and thus transaction costs. This exemplifies that ignoring transaction costs may lead to unrealistic expectations regarding the success of trading on some return predictor, such as ICC. Two important exceptions which consider transaction costs in the ICC context are Esterer and Schröder (2014) and Bielstein (2018), but the authors use only rudimentary transaction cost proxies and rely on analyst earnings forecasts. This study extends the existing body of literature by being the first to examine the relationship between model-based earnings forecasts and ICC portfolio

returns against the background of transaction costs.

Second, existing studies typically only focus on one dimension of accuracy, i.e., the average (absolute) deviation from the realization.[3] However, Hou et al. (2012) note that analyst earnings forecasts translate into much lower ICC portfolio returns than model-based earnings forecasts, despite being more accurate in the aforementioned sense. They attribute this to the fact that analyst forecasts are more biased, i.e., analysts systematically overestimate earnings. Bias resembles only one type of distortion which might influence portfolio returns. Other examples include predictions for large firms being more accurate than for small firms, as shown by e.g., Li and Mohanram (2014). I denote such characteristics as *systematic distortions* and propose a novel simulation-based metric which measures the degree to which a forecast model is subject to such systematic distortions. Importantly, this metric does not only capture bias, but also differences in accuracy across subsets of firms. Further, it allows me to separately assess the effects of general accuracy and systematic distortions on ICC portfolio returns.

My empirical findings are as follows: first, I show that the most accurate earnings forecast model considered, a machine learning model based on Hess et al. (2024), is also the least systematically distorted one. Second, the machine learning earnings forecast model yields statistically significant ICC portfolio returns, both gross and net of transactions. In contrast, traditional linear earnings forecast models fail to do so. Third, transaction costs reduce ICC portfolio returns significantly. This stresses the importance of considering them in any realistic investment analysis. Lastly, exploiting the novel metric introduced in the study, I show that both general accuracy and the degree of systematic distortions strongly impact ICC portfolio returns. However, transaction costs are neither related to general accuracy nor systematic distortions. I conclude that the absence of additional costs associated with improvements to earnings forecast models in this context, such as through novel machine learning methods, provides strong motivation for their further development.

Overall, the three essays explore the potentials of financial machine learning in two settings, i.e., portfolio optimization and earnings forecasts. More precisely, the first study provides a general machine learning solution to the portfolio optimization problem. The second study introduces a high-dimensional machine learning approach for earnings prediction and provides in-depth interpretation thereof. The third study builds on

[3]Note that typically, the literature scales this deviation by prices (e.g., Hou et al., 2012; Li and Mohanram, 2014). In the following, I will denote this definition of accuracy as *general accuracy*.

the second study and shows that better earnings forecast models translate into higher investment returns, even in the face of transaction costs.

To conclude the introduction, I outline my contributions to the three essays in this thesis. The initial research idea for the first essay was proposed by my co-authors. We collaboratively surveyed the literature and refined the concept through multiple discussions. Together with my co-authors, I developed the empirical model and conducted parts of the empirical analyses. The first draft of the paper was also jointly written. Lastly, we jointly revised the study multiple times based on feedback we got from numerous conferences and seminars that we (independently as well as jointly) presented the study on. The idea behind the second essay emerged from several discussions between my co-authors and me. I surveyed the relevant literature, conducted the majority of the empirical analyses and continuously updated the empirical models based on input from my co-authors. The first draft of the study was jointly written and revised several times based on feedback we received. The third study is solo-authored. I came up with the research idea, conducted the empirical analysis and wrote the paper.

# Chapter 2

# Deep Parametric Portfolio Policies*

## 2.1 Introduction

Consider the formidable problem of an investor who wants to choose an optimal asset allocation within her equity portfolio. The literature provides her with a few options: she can opt for a traditional Markowitz approach (Markowitz, 1952) which requires estimating expected returns, variances and covariances, with the number of moments to estimate increasing rapidly in the number of assets. At the other end of the spectrum, she might estimate a low-dimensional parametric portfolio policy (PPP) (Brandt et al., 2009) but a linear model might not provide sufficient flexibility. She can also consult a large literature that relates characteristics to expected returns but even studies that consider a multitude of firm-level characteristics (e.g., Gu et al., 2020) only investigate expected returns and do not speak to risk as perceived by different investors' objective functions.

We provide a general solution to the portfolio optimization challenge. In short, we combine the parametric portfolio policy approach that is well-suited to estimate portfolio weights for any utility function with the flexibility of feed-forward networks from the machine learning literature. The resulting approach that we label *Deep Parametric Portfolio Policy* (DPPP) is well-suited to accommodate flexible non-linear and interactive relationships between portfolio weights and stock characteristics, to integrate different

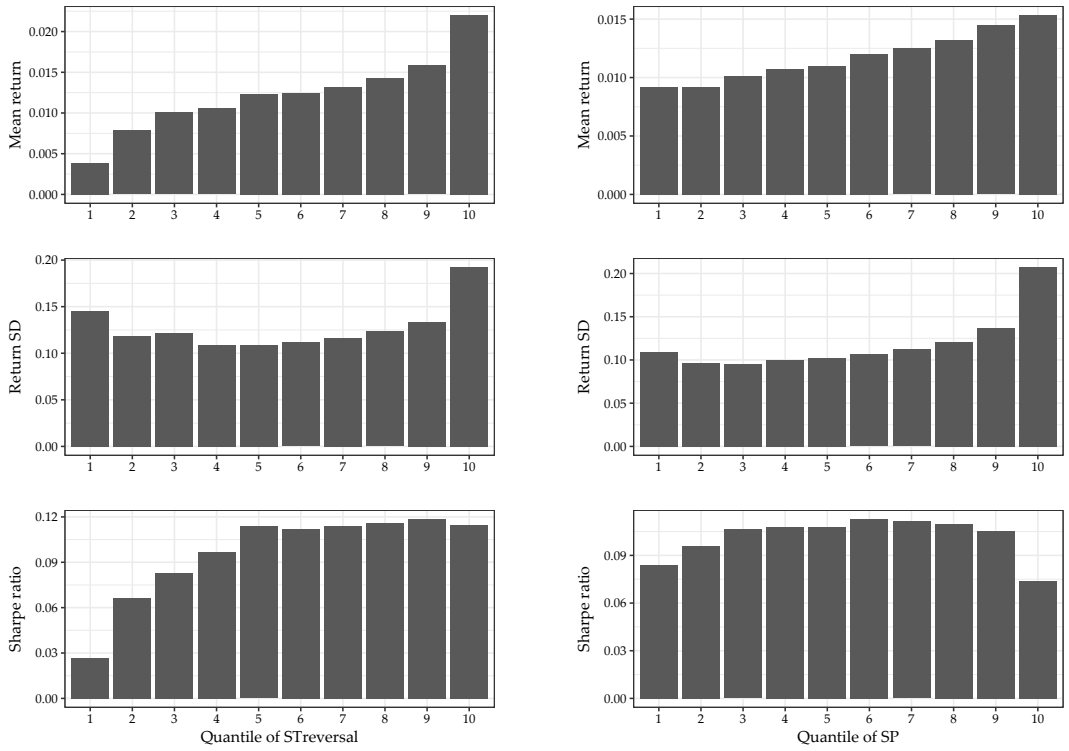
*This chapter is based on Simon et al. (2023). We thank Victor DeMiguel, Christian Fieberg, Bryan Kelly, Alexander Klos, Simon Rottke, Mark Salmon, Fabricius Somogyi (discussant), Bastidon Cécile (discussant), Heiner Beckmeyer (discussant) and seminar participants at the Research in Behavioral Finance Conference (RBFC), the Cardiff Fintech Conference, the 2022 New Zealand Finance Meeting (NZFM), the Paris Financial Management Conference (PFMC), the Theory-based Empirical Asset Pricing Research (TBEAR) Network Workshop 2023, the University of Liechtenstein, the CEQURA Conference 2023 on Advances in Financial and Insurance Risk Management, the BVI-CFR Event 2023 as well as the 4th Frontiers of Factor Investing 2024 Conference for helpful comments and suggestions.

utility functions, to deal with leverage or portfolio weight constraints, and to incorporate transaction costs.

Our results are fourfold. First, our model significantly improves over a standard linear parametric portfolio policy, with certainty equivalent gains ranging from about 75 basis points to 276 basis points, depending on the model specification and the incorporation of constraints. These gains are not limited to specific time periods, suggesting that the relationship between firm characteristics and investor utility is non-linear and complex. Second, although the DPPP consistently outperforms the linear model, the performance difference decreases with increased risk aversion or realistic portfolio constraints such as leverage or weight constraints. In particular, the benefit of model complexity decreases in an investor's risk aversion, yet remains economically significant even for highly risk-averse investors. Third, utility gains arise for a variety of investor utility functions. While our benchmark investor is a classical constant relative risk aversion (CRRA) optimizer, our setup easily accommodates other utility functions. We also investigate deep parametric portfolio policies for the case of mean-variance utility and for loss aversion, and we find substantial utility gains in all cases. Last, past return-based stock characteristics turn out to be more relevant to the portfolio policy than accounting-based characteristics. However, in line with the existing literature (DeMiguel et al., 2020; Jensen et al., 2022), the relevance of return-based characteristics decreases when we model transaction costs explicitly in the objective function.

The importance of non-linear modeling of portfolio weights becomes evident when considering an investor who trades off mean return against return volatility. The investor uses standard one-dimensional portfolio sorting techniques as pictured in Figure 2.1. Decile portfolios formed on short-term reversal or sales-to-price display monotonically increasing mean return.[1] At the same time, the standard deviations of decile portfolios are non-linear in deciles, with top and bottom decile portfolios having high standard deviations. This leads to extreme portfolios having comparatively low Sharpe ratios relative to decile portfolios in the middle of the distribution. A (long-only) investor would therefore potentially be indifferent between investing in any portfolio in the upper half of the short-term reversal distribution, and she would prefer to invest in portfolios in the middle of the sales-to-price distribution rather than investing in the extreme portfolios. Non-linear portfolio policies are able to capture these kinds of relationships.

[1] We picked these two variables for illustrative purposes as these variables are the most important return- and fundamental-based variables in Gu et al. (2020).




**(a)** Short-term reversal **(b)** Sales-to-price

**Figure 2.1: Mean returns, standard deviations and Sharpe ratios of one-dimensional portfolio sorts**

Mean returns, standard deviations and Sharpe ratios of decile portfolios sorted on short-term reversal (left panel) and sales-to-price ratio (right panel). Data is from Chen and Zimmermann (2022) and spans from 1925 to 2021.

To the best of our knowledge, our study is the first to systematically explore how the benefits of a complex and flexible model vary for investors with different levels of risk aversion or different utility functions. A natural concern with deep learning models such as ours is their potential to overfit the historical data. Overfitting leads to less reliable out-of-sample estimates and higher prediction variance. Since our deep learning approach maximizes the investor objective function directly as opposed to than minimizing a statistical objective such as the squared distance between realized and predicted returns (Moritz and Zimmermann, 2016; Gu et al., 2020), volatility of results becomes a systematic part of the optimization of the economic objective. As risk aversion increases, the variance of portfolio returns becomes more important and leans against overfitting and thus model complexity. We refer to this mechanism as "economic model regularization" (in contrast to purely statistically motivated regularization techniques), and document that, in line with the outlined mechanism, the outperformance of our model over its linear counterpart decreases with increased risk aversion (but remains

economically meaningful even for high risk aversion).

Our model can be interpreted as a generalization of the linear parametric portfolio policy approach. More specifically, we allow portfolio weights to be of one of the arguably most flexible forms - a neural network. This represents a significant conceptual departure from linear parametric portfolio policies in two ways: first, by replacing the linear specification with a neural network, we allow the relation between firm characteristics and weights to be non-linear and we allow for potential interactions of firm characteristics. Research on using machine learning methods to predict future returns shows that such flexibility is relevant to model the relationship between firm characteristics and future returns and can lead to substantial improvements over less flexible specifications (Moritz and Zimmermann, 2016; Freyberger et al., 2020; Gu et al., 2020). It is conceivable that such flexibility will also help to model the relation between portfolio weights and firm characteristics. Second, this flexibility comes at the cost of having to estimate a model with a high-dimensional parameter vector. This is a deviation from the original motivation of the parametric portfolio policy literature which aimed to reduce portfolio optimization to a low-dimensional problem with only a small number of coefficients that need to be estimated. Our benchmark model has around 5,700 parameters compared to the three parameters that must be estimated in the application of Brandt et al. (2009). However, Kelly et al. (2024) argue that model complexity is a virtue for return prediction, and our approach can be viewed as an exploration of that point in the context of parametric portfolio policies.

Building on Brandt et al. (2009), we begin with a benchmark case of a largely unrestricted portfolio policy. In the benchmark case, an investor who optimizes CRRA utility can take long and short positions with the only restriction that absolute individual stock positions cannot exceed three percent of the overall portfolio. Other aspects of the optimization remain unrestricted, in particular, the investor does not take into account transaction costs or short-selling constraints.

In the benchmark case our network-based portfolio policy improves upon the linear portfolio policy by 116 to 276 basis points in terms of monthly certainty equivalent return, depending on the degree of risk aversion. Certainty equivalent differences are larger, the lower the degree of risk aversion, consistent with the economic model regularization mechanism outlined above. This suggests that risk aversion serves as an economic regularization parameter, in the sense that it reduces model complexity, i.e., leads to

the DPPP being more similar to the (less complex) PPP. P-values for the difference in certainty equivalent between the two approaches increase with increasing risk aversion. Nonetheless, all differences are still significant at the 1% level and economically meaningful. The results further indicate that the DPPP induces twice as much monthly turnover as compared to the PPP. We show that the difference in turnover is due to the DPPP putting larger weight on past-return based characteristics which imply higher turnover, such as short-term reversal.

We then explore portfolio strategies based on networks in a more realistic setting, in which investors are subject to various restrictions. More precisely, we investigate the effects of transaction costs and leverage constraints on the optimization problem. We observe that network-based policies generate higher certainty equivalent returns than linear portfolio policies, with increases ranging from 75 to 124 basis points. The decrease in certainty equivalent differences can be attributed to the additional constraints. For constrained portfolio policies, the importance of past return-based characteristics decreases, although they remain among the most significant predictors. This is in line with the findings of DeMiguel et al. (2020), who find that more characteristics are taken into account when transaction costs are present.

Finally, we find that utility gains are not restricted to CRRA utility investors. Our approach yields similar results when considering mean-variance or loss aversion preferences. In particular, in both cases (and for various realistic risk/loss aversion parameter values) we find that a non-linear portfolio policy leads to higher utility than a standard linear policy. Benefits of model complexity decrease with risk aversion for mean-variance preferences, while benefits are more stable for loss-averse investors for different values of loss aversion.

Overall, our contribution can be summarized as providing a general solution to the parametric portfolio policy problem that combines recent advances in combining structural economic problems and machine learning methods (Farrell et al., 2021; Kelly et al., 2024). Our setup seamlessly incorporates non-linearities in parameters and across firm characteristics. We also demonstrate how constraints on leverage and transaction costs can easily be added via customization of the statistical loss function and how such constraints impact portfolios. In particular, although the DPPP consistently outperforms the linear model, we show that the benefits of a more complex model diminish as the degree of economic regularization in the form of higher risk aversion and additional

constraints on the optimization task increases.

## 2.2 Related literature

Our work relates to four different strands of the literature. First, we add to a growing literature that explores the potential of machine learning algorithms in finance (e.g., Heaton et al., 2017; Bianchi et al., 2020; Gu et al., 2020; Kelly et al., 2024). Studies in this literature typically consider a prediction task (e.g., predicting stock returns), and optimize a standard statistical loss function such as the mean squared error (or a related distance metric) between the actual and predicted values. Predicted values are used to construct portfolio weights (e.g., Gu et al., 2020). In contrast, we optimize a utility function instead of a common loss function and model portfolio weights directly as a function of firm characteristics. The use of machine learning algorithms to estimate coefficients of structural models (in our case portfolio weights) as flexible functions has also been proposed recently by Farrell et al. (2021).

Second, we extend the literature on one-step portfolio optimization. Specifically, we extend the parametric portfolio approach by Brandt et al. (2009). While Brandt et al. (2009) argue that it may be worthwhile to consider non-linear functions and interactions in weight modeling, subsequent papers that have implemented and extended parametric portfolio policies parameterize portfolio weights as a linear function of firm characteristics (e.g., Hjalmarsson and Manchev, 2012; Ammann et al., 2016). DeMiguel et al. (2020) incorporate transaction costs, a larger set of firm characteristics, and statistical regularization but also stay within the linear framework. Our deep parametric portfolio policy replaces the linear model with a feed-forward neural network that accounts for both non-linearity and possible interactions of firm characteristics. In addition, we use a larger set of firm characteristics than previous studies and explore different utility functions, constraints, and degrees of risk aversion. Alternative, (machine learning-based) one-step portfolio optimization approaches include Cong et al. (2021), Chevalier et al. (2022), Jensen et al. (2022), Butler and Kwon (2023) and Uysal et al. (2023). Each of these differs from ours in one or more aspects. Cong et al. (2021) propose a reinforcement learning-based approach (as opposed to our feed-forward framework) and connect to a related literature in computer sciences that puts additional emphasis on more technical parts of the model implementation. Our study naturally connects to the preceding

finance literature, and generalizes the approach of Brandt et al. (2009) to explicitly analyze differences between a linear and non-linear specification for different utility functions, constraints, and levels of risk aversion. Butler and Kwon (2023) show that it is possible to integrate regression-based return predictions into the portfolio optimization by means of a two-layer neural network, one layer resembling the return prediction and one layer resembling the weight optimization. However, their results are restricted to a mean-variance setting, while our approach is flexibly applicable to any type of investor preference. Moreover, our empirical analysis is about modeling portfolios of stocks based on stock characteristics, whereas they empirically assess their models on simulated data and commodity future markets. Chevalier et al. (2022) derive optimal in-sample weights based on investor preferences and subsequently predict these weights conditional on covariates. This is conceptually different from our approach, primarily because we do not require the preprocessing step of computing the optimal in-sample weights. Jensen et al. (2022) take a different approach. They specifically address the issue of integrating transaction costs into mean-variance portfolio optimization with machine learning. They assess several approaches, including a one-step ML-based approach. However, instead of extending the approach by Brandt et al. (2009) as we do, they derive a closed-form solution to the problem and implement it empirically using random feature regressions, while we stick to a feed-forward framework. Moreover, while their focus is the derivation of an efficient frontier including transaction costs, we explicitly analyze how different types of investor preferences and constraints affect the benefit of complexity in portfolio optimization.

Third, we contribute to the literature that explicitly analyzes how transaction costs and possibly other forms of constraints on the optimization impact portfolios (DeMiguel et al., 2020; Jensen et al., 2022; Detzel et al., 2023). In contrast to Jensen et al. (2022), who also assess the effect of transaction costs in a one-step optimization setting, we explicitly analyze how transaction costs and other constraints, such as the level of risk aversion, affect differences between a linear and a complex non-linear model for portfolio optimization. Moreover, they compare different approaches to derive a superior frontier with respect to transaction costs and to study variable importance in this setting. We also shed light onto how non-linearities contribute to the portfolio optimization, and how risk aversion regularizes optimization on top of and beyond the effects of transaction costs on trading behavior.

Finally, we relate to the literature that examines which firm characteristics are jointly significant in explaining expected returns (Fama and French, 2008; Green et al., 2017; Freyberger et al., 2020). While all of these studies focus on cross-sectional regression models with extensions, Gu et al. (2020) find that neural networks perform best in predicting mean returns for a large number of firm characteristics. Our portfolio approach using neural networks considers all moments of the return distribution beyond the expected return if they are relevant to an investor's utility function. Most of this literature ignores various real world constraints such as transaction costs (with Novy-Marx and Velikov (2016), DeMiguel et al. (2020) and Jensen et al. (2022) being important exceptions) or weight constraints, whereas we show how our model allows us to seamlessly integrate transaction costs or other constraints.

## 2.3 Model

### 2.3.1 Expected utility framework and Parametric Portfolio Policies

The starting point of our framework is the parametric portfolio policy model in Brandt et al. (2009). Consider a universe of $N_t$ stocks that an investor can invest in at each month $t \in T$. Each stock $i$ is associated with a vector of firm characteristics $x_{i,t}$ and a return $r_{i,t+1}$ from date $t$ to $t+1$. An investor's objective is to maximize the conditional expected utility of future portfolio returns $r_{p,t+1}$:

$$\max_{\{w_{i,t}\}_{i=1}^{N_t}} E_t\left[u(r_{p,t+1})\right] = E_t\left[u\left(\sum_{i=1}^{N_t} w_{i,t} r_{i,t+1}\right)\right], \tag{2.1}$$

where $w_{i,t}$ is the weight of stock $i$ in the portfolio at date $t$ and $u(\cdot)$ denotes the respective utility function.

Instead of directly deriving the weights $w_{i,t}$ (as e.g., following the traditional Markowitz approach), we follow Brandt et al. (2009) and parameterize the weights as a function of firm characteristics $x_{i,t}$, i.e.,

$$w_{i,t} = f(x_{i,t}; \theta), \tag{2.2}$$

where $\theta$ is the coefficient vector to be estimated.

The parameter vector $\theta$ remains constant across assets $i$ and periods $t$, i.e., it maximizes the conditional expected utility at every period $t$. This necessarily implies that $\theta$ also

maximizes the unconditional expected utility. Hence, one can estimate $\theta$ by maximizing the unconditional expected utility via the return distribution's sample analogues:

$$\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} u\left(r_{p,t+1}(\theta)\right) = \frac{1}{T} \sum_{t=1}^{T} u\left(\sum_{i=1}^{N_t} f(x_{i,t};\theta) r_{i,t+1}\right). \tag{2.3}$$

The idea behind parametric portfolio policies is that one may exploit firm characteristics in order to tilt some benchmark portfolio towards stocks that increase an investor's utility, so that $f(\cdot)$ can be expressed as

$$w_{i,t} = b_{i,t} + \frac{1}{N_t} g(\hat{x}_{i,t};\theta), \tag{2.4}$$

where $b_{i,t}$ denotes benchmark portfolio weights such as the equally weighted or value weighted portfolio and $\hat{x}_{i,t}$ denotes the characteristics of stock $i$, standardized cross-sectionally to have zero mean and unit standard deviation in each cross section $t$.[2]

Brandt et al. (2009) and the subsequent literature (e.g., DeMiguel et al., 2020) restrict firm characteristics to affect the portfolio in a linear, additive manner, such that

$$w_{i,t} = b_{i,t} + \frac{1}{N_t} \theta^T \hat{x}_{i,t}. \tag{2.5}$$

In essence, our model can be interpreted as a generalization of the linear parametric portfolio policy approach, as we allow $\hat{x}_{i,t}$ to enter the model flexibly and non-linearly. More specifically, we allow $g(\cdot)$ in Equation (2.4) to take arguably one of the most flexible forms - a feed-forward neural network. As discussed in the introduction, this represents a significant conceptual deviation from the literature in at least two respects: first, by replacing the linear specification with a neural network, we allow the relationship between firm characteristics and weights to be non-linear, and we account for potential interactions of firm characteristics, in line with the recent literature that finds that such flexibility can be important to predict returns (Moritz and Zimmermann, 2016; Freyberger et al., 2020; Gu et al., 2020). Here, our approach explores whether such flexibility also helps to model the relationship between *portfolio weights* and firm characteristics. Second, this flexibility comes at the cost of having to estimate a model with a high-dimensional

[2]The $1/N_t$ term is a normalization that allows the portfolio weight function to be applied to a time-varying number of stocks. Without this normalization, an increase in the number of stocks with an otherwise unchanged cross-sectional distribution of characteristics leads to more radical allocations, although the investment opportunities are basically unchanged.

parameter vector. Thus, it departs from the original motivation of the parametric portfolio policy literature, which aimed to reduce portfolio optimization to a low-dimensional problem where only a small number of coefficients need to be estimated. In fact, our benchmark model has about 5,700 parameters compared to the three parameters that need to be estimated when following Brandt et al. (2009).

### 2.3.2 Network architecture

We implement and compare a range of so-called feed-forward networks, a popular network structure that is commonly used in prediction contexts such as image recognition but has also recently been applied to stock return prediction. Conceptually, our feed-forward networks are structured to estimate optimal portfolio weights and as such differ from networks used in pure prediction contexts in two important ways.

First, the objective of our estimation is to maximize expected utility. Standard use of predictive modeling (with or without networks) tries to minimize some distance metric (e.g., mean squared error) between e.g., observed stock returns and predicted stock returns. For example, Gu et al. (2020) use neural networks to predict stock returns using a penalized mean squared error as the statistical loss function.

In contrast, we follow Brandt et al. (2009) and directly estimate portfolio weights. More specifically, we predict portfolio weights by maximizing the unconditional sample analogue of a utility function as given in Equation (2.3). For example, in our base case, the loss function $\mathcal{L}$ that we aim to minimize with respect to $\theta$ is the constant relative risk aversion (CRRA) utility:

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{(1+r_{p,t+1}(\theta))^{1-\gamma}}{1-\gamma}\right), \tag{2.6}$$

where $\gamma$ is the relative risk aversion parameter. Note that minimizing Equation (2.6) is equivalent to maximizing CRRA utility.

Second, our loss function requires the portfolio return per period $t$, so that we need to aggregate our outputs cross-sectionally in each period. To do so, we maintain the three-dimensional structure of our data, i.e., we do not treat it as two-dimensional as e.g., Gu et al. (2020) do. Conceptually, our models can be depicted as shown in Figure 2.2.

In Figure 2.2, the input data on the left form a cube (or 3D tensor) with dimensions time $t$, stocks $i$ and input variables $k$. Input data is fed into networks with different

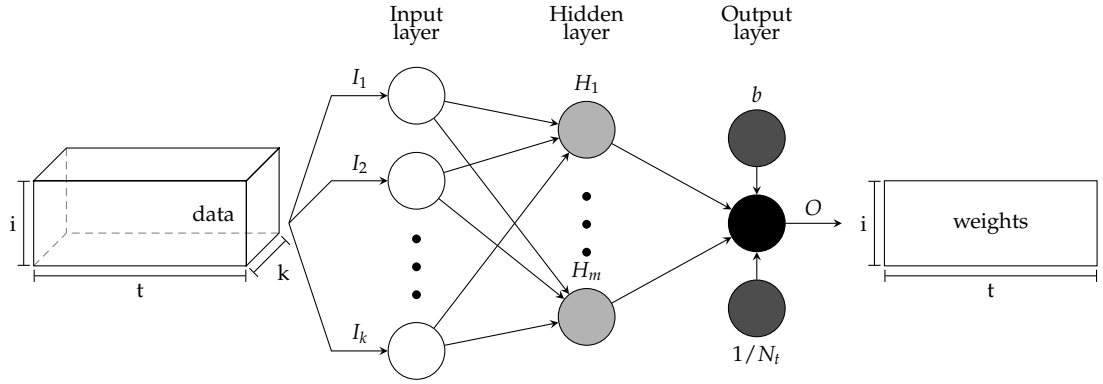



**Figure 2.2: Neural network structure**

This figure presents the core structure of our neural networks. White circles denote the input layer, grey circles denote the hidden layer and black circles denote the output layer. The data cube on the left depicts the structure of our data, i.e., we have $k$ variables across $i$ cross-sections in $t$ periods. The rectangle on the right depicts our output, i.e., weights across $i$ cross-sections in $t$ periods. The output of the neural network is normalized by $1/N_t$ and added to the benchmark portfolio $b$. The final output is labeled $O$.

numbers of hidden layers.[3] In line with Equation (2.4), the output of the neural network is then normalized by $1/N_t$ and added to the benchmark portfolio $b$. The output of the model $O$ is a two-dimensional matrix with dimensions $t \times i$ of portfolio weights for each stock and time period.

Constructing a neural network requires many design choices, including the depth (number of layers) and width (units per layer) of the model, respectively. Recent literature suggests that deeper networks can achieve higher accuracy with less width than wider models (Eldan and Shamir, 2016). However, for smaller data sets a large number of parameters can lead to overfitting and/or issues in regards to the optimization process. Selecting the best network structure is a formidable task and not our main objective.[4] Instead, we rely on the results of Gu et al. (2020) and use their most successful model as our benchmark model. We explore the robustness of our findings to changes in both network complexity and network structure in Appendix A.2.

As discussed in Section 2.3.1, the network's output needs to be normalized and can be interpreted as the deviation from a benchmark portfolio. In our application, the benchmark portfolio is the equally weighted portfolio in all models. A common alternative would be a value weighted benchmark portfolio where weights are determined by a stock's market capitalization. We stick to the equally weighted benchmark because of empirical evidence that it outperforms other benchmarks like the value weighted

[3] Following Feng et al. (2018) and Bianchi et al. (2020) we only count the number of hidden layers while excluding the output layer in the remainder of this paper.

[4] In practice, the task is often approximated by comparing a few different structures and selecting the one with the best performance.

benchmark for longer periods (DeMiguel et al., 2009).

Lastly, we control for unreasonable results and overfitting in terms of portfolio weights by ex-ante imposing an upper bound on an individual stock's absolute portfolio weight of $|3\%|$, i.e.,

$$|w_{i,t}| \leq 0.03. \tag{2.7}$$

In doing so, we ensure that the model performance does not rely too heavily on particular stocks. We employ a range of different additional regularization techniques that are standard in the deep learning literature. We give an outline of these techniques and a more detailed description of the structure of the model including its hyperparameters in Appendix A.1.

### 2.3.3 Data

We use the Open Source Asset Pricing dataset of Chen and Zimmermann (2022). The dataset contains monthly US stock-level data on 205 cross-sectional stock return predictors, covering the period from January 1925 to December 2020.

We focus on the period from January 1971 to December 2020, since comprehensive accounting data is only sparsely available in the years prior to that. In addition, we also only keep common stocks, i.e., stocks with share codes 10 and 11, and stocks that are traded on the NYSE (exchange code equal to 1) to ensure that results are not driven by small stocks. We match the data with monthly stock return data from the Center for Research in Security Prices (CRSP). We drop any observation with missing return, size and/or a return of less than -100%. We include continuous firm characteristics from Chen and Zimmermann (2022)'s categories *Price*, *Trading*, *Accounting* and *Analyst*, respectively.[5]

Finally, we follow Gu et al. (2020) and replace missing values with the cross-sectional median at each month for each stock, respectively. Additionally, similar to Gu et al. (2020) we rank all stock characteristics cross-sectionally. As in Brandt et al. (2009) and DeMiguel et al. (2020), each predictor is then standardized to have a cross-sectional mean of zero and standard deviation of one. Note that each predictor is signed so that a larger value implies a higher expected return.

[5]All characteristics are calculated at a monthly frequency. For variables that are updated at a lower frequency, the monthly value is simply the last observed value. We assume the standard lag of six months for annual accounting data availability and a lag of one quarter for quarterly accounting data availability. For IBES, we assume that earnings estimates are available by the end date of the statistical period. For other data, we follow the respective original research in regards to availability.

Our final dataset contains 157 predictors for a total of 5,154 firms. Each month, the dataset contains a minimum of 1,213, a maximum of 1,855 and an average of 1,422 firms. Table A.5 in the Appendix lists the included predictors by original paper. The three columns in the table describe the update frequency of each predictor, the predictor category and the economic category, both taken from Chen and Zimmermann (2022).

### 2.3.4 Out-of-sample testing strategy

Following Brandt et al. (2009) and Gu et al. (2020), we use an expanding window strategy to generate out-of-sample results. More specifically, we split our data into a training sample used to estimate the model, a validation sample used to tune the hyperparameters of the model and a test sample used to evaluate the out-of-sample performance of the model.

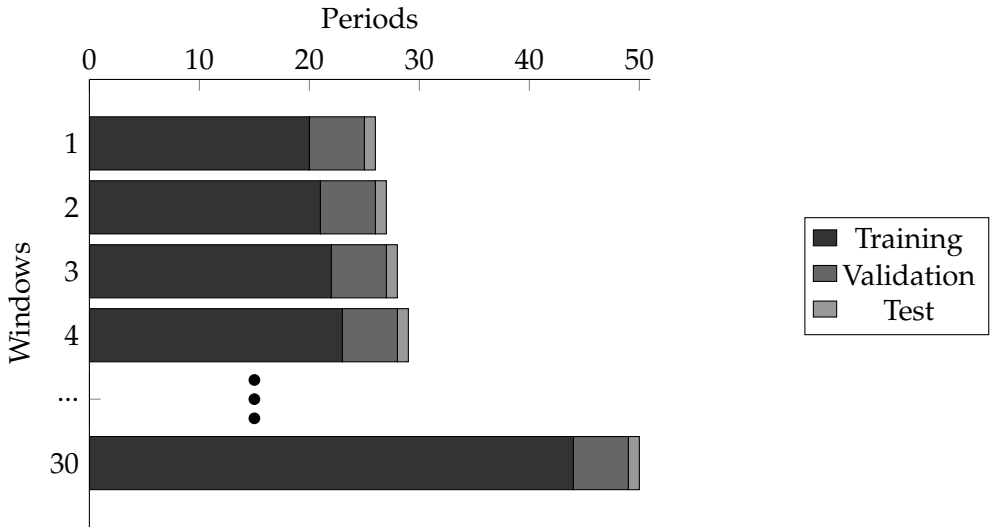



**Figure 2.3: Out-of-sample testing strategy**
This figure presents our out-of-sample testing strategy. We recursively increase our training window, presented by the black portion of each bar, while holding the validation and the test window constant, presented by the grey portions of each bar.

We initially train the model on the first 20 years of the dataset, validate it on the following five years and evaluate its out of-sample-performance on the 12 months following the validation window. We then recursively increase the training sample by one year. Each time the training sample is increased, we refit the entire model while holding the size of the validation and test window fixed. The result is a sequence of out-of-sample periods corresponding to each expanding window, in our case 25 in total. This corresponds to a total out-of-sample period of 300 months. Note that this approach

ensures that the temporal ordering of the data is maintained. The testing strategy is depicted graphically in Figure 2.3.

### 2.3.5 Model interpretation

Machine learning models are notoriously difficult to interpret and neural networks are no exception. Nevertheless, in our application, understanding the estimated relation between input (firm characteristics) and output (estimated portfolio weights) is essential in order to shed light on the relation between firm characteristics and utility. Moreover, such an understanding allows us to compare our results to the existing literature. We provide three ways of interpreting the models and of identifying the most important predictors among the plethora of variables that enter our models.

First, we calculate variable importance in the model as the decrease in model performance when a particular variable is missing from the model, as conceptually introduced by Breiman (2001). That is, for every period, we set all values of the variable of interest to zero while holding the remaining variables fixed. We then calculate the utility loss as compared to the original model in every out-of-sample period and take the average across all models. For the sake of comparability, we scale the average utility losses across all variables for each model so that they add up to one. As a result, we are able to rank the variables according to the average utility loss that occurs if they are excluded from the model.

Second, we evaluate the sensitivity of the model output to each variable. Typically, partial dependence plots provide an assessment of the variables of interest over a range of values. At each value of the variable, the model is evaluated while the remaining variables remain unchanged, and the results are then averaged across the cross-section. However, since the sum of all weights in each cross-section is equal to one and thus the mean weight prediction is always the same, applying this method to parametric portfolio policies does not yield reasonable results. To circumvent this problem, we apply our own algorithm: when assessing the sensitivity with respect to variable $k$, we set the values of the remaining variables to zero, i.e., their median. This means that effectively, we reduce our input data to the variable of interest. We then predict out-of-sample portfolio weights based on the estimated model and the manipulated data. Subsequently, we plot the weights as a function of input variable $k$. We interpret the behavior of predicted weights conditional on values of $k$ as the marginal sensitivity of weights (i.e., its partial

dependence) with respect to $k$.

Third, we evaluate the extent to which non-linearity contributes to the estimated DPPP. Put differently, we assess the extent to which different forms of non-linearity play a role when optimizing portfolios conditional on firm characteristics. To do so, we estimate a linear surrogate model in which we regress the out-of-sample weight predictions from the DPPP on all firm characteristics. This allows us to assess the extent to which a simple linear model is capable of ex-post explaining the predicted weights. In a next step, we estimate a second surrogate model, this time including all possible two-way interactions, i.e., allowing for non-linearity in variables. This allows us to assess to which extent non-linearity in variables plays a role in regards to predicting weights. We attribute the remaining unexplained portion of predicted DPPP weights to the effect of non-linearity in functional form.[6]

## 2.4 Results

### 2.4.1 Benchmark case

Table 2.1 reports the empirical results in our benchmark setting, i.e., for a CRRA-maximizing investor and not accounting for transaction costs or leverage constraints in the optimization task.[7] We compare our DPPP with its linear counterpart for different degrees of relative risk aversion.[8] Analogous to Brandt et al. (2009), we provide results as follows: we report (1) the monthly certainty equivalent return of the utility generated by each portfolio strategy, (2) the distributional properties of the monthly portfolio weights, (3) the distributional properties of the monthly portfolio returns, and (4) the monthly alphas of the strategies against a Fama-French six-factor model.

Our main finding is that for each level of risk aversion, the DPPP outperforms the PPP. The guaranteed monthly return across out-of-sample periods that an investor would require to achieve the same expected utility as the respective portfolio policy, i.e., the certainty equivalent, is higher for the DPPP than for the PPP for every level of risk aver-

[6] In addition, we report the portfolio characteristics of the ex-post fitted surrogate models during the out-of-sample periods in Table A.6 in the Appendix. Inter alia, this enables us to assess to which extent non-linearity with respect to weight predictions translates into utility differences.

[7] Results also hold compared against an equally-weighted and a value-weighted portfolio benchmark, are robust to changing the network architecture, and to the use of a long-only constraint, see Appendix A.2.

[8] To ensure comparability between the linear and the deep parametric portfolio policy we differ slightly from Brandt et al. (2009) in that the linear model includes $l_1$-regularization and early stopping, similar to the deep model. A more detailed description is given in Appendix A.1.

**Table 2.1: (D)PPP for CRRA investors**

| | $\gamma = 2$ | | $\gamma = 5$ | | $\gamma = 10$ | | $\gamma = 20$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0393 | 0.0669 | 0.0263 | 0.0492 | 0.0063 | 0.0303 | −0.0019 | 0.0097 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0001 | | 0.0002 | | 0.0079 | | 0.0065 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.5336 | 0.6897 | 0.4972 | 0.6127 | 0.3834 | 0.5211 | 0.2292 | 0.3276 |
| $max\ w_i * 100$ | 2.1781 | 1.8474 | 2.0363 | 1.7452 | 1.5531 | 1.5929 | 0.9199 | 1.1676 |
| $min\ w_i * 100$ | −2.3296 | −1.8995 | −2.1712 | −1.8709 | −1.6581 | −1.7950 | −0.9302 | −1.2923 |
| $\sum_i w_i I(w_i < 0)$ | −3.3467 | −4.4722 | −3.0841 | −3.9171 | −2.2642 | −3.2565 | −1.1521 | −1.8617 |
| $\sum_i I(w_i < 0)/N_t$ | 0.4401 | 0.4473 | 0.4351 | 0.4430 | 0.4084 | 0.4317 | 0.3672 | 0.4016 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 3.8045 | 8.7876 | 3.7816 | 7.8053 | 2.8497 | 6.5992 | 1.6268 | 4.0840 |
| Mean | 0.0489 | 0.0797 | 0.0473 | 0.0711 | 0.0368 | 0.0622 | 0.0212 | 0.0402 |
| StdDev | 0.0982 | 0.1234 | 0.0890 | 0.0982 | 0.0705 | 0.0816 | 0.0437 | 0.0548 |
| Skew | −0.1001 | 1.8314 | −0.1004 | 0.8169 | −0.1539 | 0.4023 | −0.3209 | 0.3712 |
| Kurt | 1.2734 | 14.0481 | 1.3766 | 4.9609 | 2.0482 | 1.6333 | 1.3888 | 1.8887 |
| SR | 1.7233 | 2.2382 | 1.8391 | 2.5101 | 1.8097 | 2.6422 | 1.6789 | 2.5395 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.0363 | | 0.0077 | | 0.0013 | | 0.0004 |
| $FF5 + Mom\ \alpha$ | 0.0331 | 0.0648 | 0.0324 | 0.0570 | 0.0244 | 0.0490 | 0.0116 | 0.0303 |
| $StdErr(\alpha)$ | 0.0043 | 0.0065 | 0.0040 | 0.0052 | 0.0032 | 0.0043 | 0.0019 | 0.0028 |

This table shows out-of-sample estimates of the (deep) portfolio policies optimized for a CRRA investor with relative risk aversion of 2, 5, 10 and 20, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$\gamma = 2$", "$\gamma = 5$", "$\gamma = 10$" and "$\gamma = 20$" correspond to the respective risk aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

sion considered. For example, if we set the risk aversion parameter to two, the certainty equivalent associated with the DPPP is 276 basis points higher than that of the PPP (0.0669 vs 0.0393). The differences are statistically and economically significant in every case.[9] This shows that using a more complex model that accounts for predictor interactions and non-linearities leads to significant utility gains for investors.

However, the outperformance of the DPPP compared to the PPP decreases with increasing risk aversion from around 276 ($\gamma = 2$) to 116 ($\gamma = 20$) basis points. Put differently, as risk aversion increases, the benefit of model complexity decreases. We attribute this to the fact that as risk aversion increases, the model's prediction variance is penalized to a stronger extent. In a sense, risk aversion serves as an economic regularization parameter that empirically has an effect comparable to statistical regularization methods, i.e., a reduction in model complexity, by penalizing the variance of outcomes. We provide empirical evidence for this claim in 2.4.3 when estimating partial dependence and surrogate models.

Table 2.1 provides further insight into the average distributional characteristics of portfolio weights. Regardless of the degree of risk aversion we assume, the average absolute DPPP weights are larger than the PPP weights, e.g. 0.69% versus 0.53% for $\gamma = 2$. However, for investors with a risk aversion of $\gamma = 2$ and $\gamma = 5$, the absolute maximum and minimum portfolio weights are lower in the DPPP case, while the opposite is true for investors with a higher degree of risk aversion. Nevertheless, both PPP and DPPP portfolio weights become more moderate as risk aversion increases. Consistent with this finding, portfolio leverage decreases with increasing risk aversion for both the DPPP (447% for $\gamma = 2$ to 186% for $\gamma = 20$) and the PPP (335% for $\gamma = 2$ to 115% for $\gamma = 20$). However, regardless of investor risk aversion, the DPPP approach results in higher leverage than the PPP approach. Since we do not impose any constraints on leverage or transaction costs in our benchmark setting, short-selling and portfolio turnover are unrealistically high.[10] Moreover, the average monthly turnover of the DPPP is consistently more than twice as large as that of the PPP. However, it decreases with increasing risk aversion in both cases. More precisely, the average monthly turnover of

[9] We follow DeMiguel et al. (2022) and construct one-sided p-values from 10,000 bootstrap samples using the stationary bootstrap method of Politis and Romano (1994) with an average block size of five and the procedure of Ledoit and Wolf (2008). This method is also used when assessing the statistical significance of utility and Sharpe ratio differences between the deep and the linear parametric portfolio policy hereafter.

[10] Turnover is defined as $\sum |w_{i,t} - w^+_{i,t-1}|$, where $w^+_{i,t-1}$ is the portfolio before rebalancing at time $t$, i.e. $w^+_{i,t-1} = w_{i,t-1} * (1 + r_{i,t})$.

the DPPP (PPP) ranges from 879% (380%) for the least risk-averse investor to 408% (163%) for the most risk-averse investor. We address this in section 2.4.2 by including a penalty term for transaction costs and a constraint on leverage in our objective function.

Turning to the distribution of out-of-sample portfolio returns, we find that the DPPP yields 308 to 190 basis points higher average returns than the PPP, depending on the degree of risk aversion. This comes at the cost of 10-25% higher return volatility than the PPP. These results translate into annualized Sharpe ratios of the DPPP that are 30-50% higher than the annualized Sharpe ratios of the PPP, depending on the level of risk aversion. Regardless of the level of risk aversion, the difference in the annualized Sharpe ratio is significant at the 5% level. The distribution of DPPP returns is positively skewed, while the distribution of PPP returns is negatively skewed. Thus, the DPPP has positive tails while the PPP has negative tails. As the kurtosis indicates, the distribution of DPPP returns has much fatter tails than that of PPP returns for risk aversions of $\gamma = 2$ (14.0481 versus 1.2734) and $\gamma = 5$ (4.9609 versus 1.3766). For higher degrees of risk aversion, the kurtosis of both portfolio return distributions remains at a platykurtic level below three with thin tails.

The bottom set of rows reports the alphas and their standard errors with respect to a six-factor model that adds a momentum factor to the Fama-French five-factor model. Both the DPPP and PPP alphas are highly significant for each level of risk aversion considered. However, the alphas of the DPPP are significantly larger than those of the PPP. These large unexplained returns can be partially attributed to the highly levered nature of the active portfolios. Thus, the alphas of both portfolios consistently decrease with increasing risk aversion and hence decreasing leverage. More specifically, the PPP alpha decreases from 3.3% ($\gamma = 2$) to 1.2% ($\gamma = 20$), and the DPPP alpha decreases from 6.5% ($\gamma = 2$) to 3% ($\gamma = 20$).

### 2.4.2 Transaction costs and leverage constraint

In the unconstrained benchmark setting average turnover and leverage are unreasonably high, both for the PPP and the DPPP. We next compare both approaches in a more realistic scenario that explicitly accounts for transaction costs and sets a maximum leverage constraint in the optimization task.

To account for transaction costs, we follow DeMiguel et al. (2020) and add the

following penalty term to the optimization problem:

$$TC = E_t[\sum_{i=1}^{N_t} |\kappa_{i,t}(w_{i,t} - w^+_{i,t-1})|], \tag{2.8}$$

where $w^+_{i,t-1}$ is the portfolio weight before rebalancing and $\kappa_{i,t}$ are transaction costs for stock $i$ at time $t$. Our transaction cost estimates come from Chen and Velikov (2023).[11] Thus, we define transaction costs $\kappa_{i,t}$ as the effective half bid-ask spread.

The leverage constraint is constructed analogously to our weight constraint in Equation (2.7). The penalty is constructed such that the gross leverage cannot exceed 100% in a single period in model training.[12] This constraint is formulated for every period $t$ as

$$\sum_{i=1}^{N_t} w_i I(w_i < 0) \geq -1, \tag{2.9}$$

where $I(w_i < 0)$ is a vector in which an element is one if the corresponding portfolio weight is smaller than zero and zero otherwise.

