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 17, 2024

Contents

List of Figures

List of Tables

1 Introduction

In 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 of Kingdom of Bavaria. In 1808 Finnish war broke between Kingdom of Sweden and the Russian Empire which would ultimately lead to establishment of autonomous Grand Duchy of Finland. It would still take more than 100 years for Finland to gain its independence. 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. 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 a question how to price these assets.

Major breakthrough in this topic happened in the sixties when capital asset pricing model was first developed (Sharpe, 1964; Lintner, 1965). In the eighties performance of capital asset pricing model was questioned and scholars started to come came up with variables that could explain portions of cross-sectional stock returns that 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 (Basu, 1977; Bhandari, 1988; Banz, 1981). During these times machine learning gained large interest and artificial neural networks became popular. Next big step in empirical asset pricing happened when Eugene Fama and Kenneth French combined selection of stock market factors to coherent three factor asset pricing model (Fama & French, 1993). Few years later researcher made breakthroughs to overcome limitations of the generalization of the decision trees, by ensembling multiple randomized trees, which would ultimately lead to introduction of machine learning method called random forest (Ho, 1995; Breiman, 2001).

Although three factor model was remarkable improvement compared to capital asset pricing model, it was still not able to explain variation in stock returns completely. Even after three factor model lot of new stock market factors have been discovered and in year 2015 Fama and French extended their three factor model by two additional factors (Fama & French, 2015). Characteristic for empirical asset pricing literature in recent years has been large amount 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 researches 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.

Objective of this study is to apply set of machine learning methods to well established asset pricing factors to capture abnormal stock return patterns. This study will focus on four Nordic

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

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 European union. Additionally, 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 emphazised also by Fama and French (2012), but this study focuses possibly even more integrated Nordic markets. 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 (Leivo & Pätäri, 2011). It refers to investor 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%[2]. Given the remote location of Nordic stock markets and their high share of foreign ownership it is likely that Nordic countries could be subject to periphery effect. Which again can result abnormal return patterns.

This study contributes to the existing literature in several ways. First of all it applies machine learning framework from Gu, Kelly and Xiu (2020) to a new market. Machine learning approach has been previously applied to European markets, but as mentioned this study focuses on even more coherent nordic submarket (Drobetz & Otto, 2021; Fieberg, Metko, Poddig, & Loy, 2023). Objective is to examine if investor investing only in Nordic markets could benefit from implementing the machine learning framework of Gu et al. (2020). Dataset of the study is unique in a sense that lot of machine learning approach studies have been conducted in more 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 sufficient amount of data to train complex machine learning model, but also allows us to focus on homogenous clearly defined submarket.

Additionally, existing stock market anomaly literature focusing solely to 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. Meaning that this study cannot

[2] Butt and Hogholm (2020) calculate share of foreign ownership from IMF Coordinated Direct Investment Survey CDIS data. Foreign ownership share of Butt and Hogholm is 52% for Denmark, 42% for Finland, 35% for Norway and 56% for Sweden.

only reveal the profitability of machine learning framework in Nordic market setting, but it can also reveal evidence of 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 the time. Since this study includes 23 anomalies simultaneously, it allows us also to examine performance of already discovered Nordic anomalies while controlling for many other variables. Applying sophisticated machine learning models allows us to also to control for more complex interactions.

Objective of this study is slightly more ambitious than in existing Nordic stock market anomaly literature. Existing literature mainly examines existence of anomalies by uni- or multivariate portfolio sorts. Studies predefine variables of interest and form portfolios based on these variable. 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.

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

Structure of this paper goes as follows. In second chapter introduction to related previous literature is provided. In this chapter performance of different methods and persistence of different anomalies in different regions is discussed. Third chapter introduces the data used in this study and filters applied to the data. 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. Fifth chapter focuses on benchmarking factors, showing how benchmark factors are constructed and how well they behave in Nordic markets. 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 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 majority of asset pricing studies. Despite the dominance of US markets in capturing the attraction of the academics, lot of empirical asset pricing literature has been conducted in international setting 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. Following chapter provides an overview for pioneering asset pricing anomaly literature. Focus will mainly be on literature on 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 interesting perspective for this study since in many of them Nordic countries are included.

