Master thesis submitted in partial fulfillment of the requirements for the degree
Master of Science
at Technische Universität München

Cross-sectional predictability of stock returns in Nordic stock markets using machine learning methods

Reviewer	Prof. Dr. Christoph Kaserer Department of Financial Management and Capital Markets TUM School of Management Technische Universität München
Advisor:	Noorhan Elkhayat
Study program:	TUM-BWL
Composed by:	Jesse Keränen Motorstraße 64 80809 Munich Tel.: +49 (0) 1628410926 Matriculation number: 03748837
Submitted on:	March 27, 2024

Table of Contents

List of Figures

List of Tables

List of Abbreviations

EXC.RET	Excess return	SMB	Size factor
CA	Cash-to-asset ratio	RMRF	Market factor
CTO	Capital turnover	HML	Value factor
INV	Investment	CMA	Investment factor
BEME	Book-to-market ratio	RMW	Profitability factor
CFP	Cashflow-to-price ratio	MOM	Momentum factor
DEBT	Leverage	VO	Trading volume
SP	Sales-to-price ratio	NOSH	Number of shares outstanding
EP	Earnings-to-price ratio	L.OBV	One month lagged on-balance volume
ROA	Return on assets	L.SD	One month lagged turnover
ROE	Return on equity	L.LOG.MV	One month lagged log market value
Q	Tobin's Q	L.IDVOL	One month lagged idiosyncratic volatility
MOM$_{7}$	Intermediate momentum	L.BETA	One month lagged beta
MOM$_{12}$	Momentum	UP	Unadjusted price
MOM$_{36}$	Long-term reversals	DM	Diebold-Mariano statistic
MOM$_{2}$	Short-term reversals		
MOM.IND	Industry momentum		
L.SD	One month lagged standard deviation		
L.HIGH52	One month lagged 52 week high price ratio		

List of Symbols

θ	Parameter vector
$\hat{r}$	Predicted return
$\overline{r}$	Mean return
β	Regression coefficient
α	Regression intercept
ϵ	Regression error term
σ	Standard deviation
w	Weight of the security in a portfolio
R^2_{oos}	Conservative out-of-sample r squared
$R^2_{Trad.\ oos}$	Traditional out-of-sample r squared
d	Average difference in prediction errors

1 Introduction

In the year 1808 world was in many ways different compared to what it is today. In 1808 Napoleon was the Emperor of the French Empire and Maximilian I was ruling the Kingdom of Bavaria. In 1808 Finnish war broke out between the Kingdom of Sweden and the Russian Empire which would ultimately lead to the establishment of the autonomous Grand Duchy of Finland. It would still take more than 100 years for Finland to gain its independence. In the same year began the Dano-Swedish war between Sweden and Denmark-Norway. Something historically far less remarkable, but essential for this study happened in 1808 as well. The first stock exchange in a Nordic country was opened in Copenhagen[1]. Slowly after that rest of the Nordic countries would open their own stock exchanges as well. Upon facilitated change of ownership of securities, investors were left with the question how to price these assets.

A major breakthrough in this topic happened in the sixties when the capital asset pricing model was first developed.[2] In the eighties performance of the capital asset pricing model was questioned and scholars started to come up with variables that could explain portions of cross-sectional stock returns that the capital asset pricing model could not. These so-called stock market factors would include variables such as earnings-to-price ratio, leverage and market capitalization.[3] During these times machine learning gained large interest and artificial neural networks became popular. The next big step in empirical asset pricing happened when Eugene Fama and Kenneth French combined a selection of stock market factors into a coherent three-factor asset pricing model. [4] A few years later researchers made breakthroughs to overcome limitations of the generalization of the decision trees, by ensembling multiple randomized trees, which would ultimately lead to the introduction of a machine learning method called random forest.[5]

Although the three-factor model was a remarkable improvement compared to the capital asset pricing model, it was still not able to explain variation in stock returns completely. Even after the three-factor model lot of new stock market factors have been discovered

[1] See: `https://www.nasdaqomxnordic.com/about_us?languageId=1&Instrument=SSE101`.

[2] See: Sharpe (1964), pp. 425-442; Lintner (1965), pp. 13-37.

[3] See: Basu (1977), p. 680; Bhandari (1988), p. 508; Banz (1981), p.16.

[4] See: Fama and French (1993), pp. 7-10.

[5] See: Breiman (2001), pp. 5–32.

and in the year 2015, Fama and French extended their three-factor model by two additional factors.[6] Characteristic for empirical asset pricing literature in recent years has been a large number of predictive variables. Including more than 100 possibly even strongly correlated explanatory variables, could pose serious challenges to traditional linear regression models. This has led researchers to examine other models that do not suffer from over-parameterization as much as linear regression. In recent years lot of research has applied machine learning methods to capture abnormal returns in stock markets.

The objective of this study is to apply a set of machine learning methods to well-established asset pricing factors to capture abnormal stock return patterns. This study will focus on four Nordic stock markets namely Denmark, Finland, Norway and Sweden. These four developed markets are relatively homogenous in many aspects. They are geographically close, politically stable and economically interconnected. Denmark, Finland and Sweden all belong to the European Union. Additionally, the stock exchanges of all these three countries are operated by Nasdaq, Inc. Therefore, investors could view them as a single market. European market integration is emphasized also by Fama and French, but this study focuses on possibly even more integrated Nordic markets.[7] Some of the features that are characteristic for Nordic markets make them fertile ground for stock market anomaly studies. As mentioned Nordic countries are geographically closely located in northern Europe and therefore relatively distant from large European and especially American markets.

One stock market phenomenon that could affect Nordic stock markets is the periphery effect.[8] It refers to investor's behaviour where during times of a crisis investors tend to liquidate their investments first from the markets more distant to them. This increases the volatility of periphery markets and can challenge the efficient market theory. Another common feature that Nordic markets share is the high level of foreign ownership. Share of foreign investments in Nordic stock markets can reach more than 50%[9]. Given the remote location of Nordic stock markets and their high share of foreign ownership,

[6] See: Fama and French (2015), pp. 2-3.

[7] See: Fama and French (2012), p. 459.

[8] See: Leivo and Pätäri (2011), p. 403.

it is likely that Nordic countries could be subject to periphery effect. Which again can result in abnormal return patterns.

This study contributes to the existing literature in several ways. First of all, it applies a machine learning framework from Gu, Kelly, and Xiu (2020) to a new market. The machine learning approach has been previously applied to European markets, but as mentioned this study focuses on an even more coherent Nordic submarket.[10] The objective is to examine if an investor investing only in Nordic markets could benefit from implementing the machine learning framework of Gu et al. (2020). The dataset of the study is unique in the sense that a lot of machine learning approach studies have been conducted in wider markets such as European stock markets, whereas previous Nordic stock market anomaly studies have mainly focused on single markets. Pooling the four Nordic markets ensures us a sufficient amount of data to train complex machine learning models, but also allows us to focus on homogenous clearly defined submarket.

Additionally, existing stock market anomaly literature focusing solely on Nordic markets is rather limited. Section 2.3 introduces the existing Nordic stock market anomaly literature. Characteristic for studies in this section is that they mainly focus on one or two stock market anomalies. As this study constructs 23 stock characteristic anomalies, it allows to us examine anomalies that have not been studied in Nordic stock markets previously. This means that this study cannot only reveal the profitability of the machine learning framework in the Nordic market setting, but it can also reveal evidence of the existence of certain stock market anomalies in Nordic markets. As mentioned lot of Nordic stock market anomaly research focuses only up to two anomalies at a time. Since this study includes 23 anomalies simultaneously, it allows us also to examine the performance of already discovered Nordic anomalies while controlling for many other variables. Applying sophisticated machine learning models allows us to control for more complex interactions.

The objective of this study is slightly more ambitious than in existing Nordic stock market anomaly literature. Existing literature mainly examines the existence of anomalies

[9] See: Butt and Högholm (2018), p. 3, Butt and Högholm calculate the share of foreign ownership from IMF Coordinated Direct Investment Survey CDIS data. Foreign ownership share of Butt and Högholm is 52% for Denmark, 42% for Finland, 35% for Norway and 56% for Sweden.

[10] See: Drobetz and Otto (2021), p. 510; Fieberg, Metko, Poddig, and Loy (2023), pp. 304-307.

by uni- or multivariate portfolio sorts. Studies predefine variables of interest and form portfolios based on these variables. Then the historical excess returns are investigated. This study goes one step further and attempts to predict stock level out-of-sample returns based on the predefined set of variables. This allows us to evaluate which portion of the return variability these variables are able to capture in addition to the profitability of the strategy.

The final contribution of this study is to expand the explanatory variable set. This study includes a variable called on-balance volume. On-balance volume is a technical trading indicator that has not been studied in a great extent as a cross-sectional stock return predictor. Due to the previous strong performance of momentum indicators, this study includes several momentum variables. The extended variable set allows us to examine whether an on-balance anomaly exists in Nordic stock markets or whether including on-balance volume affects the performance of well-established momentum indicators.

The structure of this paper goes as follows. In the second chapter introduction to related previous literature is provided. In this chapter performance of different methods and the persistence of different anomalies in different regions are discussed. The third chapter introduces the data used in this study and the filters applied to the data. The fourth chapter presents the methodology. It introduces the implemented models in more detail and describes the measurements applied in order to evaluate the performance of the models. The fifth chapter focuses on benchmarking factors, showing how benchmark factors are constructed and how well they perform in Nordic markets. The sixth chapter describes the results of the empirical study. The chapter is divided to discuss separately predictive accuracy, economic profitability and characteristic importance for the machine learning models. Finally, the last chapter provides a conclusion of the empirical study.

2 Stock return anomaly literature

Being the largest and most prominent stock market in the world US stock market has been subject to the majority of asset pricing studies. Despite the dominance of US markets in capturing the attraction of academics, a lot of empirical asset pricing literature has been conducted in international settings as well. Characteristic for international asset pricing literature is that instead of focusing on single countries they aggregate stock market data to a certain regional level such as Europe or Asia-Pacific. The following chapter provides an overview of pioneering asset pricing anomaly literature. The focus will mainly be on the literature on the US, European and Nordic markets. US stock markets are chosen because of their significant impact on international stock markets and because most anomalies have been discovered there and therefore majority of the initial studies of these anomalies have been conducted there. European studies provide an interesting perspective for this study since in many of them Nordic countries are included.

The chapter introduces the most important anomalies in these markets and how they have been exploited with different methods. This works as a starting point to define a set of factors that will be used in this study. It can be argued that this kind of process when the set of variables is chosen based on their performance in previous studies is one sort of forward-looking information as we later try to mimic the information set of a historical investor. On the other hand Jacobs and Müller (2020) only find a reliable post-publication decline in long/short factor returns in the US, which emphasizes the practical potential of this study. [11]

2.1 US stock market anomalies

Many of the recent cross-sectional stock return studies use the framework of Lewellen (2015) as the base model. He runs 10-year rolling Fama and MacBeth (1973) regressions using lagged firm characteristics to predict out-of-sample stock returns. He studies cross sections of US stock returns between 1964 and 2013 using different model settings up to 15 company characteristics.[12] He finds a strong positive correlation between expected returns derived from rolling Fama-MacBeth regressions and realized returns. Additionally, Lewellen shows that the spread between the realized return of the portfolio formed from stock with the lowest expected returns and the portfolio with the highest expected return is up to 2.36%. In his study logarithmic market value of equity, logarithmic book-to-market value, momentum and accruals show the strongest

[11] See: Jacobs and Müller (2020), pp. 218-219.

[12] See: Lewellen (2015), pp. 2-5.

statistical power in explaining monthly returns using lagged variables.[13]

Gu et al. (2020) contribute to the literature by applying machine learning methods to exploit the stock market anomalies. By deploying sophisticated models that do not suffer from over-parameterization as heavily as OLS Gu et al. are able to include 94 stock characteristics and their interactions as well as eight aggregated time series variables to their models.[14] Gu et al. use a large variety of statistical methods including linear regression, generalized linear models with penalization, dimension reduction via principal components regression and partial least squares, gradient-boosted regression trees, random forest and different settings of neural networks.[15] Gu et al. Out of these gradient-boosted regression trees and neural networks explain the monthly out-of-sample stock returns the best, reaching out-of-sample R^2 of 0.33% and 0.44% correspondingly whereas the ordinary least squares model only reaches out-of-sample R^2 of -3.46%.[16]

Similar to Lewellen (2015), Gu et al. construct portfolios based on the predicted return of different models. Monthly spread in realized return between portfolio constructed from decile of companies with lowest expected return and decile of stocks with highest expected return is 0.94%, 1.62% and 2.26% for models based on OLS, random forest and four-layer neural network correspondingly. [17]. Gu et al. also show that all methods they examine show somewhat similar patterns on variable importance on return predictability. The most important factors are price trends such as momentum followed by stock liquidity, stock volatility, and valuation ratios.[18]

2.2 European stock market anomalies

As mentioned, the US stock market environment is different in many ways compared to Nordic markets. Fortunately lot of stock market studies have been conducted in Europe. Since Nordic markets are usually just a subset of European markets it can be

[13] See: Lewellen (2015), pp. 14-19 and pp. 23-26.

[14] See: Gu et al. (2020), p. 2248.

[15] See: Gu et al. (2020), pp. 2232-2246.

[16] See: Gu et al. (2020), pp. 2250-2252; Gu et al. use five different settings of neural networks differing by the number of hidden layers. A neural network with three hidden layers reaches the highest R^2_{oos} and is reported here.

[17] See: Gu et al. (2020), pp. 2265-2266; Portfolio returns are average value-weighted returns.

[18] See: Gu et al. (2020), p. 2254.

beneficial to have a look at the European studies. Tobek and Hronec (2021) study machine learning-based anomaly strategies in an international setting. Their study includes 153 different equity anomalies and they only include anomalies to their data after the documented discovery of corresponding anomalies. This way they can mimic the information set an investor would have had and avoid forward-looking information. Tobek and Hronec examine five different models including weighted least squares, penalized weighted least squares, gradient boosting regression trees, random forest and neural networks. Their data set spans from 1990 to 2018.[19]

Similar to Gu et al. (2020) in the US, Tobek and Hronec find that strategy using neural networks provides the highest returns on quintile long-short portfolios. The mean return for neural network long-short portfolio in Europe was 0.7%. Interestingly penalized weighted least square method provided a mean return of 0.65% which is higher than the return of the random forest-based portfolio's return of 0.40%. Tobek and Hronec find that Industry momentum, lagged momentum, liquidity shocks and 52-week high price are the most important variables for neural networks mode.[20]

Exploiting stock market anomalies using machine learning methods is also studied by Drobetz and Otto (2021). Their data set contains all companies listed in at least one of the 19 Eurozone countries until December 2020 and spans from January 1990 to December 2020.[21] Drobetz and Otto examine the performance of ordinary least squares, penalized least squares, principal component regressions, partial least squares, random forests, gradient boosted regression trees and neural networks on predicting monthly stock-level returns exploiting a set of 22 predictors, their two-way interactions and second- and third-order polynomials. Findings of Drobetz and Otto are similar to Gu et al. (2020). They show that with a large number of explanatory variables simple linear regression is not able to explain well out-of-sample stock returns.[22]

[19] See: Tobek and Hronec (2021), pp. 3-8.

[20] See: Tobek and Hronec (2021), p. 13 and p.16; Tobek and Hronec discover possibilities for training models either only using historical data from the US, using historical data from local markets or using international historical data. Only results for models trained using local data are reported here because that is closest to the setting of this study. Additionally, Tobek and Hronec state that the difference between model trained on US data and local data is small.

[21] See: Drobetz and Otto (2021), p. 510; Finland is the only country included in the study of Drobetz and Otto that is also included in this study since it is the only country belonging to the Eurozone.

[22] See: Drobetz and Otto (2021), pp. 513-516 and p. 522.

Findings of Drobetz and Otto (2021) are also similar to Tobek and Hronec (2021) in the sense that least squares methods where dimensionality is restricted can actually perform better than tree-based methods. Like in the majority of other literature, Drobetz and Otto find out that neural networks provide a superior framework for stock return prediction models measured in both explanatory power and economic profitability. The neural network method reaches an out-of-sample R^2 value of 1.23% and the long-short portfolio formed based on expected returns derived from the neural networks model provides an average value-weighted monthly return of 1.94%.[23] Similar to Gu et al. (2020), Drobetz and Otto find that the same variables show the most importance across the different models, most notably earnings-to-price ratio and 12-month momentum.[24]

Fieberg, Metko, Poddig, and Loy (2023) study stock market anomalies in 16 European stock markets using machine learning methods over almost the same period as Drobetz and Otto (2021). Nevertheless, they choose a slightly different approach where instead of including a vast set of anomalies they only consider six prominent equity factors. Factors Fieberg et al. consider are beta, market capitalization, the book-to-market-equity ratio, momentum, investment and operating profitability. These factors correspond to the benchmark factor set of this study discussed in Section 5. Their conclusion endorses the findings of Drobetz and Otto (2021) and Tobek and Hronec (2021) as they show that more complex machine learning models beat the linear approach in terms of both economic and statistical performance. [25]

2.3 Nordic stock market anomalies

This chapter provides an overview of discovered stock market anomalies in different Nordic stock markets. Many studies in this chapter have slightly different objectives than this study. Studies show the existence of the anomalies by constructing a portfolio heavily weighted on a certain factor. Nevertheless, they do not describe the magnitude of the relationship between the factor and the expected stock return. This study has a slightly more ambitious objective and tries to derive return expectations from predefined

[23] See: Drobetz and Otto (2021), pp. 521 and p. 524.

[24] See: Drobetz and Otto (2021), p. 516.

[25] See: Fieberg et al. (2023), pp. 304-314; Dataset of Fieberg et al. contains Denmark, Finland, Norway and Sweden.

stock market factors. This literature review serves as a starting point for choosing the most promising stock market factors that have already been studied.

The magnitude of value and momentum anomalies in Nordic stock markets are examined in the paper by Grobys and Huhta-Halkola (2019). They combine information from companies listed in the main lists of Danish, Finnish, Norwegian and Swedish stock exchanges between 1991 and 2017. Grobys and Huhta-Halkola measure value with price-to-book value and momentum with past 12-month total shareholder return.[26] Grobys and Huhta-Halkola show that the momentum effect exists in Nordics markets and profitability of momentum strategy is not related to the size factor. Value factor yields also significant excess return, but according to Grobys and Huhta-Halkola it could be partly driven by the size factor, since value premium reduces when accounted for the size. Among all stocks, the monthly average equally weighted long-short return is 1.72% and 1.25% for momentum and value strategies correspondingly. Both of the excess returns are statistically highly significant. Grobys and Huhta-Halkola also test combination strategies using signals from both momentum and strategy which yield even stronger results.[27]

Value premium has shown consistency also in Finnish stock market. Davydov, Tikkanen, and Äijö (2017) examine the profitability of different value investing strategies between 1991 and 2013. Davydov et al. investigate a set of value indicators which included earnings-to-price, book-to-price, cashflow-to-price, dividends-to-price and earnings before income and taxes-to-enterprise value ratios. Additionally, they test the performance of investing strategy where portfolios are formed based on the combined ranking of the company's return on invested capital and earnings before income and taxes-to-enterprise value ratio. Davydov et al. (2017) show that returns of all of the value portfolios not only beat the market return but can also not be explained by the four-factor model of Carhart (1997). [28]

Similar to Grobys and Huhta-Halkola (2019), Leivo and Pätäri (2011) combine value anomaly with momentum anomaly in the Finnish stock market for data set between 1993

[26] See: Grobys and Huhta-Halkola (2019), pp. 6-7.

[27] See: Grobys and Huhta-Halkola (2019), p. 4.

[28] See: Davydov et al. (2017), pp. 41-42 and p. 46; Carhart (1997), p. 61.

and 2008. They show that a two-step portfolio sort that first allocates stocks to three portfolios based on their value indicators and subsequently based on the momentum indicator can capture extraordinary stock returns. Leivo and Pätäri show that including momentum further increases returns of already recognized value sorting. The strategy performs even better when authors allow for a long position in a high value high momentum portfolio and a short position in a low value low momentum portfolio. Excess returns resulting from the two-fold portfolio construction cannot be explained by the capital asset pricing model or two-factor model including also the size factor. It is not a surprise that value and momentum premiums exist in Nordic markets.[29] Value and momentum anomalies are among the most well-documented factors showing persistence in multiple international cross-sectional studies.[30]

Nordic stock markets have several characteristic features. One is that all Nordic stock markets are considered to be developed, but also small. Especially market capitalization of companies listed in Nordic stock exchanges are on average much smaller than their international counterparties. Therefore, it is reasonable to ask whether the liquidity of the stock could be a driving factor of the stock returns. The impact of illiquidity risk on stock returns in the Nordic market setting has been studied by Butt and Högholm (2018).

Butt and Högholm test a variety of different illiquidity measures and find that dollar zero returns is the most profitable illiquidity anomaly measure across all four Nordic markets. Dollar zero return measurement is calculated by dividing the number of days stocks return in US dollars is zero by the total number of trading days. Butt and Högholm construct five quintile portfolios based on the liquidity of the stocks with data spanning from April 1988 to September 2013. They show that in all Nordic markets there exists a large illiquidity premium as the annual difference in equal-weighted return of the most illiquid portfolio and least illiquid quintile portfolio is more than 18% for Finland, Norway and Sweden. For Denmark premium is slightly smaller 8.8%.[31]

Jokipii and Vähämaa (2006) investigate free cash flow anomaly in Finnish stock market

[29] See: Leivo and Pätäri (2011), p. 403, p.407 and 412

[30] See: e.g. Gu et al. (2020), p. 2254; Lewellen (2015), pp. 9-14; Drobetz and Otto (2021), p. 518; Tobek and Hronec (2021), p. 16.

[31] See: Butt and Högholm (2018), p. 5 and pp. 12-14.

between 1992 and 2002. They construct portfolios from stocks listed in the Finnish stock exchange based on predefined thresholds for free cash flow ratios. These ratios include market value to free cash flow and total debt to free cash flow ratios. A high free cash flow portfolio yields higher returns than the market on average and the excess returns can not be completely explained by weightings in Fama and French (1993) risk factors. [32]

3 Data

This section provides an overview of the dataset used in this study. The section starts by introducing overall market characteristics in Nordic stock markets. The section discusses how companies are distributed across Nordic markets and also describes the size properties of the companies in different Nordic markets. This part also describes the static and dynamic screens applied to the Datastream data in order to ensure sufficient data quality. The second part of the section describes the firm-level characteristics considered in this study. This includes stock-level excess returns as well as all independent variables. This part introduces definitions of all variables including which characteristics are included in the calculation of each variable. Descriptive statistics of the firm-level characteristics are also presented in this part of the study.