Table 2.2 shows the results of the constrained optimization process for CRRA investors with different degrees of risk aversion. Even when imposing realistic constraints, the DPPP outperforms the PPP, regardless of the level of risk aversion. The difference in monthly certainty equivalent between the two approaches is reduced to 75 to 124 basis points, depending on the degree of risk aversion. This suggests that similar to the risk aversion parameter, the transaction cost penalty and the maximum leverage constraint can be seen as additional economical regularization terms, which lead to a decrease in model complexity.[13] We provide empirical evidence for this claim in 2.4.3 when estimating partial dependencies and surrogate models. The p-values of the differences in monthly certainty equivalent increase as risk aversion increases, and for $\gamma = 20$, the difference is no longer significant at the 1% level. This is consistent with increased risk aversion leaning against model complexity and serving as an economically motivated regularization parameter as discussed above. The constraints lead to more realistic portfolios: leverage is below 100% for all portfolios and turnover is reduced significantly to 47-54% for the PPP and 111-171% for the DPPP, depending on the degree of risk

[11] We thank the authors for making an updated version of the data available.

[12] Ang et al. (2011) show that the average gross leverage of hedge fund companies amounts to 120% in the period after the financial crisis 2007-2008. We use a slightly more conservative number of a maximum leverage of 100%.

[13] Note that we report the certainty equivalent for the expected utility net of transaction costs and hence a decrease of the respective certainty equivalent trivially follows to some extent.

**Table 2.2: (D)PPP for CRRA investors incl. transaction costs and leverage constraint**

| | $\gamma = 2$ | | $\gamma = 5$ | | $\gamma = 10$ | | $\gamma = 20$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0114 | 0.0206 | 0.0084 | 0.0159 | 0.0020 | 0.0107 | −0.0125 | −0.0001 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0001 | | 0.0007 | | 0.0018 | | 0.0178 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.1238 | 0.1809 | 0.1288 | 0.1836 | 0.1195 | 0.1813 | 0.1199 | 0.1764 |
| $max\ w_i * 100$ | 0.4423 | 0.7863 | 0.4595 | 0.7337 | 0.3948 | 0.7373 | 0.4010 | 0.7527 |
| $min\ w_i * 100$ | −0.4000 | −1.0246 | −0.4337 | −1.0098 | −0.3671 | −0.9559 | −0.3538 | −0.8031 |
| $\sum_i w_i I(w_i < 0)$ | −0.3924 | −0.8042 | −0.4288 | −0.8234 | −0.3614 | −0.8072 | −0.3642 | −0.7717 |
| $\sum_i I(w_i < 0)/N_t$ | 0.2279 | 0.3242 | 0.2453 | 0.3160 | 0.1974 | 0.3202 | 0.2092 | 0.3446 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 0.5201 | 1.7149 | 0.5431 | 1.5699 | 0.4701 | 1.3921 | 0.4989 | 1.1146 |
| Mean | 0.0139 | 0.0232 | 0.0133 | 0.0214 | 0.0121 | 0.0200 | 0.0112 | 0.0174 |
| StdDev | 0.0489 | 0.0502 | 0.0424 | 0.0447 | 0.0412 | 0.0402 | 0.0392 | 0.0364 |
| Skew | −0.6865 | −0.4891 | −0.9352 | −0.7242 | −0.8990 | −0.6081 | −0.9919 | −0.7242 |
| Kurt | 3.0761 | 3.0184 | 2.5399 | 2.3413 | 2.1149 | 1.7382 | 2.5912 | 1.8450 |
| SR | 0.9825 | 1.6009 | 1.0860 | 1.6609 | 1.0208 | 1.7235 | 0.9871 | 1.6537 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.0007 | | 0.0032 | | 0.0029 | | 0.0051 |
| $FF5 + Mom\ \alpha$ | 0.0032 | 0.0116 | 0.0030 | 0.0109 | 0.0034 | 0.0101 | 0.0031 | 0.0084 |
| $StdErr(\alpha)$ | 0.0011 | 0.0017 | 0.0012 | 0.0016 | 0.0013 | 0.0015 | 0.0013 | 0.0015 |

This table shows out-of-sample estimates of the (deep) portfolio policies with the transaction costs penalty (Equation (2.8)) and leverage constraint (Equation (2.9)) optimized for a CRRA investor with relative risk aversion of 2, 5, 10 and 20, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$\gamma = 2$", "$\gamma = 5$", "$\gamma = 10$" and "$\gamma = 20$" correspond to the respective risk aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions net of transaction costs as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

aversion. Despite its larger turnover, the DPPP yields notably larger returns net of transaction costs, with similar standard deviations of portfolio returns and significantly higher Sharpe ratios. The maximum and minimum positions of both approaches are less extreme than in the unconstrained case and thus also more realistic. The alphas of the estimated models are much smaller than in the benchmark scenario, but still highly significant.

The main results in Table 2.1 and Table 2.2 are visually summarized in Figure 2.4, which shows the cumulative performance of portfolio returns over time for both the PPP and the DPPP, all degrees of risk aversion, and with and without transaction cost and leverage constraints. The figure shows that the DPPP consistently outperforms the PPP by a substantial margin in all specifications. Figure 2.4 also reveals some important additional insights. Specifically, the benchmark portfolios are more robust than their traditional counterparts during the dot-com bubble in 2000, the global financial crisis in 2008, and the COVID-19 stock market crash in 2020. Further, we observe that the returns of the higher risk aversion portfolios are more robust during these periods.

### 2.4.3 Variable importance, partial dependence and surrogate models

In this section, we analyze the estimated models with the tools discussed in section 2.3.5.

**Variable importance**

In Figure 2.5 we compare the most important clusters of variables (such as "earnings-related", or "risk-related") according to the economic category specified in the Open Source Asset Pricing data set by Chen and Zimmermann (2022).[14] The figure displays the nine most important clusters and subsumes all other clusters under "other" for the benchmark and constrained case and across all degrees of risk aversion, respectively.[15] The size of the area corresponds to the relative importance of the cluster within that specific model. We report the results for the DPPP and PPP model, respectively.

For the DPPP, we find that in both the unconstrained and the constrained setting, the majority of the most important predictors are related to past returns. Short-term reversal is the most important single variable in both models, mirroring the findings in Moritz and Zimmermann (2016) and Gu et al. (2020), while the momentum cluster is

[14]Table A.5 in the Appendix shows the economic category of each anomaly variable, based on Chen and Zimmermann (2022).

[15]The clusters are ranked according to the importance in the DPPP benchmark model.

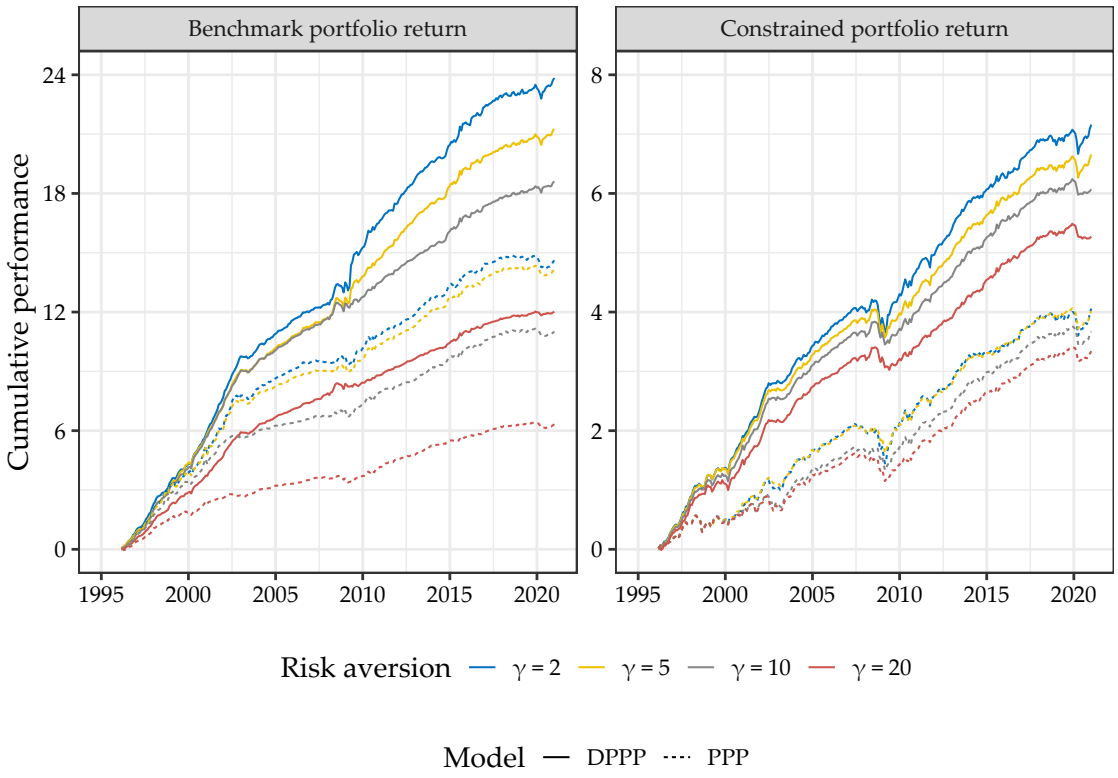



**Figure 2.4: Cumulative performance over time for CRRA investors**
The left panel shows the cumulative sum of portfolio returns for the benchmark, i.e., unconstrained, DPPP and PPP. The right panel shows the cumulative sum of portfolio returns net of trading costs for the transaction cost and leverage constrained DPPP and PPP. We show the results for each of the degrees of relative risk aversion considered and across all out-of-sample periods.

more important overall.[16]

In the unconstrained benchmark case, we find that about 75% of the total importance is associated with the top nine clusters. We also find that momentum and short-term reversal account for ∼40% of the importance, which is consistent across different degrees of risk aversion. Overall, we do not find large differences across different degrees of risk aversion in terms of cluster importance per model.

Turning to the DPPP in the constrained setting, the figure shows that the importance of short-term reversal is much lower than in the unconstrained benchmark case. This is an intuitive result, since trading conditional on short-term reversal implies high turnover. Thus, if turnover is penalized by introducing transaction costs, short-term reversal inevitably loses some of its importance, consistent with DeMiguel et al. (2020) and Jensen et al. (2022). Interestingly, other characteristics based on past returns, such as the momentum cluster, do not lose importance when constraints are included. The other

[16]Note that the short-term reversal cluster consists of the short-term reversal characteristic only.

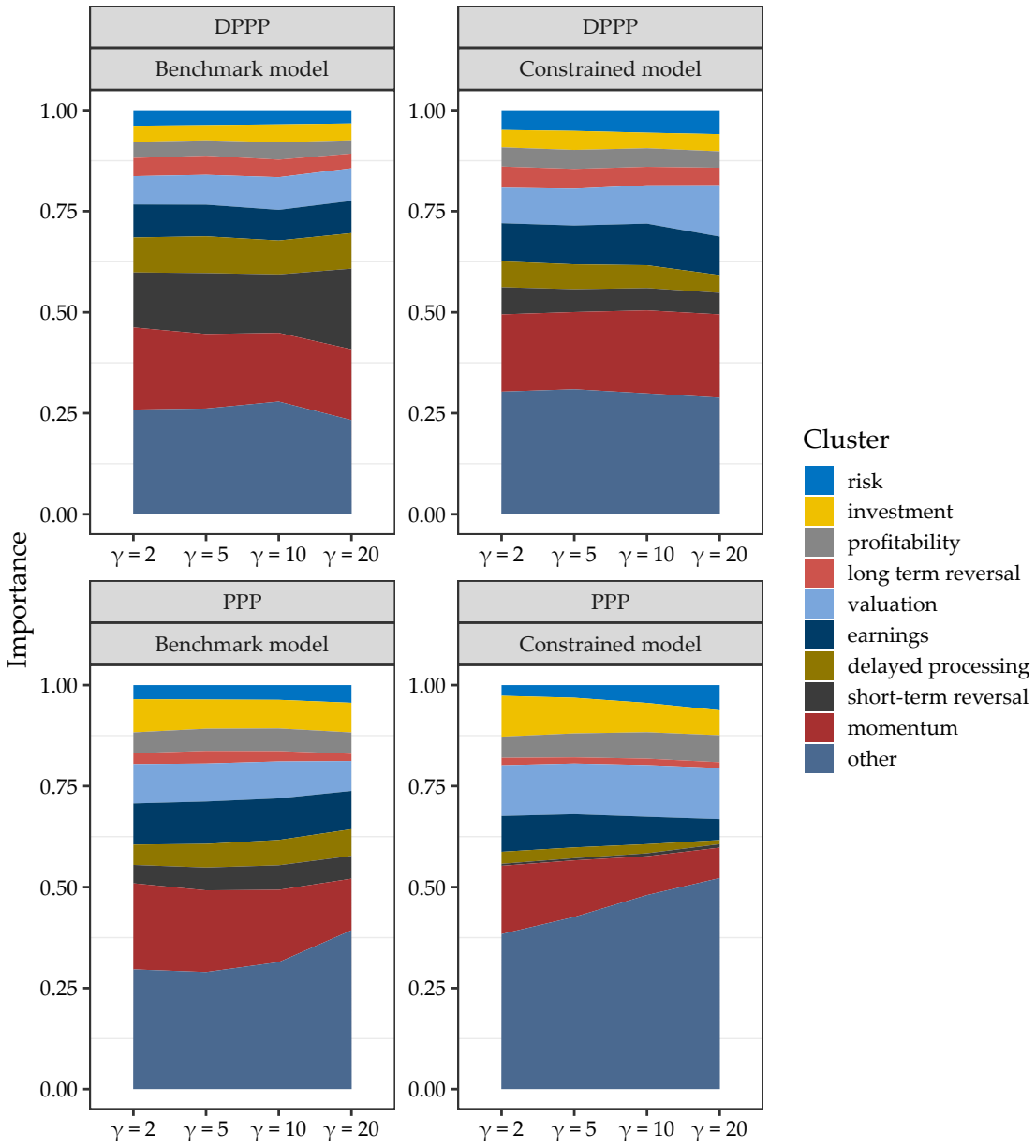



**Figure 2.5: (D)PPP variable importance per cluster without and including transaction costs and leverage constraint**

We group the variables into clusters according to the economic category specified in the Open Source Asset Pricing data set by Chen and Zimmermann (2022). Clusters are then ranked by sum of characteristic importance within the respective cluster. We display the top nine clusters and subsume all other clusters within "other". We plot the top clusters in terms of its importance across all benchmark and constrained DPPP and PPP models for different degrees of risk aversion, respectively. The filled area of a cluster corresponds to its importance.

clusters also remain similarly important in the constrained model. Again, we do not find large differences across different degrees of risk aversion in terms of cluster importance per model.

Next, we turn to the linear PPP. Again, in both settings, we find that the majority

of the most important predictors is related to past returns. Short-term reversal is the most important cluster in the unconstrained models, but it becomes the least important one when constraints are imposed. This is in contrast to the results of the non-linear DPPP, for which the short-term reversal cluster still bears notable importance in the constrained setting. We also observe that the importance of the momentum variables decreases with increasing risk aversion in both settings, albeit stronger in the constrained one. Moreover, in the constrained setting, the importance of valuation-related variables increases significantly. This is consistent with valuation-based information being less volatile than past-return based information.

Finally, Figure 2.6 shows the 40 most important individual characteristics for the deep and linear models for the benchmark and constrained cases and across all levels of risk aversion. In line with our results above, the majority of the most important predictors are related to past returns, with short-term reversal being the most important variable for both models, and more prominently so in the DPPP case. As past-return based variables typically imply higher turnover, this is consistent with the higher turnover of the DPPP as compared to the linear PPP reported above. Moreover, consistent with the results of DeMiguel et al. (2020), we find that the importance of the variables is generally much more balanced across variables for the constrained models. Table 2.2 shows that the constraints lead to a more diversified portfolio, partially reflected by the more evenly distributed importance of firm characteristics.

**Partial dependence**

Figure 2.7 depicts the marginal association between DPPP portfolio weights and input variables for the benchmark setting, the constrained setting and across different risk aversions, respectively. We examine the sensitivity with respect to three fundamental variables, namely the book-to-market ratio (BM), liquid assets (cash), and quarterly return on assets (roaq), as well as an analyst variable, namely earnings forecast revisions per share (AnalystRevision), and four past return-based variables, namely 12-month momentum (Mom12m), short-term reversal (STreversal), seasonal momentum (MomSeason), and intermediate momentum (IntMom). Recall that each predictor is signed, so that a larger value implies a higher expected return. To assess whether the marginal association of the deep model is more in line with the actual risk and return associated with each characteristic than a linear model, we include the overall Sharpe ratio for each decile portfolio sorted on each of the characteristics.

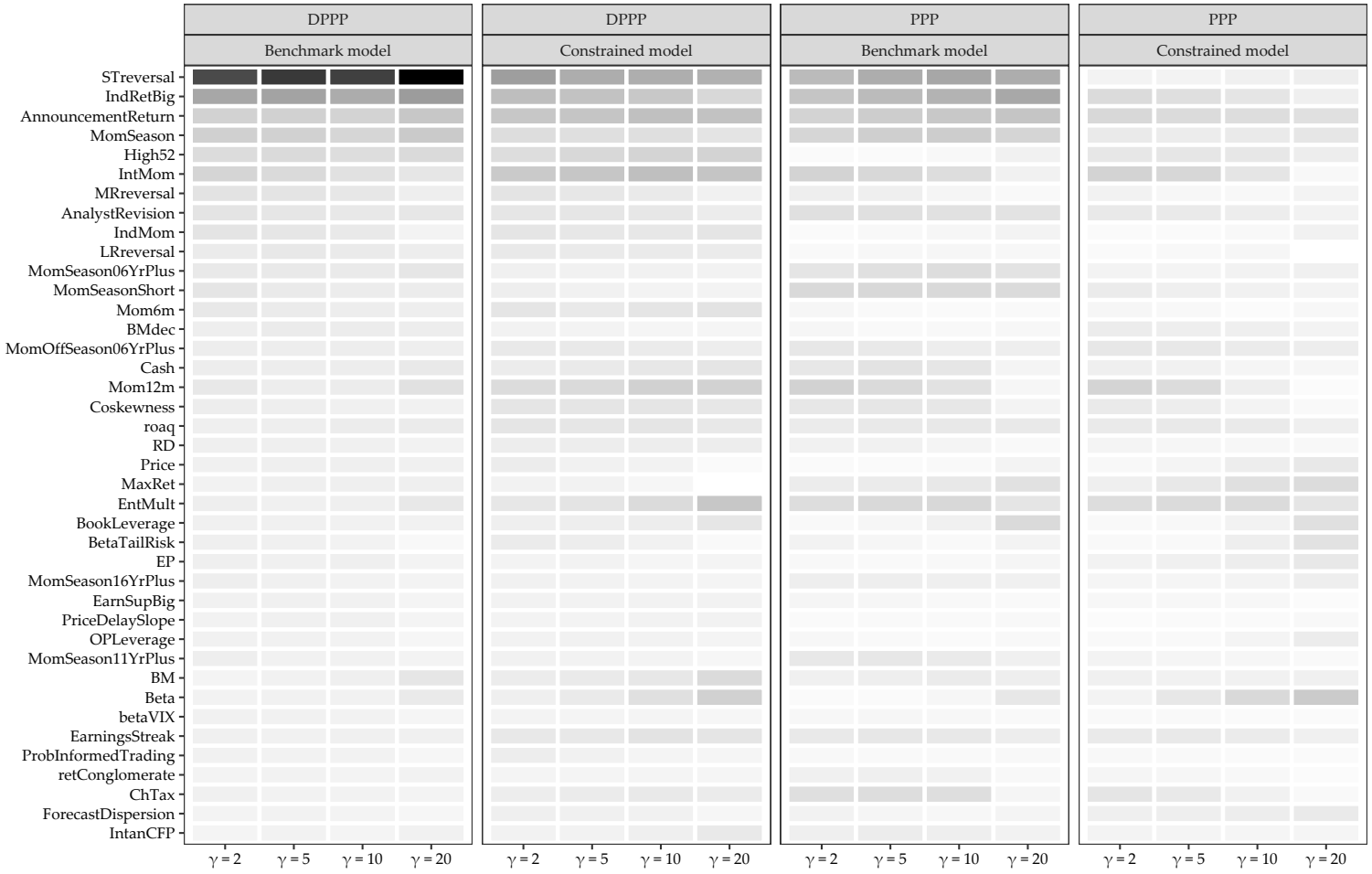



**Figure 2.6: (D)PPP variable importance without and including transaction costs and leverage constraint**

Variable importance for the 40 most influential variables in the PPP and the DPPP across model specifications and risk aversions, respectively. Variable importance is computed as the average importance over all training samples and normalized to sum to one within each model. The darker the color gradient, the higher the respective importance.

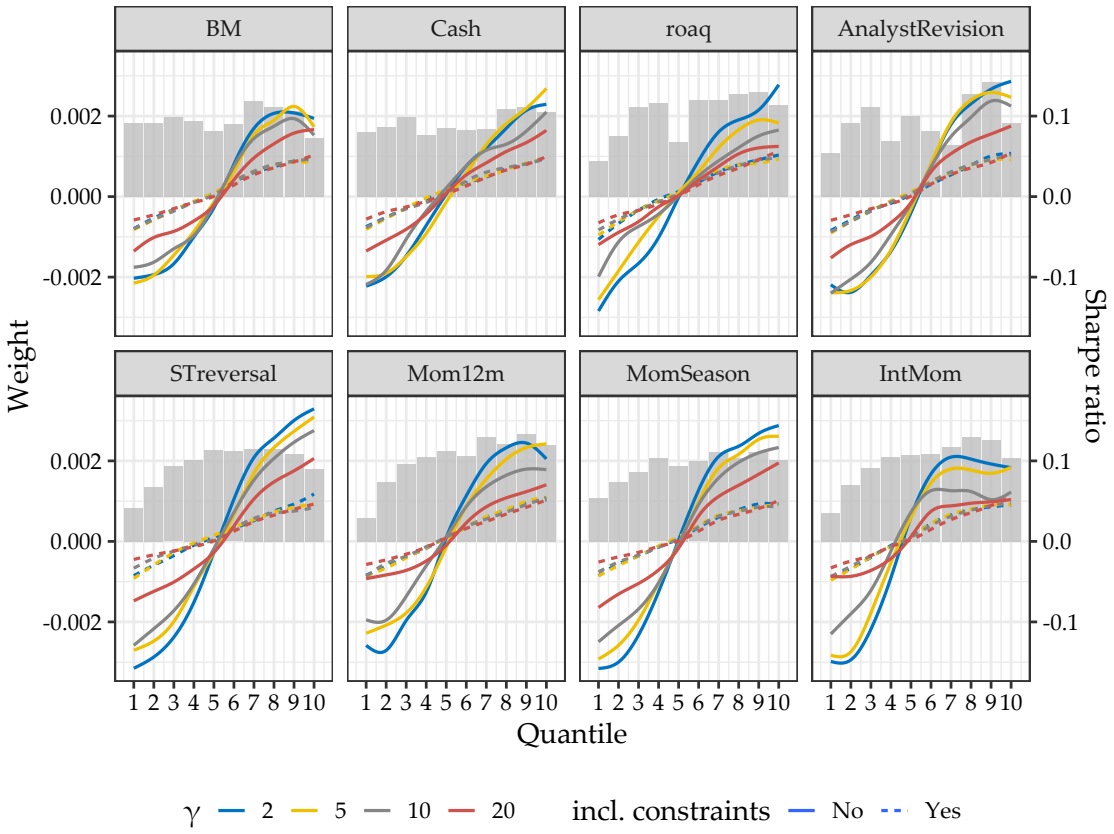



**Figure 2.7: Marginal associations of DPPP weights and characteristics without and including transaction costs and leverage constraint**

This figure shows the sensitivity of predicted weights (left vertical axis) with respect to values of the respective variable (horizontal axis) for both the unconstrained and the constrained DPPP models for different risk aversions ($\gamma$). The aforementioned relationship is depicted by curves, smoothed via spline-regressions. The figure also includes bars, depicting the Sharpe ratio (right vertical axis), per variable decile (horizontal axis).

In the unconstrained benchmark case, the DPPP weights are mostly non-linearly related to the characteristics. This is in line with the fact that Sharpe ratios are generally not linearly increasing in characteristic deciles, as this is indicative for the fact that utility is not linearly increasing in characteristic deciles. The DPPP captures these patterns. For example, weights associated with earnings forecast revisions per share (AnalystRevision) and intermediate momentum (IntMom), as well as the book-to-market ratio (BM), decrease in higher deciles as the Sharpe ratio decreases. We find a similar but less pronounced pattern for the other characteristics as well. Turning to differences across different degrees of risk aversion in the benchmark setting, we find that the degree of non-linearity in the marginal association between portfolio weights and characteristics decreases as risk aversion increases. This confirms the reasoning that increasing risk aversion leads to a decrease in model complexity. In line with the findings in regards

to importance, short-term reversal exhibits the most pronounced marginal effect, as indicated by the steepness of the depicted relationship.

When introducing transaction costs and a leverage constraint to the setting, the marginal relationships turn mostly linear. Again, this confirms the reasoning that additional constraints serve as regularization parameters which reduce model complexity, similar to increasing the degree of risk aversion. Notably, differences in the marginal relationships across different degrees of risk aversion are less pronounced in the constrained case. Consistent with the findings on importance, the differences in marginal association are less pronounced across characteristics. This serves as further evidence that more characteristics matter under transaction costs as also shown by DeMiguel et al. (2020).

In summary, these results confirm that imposing constraints and increasing risk aversion lead to a convergence of the linear PPP and the more complex DPPP. We dive deeper into this in the next step, in which we estimate surrogate models to more thoroughly disentangle the degrees to which (non-)linearity plays a role in the different settings.

**Surrogate model**

Surrogate modeling allows us to disentangle the contributions of non-linearity with respect to the predictions as well as the utility gains of the deep parametric portfolio policy as compared to the linear parametric portfolio policy. Figure 2.8 shows the adjusted $R^2$s of a linear surrogate model for the out-of-sample predicted weights of the DPPP in the different settings on the 50 most important characteristics in each model, respectively. The surrogate model with interactions is an extension to the aforementioned surrogate model which additionally includes all possible two-way interactions between the characteristics included. In line with the previous findings, the results highlight that the importance of non-linearity is less prevalent for higher degrees of risk aversion. More specifically, the simple linear surrogate model explains about 60-80% of the variation in predicted portfolio weights for $\gamma = 20$, and between 50-70% for the other degrees of risk aversion. This underscores that risk aversion acts as an economic regularization parameter, in that it reduces model complexity. Adding interactions has two effects in particular. First, the range in-between which the $R^2$ fluctuates becomes smaller, i.e., we observe less fluctuation across the periods. More importantly, however, we observe an increase of the $R^2$ of about $\sim$10% across all degrees of risk aversion.

Since performance of the linear PPP and the non-linear DPPP converges when

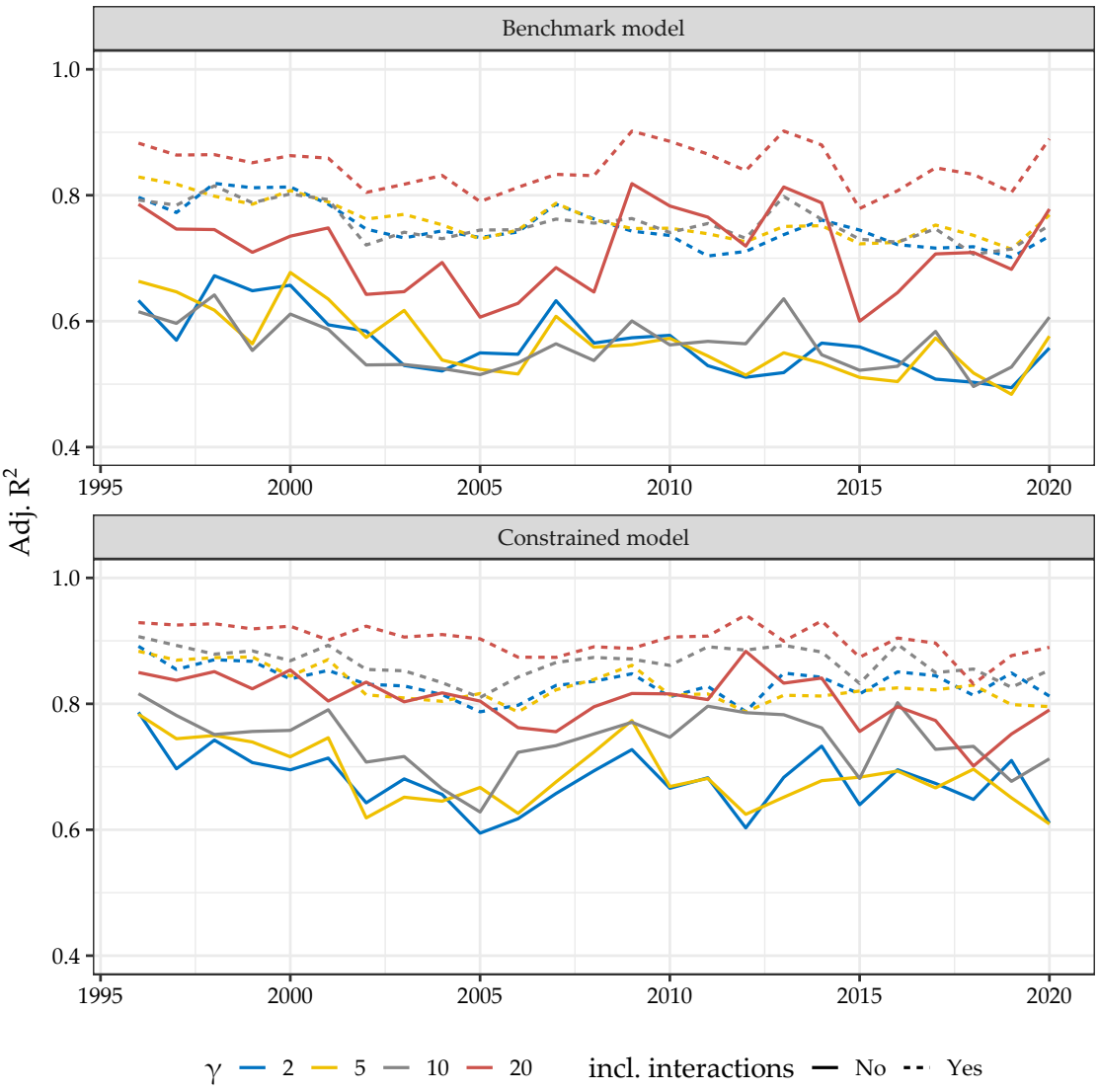



**Figure 2.8: DPPP surrogate $R^2$ without and including transaction costs and leverage constraint**
This figure depicts the adjusted $R^2$ of the surrogate models for both the unconstrained and the constrained DPPP models for different risk aversions ($\gamma$). More specifically, the lines show the adjusted $R^2$ of a linear surrogate model in which estimated DPPP weights are regressed on the 50 most important variables across all out-of-sample periods. Interactions include all possible two-way interactions between the variables.

imposing realistic constraints as shown in 2.4.2, one would expect that a linear surrogate explains a larger portion of portfolio weight predictions in the constrained setting. In fact, this is what we find empirically, i.e., the surrogate $R^2$s are generally much higher in the constrained setting as compared to the unconstrained benchmark case. More precisely, the simple linear surrogate model explains between 70% and 90% of the weights for $\gamma = 20$, while the $R^2$ ranges between 60% and 80% for the other degrees of risk aversion considered. Introducing transaction costs and a leverage constraint hence results in an increase of $\sim$10% of the simple linear surrogate $R^2$. Analogous to the unconstrained case,

adding interactions further leads to a surrogate $R^2$-increase of $\sim$10%. In fact, in 2012 and for $\gamma = 20$, the linear surrogate model including interactions nearly perfectly explains variation in weight predictions ($R^2$ of 95%).

The analysis stresses the fact that the complexity of the DPPP decreases in a realistic setting and when increasing risk aversion. Moreover, based on these numbers, we infer that between 50-90% of the underlying characteristic-weight relationship is of linear nature, depending on whether we impose constraints and the degree of risk aversion. About another 10-20% can be captured by interactions, and the remaining 5-30% can be attributed to the non-linear functional form of the DPPP model.[17]

## 2.5 Different investor utility functions

Similarly to varying the degree of risk aversion for a CRRA investor, we can account for different investor types by changing the utility function that we use to optimize the models. In particular, we explore linear and deep portfolio policies for an investor with mean-variance utility defined as

$$u(r_{p,t+1}) = r_{p,t+1} - \frac{\gamma}{2}\left(r_{p,t+1} - \frac{1}{T}\sum_{t=1}^{T} r_{p,t+1}\right)^2, \tag{2.10}$$

where $\gamma$ is the absolute risk aversion of the investor, and for a loss-averse investor (Tversky and Kahneman, 1992) with utility defined as

$$u(r_{p,t+1}) = \begin{cases} -l(\overline{W} - (1 + r_{p,t+1}))^b & \text{if } (1 + r_{p,t+1}) < \overline{W} \\ ((1 + r_{p,t+1}) - \overline{W})^b & \text{otherwise} \end{cases}, \tag{2.11}$$

where $\overline{W}$ is a reference wealth level determined in the editing stage, the parameter $l$ measures the investor's loss aversion and the parameter $b$ captures the degree of risk seeking over losses and risk aversion over gains. For simplicity, we fix the parameters $\overline{W}$ and $b$ at one and only change the loss aversion parameter $l$. We include the constraints specified in Section 2.4.2 in the optimization process for both preferences.

[17] Note that a high adjusted $R^2$ does not always translate into a similar certainty equivalent, i.e., a similar utility. In Table A.6 in the Appendix we analyze the portfolios generated by the respective surrogate models. The table shows the certainty equivalent of the portfolios generated by the surrogate models and the corresponding original DPPP. In addition, we report whether the differences between the surrogate and original certainty equivalents are statistically significant. Results are stratified by model specification and inclusion of interactions.

Table 2.3 shows the results for the linear and deep portfolio policies for a mean-variance investor with different degrees of absolute risk aversion. We report the distributional characteristics of portfolio returns net of transaction costs. Most importantly, for all degrees of risk aversion, the DPPP yields higher certainty equivalent returns than the PPP. Generally, the results for the mean-variance investor are similar to those for the CRRA investor for the DPPP. The model yields similar certainty equivalents, Sharpe ratios, and weight characteristics. In contrast, the linear model provides significantly better results for the mean-variance preference across all risk aversions. As a result, the difference in monthly certainty equivalent returns of 20-50 basis points is smaller than in the CRRA case, driven by the better performance of the linear model. In line with the previous results, the outperformance in terms of certainty equivalent difference decreases with increasing risk aversion.[18] The mean-variance utility function perfectly illustrates that the degree of absolute risk aversion determines the strength of the penalty on the variance of portfolio returns, i.e., the strength of regularization, since portfolio return variance is an explicit part of the utility function. This is supported not only by the decreasing difference in certainty equivalents with increasing risk aversion, but also by the increasing p-values for the difference. In fact, for $\gamma = 10$ we find that the difference is no longer significant at the 1% level, while for $\gamma = 20$ we find the only case where the difference is not significant for all common levels.

Next, we optimize portfolio policies for the loss-averse investor and report results in Table 2.4 similar to the mean-variance investor for different levels of loss aversion. Again, the DPPP outperforms the PPP for all degrees of loss aversion. More precisely, the outperformance of the DPPP ranges between 61 basis points and 54 basis points with all differences being significant at the 1% level.[19] An interesting feature of the loss-averse investor's preference is the fact that she cares about the size of the tail of the portfolio return distribution, rather than the mean to variance ratio, which is relevant to a mean-variance investor. The results in Table 2.4 reflect this. Both portfolios display higher skewness of returns compared to the portfolios optimized conditional on mean-variance or CRRA preferences. Most importantly, the DPPP yields significantly higher skewness than the linear analogue, explaining the higher certainty equivalent for the loss-averse

[18] The outperformance of the DPPP is amplified when we remove transaction costs and the leverage constraint, analogous to our CRRA benchmark case. We report the results for this in Table A.7 in the Appendix.

[19] Again, we show in Table A.8 in the Appendix that these findings are amplified when we remove transaction costs and the leverage constraint.

**Table 2.3: (D)PPP for MV investors incl. transaction costs and leverage constraint**

| | $\gamma = 2$ | | $\gamma = 5$ | | $\gamma = 10$ | | $\gamma = 20$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0155 | 0.0205 | 0.0139 | 0.0169 | 0.0095 | 0.0115 | 0.0024 | 0.0041 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0002 | | 0.0019 | | 0.0445 | | 0.1083 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.1703 | 0.1813 | 0.1749 | 0.1819 | 0.1777 | 0.1831 | 0.1698 | 0.1807 |
| $max\ w_i * 100$ | 0.6604 | 0.8464 | 0.6827 | 0.7866 | 0.6870 | 0.7554 | 0.6496 | 0.7271 |
| $min\ w_i * 100$ | −0.6442 | −0.9616 | −0.6817 | −0.9814 | −0.6921 | −1.0005 | −0.6387 | −0.8684 |
| $\sum_i w_i I(w_i < 0)$ | −0.7280 | −0.8072 | −0.7607 | −0.8113 | −0.7808 | −0.8201 | −0.7244 | −0.8029 |
| $\sum_i I(w_i < 0)/N_t$ | 0.3279 | 0.3348 | 0.3417 | 0.3181 | 0.3455 | 0.3144 | 0.3367 | 0.3263 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 0.8275 | 1.7662 | 0.9699 | 1.6756 | 0.9834 | 1.6001 | 0.8911 | 1.4106 |
| Mean | 0.0177 | 0.0231 | 0.0183 | 0.0223 | 0.0171 | 0.0202 | 0.0165 | 0.0186 |
| StdDev | 0.0479 | 0.0517 | 0.0422 | 0.0467 | 0.0391 | 0.0417 | 0.0375 | 0.0380 |
| Skew | −0.6768 | −0.6706 | −0.9111 | −0.7508 | −0.9562 | −0.6423 | −0.9801 | −0.7631 |
| Kurt | 2.9347 | 3.3770 | 2.6367 | 2.8915 | 2.5835 | 1.9170 | 2.6507 | 2.0241 |
| SR | 1.2811 | 1.5494 | 1.5054 | 1.6560 | 1.5153 | 1.6756 | 1.5215 | 1.6961 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.0086 | | 0.0499 | | 0.0578 | | 0.0643 |
| $FF5 + Mom\ \alpha$ | 0.0063 | 0.0113 | 0.0075 | 0.0112 | 0.0069 | 0.0101 | 0.0068 | 0.0092 |
| $StdErr(\alpha)$ | 0.0013 | 0.0017 | 0.0013 | 0.0017 | 0.0014 | 0.0017 | 0.0014 | 0.0015 |

This table shows out-of-sample estimates of the (deep) portfolio policies with the transaction costs penalty (Equation (2.8)) and leverage constraint (Equation (2.9)) optimized for a mean-variance investor with absolute risk aversion of 2, 5, 10 and 20, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$\gamma = 2$", "$\gamma = 5$", "$\gamma = 10$" and "$\gamma = 20$" correspond to the respective risk aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions net of transaction costs as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

**Table 2.4: (D)PPP for LA investors incl. transaction costs and leverage constraint**

| | $l = 1.5$ | | $l = 2$ | | $l = 2.5$ | | $l = 3$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0127 | 0.0188 | 0.0094 | 0.0150 | 0.0057 | 0.0117 | 0.0032 | 0.0086 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0001 | | 0.0007 | | 0.0009 | | 0.0082 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.1688 | 0.1806 | 0.1745 | 0.1816 | 0.1755 | 0.1821 | 0.1786 | 0.1838 |
| $max\ w_i * 100$ | 0.6587 | 0.8298 | 0.6856 | 0.8486 | 0.6877 | 0.8334 | 0.7004 | 0.8658 |
| $min\ w_i * 100$ | −0.6328 | −0.9735 | −0.6719 | −0.9619 | −0.6744 | −0.9578 | −0.6948 | −0.9557 |
| $\sum_i w_i I(w_i < 0)$ | −0.7169 | −0.8018 | −0.7580 | −0.8093 | −0.7653 | −0.8125 | −0.7878 | −0.8249 |
| $\sum_i I(w_i < 0)/N_t$ | 0.3264 | 0.3285 | 0.3418 | 0.3301 | 0.3435 | 0.3284 | 0.3475 | 0.3365 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 0.8454 | 1.8264 | 0.9550 | 1.8575 | 1.0269 | 1.8608 | 1.1131 | 1.8881 |
| Mean | 0.0175 | 0.0236 | 0.0180 | 0.0234 | 0.0178 | 0.0233 | 0.0183 | 0.0232 |
| StdDev | 0.0473 | 0.0521 | 0.0430 | 0.0480 | 0.0411 | 0.0460 | 0.0401 | 0.0439 |
| Skew | −0.6541 | −0.6393 | −0.8328 | −0.6609 | −0.9037 | −0.5521 | −0.8835 | −0.5700 |
| Kurt | 2.8837 | 3.2452 | 2.5963 | 3.1598 | 2.3513 | 2.5513 | 2.2285 | 2.2444 |
| SR | 1.2806 | 1.5689 | 1.4486 | 1.6887 | 1.5035 | 1.7531 | 1.5821 | 1.8311 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.0041 | | 0.0222 | | 0.0139 | | 0.0219 |
| $FF5 + Mom\ \alpha$ | 0.0079 | 0.0147 | 0.0089 | 0.0157 | 0.0092 | 0.0160 | 0.0100 | 0.0166 |
| $StdErr(\alpha)$ | 0.0013 | 0.0017 | 0.0013 | 0.0017 | 0.0013 | 0.0017 | 0.0014 | 0.0017 |

This table shows out-of-sample estimates of the (deep) portfolio policies with the transaction costs penalty (Equation (2.8)) and leverage constraint (Equation (2.9)) optimized for a loss-averse investor with loss aversion of 1.5, 2, 2.5, and 3, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$l = 1.5$", "$l = 2$", "$l = 2.5$" and "$l = 3$" correspond to the respective loss aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions net of transaction costs as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

investor.

In contrast to previous results, we do not find a decrease in certainty equivalent differences between the DPPP and PPP with increasing loss aversion. Furthermore, mean return and standard deviation only decrease slightly with increasing loss aversion. In contrast to the risk aversion parameter $\gamma$, the loss aversion parameter $l$ does not regularize the variance of predictions directly, but rather penalizes low skewness, reducing fat tails on the left side of the distribution. This does not translate into a similar degree of economic regularization of model complexity as risk aversion.

In line with the intuition that the investor does not care about the mean to variance ratio, the p-values of the Sharpe ratios are slightly higher and do not seem to differ significantly at the 1% level for three out of four loss aversions. Lastly, although the DPPP yields slightly higher turnover in all cases, the weight distribution of the portfolios is still very similar to that for other utility functions considered.

The main results in Table 2.3 and Table 2.4 are visually summarized in Figure 2.9, which shows the cumulative performance of portfolio returns over time for both the PPP and the DPPP, all degrees of risk aversion or loss aversion, and with transaction cost and leverage constraints. The figure shows that the DPPP consistently outperforms the PPP by a substantial margin in all specifications.

## 2.6 Conclusion

Building on the parametric portfolio policy of Brandt et al. (2009), we show that feed-forward neural networks can be used to directly optimize portfolios based on a large number of firm characteristics for different investor preferences. In essence, we do so by replacing traditional distance loss functions with context-specific utility functions when optimizing neural networks. Analogous to Brandt et al. (2009), our framework allows for integration of constraints, such as transaction cost penalties or leverage restrictions.

Our empirical results indicate that neural networks perform significantly better than linear models in regards to portfolio allocation, suggesting that firm characteristics are non-linearly related to optimal portfolio weights. This is especially true when the investor's utility preference takes into account higher moments of the resulting portfolio return distribution. Consistent with this hypothesis, we show that linear surrogate models are not able to fully explain the deep parametric portfolio weight predictions,

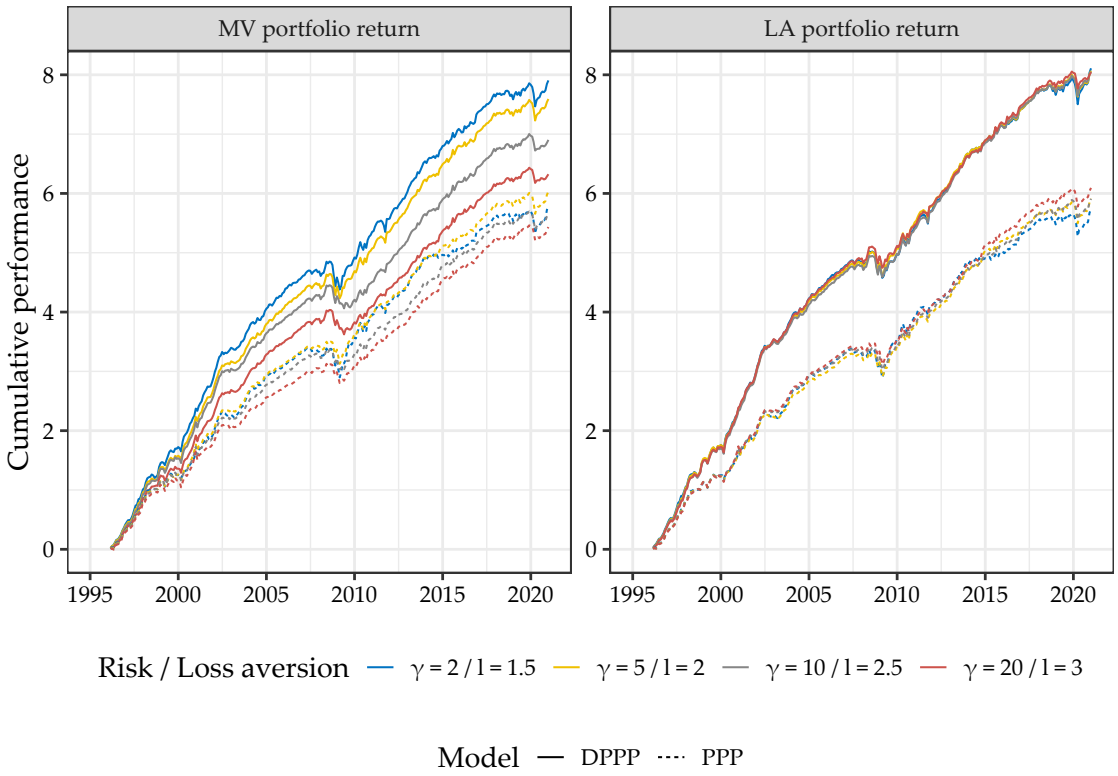



**Figure 2.9: Cumulative performance over time for MV and LA preferences**
The left panel shows the cumulative sum of portfolio returns net of trading costs for the transaction cost and leverage constrained DPPP and PPP of investors with mean-variance preferences. The right panel shows the cumulative sum of portfolio returns net of trading costs for the transaction cost and leverage constrained DPPP and PPP of investors with loss-aversion preferences. We show the results for each of the degrees of absolute risk aversion ($\gamma$) and loss aversion (l) considered and across all out-of-sample periods.

even when accounting for two-way interactions. We further shed light on the non-linear relationship between characteristics and predicted weights by depicting the sensitivity of predicted weights with respect to the input. Again, we find a clearly non-linear relation between stock characteristics and optimal portfolio weights. We further find that return-based stock characteristics resemble the most important group of predictors. However, consistent with DeMiguel et al. (2020), variable importance is more evenly distributed and puts less weight on past returns when leverage constraints and transaction costs are explicitly accounted for when deriving optimal portfolios.