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

2.1 US stock market anomalies

Many of the recent cross-sectional stock return studies use framework of Lewellen (2015) as base model. He runs 10-year rolling Fama-MacBeth regressions using lagged firm characteristics to predict out-of-sample stock returns. He studies cross sections of US stock return between 1964 and 2013 using different model settings up to 15 company characteristics. He finds strong positive correlation between expected returns derived from rolling Fama-MacBeth regressions and realised returns. Additionally, Lewellen shows that spread between realised return of portfolio formed from stock with lowest expected returns and portfolio with 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 statistical power in explaining monthly returns using lagged variables.

Gu, Jelly and Xiu (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. Gu et al. use 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. Out of these gradient boosted regression trees and neural networks[3] explain the monthly out-of-sample stock return the best reaching out-of-sample R^2 of 0.33% and 0.44% correspondingly whereas three factor OLS model introduced by Lewellen (2015) only reaches out-of-sample R^2 of -3.46%.

Similar to Lewellen (2015) Gu et al. construct portfolios based on 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[4]is 0.94%, 1.62% and 2.12% for models based on OLS, random forest and three layer neural network correspondingly. Gu et al. also show that all methods they examine show somewhat similar patterns on variable importance on return predictability. Most important factors are price trends such as momentum followed by stock liquidity, stock volatility, and valuation ratios.

2.2 European stock market anomalies

As mentioned, 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 beneficial to have a look on the European studies. Tobek and Hronec (2021) study machine learning based anomaly strategies in international setting. Their study includes 153 different equity anomalies and they only include anomalies to their data after documented discovery of corresponding anomaly. This way they can mimic the information set 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.

Similar to Gu et al. (2020) in US, Tobek and Hronec find that strategy using neural networks provides highest returns on quintile long-short portfolios. Mean return for neural network long-short portfolio in Europe was 0.7%. Interestingly penalized weighted least square method provided mean return of 0.65% which is higher than return of random forest based portfolio's return of 0.40%. Tobek and Hronec find that Industry momentum, lagged momentum, liquidity shocks, 52 week high, book-to-market value and return on equity are the most important variables for neural networks mode[5].

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 on December 2020 and spans from 1990 to 2020[6]. Drobetz and Otto examine performance of ordinary least squares, penalized least squares, principal component regressions, partial least squares, random forests, gradient boosted regression trees and neural networks on

[3] Gu et al. (2020) use five different settings of neural networks differing by number of hidden layers. Neural network with three hidden layers reaches the highest R^2_{oos} and is reported here.

[4] Portfolio returns are average value weighted returns.

[5] Tobek and Hronec (2021) discover possibilities training models either only using historical data from 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 difference between model trained on US data and local data are small.

predicting monthly stock level returns exploiting a set of 22 predictions, 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 large number of explanatory variables simple linear regression is not able to explain well out-of-sample stock returns.

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

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)[7]. Nevertheless, they choose a slightly different approach where instead of including 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 benchmark factor set of this study discussed in Section 5. Their conclusion endorses findings of Drobetz and Otto (2021) and Tobek and Hronec (2021) as they shown that more complex machine learning models beat linear approach in terms of both economic and statistical performance.

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 objective than this study. Studies show the existence of the anomalies by constructing a portfolio heavily weighted on certain factor. Nevertheless, they do not describe the magnitude of the relationship between the factor and the expected stock return. This study has slightly more ambitious objective and tries to derive return expectations from predefined stock market factors. This literature review serves as starting point for choosing most promising stock market factors that have already been studied.

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 main lists of Danish, Finnish, Norwegian and Swedish stock exchanges between 1991 and 2017.

[6] Finland is the only country included in the study of Drobetz and Otto (2021) that is also included in this study, since it is the only country belonging to Eurozone.

[7] Dataset of Fieberg et al (2023) contains Denmark, Finland, Norway and Sweden.