3.1 Nordic stock market data

The main data source for this study is Refinitv Datastream. Company fundamentals data is collected from Worldscope database. The dataset spans from 1990 January to 2022 December which is shorter than in many previous studies in the US stock markets, but is close to lot of studies conducted in Europe.[33] The reason why the period is limited to 1990 is that the amount of publicly listed companies in Nordic markets was rather low in the 1980s and finding reliable data for the period before 1990 is difficult. The dataset contains all stocks listed in primary markets of corresponding countries including also companies that went bankrupt or were de-listed for any other reason. Therefore, the

[32] See: Jokipii and Vähämaa (2006), pp. 963-964; Fama and French (1993), pp. 7-10.

[33] See: e.g. Drobetz and Otto (2021), p. 510; Tobek and Hronec (2021), pp. 3-4.

dataset is not subject to survivorship bias.

Table A.1 in the appendix shows the constituent lists used in data collection. As highlighted by Ince and Porter (2006) data from Datastream can be noisy and uncleaned data could lead to a false statistical inference.[34] Therefore, several static and dynamic screens are applied to the data. Static screens include filtering non-equity securities, securities that are not listed in the respective country and securities that are quoted in a currency other than respective country's currency. Panel A from Table A.3 shows which values are accepted for the type of instrument, ISIN code, code indicating the country of origin of the company, country where the security is listed, currency in which the security is noted and ISIN country code.

In order to filter non-common and duplicate stock affiliations keywords indicating such securities are searched from Datastream attributes NAME, ENAME and ECNAME. Panel B of Table A.3 presents the country-specific keywords. These keywords are only searched for securities from specific countries but among all the above-mentioned attributes. Keywords from Table A.4 are searched from name attributes of securities from all countries. If a keyword is found from any of the name attributes, the security will be removed from the dataset. Keyword deletion follows process of Ince and Porter (2006) and Hanauer and Windmüller (2023).[35]

As mentioned Ince and Porter (2006) argue that data quality issues in Datastream could even lead to wrong conclusions.[36] In order to avoid results being driven by extraordinary data points, which could be caused by data quality issues, dynamic screens are applied to the data. Table A.2 in the appendix presents the applied dynamic screens. Observations are removed from the dataset in case of extreme abnormal returns. Observations are also removed in case of extremely strong strong return reversals.

One special characteristic has to be taken into consideration when working with data from Datastream. In case the company is delisted for some reason Datastream returns the last available value for the remaining periods in the query. In order to only include actively traded securities these observations have to be cleaned from the dataset. This

[34] See: Ince and Porter (2006), pp. 475-479.

[35] See: Ince and Porter (2006), pp. 475-479; Hanauer and Windmüller (2023), pp. 17-19.

[36] See: Ince and Porter (2006), p. 479.

could be done with variable TIME from Datastream which shows the date of last equity price data. Nevertheless, Ince and Porter (2006) argue that the TIME attribute is not a reliable indicator of the delisting date, but propose to remove consecutive zero returns from the end of the dataset. Removal of zero returns from the end of the dataset could lead to the removal of actual zero returns, but the effect of this is considered to be smaller than the noise caused by the usage of TIME variable.[37] Therefore, all consecutive zero returns at the end of the dataset are removed for all companies.

Table 1: Country summary statistics - Own source
Table provides summary statistics for pooled Nordic market and separate country-specific Nordic markets. The minimum number of companies tells the number of companies included in the data set in a month that the value was lowest for the respective country. The maximum number of companies tells the number of companies included in the data set in a month that the value was highest for the respective country. The mean number of companies is the time series average of the monthly number of companies for each country. Total number of companies is the number of unique companies in the whole data set. Time series averages for monthly mean, median and total market values are also presented. Total market value is the sum of the market values of the respective country in each month. All market values are converted to US dollar and expressed in millions. Only companies in the final dataset are included in the calculation of the figures. The microcap stocks are excluded from the dataset. The dataset spans from January 1990 to December 2022.

	Number of companies				Market value		
Market	Min	Max	Mean	Total	Mean	Median	Total
Denmark	42	106	70	235	2400.14	810.90	141659.8
Finland	26	83	62	186	1893.78	634.60	124389.7
Norway	44	132	79	408	1520.80	506.17	124692.2
Sweden	45	256	132	593	2115.69	616.59	308958.1
Nordic	200	527	343	1422	1946.60	583.07	699699.8

On average number of companies with large market capitalization is more limited in Nordic countries than in the United States or Europe. The smallest companies can be numerous, but still only account for a fraction of total market capitalization. The liquidity of these companies is often also quite low. To avoid results being driven by such stocks, the approach of Hanauer and Kalsbach (2023) for emerging markets is applied and companies with the smallest market value that account for 3% of the aggregated market value are excluded. On the other hand, we do not want a few extremely large companies to drive the results either. Therefore, the market value of the companies is

[37] See: Ince and Porter (2006), p. 465.

winsorized monthly to 99%. If company's market value is among the 1% biggest market values in the corresponding month, the market value will be replaced by the 1% threshold value.[38]

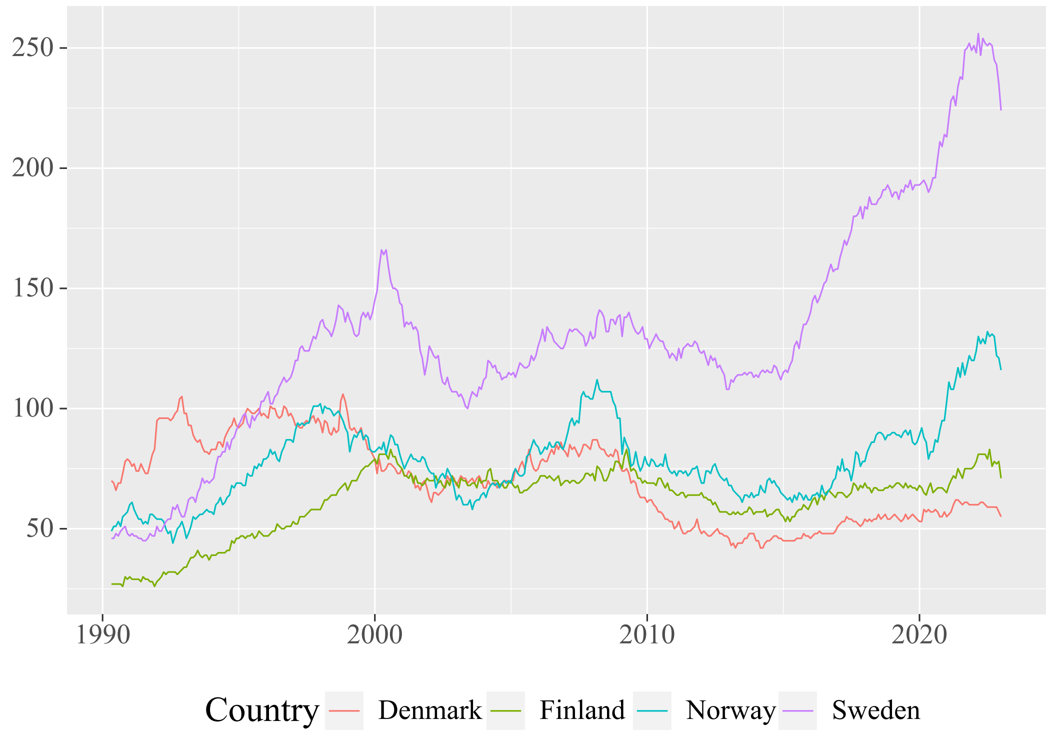
Table 1 presents the number of companies and their sizes in separate and pooled Nordic markets after applying the above-described filters. The total number of non-microcap stocks in the dataset is 1422 whereas the monthly number of stocks in the dataset is 343 on average. Figure 1 shows the development of the non-microcap company count that passed the static and dynamic screens over time. Figure C.1 in the appendix shows the development of company counts over time including microcap stocks. Comparing the two figures it can be seen that even though microcap stocks account only for 3% of the aggregated market value, they account for a remarkable share of company count also in Nordic stock markets. The maximum number of companies including microcap stocks exceeds 1000 whereas the maximum number of companies excluding microcap stocks is slightly above 500.

Sweden is clearly the biggest of the four Nordic markets both in regards to the number of companies and total market value of the companies. Even though Sweden is the biggest market it is not dominating. On average Sweden accounts for less than half of the total market value of the pooled Nordic market. In regards to average and median market value, Denmark has the biggest companies. Finland on the other hand is clearly the smallest of the markets included in the study measured both in number of companies and market value of the companies.

In this study, Nordic markets are examined as one market. In the introduction, it was argued that in the eyes of foreign investors Nordic markets can appear quite homogenous. There is also a more practical reason why Nordic markets are pooled in this study. Table 1 shows that individual Nordic stock markets hold a limited amount of large market capitalization stocks. This leads to a situation where the performance of the whole market, or portfolios formed from the market, can be driven by very few large market capitalization stocks. This could happen even after winsorizing the market values. Later in the study, we will allocate stocks to portfolios based on their expected returns

[38] See: Hanauer and Kalsbach (2023), pp. 3-4.

Figure 1: Number of non-microcap companies - Own source
Figure shows the development of the total number of securities considered in the dataset from January 1990 to December 2022 for each Nordic country. The figure counts all non-microcap securities that passed the static and dynamic screens.



and we want to ensure that there exists a reasonable amount of companies to diversify each portfolio.

Unfortunately, Nordic countries have different currencies. In order to ensure comparability of the companies from different countries, we have to convert certain variables to a common currency which in this case is the US dollar. Variables will be converted using historical currency spot rates. Variables that are converted to US dollars include for example return index and market capitalization. Development of the foreign exchange rates are shown in Figure C.2 in the appendix. The majority of the explanatory variables are some sort of ratios that can be directly calculated from the local currency values.

3.2 Company characteristics

A total of 23 characteristics are derived from the stock-level data. All the models use the same set of explanatory variables which includes book-to-market ratio (BEME), investment (INV), earnings-to-price ratio (EP), cash-to-total assets ratio (CA), capital turnover (CTO), cashflow-to-price ratio (CFP), leverage (DEBT), sales-to-price ratio (SP), return on assets (ROA), return on equity (ROE), Tobin's Q (Q), one-month momentum (MOM_2), momentum from $t-12$ to $t-7$ month (MOM_7), momentum from $t-12$ to $t-2$ (MOM_{12}), momentum from $t-36$ to $t-12$ (MOM_{36}), industry momentum (MOM.IND), log scaled market capitalization (L.LOG.MV), standard deviation (L.SD), ratio of current price and 52-week high price (L.HIGH52.RATIO), beta coefficient (L.BETA), idiosyncratic volatility (L.IDVOL), turnover (L.TO) and on balance volume (L.OBV).

Data consists of variables that are available on three different frequencies. The majority of these variables are ratios calculated from accounting data. Usually, income statement and balance sheet information are available annually and therefore majority of accounting-based variables are updated only once a year. To account for possible reporting delay associated with accounting data, accounting data from year t is considered to be available end of June $t+1$. This timeline follows the common approach of Fama and French (1993).[39] Detailed descriptions of how each of these variables is calculated are provided in Table A.5. Table A.5 also provides corresponding Datastream items used in the calculation of the variables.

The dataset contains also variables calculated from monthly data. These include momentum variables and the market value. Even though the frequency of the return prediction will be monthly, some variables are calculated from weekly data. These include standard deviation, the ratio between price and 52-week high price, beta coefficient, idiosyncratic volatility, turnover and on-balance volume. Nevertheless, also these variables are updated only monthly and therefore these variables are noted as having monthly frequency in Table A.5.

The set of explanatory variables includes seven value indicators. Book-to-market value is calculated by dividing the sum of common equity and deferred taxes by the market value of last December. Also four of the other value indicators are price ratios. Income before extraordinary items, net cash flow from operating activities and net sales are di-

[39] See: Fama and French (1993), p. 8.

vided by the market capitalization of the previous year December in order to obtain earnings-to-price, cash flow-to-price ratio and sales-to-price ratios correspondingly. Leverage is calculated by first subtracting common equity from total assets and then dividing by market capitalization from the previous year's December. The rest two of the value indicators are normalized by the total assets. Cash-to-total assets ratio is calculated by dividing cash and short-term investments by total assets and Tobin's Q is calculated by summing up total assets and market capitalization from the previous December, then subtracting cash and short-term investments and deferred taxes and finally dividing by the total assets.

The profitability of the companies is described with three indicators. Return on assets and return on equity divide earnings before extraordinary items by lagged total assets and lagged book equity. As described above, book equity is defined as the sum of common equity and deferred taxes. The third profitability indicator is capital turnover. Capital turnover is calculated by dividing net sales by the total assets. Momentum characteristics are described by five different momentum variables. Momentum variables include traditional and intermediate momentum as well as short-term and long-term reversals. Traditional momentum is defined as cumulative return from $t-12$ to $t-2$ and intermediate momentum as cumulative return from $t-12$ to $t-7$. Short-term reversal is the return of the previous month whereas long-term reversal is defined as cumulative return from $t-36$ to $t-12$. The final momentum indicator is the industry momentum. Industry momentum is defined as the 12-month cumulative equal-weighted return of an industry sector. Industries are defined by the INDG Datastream attribute.

Trading frictions are estimated by six variables. Beta coefficient and idiosyncratic volatility are calculated by regressing returns of the stocks by the excess market return. As described above, in order to pool the dataset certain variables are converted to US dollars. One of these variables is weekly unadjusted stock price which is used to calculate the weekly stock returns used in the regression. The market return is constructed as an equal-weighted weekly market return following Green, Hand, and Zhang (2017).[40] Because the returns are noted in US dollars one-month Treasury bill rate,

[40] See: Green et al. (2017), p. 4427.

Table 2: Descriptive statistics - Own source
Table provides the time series averages of cross-sectional means and standard deviations of all variables used in this study. Values are reported separately for the pooled Nordic market and four Nordic markets Denmark, Finland, Norway and Sweden. EXC.RET is the monthly excess return calculated from the total return index noted in US dollars. The risk-free rate used to calculate excess returns is a US dollar one-month Treasury bill rate. The time period spans from January 1990 to December 2022.

	Nordic		Denmark		Finland		Norway		Sweden	
Variable	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
EXC.RET	0.007	0.100	0.007	0.087	0.009	0.089	0.006	0.106	0.009	0.098
CA	0.117	0.148	0.107	0.159	0.108	0.119	0.132	0.159	0.115	0.144
CTO	0.813	0.710	0.739	0.670	0.970	0.647	0.651	0.674	0.868	0.741
INV	0.161	0.422	0.118	0.305	0.097	0.273	0.203	0.481	0.179	0.443
BEME	0.699	0.709	0.688	0.558	0.757	0.661	0.680	0.630	0.753	0.827
CFP	0.080	0.122	0.086	0.140	0.090	0.099	0.093	0.145	0.065	0.102
DEBT	2.574	5.231	3.025	5.249	2.415	4.646	3.331	6.456	2.353	4.443
SP	1.585	2.279	1.356	1.636	2.066	2.106	1.316	1.719	1.940	2.849
EP	0.045	0.143	0.044	0.114	0.051	0.106	0.026	0.163	0.058	0.160
ROA	0.045	0.103	0.043	0.096	0.053	0.072	0.029	0.110	0.052	0.112
ROE	0.114	0.238	0.119	0.211	0.118	0.183	0.095	0.290	0.120	0.228
Q	0.670	0.327	0.585	0.385	0.721	0.286	0.626	0.346	0.719	0.283
MOM$_7$	0.080	0.229	0.069	0.198	0.072	0.201	0.080	0.245	0.089	0.230
MOM$_{12}$	0.171	0.382	0.149	0.329	0.153	0.320	0.172	0.411	0.191	0.387
MOM$_{36}$	0.397	0.751	0.401	0.679	0.363	0.624	0.356	0.796	0.431	0.754
MOM$_2$	0.015	0.093	0.012	0.081	0.014	0.085	0.015	0.097	0.016	0.093
MOM.IND	1.132	0.268	1.121	0.232	1.134	0.273	1.133	0.274	1.137	0.265
L.SD	0.048	0.032	0.045	0.029	0.042	0.029	0.052	0.034	0.051	0.029
L.HIGH52	0.684	0.286	0.721	0.268	0.621	0.294	0.683	0.272	0.695	0.261
L.BETA	0.882	0.542	0.708	0.396	0.747	0509	0.941	0.562	1.030	0.527
L.IDVOL	0.046	0.029	0.044	0.025	0.039	0.026	0.051	0.033	0.047	0.026
L.LOG.MV	6.338	1.359	6.408	1.347	6.411	1.300	6.131	1.268	6.458	1.428
L.TO	0.052	0.226	0.071	0.258	0.027	0.082	0.033	0.109	0.067	0.206
L.OBV	0.294	2.392	0.720	4.802	0.105	0.451	0.184	0.640	0.264	0.988

which is obtained from Kenneth French's database[41], is used as a risk-free rate proxy.

Development of the risk-free rate throughout the examined period is shown in Figure C.3 in the appendix. The regression is run for each company separately for each month on a rolling basis. For each regression up to three years of weekly historical data is considered, but a minimum of 15 weeks of data is required. Finally, the beta is simply the sensitivity of stock returns on the market return changes and the idiosyncratic volatility is the standard deviation of the regression residuals obtained from the same regression.

In addition to beta coefficient and idiosyncratic volatility, trading frictions are also mea-

[41] See: `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

sured by turnover, standard deviation, market capitalization and 52-week high price. Turnover is calculated by dividing weekly trading volume by the number of shares outstanding. Standard deviation is calculated from up to 52 weeks of weekly unadjusted price data. The one-month lagged logarithm of market value is used as a size indicator. The 52-week high indicator is also calculated from weekly data and it is defined as the ratio between the highest unadjusted weekly price in the last 52 weeks and the current price. The investment characteristic of the companies is measured by the yearly growth rate of total assets.

This study introduces a new explanatory variables class to the cross-sectional stock return literature. On-balance volume which is traditionally used as a technical trading indicator is included in the explanatory variable set. On-balance volume is often used as a time series predictor for individual securities, but the objective of this study is to investigate whether it could contain information about the cross-section of future stock returns. Calculation of on-balance volume consists of two steps and it is also calculated from weekly data. First current weekly turnover is multiplied by the sign of the corresponding week's return. Then this product is added to the cumulative sum of the previous on-balance volumes. On-balance volume is defined as

$$OBV_{i,t} = OBV_{i,t-1} + \begin{cases} \text{trading volume}, & \text{if } r_{i,t} \geq 0 \\ \text{- trading volume}, & \text{if } r_{i,t} < 0 \end{cases} \tag{1}$$

where $OBV_{i,t}$ is the on-balance volume value of the company i at time t, $OBV_{i,t-1}$ is the on-balance volume value at $t-1$, trading volume is the weekly trading volume and $r_{i,t}$ is the return of the company i at time t.

As described above covariates included in this study can be clustered to value, trading frictions, momentum, investment and profitability. Since trading frictions include size and beta coefficient, all dimensions of the typical Fama and French framework are considered. For many of the dimensions lot of alternative indicators are also considered. For example, whereas the traditional Fama and French (1993) model only evaluates the value characteristic of a company by the book-to-market value, this study simultaneously considers six additional value indicators.[42]

Machine learning algorithms can be sensitive to outliers in the data. Therefore, all explanatory variables are winsorized between the 1st and 99th percentiles. In case the value of a certain variable is less than the 1st percentile of the corresponding month's values it will be replaced by the 1st percentile threshold value. In case the value of the certain variable is above the 99th percentile of the corresponding month's values it will be replaced by the 99th percentile threshold value. Additionally, any missing value in explanatory variables will be replaced by zero.

Table 2 provides descriptive statistics of the company characteristics. For each of the variables time series average and standard deviation of the cross-sectional mean are reported. Values are reported for pooled Nordic market as well as individual markets. The table shows that the mean monthly excess return of Nordic market during the studied period was 0.7%. The mean excess return of Sweden and Finland is slightly above that and the mean excess return of Denmark and Norway is slightly below that. Additionally, Figure C.4 in the appendix shows the time series development of cross-sectional means of the firm-level characteristics. Some trends can be seen in the time series. For example, for many of the variables, a clear change in the mean value can be seen during the financial crisis.

In this study, the excess return is defined as a spread between the return noted in US dollars and the risk-free rate. This follows the approach of Gu et al. (2020). Hanauer and Kalsbach (2023) use an alternative definition of excess market return. While training machine learning models they use stock return demeaned by the value-weighted average market return of the company's home country as the independent variable. Then they rank companies into portfolios in a country-neutral manner. Hanauer and Kalsbach conduct their study on emerging markets which can be geographically extremely scattered. Additionally, the political systems of the emerging markets can be diverse and their economies might not be tightly connected. This justifies their approach. Nevertheless, in this study, four Nordic markets are pooled and treated as one market. Therefore, a more traditional definition of the excess return is chosen for this study.[43]

[42] See: Fama and French (1993), p. 8.

[43] See: Gu et al. (2020), p. 2229; Hanauer and Kalsbach (2023), p. 5.

4 Methodology

This section provides the theoretical framework of this study. The section starts by providing the theoretical foundation of the three machine learning models used in this study. Each model has its own subsection. After machine learning methods are introduced, the following section describes how stock return predictions obtained from different models are evaluated. The section describes both prediction accuracy and economic profitability metrics used to evaluate the performance of the models.

The variable importance section presents the approach implemented in order to evaluate comparably between models the importance of different covariates to the predictive accuracy of the models. The final part of this section introduces the sample splitting scheme applied while training the machine learning models. It also describes the hyperparameters considered as well as if they are subject to optimization.

4.1 Linear regression

The benchmark model of this study is the Fama and MacBeth (1973) regression.[44] This method follows the approach of Lewellen (2015).[45] The first step of the method is to run rolling cross-sectional regressions with lagged variables. The second step of the method calculates the means of the factor loadings obtained from the cross-sectional regressions. To obtain the expected return mean of 120 historical regression coefficients is calculated. Finally, the expected stock return can be obtained by multiplying the mean factors loading with the latest available stock characteristics. The below formulas show the generalized notation of the model

$$f(x_{i,t};\theta) = x_{i,t}\theta \tag{2}$$

$$\overline{\theta}_j = \frac{1}{T}\sum_{t=1}^{T}\theta_{j,t} \tag{3}$$

[44] See: Fama and MacBeth (1973), pp. 614-617.