Exploring variations in the degree of an investor's risk aversion and utility function, we find that a more complex non-linear model yields higher utility than a linear model in all cases. These differences are not only statistically significant, but also economically meaningful. However, higher risk aversion is associated with lower gains across all specifications. In that sense, the level of risk aversion can be seen as a regularization

parameter that leans against model complexity.

Overall, we show how to generalize the original linear parametric portfolio policy of Brandt et al. (2009), and our results support the use of neural networks in solving portfolio choice problems. While other non-linear methods might show success as well, neural nets are particularly suited because of their ability to comprehensively model arbitrary functional forms. Highlighting the growing role of machine learning and non-linear models in finance, our approach thus resembles a comparably simple and flexible neural network-based model that enables practitioners and researchers alike to create reasonable portfolio allocations based on firm characteristics and preferences.

# Chapter 3

# Interpretive Earnings Forecasts via Machine Learning: A High-Dimensional Financial Statement Data Approach[†]

## 3.1 Introduction

Future earnings are of central importance against the background of deriving the intrinsic value of assets (e.g., Monahan, 2018). In particular, analysts use earnings forecasts to derive buy/sell recommendations for stocks (e.g., Schipper, 1991; Brown, 1993). Earnings are also used in corporate decision-making, as they, inter alia, represent one of the primary financial metrics for external stakeholders in general (Graham et al., 2005). Lastly, as first shown by Ball and Brown (1968), earnings are assumed to directly relate to stock returns. One may even derive return expectations directly from predicted earnings in the form of the implied cost of capital (ICC) of a company (e.g., Gordon and Gordon, 1997; Claus and Thomas, 2001; Gebhardt et al., 2001; Easton, 2004; Ohlson and Juettner-Nauroth, 2005).

In general, earnings predictions are usually either retrieved from analysts or from statistical models. Traditionally, statistical approaches have primarily been of simple, linear nature (e.g., Hou et al., 2012; Li and Mohanram, 2014).[1] With the advent of more

[†]This chapter is based on Hess et al. (2024).

[1]Hereafter, we refer to these models as traditional models.

advanced statistical models, i.e., machine learning approaches in particular, various more flexible approaches have been proposed by researchers (e.g., Cao and You, 2021; Jones et al., 2023).

As outlined by Gu et al. (2020) and Israel et al. (2020), machine learning introduces flexibility in terms of three dimensions: first, it allows for flexibility in terms of functional form. That is, in contrast to the traditional linear approaches that have been used to predict earnings, machine learning allows for complex non-linear functional forms. Second, machine learning allows the use of large conditioning information sets, thereby enabling researchers to search for relations that have been undetected thus far. Third, machine learning entails advanced optimization techniques such as regularization to avoid overfit.

To the best of our knowledge, earnings prediction research has predominantly been focused on the first dimension so far. Put differently, researchers have proposed the use of different types of machine learning algorithms to approximate the possibly complex functional form that relates predictors and future earnings (e.g., Cao and You, 2021). It is conceivable that in general, more complex machine learning models lead to more accurate earnings predictions. In fact, Kelly et al. (2024) prove that increasing complexity, i.e., the ratio of model parameters to data, is always beneficial in terms of out-of-sample prediction performance for return prediction. They explicitly recommend to use rich non-linear model specifications rather than simple linear ones. We argue that further research along this dimension thus bears little additional insights. Moreover, extant studies on earnings prediction typically only assess a rather limited set of predictor variables. An important exception is the study by Chen et al. (2022) which exploits the entirety of Extensible Business Reporting Language (XBRL) data. Yet, they restrict their analysis to the prediction of earnings changes and do not predict earnings per se.[2] The fact that machine learning allows for the use of large conditioning information sets as mentioned by Israel et al. (2020) is rather unexploited in this context thus far.

We fill this gap in the research and predict annual earnings per share conditional on a comprehensive set of variables, i.e., the entire set of financial statement variables from Compustat. This allows us to thoroughly analyze how fundamental accounting-

[2]In fact, as an extension to their main analysis, they also predict earnings levels. However, they only use a set of 24 variables for this and not the high-dimensional XBRL data set which they use for their main analysis. Surprisingly, their machine learning approach does worse than a simple random walk model. They conclude that earnings levels are hard to predict and use this as an argument for the fact that they focus on earnings changes in their primary analysis.

based information drives and relates to future earnings. To do so, we use a selection of prominent, flexible machine learning models, namely a random forest model (RF), a gradient boosted tree model (GBT), a gradient boosted tree model with dropout (DART), a feed-forward neural network (NN) and an ensemble of the aforementioned models (ENML).[3] We further show how our approach relates to the most widely used traditional linear approaches, namely a simple model only including earnings as a predictor (L) (Gerakos and Gramacy, 2012), the HVZ-model (Hou et al., 2012), the EP-model (Li and Mohanram, 2014), the RI-model (Li and Mohanram, 2014) and an ensemble of the aforementioned models (ENTD).

Importantly, we also provide extensive model interpretation. The ability to understand the inner workings of prediction models is a fundamental requirement in most asset management applications (Israel et al., 2020). However, due to their complexity, machine learning models are hard to interpret. In the machine learning earnings prediction case, model interpretation has thus far been restricted to metrics of variable importance and partial dependencies. More so, these metrics have usually been applied to a predetermined, restrictive set of predictor variables as mentioned above. Put differently, researchers typically choose or construct a set of predictor variables that they deem important before estimating the model. After estimating the models, they derive the extent to which variables from this predetermined set contribute to the predictions and assess the partial dependencies of predicted earnings in regards to them.

This study aims to broaden the limited scope of model interpretation found in the existing literature on earnings forecasts: first, we derive the relative importance of variables using SHAP (SHapley Additive exPlanations) values, an approach based on cooperative game theory (Lundberg and Lee, 2017).[4] Since we do not select only a small subset of variables or construct variables beforehand as done in comparable studies (e.g., Hansen and Thimsen, 2020; Cao and You, 2021), we are able to holistically infer which out of all the financial statement variables are important from a statistical perspective. We also derive the relative importance of different groups of financial statement variables, such as cash flow statement (CF/S) variables, income statement (I/S) variables and balance sheet (B/S) variables. Importantly, we conduct this analysis for forecast horizons

[3]The model selection is based on Bali et al. (2023).

[4]SHAP values are a way of explaining the results of any machine learning model. They are based on a game-theoretic approach that measures the contribution of each player to the final outcome. In machine learning, each variable is assigned an importance value that represents its contribution to the model's outcome (Lundberg and Lee, 2017).

of up to five years. This allows us to assess how different (groups of) variables might vary in terms of their predictive power, depending on the forecast horizon considered. Second, we analyze non-linearity in the context of earnings prediction. In addition to partial dependencies that extant studies focus on in that context (e.g., Cao and You, 2021; Chen et al., 2022), we infer the degree to which different types of non-linearity play a role. More precisely, we show the degree to which interaction effects across financial statement variables and other forms of non-linearities, i.e., non-linearity of the functional form, are important by means of surrogate modeling. This approach is completely transparent, intuitive and easy to replicate, irrespective of the model or software used. Again, we conduct these analyses for forecast horizons of up to five years.

Our results can be summarized as follows: first, we find that generally, ensembles of models perform significantly better than their component models. This holds true both for traditional linear approaches and the machine learning approaches considered. For one-year-ahead predictions, for example, the machine learning ensemble yields around 2% more accurate predictions than the best performing component machine learning model.

In terms of bias, the traditional and the machine learning approaches yield very similar results. Both types of models yield mostly unbiased predictions for forecasts horizons of up to three years. For forecasts horizons of four and five years, the models considered begin to systematically overestimate earnings. However, the machine learning models are consistently less biased in terms of levels as compared to their traditional analogues.

Mirroring previous findings on machine learning earnings predictions, we further show that our machine learning approaches constantly outperform traditional linear approaches in terms of accuracy (e.g., Cao and You, 2021). For the one-year forecast horizon, the best performing machine learning model, i.e., the ENML is around 12% more accurate than the best performing traditional model, i.e., the ENTD. Even for long forecast horizons of five years, we find that the ENML beats the ENTD in terms of accuracy by around 7%. The superiority in terms of accuracy holds similarly for both small and large firms. Furthermore, assessing accuracy differences across out-of-sample periods shows that model performance converges in the periods following the financial crisis, and diverges in favor of the machine learning approaches afterwards.

Assessing the degree to which the models are able to explain out-of-sample variation

in earnings makes an even more convincing case for the ML approaches. In fact, the average out-of-sample $R^2$ (OOS $R^2$) of the ENML is between 14% and 28% higher than that of the ENTD for the forecast horizons considered.

Lastly, we show that the more accurate predictions of the machine learning approach translate into more profitable portfolios based on ICC. Buy-and-hold returns of long-short portfolios sorted on ICC based on the best-performing machine learning model (ENML-ICC) surpass the returns of long-short portfolios sorted on ICC based on the best-performing traditional model (ENTD-ICC). Moreover, while returns stay statistically significant for portfolios based on ENML-ICC for each buy-and-hold period length considered, they are only statistically significant for one of three buy-and-hold period lengths considered in the ENTD-ICC case.

Turning to model interpretation, we find that current I/S variables, especially current earnings, are the most important group of predictors. With increasing forecast horizon, however, variable importance becomes more balanced among financial statement types. For one-year-ahead predictions, I/S variables contribute around 65% while B/S variables contribute around 20% to total importance. Total importance of these two groups of variables consistently converges with increasing forecast horizon. More precisely, for $t+5$-predictions, I/S variables contribute around 47% while B/S variables contribute around 37% to total importance. CF/S variables consistently contribute around 15% to total importance throughout the forecast horizons considered. Put differently, the longer the forecast horizon, the more important B/S information. Further disentangling the effects of different components of financial statement information reveals that certain pieces of financial statement information dominate others. For example, debt and supplemental information resemble the most important pieces of B/S information. Turning to the CF/S, we find that variables related to the operating cash flow are much more relevant than variables related to either the investing cash flow or the financing cash flow. Lastly, turning to the I/S, we find that especially the EBIT and the net income are important, contributing around 14% and 31% to total importance for one-year-ahead forecasts, respectively. Interestingly, however, the importance of net income consistently declines with increasing forecast horizon to around 16% for five-years-ahead forecasts, whereas the importance of the EBIT stays constant. This suggests that information which is less exposed to accounting manipulation gains relevance for longer-term forecasts.

We further show that for one-year-ahead predictions, a linear surrogate model is able

to explain around 80-90% of the variation of earnings predictions across out-of-sample periods. In contrast to Jones et al. (2023), we find that interactions across financial statement variables are irrelevant.[5] We attribute the remaining 10-20% of unexplained variation to other types of non-linearities, which are not captured by interactions, i.e., non-linearity of the functional form. As the forecast horizon increases, the linear surrogate model approximates the relationship between predictions and inputs slightly worse. Interestingly, interaction effects across financial statement variables stay irrelevant for all forecast horizons.

The remainder of this study is structured as follows: in Section 3.2 we outline the relevant literature that we are contributing to. In Section 3.3 we describe our empirical approach. We evaluate and compare our approach in Section 3.4. In Section 3.5 we provide extensive interpretation. Finally, Section 3.6 concludes the study.

## 3.2 Related literature

Our work relates to three strands of literature in particular. First, we contribute to the literature on machine learning applications in finance. Machine learning methods have become the prevalent way of conducting prediction exercises, primarily due to their superiority in terms of flexibility as compared to traditional econometric methods and their efficacy in regards to large sets of input data (e.g., Israel et al., 2020; Kelly et al., 2024). For example, Gu et al. (2020) show how different machine learning approaches perform in terms of predicting stock returns and Bali et al. (2023) apply machine learning to the task of predicting option returns. Our study is similar, in the sense that we predict another financial variable, i.e., earnings, with machine learning.

Second, we apply state-of-the-art techniques to interpret our machine learning predictions, thereby explicitly responding to the "need for interpretability" of financial machine learning models as formulated by Israel et al. (2020). We hence also contribute to the literature that aims to foster transparency and understanding of machine learning methods for prediction in finance research, such as e.g., Bali et al. (2023) who extensively assess machine learning option return predictions.

Third, we contribute to the literature on model-based earnings forecasts. Traditionally, researchers have suggested predicting earnings using time-series regression models (e.g.,

[5] This comes with a word of caution, as our input variables differ from theirs.

Ball and Watts, 1972; Albrecht et al., 1977; Watts and Leftwich, 1977), cross-sectional regression models (e.g., Hou et al., 2012; Li and Mohanram, 2014; Harris and Wang, 2019) or even simple random walk models (e.g., Li and Mohanram, 2014). More recently, however, several machine learning earnings prediction approaches have been implemented in the research (e.g., Hansen and Thimsen, 2020; Cao and You, 2021; Chen et al., 2022; Hendriock, 2022; Campbell et al., 2023; Jones et al., 2023; Van Binsbergen et al., 2023). However, the extant literature differs from our study in several key aspects which are outlined in the following. Hansen and Thimsen (2020) also estimate a range of machine learning methods. They use a more high-dimensional input vector than other studies on predicting level earnings. However, it is still restrictive in the sense that it is based on prior research and not as high-dimensional as our data. Moreover, in contrast to our study, no model interpretation is provided.

The study of Cao and You (2021) also encompasses different machine learning models. However, they use a more restrictive set of input variables than us and provide only limited model interpretation. Moreover, Cao and You (2021) validate their models using traditional cross-validation (and a limited hyperparameter space). This, however, destroys the temporal structure of the observations and introduces information leakage (Gu et al., 2020). We preserve the temporal ordering of the observations by using fixed training, validation and test intervals.

Chen et al. (2022) use a single model (gradient boosted trees) as opposed to our multi-model approach. Furthermore, they predict binary earnings changes, while we focus on predicting level earnings.

Hendriock (2022) suggests predicting earnings by predicting the complete conditional density function. However, he restricts his input variable space to the one as defined by traditional linear models. Moreover, he does not provide model interpretation.

Campbell et al. (2023) benchmark an extensive range of machine learning model specifications with the aim of identifying the ones which compare the best to analyst forecasts. Apart from the fact that their model choices differ from ours, they use a different, smaller set of inputs, i.e., the Wharton Research Data Services (WRDS) Financial Suite Ratios extended by some additional variables like e.g., the stock return. Furthermore, they provide limited model interpretation as their primary focus is on the aforementioned horse-race between model specifications.

Jones et al. (2023) use a single model, i.e., a gradient boosted tree model algorithm,

as opposed to our ensemble approach. In contrast to our study, their target variable is return on net operating assets and they use a set of six ratios as their predictors. Another crucial difference is that Jones et al. (2023) exclusively forecast earnings (changes) in $t+1$, while we forecast earnings (per share) for horizons $t+1$ to $t+5$. Finally, to the best of our knowledge, as the only other study in this context, they assess the impact of interactions. However, their method of doing so and their predictor variables differ from ours. Our surrogate modeling approach is easily applicable to any type of model, in any software and further allows us to explicitly determine the effect of interaction effects and non-linearity in parameters. Interestingly, our findings in regards to interactions differ strongly from the ones provided by Jones et al. (2023). They find substantial importance of interactions, while we find that interactions among financial statement variables are irrelevant.

Van Binsbergen et al. (2023) use a random forest model to predict earnings conditional on financial ratios, similar to Campbell et al. (2023). In contrast, our approach involves estimating a spectrum of machine learning models individually and in ensemble configurations. Moreover, we utilize an entirely distinct set of input data, specifically the comprehensive Compustat financial statement dataset. Finally, unlike their study, which primarily assesses analyst biases, our analysis encompasses detailed explanations of model predictions.

Summing up, to the best of our knowledge, we are the first to predict level earnings per share for forecast horizons of up to five years using the entirety of available Compustat financial statement variables. Furthermore, we contribute novel guidance for future research on earnings (per share) predictions by thoroughly interpreting our state-of-the-art machine learning approaches using model agnostic and easily applicable methods, something that has been not done extensively thus far.

## 3.3 Empirical approach

### 3.3.1 General setup

We express earnings $E$ of firm $i$ in period $t+\tau$ as the expectation in period $t$ plus an error $\epsilon$:

$$E_{i,t+\tau} = \mathbb{E}_t[E_{i,t+\tau}] + \epsilon_{i,t+\tau}. \tag{3.1}$$

Every model that aims to predict earnings can be interpreted as an attempt to derive $\mathbb{E}_t[E_{i,t+\tau}]$, i.e., the expectation. More precisely, we assume that expected earnings in $t+\tau$ are a function of a vector of inputs $\boldsymbol{X}$ of firm $i$ known at time $t$:

$$\mathbb{E}_t[E_{i,t+\tau}] = f(\boldsymbol{X}_{i,t}). \tag{3.2}$$

It becomes evident that modelling expected earnings consists of three crucial parts. First, one has to determine which inputs enter the model, that is, how $\boldsymbol{X}$ is defined. Second, one has to decide which functional form $f(\cdot)$ takes on. This corresponds to the decision about which empirical model to choose from. Lastly, one has to decide on how to estimate $f(\cdot)$, i.e., which statistical loss function to minimize. The latter is not the explicit focus of this study and thus we keep it simple. We follow the original implementations of the traditional models and estimate them using the mean squared error (MSE). In case of the machine learning models, we follow Gu et al. (2020) and estimate the models both using the MSE and the mean absolute error (MAE) and report the predictions based on the loss function that leads to more accurate forecasts according to the price scaled absolute forecast error (PAFE) at the 1-year horizon.[6]

Thus, apart from the decision regarding the loss-function, the two contrasting *extreme* approaches to predicting earnings are: (1) actively making the decision which variables and which functional form to assume ex ante, and (2) letting the data speak by selecting a model that permits flexible functional forms and providing it with the entire data (or at least a very large set of variables) available. The former corresponds to the traditional prediction approaches suggested by e.g., Hou et al. (2012). To the best of our knowledge, the latter has not yet been implemented for earnings per se, a noteworthy exception being Chen et al. (2022) who employ this approach for (binary) earnings *changes*. There are approaches that fall in between (1) and (2). In these approaches, either the functional form or the input vector is restricted significantly (e.g., Hendriock, 2022; Van Binsbergen et al., 2023).

Our study focuses on assessing the second, flexible approach (2) and comparing it to traditional approaches (1). We more thoroughly elaborate on the choice of the input vector and of the model in 3.3.2 and 3.3.3, respectively.

[6]To be precise, Gu et al. (2020) choose either the MSE or the Huber loss, depending on which performs better. We choose either the MSE or the MAE, depending on which performs better.

### 3.3.2 Data

US annual financial statement data is obtained from Compustat. Our sample period ranges from 1988 to 2021. This is due to the fact that CF/S-data is only sparsely available prior to 1988. To conduct the ICC portfolio evaluation, we add price and return data from CRSP to the Compustat data used for model estimation. Moreover, we drop observations with missing prices, prices smaller than 1$, missing common shares outstanding or missing earnings. This results in a final sample for model estimation that consists of 191,273 observations.

For our machine learning models, we use the Compustat financial statement items as predictors. We drop variables with more than 50% of observations missing or no observations in any of the cross-sections (i.e., estimation years), yielding 192 variables.[7] An overview over these variables is given in Table B.3 in the Appendix. Analogous to Chen et al. (2022), we include lags and first-order differences of these variables, resulting in a set of 576 predictor variables in total. For the traditional models, we construct input variables according to the respective models. An overview over these variables is given in Table B.1 in the Appendix. All our variables, including our target variable, are scaled by common shares outstanding as of the estimation year and winsorized at the 1% and 99% level, respectively. Finally, since neural networks are sensitive to scale differences across input variables, we standardize each variable.

### 3.3.3 Models

We estimate two groups of models. The first group consists of popular simple linear models that have been introduced in the literature thus far. All of these models assume a linear additive relation between earnings and some low-dimensional input vector, i.e.,

$$\mathbb{E}_t[E_{i,t+\tau}] = \boldsymbol{\beta} \boldsymbol{X_{i,t}}, \tag{3.3}$$

where $\boldsymbol{\beta}$ denotes a vector of coefficients. The models differ in terms of which variables the input vector consists of. A more detailed description is given in Appendix B.1. A difference to be noted is that the models also slightly differ in how they define the output. While the RI and the EP model use earnings per share, the HVZ model uses earnings.

[7] We also drop variables already scaled by shares. The reason for that is, that we scale all our variables by shares as mentioned below and hence these variables are redundant. However, this only pertains to five variables in our study.

In this study, we define earnings as "income before extraordinary items" (Compustat variable: ib). As mentioned above, we consistently scale our output as well as our input variables by common shares outstanding in all of the models. Our forecast horizon, both for this and the following group of models spans from $\tau = 1$ to $\tau = 5$.

The second group of models consists of flexible models that are able to approximate arbitrary complex functional forms. Put differently, we do not assume any specific type of functional form when estimating these models. Analogous to Bali et al. (2023), we estimate a random forest model (RF), a gradient boosted tree model with (DART) and without dropout (GBT) and a neural net (NN). Importantly, we try to restrict our input vector as little as possible. Specifically, we feed the models the 576 variables as outlined above, including the lags and first-level differences. We argue that this input vector corresponds to a proxy for the entirety of (relevant) financial statement input variables. Note that extending the input vector even further implies more computational effort.[8]

Since studies in other realms of financial forecasting have shown that using an ensemble of models may prove superior to using single models (e.g., Bali et al., 2023), we derive the equally weighted average prediction for both groups of models, i.e., we derive two ensemble model predictions:

$$\mathbb{E}_t^{(En)}[E_{i,t+\tau}] = \frac{1}{J}\Sigma_{j=1}^{J}\mathbb{E}_t^{(j)}[E_{i,t+\tau}], \tag{3.4}$$

where $j \in J$ denotes the respective single model and $En$ denotes the respective ensemble model.

The ensemble of the traditional models is denoted by *ENTD* and the ensemble of the fully flexible machine learning prediction models is denoted by *ENML*.

### 3.3.4 Out-of-sample approach

We employ a rolling window strategy to obtain our out-of-sample prediction results. Specifically, for the machine learning models, we divide our data into training, validation and test sets. For each forecast horizon $\tau$, the process for generating forecasts as of $t$ proceeds as follows: we train our models using earnings from $t-11$ to $t-2$ as output and corresponding financial statement data lagged by $\tau$ as predictors. Next,

[8] We could have also included other variables like price, analyst forecasts, etc. (e.g., Campbell et al., 2023; Van Binsbergen et al., 2023). However, the focus of our study is the thorough analysis of the relationship between fundamental accounting-based information and future earnings.

we tune the machine learning models using earnings from $t-1$ to $t$ as output and lagged financial statement data by $\tau$ as predictors. Tuning involves determining the optimal hyperparameter values for the model, as detailed in Table B.2 in the Appendix. Subsequently, earnings predictions for $t+\tau$ are derived by inputting variables from $t$ into the optimized models. This process is repeated recursively, advancing one year at a time. We follow the approach of Hou et al. (2012) and Li and Mohanram (2014), estimating models at the end of June each year, under the assumption of a reporting lag of three to fourteen months for financial statements.[9]

Traditional linear approaches, on the other hand, do not necessitate a tuning window. Hence, we only partition the data into training and test sets when estimating these models. The subsequent steps of the procedure remain unchanged. More precisely, for each forecast horizon $\tau$, models are trained using earnings from $t-11$ to $t$ as output and corresponding lagged financial statement data by $\tau$ as predictors. Earnings predictions for $t+\tau$ are then derived by utilizing predictor variables from $t$.

### 3.3.5 Evaluation

We evaluate the predictive performance of the models across a range of evaluation metrics. First, we compute the error metrics that are common in the earnings prediction literature (e.g., Hou et al., 2012). For each forecast horizon $\tau \in [1,2,3,4,5]$, these include the price scaled forecast error (PFE) or bias:

$$PFE_{i,t+\tau} = \frac{E_{i,t+\tau} - \hat{E}_{i,t+\tau}}{Price_{i,t}}, \tag{3.5}$$

and the price scaled absolute forecast error (PAFE) or accuracy:

$$PAFE_{i,t+\tau} = \frac{|E_{i,t+\tau} - \hat{E}_{i,t+\tau}|}{Price_{i,t}}, \tag{3.6}$$

where $E_{i,t+\tau}$ denotes actual earnings for firm $i$ in period $t+\tau$, $\hat{E}_{i,t+\tau}$ denotes the respective forecast and $Price_{i,t}$ is the firm's stock price at the end of June in the respective estimation year.

Second, analogous to Gu et al. (2020), we make pairwise comparisons of the individual as well as the ensemble models using an adjusted version of the Diebold and Mariano

[9] Specifically, data from April of year $t-1$ to March of year $t$ is considered the most recent fiscal year-end data available as of June in year $t$, capturing the information as of $t$.

(1995) procedure. This procedure allows for a quantitative comparison of the different forecasts. More precisely, we compare the forecast performance of model (1) and (2) using the test statistic $DM = \bar{d}/\hat{\sigma}_{\bar{d}}$, where

$$d_{t+\tau} = \frac{1}{n_{t+\tau}} \Sigma_{i=1}^{n_{t+\tau}} ((PAFE_{i,t+\tau}^{(1)} - PAFE_{i,t+\tau}^{(2)})^2), \tag{3.7}$$

with $n_{t+\tau}$ being the number of firms in the out-of-sample period $t+\tau$. $\bar{d}$ and $\hat{\sigma}_{\bar{d}}$ then denote the mean and the Newey-West adjusted standard error (Newey and West, 1987) of $d_{t+\tau}$ over the out-of-sample periods.

Third, we assess the out-of-sample $R^2$ ($OOSR^2$) of each individual as well as the ensemble forecasts, i.e., for every out-of-sample period we calculate

$$OOS\, R^2_{t+\tau} = 1 - \frac{\Sigma_{i=1}^{n_{t+\tau}} (E_{i,t+\tau} - \hat{E}_{i,t+\tau})^2}{\Sigma_{i=1}^{n_{t+\tau}} (E_{i,t+\tau} - \bar{E}_{t+\tau})^2}, \tag{3.8}$$

for each model. Here, $\bar{E}_{i,t+\tau}$ denotes average earnings of firms in period $t+\tau$. Albeit not commonly used in the earnings prediction literature (e.g., Hendriock (2022) being an exception), this evaluation metric is of particular importance for the typical use case of earnings predictions, i.e., long-short ICC portfolios. In this context, predicting cross-sectional variation is much more important than accurately predicting earnings per se, since an investor goes long (short) the stocks whose ICC are high (low) in cross-sectional comparison.

Lastly, we derive the ICC based on the two ensemble forecasts. We follow the literature and calculate ICC following the methods of Gordon and Gordon (1997), Claus and Thomas (2001), Gebhardt et al. (2001), Easton (2004) and Ohlson and Juettner-Nauroth (2005).[10] More precisely, our ICC estimates are derived as the average of the five aforementioned methods using both the ENTD and the ENML earnings predictions. We then construct equally weighted long-short zero investment portfolios based on ICC and assess their average performance across the out-of-sample periods.

### 3.3.6 Interpretation

A primary contribution of this study is the comprehensive interpretation of the machine learning approach, addressing a key issue in machine learning applications in finance and accounting (Israel et al., 2020). Our study fills a gap in the existing literature, which either

[10] A description of the models is provided in Appendix B.3.

lacks model interpretation entirely or provides only limited insights, as discussed above. More precisely, we derive the variable importance and the degree to which different types of non-linearity play a role for our best performing machine learning model, i.e., the machine learning ensemble.

**Variable importance**

We determine the importance of the different variables with respect to the earnings prediction. To do so, we compute their SHAP values, a state-of-the-art approach for assessing the importance of input variables which is based on cooperative game theory (Lundberg and Lee, 2017). In essence, SHAP values approximate how a model's prediction changes when knowing the value of a respective input variable. The approach is model-agnostic and allows us to evaluate the importance of input variables irrespective of the model used. We conduct these analyses at both the individual variable level and the grouped-variable level. Specifically, we determine the relative importance of predictors grouped into balance sheet, cash flow statement, and income statement data as well as predictors grouped into current, lagged, and difference variables. Furthermore, we provide an in-depth accounting perspective on which specific types of financial statement information are important by breaking the financial statements down into schematic components. Importantly, we conduct these analyses per forecast horizon.

**Non-linearity**

In addition to deriving the variable importance, we also evaluate the extent to which non-linearity plays a role in our machine learning model. First, we assess the degree to which non-linearity in terms of variables and non-linearity in terms of functional form play a role. In a first step, we regress predicted earnings on the 50 most important input variables using a linear Lasso-penalized regression model.[11] We add a Lasso-term to control for the overfit that would otherwise be induced by the large input vector. In a second step, we add all possible two-way interactions to the surrogate regression model from the prior step. Assuming there is some degree of non-linearity present in the fully flexible model, this allows us to disentangle the degree to which non-linearity in terms of variables (i.e., interactions across inputs) and non-linearity in terms of functional form play a role. More specifically, we assess the adjusted in-sample $R^2$s of the two

[11] We only use the 50 most important variables, because otherwise, including all possible two-way interactions in the second step requires an excessive amount of computing power.

surrogate models. The $R^2$ of the first-step surrogate model indicates the degree to which the predictions are linear. The difference between the $R^2$ of the first-step surrogate model and the second-step surrogate model indicates the degree to which non-linearity in terms of variables, i.e., (two-way) interactions across financial statement variables, plays a role. We attribute the portion that remains unexplained by the second-step surrogate model to non-linearity in functional form.[12]

Second, we assess the partial dependence of earnings with respect to the most important input variables. As typically done in the literature, we evaluate the partial dependencies graphically via so-called partial dependence plots. To do so, we fit a non-parametric lowess model (locally weighted linear regression) to the SHAP values of a predictor value of interest (the output) and the associated predictor values (the input) and plot the result. This allows us to approximate the effect of the respective predictor variable on future earnings.

Again, we conduct the surrogate modeling and the partial dependence analyses per forecast horizon.

## 3.4 Evaluation

### 3.4.1 Accuracy and bias

Chen et al. (2022) report that their machine learning approach does worse than a simple random walk type model when predicting level earnings. They explicitly state that level earnings are hard to predict and thus resort to the prediction of earnings changes in their main analysis. In contrast, mirroring findings by e.g., Cao and You (2021) our flexible machine learning models for earnings (per share) level prediction outperform the traditional linear models by a significant margin.

#### Price scaled forecast error

Table 3.1 shows the time-series averages of the median PFE for the four traditional, the four machine learning and the two ensemble models. The PFE provides insight into whether the estimated earnings are systematically over- or underestimated (biased)

[12]Theoretically, the unexplained portion also includes effects of interaction terms of order three and higher. However, we assume that these can be neglected and find evidence for this assumption in undocumented analyses.

relative to actual earnings.

For forecast horizons $t+1$-$t+2$ neither the traditional ensemble nor the machine learning ensemble yield statistically significant PFEs. Put differently, neither of the two ensemble approaches systematically over- or underestimates earnings for these forecast horizons. This is in line with extant studies on earnings prediction, which report that earnings predictions by statistical models do not exhibit biases as opposed to those by analysts (Hou et al., 2012). Nevertheless, the ENML model has a lower bias compared to the ENTD (0.0003 vs 0.0019 for $t+1$ predictions). The traditional ensemble systematically overestimates earnings in $t+3$-$t+5$. The same is true for the ENML for $t+4$-$t+5$, but the bias is again 41% lower for $t+4$ predictions (-0.0082 vs -0.0138) and 37% lower for $t+5$ predictions (-0.0146 vs -0.0230).

The results further show that at the non-ensemble level across horizons $t+1$ to $t+3$, some models yield small but statistically significant PFEs, with the most biased machine learning model for $t+1$ being the RF model (0.0045) and the most biased traditional model for $t+1$ being the RI model (0.0047). For $t+4$ and $t+5$, all non-ensemble models yield PFEs that are statistically significantly different from zero. However, all non-ensemble machine learning models score lower PFEs than their traditional counterparts.

Overall, the machine learning models generally exhibit lower bias. Moreover, earnings prediction models appear to systematically overestimate earnings as the forecast horizon increases. This effect holds, irrespective of whether we predict earnings using traditional linear or machine learning methods.

As a robustness check, we stratify our predictions by size and report the resulting PFEs in Table 3.2. More precisely, at every prediction date $t$, we split the sample into two equally sized groups based on their market capitalization as of $t$. This allows us to separately assess PFEs for small and large firms, respectively.

For large firms, we find that the all traditional models significantly underestimate the actual earnings for forecast horizons $t+1$-$t+2$, while the biases of the machine learning models are smaller and, with the exception of the RF model, not significantly different from zero for these horizons. For forecast horizons $t+3$-$t+4$, no model, except for the DART model for the $t+4$ forecast horizon, is significantly biased. For predictions in $t+5$, most of the non-ensemble models significantly overestimate earnings. Similar to the full sample case, both ensemble models somewhat equally overestimate actual earnings in $t+5$ (-0.0096 vs -0.0079).

**Table 3.1: Median PFE**

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| $L^{MSE}$ | −0.0031** | −0.0065*** | −0.0105*** | −0.0173*** | −0.0254*** |
| $HVZ^{MSE}$ | 0.0009 | −0.0023 | −0.0072** | −0.0153*** | −0.0256*** |
| $EP^{MSE}$ | 0.0041*** | 0.0019 | −0.0023 | −0.0096*** | −0.0185*** |
| $RI^{MSE}$ | 0.0047*** | 0.0025 | −0.0018 | −0.0096*** | −0.0191*** |
| ENTD | 0.0019 | −0.0013 | −0.0058* | −0.0138*** | −0.0230*** |
| $RF^{MSE}$ | 0.0045*** | 0.0034* | −0.0007 | −0.0078*** | −0.0163*** |
| $GBT^{MAE}$ | −0.0020 | −0.0035* | −0.0049** | −0.0088*** | −0.0129*** |
| $DART^{MAE}$ | −0.0013 | −0.0029 | −0.0045 | −0.0075** | −0.0133*** |
| $NN^{MAE}$ | −0.0015 | −0.0040* | −0.0044 | −0.0093*** | −0.0140*** |
| ENML | 0.0003 | −0.0017 | −0.0035 | −0.0082*** | −0.0146*** |

This table reports the time-series averages of the median price scaled forecasting errors (PFEs) for all models. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Turning to small firms, the traditional ensemble model overestimates actual earnings with increasing bias (-0.0045 to -0.0449 from $t+1$ to $t+5$). The bias is statistically significant for all horizons. The machine learning ensemble also overestimates actual earnings, but the biases are much smaller (-0.0019 to -0.0255 from $t+1$ to $t+5$). The biases are statistically significant for horizons $t+2$ to $t+5$. We conclude that the ENML performs much better than its linear analogue for small firms specifically.

**Price scaled absolute forecast error**

Turning to the next evaluation metric, Table 3.3 reports the time-series averages of the median PAFEs for the four traditional, the four machine learning and the two ensemble models. The PAFE is a measure for the accuracy of a model, with values closer to zero indicating higher accuracy.

First, we observe a positive effect of model stacking. Overall, among the traditional models, the best-performing one is the traditional ensemble, while among the machine learning models, the best-performing one is the machine learning ensemble. In fact, the

**Table 3.2: Median PFE by firm size**

| | Large firms | | | | |
|---|---|---|---|---|---|
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | 0.0054*** | 0.0048*** | 0.0027 | −0.0017 | −0.0075* |
| $HVZ^{MSE}$ | 0.0041*** | 0.0029* | 0.0006 | −0.0041 | −0.0104*** |
| $EP^{MSE}$ | 0.0058*** | 0.0054*** | 0.0030 | −0.0017 | −0.0082* |
| $RI^{MSE}$ | 0.0048*** | 0.0038** | 0.0009 | −0.0047 | −0.0116*** |
| ENTD | 0.0050*** | 0.0043** | 0.0018 | −0.0031 | −0.0096** |
| $RF^{MSE}$ | 0.0047*** | 0.0043*** | 0.0017 | −0.0030 | −0.0090*** |
| $GBT^{MAE}$ | −0.0004 | −0.0010 | −0.0016 | −0.0033 | −0.0063* |
| $DART^{MAE}$ | 0.0002 | −0.0010 | −0.0020 | −0.0039* | −0.0094*** |
| $NN^{MAE}$ | 0.0001 | −0.0001 | 0.0008 | −0.0026 | −0.0052 |
| ENML | 0.0012 | 0.0006 | −0.0003 | −0.0031 | −0.0079** |
| | Small firms | | | | |
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | −0.0239*** | −0.0315*** | −0.0378*** | −0.0491*** | −0.0621*** |
| $HVZ^{MSE}$ | −0.0049** | −0.0116*** | −0.0212*** | −0.0349*** | −0.0508*** |
| $EP^{MSE}$ | 0.0003 | −0.0048 | −0.0127*** | −0.0241*** | −0.0366*** |
| $RI^{MSE}$ | 0.0043** | 0.0006 | −0.0064* | −0.0181*** | −0.0310*** |
| ENTD | −0.0045** | −0.0111*** | −0.0195*** | −0.0320*** | −0.0449*** |
| $RF^{MSE}$ | 0.0041** | 0.0018 | −0.0049 | −0.0167*** | −0.0293*** |
| $GBT^{MAE}$ | −0.0053** | −0.0076** | −0.0112*** | −0.0188*** | −0.0241*** |
| $DART^{MAE}$ | −0.0044* | −0.0058* | −0.0086** | −0.0132** | −0.0200*** |
| $NN^{MAE}$ | −0.0046** | −0.0126*** | −0.0140*** | −0.0208*** | −0.0302*** |
| ENML | −0.0019 | −0.0056** | −0.0090*** | −0.0169*** | −0.0255*** |

This table reports the time-series averages of the median price scaled forecasting errors (PFEs) for all models, stratified by firm size. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

machine learning ensemble outperforms each of its individual component models for every forecast horizon. The same is true for the traditional ensemble for the horizons $t+1$ to $t+2$. Especially in the traditional case, this result is surprising, since the models differ very little in terms of the predictor variables. For forecast horizons $t+3$ to $t+5$,

**Table 3.3: Median PAFE**

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| $L^{MSE}$ | 0.0317*** | 0.0409*** | 0.0463*** | 0.0531*** | 0.0606*** |
| $HVZ^{MSE}$ | 0.0293*** | 0.0376*** | 0.0429*** | 0.0499*** | 0.0578*** |
| $EP^{MSE}$ | 0.0297*** | 0.0375*** | 0.0416*** | 0.0471*** | 0.0535*** |
| $RI^{MSE}$ | 0.0293*** | 0.0365*** | 0.0403*** | 0.0453*** | 0.0517*** |
| ENTD | 0.0282*** | 0.0360*** | 0.0406*** | 0.0467*** | 0.0536*** |
| $RF^{MSE}$ | 0.0273*** | 0.0349*** | 0.0402*** | 0.0463*** | 0.0523*** |
| $GBT^{MAE}$ | 0.0274*** | 0.0357*** | 0.0407*** | 0.0460*** | 0.0515*** |
| $DART^{MAE}$ | 0.0253*** | 0.0342*** | 0.0399*** | 0.0465*** | 0.0531*** |
| $NN^{MAE}$ | 0.0257*** | 0.0361*** | 0.0410*** | 0.0451*** | 0.0524*** |
| ENML | 0.0249*** | 0.0335*** | 0.0384*** | 0.0441*** | 0.0498*** |
| ENML - ENTD | −0.0033*** | −0.0026*** | −0.0022*** | −0.0026*** | −0.0038*** |

This table reports the time-series averages of the median price scaled absolute forecasting errors (PAFEs) for all models. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled PAFE for the $t+1$ horizon). The PAFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

the traditional ensemble performs slightly worse than the best performing traditional component model (0.0406 vs 0.0403 in $t+3$, 0.0467 vs 0.0453 in $t+4$ and 0.0536 vs 0.0517 in $t+5$). Second, we find that in most cases, all machine learning models, including the ensemble, outperform all traditional models, including the ensemble, for all prediction horizons. However, for forecast horizons $t+3$ to $t+5$, the RI model is more accurate than some of the machine learning component models. This stresses the benefit of model averaging for the non-linear machine learning models specifically. The difference in accuracy between the machine learning and the linear ensemble is at a statistically significant level between -0.0022 and -0.0038. This translates into a relative difference of 11.70% for $t+1$ to 7.09% for $t+5$ predictions. The ENML thus provides not only statistically significant, but also economically meaningful gains in accuracy over the traditional models.

Figure 3.1 illustrates the median PAFE of the ENML and ENTD for each out-of-sample year and for forecast horizons of $t+1$ and $t+5$, respectively. The plots illustrate that

the overall PAFE levels strongly increase with increasing forecast horizon. Furthermore, they reveal that the machine learning ensemble is more accurate in all years and for both forecast horizons, except for 2009, for which the ENTD yields slightly more accurate $t+5$ forecasts than the ENML. In general, the difference between the ENML-PAFE and the ENTD-PAFE varies across out-of-sample periods.

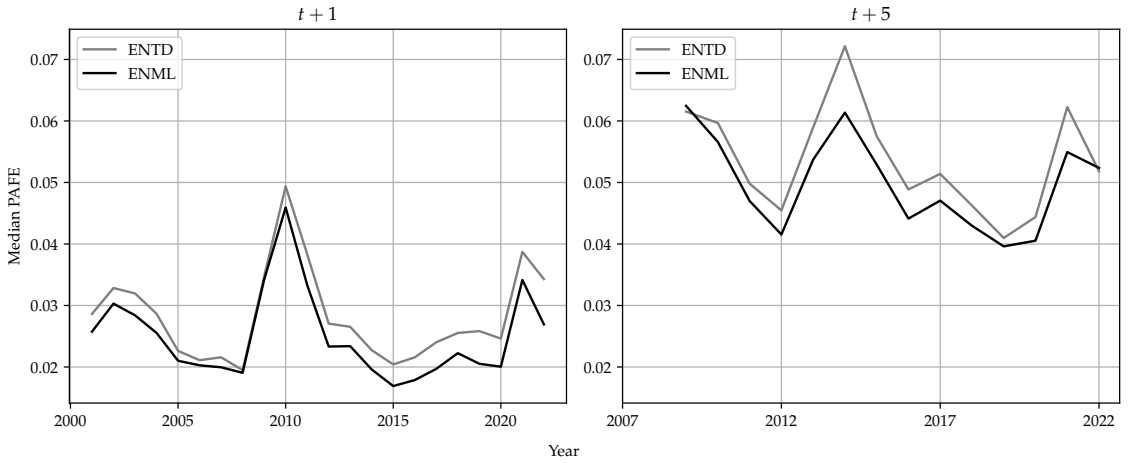



**Figure 3.1: PAFE across out-of-sample periods**
This figure shows the median price scaled absolute forecast errors (PAFEs) of the machine-learning ensemble (ENML) and the traditional ensemble (ENTD) per out-of-sample period for forecast horizons $t+1$ and $t+5$.

As a robustness check, we again stratify our predictions into those for small firms and those for large firms and report the results in Table 3.4. Consistent with the existing literature, we find that accuracy is generally much higher for large firms than for small firms (Li and Mohanram, 2014). This holds true for all models considered. Furthermore, the accuracy superiority of the machine learning ensemble over its traditional analogue is more pronounced for the small firm sample. For small firms, the accuracy difference ranges from 14.38% for $t+1$ predictions to 11.88% for $t+5$ predictions. Nonetheless, for large firms, the difference still ranges from 13.16% for $t+1$ predictions to 2.65% for $t+5$ predictions. Moreover, the difference in accuracy is statistically significant for all forecast horizons in both firm samples. Lastly, similar to the full sample case, we again observe that the traditional ensemble performs slightly worse than some of its component models for large forecast horizons.

### Diebold and Mariano forecast comparison

Table 3.5 shows the pairwise comparisons of the models using the aforementioned modified Diebold and Mariano (1995) test statistic. We restrict this analysis to predictions

**Table 3.4: Median PAFE by firm size**

| | Large firms | | | | |
|---|---|---|---|---|---|
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | 0.0201*** | 0.0266*** | 0.0300*** | 0.0346*** | 0.0390*** |
| $HVZ^{MSE}$ | 0.0200*** | 0.0266*** | 0.0304*** | 0.0350*** | 0.0399*** |
| $EP^{MSE}$ | 0.0193*** | 0.0258*** | 0.0292*** | 0.0336*** | 0.0381*** |
| $RI^{MSE}$ | 0.0190*** | 0.0253*** | 0.0287*** | 0.0331*** | 0.0381*** |
| ENTD | 0.0190*** | 0.0252*** | 0.0287*** | 0.033*** | 0.0377*** |
| $RF^{MSE}$ | 0.0182*** | 0.0245*** | 0.0285*** | 0.0329*** | 0.0373*** |
| $GBT^{MAE}$ | 0.0180*** | 0.0246*** | 0.0291*** | 0.0332*** | 0.0377*** |
| $DART^{MAE}$ | 0.0168*** | 0.0239*** | 0.0285*** | 0.0337*** | 0.0390*** |
| $NN^{MAE}$ | 0.0169*** | 0.0250*** | 0.0300*** | 0.0330*** | 0.0389*** |
| ENML | 0.0165*** | 0.0231*** | 0.0277*** | 0.0320*** | 0.0367*** |
| ENML - ENTD | −0.0025*** | −0.0021*** | −0.0009** | −0.0010*** | −0.0010* |
| | Small firms | | | | |
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | 0.0583*** | 0.0724*** | 0.0802*** | 0.0898*** | 0.1010*** |
| $HVZ^{MSE}$ | 0.0474*** | 0.0575*** | 0.0645*** | 0.0742*** | 0.0854*** |
| $EP^{MSE}$ | 0.0513*** | 0.0593*** | 0.0631*** | 0.0695*** | 0.0780*** |
| $RI^{MSE}$ | 0.0506*** | 0.0580*** | 0.0602*** | 0.0649*** | 0.0734*** |
| ENTD | 0.0466*** | 0.0554*** | 0.0613*** | 0.0693*** | 0.0791*** |
| $RF^{MSE}$ | 0.0443*** | 0.0527*** | 0.0594*** | 0.0677*** | 0.0764*** |
| $GBT^{MAE}$ | 0.0444*** | 0.0545*** | 0.0598*** | 0.0656*** | 0.0728*** |
| $DART^{MAE}$ | 0.0408*** | 0.0512*** | 0.0588*** | 0.0662*** | 0.0747*** |
| $NN^{MAE}$ | 0.0416*** | 0.0555*** | 0.0590*** | 0.0641*** | 0.0733*** |
| ENML | 0.0399*** | 0.0503*** | 0.0559*** | 0.0626*** | 0.0697*** |
| ENML - ENTD | −0.0067*** | −0.0051*** | −0.0054*** | −0.0067*** | −0.0094*** |

This table reports the time-series averages of the median price scaled absolute forecasting errors (PAFEs) for all models, stratified by firm size. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the PAFE for the $t+1$ horizon). The PAFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

for earnings in $t+1$ and exclusively use the PAFE. This is because pairwise comparisons using the PFE do not yield interpretable results, as a higher (lower) PFE of a model compared with another model may indicate both better or worse performance

of the respective model. A positive statistic indicates the column model outperforms the respective row model. The results confirm the finding that the machine learning approaches outperform their traditional counterparts. Moreover, the ENML is again the best performing model, beating every other one except for the DART model.[13]

### 3.4.2 Out-of-sample $R^2$

The next metric considered is the out-of-sample $R^2$ ($OOS\ R^2$). The results are reported in Table 3.6. The $OOS\ R^2$ allows us to assess how the models perform in terms of explaining out-of-sample variation in future earnings. As expected, the $OOS\ R^2$ decreases with increasing forecast horizon. Moreover, the results generally confirm the positive effect of model stacking. Our analysis demonstrates that the machine learning ensemble consistently outperforms its individual components across all forecast horizons considered. In contrast, there are instances in which the traditional ensemble exhibits slightly lower performance than the RI model for specific forecast horizons, i.e., $t+1$ and $t+2$. Nevertheless, the traditional ensemble consistently outperforms all of its component models for the remaining forecast horizons.