Grobys and Huhta-Halkola measure value with book-to-market value and momentum with past 12-month total shareholder return. Grobys and Huhta-Halkola show that momentum effect exists in Nordics markets and profitability of momentum strategy is not related to 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 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.

Value premium has shown consistency in Finnish stock markets. Davydov, Tikkanen and Äijö (2017) examine profitability of different value investing strategies between 1991 and 2013. Davydov et. al. investigate set of value indicators which included earnings to price, book to price, cash flow to price, dividends to price and earnings before income and taxes to enterprise value ratios. Additionally they test performance of investing strategy developed by Greenblatt (cite here) where portfolios are formed based on combined ranking of company's return on invested capital and earnings before income and taxes to enterprise value ratio. They 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 (?).

Similar to Grobys and Huhta-Halkola (2019) Leivo and Pätäri (2011) combine value anomaly with momentum anomaly in Finnish stock market for data set between 1993 and 2008. They show that two step portfolio sort that first allocates stocks to three portfolios based on their value indicators and subsequently based on momentum indicator can capture extraordinary stock returns. Leivo and Pätäri show that including momentum further increases returns of already recognised value sorting. Strategy performs even better when authors allow for long position in high value high momentum portfolio and short position on low value low momentum portfolio. Excess returns resulting from the two-fold portfolio construction can not be explained by CAP-model or two factor model including also size factor. It is not a surprise that value and momentum premium show existence in Nordic markets. Value and momentum anomalies are among the most well documented factors showing persistence in multiple cross-sectional studies (e.g, Gu, Jelly and Xiu (2020), Lewellen (2015), Drobetz and Otto (2021), Tobek and Hronec (2021)).

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 in US. Therefore, it is reasonable to ask whether liquidity of the stock could be driving factor of the stock returns. Impact of illiquidity risk to stock returns in Nordic market setting has been studied by Butt and Hogholm (2020). Butt and Hogholm test variety of different illiquidity measures and find that dollar zero returns is the most profitable illiquidity anomaly measure across all four Nordic market. Dollar zero return measurement is calculated by dividing number of days stocks return in US dollars is zero by total number of trading days. Butt and Hogholm construct five quintile portfolios based on liquidity of the stocks with data spanning from 1988 to 2013. They show that in all Nordic markets there exists large illiquidity premium as annual difference in equal weighted return of most illiquid portfolio and least illiquid portfolio is more than 18% for Finland, Norway and Sweden. For Denmark premium is slightly smaller 8.8%.

Jokipii and Vähämaa (2006) investigate free cash flow anomaly in Finnish stock markets between 1992 and 2002. They construct portfolios from stocks listed in Finnish stock exchange based on predefined thresholds for free cash flow ratios. These ratios include market value to free cashflow and total debt to free cashflow ratios. High free cashflow portfolio yields higher returns than market on average and the excess returns can not be completely explained by weightings in Fama and French (1993) risk factors.

3 Data

This section provides an overview of the dataset used in this study. Section starts by introducing overall market characteristics in Nordic stock markets. 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. 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 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

Main datasource for this study is Thomson Datastream. Company fundamentals data is collected from Worldscope database. Dataset spans from 1990 January to 2022 December which is shorter than in many previous studies. Reason why time period is limited to 1990 is that the amount of publicly listed companies in Nordic markets was rather low in the 1980's and finding reliable data for period before 1990 is difficult. 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, dataset is not subject to survivorship bias.

Table A.1 in 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. Therefore, several static and dynamic screens are applied to the data. Static screens include filtering non-equity securities, securities that are not listed in respective country and securities that are quoted in currency other than respective country's currency. Panel A from Table A.3 shows which values are accepted for 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 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 Ince and Porter (2006) and Hanauer and Windmüller (2023).

As mentioned Ince and Porter (2006) argue that data quality issues in Datastream could even lead to wrong conclusions. In order to avoid results to be driven by extraordinary datapoints, 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 return. Observations are also removed in case of extremely strong strong return reversals.

One characteristic has to be taken into consideration when working with data from Datastream. In case company is delisted for some reason Datastream returns last available value for remaining periods in the query. In order to only include actively traded securities these observations have to be cleaned from the dataset. This 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 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 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. Therefore, all consecutive zero returns at the end of the dataset are removed for all companies.