[45] See: Lewellen (2015), pp. 14-19.

$$E_t\left[r_{i,t}|x_{i,t-1}\right] = x_{i,t-1}\overline{\theta}_{t-1} \tag{4}$$

In the above formulas, x indicates the firm-level characteristics and θ the loadings obtained from the regression. Symbol T indicates the rolling window considered to calculate the historical mean factor loadings.

One advantage that linear regression models typically have is that they do not require hyperparameter tuning. Therefore data does not have to be split into three sub-samples for separate validation of hyperparameters and testing. Nevertheless, the implemented linear model is not the simplest linear regression model. One variable in the implemented Fama-MacBeth model that could be treated as a hyperparameter is the rolling window. For example Lewellen (2015) reports results also for alternative rolling windows in addition to the rolling window of 120 months. Despite the possibility for hyperparameter optimization, this study uses a predefined rolling window of 120 months, which is also the rolling window in the main focus of Lewellen.[46]

Due to their high computing cost machine learning models are usually trained only once a year and then used for the rest of the year. More precisely in this study models are trained once each July for the next 12 months. Each month most recent information is just inserted into the model. Computing requirements for linear models are far lesser than for many non-linear models. Nevertheless, to ensure comparability between different models also the linear model is trained only once per year in this study. That means that no more recent stock returns than $t-1$ are used to train the model to predict stock return t, but the gap between the predicted return and the last return used to train the model can grow up to 12 months. Since we use lagged variables, this means that for the prediction of stock return t we always use stock characteristics from $t-1$, but some factors are only updated yearly. To mimic the information set an investor would have had available in historical periods we have to account for the delay in reporting balance sheet information.

[46] See: Lewellen (2015), pp. 19-23.

4.2 Random forest

Decision trees are one example of nonparametric machine learning algorithms. The idea of the decision trees is to split data into the most homogenous groups. Decision trees can be used for both classification and regression tasks. The starting point of the decision tree is called a root node. At each iteration of the decision tree algorithm finds the optimal threshold to split the data to the nodes to minimize the objective function value. Then iteratively these nodes can be further split and the tree grows. This process is repeated until a predefined tree size, set by the user, is reached or objective function cannot be improved anymore. Regression tree nodes that are not further split are called leaves. The final prediction of the regression tree leaf is the average of the dependent variable values of training set observations inside it. Gu et al. (2020) formulate prediction of a regression tree with K leaves as

$$f(x_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k 1_{\{x_{i,t}\in C_K(L)\}} \tag{5}$$

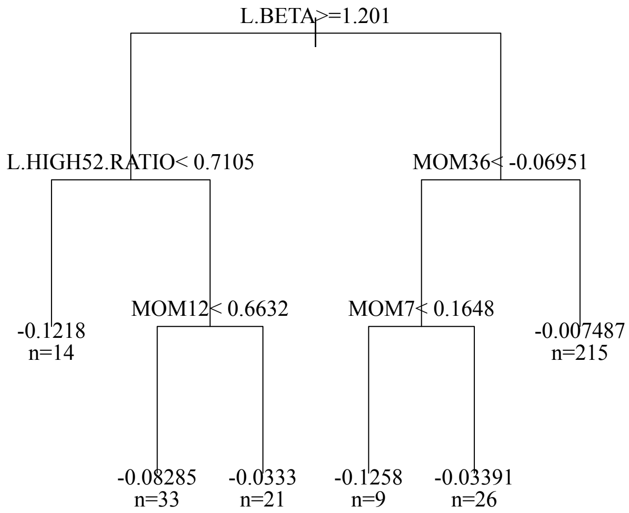
Where $C_k(L)$ represents one of the K splits the tree consists of. L is the indicator of the depth of the leaf. θ_k indicates average return within leaf k and $1_{\{x_{i,t}\in C_K(L)\}}$ indicates whether observation $x_{i,t}$ belongs to leaf k.[47] Since observation can only belong to one leaf, partition $C_K(L)$ is the product of the above partitions.

The advantage of the regression trees is that they are rather simple and intuitive, but still, they are able to model even complex interactions and non-linear relationships among the predictors. One common problem with regression trees is that they easily overfit the data and would require heavy regularization. Random forest models aim to avoid this problem by deriving the predictions from an ensemble of regression trees. As the name might suggest random forest consists of multiple decision trees.

The idea of the random forest is to randomly generate a set of decision trees and then use the average outcome of the decision trees as the final output. This way the model is less likely to overfit the data. Nevertheless, to avoid overfitting trees inside a random forest should not be too correlated and this is ensured by including randomness in the

[47] See: Gu et al. (2020), p. 2239.

Figure 2: Illustrative regression tree - Own source
Tree is trained from the actual dataset for 30th of July 2004 and then pruned to show only a few most important leaves. The figure serves only illustrative purposes and random forest models used in the study do not necessarily contain identical trees.

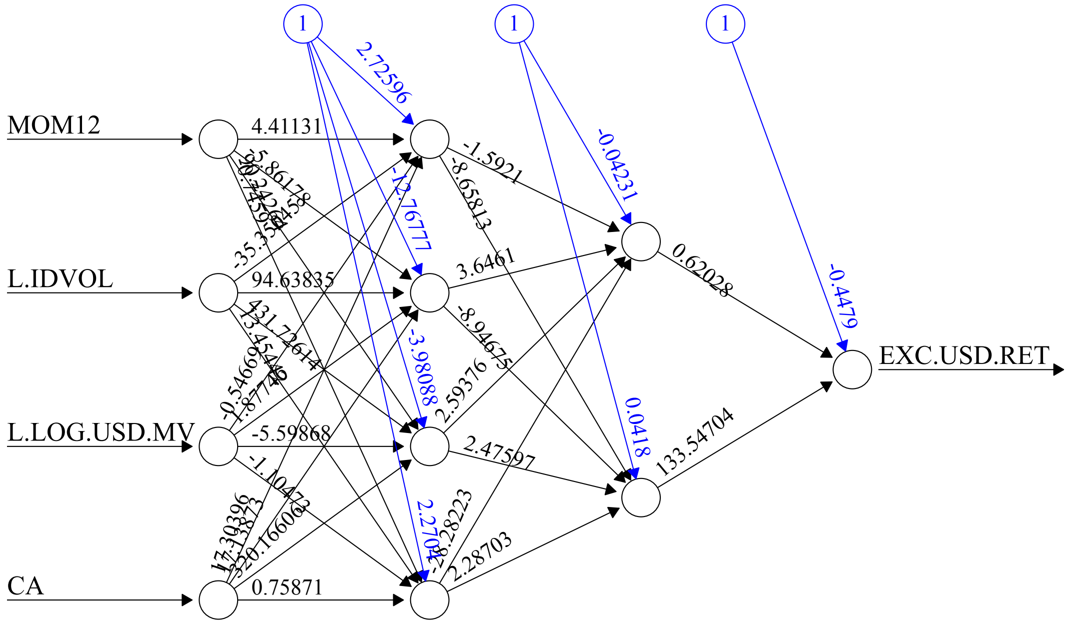


construction of the decision trees. Randomness in the generation of the decision trees is applied by restricting the set of observations used in the training of the model. The number of variables the model considers in each split as well as the maximum depth of the decision tree and the number of trees in the random forest can also be limited. Setting these parameters correctly is a crucial part of the training. These are the hyperparameters that require input from the user, but which also can be optimized for different tasks. Table 3 in Section 4.6 shows which values were considered for each hyperparameter that was optimized for random forest.

4.3 Neural networks

Artificial neural networks are a powerful machine learning method category. Currently, neural networks are a popular approach to many real-world prediction tasks. Due to their strong performance in multiple domains, neural networks are often considered state-of-the-art machine learning methods. Despite their popularity, for many users neural networks are sort of black box tools because of their complexity. Compared to linear regression and tree-based models, neural networks are far less interpretable. Another weakness of the neural networks is that they are highly parameterized and highly sensitive for the parameter initialization. Some of the parameters, such as learning rate, are usually optimized during the training of the model while others such as the architecture of the model are usually fixed.

Figure 3: Illustrative neural network - Own source
Figure serves only for illustrative purposes. The sole purpose of the figure is to visualize the structure of a neural network. Weights and biases of the neural network are obtained by training a model from the dataset of this study. Nevertheless, this model is not used in stock return predictions. Neural networks used in predictions contain more nodes in hidden layers, but a narrower architecture was chosen for better visualization.



One of the first things the user has to decide while training a neural network is the architecture of the model. This study focuses on feedforward neural networks which consist of the input layer, hidden layers and an output layer. The input layer consists of the predictive variables whereas the output layer produces the final prediction. In between there exist 1 to N hidden layers. Hidden layers again consist of so-called neurons. Similar to the number of hidden layers, the user has to also decide the number of neurons in each of the hidden layers. Number of the hidden layers is often referred as the deepness of the model whereas number of the neurons in each hidden layer is referred as the width of the model.

While a lot of previous literature simultaneously examines multiple different architectural forms, due to computing capacity in this study only one architecture will be examined.[48] The neural network architecture chosen for this study has two hidden layers. The first hidden layer has 16 neurons and following the common geometric pyramid rule second hidden layer has 8 neurons. Rather shallow and narrow architecture is chosen because they usually perform better with smaller datasets (Gu et al., 2020).[49] In order to improve and fasten the converging of the model batch normalization is implemented between all layers.

The idea of the neural network is that each neuron, using weights and biases terms, aggregates information from the previous layer and subsequently feeds the information to the activations function. The neural network model used in this study is fully connected, meaning that each neuron is connected to all neurons in the previous layer. The output of the activation function will be the input for the next layer. The neural network model is trained by optimizing these weights and bias terms. There exist many options for the activation function, which is again one choice the user has to make. The activation function used in this study is rectified linear unit

$$ReLU(x) = max(0, x) \tag{6}$$

Since the model is trained for a regression task final neuron in the output layer has a different activation function than the neurons in the hidden layers. The activation function for the output neuron is a linear function.

As mentioned neural networks include numerous hyperparameters that can be optimized during the training of the model, but training neural networks is also computationally demanding. Due to limited computing capacity, hyperparameters are not optimized for the neural network model in this study, but predefined values are used. Hyperparameters

[48] See: Gu et al. (2020), p. 2244; Hanauer and Kalsbach (2023), p. 5.; Tobek and Hronec (2021), p. 7.

[49] See: Gu et al. (2020), p. 2228

and their values are presented in Table 3. Additionally, to further limit the computational demand and simultaneously avoid overfitting early stopping algorithm is applied. Early stopping is implemented so that training of the model is terminated after five epochs where the loss function value does not reduce for the validation set. Instead of inserting the whole dataset into the model at once, data is inserted into the model in smaller subsamples so-called batches. Epoch on the other hand measures how many times the whole dataset is run through the model.

Neural networks learn by adjusting weights to the direction of the gradient. This is done in repetitive iterations. In each iteration size of the change is defined by a hyperparameter called learning rate. Since the learning rate is a hyperparameter it needs input from the user. It can also be optimized. Setting the correct learning rate is crucial since too big learning rate might prevent the algorithm from converging to the optimal solution, but too small learning rate makes converging slow. For the above-described reasons learning rate is not optimized in this study, but using a simple learning rate scheduler it is adjusted during the training of the model. In order to ensure efficient training learning rate is set to 0.001 at the beginning of the training and after ten epochs learning rate will be multiplied by 0.9 in each epoch.

Neural networks are also sensitive to the weight initialization, where the initial weights are set which the model starts to optimize. Depending on the initialization of the weights neural networks can converge to different results. To reduce model variance caused by this, an ensemble method is applied. An ensemble is implemented by training five separate models with different initial weights. The final prediction will be then the average of the predictions of the five models.

4.4 Prediction performance evaluation

This study will evaluate machine learning methods from two perspectives. Models are evaluated based on their predictive accuracy and their potential economic profitability. The profitability of the models is evaluated by backtesting portfolios that are constructed based on the stock return prediction of different models. Prediction accuracy is on the other hand evaluated by a set of statistical tests which are described in more detail below. This allows us to evaluate the relationship between predicted and realized excess returns.

The first perspective that machine learning models are evaluated is based on their prediction accuracy. Prediction accuracy will be evaluated using out-of-sample R^2_{oos} and Diebold and Mariano (1995) tests. Two out-of-sample R^2 figures are considered. Traditional out-of-sample R^2 uses historical mean return as the benchmark estimation. Traditional out-of-sample R^2 is defined as

$$R^2_{oos\ Trad.} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}(r_{i,t} - \hat{r}_{i,t})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N}(r_{i,t} - \overline{r}_{i,t})^2} \tag{7}$$

where $r_{i,t}$ is the realized return of stock i in month t, $\hat{r}_{i,t}$ is the predicted return of the same stock for month i and $\overline{r}_{i,t}$ is the historical mean return of the same stock excluding month t. Nevertheless, Gu et al. (2020) argue that the historical mean return is so noisy estimator that it underperforms compared to a static estimation of zero and therefore artificially improves the out-of-sample R^2.[50] Instead, they propose an alternative out-of-sample R^2 measure where the squared sum of returns in the denominator is not demeaned.

$$R^2_{oos} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}(r_{i,t} - \hat{r}_{i,t})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N} r_{i,t}^2} \tag{8}$$

Out-of-sample R^2 presents the prediction accuracy as a single figure, whereas Diebold and Mariano (1995) test allows for pairwise comparison of different models. Diebold-Mariano value is calculated as

[50] See: Gu et al. (2020), p. 2246.

$$\begin{aligned} d_{12,t} &= \frac{1}{N_t}\sum_{i=1}^{N}((r_{i,t} - \hat{r}_{i,t,1})^2 - (r_{i,t} - \hat{r}_{i,t,2})^2) \\ \overline{d}_{12} &= \frac{1}{T}\sum_{t=1}^{T} d_{12,t} \\ DM_{12} &= \frac{\overline{d}_{12}}{\hat{\sigma}_{d_{12}}} \end{aligned} \tag{9}$$

where $\hat{r}_{i,t,1}$ is the return prediction of first model for company i at time t, $\hat{r}_{i,t,1}$ is the return prediction of the second model for company i at time t and N_t is the number of observations in the prediction period t. Therefore, $d_{12,t}$ form a time series of differences in average squared prediction errors between model 1 and model 2. Then $\overline{d}_{12}$ is the time series mean of $d_{12,t}$ and $\hat{\sigma}_{d_{12}}$ is the Newey and West (1987) standard error of $d_{12,t}$. Diebold and Mariano (1995) test allows us to estimate the statistical significance of the prediction accuracy of two models. Under the assumption that there does not exist difference in prediction accuracy between models Diebold-Mariano statistic follows the normal distribution with a mean of 0 and standard deviation of 1, $DM \sim \mathcal{N}(0,\,1)$. The significance of the difference is reported both for traditional 5% level as well as for 3-way comparisons with Bonferroni adjustment.

In the spirit of Lewellen (2015) expected returns are also estimated by regressing realized returns with the expected returns. This regression follows

$$r_{i,t} = \alpha + \beta_1 \hat{r}_{i,t} \tag{10}$$

where $r_{i,t}$ is the realized excess return of company i at time t and $\hat{r}_{i,t}$ is the expected return of corresponding model for company i at time t.[51] For these regressions betas, t-statistics for betas and R^2 values will be reported. Ideally, the beta coefficient or the slope for the predicted return should be 1 and highly significant. The magnitude of the beta coefficients can provide information on possible over or undershooting of the models.

[51] See: Lewellen (2015), pp. 14-18.

The second perspective that is evaluated is the economic profitability of the methods. Profitability is estimated via portfolio construction following the approach of Lewellen (2015).[52] First, expected returns are derived from each model. This process is introduced in more detail in the above subsections for each model. After obtaining the expected returns, each month all stocks are distributed to ten portfolios based on the magnitude of their expected return. Allocation is univariate and does not consider any other variables than the expected return of the stock for the next month. Even though models are trained only once a year, expected returns are re-calculated every month as the most recent available data is inserted into the model. Therefore, also the portfolio allocation is repeated monthly. Each month all stocks are allocated to one of the ten expected return portfolios, but to avoid the result being mainly driven by the small stocks approach from Hanauer and Kalsbach (2023) is applied and the breakpoints for the allocation are calculated only from the large market value stocks. Large market value stocks are the biggest stocks that account for 90% of the aggregated total market value of the respective month.[53]

In addition to the ten expected return portfolios, for each method zero investment portfolio is formed. Zero investment, or long-short portfolio, is simply the spread between the return of the highest expected return portfolio and the return of the lowest expected return portfolio. Both value and equal-weighted returns will be reported for each portfolio including expected return and long-short portfolios. The performance of the machine learning portfolios is backtested and evaluated in multiple ways. For all expected return portfolios historical realized mean returns are reported together with Sharpe ratios. Sharpe ratio is defined as

$$Sharpe_i \ Ratio = \frac{\overline{r}_i}{\sigma_i} \tag{11}$$

where $\overline{r}_i$ is the time series average excess return of portfolio i and σ_i is the standard deviation of the excess returns of portfolio i. For long-short portfolios also maximum drawdown and maximum one-month loss will be reported. Maximum one-month loss

[52] See: Lewellen (2015), pp. 23-26.

[53] See: Hanauer and Kalsbach (2023), p. 6.

is simply the largest negative monthly return of each portfolio. Maximum drawdown is defined as

$$MaxDD = \max_{0 \leq t_1 \leq t_2 \leq T} (Y_{t_1} - Y_{t_2}) \tag{12}$$

where Y_t stands for cumulative return from the beginning of the period until date t. In order to examine risk-adjusted performance long-short portfolio returns will be regressed against Fama and French (2015) six-factor model factors.[54] From these regressions alphas are reported, which can be interpreted as the excess return that the models are able to generate that cannot be explained by the loadings in the six risk factors. Also, t-statistics for the alphas and R^2 values are reported. Regression formula for risk-adjusted performance is

$$\begin{gathered} r_{t,m} = \alpha + \beta_1 \; RMRF_t + \beta_2 + \; SMB_t + \beta_3 \; HML_t + \\ \beta_4 \; CMA_t + \beta_5 \; RMW_t + \beta_6 \; MOM_t + \epsilon_t \end{gathered} \tag{13}$$

where $r_{t,m}$ is the realized return of long-short portfolio from model m in time t, $RMRF$ is the excess market return, SMB is the spread in the return between small market value stocks and large market value stocks, HML is the spread in the return between high book-to-market value stocks and low book-to-market value stocks, CMA is the spread in the return between conservatively investing stocks and aggressively investing stocks, RMW is the spread in the return between stocks with robust profitability and stocks with weak profitability and MOM is the spread between return of stocks that had the highest return in period $t-1$ and the stocks that had the lowest return in period $t-1$. [55] Factors are constructed from the same dataset as machine learning portfolios, except that the microcap stocks are not excluded. Construction of these factors is described in more detail in Section 5.1.

[54] See: Fama and French (2015), .pp 2-3. Fama and French introduced the five factor model. Factors used to regress machine learning portfolio returns include five-factor model factors and momentum factor from Carhart (1997).

[55] See: Fama and French (2015), pp. 2-3; Carhart (1997), p. 61.

Finally, turnovers for the long-short portfolios are reported. In a real-world setting, investors usually are subject to some sort of trading cost. Even if the model generates excess returns but the monthly turnover remains large, it could lead to a situation where after counting for transaction costs more passive strategy would become more profitable. Therefore, turnover provides valuable information on the practical implementability of the constructed machine learning models. For month t turnover is defined as

$$Turnover_t = \sum_{i=1}^{N_t} \left(\left| w_{i,t} - \frac{w_{i,t-1}(1+r_{i,t})}{\sum_{j=1}^{N_t} w_{j,t-1}(1+r_{j,t})} \right| \right) \tag{14}$$

where N_t is number of companies in portfolio j in month t and $w_{i,t}$ is the weight of the company i in portfolio j after the reallocation. The latter part of the equation indicates the weight of the company i right before the reallocation. It considers the change in weight of company i due to the relative return in month t compared to the return of the corresponding portfolio.

4.5 Variable importance

One challenge in dealing with various statistical methods is that they lack common metrics for explanatory inference. Many of the models have metrics for variable importance, but the comparability of these metrics can be questioned. Therefore, the approach of Gu et al. (2020) is implemented to define variable importance metrics for the model applied in this study. The approach consists of the following steps. First, one variable at a time is set to zero. Then the reduced model is retrained and new predicted returns are derived using the reduced model. The process of training and predicting returns is identical to the reduced model as for the full model. After obtaining the predicted returns from the reduced model, out-of-sample R^2 values are calculated for these returns. Then the change compared to out-of-sample R^2 of the full model is calculated. Finally, to obtain the relative variable importance metric, the sum of changes in out-of-sample R^2s is normalized to one within a model. The same process is applied to each variable and all models.[56]

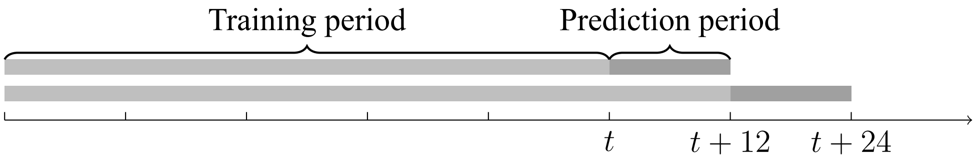
[56] See: Gu et al. (2020), p. 2247.

4.6 Sample splitting

It is common while training machine learning models to split the data into three sets. The training set will be used, as the name suggests, to train the model. In case a machine learning model includes hyperparameters these can be optimized with a validation set. Finally, the true out-of-sample predictions can be performed for testing data. Because we want to mimic the situation and information set of an investor we have to take into consideration the time series nature of the data.

In stock return prediction literature it is common to split the data as described above, but considering the chronological order. For example Fieberg et al. (2023) use a rolling ten-year scheme where they first train the model using the first seven years of the data and then optimize hyperparameters using the last three years of the rolling window.[57] Finally, they train the model with optimal hyperparameter initialization with a whole ten-year window to predict returns for the next year. Gu et al. (2020) use a slightly different approach. Instead of using a rolling window, they increase the training window size after each training period by one year.[58] Common for these two approaches is that they both train the model only once a year.