Furthermore, our findings validate the notion that machine learning approaches surpass traditional linear models in terms of predictive performance. To be more specific, when we compare the $OOS\ R^2$ of the machine learning models, including the ensemble, against that of the traditional models, including the ensemble, we find that the machine learning models outperform the traditional ones in almost every case. In fact, just assessing the best-performing models, i.e., the ensembles, we find that the machine learning ensemble beats its traditional counterpart for every forecast horizon. The difference in $OOS\ R^2$ between the ensemble models is statistically significant at the 1% level for forecast horizons $t+1$ to $t+4$. Further, while the relative PAFE difference between the ensembles decreases with increasing forecast horizon, the difference in $OOS\ R^2$ increases from 14.03% for $t+1$ predictions to 18.72% for $t+5$ predictions.

Figure 3.2 plots the $OOS\ R^2$ for the ENML and ENTD for each year and for $t+1$ and $t+5$, respectively. Again, it is evident that the machine learning ensemble outperforms the traditional ensemble in the majority of years and for both forecast horizons.

[13] Note that we adjust our p-values very conservatively via a multiple comparisons Bonferroni correction. This significantly increases the hurdles for reaching significance, hence possibly explains why the test statistic is not statistically significant.

**Table 3.5: Pairwise Diebold-Mariano test statistics**

| | $L^{MSE}$ | $HVZ^{MSE}$ | $EP^{MSE}$ | $RI^{MSE}$ | ENTD | $RF^{MSE}$ | $GBT^{MAE}$ | $DART^{MAE}$ | $NN^{MAE}$ | ENML |
|---|---|---|---|---|---|---|---|---|---|---|
| $L^{MSE}$ | | 0.0107*** | 0.0098*** | 0.0104*** | 0.0123*** | 0.0168*** | 0.0149*** | 0.0182*** | 0.0170*** | 0.0192*** |
| $HVZ^{MSE}$ | | | −0.0009 | −0.0003 | 0.0016 | 0.0061*** | 0.0042*** | 0.0075*** | 0.0062*** | 0.0085*** |
| $EP^{MSE}$ | | | | 0.0006 | 0.0024** | 0.0070*** | 0.0051*** | 0.0084*** | 0.0071*** | 0.0094*** |
| $RI^{MSE}$ | | | | | 0.0018** | 0.0064*** | 0.0045*** | 0.0078*** | 0.0065*** | 0.0088*** |
| ENTD | | | | | | 0.0045*** | 0.0027** | 0.0060*** | 0.0047*** | 0.0069*** |
| $RF^{MSE}$ | | | | | | | −0.0019** | 0.0014* | 0.0002 | 0.0024*** |
| $GBT^{MAE}$ | | | | | | | | 0.0033*** | 0.0020*** | 0.0043*** |
| $DART^{MAE}$ | | | | | | | | | −0.0013* | 0.0010 |
| $NN^{MAE}$ | | | | | | | | | | 0.0023*** |
| ENML | | | | | | | | | | |

This table reports the Diebold-Mariano (Diebold and Mariano, 1995) statistics for each model comparison. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). Positive numbers indicate the column model outperforms the respective row model. ***, **, and * denote the Bonferroni-adjusted significance levels at 10%, 5% and 1%, respectively. Standard errors used to derive statistical significance are adjusted following Newey and West (1987) assuming a lag length of three. P-values are adjusted by applying the Bonferroni procedure.

**Table 3.6: Average out-of-sample $R^2$**

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| $L^{MSE}$ | 0.3959 | 0.2511 | 0.1951 | 0.1251 | 0.0535 |
| $HVZ^{MSE}$ | 0.4495 | 0.3295 | 0.2662 | 0.1789 | 0.0899 |
| $EP^{MSE}$ | 0.4450 | 0.3219 | 0.2581 | 0.1650 | 0.0805 |
| $RI^{MSE}$ | 0.4589 | 0.3439 | 0.2813 | 0.1883 | 0.0966 |
| ENTD | 0.4519 | 0.3382 | 0.2856 | 0.2056 | 0.1298 |
| $RF^{MSE}$ | 0.4971 | 0.3755 | 0.3165 | 0.2237 | 0.1390 |
| $GBT^{MAE}$ | 0.4858 | 0.3498 | 0.2774 | 0.2089 | 0.1135 |
| $DART^{MAE}$ | 0.5091 | 0.3778 | 0.3057 | 0.2085 | 0.1144 |
| $NN^{MAE}$ | 0.4992 | 0.2075 | 0.2631 | 0.2045 | 0.0736 |
| ENML | 0.5153 | 0.3785 | 0.3200 | 0.2390 | 0.1541 |
| ENML - ENTD | 0.0634*** | 0.0403** | 0.0345*** | 0.0332*** | 0.0243 |

This table reports the time-series averages of the out-of-sample $R^2$s (*OOS* $R^2$s) for all models. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). ***, **, and * denote statistical significance at the 10%, the 5% and the 1% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. We only test for statistical significance of the difference between the ensemble models (ENML - ENTD).

Remarkably, the *OOS* $R^2$ of both ensemble models exhibits a noticeable dip in 2009, particularly pronounced in forecasts for $t+5$. Notably, for $t+5$ forecasts, it takes some years for the *OOS* $R^2$ to rebound. This result underscores the delayed integration of new information, such as the financial crisis in this case, into longer-term forecasts. Such delayed adaptation is inherent in the rolling window approach we employ.

Again, for robustness, we stratify our predictions into those for small firms and large firms, respectively and report the results in Table 3.7.[14] Consistent with our prior findings, we observe notable disparities in the *OOS* $R^2$ performance between large and small firms, with the former exhibiting higher predictive accuracy. Furthermore, our analysis demonstrates that the ENML model consistently outperforms the ENTD model across all forecast horizons and for both subsamples. In line with the overarching patterns evident in our prior results, we find that the differential in performance between the ENML and ENTD models is more pronounced for small firms. Specifically, our findings indicate

[14]Note that the *OOS* $R^2$s of both subsamples are lower than the total *OOS* $R^2$. This is because the *OOS* $R^2$ is a non-linear function.

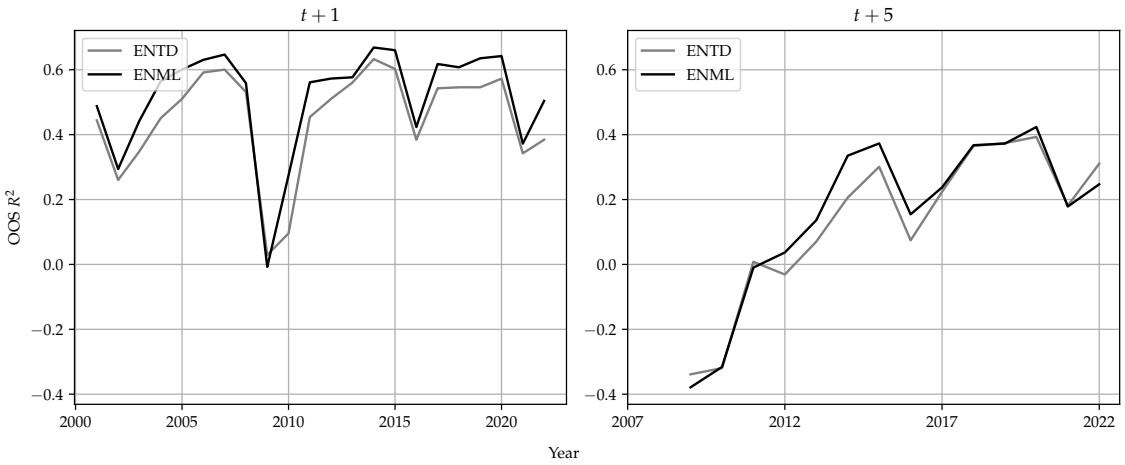



**Figure 3.2: $R^2$ across out-of-sample periods**
This figure shows the OOS $R^2$ of the machine-learning ensemble (ENML) and the traditional ensemble (ENTD) per out-of-sample period for forecast horizons $t+1$ and $t+5$.

that the ENML model exhibits a 15.16% higher *OOS* $R^2$ for large firms compared to the 23.37% higher *OOS* $R^2$ for small firms for $t+1$ predictions. This pronounced discrepancy reaffirms the consistency of our earlier observations and substantiates the argument that a traditional linear model is inadequate for capturing the intricate nuances in future earnings, particularly for smaller firms. Interestingly, the relative outperformance of the ENML in terms of *OOS* $R^2$ for small firms strictly increases with increasing forecast horizon to 35.00% for $t+5$. In the large firm sample, the relative outperformance of the ENML as compared to the ENTD also increases to 45.21% for $t+5$ predictions. The *OOS* $R^2$ difference between the ENML and the ENTD is statistically significant for most forecast horizons in both subsamples.

### 3.4.3 Implied cost of capital

A common application of earnings forecasts is the derivation of the implied cost of capital (ICC), for which earnings predictions serve as a crucial input. We restrict this analysis to the best performing traditional earnings forecast model, i.e., the traditional ensemble, as well as the best performing machine learning earnings forecast model, i.e., the machine learning ensemble.

As typically done in the literature, we evaluate the return expectations in form of ICC by evaluating the performance of long-short portfolios sorted on ICC (e.g., Hou et al., 2012; Li and Mohanram, 2014). More specifically, we sort the stocks into deciles according to ICC in $t$ and assess the realized annual geometric average return of a buy-and-hold

**Table 3.7: Average out-of-sample $R^2$ by firm size**

| | Large firms | | | | |
|---|---|---|---|---|---|
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | 0.3808 | 0.2247 | 0.1694 | 0.0999 | 0.0203 |
| $HVZ^{MSE}$ | 0.4360 | 0.3032 | 0.2379 | 0.1495 | 0.0521 |
| $EP^{MSE}$ | 0.4285 | 0.2885 | 0.2144 | 0.1090 | 0.0050 |
| $RI^{MSE}$ | 0.4455 | 0.3131 | 0.2394 | 0.1325 | 0.0225 |
| ENTD | 0.4366 | 0.3073 | 0.2484 | 0.1615 | 0.0712 |
| $RF^{MSE}$ | 0.4781 | 0.3412 | 0.2822 | 0.1877 | 0.0968 |
| $GBT^{MAE}$ | 0.4707 | 0.3167 | 0.2379 | 0.1631 | 0.0575 |
| $DART^{MAE}$ | 0.4972 | 0.3468 | 0.2628 | 0.1559 | 0.0524 |
| $NN^{MAE}$ | 0.4872 | 0.1570 | 0.2127 | 0.1558 | 0.0074 |
| ENML | 0.5028 | 0.3474 | 0.2825 | 0.1962 | 0.1034 |
| ENML - ENTD | 0.0662*** | 0.0400** | 0.0341*** | 0.0347*** | 0.0322 |
| | Small firms | | | | |
| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
| $L^{MSE}$ | 0.2585 | 0.0967 | 0.0224 | −0.0514 | −0.1111 |
| $HVZ^{MSE}$ | 0.3255 | 0.2042 | 0.1250 | 0.0320 | −0.0439 |
| $EP^{MSE}$ | 0.3243 | 0.2110 | 0.1500 | 0.0728 | 0.0246 |
| $RI^{MSE}$ | 0.3372 | 0.2362 | 0.1779 | 0.1042 | 0.0449 |
| ENTD | 0.3322 | 0.2261 | 0.1710 | 0.0991 | 0.0516 |
| $RF^{MSE}$ | 0.3983 | 0.2831 | 0.2044 | 0.1047 | 0.0311 |
| $GBT^{MAE}$ | 0.3761 | 0.2503 | 0.1654 | 0.1097 | 0.0306 |
| $DART^{MAE}$ | 0.4011 | 0.2808 | 0.2099 | 0.1227 | 0.0447 |
| $NN^{MAE}$ | 0.3861 | 0.1182 | 0.1703 | 0.1071 | −0.0028 |
| ENML | 0.4098 | 0.2832 | 0.2169 | 0.1405 | 0.0697 |
| ENML - ENTD | 0.0776*** | 0.0571*** | 0.0459*** | 0.0414*** | 0.0181 |

This table reports the time-series averages of the out-of-sample $R^2$s (*OOS* $R^2$s) for all models, stratified by firm size. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript $MAE$ ($MSE$) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. We only test for statistical significance of the difference between the ensemble models (ENML - ENTD).

portfolio that goes long the highest ICC decile and short the lowest ICC decile in $t$.

Table 3.8 shows that long-short ICC portfolios based on the ENML predictions outperform those based on the ENTD predictions for each buy-and-hold period length

considered. A buy-and-hold long-short ICC portfolio based on ENML predictions yields (geometric) average annual returns of 8.45%, 9.14% and 6.80% for buy-and-hold periods of one, two and three years, on average. The linear analogue yields corresponding average annual returns of 7.15%, 7.75% and 5.89%. Moreover, while the portfolio returns based on the ENML model are statistically significant for every buy-and-hold period length, the portfolio returns based on the ENTD model are only statistically significant when holding stocks for three years. We conclude that the improved accuracy of machine learning predictions translates into more profitable investment strategies, thereby stressing the practical importance of earnings prediction accuracy.

**Table 3.8: Long-short ICC portfolio performance**

| | | ENTD | | ENML | |
|---|---|---|---|---|---|
| $\kappa$ | Decile | ICC | Realized | ICC | Realized |
| | 1 | 0.0301 | 0.0809 | 0.0249 | 0.0660 |
| 1 | 10 | 0.2609 | 0.1524 | 0.2631 | 0.1505 |
| | 10-1 | 0.2308 | 0.0715 | 0.2382 | 0.0845* |
| | 1 | 0.0301 | 0.0899 | 0.0249 | 0.0819 |
| 2 | 10 | 0.2609 | 0.1673 | 0.2631 | 0.1732 |
| | 10-1 | 0.2308 | 0.0775 | 0.2382 | 0.0914** |
| | 1 | 0.0301 | 0.0990 | 0.0249 | 0.0933 |
| 3 | 10 | 0.2609 | 0.1583 | 0.2631 | 0.1621 |
| | 10-1 | 0.2308 | 0.0589* | 0.2382 | 0.0680** |

This table reports the average implied cost of capital (ICC) estimates as well as the average realized annual returns of buy-and-hold portfolios sorted conditionally on them for buy-and-hold periods of $\kappa \in [1, 2, 3]$ years as indicated by the first column. Annual realized returns are derived as the geometric average of portfolio returns over the respective buy-and-hold period. ENTD refers to the traditional ensemble and ENML refers to the machine learning ensemble. The 10-1 rows denote the long-short portfolios, in which an investor goes short the lowest ICC decile and long the highest ICC decile, equally weighting stocks in each decile. We test for significance of the realized buy-and-hold return of this portfolio. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

## 3.5 Interpretation

### 3.5.1 Variable importance

As outlined above, we assess the degree to which financial statement variables matter in the machine learning ensemble by computing their SHAP values.[15] More precisely, we compute SHAP values per out-of-sample period and derive their respective averages for each variable.

[15]We focus on the ensemble model as it is the best performing machine learning approach.

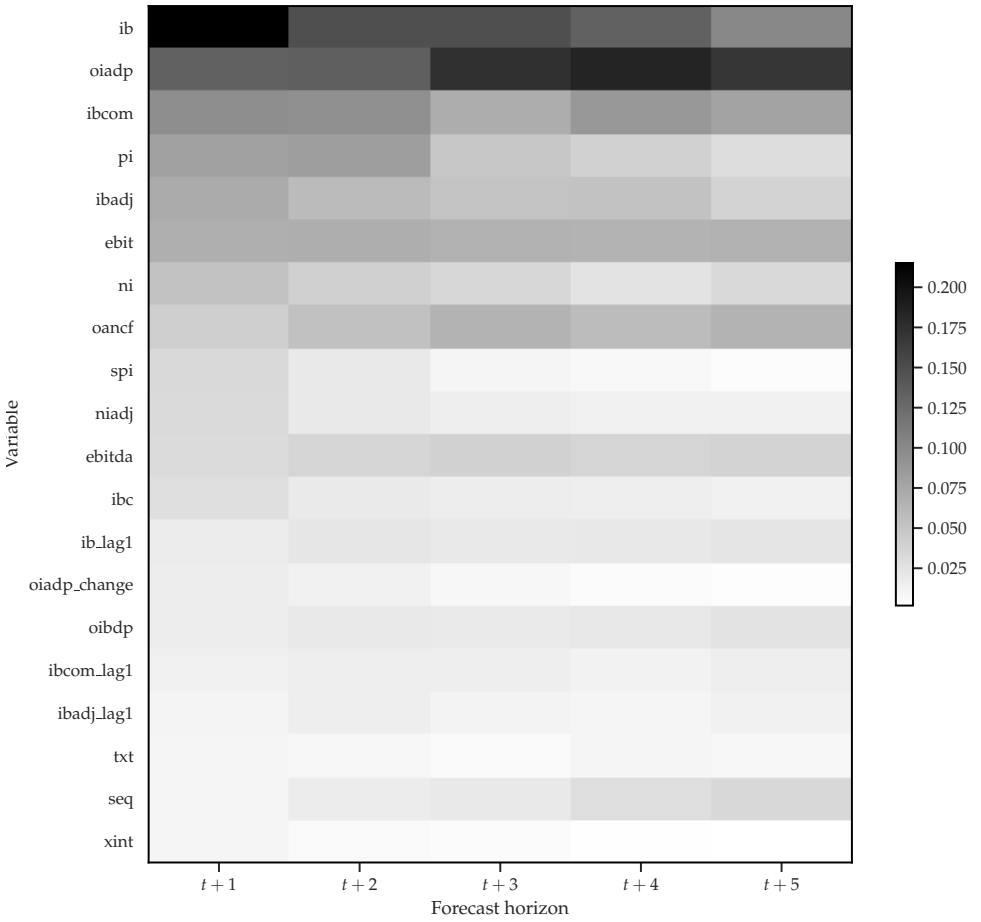



**Figure 3.3: Variable importance for the machine learning forecast ensemble**
This figure depicts the absolute SHAP values of the 20 most important variables for the machine learning ensemble, averaged over out-of-sample periods and scaled so they sum up to one within each forecast horizon. In this context, importance is defined as the ranking of the respective variable according to the aforementioned metric for forecast horizon $t + 1$.

Figure 3.3 shows the mean absolute SHAP values of the ENML, averaged over all out-of-sample periods and scaled so that variable importance per forecast horizon sums to one. We show the twenty most important variables for predicting earnings in $t + 1$ and sort them according to their importance.[16] The higher the SHAP value, the more important the variable. For $t + 1$ predictions, *ib*, i.e., current earnings, is the most important variable by far.[17] This comes as no surprise, considering that a simple model including only current earnings as a predictor performs comparably well in predicting future earnings.

[16]The variable definitions are provided in B.3

[17]Note that this is the earnings definition that we use as our target variable. Further note that all of our variables are scaled by common shares outstanding. Thus, strictly speaking, we refer to earnings per share when talking about *ib*.

A striking finding is that the remaining 19 most important variables are primarily different definitions of earnings. For example, the second most important variable *oiadp* resembles "operating income after depreciation" and the third most important variable *ibcom* resembles "earnings before extraordinary items - available for common". The only top-twenty variables that do not originate from the income statement are *oancf* and *seq*. However, they resemble comparably low importance.

In general, six to seven variables dominate across forecast horizons $t+1$ to $t+5$. Another finding regarding the different forecast horizons is that the significance of *ib* gradually diminishes with increasing horizon. Instead, *oiadp*, another earnings variable, emerges as the most important variable. Moreover, the top-twenty variables not stemming from the income statement, i.e., *oancf* and *seq*, become increasingly important with increasing forecast horizon.

Lastly, the results suggest that current data is more important than lagged data or first order differences. We revisit this claim below.

We now explore whether the most important predictor variables act as substitutes or complements. This assessment is conducted by examining the absolute Pearson correlation coefficients of the top-twenty variables reported in Figure 3.4. If the variables are substitutes, one would expect high coefficient values. In general, we find mixed results. *ibcom*, *pi*, *ibadj*, *ni* and *niadj* are correlated quite strongly with *ib* and each other. All of these are variations of earnings definitions which do not differ strongly from each other. This observation leads us to consider these variables as substitutes for *ib*, indicating that they do not necessarily possess significant independent predictive power. Other variables, that are not as closely related to *ib*, either because they explicitly exclude major income statement items, such as *oiadp*, or because they are not income statement items at all, such as *oancf*, do not correlate as strongly with earnings. We interpret this as evidence that these items are complements to *ib* and hence possess stand-alone predictive power.

### 3.5.2 Group importance

Table 3.9 reports variable importance per group. More precisely, we group the variables according to the financial statement they originate from (Panel A), whether they are current, lagged or change information (Panel B), and according to the two aforementioned categories (Panel C). Grouping the variables according to the financial statement

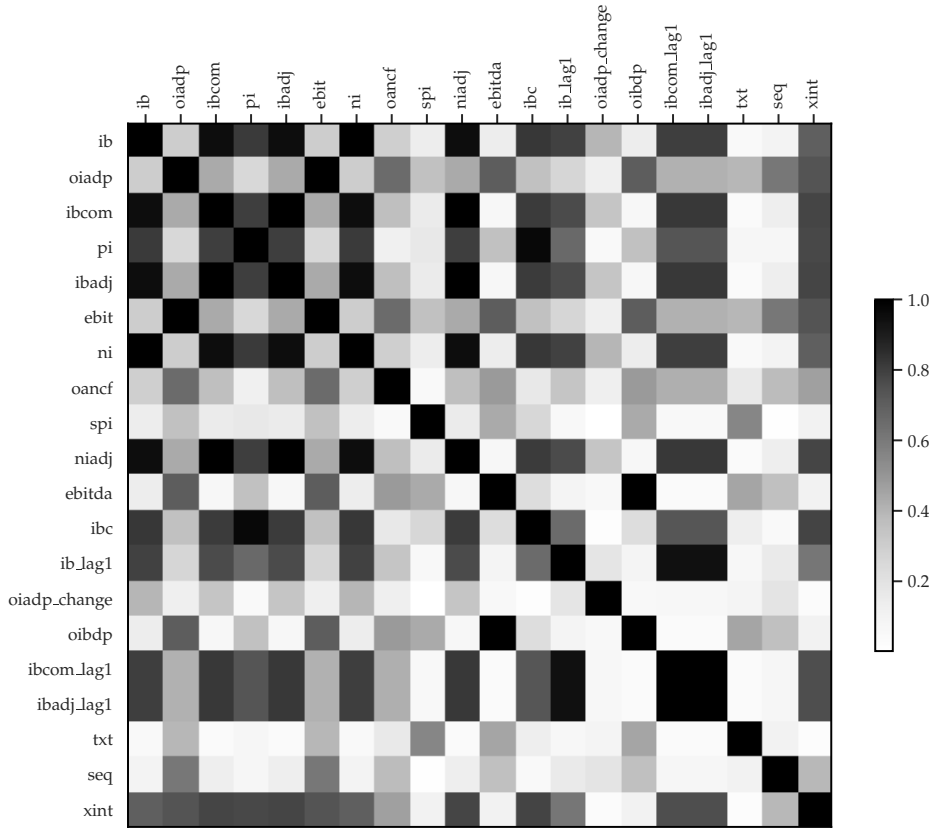



**Figure 3.4: Correlation heatmap for the most important variables**
This figure shows the absolute Pearson correlation coefficients for the 20 most important variables for the machine learning ensemble.

they originate from reveals that income statement (I/S) variables are the most important variables. On average, for one-year-ahead predictions, I/S variables contribute approximately 65% to the total importance, while balance sheet (B/S) variables and cash flow statement (CF/S) variables contribute around 20% and 15%, respectively. This finding aligns with the analysis of the most important variables, indicating that I/S variables significantly outweigh others in importance for predicting earnings. However, we also find that I/S variables become less important with increasing forecast horizon. In fact, the importance of I/S variables decreases to around 47% for $t+5$ forecasts. In contrast, B/S variables become more important with increasing forecast horizon (around 37% for $t+5$ predictions) while CF/S variables stay at a constant level.

Grouping variables according to whether they are current, lagged or difference variables in Panel B reveals that current data is by far the most important group out of these categories, contributing around 71% of total importance for $t+1$ predictions. Lagged and difference data each contribute around $14-15\%$ to total importance for

**Table 3.9: Variable importance per financial statement × variable type group**

Panel A: Financial statement type

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| B/S | 0.2023 | 0.2497 | 0.3010 | 0.3437 | 0.3733 |
| CF/S | 0.1517 | 0.1620 | 0.1619 | 0.1557 | 0.1545 |
| I/S | 0.6460 | 0.5883 | 0.5371 | 0.5006 | 0.4723 |

Panel B: Variable type

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| Current | 0.7123 | 0.6701 | 0.6506 | 0.6462 | 0.6228 |
| Lagged | 0.1422 | 0.1933 | 0.2109 | 0.2164 | 0.2472 |
| Change | 0.1455 | 0.1366 | 0.1385 | 0.1374 | 0.1299 |

Panel C: Financial statement type × variable type

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| B/S current | 0.0923 | 0.1196 | 0.1493 | 0.1767 | 0.2034 |
| B/S lagged | 0.0443 | 0.0654 | 0.0848 | 0.0973 | 0.1027 |
| B/S change | 0.0657 | 0.0648 | 0.0669 | 0.0697 | 0.0671 |
| CF/S current | 0.0942 | 0.0958 | 0.0928 | 0.0856 | 0.0828 |
| CF/S lagged | 0.0293 | 0.0392 | 0.0408 | 0.0401 | 0.0446 |
| CF/S change | 0.0282 | 0.0269 | 0.0284 | 0.0300 | 0.0271 |
| I/S current | 0.5258 | 0.4547 | 0.4086 | 0.3840 | 0.3366 |
| I/S lagged | 0.0687 | 0.0887 | 0.0854 | 0.0790 | 0.0999 |
| I/S change | 0.0516 | 0.0449 | 0.0432 | 0.0377 | 0.0357 |

Panel A reports the relative variable importance per financial statement group. B/S , CF/S and I/S denote balance sheet, cash flow statement and income statement, respectively. The variables are grouped according to Table B.3 in the Appendix. Panel B reports the relative variable importance per variable type. Panel C reports the relative variable importance per financial statement type × variable type group. Importance per group in each panel is defined as the fraction that the respective group contributes to total importance, measured as the sum of absolute SHAP values.

$t+1$ predictions. Again, importance becomes more evenly distributed among the groups with increasing forecast horizon. More precisely, current variables become less important while lagged variables become more important. This might be the case because short-term forecasts are heavily influenced by current information due to their sensitivity to recent developments. Longer-term forecasts benefit from a combination of current and lagged information to capture the interplay of short-term dynamics and longer-term trends.

Further breaking down the groups according to the two aforementioned categories stresses the findings above. Overall, current I/S variables contribute around 53% to total importance for $t+1$ forecasts and hence represent the most important group of variables by a significant margin. This is intuitive and supports the finding that simple earnings forecasts models only considering current earnings items, like the L model or the EP model, perform comparably well in predicting future earnings.

We conclude our variable importance assessment by more thoroughly analyzing

the variable importance per financial statement type in Table 3.10. More precisely, we assess the importance of each financial statement type per schematic financial statement component, such as e.g., current assets, fixed assets or equity, in the B/S case. This analysis provides an intuitive accounting perspective on which components of financial statements are important and how the importance might change across forecast horizons.

**Table 3.10: Variable importance per financial statement component**

Panel A: Balance sheet

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| Current assets | 0.0302 | 0.0372 | 0.0445 | 0.0542 | 0.0589 |
| Fixed assets | 0.0214 | 0.0243 | 0.0259 | 0.0300 | 0.0326 |
| Total assets | 0.0046 | 0.0076 | 0.0105 | 0.0118 | 0.0177 |
| Debt | 0.0513 | 0.0604 | 0.0709 | 0.0828 | 0.0925 |
| Equity | 0.0367 | 0.0474 | 0.0558 | 0.0613 | 0.0610 |
| Total debt & equity | 0.0028 | 0.0034 | 0.0052 | 0.0074 | 0.0077 |
| Supplemental | 0.0553 | 0.0695 | 0.0881 | 0.0963 | 0.1028 |
| Sum B/S importance | 0.2023 | 0.2497 | 0.3010 | 0.3437 | 0.3733 |

Panel B: Cash flow statement

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| Operating cash flow | 0.1083 | 0.1087 | 0.1049 | 0.0999 | 0.1005 |
| Investing cash flow | 0.0146 | 0.0188 | 0.0204 | 0.0190 | 0.0214 |
| Financing cash flow | 0.0245 | 0.0297 | 0.0309 | 0.0317 | 0.0284 |
| Total cash flow | 0.0042 | 0.0047 | 0.0057 | 0.0051 | 0.0043 |
| Sum CF/S importance | 0.1517 | 0.1620 | 0.1619 | 0.1557 | 0.1545 |

Panel C: Income statement

| | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|
| Sales | 0.0061 | 0.0102 | 0.0135 | 0.0140 | 0.0167 |
| Operating expenses | 0.0121 | 0.0163 | 0.0199 | 0.0245 | 0.0261 |
| EBITDA | 0.0335 | 0.0361 | 0.0350 | 0.0322 | 0.0347 |
| Depr. & Amort. | 0.0078 | 0.0066 | 0.0070 | 0.0089 | 0.0123 |
| EBIT | 0.1355 | 0.1355 | 0.1414 | 0.1326 | 0.1279 |
| Interest expenses | 0.0600 | 0.0516 | 0.0453 | 0.0393 | 0.0377 |
| EBT | 0.0517 | 0.0514 | 0.0308 | 0.0237 | 0.0178 |
| Tax expenses | 0.0224 | 0.0259 | 0.0262 | 0.0297 | 0.0306 |
| Net income | 0.3063 | 0.2403 | 0.2006 | 0.1812 | 0.1555 |
| Dividends | 0.0106 | 0.0145 | 0.0174 | 0.0146 | 0.0128 |
| Sum I/S importance | 0.6460 | 0.5883 | 0.5371 | 0.5006 | 0.4723 |

This table reports the relative variable importance per financial statement group. B/S , CF/S and I/S denote balance sheet, cash flow statement and income statement, respectively. EBITDA denotes earnings before interest, taxes and depreciation and amortization. EBIT denotes earnings before interest and taxes. EBT denotes earnings before taxes. The variables are grouped according to Table B.3 in the Appendix. Importance per financial statement component is defined as the fraction that the respective component contributes to total importance, measured as the sum of absolute SHAP values.

The table provides several key insights: first, assessing the B/S, we find that the debt and supplemental items are the most important pieces of B/S information, with both contributing around 5-6% to total importance for $t+1$ forecasts. Moreover, all pieces of

B/S information consistently increase in importance with increasing forecast horizon.

Second, variables associated with the operating cash flow resemble the most important category of CF/S variables. This comes as no surprise, since the operating cash flow closely relates to earnings. The relevance of the investing cash flow slightly increases with increasing forecast horizon. However, overall, the differences are minor.

Third, different definitions of earnings, i.e., EBITDA, EBIT, EBT and net income, resemble the most important I/S categories. Out of these categories, EBIT and net income are the most important categories with around 14% and 31% share in total importance, respectively. Interestingly, the importance of the EBIT stays somewhat constant throughout forecast horizons, whereas the importance of the net income consistently declines with increasing forecast horizon to around 16% for $t + 5$ forecasts. This dynamic might be attributable to the fact that net income is more strongly exposed to accounting manipulation than EBIT and hence less reliable in the long-term. Moreover, we find that while sales contribute very little overall, they consistently increase in importance with increasing forecast horizon. This supports the notion that items which are less exposed to discretionary accounting gain predictive value when considering longer forecast horizons. Revisiting the aforementioned finding that operating cash flow variables maintain consistent importance across forecast horizons further reinforces this notion. Unlike earnings, cash flows include no discretionary accrual items and are hence not exposed to earnings management (e.g., Jones, 1991). Consequently, their predictive value does not decrease for longer-term forecasts.

In summary, these findings suggest that the variations in importance across forecast horizons are primarily driven by the presence of earnings management. Future research endeavors could offer additional insights into these dynamics.

### 3.5.3 Non-linearity

We now shed light on the degree to which non-linearity of the functional form and non-linearity of variables, i.e., interactions among financial statement variables considered, play a role in predicting earnings.

**Surrogate model**

We find that for our flexible ENML approach, around 89% of the variation in predicted earnings can be explained by a linear surrogate model for $t + 1$ predictions on average. This does not change when including two-way interactions, suggesting that interaction

effects play virtually no role in predicting earnings using raw financial input data. Interestingly, this stands in stark contrast to the results by Jones et al. (2023) who find that interactions among the ratios which they use as predictors contribute substantially to the prediction. Apart from the fact that our target variables differ, this might be attributable to the fact that we use a significantly larger set of inputs, which might directly capture interactions among variables of a more limited set of variables. The remaining unpredictable portion of the ENML can be attributed to non-linearity of the functional form.[18]

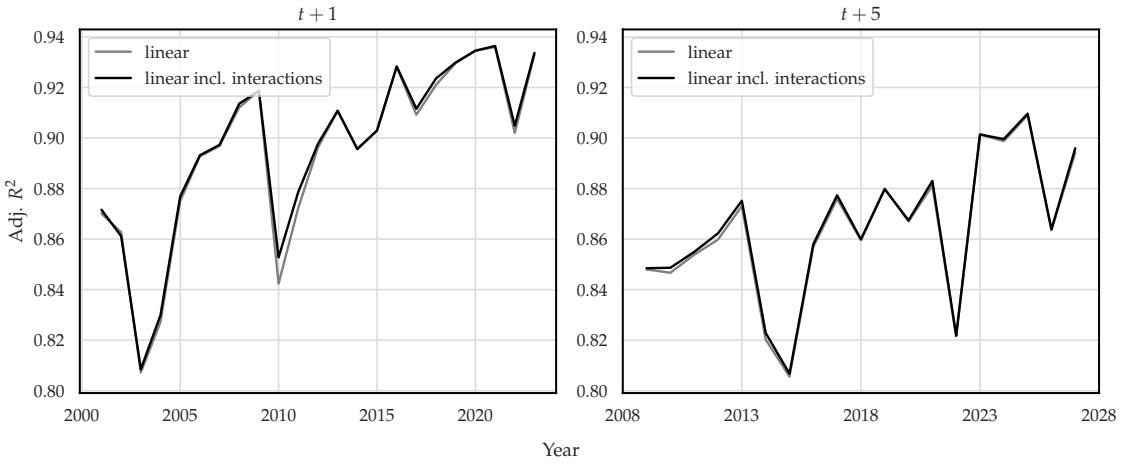



**Figure 3.5: Surrogate models**

This figure shows the adjusted $R^2$ of the surrogate models that we fit to our machine learning ensemble predictions for forecast horizons $t+1$ and $t+5$. The linear model (linear) is a simple linear model in which we regress the respective predictions on the 50 most important predictor variables according to their average absolute SHAP values for $t+1$ forecasts. The linear model including interactions (linear incl. interactions) is a linear model in which we use the same set of predictors as well as all possible two-way interactions.

We depict this graphically in Figure 3.5. The figure shows the adjusted in-sample $R^2$ per out-of-sample period, derived by regressing predicted earnings on a linear surrogate model and a linear surrogate model including two-way interactions. The figure also includes the surrogate model for predictions for $t+5$. In fact, assessing the surrogate model for $t+5$ reveals that even for longer forecast horizons, interaction effects across financial statement variables do not play a role. With increasing horizon, a slightly larger portion of the earnings-predictor relation can be attributed to non-linear functional form. Nonetheless, the linear surrogate model still explains around 87% of the predictions for $t+5$ on average.

[18]Theoretically, it can also be attributed to higher-order interactions. However, in undocumented results, we find that this is not the case.

### Partial dependence plots

We now turn to how the aforementioned degree of non-linearity is expressed at the variable-level. Figure 3.6 shows the partial dependence plots for *ib*, *oiadp*, *ibcom* and *oancf* for all forecast horizons.[19]

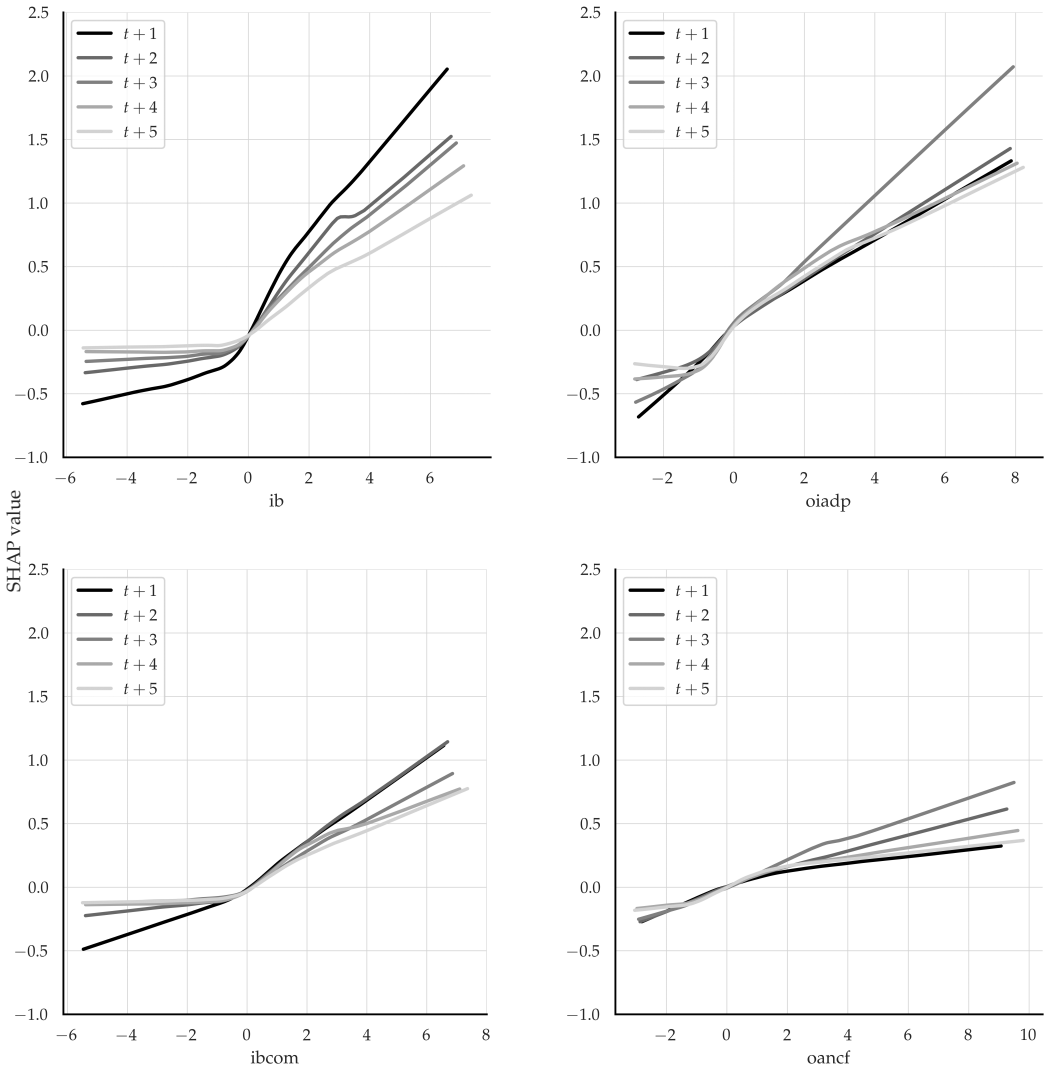



**Figure 3.6: Partial dependence plots**
The panels show the sensitivity of expected future earnings to the respective variable for all forecast horizons. More specifically, we fit a nonparametric lowess model (locally weighted linear regression) to the SHAP values of the respective variable.

The partial dependence measures the sensitivity of the predicted earnings to the individual financial statement variables. The upper-left panel shows that the *ib* effect is the strongest, i.e., the steepest. Remarkably, the sensitivity appears to be linear for

[19]We plot the partial dependence of the three most important variables and the most important variable not stemming from the income statement.

both positive and negative values of *ib*. However, there is a distinction in the slope of the line for positive and negative values, suggesting varying sensitivities of future earnings to current earnings for profit and loss firms, respectively. This may explain why the EP model and the RI model by Li and Mohanram (2014) yield comparably good forecasting results, especially for short forecast horizons. The two models include a dummy for negative earnings and the interaction between earnings and the negative earnings dummy, which essentially allows for different slopes of *ib* for profit and loss firms.

For *ibcom* in the lower-left panel we find a similar trend as for *ib*. For $oiadp$, there is also a difference in slopes between profit and loss firms, especially for longer forecast horizons. However, the kink appears slightly below zero. In contrast $oancf$ is essentially linearly related to future earnings across all forecast horizons.

## 3.6 Conclusion

We show that earnings per share predictions based on state-of-the-art machine learning approaches using high-dimensional financial statement data are more accurate than those based on traditional linear approaches. These improvements hold across all evaluation metrics assessed, i.e., commonly used error metrics, the *OOS* $R^2$ as well as the performance of long-short ICC portfolios based on the predictions.

Importantly, we provide an intuitive breakdown of how important the different pieces of fundamental accounting information are for predicting earnings. We find that current I/S variables, especially current earnings, are the most important predictors. However, with increasing forecast horizon, variable importance becomes more balanced. More precisely, B/S information becomes much more important whereas I/S information becomes less important with increasing forecast horizon. Thoroughly disentangling the different financial statements suggests that this dynamic may be attributable to earnings management.

As the first study to thoroughly decompose the effects of non-linearity in the earnings prediction context we find that especially for short term-horizons, the relationship as approximated by the best performing machine learning model, i.e., the machine learning ensemble, can still be described by a linear surrogate model to a large extent. More precisely, we find that on average around 84-89% of the variance in predictions can be

explained by a linear model, depending on the forecast horizon. Interactions between financial statement variables play virtually no role, implying that non-linearity of the functional form resembles the non-linear part of the model. As the forecast horizon increases, the linear surrogate $R^2$ decreases slightly. Nevertheless, interactions across financial statement variables remain irrelevant, thus implying that non-linearity of the functional form becomes more important with increasing forecast horizon.

Our findings provide important guidance for future research. First, we show that machine learning approaches are an excellent tool for earnings predictions. We hence argue that research which uses (model) earnings predictions in some way or another should resort to machine learning methods, if high accuracy is desired. Second, we show which financial statement variables and groups thereof are important. Future research may build upon that when building models and deciding which variables to include. Importantly, this includes the differences in terms of variable importance across forecast horizons. For example, if one is interested in an earnings prediction model including only a small number of variables for computation-related reasons, employing distinct (small) sets of variables for different forecast horizons might be beneficial. Lastly, we show that interactions among financial statement variables bear no predictive power. Again, future research may build upon this finding when deciding on an earnings prediction model.

# Chapter 4

# Model-Based Earnings Forecast Accuracy and Implied Cost of Capital Portfolio Returns[‡]

## 4.1 Introduction

Implied cost of capital (ICC) hold a pivotal role in the finance and accounting literature. They resemble the internal rate of return that relates expected future payoffs to current observable prices. Put differently, ICC are return expectations given current market prices and expected future payoffs. In contrast to ICC, which are inherently forward looking, alternative common methods of deriving return expectations, such as characteristic-based or factor-based approaches, typically rely on past return realizations. However, research suggests that return expectations derived from realized returns are prone to substantial noise (e.g., Fama and French, 1997; Elton, 1999). ICC serve as a conceptually different way of deriving expected returns, aimed at tackling this issue (Gebhardt et al., 2001).

Unsurprisingly, the academic literature provides a plethora of different ICC models describing how exactly future expected payoffs are related to current prices and the discussion on the validity of these ICC models is still ongoing. For example, Callen and Lyle (2020) show that the term structure of cost of capital is not flat for most time periods, thereby questioning the assumption of constant cost of capital, which ICC models typically rely on. Importantly, ICC also heavily rely on the accuracy of

[‡]This chapter is based on Simon (2024). I thank Dieter Hess, Simon Wolf and participants at the Finance Research Seminar at the University of Cologne for helpful comments and suggestions.

the earnings predictions in particular (Richardson et al., 2010). Studies have typically predicted earnings via analyst forecasts (Hou et al., 2012). However, a large body of literature documents biases of analyst earnings forecasts and analyst firm coverage is limited.[1] To circumvent these shortcomings, Hou et al. (2012) suggest to use model-based earnings forecasts for ICC estimation. Since then, numerous studies have proposed different earnings forecast models, aiming to achieve higher forecast accuracy (e.g., Li and Mohanram, 2014; Hess and Wolf, 2022). Originally only covering simple linear approaches, several more complex approaches have been introduced recently (e.g., Cao and You, 2021; Van Binsbergen et al., 2023).