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 fraction of total market capitalization. Liquidity of these companies is often also quite low. To avoid results to be driven by such a stocks, approach of Hanauer and Kalsbach (2023) for emerging markets is applied and companies with smallest market value that account 3% of the aggregated market value are excluded. On the other hand we do not want few extremely large companies to drive the results either. Therefore, market value of the companies is winsorized monthly to 99%. If company's market value is among 1% biggest market values in

Table 1: Country summary statistics
Table provides summary statistics for pooled Nordic market and separate country specific Nordic markets. Minimum number of companies tells the amount of companies included to the data set in a month that the value was lowest for respective country. Maximum number of companies tells the amount of companies included to the data set in a month that the value was highest for respective country. Mean number of companies is the time series average of 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 market values of respective country in each month. All marked values are converted to USD. Only companies in the final dataset are included in calculation of the figures. Micro stocks are excluded from the dataset. 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	2399.78	810.32	141657.9
Finland	26	83	62	186	1893.43	634.47	124389.2
Norway	44	132	79	408	1520.75	506.17	124689.5
Sweden	45	256	132	593	2115.65	616.59	308952.1
Nordic	200	527	343	1422	1946.46	583.02	699688.8

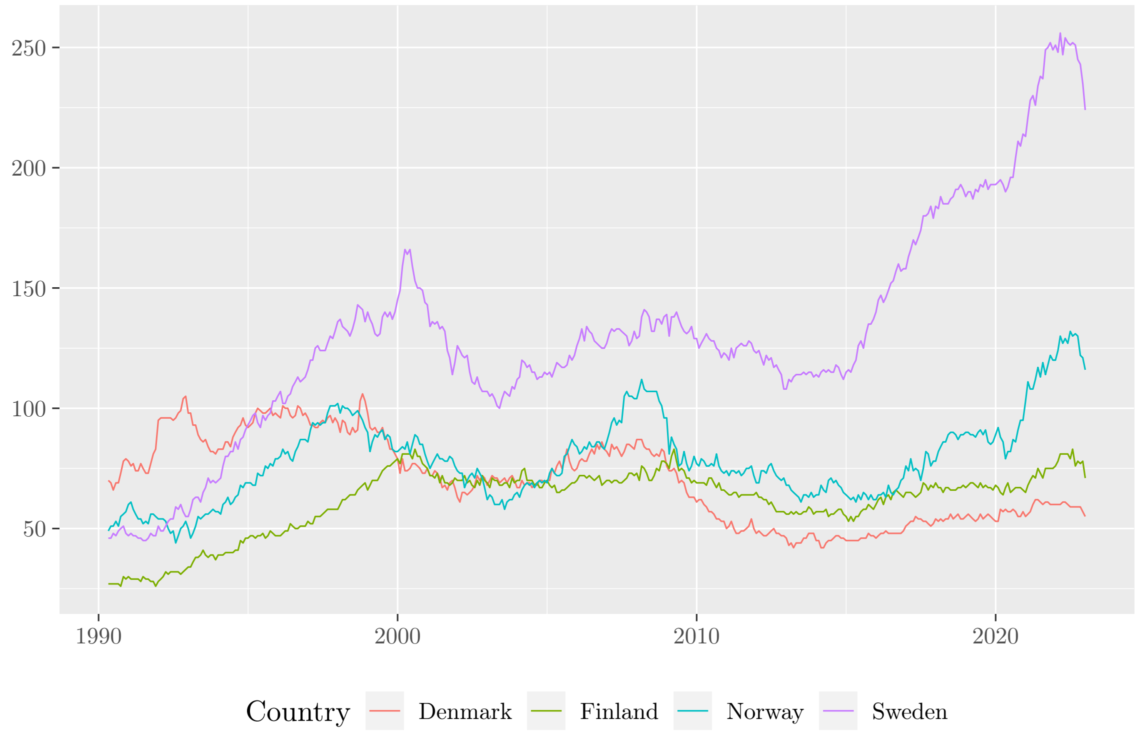
corresponding month, the market value will be replaced by the 1% threshold value.