Figure 4: Sample splitting - Own source
Illustration sample splitting used in training of the machine learning models. Machine learning models are trained once a year and after each training model is used for the next 12 months to predict the stock returns. Each year training period is extended by 12 months. The training period is further split into training data and validation data randomly allocating 80% of the data to the training set and 20% of the data to the validation set. A validation set is used to optimize hyperparameters. The minimum length of the training period is 50 months.



The sample splitting scheme applied in this study is slightly different from above described ones. The above-described approaches use disjoint time periods to mimic the out-of-sample setting in the hyperparameter optimization. In this study training and val-

[57] See: Fieberg et al. (2023), pp. 302-303.
[58] See: Gu et al. (2020), p. 2232.

idation sets are separated from the testing data based on time. The approach is closer to the approach of Gu et al. (2020) in the sense that the training data window is increased each year.[59] Nevertheless, the difference is that the data is distributed to training data and validation randomly instead of using the disjoint periods. The reason why this scheme is chosen is that our dataset is quite small. In order to have reasonable amount of data for the validation set we would need to include multiple years to the validation set, which would be away from the initial training set and could lead to suboptimal hyperparameter selection. The sample splitting scheme is illustrated in Figure 4.

Table 3: Hyperparameters - Own source
Table presents the hyperparameters that are either optimized or taken as fixed values. In case predefined values are used only one figure is indicated in the table. If hyperparameter is optimized set or list is displayed. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model.

	FM	RF	NN
Hyperparameter	Rolling window = 120	ntree $= 300$ mtry $= \in (2, 3, 5, 7)$ max.depth $= 2 \sim 6$ sample.fraction $= 0.5$	Learning rate $= 0.001$ Batch size $= 502$ Epochs $= 100$ Patience $= 5$ Ensemble $= 5$

Since linear regression does not require any hyperparameter optimization there is no need for validation set and all data can be used to train the model. For the random forest model, we actually optimize the hyperparameters. Therefore, the training window is split into training and validation data so that 80% of the data is used in the training and 20% is assigned to the validation set. For the neural network model, we do not optimize any actual hyperparameters, but we still need a validation set for the early stopping algorithm. Therefore, for also neural network 20% of the training window is assigned to the validation set. The approach of this study follows the common approach to only train the models once a year in July.

5 Benchmark factors

This section serves as a prerequisite for machine learning portfolios. In this section, well-established factors from Fama and French (2015) framework are constructed. These

[59] See: Gu et al. (2020), p. 2232.

factors include market, size, value, investment, profitability and momentum factor.[60] The first part describes the construction of the factors and the second part discusses the performance of the factors in Nordic stock markets. Later in the study, these factors are used to evaluate the risk-adjusted performance of the different machine learning models. Factors also provide interesting insight into the existence of different traditional stock market factor anomalies in Nordic markets in a bit more traditional setting.

Factors can also be used as a benchmark for the profitability of the machine learning models. Compared to machine learning portfolios, the construction of the traditional stock market factors is far more simple. They do not require such intense computing as training machine learning models. The amount of data required for the construction of traditional stock market factors is also rather limited compared to the set of predictors included in this study. In order to justify the effort required, the machine learning models should be able to exceed the performance of the traditional factors. These factors are also used later to evaluate the risk-adjusted performance of the machine learning models.

5.1 Benchmark factor construction

Benchmark factor construction follows 2×3 portfolio sort approach of Fama and French (1993, 2015) and Carhart (1997). Fama and French (1993) use NYSE breakpoints for size and book-to-market value sorts.[61] Since compared to US markets Nordic markets have fewer companies with high market value, using NYSE breakpoints could lead to highly undiversified portfolios, especially among the high market value portfolios. On the other hand, breakpoints should not be driven by the small stocks that are numerous, but only account for a small part of the total market capitalization. Therefore the approach of Fama and French (2012) is applied.[62]

At the end of each June, stocks are first distributed to two size portfolios. Companies with the biggest market value that accounts for 90% of the total market value are classified as big stocks. All the rest of the stocks are considered to be small stocks. Therefore

[60] See: Fama and French (2015), pp. 2-3; Carhart (1997), p. 61.

[61] See: Fama and French (1993), p. 8; Fama and French (2015), pp. 2-3; Carhart (1997), p. 61.

[62] See: Fama and French (2012), p. 459.

dataset used to construct benchmark factors is slightly different than the dataset used to train the machine learning models as it includes also the microcap stocks. Next stocks are allocated to three value, investment, profitability and momentum portfolios independently of the size allocation. For all of the above variables 30th and 70th percentiles are used to calculate breakpoints. Breakpoints are calculated using only big companies from the size allocation, but the breakpoints are used to allocate all stocks to a portfolio. Factor construction can be formulated as

$$\begin{aligned}
SMB_{B/M} &= \frac{1}{3}(Small.High + Small.Neutral + Small.Low) \\
&- \frac{1}{3}(Big.High + Big.Neutral + Big.Low) \\
SMB_{OP} &= \frac{1}{3}(Small.Robust + Small.Neutral_{OP} + Small.Weak) \\
&- \frac{1}{3}(Big.Robust + Big.Neutral_{OP} + Big.Weak) \\
SMB_{INV} &= \frac{1}{3}(Small.Conservative + Small.Neutral_{INV} + Small.Aggressive) \\
&- \frac{1}{3}(Big.Conservative + Big.Neutral_{INV} + Big.Aggressive) \\
SMB_{MOM} &= \frac{1}{3}((Small.Winner + Small.Neutral_{MOM} + Small.Loser) \\
&- \frac{1}{3}(Big.Winner + Big.Neutral_{MOM} + Big.Loser) \\
SMB &= \frac{1}{4}(SMB_{B/M} + SMB_{OP} + SMB_{INV} + SMB_{MOM}) \\
\\
HML &= \frac{1}{2}(Small.High + Big.High) - \frac{1}{2}(Small.Low + Big.Low) \\
RMW &= \frac{1}{2}(Small.Robust + Big.Robust) - \frac{1}{2}(Small.Weak + Big.Weak) \\
CMA &= \frac{1}{2}(Small.Conservative + Big.Conservative) \\
&- \frac{1}{2}(Small.Aggressive + Big.Aggressive) \\
MOM &= \frac{1}{2}(Small.Winner + Big.Winner) \\
&- \frac{1}{2}(Small.Loser + Big.Loser)
\end{aligned} \tag{15}$$

Book-to-market value is used as an indicator of the value characteristic of a company. Book-to-market value is calculated as the ratio between the sum of common equity and deferred taxes and market capitalization on December $t-1$. Profitability is defined as net income before extraordinary items divided by the book equity of the company.

The investment variable is calculated as an annual change in total assets. Momentum is defined as a cumulated return from $t-12$ to $t-2$. Returns are calculated using the total return index that is converted to US dollars for comparability between different countries. The market value used in size allocation as well as to weight portfolio returns is also converted to US dollars.

Equation 15 shows the formula for each factor. The abbreviation for each variable is derived from how they are calculated. The value factor is called high minus low (HML), profitability factor is called robust minus weak (RMW), investment factor is called conservative minus aggressive (CMA). Only the momentum factor is an exception to this rule and more intuitive naming is used. Portfolio allocation results in six portfolios for value, investment, profitability and momentum factors and 24 two-fold size portfolios. After portfolio construction portfolio returns are calculated as differences in value-weighted average returns on portfolios formed based on respective variables. E.g. value factor return is the difference between the average of value-weighted returns of two high book-to-market portfolios and an average of value-weighted returns of two low book-to-market portfolios. The market factor is the average value-weighted excess return of the whole market. The risk-free rate is obtained from Kenneth French's website.

5.2 Benchmark factor performance

Before jumping to the machine learning portfolios this section shows the historical performance of the Fama and French (2015) five-factor model factors augmented by the momentum factor in Nordic stock markets.[63] As mentioned, later these factors are used to evaluate the risk-adjusted performance of the machine learning portfolios, but prior to that, it is interesting to observe whether simpler factor construction shows profitability in Nordic markets.

Table 4 provides the time series averages of the factor returns, standard deviation of the factor returns, corresponding t-statistics and p-values as well as monthly minimum and maximum returns for all six factors. From Table 4 it is clear that the momentum factor

[63] See: Fama and French (2015), pp. 2-3; Carhart (1997), p. 61.

Table 4: Benchmark factor summary statistics - Own source
Table presents the mean returns and standard errors of the benchmark factors together with t-statistics and corresponding p-values. For each factor minimum and maximum monthly returns are reported. RMRF is the average value-weighted excess return of the pooled Nordic market. Portfolio returns are calculated based on 2×3 sorts on size and one other factor. HML is the difference in the average value-weighted return of two high value portfolios and the average value-weighted return of two low value portfolios. RMW, CMA and MOM are calculated in a similar manner, but portfolio sorts are done based on investment, profitability and momentum factors. SMB is the average of the value-weighted returns of the 12 portfolios of small stocks minus the average of the value-weighted returns of the 12 portfolios of big stocks. Returns are calculated in US dollars. The risk-free rate used to calculate excess returns is the US dollar one-month Treasury bill rate. The time period spans from January 1990 to December 2022.

	Mean	SE	t-stat.	p-value	Min	Max
HML	0.0016	0.0022	0.7368	0.4617	-0.2517	0.2324
RMW	0.0016	0.0015	1.0298	0.3038	-0.1223	0.1607
CMA	0.0014	0.0016	0.8725	0.3835	-0.1605	0.1828
MOM	0.0095	0.0022	4.3894	0.0000	-0.1705	0.1867
SMB	-0.0003	0.0014	-0.2470	0.8051	-0.1185	0.1029
RMRF	0.0073	0.0032	2.3006	0.0219	-0.2574	0.2070

shows the strongest performance measured both by the magnitude of the return as well as the statistical significance of the return. The monthly momentum factor return is 0.9% with t-statistic of 4.3. Table B.1 in the appendix shows the correlations between the factor returns. In Nordic markets, the correlation of the momentum factor with other factors is only minor. Interestingly in Nordic markets momentum factor seems to negatively correlate with the market factor.

The strong performance of the momentum factor is in line with previous literature. Many previous studies have documented excess momentum returns either in pooled or individual Nordic markets.[64] Slightly more surprising is the poor performance of the value factor. The average return of the value factor is 0.14% and it is not statistically significant. Some of the previous studies document value premiums in Nordic markets. Grobys and Huhta-Halkola (2019) find statistically significant value premium in Nordic markets, but Grobys and Huhta-Halkola construct equal-weighted portfolios whereas benchmark factors reported here are formed from value-weighted portfolios.[65]

[64] See: e.g. Grobys and Huhta-Halkola (2019), pp. 10-11; Leivo and Pätäri (2011), pp. 407-411.

[65] See: Grobys and Huhta-Halkola (2019), p. 9.

Figure 5: Benchmark factor performance

Plot presents the cumulative return of the benchmark factors. RMRF is the average value-weighted excess return of the pooled Nordic market. Portfolio returns are calculated based on 2 × 3 sorts on size and one other factor. HML is the difference in the average value-weighted return of two high value portfolios and the average value-weighted return of two low value portfolios. RMW, CMA and MOM are calculated in a similar manner, but portfolio sorts are done based on investment, profitability and momentum factors. SMB is the average of the value-weighted returns of the 12 portfolios of small stocks minus the average of the value-weighted returns of the 12 portfolios of big stocks. Returns are calculated in US dollars. The risk-free rate used to calculate excess returns is the US dollar one-month Treasury bill rate.

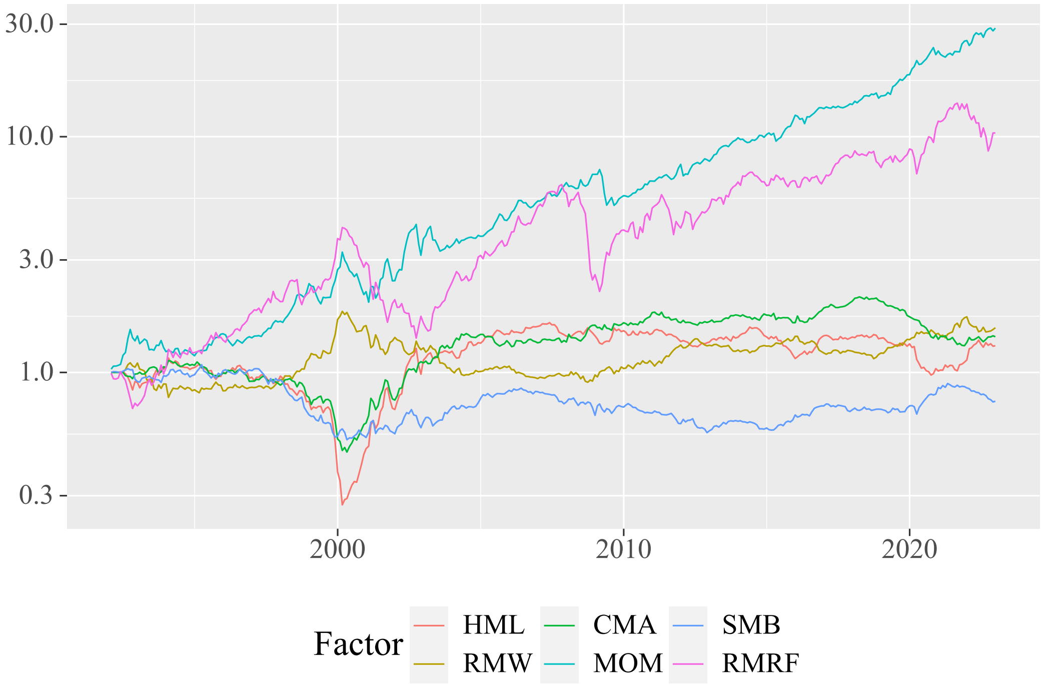


Figure 5 shows the historical performance of the benchmark factors. It shows that until 2008 even the momentum factor could not exceed the market return, but still seemed to be slightly less volatile. In the post-financial crisis period, the momentum factor has been clearly the strongest performing factor. Interestingly late nineties, which was a strong period for market return, was an extremely difficult period for value factor. Contrary in the early 2000's value factor performed strongly when the market factor was decreasing steeply. Until 2003 performance of the value factor was almost opposite to the market factor, but after that value factor has not been able to generate significant value.

Figure 5 also shows that the performance of the size, investment and profitability factors has been poor throughout the examined period. Return of the size factor is even negative indicating reverse size premium. This would indicate that on average large market capitalization stocks have performed better than small market capitalization stocks. Nevertheless, the negative return of the size factor is small and not statistically significant. Overall the performance of the benchmark factor indicates that momentum indicators could hold profitable information about Nordic stock market returns.

6 Empirical results on Nordic equities

The performance of the machine learning models will be evaluated from two aspects. Prediction accuracy is evaluated by the out-out sample R^2 values and Diebold and Mariano (1995) tests. Additionally, in the spirit of Lewellen (2015) relationship between expected and realized excess returns is examined by regressing the realized excess returns with the individual stock level predicted returns.[66] Then the economic profitability of the models will be evaluated by investigating the performance of univariate expected return sort portfolios. The construction of these portfolios is explained in more detail in Section 4.4. Finally, the variable importance for different methods is calculated to see the effect of each explanatory variable on the prediction accuracy of the model. The process to define variable importance is described in Section 4.5.

6.1 Prediction accuracy

First interesting remark from the results is that at least for the random forest model more complex model specification tend to perform better. Figure C.10 in the appendix shows the optimal hyperparameters for the random forest model for all retraining periods. For most of the periods the model tends to favour the most complex setting both in regards of number of variables considered in each split as well as maximum depth of the trees. Even though model tends to converge to the highest allowed hyperparameter values, these limits of the hyperparameters are not increased to avoid too high correlation and overfitting of the regression trees inside the random forest model.

[66] See: Lewellen (2015), pp. 14-18.

Panel A of Table 5 presents the out-of-sample R^2 values for all models whereas panel B presents the pairwise Diebold-Mariano statistics. It can clearly be seen from Table 5 that the random forest model produces the most accurate out-of-sample predictions. Out of the three models, it is the only one that produces a positive out-of-sample R^2 value of 0.3% even with the more conservative definition where the benchmark model is a prediction of zero. While linear and neural network models both produce negative out-of-sample R^2 values of -1.95% and -0.27%, it can still be seen that the neural network model performs better than the linear model in regards to predictive accuracy.

Another interesting insight Table 5 provides is the relationship between traditional out-of-sample R^2 and the conservative out-of-sample R^2. It confirms the hypothesis of Gu et al. (2020) about traditional out-of-sample R^2 metric being too loose and showing unrealistically strong results.[67] The traditional out-of-sample R^2 metric is positive for all three models, which means that compared to the more strict out-of-sample R^2 metric, the sign of the metric changes for linear and neural network models. Nevertheless, the order between models does not change based on the definition of the out-of-sample R^2 metric. In this regard, results are in line with findings of Fieberg et al. (2023).[68]

Figure C.5 in the appendix shows the out-of-sample R^2 values as a time series of the prediction periods. The figure shows that the same overall trends can be seen from all of the methods. It also reveals that in the prediction period starting from July 2011 neural network model produced extremely bad out-of-sample predictions. This period is difficult for other methods as well, but not on the same scale as for the neural network model. The conservative out-of-sample R^2 value for the neural network model reaches -10% during this period.

Inspecting the time series of the out-of-sample R^2 values produced by the two definitions further supports the argumentation that the traditional out-of-sample R^2 definition is too optimistic metric to evaluate the goodness of the stock return prediction model. Figure C.5 shows how the traditional out-of-sample R^2 values are not only sifted upwards, but also the variation of the out-of-sample R^2 vales is smaller. This again supports the argument of Gu et al. (2020) that the historical mean as an estimator

[67] See: Gu et al. (2020), p. 2246.

[68] See: Fieberg et al. (2023), p. 303.

Table 5: Prediction accuracy - Own source
Table presents the prediction accuracy metrics for different machine learning models. Panel A presents two out-of-sample R^2 values. The first one uses zero prediction as a benchmark model. This means that the denominator in the calculation of the metric is the squared excess return. The second out-of-sample R^2 figure follows the traditional definition and the realized excess return is demeaned by the historical mean return. Panel B of the table presents the pairwise Diebold-Mariano statistics for all the methods. Positive number indicates that the out-of-sample prediction accuracy of the model indicated in the columns is better than the prediction accuracy of the model indicated in the rows. The bolded figure indicated significance at 5% level, whereas the asterisk indicates significance at 5% level after three-way Bonferroni adjustment. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model.

Panel A: Out-of-sample R^2

	FM	RF	NN
R^2_{oos}	-0.0195	0.0032	-0.0027
$R^2_{oos\ Trad.}$	0.0299	0.0502	0.0447

Panel B: Diebold-Mariano statistics

	FM	RF	NN
FM		**7.8155*** (0.0000)	**4.2782*** (0.0001)
RF			**-12.1227*** (0.0000)

of future stock return contains so much noise that it actually artificially improves the out-of-sample R^2 values.[69]

As mentioned, the overall trends in the development of out-of-sample R^2 can be seen for all models as shown in Figure C.5 in the appendix. At the beginning of the prediction periods models are able to produce rather positive out-of-sample R^2 values, but coming to the 2000s models struggle more. Then until the period of financial crisis predictive performance of all models improves, ultimately leading to more than five percent out-of-sample R^2 values for all models. Then in post post-financial crisis period predictive performance of the models is quite volatile for all models, fluctuating on both sides of zero. As these trends can be seen for all three methodologically different models it could be argued that the underlying predictive power of the predictors included in this study is varying over time.

In this study, machine learning models predict the stock returns between July 1994 and

[69] See: Gu et al. (2020), p. 2246.

November 2022. Meaning that the period of financial crisis in 2008 is exactly in the middle of the prediction period. Interestingly in the first half of the prediction window neural network model shows the strong predictive accuracy with conservative out-of-sample R^2 value of 0.7%. During that time the out-of-sample R^2 of the random forest model is 0.8% whereas the predictive accuracy of the linear model remains negative with out-of-sample R^2 of -3.7%. The predictive accuracy of the linear model is even worse in the first half of the prediction window than in the whole window. This also shows that negative out-of-sample R^2 value of the neural network model is driven by the poor predictive accuracy of the model in the second half of the prediction window.

Results of the predictive performance of different models measured by the conservative out-of-sample R^2 values are partially in line with previous literature. Results are in line with findings of Drobetz and Otto (2021) study in European stock markets in a sense that the linear model offers the worst out-of-sample predictive power when measured by the out-of-sample R^2 introduced in Equation 8. They also find negative out-of-sample R^2 value for the linear model. Results are also in line in the sense that the random forest model shows strong predictive performance.[70]

Results of Fieberg et al. (2023) are slightly more contradictory since they show positive out-of-sample R^2 value also for linear model.[71] Both studies of Drobetz and Otto and Fieberg et al. are conducted in European stock markets, which partially overlap with markets of this study, but the difference is that Drobetz and Otto use twenty-two characteristics as well as their second- and third-order polynomials and two-way interactions whereas Fieberg et al. only use six characteristics. Variable selection of this study is in between these two since we include more variables than Fieberg et al., but we do not include second- and third-order polynomials or two-way interactions like Drobetz and Otto.[72]

Where the results clearly deviate from previous literature is the predictive performance

[70] See: Drobetz and Otto (2021), p. 521-522. Drobetz and Otto Report negative out-of-sample R^2 for linear model using all predictors.

[71] See: Fieberg et al. (2023), p. 307. Fieberg et al. report the results for multiple subsets where companies are filtered based on their market capitalization. The linear model produces negative out-of-sample R^2 values when only the biggest 20% of the stocks are included, but this is not the setting of this study.

[72] See: Fieberg et al. (2023), pp. 305-306; Drobetz and Otto (2021), p. 510.

of the neural network model. In studies of Drobetz and Otto (2021) and Fieberg et al. (2023) neural network model is among the most accurate models producing clearly positive, out-of-sample R^2 values.[73] Naturally studies of Drobetz and Otto and Fieberg et al. are not directly comparable to this study since the variable set differs and the datasets of Drobetz and Otto and Fieberg et al. are much wider since they include more countries. The size of the dataset could at least partially explain the relatively poor performance of the neural network model, since usually neural network models require a lot of data.