Despite their limitations and the ongoing pursuit of the most accurate earnings prediction model, ICC have been extensively utilized in empirical studies. For example ICC have been used to study the relationship between risk and returns (e.g., Botosan and Plumlee, 2005; Pástor et al., 2008) and return predictability (e.g., Li et al., 2013). Another stream of literature employs ICC to study various issues in corporate finance, such as the impact of corporate governance on cost of capital (e.g., Botosan, 1997; Francis et al., 2005; Ashbaugh-Skaife et al., 2009), the relation between earnings attributes and cost of capital (Francis et al., 2004), the impact of accounting restatements on cost of capital (Hribar and Jenkins, 2004), and the relationship between investment and cost of capital (Frank and Shen, 2016).

Considering ICC serve as a proxy for expected returns, they may also be used for investment purposes. However, there is only little literature that explicitly focuses on the investment aspects of ICC. Typically, studies that aim to improve model-based earnings forecast accuracy in the context of ICC estimation briefly cover long-short portfolio strategies based on ICC and show that the more accurate earnings forecast model translates into higher average spread returns (e.g., Hou et al., 2012; Li and Mohanram, 2014; Hess and Wolf, 2022). However, these studies do not account for transaction costs, a factor that is crucial when evaluating actual investment performance (e.g., Chen and Velikov, 2023; Detzel et al., 2023). Notable exceptions are Esterer and Schröder (2014) and Bielstein (2018), but these studies resort to analyst earnings forecasts and only use rudimentary transaction cost proxies. This study aims to fill this gap in the literature by revisiting the relationship between model-based earnings forecast accuracy and ICC investment performance while accounting for transaction costs. More precisely, I provide

[1] See e.g., Kothari et al. (2016) for an extensive review of the literature.

an answer to the following novel research question: *What is the relationship between model-based earnings forecast accuracy and ICC portfolio performance against the background of transaction costs*?

In addition to explicitly accounting for transaction costs, one of the key contributions of this study is the fact that I consider two dimensions of accuracy, or more generally - model performance, separately: general accuracy and the degree to which forecasts are exposed to systematic distortions. In general terms, I define general accuracy as the absolute deviation of earnings forecasts from earnings realizations. General accuracy is thus captured by standard error metrics such as e.g., the mean squared error (MSE) or the mean absolute error (MAE). I define systematic distortions as deviations of the actual forecast error distribution from a corresponding distribution of forecast errors with mean zero and constant standard deviation which yields the same level of general accuracy.[2] Systematic distortions hence include forecast characteristics like bias, i.e., a distribution of forecast errors with mean unequal to zero, and accuracy differences for different subsets of firms, i.e., varying mean/standard deviation of error terms across subsets of firms. Existing studies typically focus on improving general accuracy as measured by e.g., the price scaled absolute forecast error (PAFE). Some studies also shed light onto the degree to which specific forms of systematic distortions are present in the respective forecast model. One example is given by Hou et al. (2012), who show that their model forecasts are less accurate, but also less biased than analyst forecasts. Despite being less accurate, their model forecasts translate into more profitable ICC portfolios. They attribute this to the difference in bias. Another example of such a systematic distortion is provided by e.g., Li and Mohanram (2014), who show that model-based earnings forecasts are less accurate for smaller firms. They do not, however, explicitly relate this characteristic to portfolio returns. I summarize all such distortions by denoting them as systematic distortions and provide a novel way of measuring them using the Kolmogorov-Smirnov test statistic. In addition to measuring systematic distortions of a forecast model, I follow the literature and measure general accuracy via the PAFE.

I start off the empirical analysis by estimating a selection of common earnings forecast models. First, I predict earnings using common linear traditional approaches. More precisely, I employ the well-known models by Hou et al. (2012) and Li and Mohanram (2014). Second, I predict earnings using a nonlinear machine learning (ML) model based

[2]Furthermore, I assume that forecast errors of undistorted forecasts are normally distributed.

on Hess et al. (2024).[3] Consistent with findings from Hess et al. (2024), I confirm that the ML model attains the highest general accuracy. Furthermore, utilizing the novel metric introduced in this study, I demonstrate that the ML model also showcases the least amount of systematic distortions.

Turning to ICC portfolios based on the different model forecasts, I find that the less accurate and more distorted models fail to produce statistically significant average gross long-short spread returns. In contrast, the ML model leads to statistically significant positive average return spreads. These findings remain robust across varying quantile sizes, aligning with existing literature suggesting that higher earnings forecast accuracy generally corresponds to higher gross return spreads (e.g., Hou et al., 2012; Li and Mohanram, 2014; Hess and Wolf, 2022; Hess et al., 2024). Consistent with the literature on the effect of transaction costs on return anomalies (e.g., Novy-Marx and Velikov, 2016; Chen and Velikov, 2023; Detzel et al., 2023), I demonstrate that the introduction of transaction costs substantially reduces ICC portfolio returns, attributable to high turnover levels, even with just an annual rebalancing frequency. This bears important implications for studies that evaluate ICC long-short portfolio return spreads as it shows that transaction costs significantly alter results. Nonetheless, the ML portfolio still yields statistically significant spread returns and clearly outperforms the portfolios based on less accurate earnings forecast models.

Since the ML forecasts are both more accurate and less distorted, it is not possible to explicitly attribute return gains to either of these two characteristics. In fact, existing studies typically only cover general accuracy, while some studies also assess specific systematic distortions like bias or accuracy differences across subsets of firms (e.g., Hou et al., 2012; Li and Mohanram, 2014). However, to the best of my knowledge, no study explicitly disentangles the two dimensions of model performance and their effect on portfolio performance. By introducing simple simulation frameworks, I separately assess the impact of general forecast accuracy and systematic distortions. First, I assess the impact of completely mitigating systematic distortions versus achieving perfect accuracy. I then dig deeper into the underlying dynamics and assess how changes in systematic distortions while keeping general accuracy constant impacts portfolio returns, and vice versa. More precisely, I show how e.g., improving the general accuracy by 50% while keeping the level of systematic distortions constant affects ICC portfolio returns.

[3] I argue that this model serves as a proxy for the most accurate class of models which the literature in this context currently provides, i.e., nonlinear ML-based approaches utilizing a large input vector.

The results reveal that both general accuracy and systematic distortions correlate with investment performance. In fact, completely mitigating systematic distortions while maintaining the same level of general accuracy improves average gross (net) portfolio returns by around 81% (88%) to 213% (261%), depending on the model, i.e., the level of general accuracy, and the quantile split considered. Achieving perfect general accuracy improves average gross (net) portfolio returns by another 17% (18%) to 33% (35%). The assessment of how isolated changes of general accuracy and the level of systematic distortions affect ICC portfolio returns further stresses that both aspects of model performance are correlated with average returns, but not with transaction costs.

Overall, the contribution of this study can be summarized as showing that generally, improving model-based earnings forecasts translates into higher ICC portfolio returns. More precisely, I show that both aspects of forecast performance, i.e., general accuracy and the degree to which forecasts are systematically distorted, are related to portfolio returns. Importantly, transaction costs significantly reduce portfolio returns in general, but do not strongly correlate with either aspect of predictive performance considered. These findings provide strong motivation for future research endeavors focused on enhancing earnings forecast models. They specifically underscore the potential benefits of mitigating systematic distortions, thereby highlighting promising directions for further model improvements.

The remainder of this study is structured as follows: in Section 4.2 I review the literature that this study contributes to. In Section 4.3 I describe the empirical research design underlying the core part of this study. I then summarize the empirical results of this part in Section 4.4. In Section 4.5 I propose simple simulation frameworks which highlight the different effects of earnings forecast model improvement on portfolio performance. Finally, Section 4.6 concludes the study.

## 4.2 Related literature

This study primarily relates to two strands of literature. First, it contributes to the literature which empirically evaluates ICC as an expected return proxy. Research that adds to this literature stream typically analyzes the extent to which ICC predict future realized returns (e.g., Gebhardt et al., 2001; Gode and Mohanram, 2003; Easton and Monahan, 2005; Guay et al., 2011; Hou et al., 2012; Li et al., 2013; Li and Mohanram,

2014; Ang and Aadka, 2017; Callen and Lyle, 2020). The rationale for investigating this relationship is grounded in the notion that a reliable expected return proxy should correlate with future realized returns, considering that any unexpected shock is not predictable by definition (Lee et al., 2020). These studies use various different ICC models, different methods of deriving the crucial ICC input, i.e., expected earnings, different samples and assess returns on different levels, including firm-, portfolio-, or market-level analyses. However, generally, the evidence on the relation between realized returns and ICC is mixed. Some studies find a positive relationship (e.g., Gode and Mohanram, 2003; Hou et al., 2012; Li et al., 2013; Li and Mohanram, 2014; Callen and Lyle, 2020), some studies find no significant relationship (e.g., Gebhardt et al., 2001; Easton and Monahan, 2005; Guay et al., 2011) and Ang and Aadka (2017) even find a negative relationship. On a different note, Hou et al. (2012) show that less biased earnings forecasts translate into more accurate return predictions via ICC. They generate such forecasts by deriving earnings predictions via a statistical model instead of using analyst forecasts. Building on Hou et al. (2012), several studies have shown more accurate model-based earnings forecasts lead to more accurate ICC return predictions. In fact, yielding better expected return proxies is oftentimes the primary motivation behind deriving more accurate earnings forecast models.

Since finance theory generally predicts a positive relationship between risk and return (Pástor et al., 2008), other studies evaluate ICC as an expected return proxy by assessing their relation to risk and risk factors, both in the cross-section (e.g., Botosan and Plumlee, 2005; Hou et al., 2012; Li and Mohanram, 2014; Lee et al., 2020) and in the time series (e.g., Pástor et al., 2008).

In yet another conceptually different attempt at evaluating, inter alia, ICC, Lee et al. (2020) propose a theoretically grounded testing framework for expected return proxies in general. The framework builds on a simple decomposition of returns and is primarily constructed such that the respective expected return proxy is evaluated with respect to its usefulness for deriving treatment effects. Empirically applying their framework shows that ICC outperform characteristic- and factor-based expected return proxies in the time series, while characteristic-based proxies perform best in the cross-section.

This paper falls into the first of the aforementioned categories, i.e., it complements literature that evaluates the predictive capabilities of ICC. More precisely, I evaluate the impact of model-based earnings forecast accuracy on the predictive capabilities of ICC. I

extend the findings by Hou et al. (2012) and similar studies in two key aspects: first I revisit the earnings forecast-ICC-future return relationship against the background of transaction costs. Existing research in this body of literature typically does not take into account transaction costs when assessing ICC investment performance (e.g., Hou et al., 2012; Li and Mohanram, 2014).[4] Second, I differentiate between two aspects of accuracy, i.e., general accuracy and systematic distortions, such as bias or accuracy differences across firms.

The second major literature stream I contribute to is the discussion regarding the effects of transaction costs on investment performance in general. This literature has gained particular traction recently, since transaction costs significantly alter findings regarding return anomalies (e.g., Novy-Marx and Velikov, 2016; DeMiguel et al., 2020; Chen and Velikov, 2023; Detzel et al., 2023) and lead to a different, actually implementable efficient frontier (Jensen et al., 2022). However, to the best of my knowledge, neither the effect of transaction costs on ICC investment performance, nor the relationship between model-based earnings forecast accuracy and transaction costs associated with the respective ICC long-short portfolios have been studied extensively thus far. The two notable exceptions are Esterer and Schröder (2014) and Bielstein (2018). Yet, these studies do not shed light on the latter of the aforementioned points of interest. Moreover, both studies employ rather rudimentary transaction cost proxies and resort to analyst earnings forecasts.

## 4.3 Empirical approach

### 4.3.1 Implied cost of capital

The central metric of this study are the ICC. Conceptually, ICC solve

$$P_{i,t} = \sum_{\tau=1}^{\infty} \frac{\mathbb{E}_t(X_{i,t+\tau})}{(1+ICC_{i,t})^{\tau}}, \tag{4.1}$$

where $P_{i,t}$ denotes the market value of equity of firm $i$ at time $t$, $\mathbb{E}_t(X_{i,t+\tau})$ is the conditionally expected pay-off of firm $i$ $\tau$ periods ahead, and ICC is the implied cost of equity. Note that the ICC varies for each $i$ and $t$, but not across $\tau$. Put differently, for a

[4] Important exceptions are Esterer and Schröder (2014) and Bielstein (2018), but they do not explicitly assess the impact of earnings forecast accuracy on ICC and ICC-based investment performance.

given firm $i$ at time $t$, one assumes constant cost of equity throughout future periods.[5]

As such, ICC are an *ex ante* measure of equity return expectations of the market, under the assumption that the respective assumed relation between future payoffs and current prices holds.[6] ICC thus represent an alternative to other methods of deriving the cost of equity capital, such as factor-based or characteristic-based proxies, which typically rely on return realizations. Many issues with using realized returns for cost of capital computation have been documented in the literature (e.g. Elton, 1999), highlighting ICC as a potential conceptual enhancement in this regard.

The literature provides numerous models to estimate ICC.[7] Following e.g., Echterling et al. (2015), these models may be grouped into those derived from dividend discount models (DDMs), residual income models (RIMs) and abnormal earnings growth models (AEGs). Yet, under the assumptions inherent to all of them, the models simply resemble algebraic reformulations of each other (Hendriock, 2022).

I follow the literature and compute ICC as the average of commonly used ICC models (e.g., Hou et al., 2012; Li and Mohanram, 2014; Hess et al., 2019). More specifically, I estimate one DDM, namely the GG model by Gordon and Gordon (1997). I further estimate two RIMs, namely the GLS model by Gebhardt et al. (2001) and the CT model by Claus and Thomas (2001). Lastly, the two AEGMs estimated in this study are the MPEG model by Easton (2004) and the OJ model by Ohlson and Juettner-Nauroth (2005). I exclude ICC below 0% and above 100%. Furthermore, I drop observations for which any of the composite ICC values is missing due to the aforementioned constraint.[8] An overview over the models is given in Appendix C.1. In addition to prices, which are observable, the crucial input to all of these models are expected earnings. The methodology employed to derive these expectations is detailed in the next subsection.

### 4.3.2 Earnings forecasts and model estimation

#### Earnings forecasts

I cover a selection of different earnings forecast models. First, I predict earnings using

[5]This is quite a restrictive assumption. In fact, Callen and Lyle (2020) show that the term structure of ICC is not flat throughout most periods. In this context, Penman et al. (2023) argue that ICC are similar to the yield-to-maturity for bonds and do not necessarily quantify the return for risk. However, a critique of ICC per se is beyond the scope of this study.

[6]Note that despite being called implied cost of capital, ICC exclusively relate to equity capital.

[7]See e.g. Echterling et al. (2015) for an overview.

[8]This ensures an equal amount of ICC estimates per forecast model. However, in undocumented results I confirm that the results remain robust even when this constraint is not imposed.

the three most common traditional linear models, namely the HVZ model (Hou et al., 2012), the EP model (Li and Mohanram, 2014) and the RI model (Li and Mohanram, 2014).[9] All of these models are simple linear models of the following form:

$$\mathbb{E}_t[E_{i,t+\tau}] = \boldsymbol{\beta} \boldsymbol{X}_{i,t}, \tag{4.2}$$

where $\mathbb{E}_t[E_{i,t+\tau}]$ denotes expected $\tau$-period ahead earnings of firm $i$, $\boldsymbol{\beta}$ denotes a vector of coefficients and $\boldsymbol{X}_{i,t}$ denotes a vector of predictor variables of firm $i$ at time $t$.

The models differ in terms of which variables the input vector consists of. However, in all three cases, it is low-dimensional, i.e., only a few predictor variables are utilized. An overview over the models and their respective input variables is given in Table C.2 in the Appendix.

In addition, I predict earnings using a flexible machine learning approach based on multiple models. The models used in this study allow for complex functional forms $f(\cdot)$ and high-dimensional input vectors. Formally, the models may generally be described as follows:

$$\mathbb{E}_t[E_{i,t+\tau}] = f(\boldsymbol{X_{i,t}}). \tag{4.3}$$

More precisely, I estimate an equally weighted ensemble of a random forest (RF), a gradient-boosted tree model (GBT) and a gradient-boosted tree model with dropout (DART).[10] Crucially, these models leverage a broader and more extensive set of inputs compared to traditional linear approaches. Closely following Hess et al. (2024), I provide the models with the complete financial statement data from Compustat.[11] An overview over these financial statement items is given in Table C.4 in the Appendix.

Note that although a random walk type forecasting approach achieves similar accuracy as many of the models employed in the literature (e.g., Gerakos and Gramacy, 2012; Li and Mohanram, 2014), I do not include such a model in this study.[12] The reason behind this lies in its conceptual mismatch for ICC calculation. To elaborate, random

[9] These models are named *traditional models* hereafter.

[10] I exclusively resort to tree-based methods in this study due to their ease of implementation and restrictions regarding computational power. Building on, inter alia, Hess et al. (2024), who show that model averaging leads to increases in model performance and on the literature which uses tree-based models to proxy for ML approaches in general (e.g., Van Binsbergen et al., 2023), I argue that this tree-based ensemble approach resembles a valid proxy for flexible ML models.

[11] Note that certain restrictions apply. A comprehensive discussion is provided in section 4.3.5 below.

[12] A random walk model equates earnings forecasts for any forecast horizon with current earnings.

walk forecasts remain constant across forecast horizons, consequently impeding any possibility of earnings growth (Li and Mohanram, 2014).

**Model estimation**

I employ a rolling window strategy to estimate the models and retrieve out-of-sample earnings predictions which are then used to compute ICC. More specifically, in the case of the ML approaches, I split the data into training, tuning and test data. For each forecast horizon $\tau$, the procedure for generating forecasts $t + \tau$ is as follows: I train the models using earnings from $t - 11$ to $t - 2$ as output and respective financial statement data lagged by $\tau$ as input. I then tune the models using earnings from $t - 1$ to $t$ as output and respective financial statement data lagged by $\tau$ as input. By tuning, I refer to finding the set of model-specific hyperparameter values that correspond to the best predictions for the tuning data. A summary of these model-specific hyperparameters is given in Appendix C.3. I subsequently derive earnings predictions for $t + \tau$ by feeding the optimized models the input variables from $t$. I then move forward one year and repeat the procedure. Importantly, I follow Hou et al. (2012) and Li and Mohanram (2014) and estimate the models at the end of June each year, assuming a three to fourteen month reporting lag for financial statements.[13]

The traditional linear approaches do not require hyperparameter tuning. In these cases, I only split the data into training and test data. The rest of the procedure remains the same. More precisely, for each forecast horizon $\tau$, I train the models using earnings from $t - 11$ to $t$ as output and respective financial statement data lagged by $\tau$ as input. I subsequently derive earnings predictions for $t + \tau$ by feeding the estimated models the input variables from $t$.

### 4.3.3 Predictive performance

One of the key contributions of this study is that I differentiate between general accuracy and systematic distortions. The latter includes attributes such as bias or varying accuracy across subsets of firms.[14] In the following, I describe how I measure general accuracy and the degree to which systematic distortions are present in a model forecast.

[13] More precisely, data from April of year $t - 1$ to March of year $t$ is considered to be the most recent fiscal year-end data available as of June in $t$ and hence comprises the information as of $t$.

[14] Existing studies typically only focus on specific forms of systematic distortions. For example, Hou et al. (2012) only determine the bias of their forecast model.

**General accuracy**

I define general accuracy as the price scaled absolute forecast error (PAFE), the primary method of evaluating earnings forecasts in the literature (e.g., Hou et al., 2012; Li and Mohanram, 2014; Hess et al., 2019). The PAFE is calculated as follows:

$$PAFE_{i,t+\tau} = \frac{|E_{i,t+\tau} - \hat{E}_{i,t+\tau}|}{Price_{i,t}}, \tag{4.4}$$

where $E_{i,t+\tau}$ denotes actual earnings for firm $i$ in period $t+\tau$, $\hat{E}_{i,t+\tau}$ denotes the corresponding forecast as of $t$ and $Price_{i,t}$ is the firm's stock price at the end of June in the respective estimation year. I compute the general accuracy of a given forecast model as its median PAFE across all out-of-sample periods.

**Systematic distortions**

To determine the degree to which forecasts are exposed to systematic distortions, I propose a novel, simulation-based metric. Generally, the metric captures the difference between the forecasts of the respective model and corresponding forecasts which exhibit zero systematic distortions and the same level of general accuracy. I derive the undistorted forecasts via simulation. More precisely, I simulate such earnings forecasts as follows: let $E_{i,t+\tau}$ denote earnings to be predicted for firm $i$ in future period $t+\tau$. Moreover, consider some median PAFE denoted by $\hat{PAFE}^{model}_{t+\tau}$.[15] This PAFE resembles the respective model's median PAFE which the corresponding undistorted forecasts should match. I then draw from a normal distribution centered around 0 with standard deviation $\sigma$, repeating the process for each firm $i$ and each forecast period $t+\tau$. Importantly, $\sigma$ depends on the desired median accuracy $\hat{PAFE}^{model}_{t+\tau}$. More precisely, for a given level of general accuracy, I solve for the respective $\sigma$ leading to $\hat{PAFE}^{model}_{t+\tau}$ via numerical optimization. This yields firm/forecast period specific deviations $\Gamma_{i,t+\tau}$. I then derive undistorted earnings predictions $\hat{E}^{ud}_{i,t+\tau}$ for a given level of general accuracy by setting

$$\hat{E}^{ud}_{i,t+\tau} = E_{i,t+\tau} + \Gamma_{i,t+\tau}. \tag{4.5}$$

Note that this method ensures equal treatment of every firm, as the deviation distribution remains consistent across all firms, while matching the general accuracy of the respective forecast model considered. In essence, each firm, regardless of size or other characteristics, is expected to exhibit the same level of general accuracy. Moreover,

[15] The method is agnostic towards the evaluation metric used. Hence, one may also use, e.g., the OOS $R^2$.

since the deviation distribution's mean is zero, there is no inherent bias, eliminating any systematic tendencies toward overestimation or underestimation. Consequently, the forecasts are absent of any form of systematic distortion, as each forecast is equally unbiased and reliable across the entire spectrum of firms.

Equipped with a set of undistorted forecasts with equal general accuracy as the respective forecast model of interest, I derive the degree to which a forecast is exposed to systematic distortions via the Kolmogorov-Smirnov (KS) test for two samples. More specifically, I measure the difference between the distribution of model forecasts and the corresponding distribution of undistorted forecasts with the same level of general accuracy by deriving the two-sample KS test statistic for these two sets of forecasts. The larger the size of the KS test statistic, the more dissimilar the forecasts, i.e., the larger the degree of systematic distortions. In the following, I will refer to the KS test statistic for the difference between actual forecasts and corresponding undistorted forecasts as SYSD.

### 4.3.4 Portfolio analysis and transaction costs

Following the literature, I construct long-short portfolios conditional on ICC to evaluate the investment performance (e.g., Hou et al., 2012; Li and Mohanram, 2014; Hess et al., 2019).[16]

More precisely, the procedure is as follows: at the end of June in a given year, I derive the cross-section of ICC. I then sort firms according to their ICC and go long the highest quantile and short the lowest quantile of stocks, equally weighting each stock within the respective quantile. To enhance the robustness of the results, I evaluate various common quantile splits, including decile, quintile, and tercile splits. Finally, I observe the realized returns in the following year, i.e., from July in the respective year to (including) June of the following year. The procedure is then repeated in the following year, such that the investment portfolio is rebalanced annually at the end of June. In total, this procedure leads to 42 out-of-sample investment periods.

In contrast to the majority of existing studies, I explicitly account for transaction costs. Transaction costs ($TC$) associated with the rebalancing of the weight $w$ of stock $i$ in period $t$ are derived as follows:

[16]Note that one can use ICC as an expected return proxy in any alternative approach. For example, Bielstein (2018) uses ICC in a Markowitz portfolio optimization scenario.

$$TC_{i,t} = \kappa_{i,t}|w_{i,t} - (1 + r_{i,t-1})w_{i,t-1}|, \tag{4.6}$$

where $r_{i,t-1}$ is the return that stock $i$ generated in period $t-1$ and $\kappa_{i,t}$ is a transaction cost parameter.

Following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020), I define the transaction cost parameter $\kappa_{i,t}$ associated with stock $i$ at time $t$ as

$$\kappa_{i,t} = y_t z_{i,t}, \tag{4.7}$$

where $z_{i,t} = 0.006 - 0.0025 me_{i,t}$, with $me_{i,t}$ as the cross-sectionally normalized market capitalization of firm $i$ at time $t$. $y_t$ is a vector consisting of values which decrease linearly from 3.3 in January 1980 to 1.0 in January 2002, and remain 1.0 afterwards. Compared to Esterer and Schröder (2014) and Bielstein (2018), who use flat percentages, this represents a more granular and realistic approach of modeling transaction costs that is in line with the current literature.

### 4.3.5 Data

The sample resembles an intersection of Compustat and CRSP data. Earnings as well as other financial statement items used to derive earnings forecasts and ICC are retrieved from Compustat. Returns and prices are retrieved from CRSP. The sample covers securities with share codes 10 or 11 which are listed on the NYSE, Amex or Nasdaq. I drop observations with missing or $< 1\$$ price, missing common shares outstanding or missing earnings. The Compustat data spans from 1969 to 2021. Considering the assumed reporting lag mentioned above, this implies that the most recent estimation date for ICC is June 2022. As returns are collected in the year following the respective ICC estimation and hence portfolio formation date, the CRSP data extends to 2023.

The ML model is estimated using the Compustat financial statement items that remain after dropping variables with more than 50% of observations missing or no observations in any of the cross-sections (i.e., estimation years).[17] I also include the subset of traditional model variables which is not included in the aforementioned group of financial statement variables. In total, this amounts to a predictor set of 109 variables which the ML model

[17] I also drop variables already scaled by shares. The reason for that is, that I scale all variables by shares and hence these variables are redundant. Further, due to computational restrictions, I do not include lags and first-order differences in contrast to the approach taken by Hess et al. (2024).

utilizes. An overview over these variables is given in Table C.4 in the Appendix. An overview over the variables used to estimate the traditional models is given in Table C.2 in the Appendix.

To ensure consistency, earnings are defined as Compustat income before extraordinary items and discontinued operations (ib), irrespective of the earnings forecast model.[18] All variables, including the target variable, are scaled by common shares outstanding and winsorized at the 1% and 99% level, respectively.

## 4.4 Results

### 4.4.1 Evaluating earnings forecasts

The results in regards to predictive performance are reported in Table 4.1. The ML model outperforms the traditional models in terms of general accuracy and degree to which systematic distortions are present. For $t+1$ predictions, for example, the ML model achieves a median PAFE which is around 10% lower than that of the second-best performing model, i.e., the RI model. Interestingly, the second-best performing model in terms of SYSD is the HVZ model, which the ML model beats by around 14%. The percentage differences in median PAFE show a slight decrease as the forecast horizon increases. Furthermore, the percentage differences in SYSD converge towards zero with increasing forecast horizon. In other words, for longer-term forecasts, all models display similar levels of systematic distortions. These results indicate that the performance of earnings forecast models converges with increasing forecast horizon in terms of both dimensions of accuracy considered.

Overall, the results stress that the ML model is not only generally more accurate, but also exhibits a lower amount of systematic distortions than common traditional earnings forecast models. It hence improves over common traditional models in both aspects of predictive performance considered. However, forecast performance converges as the forecast horizon increases.

### 4.4.2 Implied cost of capital

Table 4.2 gives an overview over the ICC estimates based on each forecast model considered. In general, the traditional models and the ML model lead to similar ICC in

[18] Note that this might lead to deviations from the original models.

**Table 4.1: Model-based earnings forecast accuracy**

| | | $E_{t+1}$ | $E_{t+2}$ | $E_{t+3}$ | $E_{t+4}$ | $E_{t+5}$ |
|---|---|---|---|---|---|---|
| HVZ | PAFE | 0.0318*** | 0.0453*** | 0.0547*** | 0.0638*** | 0.0733*** |
| | SYSD | 0.1497*** | 0.2386*** | 0.2978*** | 0.3451*** | 0.3815*** |
| EP | PAFE | 0.0308*** | 0.0435*** | 0.0524*** | 0.0608*** | 0.0696*** |
| | SYSD | 0.1520*** | 0.2403*** | 0.2988*** | 0.3457*** | 0.3813*** |
| RI | PAFE | 0.0307*** | 0.0427*** | 0.0507*** | 0.0580*** | 0.0664*** |
| | SYSD | 0.1498*** | 0.2362*** | 0.2928*** | 0.3383*** | 0.3737*** |
| ML | PAFE | 0.0277*** | 0.0392*** | 0.0476*** | 0.0554*** | 0.0621*** |
| | SYSD | 0.1283*** | 0.2153*** | 0.2832*** | 0.3367*** | 0.3789*** |

This table reports the general accuracy as well as the level of systematic distortions for each model considered. General accuracy is measured as the median price scaled absolute forecast error (PAFE). Systematic distortions are measured as the Kolmogorov-Smirnov test statistic of the comparison between simulated forecasts exposed to no systematic distortions and actual model forecasts exposed to systematic distortions (SYSD). The simulated forecasts are derived such that their general accuracy (measured via the median PAFE) is equal to the general accuracy of the actual model forecasts for each forecast horizon. $E_{t+1}$ to $E_{t+5}$ denote one- to five-year ahead earnings. HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), and ML is the machine learning approach based on Hess et al. (2024). ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance of the median PAFE are adjusted following Driscoll and Kraay (1998) assuming a lag length of three years.

terms of descriptive statistics. ICC estimates of each model considered have an average of 10.07-10.93% and a standard deviation of 6.73-7.87%. The remaining statistics also exhibit minimal variation across the forecast models under consideration.

**Table 4.2: Implied cost of capital: descriptive statistics**

| | N | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| HVZ | 212 434 | 0.1093 | 0.0787 | 0.0005 | 0.0608 | 0.0894 | 0.1317 | 0.9933 |
| EP | 212 434 | 0.1037 | 0.0733 | 0.0004 | 0.0593 | 0.0838 | 0.1234 | 0.9968 |
| RI | 212 434 | 0.1007 | 0.0673 | 0.0000 | 0.0599 | 0.0841 | 0.1201 | 0.9532 |
| ML | 212 434 | 0.1012 | 0.0684 | 0.0001 | 0.0583 | 0.0849 | 0.1230 | 0.9822 |

This table reports descriptive statistics of the ICC based on the different earnings forecast models for the period during 1981-2022. The ICC are derived as the average of the GG model (Gordon and Gordon, 1997), the GLS model (Gebhardt et al., 2001), the CT model (Claus and Thomas, 2001), the MPEG model (Easton, 2004) and the OJ model (Ohlson and Juettner-Nauroth, 2005). HVZ refers to ICC based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer ICC based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to ICC based on ML model earnings forecasts (Hess et al., 2024). The table reports the number of ICC estimated (N), the average ICC (Mean), the standard deviation of ICC (Std), the minimum (Min), the 25%-percentile (25%), the median (50%), the 75%-percentile (75%) as well as the maximum (Max) of the ICC estimates.

In Figure 4.1 I plot the time-series of the ICC estimates based on the different forecast models. It becomes evident that all ICC move in similar patterns, i.e., they react similarly to different market states. This is expected, as all forecast models share the most important future earnings predictor, i.e., current earnings. Mirroring the findings by Hou

et al. (2012), the ICC estimates generally seem to decrease over the periods considered.

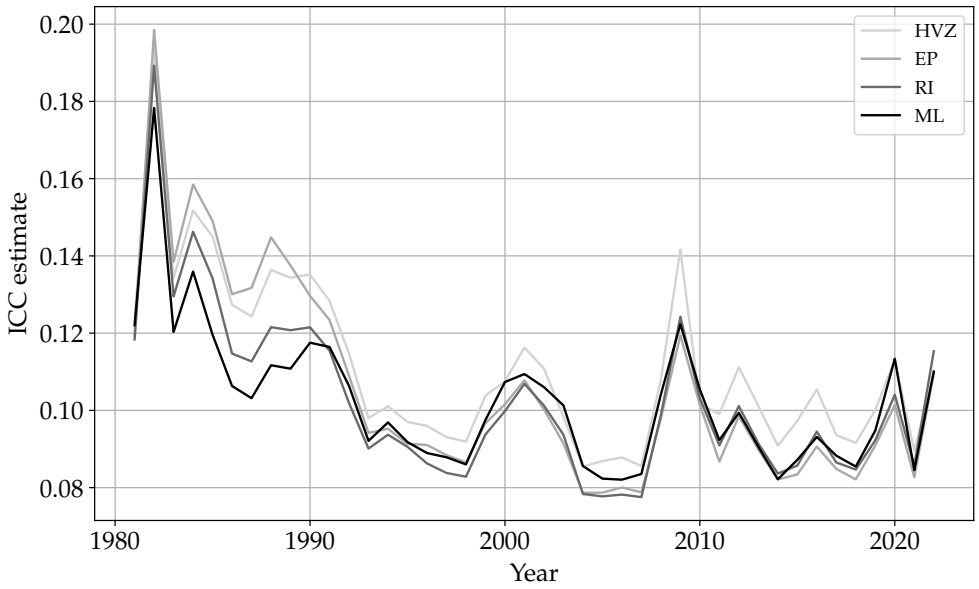



**Figure 4.1: Time-series of ICC**

This figure plots the average ICC estimates per estimation year and per earnings forecast model considered. The ICC are derived as the average of the GG model (Gordon and Gordon, 1997), the GLS model (Gebhardt et al., 2001), the CT model (Claus and Thomas, 2001), the MPEG model (Easton, 2004) and the OJ model (Ohlson and Juettner-Nauroth, 2005). HVZ refers to ICC based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer ICC based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to ICC based on ML model earnings forecasts (Hess et al., 2024).

I quantify the fact that ICC behave similarly by reporting the Pearson correlation coefficients in Table 4.3. ICC based on the three traditional models correlate strongly with each other, with correlation coefficients ranging from around 0.85 to 0.88. Conversely, ICC based on the ML model display lower correlation with those based on the other models, with correlation coefficients ranging from approximately 0.69 to 0.71.

**Table 4.3: Implied cost of capital: Pearson correlation coefficients**

| | HVZ | EP | RI | ML |
|---|---|---|---|---|
| HVZ | 1.0000 | | | |
| EP | 0.8590 | 1.0000 | | |
| RI | 0.8462 | 0.8811 | 1.0000 | |
| ML | 0.6998 | 0.6854 | 0.7146 | 1.0000 |

This table reports the Pearson correlation coefficients of the ICC based on the different earnings forecast models. The ICC are derived as the average of the GG model (Gordon and Gordon, 1997), the GLS model (Gebhardt et al., 2001), the CT model (Claus and Thomas, 2001), the MPEG model (Easton, 2004) and the OJ model (Ohlson and Juettner-Nauroth, 2005). HVZ refers to ICC based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer ICC based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to ICC based on ML model earnings forecasts (Hess et al., 2024).

### 4.4.3 Portfolio returns

I now turn to the main research focus of this study by assessing the relation between forecast model performance and ICC portfolio returns against the background of transaction costs.

Table 4.4 reports average returns of the long-/short-legs as well as the long-short spread returns of the three different quantile splits with equal weighting within the respective leg. In the following, I will refer to the ICC long-short portfolio based on the respective earnings forecast model as *model* portfolio. For example, the ICC long-short portfolio based on the HVZ model is simply referred to as the HVZ portfolio.

**Table 4.4: Average returns of portfolio sorts on implied cost of capital**

| | | HVZ | EP | RI | ML |
|---|---|---|---|---|---|
| Deciles | 1 | 0.1188 | 0.1139 | 0.1133 | 0.0850 |
| | 10 | 0.1476 | 0.1474 | 0.1528 | 0.1515 |
| | 10-1 | 0.0288 | 0.0335 | 0.0394 | 0.0665** |
| | 10-1 net | 0.0149 | 0.0192 | 0.0253 | 0.0520* |
| Quintiles | 1 | 0.1124 | 0.1110 | 0.1093 | 0.0994 |
| | 5 | 0.1410 | 0.1392 | 0.1458 | 0.1502 |
| | 5-1 | 0.0287 | 0.0282 | 0.0365 | 0.0509** |
| | 5-1 net | 0.0160 | 0.0152 | 0.0237 | 0.0376* |
| Terciles | 1 | 0.1159 | 0.1146 | 0.1123 | 0.1097 |
| | 3 | 0.1410 | 0.1428 | 0.1462 | 0.1492 |
| | 3-1 | 0.0252 | 0.0282 | 0.0339* | 0.0395** |
| | 3-1 net | 0.0140 | 0.0169 | 0.0226 | 0.0280 |

This table reports both the average gross and net returns per ICC long-short portfolio. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024). ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. Statistical significance is only derived for the long-short returns.

Irrespective of the number of quantiles chosen, the ML portfolio outperforms the other portfolios by a significant margin, both gross and net of transaction costs.

Gross of transaction costs, the ML portfolio achieves a statistically significant average annual spread return of 6.65% in the decile split scenario. It consistently outperforms the other portfolios by a margin ranging from 2.71 to 3.77 percentage points. This translates into the ML portfolio achieving between 68.78% and 130.93% higher average gross return spreads compared to the other portfolios. The remaining portfolios achieve statistically

insignificant average gross return spreads of 2.88-3.94%.

Upon factoring in transaction costs, the rankings of the portfolios in terms of average return spread remain unchanged. Yet, returns decline by a significant margin. The ML portfolio still clearly outperforms the other portfolios, achieving a statistically significant average net return spread of 5.20%. In absolute terms, the percentage differences decrease only slightly, with differences as compared to the other portfolios ranging from 2.67 to 3.71 percentage points.

Turning to the quintile as well as the tercile split scenarios reveals two key insights. First, the bigger the quantile, the lower the average return spread, both gross and net of transaction costs. The average gross ML portfolio return, for example, decreases to 5.09% in the quintile case and to 3.95% in the tercile case. The same dynamic holds true for the remaining portfolios. Second, the ML portfolio clearly outperforms the other portfolios for each quantile size considered. However, the ranking among the other portfolios changes with changing quantile size. More precisely, the EP portfolio is the third-best performing portfolio in the decile as well as the tercile split scenario, but falls short of the HVZ portfolio in the quintile scenario. The ML portfolio demonstrates statistically significant gross and net returns in both the decile as well as the quintile split scenario, and statistically significant gross returns in the tercile split scenario. With the exception of gross returns of the tercile RI portfolio, the remaining portfolios fail to yield statistically significant average returns.

Assessing the returns across the out-of-sample periods in Figure 4.2 reveals additional insights and stresses the economic dimensions of the aforementioned differences in portfolio performance.

First, as the results regarding average performance suggest, the ML portfolio clearly outperforms the other portfolios. When considering decile splits, the ML portfolio achieves a cumulative gross return of more than 610% throughout out-of-sample periods, clearly outperforming the traditional model portfolios, which achieve cumulative returns of around 23-108% by the end of the out-of-sample time frame. Second, the less granular the split approach, the lower the margin by which the ML portfolio outperforms the other portfolios. This suggests that the ML-based ICC relate particularly well to future returns in the most extreme quantiles, i.e., the very highest and the very lowest next-period returns. Yet, even when considering tercile splits, the ML portfolio still outperforms the second best portfolio, i.e., the RI portfolio, by more than around 69 percentage points

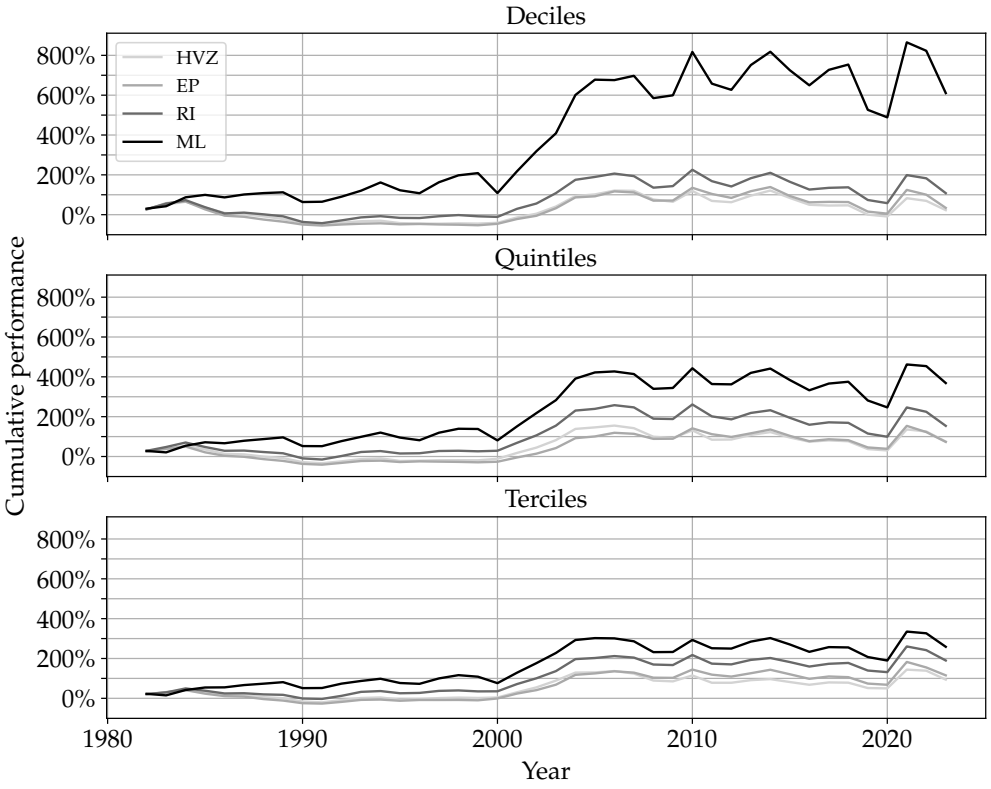



**Figure 4.2: Cumulative gross returns**
This figure plots the cumulative annual gross returns of the ICC long-short portfolios. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024).

in terms of cumulative returns. Third, irrespective of the quantile split scenario, the portfolios are at similar and comparably low cumulative return levels up until 2000. Afterwards there is a sharp increase in cumulative returns up until 2010, before they stay on somewhat similar levels and even decrease at the end of the out-of-sample time-frame.

Figure 4.3 shows that the net return time-series patterns are identical to the gross return patterns. However, cumulative portfolio returns diminish significantly upon the inclusion of transaction costs. For instance, the ML portfolio achieves a cumulative return of around 295% in the decile split scenario. The other models yield cumulative returns of around zero or even below zero. Once again, this stresses the importance of considering transaction costs when assessing ICC portfolios, as it significantly influences results. Notably, this also means that the gap between the ML portfolio and the second-best portfolio narrows to around 281 percentage points when considering decile splits. This is attributable to the fact that the exponential effect of compounding necessarily leads to

a convergence of cumulative returns when overall return levels decrease. Interestingly and contrary to the gross return scenario, the cumulative net returns of the traditional model portfolios increase with increasing quantile size. This is attributable to decreasing turnover levels with increasing quantile size, as I show below. Moreover, this stresses that the portfolio split considered can strongly influence cumulative returns via transaction costs. In the case of the ML portfolio, cumulative net returns once again decrease with increasing quantile size.

To summarize, I find that the best-performing forecast model in terms of median PAFE and SYSD leads to the highest ICC long-short portfolio returns. Moreover, transaction costs significantly impact portfolio performance. However, the results suggest that transaction costs do not strongly vary between different ICC portfolios on average.

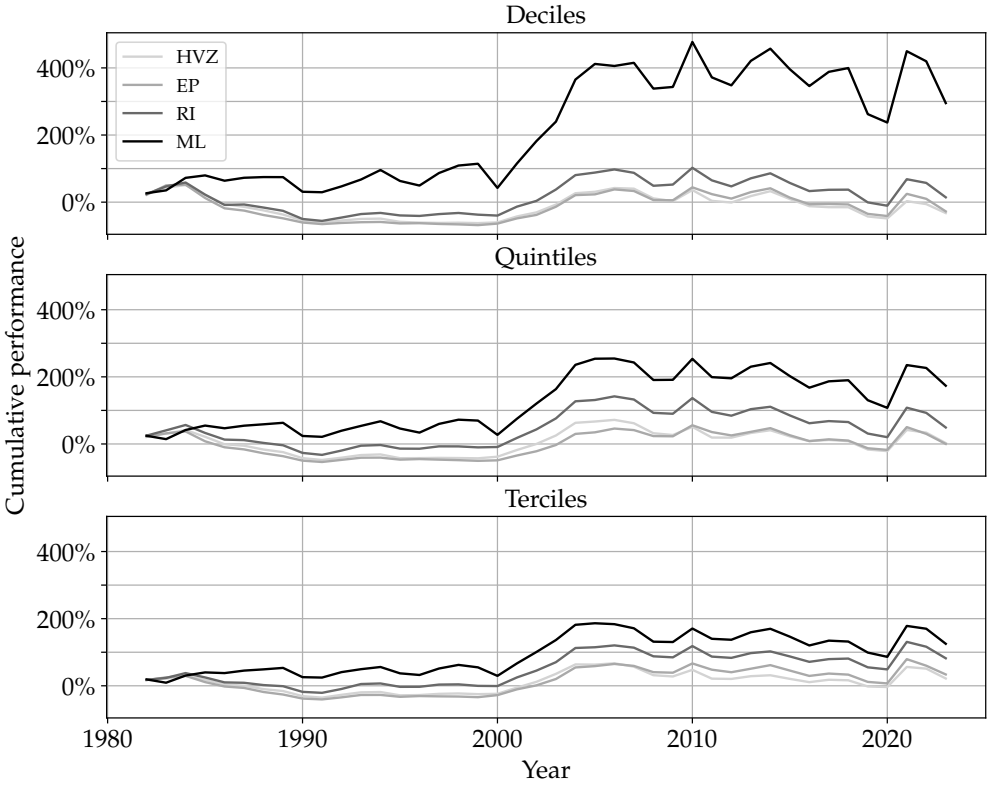



**Figure 4.3: Cumulative net returns**

This figure plots the cumulative annual net returns of the ICC long-short portfolios. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024). Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020).

### 4.4.4 Return spread differences through a factor lense

At the very heart of finance lies the question of how risk and return are related. Consequently, it becomes compelling to scrutinize the systematic risk factors driving the differences in return spreads. To this end, I regress the return spread differences on the five factors posited by Fama and French (2015). More precisely, I focus on the return spread differences between the ML portfolio and each of the remaining portfolios.

Table 4.5 summarizes the results and reveals several key insights. First, in terms of risk loadings, no stark differences between gross and net return differences across different quantile split scenarios exist. Second, the ML portfolio differs from the other portfolios in terms of risk factor loadings. More precisely, the portfolios based on traditional forecast models exhibit stronger small-minus-big (SMB) and conservative-minus-aggressive (CMA) factor loadings than the ML portfolio. The corresponding coefficients are negative and statistically significant in almost every case.[19] In contrast, the ML portfolio loads stronger on the robust-minus-weak-profitability (RMW) as well as the high-minus-low (HML) factor. The corresponding coefficients are positive and statistically significant in every case.