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

Sweden is clearly the biggest of the four Nordic markets both in regards off 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 less than half of the total market value of the pooled Nordic market. In regards of 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 more practical reason why Nordic markets are pooled in this study. Table 1 shows that individual Nordic stock markets hold limited amount of large market capitalization stocks. This leads to situation where the performance of the whole market, or portfolios formed from the market,

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



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 and we want to ensure that there exist 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 common currency which in this case is 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. Majority of the explanatory variables are some sort of ratios which can be directly calculated from the local currency values.

3.2 Company characteristics

Total 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. 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 $t+1$. Detailed descriptions how each of these variables is calculated is provided in Table 2. Table 2 also provides corresponding Datastream items used in the calculation of the variables.

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, ratio between price and 52-week high price, beta coefficient, idiosyncratic volatility, turnover and on-balance volume. Standard deviation is calculated as rolling 52-week standard deviation of the stock returns. Nevertheless, also these variables are updated only monthly and therefore these variables are noted as having monthly frequency in Table 2.

Table 2: Variable definitions
Tables 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. Table also provides the direct formulas and relevant Datastream items used to calculated 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 end of December in year $t-1$. Frequency of the variable is indicated after the variable name.

Variable	Author	Definition
Cash-to-Assets *Yearly*	Palazzo (2012)	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)	Capital turnover is calculated by dividing total sales by one year lagged total assets. CTO$_t$ = WC01001$_t$ / WC02999$_{t-1}$
Investment *Yearly*	Cooper, Gulen and Schill (2008)	Investments are defined as a yearly change in total assets. INV$_t$ = (WC02999$_t$ - WC02999$_{t-12}$) / WC02999$_{t-12}$

Table continues in the next page . . .

... Continued from the previous page

Variable	Author	Definition and affected Datastream items
Book-to-Market Equity *Yearly*	Davis, Fama and French (2000)	Book-to-Market value is calculated by dividing company's book value of equity by company's market capitalization end of previous year. Book value of equity is calculated by summing common equity and deferred taxes of the company. $\text{BEME}_t = (\text{WC03501}_t + \text{WC03263}_t) / \text{MV}_{Dec}$
Cash Flow-to-Price *Yearly*	Lakonishok, Shleifer and Vishny (1994)	Cash flow to price ratio is calculated by dividing company's cash flow from operating activities by the asset's market capitalization end of previous year. $\text{CFP}_t = \text{WC04860}_t / \text{MV}_{Dec}$
Debt-to-Price *Yearly*	Bhandari (1988)	Debt-to-price value is calculated as difference between total assets and common equity divided by the asset's market capitalization end of previous year. $\text{DEBT}_t = (\text{WC02999}_t - \text{WC03501}_t) / \text{MV}_{Dec}$
Sales-to-Price *Yearly*	Lewellen (2015)	Sales-to-price ratio is calculated by dividing total sales by asset's market capitaliztion end of previous year. $\text{SP}_t = \text{WC01001}_t / \text{MV}_{Dec}$
Earnings-to-Price *Yearly*	Basu (1977)	Earnings-to-price ratio is calculated by dividing net income before extra Items and preferred dividends by asset's market capitalization end of previous year. $\text{EP}_t = \text{WC01551}_t / \text{MV}_{Dec}$
Return-on-Assets *Yearly*	Balakrishnan, Bartov and Faurel (2010)	Return-on-assets is calculated as net income before extra items and preferred dividends divided by one year lagged total assets. $\text{ROA}_t = \text{WC01551}_t / \text{WC02999}_{t-12}$
Return-on-Equity *Yearly*	Haugen and Baker (1996)	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 definition of book value of equity. $\text{ROE}_t = \text{WC01551}_t / \text{BE}_{t-12}$
Tobin's Q *Yearly*	Freyberger, Neuhierl and Weber (2020)	Tobin's Q is calculated by summing up total assets and market capitalization from previous December, then subtracting cash and short-term investments and deferred taxes. Finally result is divided by the total assets. $\text{Q}_t = (\text{WC02999}_t + \text{MV}_{Dec} - \text{WC02001}_t - \text{WC03263}_t) / \text{WC02999}_t$
Momentum_7 *Monthly*	Novy-Marx (2012)	MOM7 is defined as cumulative return in US dollars between $t-7$ and $t-12$ months.
Momentum_{12} *Monthly*	Jegadeesh and Titman (1993)	MOM12 is defined as cumulative return in US dollars between $t-2$ and $t-12$ months.
Momentum_{36} *Monthly*	De Bondt and Thaler (1985)	MOM36 is defined as cumulative return in US dollars between $t-12$ and $t-36$ months.
Momentum_2 *Monthly*	Jegadeesh (1990)	MOM2 is defined as prior month return in US dollars.
Industry Momentum *Monthly*	Moskowitz and Grinblatt (1999)	MOM.IND is defined as 12 month cumulative equal weighted industry return. Industry is defined using INDG attribute from Datastream.