Panel B of Table 5 presents the pairwise Diebold and Mariano (1995) test statistics for all the models. Calculation of the statistics is described in Section 4.4. The Diebold-Mariano statistics are reported together with corresponding p-values. Bolding of the Diebold-Mariano figure implies significance in a normal 5% confidence level whereas the asterisk implies a more conservative 5% confidence level which is Bonferroni adjusted for three-way comparisons. The three-way Bonferroni adjusted critical one-sided Diebold-Mariano value is 2.13. The Diebold-Mariano statistics support the results from the panel A of Table 5. Both neural network and random forest models produce more accurate predictions than the linear regression model as they have positive significant Diebold-Mariano statistic values. Furthermore, predictions of the random forest model are more precise than the predictions of the neural network model. All the differences in the prediction accuracies are statistically significant even in more conservative Bonferroni adjusted 5% significance level.

Next, the prediction accuracy is examined by following the approach of Lewellen (2015) by regressing the realized excess returns by the return predictions from different models.[74] Table 6 presents the summary statistics for these regressions. The left side of the table presents the univariate properties of expected returns and the right side of the table presents the regression statistics. Comparing univariate properties of the expected returns from Table 6 to descriptive statistics in Table 2 shows that the mean expected return is really close to the actual realized mean excess return for neural network and random forest model. Both mean expected and realized return are calculated as time

[73] See: Fieberg et al. (2023), pp. 307-308; Drobetz and Otto (2021), p. 522.

[74] See: Lewellen (2015), pp.14-18.

series averages of cross-sectional means. Linear model on the other hand seems to predict larger returns on average than what is actually realized. Another remark from Table 6 is that the standard deviation for the return predictions produced by the linear model and neural network model is higher than the standard deviation of the realized excess returns. This indicates that the variation in expected returns from these models is larger than the variation in the realized excess returns.

Table 6: Expected return regression summaries - Own source
Table provides univariate properties of the return predictions for all models and summary statistics for regressions where realized excess returns are regressed with expected returns. Mean and standard deviation are reported for expected returns. The mean value reported is the time series average of the cross-sectional means and standard deviation is the time series average of cross-sectional standard deviations. The right side of the table reports the regression coefficients, standard errors of the coefficients, corresponding t-statistics and the R^2 values. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model. The prediction period spans from July 1994 to November 2022.

	Univariate properties		Predictive ability			
	Mean	Std.	Slope.	SE	t-stat	R^2
FM	0.0127	0.0123	0.1454	0.0221	6.5842	0.0004
RF	0.0076	0.0093	0.4658	0.0302	15.4355	0.0020
NN	0.0066	0.0130	0.3254	0.0212	15.3200	0.0019

Results from the right side of Table 6 support the remarks from out-of-sample predictive performance and univariate properties. For all of the models, there exists a statistically highly significant positive relationship between expected returns and realized returns. If the expected returns would reflect realized excess returns perfectly regression slope shown in Table 6 should be one. The random forest model has the highest predictive slope of 0.47, which means that a 1% change in expected return responds to a 0.47% change in the realized return. Neural network and linear models have slightly smaller slopes of 0.32 and 0.15 respectively. R^2 values from the regressions are also presented as a third alternative out-of-sample prediction accuracy metric in addition to two previously introduced out-of-sample R^2 metrics. The third R^2 metric further confirms the message of the first two as the neural network and random forest perform better than the linear model. With the regression-based R^2 the relative order of the models remains the the same as for previous two out-of-sample R^2 metrics.

Given that the mean predicted return matches quite well the mean realized excess return, but the standard deviation is higher and the slope is slower, it seems that models seem to overshoot in their predictions. Especially, It seems that the neural network and random forest models are able to predict the returns correctly on average, but exaggerate the extreme returns. This could at least partially explain rather low out-of-sample R^2 values discovered in Table 5. Further insight into this will be provided in the next section where the performance of expected return sorted portfolios is examined.

6.2 Portfolio performance

This section focuses on backtesting the machine learning portfolios, which are formed based on the expected returns produced by the different models. The approach attempts to mimic of information set of a historical investor and the section describes the historical realized returns for investment strategies built based on the machine learning models. The first part of the section mainly focuses on evaluating the performance of the expected return portfolios whereas the second part discusses the performance of the long-short portfolios in more detail. Formation of expected return portfolios is described in Section 4. Results are reported separately for value and equal-weighted portfolios.

Table 7 presents the performance statistics for all ten expected return portfolios for all three models. Average predicted return, average realized excess return, standard deviation of the realized excess return, corresponding t-statistic and Sharpe ratio are reported for each portfolio. The left side of the table presents the values for equal-weighted portfolios, whereas the right side of the table presents the values for the value-weighted portfolios. Panel A of the table shows the results for the linear regression model, panel B shows the results for the random forest model and panel C shows the results for the neural network model. Numbers on the first column of the table indicate the expected return decile of the corresponding portfolio. H-L is a portfolio formed from a short position on the lowest expected return portfolio and a long position on the highest expected return portfolio.

Table 7: Machine learning portfolio performance - Own source
Table reports performance metrics for portfolios formed based on univariate expected return sort. Each month all stocks are allocated to ten portfolios based on their expected returns. Breakpoints for the allocation are calculated only from big stocks, which are the biggest stocks that in the current month account for 90% of the cumulative market value of all stocks in the dataset. H-L is the zero investment portfolio which consists of a short position in the portfolio formed from stocks with the lowest expected return and a long position in the portfolio formed from stocks with the highest expected return. The time series average of predicted return and realized excess return of each portfolio is reported for each model together with the standard deviation of realized excess return. Additionally, Sharpe ratios are reported. The left side of the table reports the results for equally weighted portfolios and the right side reports results for portfolios where each stock in the portfolio is weighted by its market value. The prediction period spans from July 1994 to November 2022.

Panel A: Linear regression

	Equal-weighted					Value-weighted				
	Pred.	Avg.	Std.	t-stat	SR	Pred.	Avg.	Std.	t-stat	SR
Low	-0.0024	0.0046	0.0788	1.0743	0.0583	-0.0012	0.0082	0.0797	1.9058	0.1034
2	0.0051	0.0039	0.0680	1.0596	0.0575	0.0051	0.0056	0.0680	1.5061	0.0817
3	0.0081	0.0060	0.0634	1.7332	0.0940	0.0081	0.0068	0.0629	2.0032	0.1086
4	0.0102	0.0069	0.0626	2.0253	0.1098	0.0102	0.0064	0.0634	1.8505	0.1004
5	0.0119	0.0100	0.0621	2.9585	0.1604	0.0119	0.0091	0.0665	2.5295	0.1372
6	0.0136	0.0093	0.0587	2.9221	0.1585	0.0136	0.0078	0.0622	2.3015	0.1248
7	0.0153	0.0087	0.0613	2.6052	0.1413	0.0153	0.0095	0.0642	2.7171	0.1474
8	0.0174	0.0094	0.0622	2.7761	0.1506	0.0174	0.0084	0.0644	2.4068	0.1305
9	0.0206	0.0135	0.0637	3.9000	0.2115	0.0207	0.0091	0.0634	2.6441	0.1434
High	0.0330	0.0170	0.0676	4.6416	0.2517	0.0348	0.0127	0.0683	3.4361	0.1863
H-L	0.0354	0.0124	0.0512	4.4708	0.2425	0.0360	0.0045	0.0576	1.4328	0.0777

Panel B: Random forest

	Equal-weighted					Value-weighted				
	Pred.	Avg.	Std.	t-stat	SR	Pred.	Avg.	Std.	t-stat	SR
Low	-0.0059	-0.0005	0.0752	-0.1187	-0.0064	-0.0046	0.0025	0.0768	0.6022	0.0327
2	0.0012	0.0061	0.0639	1.7496	0.0949	0.0012	0.0047	0.0679	1.2870	0.0679
3	0.0038	0.0069	0.0644	1.9717	0.1069	0.0038	0.0087	0.0660	2.4240	0.1315
4	0.0060	0.0092	0.0613	2.7523	0.1493	0.0060	0.0081	0.0657	2.2745	0.1234
5	0.0080	0.0084	0.0609	2.5298	0.1372	0.0080	0.0091	0.0611	2.7327	0.1482
6	0.0099	0.0096	0.0615	2.8674	0.1555	0.0099	0.0072	0.0641	2.0657	0.1120
7	0.0116	0.0106	0.0630	3.1103	0.1687	0.0117	0.0085	0.0676	2.3144	0.1255
8	0.0134	0.0120	0.0627	3.5243	0.1911	0.0134	0.0091	0.0650	2.5913	0.1405
9	0.0154	0.0139	0.0613	4.1967	0.2276	0.0154	0.0120	0.0640	3.4573	0.1875
High	0.0220	0.0166	0.0690	4.4457	0.2411	0.0208	0.0134	0.0750	3.2818	0.1780
H-L	0.0279	0.0171	0.0406	7.7722	0.4215	0.0254	0.0108	0.0528	3.7846	0.2053

Panel C: Neural network

	Equal-weighted					Value-weighted				
	Pred.	Avg.	Std.	t-stat	SR	Pred.	Avg.	Std.	t-stat	SR
Low	-0.0152	0.0022	0.0815	0.4890	0.0265	-0.0131	0.0066	0.0862	1.4161	0.0768

Table continues in the next page …

… Continued from the previous page

2	-0.0022	0.0040	0.0643	1.1578	0.0628	-0.0021	0.0050	0.0693	1.3212	0.0717
3	0.0016	0.0060	0.0619	1.7802	0.0965	0.0016	0.0048	0.0640	1.3709	0.0743
4	0.0042	0.0097	0.0614	2.9211	0.1584	0.0042	0.0096	0.0643	2.7675	0.1501
5	0.0063	0.0090	0.0605	2.7476	0.1490	0.0063	0.0105	0.0643	3.0086	0.1632
6	0.0083	0.0092	0.0591	2.8571	0.1549	0.0083	0.0092	0.0592	2.8522	0.1547
7	0.0103	0.0103	0.0598	3.1642	0.1716	0.0103	0.0088	0.0648	2.5037	0.1358
8	0.0127	0.0111	0.0611	3.3522	0.1818	0.0127	0.0088	0.0629	2.5904	0.1405
9	0.0158	0.0114	0.0629	3.3426	0.1813	0.0157	0.0094	0.0669	2.5998	0.1410
High	0.0248	0.0141	0.0680	3.8284	0.2076	0.0245	0.0122	0.0719	3.1191	0.1692
H-L	0.0400	0.0119	0.0438	5.0331	0.2730	0.0376	0.0055	0.0543	1.8818	0.1021

Looking at the equal-weighted part of Table 7 provides quite a clear message. Even though out-of-sample R^2 remained rather low, especially for the linear and neural network model, the examined variable set seems to contain information about the cross-section of future stock returns. For all models, there is a clear rising trend of realized excess returns across expected return portfolios. The linear model has a couple of outliers where the return of a lower expected return portfolio actually exceeds the return of a higher expected return portfolio. Random forest and neural network models both only have one such an outlier. The spread of average realized excess return between the minimum expected return portfolio and maximum expected return portfolio is more than one percent for all models.

Models struggle more with value-weighted portfolios. The increasing trend among value-weighted is not as smooth and more outliers exist than among equal-weighted portfolios. Nevertheless, for all models on average five smallest expected return portfolios generate lower realized returns than the five largest expected return portfolios. For all models portfolio with the highest expected return also has the highest realized excess return. Simultaneously, for all models, the return of the four highest expected return portfolios is above average market return with a clear premium.

It is quite expected that the predictability of the value-weighted portfolios is lower than the equal-weighted portfolios. One reason for this is that the stocks are divided into ten expected return portfolios. From Table 1 it can be seen that on an average month dataset contains 343 stocks. This would mean that even if stocks were allocated to portfolios evenly, each portfolio would on average contain 34 stocks. This can be considered sufficient diversification for an equal-weighted portfolio. Nevertheless, typical to stock markets also Nordic stock markets have few extremely large market capitalization companies. The performance of these companies can even after winsorizing the market value drive the performance of the whole portfolio if the number of stocks inside the portfolio is limited.

Additionally, it is not guaranteed that each expected return portfolio would consist of the same amount of stocks. This is because breakpoint expected returns for the portfolio allocation are calculated from the expected returns of large market capitalization stocks. The distribution of the expected returns of the small market value companies does not necessarily follow the expected return distribution of the large market value companies, which could lead to unbalanced portfolios. This can further lower the diversification of the portfolios. One alternative to ensure diversification of the machine learning portfolios would be to allocate stocks to only five expected return portfolios instead of ten.

Another interesting remark from Table 7 is that the standard deviation of the realized excess portfolio returns does not increase together with average realized returns. For this reason, Sharpe ratios increase together with expected returns. For all models, the equal-weighted portfolios portfolio with the highest expected return has also the highest Sharpe ratio if high-low portfolios are not considered. Among the value-weighted portfolios, this is true for the linear regression and the neural network model. Since the volatility of the returns does not increase together with the magnitude of the returns, it means that machine learning models are able to generate excess returns without simply investing in more volatile stocks. Naturally, the correlation of the prices of the stocks inside the portfolio also affects the volatility of the portfolio returns, but this is already the first indication of the risk-adjusted performance of the machine learning portfolios. Risk-adjusted performance is discussed in more detail for high-low portfolios later.

Results from Table 7 further support the findings of Section 6.1 that models overshoot in their predictions. The table shows that average predicted returns for middle expected returns portfolios are close to mean return from Table 2 especially for random forest and neural network models. On the other hand, realized excess returns of the predicted return portfolios between the third and fourth decile land the closest to the market mean return. On the other hand, expected returns for minimum and maximum expected return portfolios are rather extreme. For example for the neural network model spread between the average expected returns of the two extreme portfolios is around four percent for both equal and value-weighted portfolios.

Given that there is a clear trend of increasing realized excess returns among the predicted returns while the expected and realized return of the middle predicted return portfolios quite well match the mean return of the market, it seems that the models do a pretty good job on allocating companies to return clusters, but produce too extreme predictions for lowest and highest predicted return portfolios. On average models are able to find which stocks that produce the highest returns, but are too optimistic in their predictions. There seems to be a similar situation with the lowest expected return portfolios as well. For example, realized average excess return of the lowest expected return portfolios from the neural network model are clearly below average market return, but still positive. As the average predicted return for these portfolios is between -1.52% and -0.24% percent, the spread between realized and expected returns is rather large. This phenomenon could at least partially explain the low out-of-sample R^2 values seen in Section 6.1.

A similar kind of overshooting can be seen in the study of Drobetz and Otto (2021). Nevertheless, in the study of Drobetz and Otto overshooting seems to mainly happen for the linear regression model.[75] For other models predicted and realized excess returns are quite well on the same scale. Interestingly with less predictors Fieberg et al. (2023) do not show the overshooting phenomenon even for the linear regression model.[76]

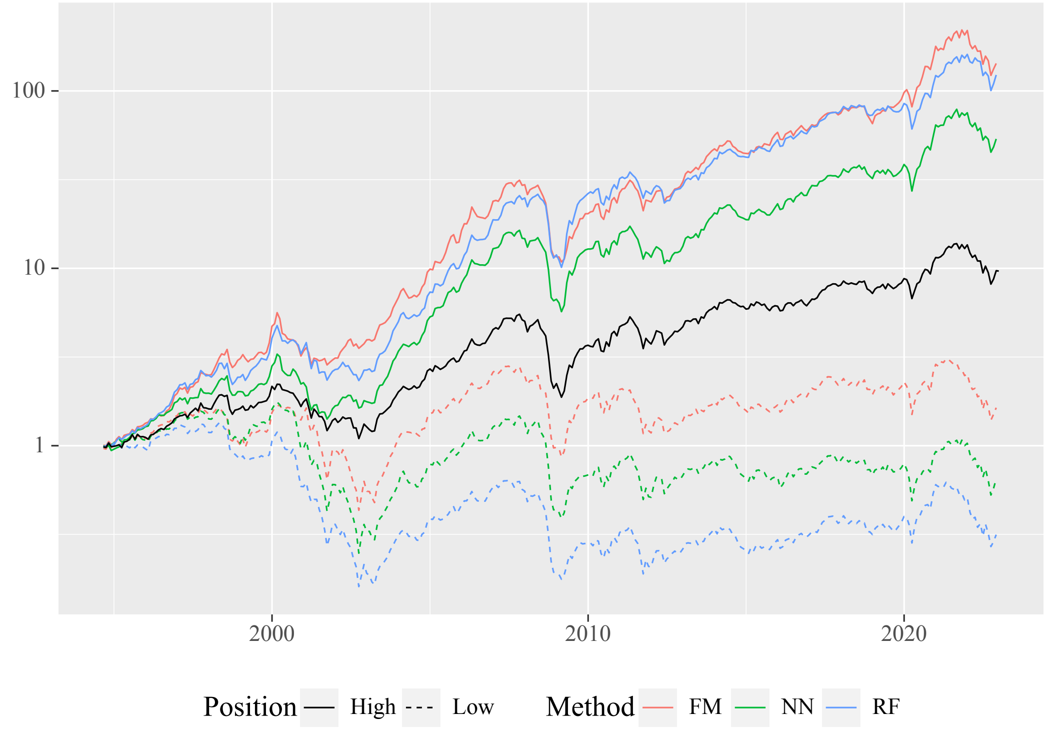
Figure 6 and Figure C.6 in the appendix show the historical cumulative return of the highest and lowest expected return portfolios for all models and for equal and value-weighted portfolios correspondingly. The solid line shows the cumulative excess return for the highest expected return portfolio, whereas the dashed line indicates the cumulative excess return of the lowest expected return portfolio for each model. Figures are in line with the results from Table 7. Overall market trends can be seen from all portfolios, but there is a clear spread between low and high expected return portfolios. Compared to Table 7 Figures 6 and C.6 provide the time series dimension of the re-

[75] See: Drobetz and Otto (2021), pp. 522-525.

[76] See: Fieberg et al. (2023), pp. 309-310.

Figure 6: Cumulative return of equal-weighted machine learning portfolios - Own source

Figure plots the realized historical cumulative excess return of the out-of-sample predictions. The figure shows the performance of portfolios that are formed by allocating all except microcap stocks to ten portfolios based on their expected returns. Re-allocation is done monthly. Section 4 describes how expected returns are derived for different models. FM stands for linear regression model, RF stands for random forest model and NN stands for neural network model. The solid line plots the cumulative performance of the highest expected return portfolio whereas the dashed line plots the cumulative excess return for the lowest expected return portfolio. All portfolios are equal-weighted. The solid black line shows the value-weighted market return. All returns are converted to US dollars. The prediction period spans from July 1994 to November 2022.



turns. For equal-weighted portfolios, Figure 6 reveals a rather constant spread between high and low expected return portfolios, which results in divergent cumulative returns. Despite the average realized return being slightly positive for the lowest expected return portfolios for the random forest model, the cumulative return of the portfolios is negative.

The same overall market trends can be seen in Figure C.6 for value-weighted portfolios. Compared to equal-weighted portfolios, value-weighted portfolios show more seasonality. Among value-weighted portfolios there are clear differences especially on the performance of the low expected return portfolio. Performance of the market value-weighted portfolios for both the highest expected return portfolio as well as the lowest expected return portfolio is more modest than among equal-weighted portfolios. Even the random forest model is not able to generate clear negative cumulative return for the lowest expected return portfolio.

The next part of the paper focuses on evaluating the performance of the long-short portfolios. Table 8 reports a set of performance metrics for both equal and value-weighted portfolios. The riskiness of the portfolios is evaluated by maximum drawdown and maximum one-month loss. Additionally, risk-adjusted performance is evaluated by regressing the realized returns of each of the portfolios by the benchmark factors. From these regressions alphas, t-statistics for alphas and R^2 values are reported. Also, Sharpe values are reported. Finally, the table reports the turnover for the long side of the long-short portfolios.

Table 8: Zero investment portfolio performance metrics - Own source
Table shows different performance metrics for the spread in realized excess return of highest and lowest expected return portfolios for each model. The left side of the table shows the results for equal-weighted portfolios and the right side for market value-weighted. Loss metrics reported in the table include maximum drawdown and maximum one-month loss. The table also reports risk-adjusted performance metrics. These include excess return that cannot be explained by regressing realized returns of the portfolios by benchmark factors indicated by alpha. Additionally, t-statistic for the alpha and R^2 values are reported. The table also shows Sharpe ratios for each of the long-short portfolios. The last row of the table shows the turnovers of the long-short portfolios. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model. The prediction period spans from July 1994 to November 2022.

	Equal-weighted			Value-weighted		
	FM	RF	NN	FM	RF	NN
Max DD(%)	-0.4001	-0.2497	-0.2943	-0.6360	-0.5456	-0.5347
Max 1month Loss(%)	-0.2096	-0.1232	-0.2369	-0.3102	-0.2218	-0.4273
FF Alpha	0.0028	0.0138	0.0071	-0.0049	0.0064	-0.0004
t-stats	1.4479	6.4431	3.3433	-1.9984	2.2174	-0.1654
R^2	0.5853	0.1809	0.3075	0.4650	0.1125	0.2567
Sharpe ratio	0.2425	0.4215	0.2730	0.0777	0.2053	0.1021
Turnover (%)	0.4456	0.4516	0.4703	0.4842	0.6194	0.5883

Message from Table 8 is in line with the previous part. The overall performance of equal-weighted long-short portfolios is stronger than value-weighted long-short portfolios. Both maximum drawdown and maximum one-month loss are larger for value-weighted long-short portfolios than their equal-weighted counterparties for each model. This indicates that the value-weighted strategy is more risky than the equal-weighted strategy. On the other hand this could also further indicate the insufficient diversification of the value-weighted portfolios. The random forest model shows the strongest performance among the equal-weighted portfolios with a maximum drawdown of -24% and a maximum one-month loss of -12%. Among the value-weighted portfolios two risk measurements give a slightly inconsistent message linear regression model has the highest maximum drawdown of -64%, but simultaneously neural network model has the highest maximum one-month the loss of -43%.

Table 7 showed the strong positive returns of the long-short portfolios, which was statistically significantly different from zero for all equal-weighted portfolios. Looking at the Sharpe ratios revealed that the returns were not driven by a difference in volatility of long and short portfolios. Next, we try to evaluate whether the positive return of the long-short portfolios can be explained by loadings in the six benchmark factors. This is done by regression of the long-short portfolio returns by the benchmark factors as explained in Section 4.4. Significant positive alpha from these regressions would indicate risk-adjusted excess returns.

Table 8 shows that the excess returns of the equal-weighted long-short portfolios from random forest and neural network models cannot be explained by the benchmark factors. Alpha for the linear model is also positive, but statistically not significant. Additionally, benchmark factors are able to explain 58% of the variation in returns of the equal-weighted long-short portfolio from the linear regression model, whereas for random forest and neural network models portion is only 17% and 32%. None of the alphas of the value-weighted long-short portfolios is positive and statistically significant. The only statistically significant alpha is the negative alpha of the linear regression model portfolio.