In sum, the results indicate that systematic differences between the ML portfolio and the other portfolios in terms of risk (factor loadings) exist. The ML portfolio appears to load more heavily on the RMW as well as the HML factor, whereas the other portfolios load more heavily on the SMB as well as the CMA factor, in comparison.

### 4.4.5 Dissecting transaction costs

In general, the comparison between gross and net returns suggests that there are no significant differences among the portfolios in terms of transaction costs. However, the ML portfolio appears to incur a slightly higher amount of transaction costs on average. Panel A in Table 4.6 confirms these results. The Panel gives an overview of the average transaction costs per long-/short-leg as well as the respective sum.

Generally, average annual transaction costs range from around 1.12-1.15% for the tercile portfolios to around 1.39-1.45% for the decile portfolios. Considering the return levels of the portfolios, these are quite large values. Remarkably, in each quantile split scenario examined, the ML approach consistently incurs the highest average transaction

[19]The SMB coefficients are statistically significant in every case. The CMA coefficients are statistically significant for every decile portfolio difference and most quintile portfolio differences.

**Table 4.5: Factor loadings of ICC portfolio return differences**

| | | $\alpha$ | $\beta_{MKT}$ | $\beta_{SMB}$ | $\beta_{HML}$ | $\beta_{RMW}$ | $\beta_{CMA}$ | Adj. $R^2$ |
|---|---|---|---|---|---|---|---|---|
| Deciles | $\Delta^{HVZ}$ | 0.0171 | 0.0091 | −0.6856*** | 0.2007** | 0.6117*** | −0.3588* | 0.4662 |
| | $\Delta^{HVZ,net}$ | 0.0165 | 0.0085 | −0.6858*** | 0.2008** | 0.6113*** | −0.3564* | 0.4673 |
| | $\Delta^{EP}$ | 0.0115 | −0.0036 | −0.9435*** | 0.3574*** | 0.6878*** | −0.5040** | 0.5277 |
| | $\Delta^{EP,net}$ | 0.0113 | −0.0042 | −0.9429*** | 0.3572*** | 0.6871*** | −0.5025** | 0.5278 |
| | $\Delta^{RI}$ | 0.0093 | 0.0056 | −0.5225*** | 0.1536* | 0.5455*** | −0.3235* | 0.4545 |
| | $\Delta^{RI,net}$ | 0.0090 | 0.0055 | −0.5217*** | 0.1537* | 0.5449*** | −0.3228* | 0.4541 |
| Quintiles | $\Delta^{HVZ}$ | 0.0042 | 0.0419 | −0.4937*** | 0.2782*** | 0.4681*** | −0.4174** | 0.5203 |
| | $\Delta^{HVZ,net}$ | 0.0036 | 0.0420 | −0.4926*** | 0.2778*** | 0.4684*** | −0.4164** | 0.5214 |
| | $\Delta^{EP}$ | 0.0035 | 0.0207 | −0.6627*** | 0.3304*** | 0.3913*** | −0.2418 | 0.5657 |
| | $\Delta^{EP,net}$ | 0.0032 | 0.0205 | −0.6609*** | 0.3300*** | 0.3914*** | −0.2420 | 0.5654 |
| | $\Delta^{RI}$ | 0.0017 | −0.0005 | −0.4398*** | 0.2451*** | 0.4217*** | −0.3761** | 0.5761 |
| | $\Delta^{RI,net}$ | 0.0013 | −0.0006 | −0.4387*** | 0.2453*** | 0.4216*** | −0.3768** | 0.5772 |
| Terciles | $\Delta^{HVZ}$ | 0.0046 | 0.0148 | −0.3261*** | 0.1913*** | 0.2354*** | −0.2078 | 0.4079 |
| | $\Delta^{HVZ,net}$ | 0.0042 | 0.0150 | −0.3257*** | 0.1908*** | 0.2357*** | −0.2068 | 0.4080 |
| | $\Delta^{EP}$ | −0.0047 | 0.0100 | −0.5464*** | 0.2270*** | 0.3083*** | −0.1176 | 0.5516 |
| | $\Delta^{EP,net}$ | −0.0047 | 0.0097 | −0.5455*** | 0.2267*** | 0.3084*** | −0.1176 | 0.5512 |
| | $\Delta^{RI}$ | −0.0069 | 0.0318 | −0.3128*** | 0.1509** | 0.2338*** | −0.1361 | 0.4376 |
| | $\Delta^{RI,net}$ | −0.0071 | 0.0320 | −0.3125*** | 0.1506** | 0.2337*** | −0.1354 | 0.4381 |

This table reports regression results for the return differences between the ML portfolio and the other portfolios on the factors provided by Fama and French (2015). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`). I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). HVZ refers to the ICC portfolio based on HVZ model earnings forecasts Hou et al. (2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts Li and Mohanram (2014), and ML refers to the ICC portfolio based on ML model forecasts Hess et al. (2024). Return differences are derived by subtracting the respective portfolio return from the ML portfolio return. Each row refers to one return difference regression. For example, $\Delta^{HVZ}$ ($\Delta^{HVZ,net}$) refers to the regression of gross (net) return differences between the ML portfolio and the HVZ portfolio on the factors provided by Fama and French (2015). ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

**Table 4.6: Transaction costs of portfolio sorts on implied cost of capital**

Panel A: Average transaction costs

| | | HVZ | EP | RI | ML |
|---|---|---|---|---|---|
| Deciles | 1 | 0.0069 | 0.0072 | 0.0071 | 0.0074 |
| | 10 | 0.0070 | 0.0071 | 0.0070 | 0.0071 |
| | 10-1 | 0.0139 | 0.0142 | 0.0141 | 0.0145 |
| Quintiles | 1 | 0.0063 | 0.0065 | 0.0065 | 0.0068 |
| | 5 | 0.0063 | 0.0064 | 0.0064 | 0.0064 |
| | 5-1 | 0.0126 | 0.0130 | 0.0128 | 0.0133 |
| Terciles | 1 | 0.0057 | 0.0058 | 0.0058 | 0.0061 |
| | 3 | 0.0055 | 0.0056 | 0.0055 | 0.0054 |
| | 3-1 | 0.0112 | 0.0114 | 0.0113 | 0.0115 |

Panel B: Average turnover

| | | HVZ | EP | RI | ML |
|---|---|---|---|---|---|
| Deciles | 1 | 0.7443 | 0.7771 | 0.7753 | 0.7993 |
| | 10 | 0.7778 | 0.7828 | 0.7766 | 0.7932 |
| | 10-1 | 1.5221 | 1.5599 | 1.5519 | 1.5925 |
| Quintiles | 1 | 0.6742 | 0.7045 | 0.6949 | 0.7296 |
| | 5 | 0.6976 | 0.7095 | 0.7044 | 0.7179 |
| | 5-1 | 1.3718 | 1.4140 | 1.3993 | 1.4475 |
| Terciles | 1 | 0.5987 | 0.6150 | 0.6169 | 0.6388 |
| | 3 | 0.5995 | 0.6141 | 0.6046 | 0.6009 |
| | 3-1 | 1.1982 | 1.2290 | 1.2216 | 1.2397 |

Panel C: Average transaction cost estimate

| | | HVZ | EP | RI | ML |
|---|---|---|---|---|---|
| Deciles | 1 | 0.9007 | 0.8992 | 0.9009 | 0.9048 |
| | 10 | 0.9087 | 0.9084 | 0.9080 | 0.9080 |
| | 10-1 | 0.9047 | 0.9038 | 0.9045 | 0.9064 |
| Quintiles | 1 | 0.9016 | 0.9007 | 0.9017 | 0.9040 |
| | 5 | 0.9078 | 0.9080 | 0.9073 | 0.9070 |
| | 5-1 | 0.9047 | 0.9044 | 0.9045 | 0.9055 |
| Terciles | 1 | 0.9016 | 0.9015 | 0.9023 | 0.9036 |
| | 3 | 0.9069 | 0.9071 | 0.9066 | 0.9059 |
| | 3-1 | 0.9043 | 0.9043 | 0.9045 | 0.9047 |

This table reports and dissects the average annual transaction costs per (long/short leg of the) ICC long-short portfolio. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024). Panel A shows the average annual transaction costs per (long/short leg of the) ICC long-short portfolio. In a given year, the transaction costs are derived as the sum of the products of stock-specific turnover times the respective stock-specific transaction costs. Stock-specific transaction costs are estimated following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). Values are denoted in decimals, i.e., 0.01 denotes 1%. Panel B shows the average annual turnover per (long/short leg of the) ICC long-short portfolio. In a given year, turnover is derived as the absolute sum of stock weights minus the respective lagged stock weights, adjusted for the respective returns. Values are denoted in decimals, i.e., 0.01 denotes 1%. Panel C shows the average annual transaction costs per stock included in the respective (long/short leg of the) ICC long-short portfolio. Stock-specific transaction costs are estimated for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). Values are denoted in percentage points, i.e., 1 denotes 1%.

costs. This suggests that while the ML model achieves accuracy gains, these come at the expense of slightly elevated transaction costs. However, these differences compared to the other models are minimal, with a maximum deviation of only around 0.06%. The table further shows that in general, the transaction costs are slightly more concentrated in the short-legs of the portfolio strategies.

Total transaction costs are the product of turnover and the respective cost parameter. Breaking them down accordingly helps to determine the causes of differences and reveals potentially important insights in general. Panel B in Table 4.6 shows average annual turnover per portfolio. Turnover decreases with quantile size, from around 152-159% for the decile portfolios to around 120-124% for the tercile portfolios.[20] Regardless of the quantile split examined, the ML portfolio consistently experiences the highest average annual turnover, exhibiting differences of approximately 3.26-7.04 percentage points compared to the other portfolios in the decile split scenario. Panel C in Table 4.6 contains the average transaction cost parameter estimates. Interestingly, the overall average transaction cost parameter of the ML portfolio is the highest for each quantile split considered. As outlined above, the parameter is a function of time and firm size. Trivially, time does not vary among quantiles or portfolios. Any differences in average cost estimates are thus due to differences in firm sizes. Consequently, the results imply that the ML portfolio invests into smaller firms to a larger extent than the other portfolios, on average.

Overall, the results of the empirical analysis stress the fact that more accurate and less distorted forecasts lead to more profitable ICC investment strategies. Importantly, however, this holds true, even when accounting for transaction costs. In fact, the portfolio based on the best performing forecast model, the ML portfolio, only incurs a minor increase in transaction costs as compared to the other portfolios based on less accurate and more distorted earnings forecast models on average.

## 4.5 Extending the empirical analysis

### 4.5.1 The idea

Consistent with prior studies, I observe a positive relationship between the predictive performance of model-based earnings forecasts and the performance of the resulting

[20] Both the long- and the short-leg of a portfolio can incur 100% turnover each.

ICC portfolio. Notably, this relationship persists even after accounting for transaction costs. Nevertheless, it is important to acknowledge limitations in the analysis, primarily stemming from the fact that the main empirical analysis of this study does not differentiate between the relation of general accuracy and portfolio returns and the relation of systematic distortions and portfolio returns. The ML model beats the traditional models in both aspects, making it impossible to disentangle the effects. To provide an example which clarifies the difference: one of the key advantages of model-based forecasts over analyst-based forecasts is the fact that they are less biased (Hou et al., 2012). Studies further show that despite being less accurate than analyst-based forecasts, model-based forecasts translate into significantly more profitable gross return spreads (e.g., Hou et al., 2012; Hess and Wolf, 2022). This shows that it is not only general accuracy, but also other characteristics of forecasts that determine the success of investment strategies conditional on them. One further well-known example of such a characteristic of (model-based) earnings forecasts which might impact portfolio performance, is that they are far less accurate for small firms than for large firms.

Some studies developing earnings forecast models cover specific forms of systematic distortion such as bias or the aforementioned accuracy difference between small and large firms in addition to general accuracy (e.g., Hou et al., 2012; Li and Mohanram, 2014). However, to the best of my knowledge, no study aims to measure systematic distortions and their impact on portfolio performance in general. I aim to fill this gap and, based on a simulation framework, study the effect of achieving both zero median PAFE (perfect general accuracy) and a SYSD of zero (no systematic distortions) on ICC portfolio returns.[21] This analysis indicates the return-decreasing effects that both dimension of accuracy have and shows how completely mitigating them would theoretically affect ICC portfolio performance.

By introducing a second simulation framework, I then dig deeper into the relations between the two dimensions of model performance and ICC portfolio returns. More precisely, I approximate the effect of gradually de-/increasing the median PAFE while holding the SYSD constant, and vice versa. By doing so, I do not only examine how achieving perfect accuracy along either of the two dimensions, but also how e.g., reducing

[21] de Azevedo (2018) also assesses ICC portfolios based on perfect forecasts. However, first and foremost, he does not consider transaction costs and their relation to accuracy. Second, he only considers perfect general accuracy and does not assess perfectly undistorted forecasts. Third, he analyzes such portfolios in the context of deriving the effect of analyst forecast errors on ICC. In contrast, this study is concerned with the effect of model-based forecast performance on ICC.

the median PAFE by 50% while holding the SYSD constant impacts ICC portfolio returns.

In sum, these analyses provide evidence for how exactly the two dimensions of accuracy, i.e., general accuracy and systematic distortions, relate to ICC portfolio returns.

### 4.5.2 Systematic distortions versus general accuracy

I propose a straightforward methodology for determining the impact of completely mitigating systematic distortions versus achieving perfect general accuracy.[22] In a first step, I generate earnings forecasts that exhibit uniform accuracy and no bias across all types of firms. More precisely, for each of the four forecast models considered, I generate a set of forecasts which matches the respective level of general accuracy for each forecast horizon, but exhibits zero systematic distortions. I do so by following the same procedure as in the case of the newly introduced metric measuring systematic distortions above (SYSD). I then assess the average spread returns of ICC portfolios based on these forecasts. The disparity in spreads between portfolios derived from actual model forecasts and their corresponding undistorted forecasts, which align with comparable levels of general accuracy, illustrates the potential enhancement in portfolio performance attainable through the complete mitigation of systematic distortions, while maintaining consistent levels of general accuracy.

In a second step, I derive forecasts which are not exposed to systematic distortions and resemble perfect general accuracy. Trivially, these are the actual future earnings realizations. I again compute the average returns of ICC portfolios based on these forecasts. The difference between these average returns and the average returns based on undistorted forecasts with some level of general accuracy present indicates the increase in portfolio performance that may be achieved by completely mitigating general inaccuracies while holding the level of systematic distortions constant (at a level of zero).

In sum, I disentangle the effects of achieving perfect accuracy while holding the degree of systematic distortions constant, and vice versa. An important consideration is that this substantially alters the sample for which ICC can be computed. This stems from the necessity of future earnings data for simulating earnings forecasts with both zero SYSD and zero median PAFE. However, I address this concern in my subsequent comparison by restricting the analysis to firm-year observations for which simulated forecasts exist. Put differently, I reduce my actual model forecast sample to the firm-year

[22] Note that the latter implies the former, but the former does not imply the latter.

observations for which simulated forecasts exist. This ensures a fair and meaningful comparison while mitigating any potential biases introduced by the alteration in the sample composition.

Figure 4.4 depicts the result.[23] The figure shows the average ICC portfolio return spreads based on the actual forecasts, based on simulated undistorted forecasts that replicate the median PAFE of the respective actual forecast model for each forecast horizon, and based on undistorted and perfectly accurate forecasts. First, it becomes evident that the aforementioned sample change influences average return spreads. In fact, return spreads based on actual earnings forecasts increase significantly as compared to the spreads for the original sample. This is attributable to the survivorship bias that the simulation introduces. Surprisingly, the ML portfolio is not the most profitable portfolio anymore. However, this does not devalidate the results in section 4.4, as the sample in this exercise is unrepresentative and biased. More so, it implies that the ML forecast model did particularly well for those firms that are excluded in the simulation sample.

Second, albeit general return levels are distorted due to the sample changing, the results reveal an important insight: systematic distortions of model-based forecasts, such as imbalanced accuracy across firm sizes or bias, exert a substantial influence on investment performance. When simulating undistorted earnings forecasts with the same level of general accuracy as their respective actual forecast counterparts, gross (net) average return spreads increase by around 81% (88%) to 213% (261%), depending on the model, i.e., the level of general accuracy, and the quantile split considered. This effect is very pronounced for every forecast model, making it conceivable that it translates into an unbiased sample.[24] Achieving perfect general accuracy yields an additional gross (net) average return spread increase of around 17% (18%) to 33% (35%). Lastly and importantly, transaction costs do not change significantly when completely mitigating systematic distortions or when achieving perfect general accuracy.

In line with the findings by Hou et al. (2012) regarding the impact of forecast bias, the results illustrate the major impact of systematic distortions on ICC portfolio returns. Model forecasts which exhibit some median PAFE but are not systematically distorted, i.e., exhibit no bias and perform equally for every type of firm, lead to stellar ICC

[23] Table C.5 in the Appendix provides a tabular summary of the results.

[24] This is an assumption which I consider conceivable due to the strength of the effect and its consistency across models. Unfortunately, there is no way to test this for an unbiased sample. Apart from developing actual models that exhibit the same exact general accuracy of existing models while not being affected by systematic distortions, going the simulation route is the only way to empirically test this.

portfolio improvements. In fact, such forecasts come close to perfect forecasts in terms of ICC portfolio performance based on them.

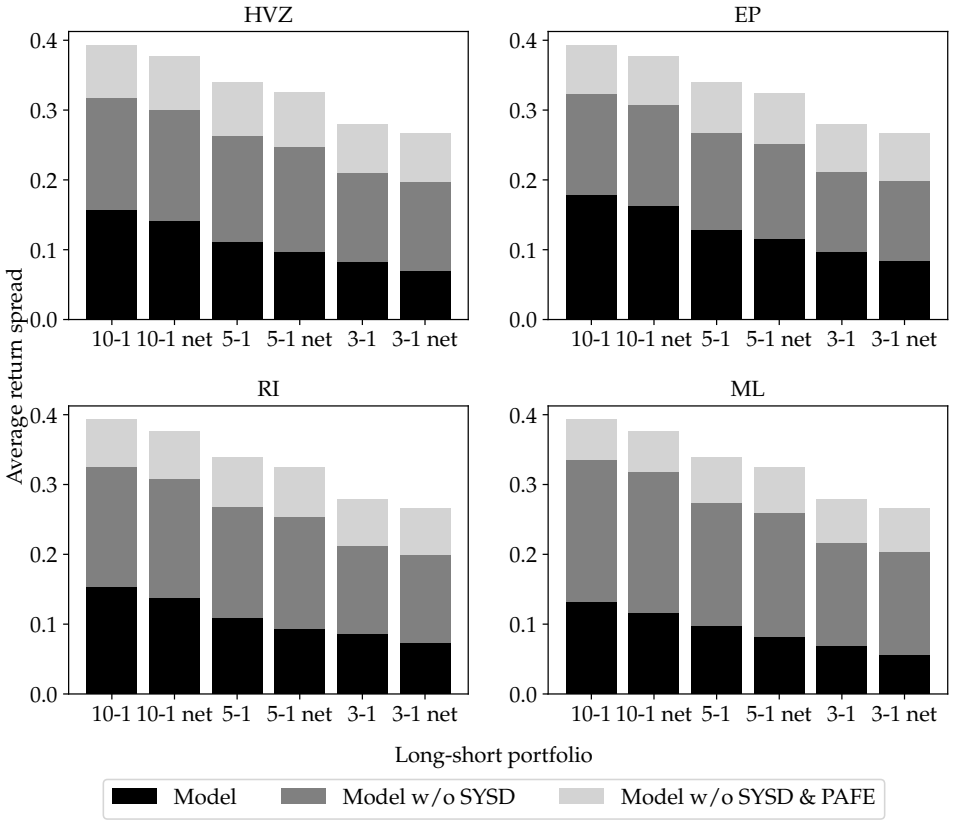



**Figure 4.4: No systematic distortions versus perfect general accuracy**
This figure illustrates the effect of forecasting earnings with zero systematic distortions (SYSD) versus zero median price scaled absolute forecast error (PAFE) on both gross and net long-short ICC portfolio returns based on the respective set of forecasts. The sample only includes observations for which simulated ICC are available. I consider three quantile splits, i.e., decile splits (10-1), quintile splits (5-1), and tercile splits (3-1). HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024). Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). The undistorted forecasts are derived by simulating forecasts which match the respective model's median PAFE for each forecast horizon. The undistorted and fully accurate forecasts are derived by simulating forecasts which are equivalent to the future earnings realizations for each forecast horizon.

### 4.5.3 Approximating the relations

The effect of both completely mitigating systematic distortions and achieving perfect general accuracy is consistent across all forecast models considered. However, apart from being simulation-based, the analysis above faces the drawback of only comparing two scenarios for both dimension of accuracy: general accuracy and systematic distortions on some model-level and perfect general accuracy and/or zero systematic distortions,

respectively. In other words, I only assess the effect of theoretically being able to completely mitigate either or both types of inaccuracies. However, this does not necessarily reveal how e.g., decreasing the median PAFE or the SYSD by 20% affects ICC portfolio returns. I provide a simple framework which allows me to approximate the exact relation between general accuracy as well as systematic distortions and ICC portfolio returns. More precisely, I simulate earnings forecasts with varying median PAFE while holding the SYSD constant, and vice versa.

To achieve varying median PAFEs while fixing the SYSD, I simulate forecasts as above, i.e., I simulate forecasts with SYSD of zero and some fixed median PAFE. More precisely, I simulate zero SYSD forecasts with median PAFEs matching the ML median PAFEs for each forecast horizon multiplied by a weighting scalar $w_1 \in [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$.[25] The weight $w_1$ is linearly related to the median PAFE. Thus, if e.g., $w_1 = 0.5$, I simulate earnings forecasts that exhibit half of the median PAFE of the ML model for forecast horizons $\tau = 1$ to $\tau = 5$. Note that by construction, every simulated forecast has a SYSD of zero. Put differently, the SYSD is fixed at a level of zero, while I gradually increase median PAFE from zero to a level which is equal to that of the ML model.[26]

The results are depicted in Figure 4.5.[27] The simulation suggests that the average ICC portfolio return is somewhat linearly related to the level of general accuracy of the respective earnings forecast model. In fact, if the median PAFE increases by 1 percentage point, the average gross (net) ICC portfolio return spread decreases by around 0.06 (0.06) to 0.07 (0.07) percentage points, depending on the split scenario considered. Again, the results indicate that transaction costs do not correlate with general accuracy.

To achieve varying SYSDs while fixing the median PAFE, I simulate forecasts as follows: let $\hat{E}^{ml}_{i,t+\tau}$ denote the ML model forecast for firm $i$ and period $t+\tau$, and let $\hat{E}^{ud}_{i,t+\tau}$ denote the corresponding undistorted forecast, derived via simulation as outlined above.[28] I then derive earnings forecasts with varying degrees of systematic distortions while fixing the median PAFE at the ML model level ($\hat{E}^{sysd}_{i,t+\tau}$) by setting

$$\hat{E}^{sysd}_{i,t+\tau} = w_2 \hat{E}^{ml}_{i,t+\tau} + (1 - w_2)\hat{E}^{ud}_{i,t+\tau}. \tag{4.8}$$

[25] I choose a model median PAFE and multiply it with some scalar to simulate a realistic relative median PAFE increase for increasing forecast horizons. However, the choice of which model to use as a baseline, in this case the ML model, is arbitrary.

[26] Using this approach, I can only fix the SYSD at zero. This might be a limitation as the relation between the median PAFE and ICC portfolio returns might change for different SYSD levels.

[27] Table C.6 in the Appendix provides a tabular summary of the results.

[28] The results are robust across earnings forecast models, i.e., fixing the median PAFE at different levels.

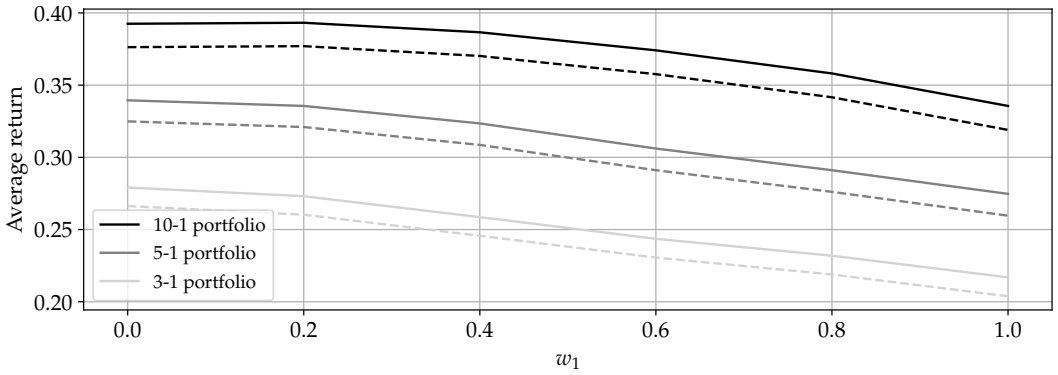



**Figure 4.5: General accuracy and average returns**

This figure shows both gross and net average return spreads of ICC long-short portfolios based on simulated forecasts with varying median PAFEs and fixed SYSD. The sample only includes observations for which simulated ICC are available. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Straight (dashed) lines resemble average gross (net) returns. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). The simulated forecasts are derived by simulating undistorted forecasts which match the median ML model PAFE for each forecast horizon, multiplied by a scalar $w_1$. The relationship between $w_1$ and the median PAFE is linear. For example, if $w_1 = 0.5$, the respective simulated model median PAFEs are half of the median PAFEs of the ML model. Since each of the simulated forecasts is undistorted by construction, the SYSD is fixed at zero.

$w_2 \in [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$ denotes the weight that I put on the ML model forecast. In Table C.6 in the Appendix I show that the weight $w_2$ and SYSD are approximately linearly related. In other words, setting $w_2$ to, for instance, 0.5, results in a SYSD which is around half the SYSD of the respective forecast model considered.[29]

The results are depicted in Figure 4.6.[30] The relationship appears to be concave. This might be attributable to the fact that the median PAFE is convexly related to $w_2$. Nonetheless, the results once again stress that systematic distortions have a significant impact on ICC portfolio returns. More precisely, the figure suggests that not only complete mitigation, but also partial reduction of systematic distortion yields substantial ICC portfolio return gains. Importantly, transaction costs do not appear to correlate with the degree of systematic distortions.

In summary, the simulation exercise yields two key findings: first, improving earnings forecasts both in terms of general accuracy and systematic distortions substantially improves ICC portfolio returns. Put differently and as the findings by Hou et al. (2012) suggest, general accuracy of earnings forecasts is not the only forecast characteristic which matters in the ICC portfolio context. And second, transaction costs are neither

[29]The median PAFE is also only approximately constant due to the positive effect of model stacking. To be precise, the median PAFE of a linear combination of two models with the same median PAFE is equal to or lower than a corresponding linear combination of the two respective median PAFEs. Put differently, the relationship between the weight $w_2$ and the median PAFE is slightly convex. However, the effect is negligible in this context.

[30]Table C.6 in the Appendix provides a tabular summary of the results.

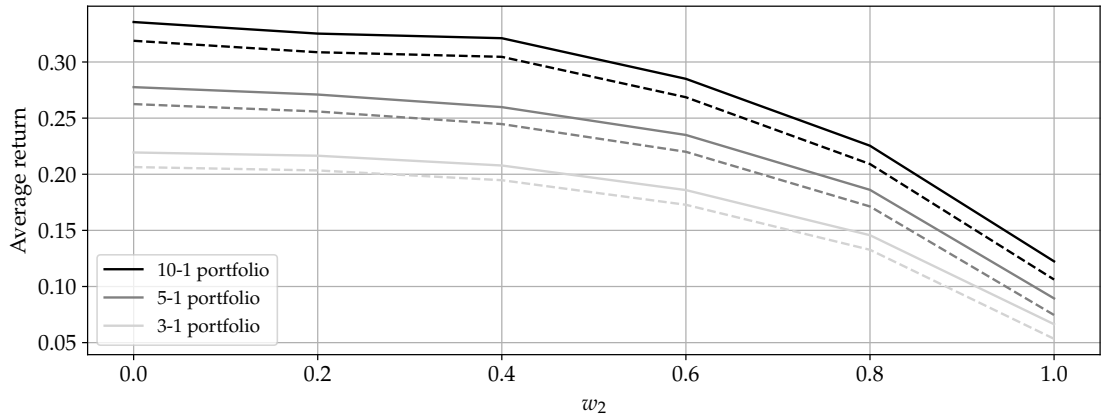



**Figure 4.6: Systematic distortions and average returns**

This figure shows both gross and net average return spreads for the portfolios based on simulated forecasts with varying SYSD and (approximately) fixed median PAFE. The sample only includes observations for which simulated ICC are available. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Straight (dashed) lines resemble average gross (net) returns. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). The simulated forecasts are derived by computing a weighted average of the actual ML model forecast and the corresponding simulated undistorted forecast with matching median PAFE for each forecast horizon. The weight given to the ML model forecast is equal to $w_2$ and the weight given to the corresponding undistorted forecast is equal to $1 - w_2$. The relationship between $w_2$ and SYSD is approximately linear. For example, if $w_2 = 0.5$, the respective simulated model SYSD is approximately half of the SYSD of the ML model. The relationship between the median PAFE and $w_2$ is slightly convex to due to the effect of model stacking.

related to general accuracy nor systematic distortions.

These results bear important implications for research efforts aimed at improving earnings forecasts against the background of using them for investment strategies. First, improving earnings forecast models is a promising research direction, as such improvements translate into significant investment performance gains. Second, researchers should assess the degree to which models are exposed to systematic distortions, as this appears to be a core driver behind the success of ICC long-short portfolios. Unsupervised learning methods, such as clustering, and subsequent separate model estimation for each of these clusters, might prove useful in mitigating systematic distortions. However, it has to be noted that systematic distortions are not necessarily mitigable to the full extent. This is because e.g., accuracy differences across different subsets of firms may not occur due to the model being systematically flawed but due to different subsets of firms exhibiting different degrees of unpredictable noise in future earnings.

## 4.6 Conclusion

This study is the first to provide a detailed assessment of the relation between model-based earnings forecast accuracy and ICC investment performance against the back-

ground of transaction costs.

I explicitly differentiate between two aspects of model accuracy, i.e., general accuracy and systematic distortions, providing a novel metric for measuring the latter. The results of the empirical analysis reveal that the most accurate earnings forecast model, a ML model, is also the least systematically distorted model. In line with the extant literature on earnings forecast models, the best performing earnings forecast model yields the highest gross return spreads. Importantly, this holds true even when accounting for transaction costs. Nevertheless, in line with the extant literature on return anomalies, transaction costs lead to significantly lower average return spreads, thereby stressing the importance of incorporating transaction costs to future research which involves ICC portfolios (e.g., Chen and Velikov, 2023). Interestingly, transaction costs do not strongly differ between the portfolios based on the different earnings forecast models.

By leveraging simple simulation frameworks, I assess the effect of general accuracy and systematic distortions separately. The analysis reveals that both systematic distortions of model-based earnings forecasts and general accuracy strongly impact average investment performance. In other words, improvements along both dimensions of model performance lead to substantial ICC portfolio return gains. Mirroring the aforementioned results, the findings further indicate that transaction costs do not change with varying levels of systematic distortions or general accuracy.

To conclude, the study provides strong evidence for the fact that better earnings forecast models translate into more profitable investment strategies. Furthermore, I show that not only general accuracy, but also systematic model distortions such as unequal accuracy across subsets of firms or systematic over- or underestimation of future earnings, strongly influence investment performance conditional on earnings forecasts. Transaction costs significantly alter investment performance in general, but are neither correlated with general accuracy nor systematic distortions.

# Appendix A

# Appendix to Chapter 2

## A.1 Neural network configuration

Our benchmark model consists of an input layer, three hidden layers and an output layer. We apply the geometric pyramid rule (Masters, 1993), i.e., the first hidden layer consists of 32 nodes, the second hidden layer consists of 16 nodes and the third hidden layer consists of eight nodes. We consider different network architectures in Appendix A.2.

At each node of the network, a linear transformation of the preceding outputs is fed into an activation function. We choose to use the leaky rectified linear unit (leaky ReLU) activation function at every node (e.g., Jarrett et al., 2009):

$$R(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha z & \text{otherwise} \end{cases}, \tag{A.1}$$

where $z$ denotes the input and $\alpha$ denotes some small non-zero constant, in our case 0.01. ReLU is the most popular activation function because it is cheap to compute, converges fast and is sparsely activated. The disadvantage of transforming all negative values to zero is a problem called "dying ReLU". A ReLU neuron is "dead" if it is stuck in the negative range and always outputs zero. Since the slope of ReLU in the negative range is also zero, it is unlikely that a neuron will recover once it goes negative. Such neurons play no role in discriminating inputs and are essentially useless. Over time, a large part of the network may do nothing. Leaky ReLU fixes this problem because it has small slope for negative values instead of a flat slope. Moreover, we shift the activation function at every node in every hidden layer by adding a constant. This is commonly referred to as

bias in the machine learning literature.

Our benchmark network is estimated by minimizing the loss function (utility function) given in Equation (2.6). To do so, we apply the commonly used ADAM stochastic gradient descent optimization technique developed by Kingma and Ba (2014).

To control for the non-linearity and heavy parametrization of the model, we employ different regularization techniques to prevent overfitting: first, as mentioned above, we impose a constraint on an individual stock's absolute portfolio weight of $|3\%|$.

Second, we add a Lasso ($l_1$) penalty term to the loss function to be minimized. Adding the penalty implies a potential shrinkage of coefficients towards 0. This in turn reduces the variance of the prediction, i.e., prevents overfit of the model.

Third, we employ early stopping on the validation data. Early stopping refers to a very general regularization technique. At each new iteration, predictions are estimated for the validation sample, and the loss (utility) is derived. The optimization is terminated when the validation sample loss starts to increase by some small specified number (tolerance) over a specified number of iterations (patience). Typically, the termination occurs before the loss is minimized in the training sample. Early stopping is a popular regularization tool because it reduces the computational cost.

Fourth, we implement a dropout layer before the first hidden layer (Srivastava et al., 2014). As Srivastava et al. (2014) state, the basic idea of dropout is to randomly remove neurons (and their connections) from the neural network during training. This promotes individual feature learning and the network becomes less sensitive to specific neuron weights. When generating predictions, the neuron weights are scaled down by the dropout rate, i.e., the chance by which neurons are randomly getting dropped out, to account for the fact that each neuron essentially gets trained less due to dropping them out randomly. The combination of a dropout layer, $l_1$-regularization and early stopping tremendously helps to reduce overfitting and model complexity.

Fifth, we adopt an ensemble approach in training our neural network (e.g., Hansen and Salamon, 1990). In particular, we initialize five neural networks with different random seeds and construct predictions by averaging the predictions from all networks. This reduces the variance across predictions since different seeds produce different predictions due to the stochastic nature of the optimization process.

Finally, we adopt our own version of the batch normalization method by Ioffe and Szegedy (2015). In general, training deep neural networks is complicated by the fact

that the distribution of inputs to each layer changes during training as the parameters of the previous layers change. This is referred to as internal covariate shift and can be remedied by normalizing the layer inputs. The standard batch normalization following Ioffe and Szegedy (2015) makes this normalization part of the model architecture and performs it for each training mini-batch. Batch normalization allows much higher learning rates to be used and less care to be taken in initialization of the network (Ioffe and Szegedy, 2015). Brandt et al. (2009) standardize characteristics cross-sectionally to have zero mean and unit standard deviation across all stocks at date $t$. Hence, the model predictions represent deviations from the benchmark portfolio. However, applying the aforementioned activation function destroys this structure. In our model each observation can be interpreted as a complete cross-section (e.g., a batch size of 12 refers to 12 complete cross-sections of data). The model of Brandt et al. (2009) hence requires normalization on a cross-sectional level instead of a batch level. To account for that, we standardize cross-sectionally after applying the activation function in each hidden layer, such that the output of each node in the hidden layer has zero mean and unit standard deviation across all stocks at the respective date $t$. Thus, the output of each node in each hidden layer can also be interpreted as a deviation from the benchmark portfolio. We provide a summary of the relevant hyperparameters in Table A.1.

**Table A.1: Hyperparameters**

| | PPP | DPPP |
|---|---|---|
| L1 penalty | $l_1 \in \{0, 10^{-5}, 10^{-3}\}$ | $l_1 \in \{0, 10^{-5}, 10^{-3}\}$ |
| Learning rate | 0.001 | 0.001 |
| Dropout | 0 | $D \in \{0, 0.2, 0.4\}$ |
| Batch size | 12 | 12 |
| Epochs | 200 | 200 |
| Patience | 20 | 20 |
| Ensemble | 0 | 5 |
| Leaky ReLU | – | 0.01 |

This table gives the hyperparameters that we tune. The first column shows the hyperparameters for the linear parametric portfolio policy (PPP). The second column shows the hyperparameters for the deep parametric portfolio policy (DPPP).

## A.2 Robustness checks

### A.2.1 Benchmark comparison

For robustness, we compare the (D)PPP for a CRRA investor with a relative risk aversion of $\gamma = 5$ to an equally (EW) and value-weighted (VW) benchmark portfolio.

Table A.2 presents the comparison between the different portfolios based on their utility, weights and return characteristics. The first row reports the certainty equivalent of the realized utility across out-of-sample periods for a CRRA investor with relative risk aversion of five. The equally weighted and value weighted portfolio yield a certainty equivalent of 0.0015 and 0.0022, respectively. The standard PPP substantially outperforms the simple portfolios, yielding a certainty equivalent of 0.0263. However, the DPPP yields a certainty equivalent of 0.0492, almost twice as large as the certainty equivalent derived from the PPP.

The next set of rows gives insight into the distribution of the respective portfolio weights. The active portfolios take comparably large positions, with the average absolute weight of the deep portfolio policy being almost nine times as large as in the case of the equally weighted and value weighted portfolio, respectively. However, due to the weight constraint shown in Equation (2.7) these positions remain below 3% in absolute terms. As Ang et al. (2011) show, average gross leverage of hedge fund companies amounts to 120% in the period after the financial crisis 2007-2008. This indicates that both the linear and the deep portfolio policies are rather unrealistic in the benchmark case. We address this in Section 2.4.2 by including a penalty term for transaction costs and a constraint for leverage in our objective function.

The monthly mean returns of 4.7% and 7.1% in the linear and deep policy case are much higher than the mean returns of around 1.1% in the equally weighted and value weighted portfolio cases due to their highly levered nature. In fact, both models substantially outperform the market porfolios with more than twice as large Sharpe ratios. In terms of skewness and kurtosis the DPPP stands out as compared to the other portfolios. In particular, the portfolio exhibits a positive skewness (0.82) and high kurtosis (4.96). The bottom set of rows reports the alphas and its standard errors with respect to a six-factor model that appends a momentum factor to the Fama-French five-factor model. The market portfolio alphas are both not significantly different from zero. The linear policy alpha is 3.2%. The deep policy alpha is even higher, amounting to 5.7%. Both

alphas are highly statistically significant. These large unexplained returns can partially be attributed to the highly levered nature of the active portfolios, as we show in the following sections.

**Table A.2: (D)PPP versus market portfolios**

| | EW | VW | PPP | DPPP |
|---|---|---|---|---|
| CE | 0.0015 | 0.0022 | 0.0263 | 0.0492 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | | | 0.0002 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.0694 | 0.0694 | 0.4972 | 0.6127 |
| $max\ w_i * 100$ | 0.0704 | 0.1113 | 2.0363 | 1.7452 |
| $min\ w_i * 100$ | 0.0704 | 0.0410 | −2.1712 | −1.8709 |
| $\sum_i w_i I(w_i < 0)$ | 0.0000 | 0.0000 | −3.0841 | −3.9171 |
| $\sum_i I(w_i < 0)/N_t$ | 0.0000 | 0.0000 | 0.4351 | 0.4430 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 0.0931 | 0.0779 | 3.7816 | 7.8053 |
| Mean | 0.0110 | 0.0105 | 0.0473 | 0.0711 |
| StdDev | 0.0587 | 0.0552 | 0.0890 | 0.0982 |
| Skew | −0.3716 | −0.5039 | −0.1004 | 0.8169 |
| Kurt | 3.6591 | 3.3455 | 1.3766 | 4.9609 |
| SR | 0.6461 | 0.6609 | 1.8391 | 2.5101 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | | | 0.0075 |
| $FF5 + Mom\ \alpha$ | −0.0002 | −0.0003 | 0.0324 | 0.0570 |
| $StdErr(\alpha)$ | 0.0007 | 0.0006 | 0.0040 | 0.0052 |

This table shows out-of-sample estimates of the (deep) portfolio policies optimized for a CRRA investor with relative risk aversion of five, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "EW", "VW", "PPP" and "DPPP" show the statistics of the equal-weighted portfolio, value-weighted portfolio, parametric portfolio policy, and deep parametric portfolio policy, respectively. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

### A.2.2 Long-only

A large majority of equity portfolios face restrictions on short selling. We incorporate short-sale constraints as in Brandt et al. (2009), i.e., we truncate portfolios weights at zero (and still keep the cap of 3% per stock). In particular, to make sure that portfolio weights still sum up to one, we add the following portfolio rebalancing term to the end of our

optimization process:

$$w_{i,t}^{*} = \frac{max[0, w_{i,t}]}{\sum_{j=1}^{N_t} max[0, w_{i,t}]}. \tag{A.2}$$

Table A.3, shows results from estimating a long-only portfolio for CRRA investor with relative risk aversion of $\gamma = 5$. Again, the deep parametric portfolio policy yields the highest certainty equivalent, although it is markedly lower than in the unconstrained case. Still, the certainty equivalent of the deep parametric portfolio policy is around five times higher than the certainty equivalent of the market portfolios and around 43% higher than the certainty equivalent of the linear parametric portfolio policy. The difference between the utility of the deep and the linear parametric portfolio policy is statistically significant at the 0.1% level.

Both active portfolios result in a much higher turnover than the market portfolios, and the deep portfolio policy produces a higher turnover than the linear portfolio policy (124% versus 60%). Different from the unconstrained benchmark results in Table A.2, here we report the fraction of weights that are equal to zero. Interestingly, on average the deep portfolio policy does not include 10% of stocks, while the linear portfolio policy does not include 27% of the available stocks. Thus, the deep portfolio policy invests in more stocks but also has a higher individual maximum weight (1.57% vs 0.36%), indicating that many weights are possibly very low.

The DPPP yields higher expected returns than the PPP, with a moderate increase in volatility resulting in a Sharpe ratio that is around 20% higher than the Sharpe ratio of the linear portfolio policy. This difference is statistically significant at the 0.1% level. Interestingly, the third and fourth moments of all portfolio policies are similar and the portfolio return distributions are not heavily skewed or tailed. Lastly, the alphas of the Fama-French model are a lot smaller compared to the benchmark models, while still being highly significant in both the linear and the deep portfolio policy case. Without the ability to take (potentially extreme) short positions, the estimated parametric portfolios appear to be much more realistic. Nonetheless, the deep portfolio policy still outperforms the other portfolios in terms of realized out-of-sample utility.

### A.2.3 Model complexity

Our benchmark model is a relatively shallow neural net with only three hidden layers. It is conceivable that a more complex model can achieve even higher utility gains over a

linear model. For example, Goodfellow et al. (2016) observe that neural nets with more hidden layers tend to outperform neural nets with fewer hidden layers but more nodes per layer. Kelly et al. (2024) report evidence in support of complex models in the context of forecasting aggregate stock market returns.

We extend our benchmark model to include between two and five hidden layers. All models start with 32 nodes in the first hidden layer and then halve the number of nodes in each subsequent layer. The number of parameters across models therefore varies between 5,600 and 5,768. Additionally, we increase the number of hyperparameters by adding different possible learning rates to our hyperparameter tuning and increasing the number of epochs and patience for early stopping, to account for the different complexities of the models and to ensure that more complex models also reach their respective potential. More specifically, the learning rate is now given by $LR \in \{0.0001, 0.001, 0.01\}$, the number of maximum epochs for which we train is set to 300, and the patience is increased to 30.

Table A.4 shows the results. The second model is our original benchmark model that we added for comparison.[1] The remaining columns contain results based on networks with two, four or five hidden layers. We observe that reducing the number of hidden layers to two reduces the certainty equivalent. This reduction in certainty equivalent is significant at the 5% level. In contrast, increasing the number of hidden layers to four or five, respectively, does not yield statistically significant differences in certainty equivalent. We thus conclude that in general, reasonable complexity adjustments in terms of the number of hidden layers do not lead to significantly different outcomes. However, we note that the testing of more hyperparameter specifications may lead to significant improvements of the DPPP.

[1]Note that the certainty equivalent is higher compared to our benchmark in Section 2.4.1. This is due to the aforementioned fact that we add different possible learning rates as well as increase the number of epochs and patience for early stopping. We do so not only for the model variations, but also for our benchmark to ensure consistency across models.