Table continues in the next page ...

... Continued from the previous page

Variable	Author	Definition and affected Datastream items
Standard deviation *Monthly*	Ang, Hodrick, Xing and Zhang (2006)	L.SD is defined as standard deviation of unadjusted weekly price for last 52 weeks.
52-week high price *Monthly*	George and Hwang (2004)	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)	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)	L.IDVOL is estimated by standard deviation of regression residuals from regressing unadjusted weekly US dollar returns by equally weighted market return.
Log. market value *Monthly*	Banz (1981)	Natural logarithm of the market value of the company end of previous month. L.LOG.USD.MV$_t$ = ln(USD.MV$_{t-1}$)
Turnover *Monthly*	Datar, Naik and Radcliffe (1998)	Turnover is defines as end of previous month weekly trading volume divided by the shares outstanding. L.TO$_t$ = VO$_{t-1}$ / NOSH$_{t-1}$
On-balance volume *Monthly*	Tsang and Chong (2009)	L.OBV is calculated with following process. First on-balance volume is the weekly trading volume multiplied by corresponding week return's sign. Following on-balance volumes are calculated by adding the product of trading volume and sign of return to previous on-balance volume.

Set of explanatory variables includes seven value indicators. Book-to-market value is calculated by dividing 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 divided by market capitalization of previous year December in order to obtain earnings-to-price, cashflow-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 previous year's December. Rest two of the value indicators are normalized by the total assets. Cash-to-total assets 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 previous December, then subtracting cash and short-term investments and deferred taxes and finally dividing by the total assets.

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. 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 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 $t-12$. Final momentum indicator is the industry momentum. Industry momentum is defined as 12-month cumulative equal weighted return of an industry sector. Industries are defined by the INDG Datatsream attribute.

Table 3: Descriptive statistics
Table provides time series averages of cross-sectional means and standard deviations of all variables used in this study. Values are reported separately for pooled Nordic market and four Nordic markets Denmark, Finland, Norway and Sweden. EXC.RET is the monthly excess return calculated from total return index noted in US dollars. Risk free rate used to calculate excess returns is US dollars one-month Treasury bill rate. 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	0.088	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	3.330	6.456	2.353	4.443
SP	1.585	2.279	1.356	1.636	2.066	2.105	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	0.088	0.224	0.095	0.203	0.103	0.158	0.058	0.280	0.093	0.212
Q	0.696	0.326	0.597	0.391	0.735	0.286	0.655	0.345	0.759	0.273
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	0.171	0.409	0.190	0.387
MOM_{36}	0.397	0.751	0.400	0.679	0.363	0.625	0.357	0.797	0.432	0.754
MOM_2	0.015	0.093	0.013	0.081	0.014	0.085	0.015	0.097	0.017	0.093
MOM.IND	1.144	0.284	1.132	0.248	1.142	0.285	1.148	0.293	1.148	0.279
L.SD	0.047	0.030	0.045	0.028	0.042	0.028	0.051	0.030	0.051	0.029
L.HIGH52.RATIO	0.684	0.284	0.722	0.266	0.621	0.294	0.684	0.271	0.695	0.261
L.BETA	0.863	0.525	0.703	0.391	0.732	0.495	0.909	0.530	0.999	0.509
L.IDVOL	0.046	0.030	0.044	0.025	0.040	0.028	0.051	0.032	0.048	0.029
L.LOG.MV	6.331	1.340	6.400	1.328	6.402	1.272	6.124	1.246	6.452	1.414
L.TO	0.043	0.109	0.053	0.126	0.025	0.065	0.028	0.073	0.056	0.119
L.OBV	0.170	0.507	0.151	0.544	0.098	0.390	0.174	0.529	0.216	0.524