Figures C.7 and C.8 in the appendix show the cumulative return of the equal and value-weighted long-short portfolios from July 1994 to November 2022. As a benchmark value-weighted market value is plotted to these figures as well. As can be seen from

Figure 6 among equal-weighted portfolios random forest approach provides the largest spread between the lowest and highest expected return portfolios. For linear regression and neural network models the spread is almost identical. Figure 6 shows that the high expected return portfolio of the linear model performs better than the high expected return portfolio of the neural network, but the neural network model does a better job picking low expected return companies.

An interesting remark from Figure 6 is that for equal-weighted long-short portfolios overall market trends are not visible. This means that the difference in realized excess return between the highest and lowest expected return machine learning portfolios is not affected by the overall stock market distress. For example, the effect of the financial crisis around 2008 can be clearly seen from the cumulative market return, but the cumulative return of any long-short machine learning portfolio has not remarkably changed. Figure 6 shows that the equal-weighted long-short machine learning portfolios provide larger and smoother returns than the market return.

Looking at Figure C.8 shows that none of the value-weighted long-short portfolios are able to remarkably exceed the market return. The cumulative return of the random forest model ends up slightly above market return whereas the cumulative return of the neural network model ends up slightly below it. Performance of the linear model and the neural network model is poor throughout the period whereas random forest performs strongly in the first 16 years of the period. In the last twelve years of the prediction period, the cumulative return of the value-weighted long-short portfolio from the random forest model is actually negative.

Finally, the turnover of the long-short portfolios is surprisingly low. A lot of previous studies report turnover exceeding 100% for long-short portfolios.[77] Due to the nature of long-short portfolios and the definition of the turnover in Section 4.4 turnover can have values above 100%. This is nevertheless, not the case in this study. Table 8 shows that the average monthly turnover of the long-short portfolios is around 50%. The average monthly turnover for value-weighted long-short portfolios varies between 49% for the linear model and 61% for the random forest model. This is slightly more than the

[77] See: e.g. Gu et al. (2020), pp. 2267-2268; Tobek and Hronec (2021), p.14 and p. 20.

monthly turnover of equal-weighted long-short portfolios which are 45% for the linear random forest model and 47% for the neural network model. Figure C.9 in the appendix shows the average turnover for all training periods. Turnovers mainly fluctuate between 40% and 60% throughout the period.

Partially low turnover could be a result of the frequency of the predictor variables. This study includes a lot of covariates with a yearly frequency. In addition, all the models are trained only once a year. This means that if models weight in their predictions more the annually updated variables, then the predicted returns for a company should be rather stable for the next 12 months. In a situation where a set of companies would be constant throughout the whole period and only annual predictors would be included, the order of the predicted returns of the companies would remain the same between model re-trainings. This would mean that the turnover between the trainings would also be zero.

Actually, low turnover is a positive feature for an investor. Real-world investors are usually subject to transaction costs, which naturally reduce the return of the investment. An investment strategy that requires the investor to monthly change position in the majority of invested capital is often not implementable in a real-world setting due to increased transaction costs compared to a more passive investment strategy.

6.3 Predictive characteristic importance

This section sheds light on which covariates contribute the most to the predictive accuracy of the machine learning models. Figure C.11 in the appendix shows the relative importance of the explanatory variables for all models. Results are shown separately for different definitions of out-of-sample R^2. For the variable importance reduction out-of-sample R^2 is calculated separately for each retraining period. The final variable importance figure is then the time series average reduction in out-of-sample R^2. Therefore, variable importance measured using a different definitions of the out-of-sample R^2 can result in slightly different results, but as can be seen from Figure C.11 differences are minor.

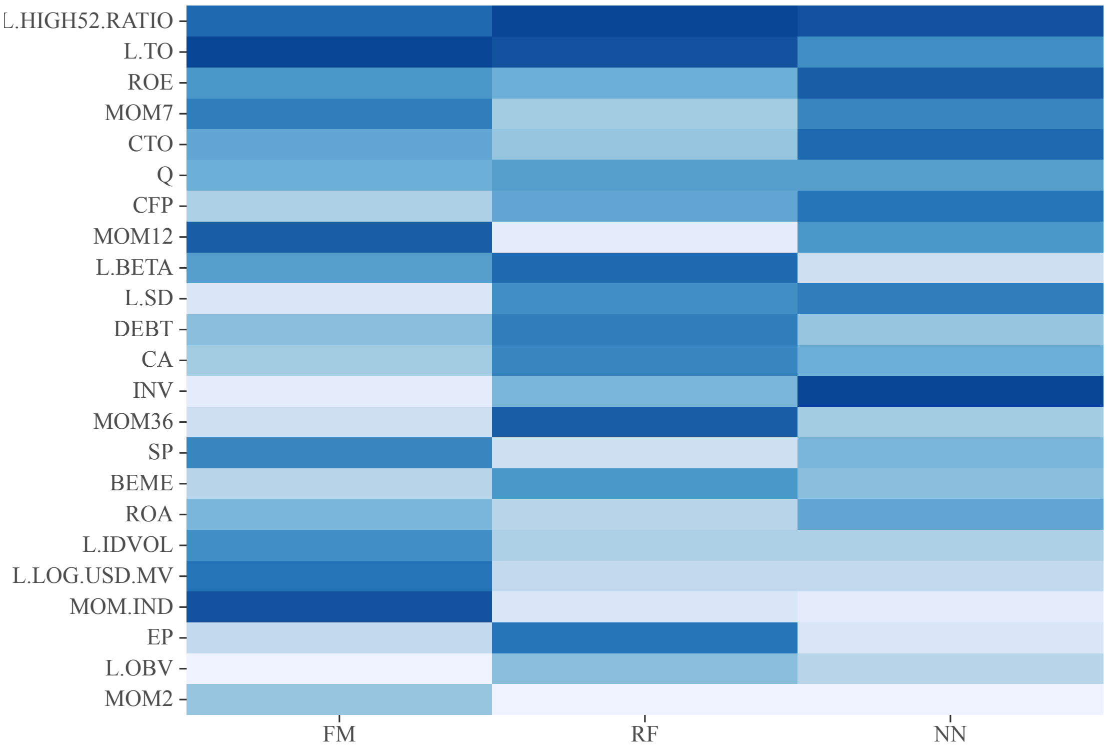
Figure C.11 shows that for linear regression model turnover (L.TO), industry momentum (MOM.IND) and 12-month momentum (MOM12) are the most influential characteristics. Also exclusion of the 52-week high price (L.HIHG52.RATIO) and log market value (L.LOG.USD.MV) results clear reduction in the prediction accuracy of the model. The rest of the variables seem to have only a minor effect on the prediction accuracy of the model or the effect is even negative. Another remarkable result from Figure C.11 is that on-balance volume (L.OBV) has a clear negative effect on the prediction performance of the linear model and predictions of the model are more precise when the variable is excluded.

The variable importance of the random forest model is extremely skewed. Only few characteristics seem to have positive impact on the predictive accuracy of the random forest model. Most notably the 52-week high price (L.HIHG52.RATIO) and lagged turnover (L.TO). Interestingly lot of the variables have even negative impact to the predictive performance of the random forest model. The variables that have strongest negative impact seem to be the momentum variables as the three least important variables are short-term reversal (MOM2), 12-month momentum (MOM12) and industry momentum (MOM.IND).

Variable importance for the neural network model shown in Figure C.11 is a bit similar as for the neural network model. A few variables have positive impact to the predictive accuracy of the neural network model. These include investments (INV), return on equity (ROE) and the 52-week high price (L.HIHG52.RATIO). Similar to the random forest model also for the neural network model momentum variables have clear negative impact on the out-of-sample R^2 values. Two least important variables for the neural network model are short-term reversal (MOM2) and industry momentum (MOM.IND).

Figure 7 provides further insight into how aligned the variable importances of different firm characteristics are between different models as it shows the variable importance ranks of the models. Each model variable with the highest variable importance gets a variable importance rank of one. Therefore, the darker color in Figure 7 indicates a higher variable importance and the lighter color lower variable importance. The figure shows some dispersion in the variable importance of the different models. Especially the linear regression model deviates from random forest and neural network models as it weights the momentum variables a lot. Where the models agree is the relatively high importance of the 52-week high price (L.HIHG52.RATIO) and low importance of the lagged on-balance volume (L.OBV) and the short-term reversal (MOM2).

Figure 7: Variable importance - Own source
Figure plots the importance of the explanatory variables to the predictive performance of the three machine learning models. Variable importance is defined as a reduction in out-of-sample R^2 when the corresponding variable is replaced by zero before each training process. The definition of the out-of-sample R^2 is described in Section 4. A darker color indicates higher variable importance. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model. The prediction period spans from July 1994 to November 2022.



The strong importance is in line with some of the previous literature. For example Hanauer and Kalsbach (2023) report the 52-week high price ratio as one of the most important characteristics for most of their models in their study on emerging markets. Similarly Tobek and Hronec (2021) show strong importance of the 52-week high price ratio in European setting across the models.[78] What is more surprising is the weak performance of the traditional momentum factors. Momentum factor showed strongest performance among the benchmark factors in Section 5.2. The momentum effect is also

[78] See: Hanauer and Kalsbach (2023), p. 8; Tobek and Hronec (2021), p. 16.

well documented in previous Nordic stock market anomaly literature.[79] Including the on-balance volume even reduces the prediction accuracy for all of the models as can be seen from Figure C.11. One objective of this study was to examine whether on-balance volume would contain information about the future cross-section of stock returns. At least in the setting of this study information that on-balance volume can provide about future stock returns is limited.

Sections 6.1 and 6.2 showed evidence that models could be overshooting in their predictions. This phenomenon could also affect variable importance. Especially, this could be an important factor for the variable importance of the linear regression and neural network models with clearly negative out-of-sample R^2. Hypothetically if the exclusion of a characteristic would demean all the predictions of the model to 0, variable importance of this variable would be clearly positive. This is because then in the described case out-of-sample R^2 would be zero and since the out-of-sample R^2 of the full model is negative there would be a positive change in out-of-sample R^2. Nevertheless, after excluding this hypothetical variable model would become useless for an investor, since constant prediction would not provide any information about the cross-section of future stock returns. Therefore, the investor would not have any indicator for the portfolio construction from Section 6.2.

The above-described situation is only hypothetical, since even if the only variable containing information of the future returns would be excluded linear regression would not necessarily predict zeros, but rather the cross-sectional averages. For example, the average excess return for Nordic stocks in the time period of this study is 0.7%. Still, at least partially variable importance of certain variables could be driven by the fact that their removal just reduces the overshooting of the model. One option to investigate this would be to reproduce the expected return portfolios after the exclusion of a variable. This way it could be seen if including a variable brings the realized excess return of an expected return portfolio closer to its expected return. Additionally, reasonability of the relative variable importance measure can be questioned as we have some variables with negative variable importance. This reduces the sum used to scale the variable importance measures and can make the changes look larger than they actually are.

[79] See: Grobys and Huhta-Halkola (2019), pp. 10-11; Leivo and Pätäri (2011), pp. 407-411.

7 Conclusion

In this study I examine the cross-sectional predictability of stock returns in Nordic stock markets using machine learning methods. I construct 23 firm-level characteristics for all publicly listed companies in four Nordic stock markets between January 1990 and December 2022. Main contribution of this study to existing literature is two-fold. First, machine learning approach from Gu et al. (2020) is applied to a new submarket.[80] Second, existing stock market anomaly literature is extended as explanatory variable set used in this study contains characteristics that have not been studied in Nordic stock markets in great extent.

This study has shown that firm level characteristics contain information about future cross-section of stock returns in Nordic stock markets. Additionally, this study shows that implementing predictive machine learning models can result in more precise stock-level return predictions compared to linear regression model. Especially random forest model shows strong performance in the Nordic market setting measured in both accuracy and profitability of the predictions. Nevertheless all of the examined models seem to suffer from overshooting, producing stock returns with higher variance as the actual realized returns.

Hyperparameter optimization in this study is limited. Therefore, it could be of interest in future research to examine how more comprehensive hyperparameter optimization would affect the results. As shown in this study there seems to be evidence of the time varying predictive power of the firm-level characteristics. Therefore it could be justified to examine different lengths of the rolling window for the linear regression model or different sample splitting schemes for the machine learning models.

[80] See: Gu et al. (2020), p. 2230-2246.

Appendix A Data collection and variable definitions

Data for this study is collected from Datastream. Raw data is collected using constitute sets introduced in Table A.1. For each country research, Worldscope and dead constitute lists are considered. Including dead lists allows us to avoid survivorship bias. As shown by Ince and Porter (2006) in order to ensure data quality, data from Datastream requires cleaning. Tables A.2, A.3 and A.4 present the dynamic and static screens used in the data cleaning.

Dynamic screens from Table A.2 result deletion of observations from the dataset. If an observation is deleted due to a dynamic screen, corresponding security is not necessarily completely excluded from the dataset. Tables A.3 and A.4 show the static screen. The objective of these screens is to clean the dataset from non-common and duplicate stock affiliations. Panel A of Table A.3 shows which values are accepted for certain attributes, whereas panel B of Table A.3 and Table A.4 introduce maleficent keywords that are searched among Datastream attributes NAME, ENAME and ECNAME. In case a requirement from panel A of Table A.3 is not met or if a maleficent keyword is found in the name of the security, security is excluded from the dataset completely.

Table A.1: Constituent lists and keywords
Modified taken from Ince and Porter (2006) and Hanauer and Windmüller (2023). The table provides the constituent lists used in data collection.

Denmark	Finland	Norway	Sweden
FDEN WSCOPEDK DEADDK	FFIN WSCOPEFN DEADFN	FNOR WSCOPENW DEADNW	FSWD WSCOPESD DEADSD FAKTSWD

Table A.2: Dynamic screens
Modified taken from Ince and Porter (2006) and Hanauer and Windmüller (2023). The table provides the dynamics screens used in the data cleaning.

Affected attribute	Applied screen
RI	Observations where the one-month return is larger than 990% are removed.
RI	Observation is removed if return in r_t or r_{t-1} exceed 300% and $(1 + r_t)(1 + r_{t-1}) - 1$ is less than 0.5.
RI	For periods after the delisting of a security Datastream returns the last available value. Therefore, by removing all consecutive zero returns at the end of the dataset for all securities.

Table A.3: Static screens
Modified taken from Ince and Porter (2006) and Hanauer and Windmüller (2023). Panel A of the table provides the Datastream items that are considered for the filtering and accepted values for each country separately. Panel B of the table provides keywords that were used to delete entries from each market separately. The same logic is applied to remove both country-specific and generic keywords. A keyword is searched from Datastream attributes NAME, ENAME and ECNAME. In case at least one of these attributes contains the keyword security is deleted from the dataset. To avoid deleting proper entries, security is only deleted if the keyword occurs at the beginning of the name, at the end of the name or as a separate word in the name.

Panel A: Static screens.

	Denmark	Finland	Norway	Sweden
MAJOR	Y	Y	Y	Y
TYPE	EQ	EQ	EQ	EQ
ISINID	P	P	P	P
GEOGN	DENMARK	FINLAND	NORWAY	SWEDEN
GEOLN	DENMARK	FINLAND	NORWAY	SWEDEN
PCUR	DK	FI, MK	NK	SK
GGSIN	DK	FI	NO	SE

Panel B: Country specific keywords.

	Denmark	Finland	Norway	Sweden
NAME ENAME ECNAME	\\)CSE \\	USE		CONVERTED INTO, USE, CONVERTED-, CONVERTED - SEE

Table A.4: Common keywords
Modified taken from Ince and Porter (2006) and Hanauer and Windmüller (2023). The table shows the general keywords that were used to delete entries from all markets. The same keyword deletion logic is applied to remove both country-specific and generic keywords. The keyword is searched from Datastream attributes NAME, ENAME and ECNAME. In case at least one of these attributes contains the keyword security is deleted from the dataset. To avoid deleting proper entries, security is only deleted if the keyword occurs at the beginning of the name, at the end of the name or as a separate word in the name.

Security class	Keywords
Duplicates	1000DUPL, DULP, DUP, DUPE, DUPL, DUPLI, DUPLICATE, XSQ, XETa
Depository receipts	ADR, GDR
Preferred stock	PF, 'PF', PFD, PREF, PREFERRED, PRF
Warrants	WARR, WARRANT, WARRANTS, WARRT, WTS, WTS2
Debt	%, DB, DCB, DEB, DEBENTURE, DEBENTURES, DEBT
Unit trusts	.IT, .ITb, TST, INVESTMENTTRUST, RLSTIT, TRUST, TRUSTUNIT, TRUSTUNITS, TST, TSTUNIT, TST UNITS, UNIT, UNITTRUST, UNITS, UNT, UNTTST, UT
ETFs	AMUNDI, ETF, INAV, ISHARES, JUNGE, LYXOR, X-TR
Expired securities	EXPD, EXPIRED, EXPIRY, EXPY
Miscellaneous	ADS, BOND, CAP.SHS, CONV, DEFER, DEP, DEPY, ELKS, FD, FUND, GW.FD, HI.YIELD, HIGHINCOME, IDX, INC. &GROWTH, INC.&GW, INDEX, LP, MITS, MITT, MPS, NIKKEI, OPCVM, ORTF, PERQS, PFC, PFCL, PINES, PRTF, PTNS, PTSHP, QUIBS, QUIDS, RATE, RCPTS, REALEST, RECEIPTS, REIT, RESPT, RETUR, RIGHTS, RST, RTN.INC, RTS, SBVTG, SCORE, SPDR, STRYPES, TOPRS, UTS, VCT, VTG.SAS, XXXXX, YIELD, YLD, PF.SHS.

Table A.5: Variable definitions - Own source
Table provides definitions and initial authors for all anomalies considered in this study. Construction of variables follows mainly Green et al. (2017) and Hanauer and Kalsbach (2023) and can deviate from variable definitions of initial authors. The table also provides the direct formulas and relevant Datastream items used to calculate the variables. Abbreviations used to indicate different variables later in the study are also displayed in the table. MV_{Dec} indicates market value as of the end of December in year $t-1$. The frequency of the variable is indicated after the variable name.

Variable	Author	Definition
Cash-to-Assets *Yearly*	Palazzo (2012), pp. 164-168	Cash-to-Asset ratio is calculated by dividing cash and short-term investments by total assets. CA_t = WC02001_t / WC02999_t
Capital Turnover *Yearly*	Haugen and Baker (1996), p. 406	Capital turnover is calculated by dividing total sales by one year lagged total assets. CTO_t = WC01001_t / WC02999_{t-1}

Table continues in the next page …

… Continued from the previous page

Variable	Author	Definition and affected Datastream items
Investment *Yearly*	Cooper, Gulen and Schill (2008), p. 161	Investments are defined as a yearly change in total assets. INV_t = (WC02999_t - WC02999_{t-12}) / WC02999_{t-12}
Book-to-Market Equity *Yearly*	Davis, Fama and French (2000), .p 393	Book-to-Market value is calculated by dividing the company's book value of equity by the company's market capitalization end of the previous year. The book value of equity is calculated by summing common equity and deferred taxes of the company. BEME_t = (WC03501_t + WC03263_t) / MV_{Dec}
Cash Flow-to-Price *Yearly*	Lakonishok, Shleifer and Vishny (1994), pp. 1545-1546	Cash flow to price ratio is calculated by dividing the company's cash flow from operating activities by the asset's market capitalization end of the previous year. CFP_t = WC04860_t / MV_{Dec}
Debt-to-Price *Yearly*	Bhandari (1988), .p 509	Debt-to-price value is calculated as the difference between total assets and common equity divided by the asset's market capitalization end of the previous year. DEBT_t = (WC02999_t - WC03501_t) / MV_{Dec}
Sales-to-Price *Yearly*	Lewellen (2015), .p 7	Sales-to-price ratio is calculated by dividing total sales by the asset's market capitalization end of the previous year. SP_t = WC01001_t / MV_{Dec}
Earnings-to-Price *Yearly*	Basu (1977), pp. 664-665	Earnings-to-price ratio is calculated by dividing net income before extraordinary items by asset's market capitalization end of the previous year. EP_t = WC01551_t / MV_{Dec}
Return-on-Assets *Yearly*	Balakrishnan, Bartov and Faurel (2010), p. 23	Return-on-assets is calculated as net income before extra items and preferred dividends divided by one year lagged total assets. ROA_t = WC01551_t / WC02999_{t-12}
Return-on-Equity *Yearly*	Haugen and Baker (1996), .p 406	Return-on-equity is calculated as net income before extra Items and preferred dividends divided by one year lagged book value of equity. See book-to-market equity for the definition of the book value of equity. ROE_t = WC01551_t / BE_{t-12}
Tobin's Q *Yearly*	Freyberger, Neuhierl and Weber (2020), p. A3	Tobin's Q is calculated by summing up total assets and market capitalization from the previous December, then subtracting cash and short-term investments and deferred taxes. Finally, the result is divided by the total assets. Q_t = (WC02999_t + MV_{Dec} - WC02001_t - WC03263_t) / WC02999_t
Momentum_7 *Monthly*	Novy-Marx (2012), pp. 431-432	MOM7 is defined as cumulative return in US dollars between $t-7$ and $t-12$ months.