## A.3 Supplementary tables

**Table A.3: Long-only (D)PPP**

| | EW | VW | PPP | DPPP |
|---|---|---|---|---|
| CE | 0.0015 | 0.0022 | 0.0075 | 0.0107 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | | | 0.0008 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.0694 | 0.0694 | 0.0694 | 0.0694 |
| $max\ w_i * 100$ | 0.0704 | 0.1113 | 0.3578 | 1.5865 |
| $min\ w_i * 100$ | 0.0704 | 0.0410 | 0.0000 | 0.0000 |
| $\sum_i w_i I(w_i < 0)$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| $\sum_i I(w_i = 0)/N_t$ | 0.0000 | 0.0000 | 0.2667 | 0.0972 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 0.0931 | 0.0779 | 0.6019 | 1.2433 |
| Mean | 0.0110 | 0.0105 | 0.0145 | 0.0200 |
| StdDev | 0.0587 | 0.0552 | 0.0506 | 0.0583 |
| Skew | −0.3716 | −0.5039 | −0.6840 | −0.3391 |
| Kurt | 3.6591 | 3.3455 | 3.1303 | 4.3683 |
| SR | 0.6461 | 0.6609 | 0.9931 | 1.1871 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | | | 0.0007 |
| $FF5 + Mom\ \alpha$ | −0.0002 | −0.0003 | 0.0043 | 0.0095 |
| $StdErr(\alpha)$ | 0.0007 | 0.0006 | 0.0007 | 0.0012 |

This table shows out-of-sample estimates of the (deep) portfolio policies including a long-only constraint optimized for a CRRA investor with relative risk aversion of five, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "EW", "VW", "PPP" and "DPPP" show the statistics of the equal-weighted portfolio, value-weighted portfolio, parametric portfolio policy, and deep parametric portfolio policy, respectively. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

**Table A.4: (D)PPP with different numbers of hidden layers**

| | Layer 2 | Layer 3 | Layer 4 | Layer 5 |
|---|---|---|---|---|
| CE | 0.0386 | 0.0633 | 0.0674 | 0.0647 |
| p-value$(CE_{L_i} - CE_{L3})$ | 0.0364 | | 0.1716 | 0.3402 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 1.2431 | 1.1550 | 1.1395 | 0.8481 |
| $max\ w_i * 100$ | 2.2951 | 2.1522 | 2.3668 | 2.2394 |
| $min\ w_i * 100$ | −2.3218 | −2.1872 | −2.3921 | −2.2716 |
| $\sum_i w_i I(w_i < 0)$ | −8.4616 | −7.8263 | −7.7149 | −5.6143 |
| $\sum_i I(w_i < 0)/N_t$ | 0.4757 | 0.4717 | 0.4675 | 0.4568 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 15.5297 | 14.2088 | 14.4381 | 11.0562 |
| Mean | 0.1102 | 0.1108 | 0.1260 | 0.1063 |
| StdDev | 0.1604 | 0.1428 | 0.1695 | 0.1497 |
| Skew | 0.2956 | 0.3956 | 1.1144 | 1.8729 |
| Kurt | 1.7233 | 1.1903 | 4.5579 | 10.5177 |
| SR | 2.3813 | 2.6886 | 2.5756 | 2.4600 |
| p-value$(SR_{L_i} - SR_{L3})$ | 0.0003 | | 0.1130 | 0.0460 |
| $FF5 + Mom\ \alpha$ | 0.0923 | 0.0927 | 0.1091 | 0.0934 |
| $StdErr(\alpha)$ | 0.0088 | 0.0078 | 0.0095 | 0.0086 |

This table shows out-of-sample estimates of the (deep) portfolio policies with different numbers of hidden layers optimized for a CRRA investor with relative risk aversion of five, using 157 firm characteristics. The deep models are feed-forward neural networks with two (32, 16), three (32, 16, 8), four (32, 16, 8, 4) and five (32, 16, 8, 4, 2) hidden layers (nodes), respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "Layer 2", "Layer 3", "Layer 4" and "Layer 5" show the statistics of the deep parametric portfolio policy with two, three, four and five hidden layers, respectively. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between the model with three layers and the other models. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between the model with three layers and the other models. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

**Table A.5: Predictor variables for the (D)PPP**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| ChInvIA | Change in capital inv (ind adj) | Abarbanell and Bushee | 1998, AR | yearly | Accounting | investment growth |
| GrSaleToGrInv | Sales growth over inventory growth | Abarbanell and Bushee | 1998, AR | yearly | Accounting | sales growth |
| GrSaleToGrOverhead | Sales growth over overhead growth | Abarbanell and Bushee | 1998, AR | yearly | Accounting | sales growth |
| IdioVolAHT | Idiosyncratic risk (AHT) | Ali, Hwang, and Trombley | 2003, JFE | monthly | Price | volatility |
| EarningsConsistency | Earnings consistency | Alwathainani | 2009, BAR | yearly | Accounting | earnings |
| Illiquidity | Amihud's illiquidity | Amihud | 2002, JFM | monthly | Trading | liquidity |
| BidAskSpread | Bid-ask spread | Amihud and Mendelsohn | 1986, JFE | monthly | Trading | liquidity |
| grcapx | Change in capex (two years) | Anderson and Garcia-Feijoo | 2006, JF | yearly | Accounting | investment growth |
| grcapx3y | Change in capex (three years) | Anderson and Garcia-Feijoo | 2006, JF | yearly | Accounting | investment growth |
| betaVIX | Systematic volatility | Ang et al. | 2006, JF | monthly | Price | volatility |
| IdioRisk | Idiosyncratic risk | Ang et al. | 2006, JF | monthly | Price | volatility |
| IdioVol3F | Idiosyncratic risk (3 factor) | Ang et al. | 2006, JF | monthly | Price | volatility |
| CoskewACX | Coskewness using daily returns | Ang, Chen and Xing | 2006, RFS | monthly | Price | risk |
| Mom6mJunk | Junk Stock Momentum | Avramov et al | 2007, JF | monthly | Price | momentum |
| OrderBacklogChg | Change in order backlog | Baik and Ahn | 2007, Other | yearly | Accounting | accruals |
| roaq | Return on assets (qtrly) | Balakrishnan, Bartov and Faurel | 2010, JAE | quarterly | Accounting | profitability |
| MaxRet | Maximum return over month | Bali, Cakici, and Whitelaw | 2010, JF | monthly | Price | volatility |
| ReturnSkew | Return skewness | Bali, Engle and Murray | 2015, Book | monthly | Price | risk |
| ReturnSkew3F | Idiosyncratic skewness (3F model) | Bali, Engle and Murray | 2015, Book | monthly | Price | risk |
| CBOperProf | Cash-based operating profitability | Ball et al. | 2016, JFE | yearly | Accounting | profitability |
| OperProfRD | Operating profitability R&D adjusted | Ball et al. | 2016, JFE | yearly | Accounting | profitability |
| Size | Size | Banz | 1981, JFE | monthly | Price | size |
| SP | Sales-to-price | Barbee, Mukherji and Raines | 1996, FAJ | yearly | Accounting | valuation |
| EP | Earnings-to-Price Ratio | Basu | 1977, JF | monthly | Price | valuation |
| InvGrowth | Inventory Growth | Belo and Lin | 2012, RFS | yearly | Accounting | profitability |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| BrandInvest | Brand capital investment | Belo, Lin and Vitorino | 2014, RED | yearly | Accounting | investment |
| Leverage | Market leverage | Bhandari | 1988, JFE | monthly | Price | leverage |
| ResidualMomentum | Momentum based on FF3 residuals | Blitz, Huij and Martens | 2011, JEmpFin | monthly | Price | momentum |
| Price | Price | Blume and Husic | 1972, JF | monthly | Price | other |
| NetPayoutYield | Net Payout Yield | Boudoukh et al. | 2007, JF | monthly | Price | valuation |
| PayoutYield | Payout Yield | Boudoukh et al. | 2007, JF | monthly | Price | valuation |
| NetDebtFinance | Net debt financing | Bradshaw, Richardson, Sloan | 2006, JAE | yearly | Accounting | external financing |
| NetEquityFinance | Net equity financing | Bradshaw, Richardson, Sloan | 2006, JAE | yearly | Accounting | external financing |
| XFIN | Net external financing | Bradshaw, Richardson, Sloan | 2006, JAE | yearly | Accounting | external financing |
| DolVol | Past trading volume | Brennan, Chordia, Subra | 1998, JFE | monthly | Trading | volume |
| FEPS | Analyst earnings per share | Cen, Wei, and Zhang | 2006, WP | monthly | Analyst | profitability |
| AnnouncementReturn | Earnings announcement return | Chan, Jegadeesh and Lakonishok | 1996, JF | monthly | Price | earnings |
| REV6 | Earnings forecast revisions | Chan, Jegadeesh and Lakonishok | 1996, JF | monthly | Analyst | earnings |
| AdExp | Advertising Expense | Chan, Lakonishok and Sougiannis | 2001, JF | monthly | Accounting | R&D |
| RD | R&D over market cap | Chan, Lakonishok and Sougiannis | 2001, JF | monthly | Accounting | R&D |
| CashProd | Cash Productivity | Chandrashekar and Rao | 2009, WP | yearly | Accounting | profitability |
| std_turn | Share turnover volatility | Chordia, Subra, Anshuman | 2001, JFE | monthly | Trading | liquidity |
| VolSD | Volume Variance | Chordia, Subra, Anshuman | 2001, JFE | monthly | Trading | liquidity |
| retConglomerate | Conglomerate return | Cohen and Lou | 2012, JFE | monthly | Price | delayed processing |
| RDAbility | R&D ability | Cohen, Diether and Malloy | 2013, RFS | yearly | Accounting | other |
| AssetGrowth | Asset growth | Cooper, Gulen and Schill | 2008, JF | yearly | Accounting | investment |
| EarningsForecastDisparity | Long-vs-short EPS forecasts | Da and Warachka | 2011, JFE | monthly | Analyst | earnings |
| CompEquIss | Composite equity issuance | Daniel and Titman | 2006, JF | monthly | Accounting | external financing |
| IntanBM | Intangible return using BM | Daniel and Titman | 2006, JF | yearly | Accounting | long term reversal |
| IntanCFP | Intangible return using CFtoP | Daniel and Titman | 2006, JF | yearly | Accounting | long term reversal |
| IntanEP | Intangible return using EP | Daniel and Titman | 2006, JF | yearly | Accounting | long term reversal |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| IntanSP | Intangible return using Sale2P | Daniel and Titman | 2006, JF | yearly | Accounting | long term reversal |
| ShareIss5Y | Share issuance (5 year) | Daniel and Titman | 2006, JF | monthly | Accounting | external financing |
| LRreversal | Long-run reversal | De Bondt and Thaler | 1985, JF | monthly | Price | long term reversal |
| MRreversal | Medium-run reversal | De Bondt and Thaler | 1985, JF | monthly | Price | long term reversal |
| EquityDuration | Equity Duration | Dechow, Sloan and Soliman | 2004, RAS | yearly | Price | valuation |
| cfp | Operating Cash flows to price | Desai, Rajgopal, Venkatachalam | 2004, AR | yearly | Accounting | valuation |
| ForecastDispersion | EPS Forecast Dispersion | Diether, Malloy and Scherbina | 2002, JF | monthly | Analyst | volatility |
| ExclExp | Excluded Expenses | Doyle, Lundholm and Soliman | 2003, RAS | quarterly | Analyst | composite accounting |
| ProbInformedTrading | Probability of Informed Trading | Easley, Hvidkjaer and O'Hara | 2002, JF | yearly | Trading | liquidity |
| OrgCap | Organizational capital | Eisfeldt and Papanikolaou | 2013, JF | yearly | Accounting | R&D |
| sfe | Earnings Forecast to price | Elgers, Lo and Pfeiffer | 2001, AR | monthly | Analyst | valuation |
| GrLTNOA | Growth in long term operating assets | Fairfield, Whisenant and Yohn | 2003, AR | yearly | Accounting | investment |
| AM | Total assets to market | Fama and French | 1992, JF | yearly | Accounting | valuation |
| BMdec | Book to market using December ME | Fama and French | 1992, JPM | yearly | Accounting | valuation |
| BookLeverage | Book leverage (annual) | Fama and French | 1992, JF | yearly | Accounting | leverage |
| OperProf | operating profits / book equity | Fama and French | 2006, JFE | yearly | Accounting | profitability |
| Beta | CAPM beta | Fama and MacBeth | 1973, JPE | monthly | Price | risk |
| EarningsSurprise | Earnings Surprise | Foster, Olsen and Shevlin | 1984, AR | quarterly | Analyst | earnings |
| AnalystValue | Analyst Value | Frankel and Lee | 1998, JAE | monthly | Analyst | valuation |
| AOP | Analyst Optimism | Frankel and Lee | 1998, JAE | monthly | Analyst | other |
| PredictedFE | Predicted Analyst forecast error | Frankel and Lee | 1998, JAE | monthly | Accounting | earnings |
| FR | Pension Funding Status | Franzoni and Marin | 2006, JF | monthly | Accounting | composite accounting |
| BetaFP | Frazzini-Pedersen Beta | Frazzini and Pedersen | 2014, JFE | monthly | Price | other |
| High52 | 52 week high | George and Hwang | 2004, JF | monthly | Price | momentum |
| IndMom | Industry Momentum | Grinblatt and Moskowitz | 1999, JFE | monthly | Price | momentum |
| PctAcc | Percent Operating Accruals | Hafzalla, Lundholm, Van Winkle | 2011, AR | yearly | Accounting | accruals |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| PctTotAcc | Percent Total Accruals | Hafzalla, Lundholm, Van Winkle | 2011, AR | yearly | Accounting | accruals |
| tang | Tangibility | Hahn and Lee | 2009, JF | yearly | Accounting | asset composition |
| Coskewness | Coskewness | Harvey and Siddique | 2000, JF | monthly | Price | risk |
| RoE | net income / book equity | Haugen and Baker | 1996, JFE | yearly | Accounting | profitability |
| VarCF | Cash-flow to price variance | Haugen and Baker | 1996, JFE | monthly | Accounting | cash flow risk |
| VolMkt | Volume to market equity | Haugen and Baker | 1996, JFE | monthly | Trading | volume |
| VolumeTrend | Volume Trend | Haugen and Baker | 1996, JFE | monthly | Trading | volume |
| AnalystRevision | EPS forecast revision | Hawkins, Chamberlin, Daniel | 1984, FAJ | monthly | Analyst | earnings |
| Mom12mOffSeason | Momentum without the seasonal part | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomOffSeason | Off season long-term reversal | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomOffSeason06YrPlus | Off season reversal years 6 to 10 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomOffSeason11YrPlus | Off season reversal years 11 to 15 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomOffSeason16YrPlus | Off season reversal years 16 to 20 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomSeason | Return seasonality years 2 to 5 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomSeason06YrPlus | Return seasonality years 6 to 10 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomSeason11YrPlus | Return seasonality years 11 to 15 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomSeason16YrPlus | Return seasonality years 16 to 20 | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| MomSeasonShort | Return seasonality last year | Heston and Sadka | 2008, JFE | monthly | Price | momentum |
| NOA | Net Operating Assets | Hirshleifer et al. | 2004, JAE | yearly | Accounting | asset composition |
| dNoa | change in net operating assets | Hirshleifer, Hou, Teoh, Zhang | 2004, JAE | yearly | Accounting | investment |
| EarnSupBig | Earnings surprise of big firms | Hou | 2007, RFS | quarterly | Accounting | delayed processing |
| IndRetBig | Industry return of big firms | Hou | 2007, RFS | monthly | Price | delayed processing |
| PriceDelayRsq | Price delay r square | Hou and Moskowitz | 2005, RFS | monthly | Price | delayed processing |
| PriceDelaySlope | Price delay coeff | Hou and Moskowitz | 2005, RFS | monthly | Price | delayed processing |
| PriceDelayTstat | Price delay SE adjusted | Hou and Moskowitz | 2005, RFS | monthly | Price | delayed processing |
| STreversal | Short term reversal | Jegadeesh | 1989, JF | monthly | Price | short-term reversal |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| RevenueSurprise | Revenue Surprise | Jegadeesh and Livnat | 2006, JFE | quarterly | Accounting | sales growth |
| Mom12m | Momentum (12 month) | Jegadeesh and Titman | 1993, JF | monthly | Price | momentum |
| Mom6m | Momentum (6 month) | Jegadeesh and Titman | 1993, JF | monthly | Price | momentum |
| ChangeInRecommendation | Change in recommendation | Jegadeesh et al. | 2004, JF | monthly | Analyst | recommendation |
| OptionVolume1 | Option to stock volume | Johnson and So | 2012, JFE | monthly | Trading | volume |
| OptionVolume2 | Option volume to average | Johnson and So | 2012, JFE | monthly | Trading | volume |
| BetaTailRisk | Tail risk beta | Kelly and Jiang | 2014, RFS | monthly | Price | risk |
| fgr5yrLag | Long-term EPS forecast | La Porta | 1996, JF | monthly | Analyst | earnings |
| CF | Cash flow to market | Lakonishok, Shleifer, Vishny | 1994, JF | monthly | Accounting | valuation |
| MeanRankRevGrowth | Revenue Growth Rank | Lakonishok, Shleifer, Vishny | 1994, JF | yearly | Accounting | sales growth |
| RDS | Real dirty surplus | Landsman et al. | 2011, AR | yearly | Accounting | composite accounting |
| Tax | Taxable income to income | Lev and Nissim | 2004, AR | yearly | Accounting | tax |
| RDcap | R&D capital-to-assets | Li | 2011, RFS | yearly | Accounting | asset composition |
| zerotrade | Days with zero trades | Liu | 2006, JFE | monthly | Trading | liquidity |
| zerotradeAlt1 | Days with zero trades | Liu | 2006, JFE | monthly | Trading | liquidity |
| zerotradeAlt12 | Days with zero trades | Liu | 2006, JFE | monthly | Trading | liquidity |
| ChEQ | Growth in book equity | Lockwood and Prombutr | 2010, JFR | yearly | Accounting | investment |
| EarningsStreak | Earnings surprise streak | Loh and Warachka | 2012, MS | monthly | Accounting | earnings |
| NumEarnIncrease | Earnings streak length | Loh and Warachka | 2012, MS | quarterly | Accounting | earnings |
| GrAdExp | Growth in advertising expenses | Lou | 2014, RFS | yearly | Accounting | investment |
| EntMult | Enterprise Multiple | Loughran and Wellman | 2011, JFQA | monthly | Accounting | valuation |
| CompositeDebtIssuance | Composite debt issuance | Lyandres, Sun and Zhang | 2008, RFS | yearly | Accounting | external financing |
| InvestPPEInv | change in ppe and inv/assets | Lyandres, Sun and Zhang | 2008, RFS | yearly | Accounting | investment |
| Frontier | Efficient frontier index | Nguyen and Swanson | 2009, JFQA | yearly | Accounting | valuation |
| GP | gross profits / total assets | Novy-Marx | 2013, JFE | yearly | Accounting | profitability |
| IntMom | Intermediate Momentum | Novy-Marx | 2012, JFE | monthly | Price | momentum |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| OPLeverage | Operating leverage | Novy-Marx | 2010, ROF | yearly | Accounting | other |
| Cash | Cash to assets | Palazzo | 2012, JFE | quarterly | Accounting | asset composition |
| BetaLiquidityPS | Pastor-Stambaugh liquidity beta | Pastor and Stambaugh | 2003, JPE | monthly | Price | liquidity |
| BPEBM | Leverage component of BM | Penman, Richardson and Tuna | 2007, JAR | monthly | Accounting | leverage |
| EBM | Enterprise component of BM | Penman, Richardson and Tuna | 2007, JAR | monthly | Accounting | valuation |
| NetDebtPrice | Net debt to price | Penman, Richardson and Tuna | 2007, JAR | monthly | Accounting | leverage |
| PS | Piotroski F-score | Piotroski | 2000, AR | yearly | Accounting | composite accounting |
| ShareIss1Y | Share issuance (1 year) | Pontiff and Woodgate | 2008, JF | monthly | Accounting | external financing |
| DelDRC | Deferred Revenue | Prakash and Sinha | 2012, CAR | yearly | Accounting | investment |
| OrderBacklog | Order backlog | Rajgopal, Shevlin, Venkatachalam | 2003, RAS | yearly | Accounting | sales growth |
| DelCOA | Change in current operating assets | Richardson et al. | 2005, JAE | yearly | Accounting | investment |
| DelCOL | Change in current operating liabilities | Richardson et al. | 2005, JAE | yearly | Accounting | external financing |
| DelEqu | Change in equity to assets | Richardson et al. | 2005, JAE | yearly | Accounting | investment |
| DelFINL | Change in financial liabilities | Richardson et al. | 2005, JAE | yearly | Accounting | external financing |
| DelLTI | Change in long-term investment | Richardson et al. | 2005, JAE | yearly | Accounting | investment |
| DelNetFin | Change in net financial assets | Richardson et al. | 2005, JAE | yearly | Accounting | investment |
| TotalAccruals | Total accruals | Richardson et al. | 2005, JAE | yearly | Accounting | investment |
| BM | Book to market using most recent ME | Rosenberg, Reid, and Lanstein | 1985, JF | monthly | Accounting | valuation |
| Accruals | Accruals | Sloan | 1996, AR | yearly | Accounting | accruals |
| ChAssetTurnover | Change in Asset Turnover | Soliman | 2008, AR | yearly | Accounting | sales growth |
| ChNNCOA | Change in Net Noncurrent Op Assets | Soliman | 2008, AR | yearly | Accounting | investment |
| ChNWC | Change in Net Working Capital | Soliman | 2008, AR | yearly | Accounting | investment |
| ChInv | Inventory Growth | Thomas and Zhang | 2002, RAS | yearly | Accounting | investment |
| ChTax | Change in Taxes | Thomas and Zhang | 2011, JAR | quarterly | Accounting | tax |
| Investment | Investment to revenue | Titman, Wei and Xie | 2004, JFQA | yearly | Accounting | investment |
| realestate | Real estate holdings | Tuzel | 2010, RFS | yearly | Accounting | asset composition |

Continued on next page

**Table A.5: Predictor variables for the machine learning model (continued)**

| Acronym | Description | Author(s) | Year, Journal | Frequency | Cat. data | Cat. economic |
|---|---|---|---|---|---|---|
| AbnormalAccruals | Abnormal Accruals | Xie | 2001, AR | yearly | Accounting | accruals |
| FirmAgeMom | Firm Age - Momentum | Zhang | 2004, JF | monthly | Price | momentum |

The table shows all available characteristics used, the author(s), the year and the journal of publication. In addition, this table shows the update frequency, the data category as well as the economic category.

**Table A.6: DPPP (CRRA) surrogate models**

| | $\gamma = 2$ | $\gamma = 10$ | $\gamma = 10$ | $\gamma = 20$ |
|---|---|---|---|---|
| $R^2$ | 0.5513 | 0.5537 | 0.5561 | 0.6914 |
| Sur. CE | 0.0477 | 0.0342 | 0.0200 | 0.0008 |
| Orig. CE | 0.0669 | 0.0492 | 0.0303 | 0.0097 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | 0.0001 | 0.0001 | 0.0003 | 0.0004 |
| incl. TC | No | No | No | No |
| incl. interactions | No | No | No | No |
| $R^2$ | 0.7606 | 0.7712 | 0.7706 | 0.8472 |
| Sur. CE | 0.0548 | 0.0382 | 0.0202 | 0.0004 |
| Orig. CE | 0.0669 | 0.0492 | 0.0303 | 0.0097 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | 0.0001 | 0.0001 | 0.0001 | 0.0008 |
| incl. TC | No | No | No | No |
| incl. interactions | Yes | Yes | Yes | Yes |
| $R^2$ | 0.6762 | 0.6841 | 0.7415 | 0.8039 |
| Sur. CE | −0.1009 | −0.0841 | −0.0669 | −0.0489 |
| Orig. CE | −0.1218 | −0.0980 | −0.0756 | −0.0536 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | 0.0001 | 0.0001 | 0.0017 | 0.0387 |
| incl. TC | Yes | Yes | Yes | Yes |
| incl. interactions | No | No | No | No |
| $R^2$ | 0.8454 | 0.8395 | 0.8757 | 0.9081 |
| Sur. CE | −0.1154 | −0.0949 | −0.0739 | −0.0516 |
| Orig. CE | −0.1218 | −0.0980 | −0.0756 | −0.0536 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | 0.0001 | 0.0001 | 0.0014 | 0.2514 |
| incl. TC | Yes | Yes | Yes | Yes |
| incl. interactions | Yes | Yes | Yes | Yes |

This table compares the monthly certainty equivalents of the linear surrogate models presented in Section 2.4.3 to the corresponding deep portfolio policies optimized for a CRRA investor with relative risk aversion of 2, 5, 10 and 20, respectively. The deep models are the feed-forward neural networks presented in Section 2.4.1 and Section 2.4.2, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$\gamma = 2$", "$\gamma = 5$", "$\gamma = 10$" and "$\gamma = 20$" correspond to the respective risk aversions. The rows represent the mean adjusted $R^2$ across all periods, the resulting monthly certainty equivalent of the weights predicted by the surrogate model, the monthly certainty equivalent of the corresponding deep model and lastly, the p-value for the difference in the certainty equivalents. The next two rows "incl. TC" and "incl. interactions" stratify the results across the model specification and the inclusion of interactions in the surrogate model.

**Table A.7: (D)PPP for MV investors**

| | $\gamma = 2$ | | $\gamma = 5$ | | $\gamma = 10$ | | $\gamma = 20$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0392 | 0.0662 | 0.0267 | 0.0469 | 0.0140 | 0.0290 | −0.0017 | 0.0053 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0001 | | 0.0002 | | 0.0066 | | 0.1182 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.5361 | 0.6749 | 0.5060 | 0.6057 | 0.4373 | 0.5295 | 0.2939 | 0.3847 |
| $max\ w_i * 100$ | 2.1772 | 1.8125 | 2.0748 | 1.7260 | 1.8184 | 1.6331 | 1.1825 | 1.2971 |
| $min\ w_i * 100$ | −2.3513 | −1.8523 | −2.2097 | −1.8370 | −1.8924 | −1.8039 | −1.2239 | −1.3872 |
| $\sum_i w_i I(w_i < 0)$ | −3.3646 | −4.3656 | −3.1475 | −3.8665 | −2.6527 | −3.3171 | −1.6188 | −2.2737 |
| $\sum_i I(w_i < 0)/N_t$ | 0.4402 | 0.4451 | 0.4334 | 0.4411 | 0.4204 | 0.4344 | 0.3761 | 0.4171 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 3.8594 | 8.5704 | 3.9370 | 7.6984 | 3.5980 | 6.7283 | 2.2396 | 4.8273 |
| Mean | 0.0489 | 0.0786 | 0.0468 | 0.0701 | 0.0430 | 0.0628 | 0.0303 | 0.0482 |
| StdDev | 0.0987 | 0.1115 | 0.0897 | 0.0965 | 0.0764 | 0.0824 | 0.0566 | 0.0656 |
| Skew | −0.1627 | 1.3035 | −0.1451 | 1.0537 | −0.0254 | 0.3598 | −0.0473 | 0.5061 |
| Kurt | 1.5433 | 8.2253 | 1.8391 | 6.5084 | 2.0479 | 0.9416 | 3.0808 | 1.3940 |
| SR | 1.7149 | 2.4408 | 1.8070 | 2.5170 | 1.9518 | 2.6402 | 1.8548 | 2.5443 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.0035 | | 0.0077 | | 0.0014 | | 0.0012 |
| $FF5 + Mom\ \alpha$ | 0.0332 | 0.0626 | 0.0323 | 0.0559 | 0.0299 | 0.0492 | 0.0193 | 0.0368 |
| $StdErr(\alpha)$ | 0.0043 | 0.0058 | 0.0040 | 0.0051 | 0.0035 | 0.0043 | 0.0026 | 0.0033 |

This table shows out-of-sample estimates of the (deep) portfolio policies with 157 firm characteristics optimized for a mean-variance investor with absolute risk aversion of 2, 5, 10 and 20, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$\gamma = 2$", "$\gamma = 5$", "$\gamma = 10$" and "$\gamma = 20$" correspond to the respective risk aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

**Table A.8: (D)PPP for LA investors**

| | $l = 1.5$ | | $l = 2$ | | $l = 2.5$ | | $l = 3$ | |
|---|---|---|---|---|---|---|---|---|
| | PPP | DPPP | PPP | DPPP | PPP | DPPP | PPP | DPPP |
| CE | 0.0406 | 0.0738 | 0.0332 | 0.0631 | 0.0266 | 0.0574 | 0.0194 | 0.0476 |
| p-value$(CE_{DPPP} - CE_{PPP})$ | | 0.0005 | | 0.0001 | | 0.0002 | | 0.0001 |
| $\sum_i \lvert w_i \rvert / N_t * 100$ | 0.5354 | 0.6784 | 0.5069 | 0.6630 | 0.5034 | 0.6468 | 0.4940 | 0.5899 |
| $max\ w_i * 100$ | 2.2067 | 1.8606 | 2.0872 | 1.7858 | 2.0743 | 1.7618 | 2.0345 | 1.6638 |
| $min\ w_i * 100$ | −2.3124 | −1.8713 | −2.1707 | −1.8517 | −2.1577 | −1.7841 | −2.1116 | −1.6680 |
| $\sum_i w_i I(w_i < 0)$ | −3.3600 | −4.3905 | −3.1542 | −4.2795 | −3.1290 | −4.1627 | −3.0616 | −3.7529 |
| $\sum_i I(w_i < 0)/N_t$ | 0.4403 | 0.4524 | 0.4332 | 0.4509 | 0.4307 | 0.4490 | 0.4286 | 0.4467 |
| $\sum_i \lvert w_{i,t} - w^+_{i,t-1} \rvert$ | 3.7083 | 8.7386 | 3.6546 | 8.5511 | 3.7464 | 8.3677 | 3.7305 | 7.6941 |
| Mean | 0.0490 | 0.0824 | 0.0478 | 0.0789 | 0.0473 | 0.0783 | 0.0458 | 0.0721 |
| StdDev | 0.0977 | 0.1575 | 0.0906 | 0.1329 | 0.0871 | 0.1359 | 0.0829 | 0.1108 |
| Skew | 0.0347 | 3.5193 | 0.1242 | 1.8141 | 0.0996 | 3.5153 | 0.1404 | 1.3095 |
| Kurt | 1.0871 | 32.9589 | 0.9407 | 13.0823 | 0.8451 | 33.2542 | 0.7114 | 7.6654 |
| SR | 1.7375 | 1.8130 | 1.8270 | 2.0574 | 1.8789 | 1.9963 | 1.9149 | 2.2548 |
| p-value$(SR_{DPPP} - SR_{PPP})$ | | 0.4763 | | 0.1878 | | 0.4242 | | 0.0916 |
| $FF5 + Mom\ \alpha$ | 0.0336 | 0.0658 | 0.0338 | 0.0633 | 0.0338 | 0.0624 | 0.0327 | 0.0578 |
| $StdErr(\alpha)$ | 0.0043 | 0.0076 | 0.0041 | 0.0065 | 0.0040 | 0.0067 | 0.0039 | 0.0056 |

This table shows out-of-sample estimates of the (deep) portfolio policies optimized for a loss-averse investor with loss aversion of 1.5, 2, 2.5, and 3, using 157 firm characteristics. The regular portfolio policy is a linear model, while the deep model is a feed-forward neural network with three hidden layers and 32, 16, and 8 nodes, respectively. We use data from the Open Source Asset Pricing data set (Chen and Zimmermann, 2022) from January 1971 to December 2020. The columns labeled "$l = 1.5$", "$l = 2$", "$l = 2.5$" and "$l = 3$" correspond to the respective loss aversions. We closely follow Brandt et al. (2009) in terms of the results presented. The first rows show the monthly certainty equivalent of the investor as well as the bootstrapped one-sided p-value for the difference in monthly certainty equivalent between DPPP and PPP. The second set of rows shows statistics on portfolio weights averaged over months $t$. These statistics include the average absolute portfolio weight, the average maximum and minimum portfolio weights, the average sum of negative weights in the portfolio, the average proportion of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows shows the first four moments of the final portfolio return distributions as well as the annualized Sharpe ratios and the bootstrapped one-sided p-value for the difference in Sharpe ratios between DPPP and PPP. The bottom panel shows the alphas and their standard errors with respect to the Fama-French five-factor model (Fama and French, 2015), extended to include the momentum factor (Carhart, 1997). Factors are retrieved from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

# Appendix B

# Appendix to Chapter 3

## B.1 Traditional earnings prediction models

**Table B.1: Traditional earnings models**

Panel A: Traditional model specifications

| Name | Model | Source |
|---|---|---|
| L | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t}$ | Gerakos and Gramacy (2012) |
| HVZ | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 A_{i,t} + \beta_3 D_{i,t} + \beta_4 DD_{i,t} + \beta_5 NegE_{i,t} + \beta_6 ACC_{i,t}$ | Hou et al. (2012) |
| EP | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t}$ | Li and Mohanram (2014) |
| RI | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t} + \beta_4 B_{i,t} + \beta_5 ACC_{i,t}$ | Li and Mohanram (2014) |

Panel B: Traditional model variables

| Variable | Definition |
|---|---|
| $E$ | Income before extraordinary items (ib) / Common shares outstanding (csho) |
| $A$ | Total assets (at) / csho |
| $D$ | Dividends total (dvt) / csho |
| $DD$ | 1 if dvt > 0; 0 else |
| $NegE$ | 1 if ib < 0; 0 else |

Continued on next page

**Table B.1: Traditional earnings models (continued)**

| Variable | Definition |
|---|---|
| *ACC* | (ib - Operating activities - net cash flow (oancf)) / csho |
| *B* | Common/ordinary equity- total (ceq) / csho |

Panel A reports the traditional earnings models estimated. $\mathbb{E}_t[E_{i,t+\tau}]$ denotes the expectation for earnings $E$ of firm $i$ in period $t+\tau$ as of $t$. $\beta_0$-$\beta_5$ are the model coefficients. Panel B reports the variable definitions for the traditional models. Compustat variable names are provided in parentheses. Note the slight changes as opposed to the original papers. More precisely, we scale all variables by common shares outstanding and use a consistent earnings as well as accruals definition.

## B.2 Machine learning earnings prediction models

**Table B.2: Hyperparameters for the machine learning models**

| | | |
|---|---|---|
| | RF, GBT & Dart | |
| | Maximum number of trees | 512 |
| | Learning rate | $\in [0.001, 0.01, 0.1, 1]$ |
| | Maximum depth | $U^{int}(2, 10)$ |
| | Maximum number of leaves | $U^{int}(2, 512)$ |
| | L1-regularization | $U(0, 0.1)$ |
| | L2-regularization | $U(0, 0.1)$ |
| | Feature fraction | $U(0.25, 1)$ |
| | Bagging fraction | $U(0.25, 1)$ |
| | Bagging frequency | $\in (1, 10, 50)$ |
| Dart | Dropout rate | $\in (0.05, 0.1, 0.15)$ |
| Dart | Probability of skipping dropout | $\in (0.25, 0.5)$ |
| | NN | |
| | Learning rate | $\in [0.001, 0.01, 0.1, 1]$ |
| | L1-regularization | $U(0, 0.1)$ |
| | Dropout | $U(0, 0.5)$ |
| | Number of hidden layers | $\in [1, 2, 3, 4, 5]$ |
| | First layer size | $\in [32, 64, 128]$ |
| | Batch size | $\in [2^{11}, 2^{12}, 2^{13}, 2^{14}]$ |

This table gives the hyperparameters that we tune and their respective boundaries. $U$ ($U^{int}$) means drawing from a uniform (integer-wise uniform) distribution. Our choice of hyperparameters and their respective boundaries is based on Bali et al. (2023). We use the *Ray* Python framework to efficiently optimize the hyperparameters (Liaw et al., 2018).

**Table B.3: Predictor variables for the machine learning models**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 1 | aco | Current assets - other - total | Balance sheet | Current assets |
| 2 | acox | Current assets - other - sundry | Balance sheet | Current assets |
| 3 | act | Current assets - total | Balance sheet | Current assets |
| 4 | am | Amortization of intangibles | Income statement | Depreciation and amortization |
| 5 | ao | Assets - other | Balance sheet | Fixed assets |
| 6 | aoloch | Assets and liabilities - other - net change | Cash flow statement | Operating cash flow |
| 7 | aox | Assets - other - sundry | Balance sheet | Fixed assets |
| 8 | ap | Accounts payable - trade | Balance sheet | Liabilities |
| 9 | apalch | Accounts payable and accrued liabilities - increase/(decrease) | Cash flow statement | Operating cash flow |
| 10 | aqc | Acquisitions | Cash flow statement | Investing cash flow |
| 11 | aqi | Acquisitions - income contribution | Income statement | Interest and other |
| 12 | aqs | Acquisitions - sales contribution | Income statement | Sales |
| 13 | at | Assets - total | Balance sheet | Total assets |
| 14 | caps | Capital surplus/share premium reserve | Balance sheet | Equity |
| 15 | capx | Capital expenditures | Cash flow statement | Investing cash flow |
| 16 | capxv | Capital expend property, plant and equipment schd v | Cash flow statement | Investing cash flow |
| 17 | ceq | Common/ordinary equity - total | Balance sheet | Equity |
| 18 | ceql | Common equity - liquidation value | Balance sheet | Supplemental |
| 19 | ceqt | Common equity - tangible | Balance sheet | Supplemental |
| 20 | ch | Cash | Balance sheet | Current assets |
| 21 | che | Cash and short-term investments | Balance sheet | Current assets |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 22 | chech | Cash and cash equivalents - increase/(decrease) | Cash flow statement | Total cash flow |
| 23 | cld2 | Capitalized leases - due in 2nd year | Balance sheet | Supplemental |
| 24 | cld3 | Capitalized leases - due in 3rd year | Balance sheet | Supplemental |
| 25 | cld4 | Capitalized leases - due in 4th year | Balance sheet | Supplemental |
| 26 | cld5 | Capitalized leases - due in 5th year | Balance sheet | Supplemental |
| 27 | cogs | Cost of goods sold | Income statement | Operating expenses |
| 28 | cstk | Common/ordinary stock (capital) | Balance sheet | Equity |
| 29 | cstkcv | Common stock-carrying value | Balance sheet | Supplemental |
| 30 | cstke | Common stock equivalents - dollar savings | Income statement | Interest and other |
| 31 | dc | Deferred charges | Balance sheet | Fixed assets |
| 32 | dclo | Debt - capitalized lease obligations | Balance sheet | Liabilities |
| 33 | dcpstk | Convertible debt and preferred stock | Balance sheet | Supplemental |
| 34 | dcvsr | Debt - senior convertible | Balance sheet | Liabilities |
| 35 | dcvsub | Debt - subordinated convertible | Balance sheet | Liabilities |
| 36 | dcvt | Debt - convertible | Balance sheet | Liabilities |
| 37 | dd | Debt - debentures | Balance sheet | Liabilities |
| 38 | dd1 | Long-term debt due in one year | Balance sheet | Liabilities |
| 39 | dd2 | Debt - due in 2nd year | Balance sheet | Liabilities |
| 40 | dd3 | Debt - due in 3rd year | Balance sheet | Liabilities |
| 41 | dd4 | Debt - due in 4th year | Balance sheet | Liabilities |
| 42 | dd5 | Debt - due in 5th year | Balance sheet | Liabilities |
| 43 | dlc | Debt in current liabilities - total | Balance sheet | Liabilities |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 44 | dltis | Long-term debt - issuance | Cash flow statement | Financing cash flow |
| 45 | dlto | Other long-term debt | Balance sheet | Liabilities |
| 46 | dltp | Long-term debt - tied to prime | Balance sheet | Liabilities |
| 47 | dltr | Long-term debt - reduction | Cash flow statement | Financing cash flow |
| 48 | dltt | Long-term debt - total | Balance sheet | Liabilities |
| 49 | dm | Debt - mortgages and other secured | Balance sheet | Liabilities |
| 50 | dn | Debt - notes | Balance sheet | Liabilities |
| 51 | do | Discontinued operations | Income statement | Interest and other |
| 52 | dp | Depreciation and amortization | Income statement | Depreciation and amortization |
| 53 | dpact | Depreciation, depletion and amortization (accumulated) | Balance sheet | Fixed assets |
| 54 | dpc | Depreciation and amortization (cash flow) | Cash flow statement | Operating cash flow |
| 55 | dpvieb | Depreciation (accumulated) - ending balance (schedule vi) | Balance sheet | Supplemental |
| 56 | ds | Debt-subordinated | Balance sheet | Liabilities |
| 57 | dudd | Debt - unamortized debt discount and other | Balance sheet | Liabilities |
| 58 | dv | Cash dividends (cash flow) | Cash flow statement | Financing cash flow |
| 59 | dvc | Dividends common/ordinary | Income statement | Dividends |
| 60 | dvp | Dividends - preferred/preference | Income statement | Dividends |
| 61 | dvpa | Preferred dividends in arrears | Balance sheet | Supplemental |
| 62 | dvt | Dividends - total | Income statement | Dividends |
| 63 | dxd2 | Debt (excl capitalized leases) - due in 2nd year | Balance sheet | Supplemental |
| 64 | dxd3 | Debt (excl capitalized leases) - due in 3rd year | Balance sheet | Supplemental |
| 65 | dxd4 | Debt (excl capitalized leases) - due in 4th year | Balance sheet | Supplemental |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 66 | dxd5 | Debt (excl capitalized leases) - due in 5th year | Balance sheet | Supplemental |
| 67 | ebit | Earnings before interest and taxes | Income statement | EBIT |
| 68 | ebitda | Earnings before interest | Income statement | EBITDA |
| 69 | esub | Equity in earnings - unconsolidated subsidiaries | Income statement | Interest and other |
| 70 | esubc | Equity in net loss - earnings | Cash flow statement | Operating cash flow |
| 71 | exre | Exchange rate effect | Cash flow statement | Total cash flow |
| 72 | fatb | Property, plant, and equipment - buildings at cost | Balance sheet | Supplemental |
| 73 | fatc | Property, plant, and equipment - construction in progress at cost | Balance sheet | Supplemental |
| 74 | fate | Property, plant, and equipment - machinery and equipment at cost | Balance sheet | Supplemental |
| 75 | fatl | Property, plant, and equipment - leases at cost | Balance sheet | Supplemental |
| 76 | fatn | Property, plant, and equipment - natural resources at cost | Balance sheet | Supplemental |
| 77 | fato | Property, plant, and equipment - other at cost | Balance sheet | Supplemental |
| 78 | fatp | Property, plant, and equipment - land and improvements at cost | Balance sheet | Supplemental |
| 79 | fiao | Financing activities - other | Cash flow statement | Financing cash flow |
| 80 | fincf | Financing activities - net cash flow | Cash flow statement | Financing cash flow |
| 81 | fopo | Funds from operations - other | Cash flow statement | Operating cash flow |
| 82 | gp | Gross profit | Income statement | Operating expenses |
| 83 | ib | Income before extraordinary items | Income statement | Net income |
| 84 | ibadj | Income before extraordinary items - adjusted for common stock equivalents | Income statement | Net income |
| 85 | ibc | Income before extraordinary items (cash flow) | Cash flow statement | Operating cash flow |
| 86 | ibcom | Income before extraordinary items - available for common | Income statement | Net income |
| 87 | icapt | Invested capital - total | Balance sheet | Supplemental |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 88 | idit | Interest and related income - total | Income statement | Interest and other |
| 89 | intan | Intangible assets - total | Balance sheet | Fixed assets |
| 90 | intc | Interest capitalized | Income statement | Interest and other |
| 91 | intpn | Interest paid - net | Cash flow statement | Operating cash flow |
| 92 | invch | Inventory - decrease (increase) | Cash flow statement | Operating cash flow |
| 93 | invfg | Inventories - finished goods | Balance sheet | Current assets |
| 94 | invo | Inventories - other | Balance sheet | Current assets |
| 95 | invrm | Inventories - raw materials | Balance sheet | Current assets |
| 96 | invt | Inventories - total | Balance sheet | Current assets |
| 97 | invwip | Inventories - work in process | Balance sheet | Current assets |
| 98 | itcb | Investment tax credit (balance sheet) | Balance sheet | Liabilities |
| 99 | itci | Investment tax credit (income account) | Income statement | Taxes |
| 100 | ivaco | Investing activities - other | Cash flow statement | Investing cash flow |
| 101 | ivaeq | Investment and advances - equity | Balance sheet | Fixed assets |
| 102 | ivao | Investment and advances - other | Balance sheet | Fixed assets |
| 103 | ivch | Increase in investments | Cash flow statement | Investing cash flow |
| 104 | ivncf | Investing activities - net cash flow | Cash flow statement | Investing cash flow |
| 105 | ivst | Short-term investments - total | Balance sheet | Current assets |
| 106 | ivstch | Short-term investments - change | Cash flow statement | Investing cash flow |
| 107 | lco | Current liabilities - other - total | Balance sheet | Liabilities |
| 108 | lcox | Current liabilities - other - sundry | Balance sheet | Liabilities |
| 109 | lct | Current liabilities - total | Balance sheet | Liabilities |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 110 | lifr | Lifo reserve | Balance sheet | Supplemental |
| 111 | lo | Liabilities - other - total | Balance sheet | Liabilities |
| 112 | lse | Liabilities and stockholders equity - total | Balance sheet | Total liabilities and equity |
| 113 | lt | Liabilities - total | Balance sheet | Liabilities |
| 114 | mib | Noncontrolling interest (balance sheet) | Balance sheet | Liabilities |
| 115 | mii | Noncontrolling interest (income account) | Income statement | Interest and other |
| 116 | mrc1 | Rental commitments - minimum - 1st year | Balance sheet | Supplemental |
| 117 | mrc2 | Rental commitments - minimum - 2nd year | Balance sheet | Supplemental |
| 118 | mrc3 | Rental commitments - minimum - 3rd year | Balance sheet | Supplemental |
| 119 | mrc4 | Rental commitments - minimum - 4th year | Balance sheet | Supplemental |
| 120 | mrc5 | Rental commitments - minimum - 5th year | Balance sheet | Supplemental |
| 121 | mrct | Rental commitments - minimum - 5 year total | Balance sheet | Supplemental |
| 122 | msa | Marketable securities adjustment | Balance sheet | Supplemental |
| 123 | ni | Net income (loss) | Income statement | Net income |
| 124 | niadj | Net income adjusted for common/ordinary stock (capital) equivalents | Income statement | Net income |
| 125 | nopi | Nonoperating income (expense) | Income statement | Interest and other |
| 126 | nopio | Nonoperating income (expense) - other | Income statement | Interest and other |
| 127 | np | Notes payable - short-term borrowings | Balance sheet | Liabilities |
| 128 | oancf | Operating activities - net cash flow | Cash flow statement | Operating cash flow |
| 129 | oiadp | Operating income after depreciation | Income statement | EBIT |
| 130 | oibdp | Operating income before depreciation | Income statement | EBITDA |
| 131 | pi | Pretax income | Income statement | EBT |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 132 | ppegt | Property, plant and equipment - total (gross) | Balance sheet | Fixed assets |
| 133 | ppent | Property, plant and equipment - total (net) | Balance sheet | Fixed assets |
| 134 | ppeveb | Property, plant, and equipment - ending balance (schedule v) | Balance sheet | Supplemental |
| 135 | prstkc | Purchase of common and preferred stock | Cash flow statement | Financing cash flow |
| 136 | pstk | Preferred/preference stock (capital) - total | Balance sheet | Equity |
| 137 | pstkc | Preferred stock - convertible | Balance sheet | Equity |
| 138 | pstkl | Preferred stock - liquidating value | Balance sheet | Supplemental |
| 139 | pstkn | Preferred/preference stock - nonredeemable | Balance sheet | Equity |
| 140 | pstkr | Preferred/preference stock - redeemable | Balance sheet | Equity |
| 141 | pstkrv | Preferred stock - redemption value | Balance sheet | Supplemental |
| 142 | re | Retained earnings | Balance sheet | Equity |
| 143 | rea | Retained earnings - restatement | Balance sheet | Supplemental |
| 144 | reajo | Retained earnings - other adjustments | Balance sheet | Supplemental |
| 145 | recch | Accounts receivable - decrease (increase) | Cash flow statement | Operating cash flow |
| 146 | recco | Receivables - current - other | Balance sheet | Current assets |
| 147 | recd | Receivables - estimated doubtful | Balance sheet | Current assets |
| 148 | rect | Receivables - tota | Balance sheet | Current assets |
| 149 | recta | Retained earnings - cumulative translation adjustment | Balance sheet | Supplemental |
| 150 | rectr | Receivables - trade | Balance sheet | Current assets |
| 151 | reuna | Retained earnings - unadjusted | Balance sheet | Equity |
| 152 | revt | Revenue - total | Income statement | Sales |
| 153 | sale | Sales/turnover (net) | Income statement | Sales |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 154 | seq | Stockholders equity - parent | Balance sheet | Equity |
| 155 | siv | Sale of investments | Cash flow statement | Investing cash flow |
| 156 | spi | Special items | Income statement | Interest and other |
| 157 | sppe | Sale of property | Cash flow statement | Operating cash flow |
| 158 | sppiv | Sale of property, plant and equipment and investments - gain (loss) | Cash flow statement | Operating cash flow |
| 159 | sstk | Sale of common and preferred stock | Cash flow statement | Financing cash flow |
| 160 | tlcf | Tax loss carry forward | Balance sheet | Supplemental |
| 161 | tstk | Treasury stock - total (all capital) | Balance sheet | Equity |
| 162 | tstkc | Treasury stock - common | Balance sheet | Equity |
| 163 | tstkp | Treasury stock - preferrred | Balance sheet | Equity |
| 164 | txach | Income taxes - accrued - increase/(decrease) | Cash flow statement | Operating cash flow |
| 165 | txc | Income taxes - current | Income statement | Taxes |
| 166 | txdb | Deferred taxes (balance sheet) | Balance sheet | Liabilities |
| 167 | txdc | Deferred taxes (cash flow) | Income statement | Operating cash flow |
| 168 | txdfed | Deferred taxes-federal | Income statement | Taxes |
| 169 | txdfo | Deferred taxes-foreign | Income statement | Taxes |
| 170 | txdi | Income taxes - deferred | Income statement | Taxes |
| 171 | txditc | Deferred taxes and investment tax credit | Balance sheet | Liabilities |
| 172 | txds | Deferred taxes-state | Income statement | Taxes |
| 173 | txfed | Income taxes - federal | Income statement | Taxes |
| 174 | txfo | Income taxes - foreign | Income statement | Taxes |
| 175 | txo | Income taxes - other | Income statement | Taxes |

Continued on next page

**Table B.3: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 176 | txp | Income taxes payable | Balance sheet | Liabilities |
| 177 | txpd | Income taxes paid | Cash flow statement | Operating cash flow |
| 178 | txr | Income tax refund | Balance sheet | Current assets |
| 179 | txs | Income taxes - state | Income statement | Taxes |
| 180 | txt | Income taxes - total | Income statement | Taxes |
| 181 | txw | Excise taxes | Income statement | Taxes |
| 182 | wcap | Working capital (balance sheet) | Balance sheet | Supplemental |
| 183 | xacc | Accrued expenses | Balance sheet | Liabilities |
| 184 | xi | Extraordinary items | Income statement | Interest and other |
| 185 | xido | Extraordinary items and discontinued operations | Income statement | Interest and other |
| 186 | xidoc | Extraordinary items and discontinued operations (cash flow) | Cash flow statement | Operating cash flow |
| 187 | xint | Interest and related expense - total | Income statement | Interest and other |
| 188 | xopr | Operating expenses - total | Income statement | Operating expenses |
| 189 | xpp | Prepaid expenses | Balance sheet | Current assets |
| 190 | xpr | Pension and retirement expense | Income statement | Operating expenses |
| 191 | xrent | Rental expense | Income statement | Operating expenses |
| 192 | xsga | Selling, general and administrative expense | Income statement | Operating expenses |

This table reports the input variables used in our machine learning models. We also report the Compustat description, the financial statement group as well as the financial statement component group we assign to the respective variable. EBITDA denotes earnings before interest, taxes, depreciation and amortization. EBIT denotes earnings before interest and taxes. EBT denotes earnings before taxes. We scale all variables by common shares outstanding.