Table continues in the next page …

… Continued from the previous page

Variable	Author	Definition and affected Datastream items
Momentum$_{12}$ *Monthly*	Jegadeesh and Titman (1993), p. 68	MOM12 is defined as cumulative return in US dollars between $t-2$ and $t-12$ months.
Momentum$_{36}$ *Monthly*	De Bondt and Thaler (1985), pp. 797-798	MOM36 is defined as cumulative return in US dollars between $t-12$ and $t-36$ months.
Momentum$_{2}$ *Monthly*	Jegadeesh (1990), p.882	MOM2 is defined as the prior month's return in US dollars.
Industry Momentum *Monthly*	Moskowitz and Grinblatt (1999), p.1252	MOM.IND is defined as 12-month cumulative equal-weighted industry return. The industry is defined using INDG attribute from Datastream.[81]
Standard deviation *Monthly*	Ang, Hodrick, Xing and Zhang (2006), p. 264	L.SD is defined as the standard deviation of unadjusted weekly price for the last 52 weeks.[82]
52-week high price *Monthly*	George and Hwang (2004), p. 2149	52-week high price ratio is calculated from weekly unadjusted prices by dividing current price by past 52-week high price. L.HIGH52.RATIO$_t$ = UP$_{t-1}$ / UP$_{52weekhigh}$
Beta *Monthly*	Fama and MacBeth (1973), pp. 609-610	L.BETA is estimated by beta coefficients obtained by regressing unadjusted weekly returns noted in US dollars with equally weighted market returns. Minimum 15 observations is required.
Idiosyncratic volatility *Monthly*	Ali, Hwang and Trombley (2003), p.356	L.IDVOL is estimated by the standard deviation of regression residuals from regressing unadjusted weekly US dollar returns by equally weighted market return.
Log. market value *Monthly*	Banz (1981), p. 4	Log. market value is defined as natural logarithm of the market value of the company end of the previous month. L.LOG.USD.MV$_t$ = ln(USD.MV$_{t-1}$)
Turnover *Monthly*	Datar, Naik and Radcliffe (1998), p. 208	Turnover is defined as the end of the previous month's weekly trading volume divided by the shares outstanding. L.TO$_t$ = VO$_{t-1}$ / NOSH$_{t-1}$[83]
On-balance volume *Monthly*	Tsang and Chong (2009), p. 2	L.OBV is calculated with the following process. First on-balance volume is the weekly trading volume multiplied by the corresponding week's return's sign. Following on-balance volumes are calculated by adding the product of trading volume and the sign of the return to the previous on-balance volume.

[81] Moskowitz and Grinblatt (1999) Use calculate value weighted industry returns whereas in this study equal weighted industry returns are used.

[82] Ang et al. (2006) examine how market volatility risk is affects the cross-section of the stock returns. Therefore our definition is closer to Green et al. (2017)

[83] Datar et al. (1998) Use average of last three months trading volume, where as in this study latest weekly value is used.

Appendix B Benchmark factor properties

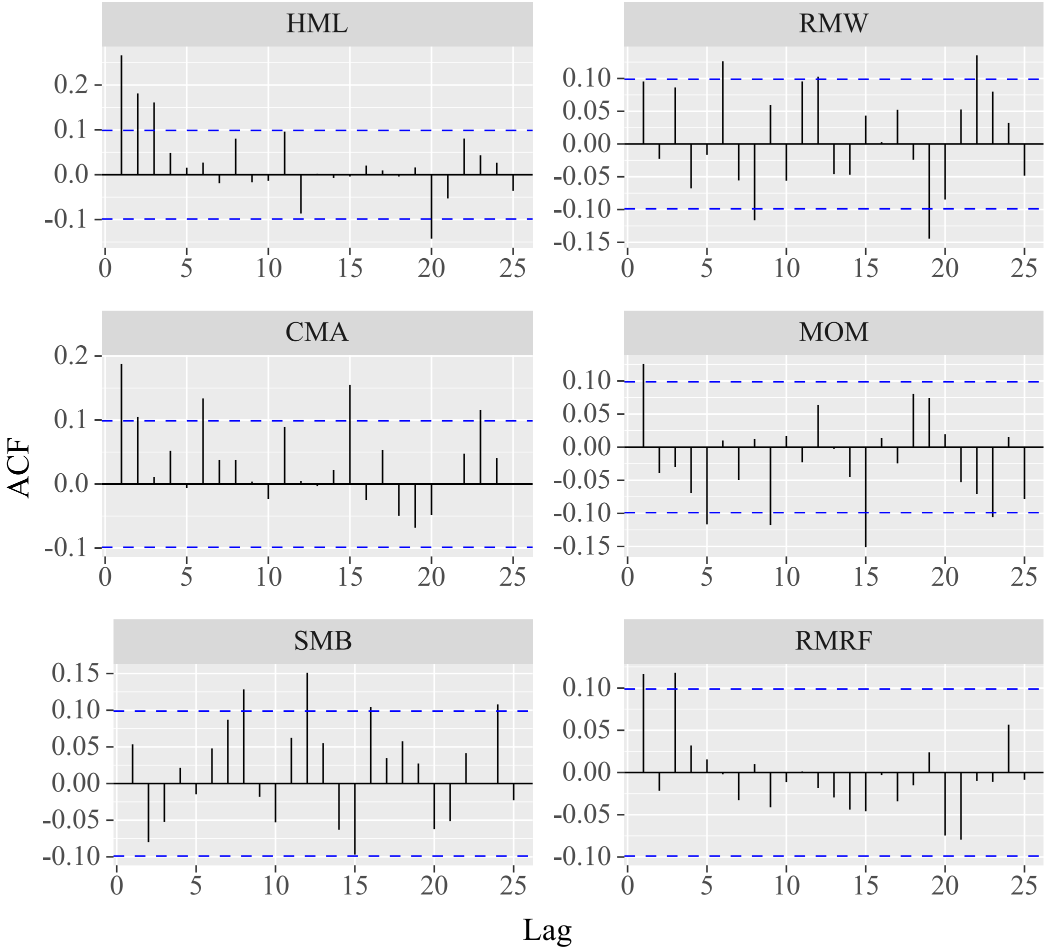
Supplementary materials for the benchmark factors.

Table B.1: Benchmark factor correlation matrix - Own source
Table shows the correlations among the benchmark factors. RMRF is the average value return of the pooled Nordic market. Portfolio returns are calculated based on 2 × 3 sorts on size and one other factor. HML is the difference in the average of the value-weighted return of two high value portfolios and the average of the value-weighted return of two low value portfolios. RMW, CMA and MOM are calculated in a similar manner, but portfolio sorts are done based on investment, profitability and momentum factors. SMB is the average of the value-weighted returns of the 12 portfolios of small stocks minus the average of the value-weighted returns of the 12 portfolios of big stocks. Returns are calculated in US dollars.

	HML	RMW	CMA	MOM	SMB	RMRF
HML	1					
RMW	-0.5814	1				
CMA	0.5889	-0.6194	1			
MOM	0.0904	0.0998	-0.0810	1		
SMB	0.2737	-0.2453	0.1563	0.1674	1	
RMRF	-0.2674	0.0631	-0.2394	-0.1976	-0.2704	1

Figure B.1: Factor autocorrelation - Own source

Figure plots the the benchmark factors autocorrelations. RMRF is the average value return of the pooled Nordic market. Portfolio returns are calculated based on 2 × 3 sorts on size and one other factor. HML is the difference in the average value-weighted return of two high value portfolios and the average value-weighted return of two low value portfolios. RMW, CMA and MOM are calculated in a similar manner, but portfolio sorts are done based on investment, profitability and momentum factors. SMB is the average of the value-weighted returns of the 12 portfolios of small stocks minus the average of the value-weighted returns of the 12 portfolios of big stocks. Returns are calculated in US dollars.



Appendix C Additional information

Figure C.1: Number of companies - Own source
Figure shows the development of the total number of securities considered in the dataset from 1990 to 2022 for each Nordic country. The figure counts all securities that passed the static screens.

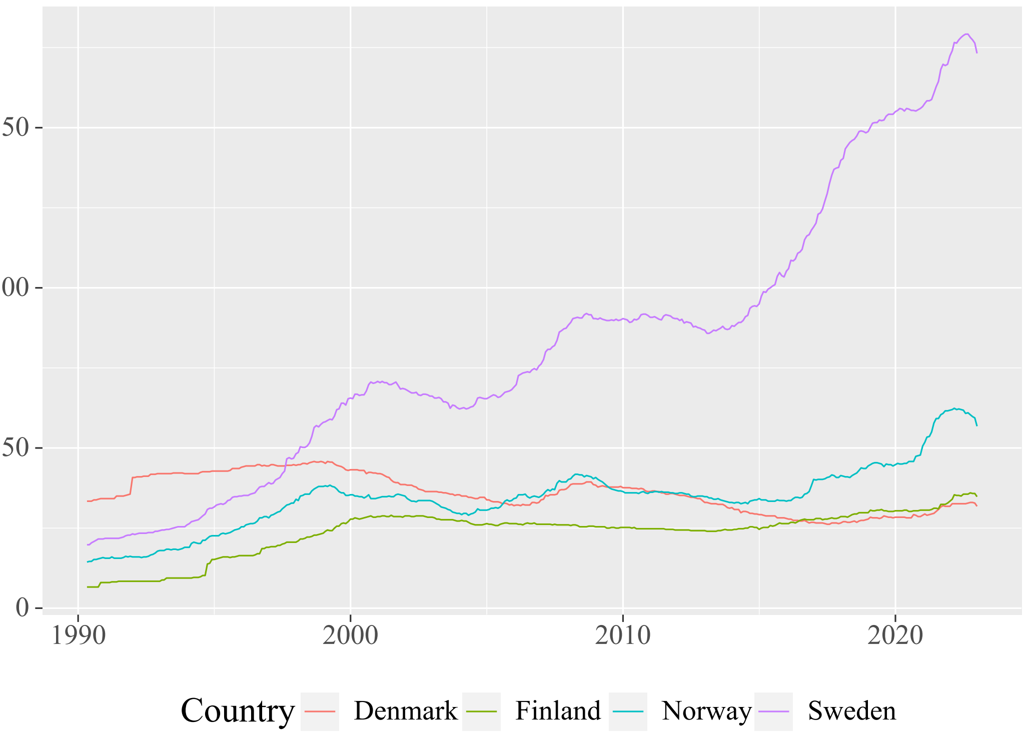


Figure C.2: Exchange rates - Own source
Figure shows the development of currency rates compared to US dollars. DK stands for Danish krone, E stands for Euro, MK stands for Finnish markka, NK stands for Norwegian krone and SK stands for Swedish krona.

Figure C.3: US dollars one-month Treasury bill rate - Own source
Figure shows the development of US dollars one-month Treasury bill rate, which is used as the risk-free rate in this study.

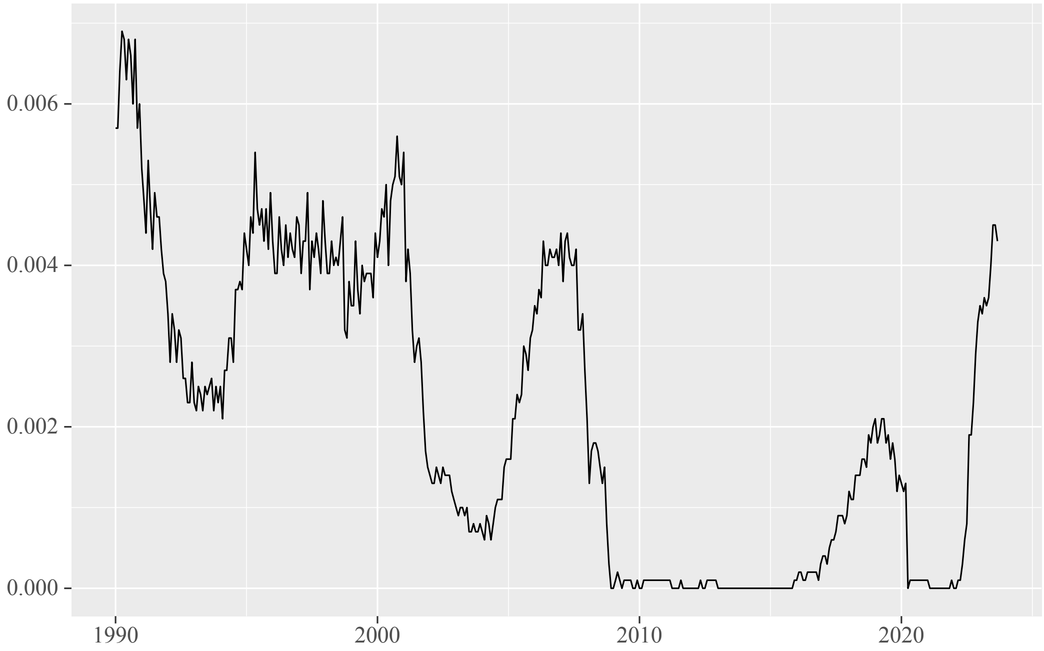


Figure C.4: Time series of the mean cross-sectional properties - Own source
Figure plots development of the firm characteristics across time. The values shown are the monthly cross-sectional averages. The construction of each variable is explained in detail in Section 3.

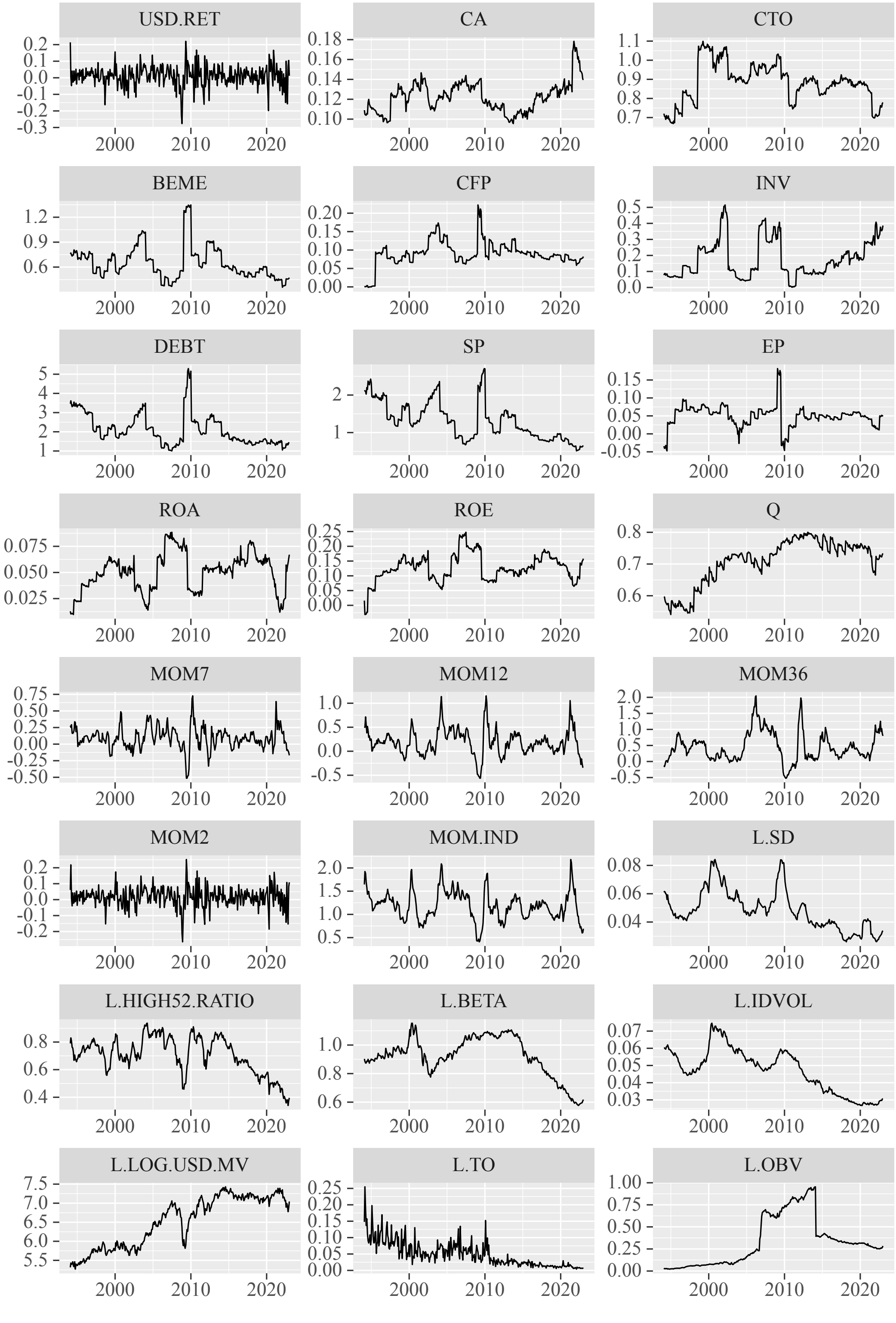


Figure C.5: Time series of out-of-sample R^2s - Own source
Figures present the out-of-sample predictive performance of different machine learning models. Left side graphs show the out-of-sample R^2 values with a benchmark prediction of zero. This method is described in Section 4. Additionally, traditional out-of-sample R^2s are displayed. In traditional out-of-sample R^2 benchmark prediction is the historical mean of corresponding stocks return. R^2s are calculated for each retraining period.

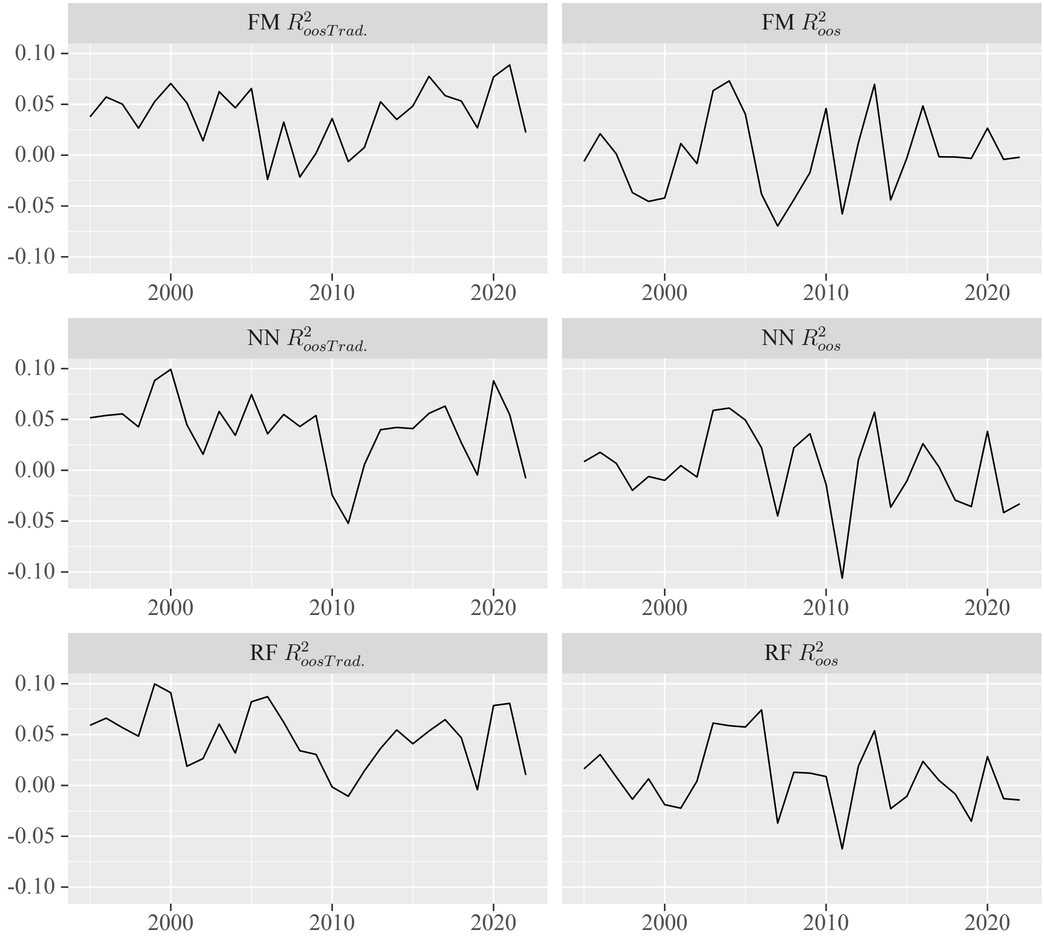


Figure C.6: Cumul. return of value-weighted machine learning portfolios - Own source

Figure plots the realized historical cumulative excess return of the out-of-sample predictions. The figure shows the performance of portfolios that are formed allocating all except microcap stocks to ten portfolios based on their expected returns. Re-allocation is done monthly. Section 4 describes how expected returns are derived for different models. FM stands for linear regression model, RF stands for random forest model and NN stands for neural network model. The solid line plots the cumulative performance of the highest expected return portfolio whereas the dashed line plots the cumulative excess return for the lowest expected return portfolio. All portfolios are value-weighted. The solid black line shows the value-weighted market return. All returns are converted to US dollars. The prediction period spans from July 1994 to November 2022.

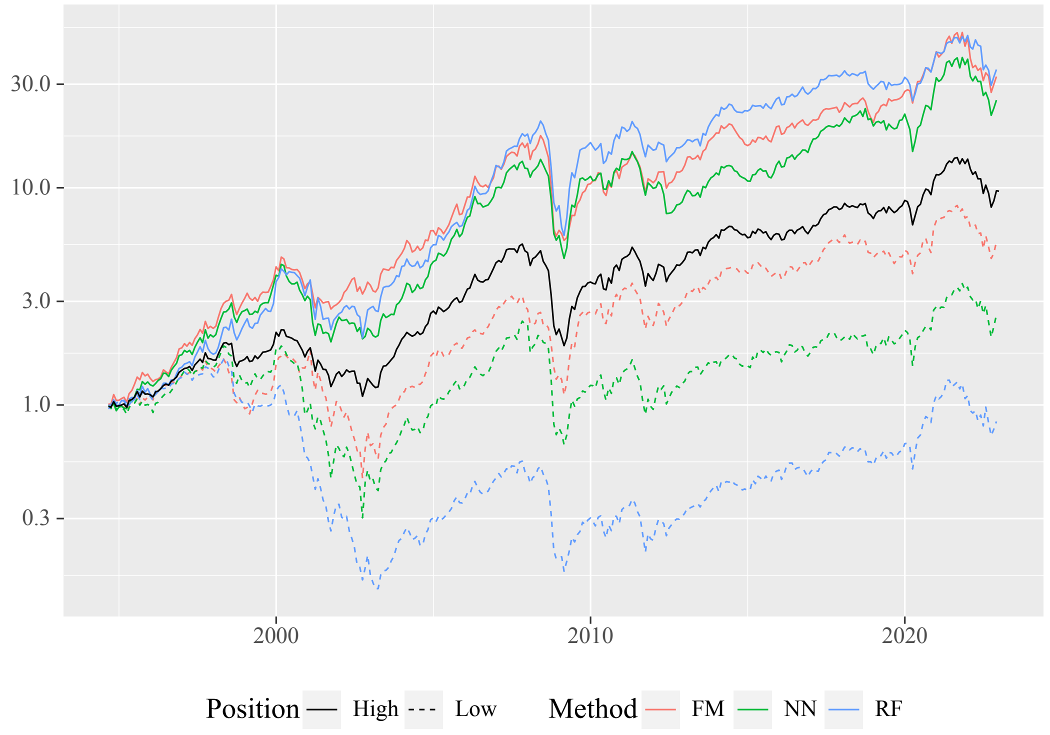


Figure C.7: Cumul. return of equal-weighted long-short portfolios - Own source
Figure presents the realized cumulative spread return between highest expected return portfolio and lowest expected return portfolio. Re-allocation of stocks to portfolios is done monthly. Section 4 describes how expected returns are derived for different models. Both high and low expected return portfolios are equal-weighted. FM stands for linear regression model, RF stands for random forest model and NN stands for neural network model. The solid black line shows the value-weighted market return. All returns are converted to US dollars. The prediction period spans from July 1994 to November 2022.