## B.3 Implied cost of capital models

**Table B.4: Implied cost of capital models**

| Name | Model/Description |
|---|---|
| GLS | $P_t = B_t + \Sigma_{\tau=1}^{11} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{GLS}) \cdot B_{t+\tau-1}]}{(1+ICC_{GLS})^\tau} + \frac{\mathbb{E}_t[ROE_{t+12} - ICC_{GLS}) \cdot B_{t+11}]}{(1+ICC_{GLS})^{11} \cdot ICC_{GLS}}$<br><br>This model is given by Gebhardt et al. (2001). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{GLS}$ denotes the implied cost of capital (ICC), $B_{t+\tau-1}$ denotes the book value of equity per share in $t+\tau-1$ and $ROE_{t+\tau}$ is the (forecasted) return on equity in $t+\tau$. $B_{t+\tau-1}$ is calculated using the clean surplus relation following Hou et al. (2012). $ROE_{t+\tau}$ is calculated by dividing earnings per share forecasts for $t+\tau$ by $B_{t+\tau-1}$. For $ROE_{t+\tau}$ up to $\tau = 3$ we use the respective models earnings per share forecast. Afterwards, we assume $ROE_{t+\tau}$ to revert to the historical industry median by $\tau = 11$ (e.g., Hou et al., 2012). The industry median of ROE is derived using 10 years of data while excluding loss firms (e.g., Gebhardt et al., 2001). We expect $ROE_{t+\tau}$ to be constant after $\tau = 11$. |
| CT | $P_t = B_t + \sum_{\tau=1}^{5} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{CT}) \cdot B_{t+\tau-1}]}{(1+ICC_{CT})^\tau} + \frac{\mathbb{E}_t[(ROE_{t+5} - ICC_{CT}) \cdot B_{t+4}] \cdot (1+g)}{(1+ICC_{CT})^5 \cdot (ICC_{CT} - g)}$<br><br>This model is given by Claus and Thomas (2001). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{CT}$ denotes the implied cost of capital (ICC), $B_{t+\tau-1}$ denotes the book value of equity per share in $t+\tau-1$, $ROE_{t+\tau}$ is the (forecasted) return on equity in $t+\tau$ and $g$ is the perpetuity growth rate. $B_{t+\tau-1}$ is calculated using the clean surplus relation following Hou et al. (2012). $ROE_{t+\tau}$ is calculated by dividing earnings per share forecasts for $t+\tau$ by $B_{t+\tau-1}$. $g$ is calculated as the current risk-free rate minus 3% (e.g., Hou et al., 2012). |
| OJ | $P_t = \frac{\mathbb{E}_t[E_{t+1}] \cdot (g_{st} - (\gamma-1))}{(R-A)^- A^2}$, with<br>$A = 0.5((\gamma-1)\frac{\mathbb{E}_t[E_{t+1}] \cdot payout}{P_t})$, $g_{st} = 0.5(\frac{\mathbb{E}_t[E_{t+3}] - \mathbb{E}_t[E_{t+2}]}{\mathbb{E}_t[E_{t+2}]} - \frac{\mathbb{E}_t[E_{t+5}] - \mathbb{E}_t[E_{t+4}]}{\mathbb{E}_t[E_{t+4}]})$<br><br>This model is given by Ohlson and Juettner-Nauroth (2005). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{OJ}$ denotes the implied cost of capital (ICC), $E_{t+\tau}$ denotes (forecasted) earnings in $t+\tau$, $g_{st}$ is the short-term growth rate, $\gamma$ is the perpetual growth rate and *payout* is the current payout ratio. $g_{st}$ is calculated as the mean of forecasted earnings growth in $\tau = 3$ and $\tau = 5$ (e.g., Hou et al., 2012). $\gamma$ is the current risk-free rate minus 3% (e.g., Hou et al., 2012). *payout* is calculated as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al., 2012). |
| MPEG | $P_t = \frac{\mathbb{E}_t[E_{t+2}] + (ICC_{MPEG} \cdot payout - 1) \cdot \mathbb{E}_t[E_{t+1}]}{ICC_{MPEG}^2}$<br><br>This model is given by Easton (2004). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{MPEG}$ denotes the implied cost of capital (ICC) and $E_{t+\tau}$ denotes the (forecasted) earnings per share in $t+\tau$. *payout* is derived as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al., 2012). |
| GG | $P_t = \frac{\mathbb{E}_t[E_{t+1}]}{ICC_{GG}}$<br><br>This model is given by Gordon and Gordon (1997). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{GG}$ denotes the implied cost of capital (ICC) and $E_{t+1}$ denotes the (forecasted) earnings per share in $t+1$. |

This table shows implied cost of capital (ICC) models that we base our composite ICC on. The presentation of the models closely follows the one provided by Hess and Wolf (2022). For simplicity, we drop the firm index $i$. The composite ICC that we use is derived as the average of these ICCs.

# Appendix C

# Appendix to Chapter 4

## C.1 Implied cost of capital models

**Table C.1: Implied cost of capital models**

| Name | Model/Description |
|---|---|
| GLS | $P_t = B_t + \Sigma_{\tau=1}^{11} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{GLS}) \cdot B_{t+\tau-1}]}{(1+ICC_{GLS})^{\tau}} + \frac{\mathbb{E}_t[ROE_{t+12} - ICC_{GLS}) \cdot B_{t+11}]}{(1+ICC_{GLS})^{11} \cdot ICC_{GLS}}$<br><br>This model is given by Gebhardt et al. (2001). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{GLS}$ denotes the implied cost of capital (ICC), $B_{t+\tau-1}$ denotes the book value of equity per share in $t+\tau-1$ and $ROE_{t+\tau}$ is the (forecasted) return on equity in $t+\tau$. $B_{t+\tau-1}$ is calculated using the clean surplus relation following Hou et al. (2012). $ROE_{t+\tau}$ is calculated by dividing earnings per share forecasts for $t+\tau$ by $B_{t+\tau-1}$. For $ROE_{t+\tau}$ up to $\tau = 3$ I use the respective models earnings per share forecast. Afterwards, I assume $ROE_{t+\tau}$ to revert to the historical industry median by $\tau = 11$ (e.g., Hou et al., 2012). The industry median of ROE is derived using 10 years of data while excluding loss firms (e.g., Gebhardt et al., 2001). I expect $ROE_{t+\tau}$ to be constant after $\tau = 11$. |
| CT | $P_t = B_t + \sum_{\tau=1}^{5} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{CT}) \cdot B_{t+\tau-1}]}{(1+ICC_{CT})^{\tau}} + \frac{\mathbb{E}_t[(ROE_{t+5} - ICC_{CT}) \cdot B_{t+4}] \cdot (1+g)}{(1+ICC_{CT})^{5} \cdot (ICC_{CT} - g)}$<br><br>This model is given by Claus and Thomas (2001). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{CT}$ denotes the implied cost of capital (ICC), $B_{t+\tau-1}$ denotes the book value of equity per share in $t+\tau-1$, $ROE_{t+\tau}$ is the (forecasted) return on equity in $t+\tau$ and $g$ is the perpetuity growth rate. $B_{t+\tau-1}$ is calculated using the clean surplus relation following Hou et al. (2012). $ROE_{t+\tau}$ is calculated by dividing earnings per share forecasts for $t+\tau$ by $B_{t+\tau-1}$. $g$ is calculated as the current risk-free rate minus 3% (e.g., Hou et al., 2012). |
| OJ | $P_t = \frac{\mathbb{E}_t[E_{t+1}] \cdot (g_{st} - (\gamma - 1))}{(R-A)^{-} A^2}$, with<br>$A = 0.5((\gamma - 1)\frac{\mathbb{E}_t[E_{t+1}] \cdot payout}{P_t})$, $g_{st} = 0.5(\frac{\mathbb{E}_t[E_{t+3}] - \mathbb{E}_t[E_{t+2}]}{\mathbb{E}_t[E_{t+2}]} - \frac{\mathbb{E}_t[E_{t+5}] - \mathbb{E}_t[E_{t+4}]}{\mathbb{E}_t[E_{t+4}]})$<br><br>This model is given by Ohlson and Juettner-Nauroth (2005). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{OJ}$ denotes the implied cost of capital (ICC), $E_{t+\tau}$ denotes (forecasted) earnings in $t+\tau$, $g_{st}$ is the short-term growth rate, $\gamma$ is the perpetual growth rate and *payout* is the current payout ratio. $g_{st}$ is calculated as the mean of forecasted earnings growth in $\tau = 3$ and $\tau = 5$ (e.g., Hou et al., 2012). $\gamma$ is the current risk-free rate minus 3% (e.g., Hou et al., 2012). *payout* is calculated as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al., 2012). |
| MPEG | $P_t = \frac{\mathbb{E}_t[E_{t+2}] + (ICC_{MPEG} \cdot payout - 1) \cdot \mathbb{E}_t[E_{t+1}]}{ICC_{MPEG}^2}$<br><br>This model is given by Easton (2004). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{MPEG}$ denotes the implied cost of capital (ICC) and $E_{t+\tau}$ denotes the (forecasted) earnings per share in $t+\tau$. *payout* is derived as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al., 2012). |
| GG | $P_t = \frac{\mathbb{E}_t[E_{t+1}]}{ICC_{GG}}$<br><br>This model is given by Gordon and Gordon (1997). $P_t$ denotes the stock price as of the estimation date in $t$, $ICC_{GG}$ denotes the implied cost of capital (ICC) and $E_{t+1}$ denotes the (forecasted) earnings per share in $t+1$. |

This table reports the ICC models which I employ to derive the composite ICC estimate. The presentation of the models closely follows the one provided by Hess and Wolf (2022). For simplicity, I drop the firm index $i$. The composite ICC that I use is derived as the average of these ICC.

## C.2 Traditional earnings prediction models

**Table C.2: Traditional earnings models**

Panel A: Traditional model specifications

| Name | Model | Source |
|---|---|---|
| HVZ | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 A_{i,t} + \beta_3 D_{i,t} + \beta_4 DD_{i,t} + \beta_5 NegE_{i,t} + \beta_6 ACC_{i,t}$ | Hou et al. (2012) |
| EP | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t}$ | Li and Mohanram (2014) |
| RI | $\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t} + \beta_4 B_{i,t} + \beta_5 ACC_{i,t}$ | Li and Mohanram (2014) |

Panel B: Variable definitions

| Variable | Definition |
|---|---|
| $E$ | Income before extraordinary items (ib) / common shares outstanding (csho) |
| $A$ | Total assets (at) / csho |
| $D$ | Dividends total (dvt) / csho |
| $DD$ | 1 if dvt > 0; 0 else |
| $NegE$ | 1 if ib < 0; 0 else |
| $ACC$ | Δ(Current assets (act) - cash and cash equivalents (che)) - Δ(current liabilities (lct) - debt in current liabilities - total (dlc) - income taxes payable (txp)) - depreciation and amortization (dp) |
| $B$ | Common/ordinary equity - total (ceq) / csho |

Panel A reports the traditional earnings models estimated. $\mathbb{E}_t[E_{i,t+\tau}]$ denotes the expectation for earnings $E$ of firm $i$ in period $t+\tau$ as of $t$. $\beta_0$-$\beta_5$ are the model coefficients. Panel B reports the variable definitions for the traditional models. Compustat variable names are provided in parentheses. Note the slight changes as opposed to the original papers. More precisely, I scale all variables by common shares outstanding and use a consistent earnings as well as accruals definition. All models are estimated by minimizing the mean squared error.

## C.3 Machine learning earnings prediction model

**Table C.3: Hyperparameters for the machine learning model**

| | RF, GBT & DART | |
|---|---|---|
| | Loss function | Mean absolute error |
| | Maximum number of trees | 512 |
| | Learning rate | $\in [0.001, 0.01, 0.1, 1]$ |
| | Maximum depth | $U^{int}(2, 10)$ |
| | Maximum number of leaves | $U^{int}(2, 512)$ |
| | L1-regularization | $U(0, 0.1)$ |
| | L2-regularization | $U(0, 0.1)$ |
| | Feature fraction | $U(0.25, 1)$ |
| | Bagging fraction | $U(0.25, 1)$ |
| | Bagging frequency | $\in (1, 10, 50)$ |
| DART | Dropout rate | $\in (0.05, 0.1, 0.15)$ |
| DART | Probability of skipping dropout | $\in (0.25, 0.5)$ |

This table reports the hyperparameters that I tune and their respective boundaries. $U$ ($U^{int}$) means drawing from a uniform (integer-wise uniform) distribution. The choice of hyperparameters and their respective boundaries is based on Bali et al. (2023). I use the *Ray* Python framework to efficiently optimize the hyperparameters (Liaw et al., 2018).

**Table C.4: Predictor variables for the machine learning model**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 1 | aco | Current assets - other - total | Balance sheet | Current assets |
| 2 | acox | Current assets - other - sundry | Balance sheet | Current assets |
| 3 | act | Current assets - total | Balance sheet | Current assets |
| 4 | ao | Assets - other | Balance sheet | Fixed assets |
| 5 | aox | Assets - other - sundry | Balance sheet | Fixed assets |
| 6 | ap | Accounts payable - trade | Balance sheet | Liabilities |
| 7 | at | Assets - total | Balance sheet | Total assets |
| 8 | caps | Capital surplus/share premium reserve | Balance sheet | Equity |
| 9 | capx | Capital expenditures | Cash flow statement | Investing cash flow |
| 10 | capxv | Capital expend property, plant and equipment schd v | Cash flow statement | Investing cash flow |
| 11 | ceq | Common/ordinary equity - total | Balance sheet | Equity |
| 12 | ceql | Common equity - liquidation value | Balance sheet | Supplemental |
| 13 | ceqt | Common equity - tangible | Balance sheet | Supplemental |
| 14 | ch | Cash | Balance sheet | Current assets |
| 15 | che | Cash and short-term investments | Balance sheet | Current assets |
| 16 | cogs | Cost of goods sold | Income statement | Operating expenses |
| 17 | cstk | Common/ordinary stock (capital) | Balance sheet | Equity |
| 18 | cstke | Common stock equivalents - dollar savings | Income statement | Interest and other |
| 19 | dc | Deferred charges | Balance sheet | Fixed assets |
| 20 | dclo | Debt - capitalized lease obligations | Balance sheet | Liabilities |
| 21 | dcpstk | Convertible debt and preferred stock | Balance sheet | Supplemental |

Continued on next page

**Table C.4: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 22 | dcvt | Debt - convertible | Balance sheet | Liabilities |
| 23 | dd | Debt - debentures | Balance sheet | Liabilities |
| 24 | dd1 | Long-term debt due in one year | Balance sheet | Liabilities |
| 25 | dlc | Debt in current liabilities - total | Balance sheet | Liabilities |
| 26 | dlto | Other long-term debt | Balance sheet | Liabilities |
| 27 | dltt | Long-term debt - total | Balance sheet | Liabilities |
| 28 | dn | Debt - notes | Balance sheet | Liabilities |
| 29 | do | Discontinued operations | Income statement | Interest and other |
| 30 | dp | Depreciation and amortization | Income statement | Depreciation and amortization |
| 31 | dpact | Depreciation, depletion and amortization (accumulated) | Balance sheet | Fixed assets |
| 32 | ds | Debt-subordinated | Balance sheet | Liabilities |
| 33 | dudd | Debt - unamortized debt discount and other | Balance sheet | Liabilities |
| 34 | dvc | Dividends common/ordinary | Income statement | Dividends |
| 35 | dvp | Dividends - preferred/preference | Income statement | Dividends |
| 36 | dvt | Dividends - total | Income statement | Dividends |
| 37 | ebit | Earnings before interest and taxes | Income statement | EBIT |
| 38 | ebitda | Earnings before interest | Income statement | EBITDA |
| 39 | esub | Equity in earnings - unconsolidated subsidiaries | Income statement | Interest and other |
| 40 | gp | Gross profit | Income statement | Operating expenses |
| 41 | ib | Income before extraordinary items | Income statement | Net income |
| 42 | ibadj | Income before extraordinary items - adjusted for common stock equivalents | Income statement | Net income |
| 43 | ibcom | Income before extraordinary items - available for common | Income statement | Net income |

Continued on next page

**Table C.4: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 44 | icapt | Invested capital - total | Balance sheet | Supplemental |
| 45 | idit | Interest and related income - total | Income statement | Interest and other |
| 46 | intan | Intangible assets - total | Balance sheet | Fixed assets |
| 47 | intc | Interest capitalized | Income statement | Interest and other |
| 48 | invt | Inventories - total | Balance sheet | Current assets |
| 49 | itcb | Investment tax credit (balance sheet) | Balance sheet | Liabilities |
| 50 | itci | Investment tax credit (income account) | Income statement | Taxes |
| 51 | ivaeq | Investment and advances - equity | Balance sheet | Fixed assets |
| 52 | ivao | Investment and advances - other | Balance sheet | Fixed assets |
| 53 | ivst | Short-term investments - total | Balance sheet | Current assets |
| 54 | lco | Current liabilities - other - total | Balance sheet | Liabilities |
| 55 | lcox | Current liabilities - other - sundry | Balance sheet | Liabilities |
| 56 | lct | Current liabilities - total | Balance sheet | Liabilities |
| 57 | lo | Liabilities - other - total | Balance sheet | Liabilities |
| 58 | lse | Liabilities and stockholders equity - total | Balance sheet | Total liabilities and equity |
| 59 | lt | Liabilities - total | Balance sheet | Liabilities |
| 60 | mib | Noncontrolling interest (balance sheet) | Balance sheet | Liabilities |
| 61 | mii | Noncontrolling interest (income account) | Income statement | Interest and other |
| 62 | ni | Net income (loss) | Income statement | Net income |
| 63 | niadj | Net income adjusted for common/ordinary stock (capital) equivalents | Income statement | Net income |
| 64 | nopi | Nonoperating income (expense) | Income statement | Interest and other |
| 65 | nopio | Nonoperating income (expense) - other | Income statement | Interest and other |

Continued on next page

**Table C.4: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 66 | np | Notes payable - short-term borrowings | Balance sheet | Liabilities |
| 67 | oiadp | Operating income after depreciation | Income statement | EBIT |
| 68 | oibdp | Operating income before depreciation | Income statement | EBITDA |
| 69 | pi | Pretax income | Income statement | EBT |
| 70 | ppegt | Property, plant and equipment - total (gross) | Balance sheet | Fixed assets |
| 71 | ppent | Property, plant and equipment - total (net) | Balance sheet | Fixed assets |
| 72 | ppeveb | Property, plant, and equipment - ending balance (schedule v) | Balance sheet | Supplemental |
| 73 | pstk | Preferred/preference stock (capital) - total | Balance sheet | Equity |
| 74 | pstkc | Preferred stock - convertible | Balance sheet | Equity |
| 75 | pstkl | Preferred stock - liquidating value | Balance sheet | Supplemental |
| 76 | pstkn | Preferred/preference stock - nonredeemable | Balance sheet | Equity |
| 77 | pstkrv | Preferred stock - redemption value | Balance sheet | Supplemental |
| 78 | re | Retained earnings | Balance sheet | Equity |
| 79 | recco | Receivables - current - other | Balance sheet | Current assets |
| 80 | rect | Receivables - tota | Balance sheet | Current assets |
| 81 | rectr | Receivables - trade | Balance sheet | Current assets |
| 82 | revt | Revenue - total | Income statement | Sales |
| 83 | sale | Sales/turnover (net) | Income statement | Sales |
| 84 | seq | Stockholders equity - parent | Balance sheet | Equity |
| 85 | spi | Special items | Income statement | Interest and other |
| 86 | tlcf | Tax loss carry forward | Balance sheet | Supplemental |
| 87 | tstkp | Treasury stock - preferrred | Balance sheet | Equity |

Continued on next page

**Table C.4: Predictor variables for the machine learning models (continued)**

| | Variable | Compustat description | Financial statement | Component |
|---|---|---|---|---|
| 88 | txc | Income taxes - current | Income statement | Taxes |
| 89 | txdb | Deferred taxes (balance sheet) | Balance sheet | Liabilities |
| 90 | txdi | Income taxes - deferred | Income statement | Taxes |
| 91 | txditc | Deferred taxes and investment tax credit | Balance sheet | Liabilities |
| 92 | txfed | Income taxes - federal | Income statement | Taxes |
| 93 | txp | Income taxes payable | Balance sheet | Liabilities |
| 94 | txr | Income tax refund | Balance sheet | Current assets |
| 95 | txt | Income taxes - total | Income statement | Taxes |
| 96 | wcap | Working capital (balance sheet) | Balance sheet | Supplemental |
| 97 | xacc | Accrued expenses | Balance sheet | Liabilities |
| 98 | xi | Extraordinary items | Income statement | Interest and other |
| 99 | xido | Extraordinary items and discontinued operations | Income statement | Interest and other |
| 100 | xint | Interest and related expense - total | Income statement | Interest and other |
| 101 | xopr | Operating expenses - total | Income statement | Operating expenses |
| 102 | xpp | Prepaid expenses | Balance sheet | Current assets |
| 103 | xpr | Pension and retirement expense | Income statement | Operating expenses |
| 104 | xrent | Rental expense | Income statement | Operating expenses |
| 105 | xsga | Selling, general and administrative expense | Income statement | Operating expenses |

This table reports the input variables used for the machine learning model. I also report the Compustat description, the financial statement group as well as the financial statement component group we assign to the respective variable. EBITDA denotes earnings before interest, taxes, depreciation and amortization. EBIT denotes earnings before interest and taxes. EBT denotes earnings before taxes. I scale all variables by common shares outstanding. In addition to these variables, the machine learning model includes the traditional variables which are not already included in this list. The traditional variables are defined in Table C.2.

## C.4 General accuracy versus systematic distortions

**Table C.5: General accuracy versus systematic distortions: simulation I**

| | | HVZ | HVZ$^{ud}$ | EP | EP$^{ud}$ | RI | RI$^{ud}$ | ML | ML$^{ud}$ | PERF |
|---|---|---|---|---|---|---|---|---|---|---|
| Deciles | 10-1 | 0.1569*** | 0.3174*** | 0.1794*** | 0.3241*** | 0.1532*** | 0.3249*** | 0.1318*** | 0.3356*** | 0.3929*** |
| | 10-1 net | 0.1415*** | 0.3007*** | 0.1636*** | 0.3075*** | 0.1375*** | 0.3083*** | 0.1157*** | 0.3190*** | 0.3767*** |
| Quintiles | 5-1 | 0.1112*** | 0.2625*** | 0.1294*** | 0.2674*** | 0.1084*** | 0.2684*** | 0.0973*** | 0.2747*** | 0.3395*** |
| | 5-1 net | 0.0971*** | 0.2474*** | 0.1149*** | 0.2523*** | 0.0941*** | 0.2533*** | 0.0826*** | 0.2596*** | 0.3249*** |
| Terciles | 3-1 | 0.0832*** | 0.2104*** | 0.0973*** | 0.2109*** | 0.0855*** | 0.2124*** | 0.0694*** | 0.2169*** | 0.2792*** |
| | 3-1 net | 0.0706*** | 0.1973*** | 0.0844*** | 0.1978*** | 0.0728*** | 0.1993*** | 0.0565*** | 0.2038*** | 0.2664*** |

This table reports both the average gross and net returns of the ICC long-short portfolios based on actual model forecasts and corresponding simulated undistorted forecasts with the same general accuracy. The sample only includes observations for which simulated ICC are available. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). HVZ refers to the ICC portfolio based on HVZ model earnings forecasts (Hou et al., 2012), EP and RI refer to ICC portfolios based on EP and RI model earnings forecasts (Li and Mohanram, 2014), and ML refers to the ICC portfolio based on ML model earnings forecasts (Hess et al., 2024). The superscript *ud* denotes the undistorted forecast which matches the level of general accuracy of the respective model forecast for each forecast horizon. PERF is the perfect forecast with zero systematic distortions and perfect general accuracy. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

**Table C.6: General accuracy versus systematic distortions: simulation II**

Panel A: Varying median PAFE

| $w_1$ | PAFE | SYSD | 10-1 | 10-1 net | 5-1 | 5-1 net | 3-1 | 3-1 net |
|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0000 | 0.0000 | 0.3929*** | 0.3767*** | 0.3395*** | 0.3249*** | 0.2792*** | 0.2664*** |
| 0.2 | 0.0093 | 0.0000 | 0.3932*** | 0.3770*** | 0.3356*** | 0.3210*** | 0.2731*** | 0.2603*** |
| 0.4 | 0.0185 | 0.0000 | 0.3866*** | 0.3702*** | 0.3235*** | 0.3086*** | 0.2586*** | 0.2457*** |
| 0.6 | 0.0278 | 0.0000 | 0.3741*** | 0.3576*** | 0.3061*** | 0.2911*** | 0.2436*** | 0.2306*** |
| 0.8 | 0.0371 | 0.0000 | 0.3581*** | 0.3416*** | 0.2911*** | 0.2761*** | 0.2319*** | 0.2189*** |
| 1.0 | 0.0463 | 0.0000 | 0.3356*** | 0.3190*** | 0.2747*** | 0.2596*** | 0.2169*** | 0.2038*** |

Panel B: Varying SYSD

| $w_2$ | PAFE | SYSD | 10-1 | 10-1 net | 5-1 | 5-1 net | 3-1 | 3-1 net |
|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0460 | 0.0265 | 0.3356*** | 0.3189*** | 0.2776*** | 0.2625*** | 0.2194*** | 0.2064*** |
| 0.2 | 0.0405 | 0.0344 | 0.3253*** | 0.3087*** | 0.2710*** | 0.2559*** | 0.2165*** | 0.2034*** |
| 0.4 | 0.0387 | 0.0576 | 0.3212*** | 0.3046*** | 0.2598*** | 0.2447*** | 0.2078*** | 0.1947*** |
| 0.6 | 0.0388 | 0.0902 | 0.2850*** | 0.2686*** | 0.2350*** | 0.2200*** | 0.1859*** | 0.1728*** |
| 0.8 | 0.0408 | 0.1246 | 0.2254*** | 0.2091*** | 0.1861*** | 0.1713*** | 0.1457*** | 0.1327*** |
| 1.0 | 0.0464 | 0.1480 | 0.1224*** | 0.1062*** | 0.0894*** | 0.0746*** | 0.0664*** | 0.0534*** |

This table reports both the average gross and net returns of the ICC long-short portfolios based on simulated forecasts with varying median PAFE and fixed SYSD, and vice versa. The sample only includes observations for which simulated ICC are available. I consider three quantile splits, i.e., decile splits, quintile splits, and tercile splits. Transaction costs are accounted for following Brandt et al. (2009), Hand and Green (2011) and DeMiguel et al. (2020). The PAFE and SYSD columns in both panels show the averaged median PAFE and averaged SYSD computed across forecast horizons $t+1$ to $t+5$ of the respective simulated forecast. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. I only test for statistical significance of the portfolio returns. Panel A reports average portfolio returns based on simulated forecasts with varying median PAFE and constant SYSD of zero. The forecasts are derived by simulating undistorted forecasts which match the ML model PAFEs for each forecast horizon, multiplied by a scalar $w_1$. Panel B reports average portfolio returns based on simulated forecasts with varying SYSD and (approximately) constant median PAFE equal to that of the ML model for each forecast horizon. The forecasts are derived by computing a weighted average of the actual ML model forecast and the corresponding simulated undistorted forecast with the same median PAFE. The weight given to the ML model forecast is given by $w_2$ and the weight given to the corresponding undistorted forecast is given by $1 - w_2$.

# Bibliography


Albrecht, W. S., L. L. Lookabill, and J. C. McKeown (1977). The Time-Series Properties of Annual Earnings. *Journal of Accounting Research 15*(2), 226–244.

Ammann, M., G. Coqueret, and J.-P. Schade (2016). Characteristics-Based Portfolio Choice with Leverage Constraints. *Journal of Banking & Finance 70*, 23–37.

Ang, A., S. Gorovyy, and G. B. van Inwegen (2011). Hedge Fund Leverage. *Journal of Financial Economics 102*(1), 102–126.

Ang, N. K. and R. S. Aadka (2017). Implied Cost of Capital in the Cross-Section of Stocks. *Working Paper*.

Ashbaugh-Skaife, H., D. W. Collins, W. R. Kinney, and R. Lafond (2009). The Effect of SOX Internal Control Deficiencies on Firm Risk and Cost of Equity. *Journal of Accounting Research 47*(1), 1–43.

Bali, T. G., H. Beckmeyer, M. Mörke, and F. Weigert (2023). Option Return Predictability with Machine Learning and Big Data. *The Review of Financial Studies 36*(9), 3548–3602.

Ball, R. and P. Brown (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research 6*(2), 159–178.

Ball, R. and R. Watts (1972). Some Time Series Properties of Accounting Income. *The Journal of Finance 27*(3), 663–681.

Best, M. J. and R. R. Grauer (1991). On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results. *The Review of Financial Studies 4*(2), 315–42.

Bianchi, D., M. Büchner, and A. Tamoni (2020). Bond Risk Premiums with Machine Learning. *The Review of Financial Studies 34*(2), 1046–1089.

Bielstein, P. (2018). International Asset Allocation Using the Market Implied Cost of Capital. *Financial Markets and Portfolio Management 32*(1), 17–51.

Botosan, C. A. (1997). Disclosure Level and the Cost of Equity Capital. *The Accounting Review 72*(3), 323–349.

Botosan, C. A. and M. A. Plumlee (2005). Assessing Alternative Proxies for the Expected Risk Premium. *The Accounting Review 80*(1), 21–53.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns. *The Review of Financial Studies 22*(9), 3411–3447.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Brown, L. D. (1993). Earnings Forecasting Research: Its Implications for Capital Markets Research. *International Journal of Forecasting 9*(3), 295–320.

Butler, A. and R. H. Kwon (2023). Integrating prediction in mean-variance portfolio optimization. *Quantitative Finance 23*(3), 429–452.

Callen, J. L. and M. R. Lyle (2020). The Term Structure of Implied Costs of Equity Capital. *Review of Accounting Studies 25*(1), 342–404.

Campbell, J., H. Ham, Z. G. Lu, and K. Wood (2023). Expectations Matter: When (not) to Use Machine Learning Earnings Forecasts. *Working Paper*.

Cao, K. and H. You (2021). Fundamental Analysis via Machine Learning. *Working Paper*.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chen, A. Y. and M. Velikov (2023). Zeroing In on the Expected Returns of Anomalies. *Journal of Financial and Quantitative Analysis 58*(3), 968–1004.

Chen, A. Y. and T. Zimmermann (2022). Open Source Cross-Sectional Asset Pricing. *Critical Finance Review 27*(2), 207–264.

Chen, X., Y. H. T. Cho, Y. Dou, and B. Lev (2022). Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data. *Journal of Accounting Research 60*(2), 467–515.

Chevalier, G., G. Coqueret, and T. Raffinot (2022). Supervised Portfolios. *Quantitative Finance 22*(12), 2275–2295.

Claus, J. and J. Thomas (2001). Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets. *The Journal of Finance 56*(5), 1629–1666.

Cong, L., K. Tang, J. Wang, and Y. Zhan (2021). AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI. *Working Paper*.

de Azevedo, V. G. (2018). The Role of Earnings Forecasts in Asset Pricing Models and Estimates of the Cost of Capital. *Doctoral Thesis. Technical University of Munich*.

DeMiguel, V., L. Garlappi, and R. Uppal (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies 22*(5), 1915–1953.

DeMiguel, V., A. Martin-Utrera, and R. Uppal (2022). A Multifactor Perspective on Volatility-Managed Portfolios. *Working Paper*.

DeMiguel, V., A. Martín-Utrera, F. J. Nogales, and R. Uppal (2020). A Transaction-Cost Perspective on the Multitude of Firm Characteristics. *The Review of Financial Studies 33*(5), 2180–2222.

Detzel, A., R. Novy-Marx, and M. Velikov (2023). Model Comparison with Transaction Costs. *The Journal of Finance 78*(3), 1743–1775.

Diebold, F. and R. Mariano (1995). Comparing Predictive Accuracy. *Journal of Business Economic Statistics 13*(3), 253–63.

Driscoll, J. C. and A. C. Kraay (1998). Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *The Review of Economics and Statistics 80*(4), 549–560.

Easton, P. D. (2004). PE Ratios, PEG Ratios, and Estimating the Implied Expected Rate of Return on Equity Capital. *The Accounting Review 79*(1), 73–95.

Easton, P. D. and S. J. Monahan (2005). An Evaluation of Accounting-Based Measures of Expected Returns. *The Accounting Review 80*(2), 501–538.

Echterling, F., B. Eierle, and S. Ketterer (2015). A Review of the Literature on Methods of Computing the Implied Cost of Capital. *International Review of Financial Analysis 42*, 235–252.

Eldan, R. and O. Shamir (2016). The Power of Depth for Feedforward Neural Networks. In V. Feldman, A. Rakhlin, and O. Shamir (Eds.), *29th Annual Conference on Learning Theory*, Volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940.

Elton, E. J. (1999). Presidential Address: Expected Return, Realized Return, and Asset Pricing Tests. *The Journal of Finance 54*(4), 1199–1220.

Esterer, F. and D. Schröder (2014). Implied Cost of Capital Investment Strategies: Evidence from International Stock Markets. *Annals of Finance 10*(2), 171–195.

Fama, E. F. and K. R. French (1997). Industry costs of equity. *Journal of Financial Economics 43*(2), 153–193.

Fama, E. F. and K. R. French (2008). Dissecting anomalies. *The Journal of Finance 63*(4), 1653–1678.

Fama, E. F. and K. R. French (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics 116*(1), 1–22.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep Learning for Individual Heterogeneity: An Automatic Inference Framework. *Working Paper*.

Feng, G., J. He, and N. G. Polson (2018). Deep Learning for Predicting Asset Returns. *Working Paper*.

Francis, J., R. LaFond, P. M. Olsson, and K. Schipper (2004). Costs of Equity and Earnings Attributes. *The Accounting Review 79*(4), 967–1010.

Francis, J. R., I. K. Khurana, and R. Pereira (2005). Disclosure Incentives and Effects on Cost of Capital around the World. *The Accounting Review 80*(4), 1125–1162.

Frank, M. Z. and T. Shen (2016). Investment and the Weighted Average Cost of Capital. *Journal of Financial Economics 119*(2), 300–315.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Gebhardt, W. R., C. M. C. Lee, and B. Swaminathan (2001). Toward an Implied Cost of Capital. *Journal of Accounting Research 39*(1), 135–176.

Gerakos, J. J. and R. B. Gramacy (2012). Regression-Based Earnings Forecasts. *Chicago Booth Research Paper No. 12-26*.

Gode, D. and P. Mohanram (2003). Inferring the Cost of Capital Using the Ohlson–Juettner Model. *Review of Accounting Studies 8*, 399–431.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Gordon, J. R. and M. J. Gordon (1997). The Finite Horizon Expected Return Model. *Financial Analysts Journal 53*(3), 52–61.

Graham, J. R., C. R. Harvey, and S. Rajgopal (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics 40*(1), 3–73.

Green, J., J. R. M. Hand, and X. F. Zhang (2017, 03). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies 30*(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Guay, W., S. Kothari, and S. Shu (2011). Properties of Implied Cost of Capital Using Analysts' Forecasts. *Australian Journal of Management 36*(2), 125–149.

Hand, J. R. M. and J. Green (2011). The Importance of Accounting Information in Portfolio Optimization. *Journal of Accounting, Auditing & Finance 26*(1), 1–34.

Hansen, J. W. and C. Thimsen (2020). Forecasting Corporate Earnings with Machine Learning. *Working Paper*.

Hansen, L. and P. Salamon (1990). Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(10), 993–1001.

Harris, R. D. F. and P. Wang (2019). Model-Based Earnings Forecasts vs. Financial Analysts' Earnings Forecasts. *British Accounting Review 51*(4), 424–437.

Heaton, J. B., N. G. Polson, and J. H. Witte (2017). Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry 33*(1), 3–12.

Hendriock, M. (2022). Forecasting Earnings with Predicted, Conditional Probability Distribution Density Functions. *Working Paper*.

Hess, D., M. Meuter, and A. Kaul (2019). The Performance of Mechanical Earnings Forecasts. *Working Paper*.

Hess, D., F. Simon, and S. Weibels (2024). Interpretive Earnings Forecasts via Machine Learning: A High-Dimensional Financial Statement Data Approach. *Working Paper*.

Hess, D. and S. Wolf (2022). Quarterly Earnings Information: Implications for Annual Earnings Forecast Models. *Working Paper*.

Hjalmarsson, E. and P. Manchev (2012). Characteristic-Based Mean-Variance Portfolio Choice. *Journal of Banking & Finance 36*(5), 1392–1401.

Hou, K., M. A. van Dijk, and Y. Zhang (2012). The Implied Cost of Capital: A New Approach. *Journal of Accounting and Economics 53*(3), 504–526.

Hribar, P. and N. T. Jenkins (2004). The Effect of Accounting Restatements on Earnings Revisions and the Estimated Cost of Capital. *Review of Accounting Studies 9*, 337 – 356.

Ioffe, S. and C. Szegedy (2015, 07–09 Jul). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning 37*, 448–456.

Israel, R., B. Kelly, and T. Moskowitz (2020). Can Machines "Learn" Finance? *Journal of Investment Management 18*(2).

Jarrett, K., K. Kavukcuoglu, M. Ranzato, and Y. LeCun (2009). What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153.

Jensen, T. I., B. T. Kelly, S. Malamud, and L. H. Pedersen (2022). Machine Learning and the Implementable Efficient Frontier. *Swiss Finance Institute Research Paper No. 22-63*.

Jones, J. J. (1991). Earnings Management During Import Relief Investigations. *Journal of Accounting Research 29*(2), 193–228.

Jones, S., W. J. Moser, and M. M. Wieland (2023). Machine learning and the prediction of changes in profitability. *Contemporary Accounting Research 40*(4), 2643–2672.

Kelly, B., S. Malamud, and K. Zhou (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance 79*(1), 459–503.

Kelly, B. and D. Xiu (2023). Financial machine learning. *Foundations and Trends® in Finance 13*(3-4), 205–363.

Kingma, D. P. and J. Ba (2014). Adam: A Method for Stochastic Optimization. *Working Paper*.

Kothari, S., E. So, and R. Verdi (2016). Analysts' Forecasts and Asset Pricing: A Survey. *Annual Review of Financial Economics 8*(1), 197–219.

Ledoit, O. and M. Wolf (2008). Robust Performance Hypothesis Testing with the Sharpe Ratio. *Journal of Empirical Finance 15*(5), 850–859.

Lee, C. M. C., E. C. So, and C. C. Y. Wang (2020). Evaluating Firm-Level Expected-Return Proxies: Implications for Estimating Treatment Effects. *The Review of Financial Studies 34*(4), 1907–1951.

Li, K. K. and P. Mohanram (2014). Evaluating Cross-Sectional Forecasting Models for Implied Cost of Capital. *Review of Accounting Studies 13*(3), 1152–1185.

Li, Y., D. T. Ng, and B. Swaminathan (2013). Predicting Market Returns Using Gggregate Implied Cost of Capital. *Journal of Financial Economics 110*(2), 419–436.

Liaw, R., E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica (2018). Tune: A Research Platform for Distributed Model Selection and Training. *Working Paper*.

Lundberg, S. M. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance 7*(1), 77–91.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Academic Press Professional, Inc.

Merton, R. C. (1980). On Estimating the Expected Return on the Market: An Exploratory Investigation. *Journal of Financial Economics 8*(4), 323–361.

Michaud, R. O. (1989). The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Financial Analysts Journal 45*(1), 31–42.

Mitchell, T. M. (1997). *Machine Learning*, Volume 1. McGraw-hill New York.

Monahan, S. J. (2018). Financial Statement Analysis and Earnings Forecasting. *Foundation and Trends in Accounting 12*, 105–215.

Moritz, B. and T. Zimmermann (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *Working Paper*.

Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*(3).

Novy-Marx, R. and M. Velikov (2016, 1). A Taxonomy of Anomalies and Their Trading Costs. *The Review of Financial Studies 29*(1), 104–147.

Ohlson, J. A. and B. E. Juettner-Nauroth (2005). Expected EPS and EPS Growth as Determinantsof Value. *Review of Accounting Studies 10*, 349–365.

Penman, S., J. Zhu, and H. Wang (2023). The Implied Cost of Capital: Accounting for Growth. *Review of Quantitative Finance and Accounting 61*(3), 1029–1056.

Politis, D. N. and J. P. Romano (1994). The Stationary Bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Pástor, L., M. Sinha, and B. Swaminathan (2008). Estimating the Intertemporal Risk-Return Tradeoff Using the Implied Cost of Capital. *The Journal of Finance 63*(6), 2859–2897.

Richardson, S., İrem Tuna, and P. Wysocki (2010). Accounting Anomalies and Fundamental Analysis: A Review of Recent Research Advances. *Journal of Accounting and Economics 50*(2), 410–454.

Schipper, K. (1991). Analysts' Forecasts. *Accounting Horizons 5*(4), 105–121.

Simon, F. (2024). Model-Based Earnings Forecast Accuracy and Implied Cost of Capital Portfolio Returns. *Working Paper*.

Simon, F., S. Weibels, and T. Zimmermann (2023). Deep Parametric Portfolio Policies. *Working Paper*.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research 15*(56), 1929–1958.

Tversky, A. and D. Kahneman (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty 5*(4), 297–323.

Uysal, A. S., X. Li, and J. M. Mulvey (2023). End-to-End Risk Budgeting Portfolio Optimization with Neural Networks. *Annals of Operations Research*.

Van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2023). Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. *Review of Financial Studies 36*(6), 2361–2396.

Watts, R. L. and R. W. Leftwich (1977). The Time Series of Annual Accounting Earnings. *Journal of Accounting Research 15*(2), 253–271.

# Lebenslauf

| | |
|---|---|
| Name: | Frederik André Heinrich Simon |
| Adresse: | Melchiorstraße 23, 50670 Köln |
| Geburtstag: | 15.06.1996 |
| Geburtsort: | Boston, USA |

| | |
|---|---|
| 08/2005 - 01/2010: | Gymnasium Ernestinum, Celle |
| 02/2010 - 07/2014: | Gymnasium Kronshagen, Kronshagen<br>Abschluss: Abitur |
| 10/2010 - 02/2018: | Studium der Betriebswirtschaftslehre, Universität Bayreuth<br>Abschluss: Bachelor of Science |
| 10/2018 - 08/2020: | Studium der Betriebswirtschaftslehre, Universität zu Köln<br>Abschluss: Master of Science |
| 10/2020 - aktuell: | Wissenschaftlicher Mitarbeiter am Seminar für ABWL und Unternehmensfinanzen, Universität zu Köln<br>Thema der Dissertation: *Essays in Empirical Financial Research*<br>Abschluss: Promotion |

Ort, Datum

Unterschrift

# Eidesstattliche Erklärung

Hiermit versichere ich an Eides Statt, dass ich die vorgelegte Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Aussagen, Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich geholfen:

Prof. Dr. Dieter Hess, Prof. Dr. Tom Zimmermann und Sebastian Weibels (gemeinsame Projektarbeit für die Inhalte in Kapitel 2 und Kapitel 3).

Weitere Personen, neben den ggf. in der Einleitung der Arbeit aufgeführten Koautorinnen und Koautoren, waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Ort, Datum

Unterschrift