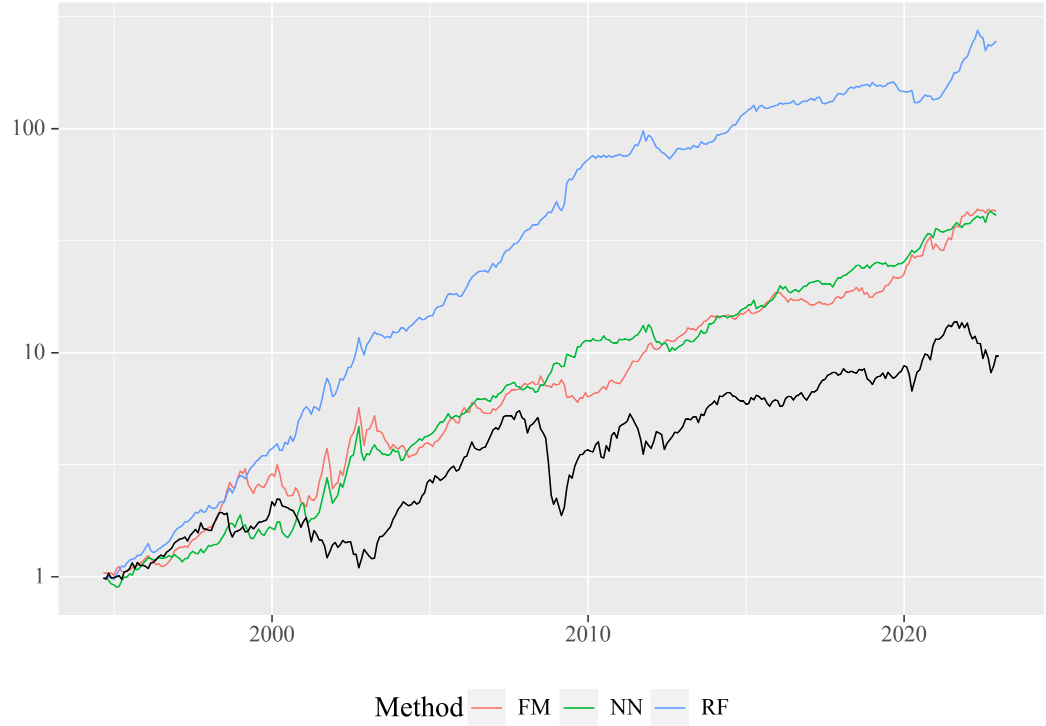


Figure C.8: Cumul. return of value-weighted long-short portfolios - Own source
Figure presents the realized cumulative spread return between highest expected return portfolio and lowest expected return portfolio. Re-allocation of stocks to portfolios is done monthly. Section 4 describes how expected returns are derived for different models. Both high and low expected return portfolios are value-weighted. FM stands for linear regression model, RF stands for random forest model and NN stands for neural network model. The solid black line shows the value-weighted market return. All returns are converted to US dollars. The prediction period spans from July 1994 to November 2022.

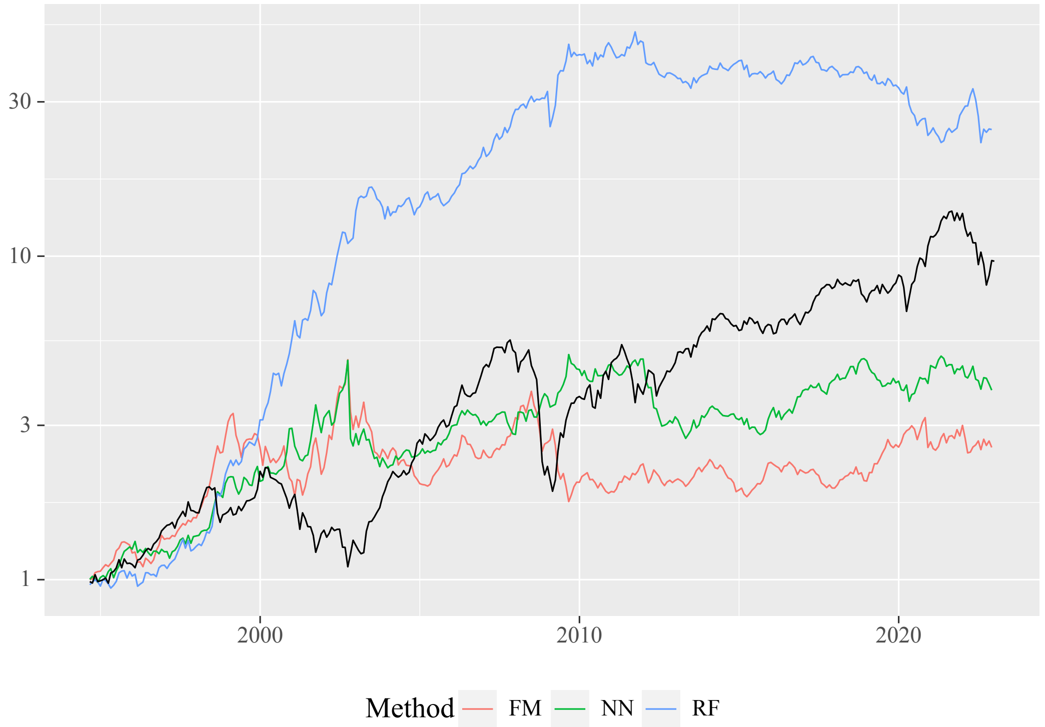


Figure C.9: Turnover of long-short machine learning portfolios - Own source
Figure plots time series of turnover of the long-short portfolios. Values shown are mean turnovers of prediction periods. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model.

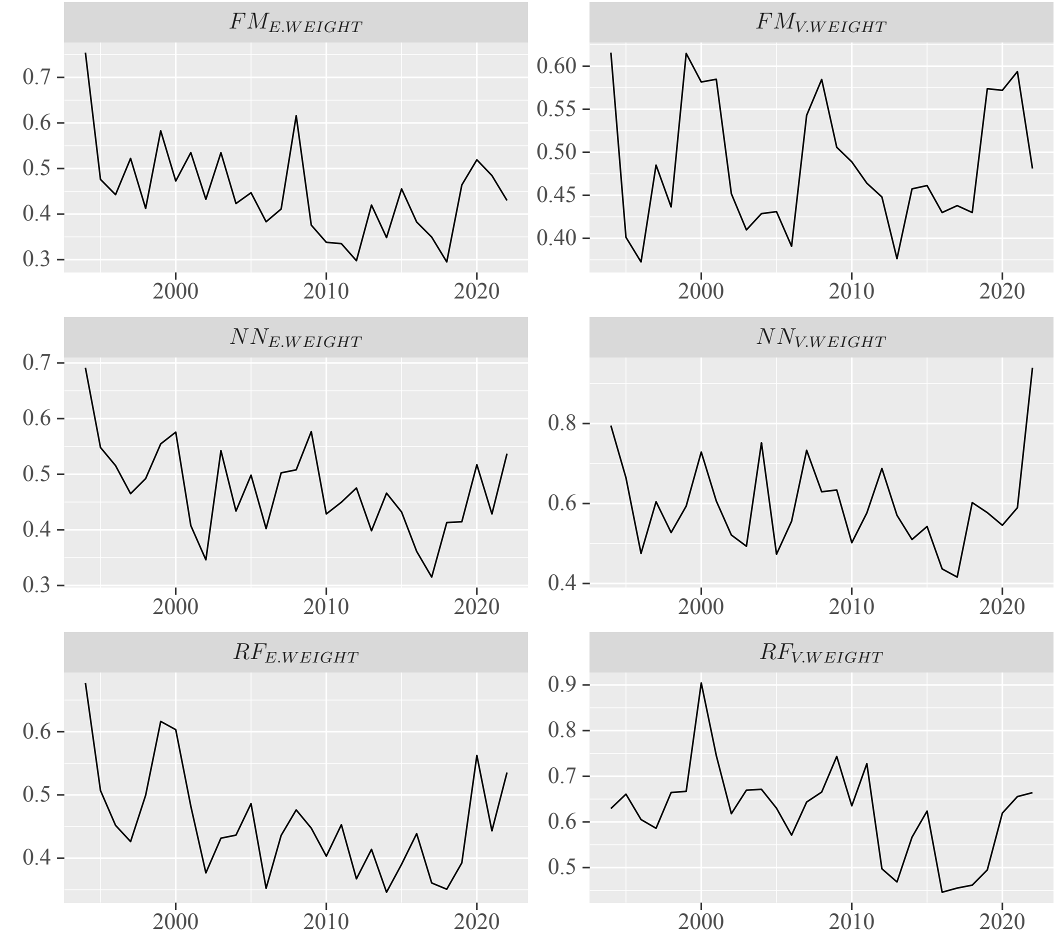


Figure C.10: Random forest optimized hyperparameters - Own source
Time series of optimal hyperparameters for random forest model. Mtry stands for the number of features to possibly split at in each node and max.depth stands for the maximum depth of the regression trees in the random forest model.

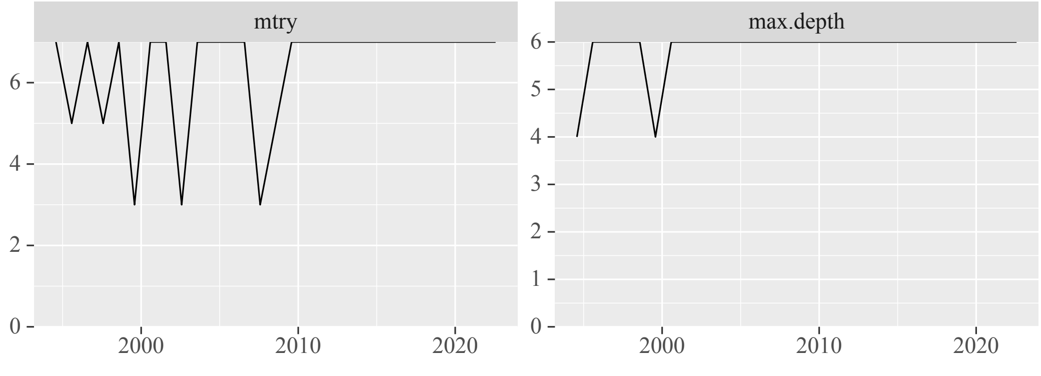
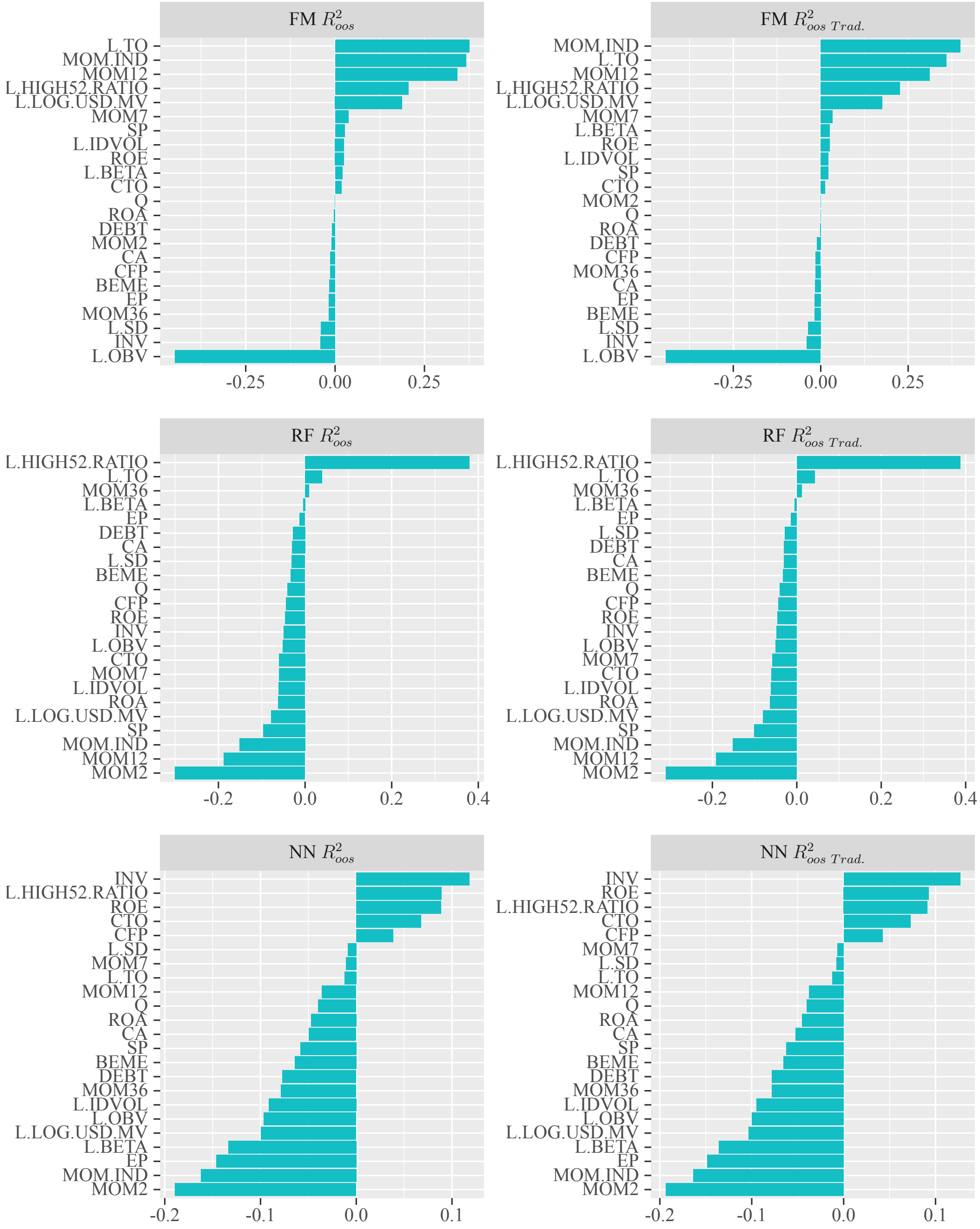


Figure C.11: Relative variable importance - Own source

Figure plots the relative importance of the explanatory variables to the predictive performance of the three machine learning models. Variable importance is defined as a reduction in out-of-sample R^2 when the corresponding variable is replaced by zero before each training process. The definition of the out-of-sample R^2 is described in Section 4. In order to obtain relative variable importance measures, reductions in out-of-sample R^2 compared to the full model are normalized to sum to one within one model. FM stands for linear regression model, RF stands for random forest model and NN stands for neural networks model. The prediction period spans from July 1994 to November 2022.



References

Ali, A. / Hwang, L.-S. / Trombley, M. A. (2003): Arbitrage risk and the book-to-market anomaly, in: *Journal of Financial Economics*, 69 (2), pp. 355–373.

Ang, A. / Hodrick, R. J. / Xing, Y. / Zhang, X. (2006): The cross-section of volatility and expected returns, in: *The Journal of Finance*, 61 (1), pp. 259–299.

Balakrishnan, K. / Bartov, E. / Faurel, L. (2010): Post loss/profit announcement drift, in: *Journal of Accounting and Economics*, 50 (1), pp. 20–41.

Banz, R. W. (1981): The relationship between return and market value of common stocks, in: *Journal of Financial Economics*, 9 (1), pp. 3–18.

Basu, S. (1977): Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis, in: *The Journal of Finance*, 32 (3), pp. 663–682.

Bhandari, L. C. (1988): Debt/equity ratio and expected common stock returns: Empirical evidence, in: *The Journal of Finance*, 43 (2), pp. 507–528.

Bondt, W. F. M. D. / Thaler, R. (1985): Does the stock market overreact?, in: *The Journal of Finance*, 40 (3), pp. 793–805.

Breiman, L. (2001): Random forests, in: *Machine Learning*, 45 (1), pp. 5 – 32. Cited by: 78308.

Butt, H. A. / Högholm, K. (2018): The impact of illiquidity risk for the nordic markets, in: *Forthcoming in Spanish Journal of Finance and Accounting*.

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

Cooper, M. J. / Gulen, H. / Schill, M. J. (2008): Asset growth and the cross-section of stock returns, in: *The Journal of Finance*, 63 (4), pp. 1609–1651.

Datar, V. T. / Y. Naik, N. / Radcliffe, R. (1998): Liquidity and stock returns: An alternative test, in: *Journal of Financial Markets*, 1 (2), pp. 203–219.

Davis, J. L. / Fama, E. F. / French, K. R. (2000): Characteristics, covariances, and average returns: 1929 to 1997, in: *The Journal of Finance*, 55 (1), pp. 389–406.

Davydov, D. / Tikkanen, J. / Äijö, J. (2017): Magic formula vs. traditional value investment strategies in the finnish stock market, in: *Nordic Journal of Business*, 65 (3-4), pp. 38–45.

Diebold, F. X. / Mariano, R. S. (1995): Comparing predictive accuracy, in: *Journal of Business & Economic Statistics*, 13 (3), pp. 253–263.

Drobetz, W. / Otto, T. (2021): Empirical asset pricing via machine learning: evidence from the european stock market, in: *Journal of Asset Management*, 22 (7), pp. 507–538.

Fama, E. F. / French, K. R. (1993): Common risk factors in the returns on stocks and bonds, in: *Journal of Financial Economics*, 33 (1), pp. 3–56.

Fama, E. F. / French, K. R. (2012): Size, value, and momentum in international stock returns, in: *Journal of Financial Economics*, 105 (3), pp. 457–472.

Fama, E. F. / French, K. R. (2015): A five-factor asset pricing model, in: *Journal of Financial Economics*, 116 (1), pp. 1–22.

Fama, E. F. / MacBeth, J. D. (1973): Risk, return, and equilibrium: Empirical tests, in: *Journal of Political Economy*, 81 (3), pp. 607–636.

Fieberg, C. / Metko, D. / Poddig, T. / Loy, T. (2023): Machine learning techniques for cross-sectional equity returns'prediction, in: *OR Spectrum*, 45 (1), pp. 289–323.

Freyberger, J. / Neuhierl, A. / Weber, M. (2020): Dissecting characteristics nonparametrically, in: *The Review of Financial Studies*, 33 (5), pp. 2326–2377.

George, T. J. / Hwang, C.-Y. (2004): The 52-week high and momentum investing, in: *The Journal of Finance*, 59 (5), pp. 2145–2176.

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

Grobys, K. / Huhta-Halkola, T. (2019): Combining value and momentum: evidence from the nordic equity market, in: *Applied Economics*, 51 (26), pp. 2872–2884.

Gu, S. / Kelly, B. / Xiu, D. (2020): Empirical asset pricing via machine learning, in: *The Review of Financial Studies*, 33 (5), pp. 2223–2273.

Hanauer, M. X. / Kalsbach, T. (2023): Machine learning and the cross-section of emerging market stock returns, in: *Emerging Markets Review*, 55, pp. 101022.

Hanauer, M. X. / Windmüller, S. (2023): Enhanced momentum strategies, in: *Journal of Banking & Finance*, 148, pp. 106712.

Haugen, R. A. / Baker, N. L. (1996): Commonality in the determinants of expected stock returns, in: *Journal of Financial Economics*, 41 (3), pp. 401–439.

Ince, O. S. / Porter, R. B. (2006): Individual equity return data from thomson datastream: Handle with care!, in: *Journal of Financial Research*, 29 (4), pp. 463–479.

Jacobs, H. / Müller, S. (2020): Anomalies across the globe: Once public, no longer existent?, in: *Journal of Financial Economics*, 135 (1), pp. 213–230.

Jegadeesh, N. (1990): Evidence of predictable behavior of security returns, in: *The Journal of Finance*, 45 (3), pp. 881–898.

Jegadeesh, N. / Titman, S. (1993): Returns to buying winners and selling losers: Implications for stock market efficiency, in: *The Journal of Finance*, 48 (1), pp. 65–91.

Jokipii, A. / Vähämaa, S. (2006): The free cash flow anomaly revisited: Finnish evidence, in: *Journal of Business Finance & Accounting*, 33 (7-8), pp. 961–978.

Lakonishok, J. / Shleifer, A. / Vishny, R. W. (1994): Contrarian investment, extrapolation, and risk, in: *The Journal of Finance*, 49 (5), pp. 1541–1578.

Leivo, T. H. / Pätäri, E. J. (2011): Enhancement of value portfolio performance using momentum and the long-short strategy: The finnish evidence, in: *Journal of Asset Management*, 11 (6), pp. 401–416.

Lewellen, J. (2015): The cross-section of expected stock returns, in: *Critical Finance Review*, 4 (1), pp. 1–44.

Lintner, J. (1965): The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, in: *The Review of Economics and Statistics*, 47 (1), pp. 13–37.

Moskowitz, T. J. / Grinblatt, M. (1999): Do industries explain momentum?, in: *The Journal of Finance*, 54 (4), pp. 1249–1290.

Newey, W. K. / West, K. D. (1987): A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, in: *Econometrica*, 55 (3), pp. 703–708.

Novy-Marx, R. (2012): Is momentum really momentum?, in: *Journal of Financial Economics*, 103 (3), pp. 429–453.

Palazzo, B. (2012): Cash holdings, risk, and expected returns, in: *Journal of Financial Economics*, 104 (1), pp. 162–185.

Sharpe, W. F. (1964): Capital asset prices: A theory of market equilibrium under conditions of risk, in: *The Journal of Finance*, 19 (3), pp. 425–442.

Tobek, O. / Hronec, M. (2021): Does it pay to follow anomalies research? machine learning approach with international evidence, in: *Journal of Financial Markets*, 56, pp. 100588.

Tsang, W. W. H. / Chong, T. T. L. (2009): Profitability of the on-balance volume indicator, in: *Economics Bulletin*, 29, pp. 2424–2431.

Declaration of Academic Integrity

Hereby, I declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

This paper has neither been previously submitted to another authority nor has it been published yet.

Munich, 26.03.2024

Signature