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. Market return is constructed as equal weighted weekly market return following Green, Hand and Zhang (2017). Because the returns are noted in US dollars one-month Treasury bill rate, which is obtained from Kenneth French's database[8], is used as a risk free rate proxy. The regression is run for each company separately for each month on rolling basis. For each regression up to three years of weekly historical data is considered, but minimum

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 measured by turnover, standard deviation, market capitalization and 52 week high price. Turnover is calculated by dividing weekly trading volume by number of shares out standing. Standard deviation is calculated from up to 52 weeks of weekly unadjusted price data. One-month lagged logarithm of market value is used as a size indicator. 52 week high indicator is also calculated from weekly data and it is defined as ratio between highest unadjusted weekly price in last 52 weeks and current price. 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 technical trading indicator is included to explanatory variable set. On-balance volume is often used as time series predictor for individual securities, but objective of this study is to investigate whether it could contain information about cross section of 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 corresponding week's return. Then this product is added to the cumulative sum of the historical on previous on-balance volumes. On-balance volume is defined as

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

where OBV_t is the on-balance volume value at time t, OBV_{t-1} is the on-balance volume value at $t-1$, trading volume is the weekly trading volume and r_t is the return of the corresponding stock at time t.

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

Machine learning algorithms can be sensitive to outliers in the data. Therefore, all explanatory variables are winsorized between 1st and 99th percentiles. In case value of the certain variable is less than 1st percentile of the corresponding months values it will be replaced by the 1st

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

percentile threshold value. In case value of the certain variable is above 99th percentile of the corresponding months values it will be replaced by the 99th percentile threshold value. Additionally, any missing value in explanatory variables will be replaced by zero.

Table 3 provides descriptive statistics of company characteristics. For each explanatory variable time series average and standard deviation of the cross-sectional mean is reported. Values are reported for pooled nordic markets as well as individual markets. Table shows that mean monthly excess return of Nordic markets during the studied period was 0.7%. Mean excess return of Sweden and Finland is slightly above that and mean excess return of Denmark and Norway is slightly below that. Additionally, Figure C.4 in 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 clear change in the mean value can be seen during financial crisis.

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

4 Methodology

This section provides the theoretical framework of this study. 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, following section describes how stock return predictions obtained from different models are evaluated. Section describes both prediction accuracy and economic profitability metrics used to evaluate the performance of the models.

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. 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

Benchmark model of this study is the Fama-MacBeth (1973) regression. This method follows the approach of Lewellen (2015). First step of the method is to run rolling cross-sectional regressions with lagged variables. Second step of the method calculates means of the factor loadings obtained from the cross-sectional regressions. Finally expected stock return can be obtained by multiplying the mean factors loading with latest available stock characteristics. Below formulas show the generalized notation of the model

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

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

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

In 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 to three sub-samples for separate validation of hyperparameters and testing. To obtain the expected return mean of 120 historical regression coefficients is calculated. Nevertheless, implemented linear model is not the simplest linear model. One variable in implemented Fama-MacBeth model that could be treated as hyperparameter is the rolling window. For example Lewellen (2015) reports results also for alternative rolling windows in addition to rolling window of 120 months. Despite the possibility for hyperparameter optimization, this study uses predefined rolling window of 120 moths, which is also the rolling window in main focus of Lewellen (2015).

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 for next 12 months. Each month most recent information is just inserted to the model. Computing requirements for linear model is far lesser than for non-linear models. Nevertheless, to ensure comparability between different models also the linear model is trained only once per year. 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 predicted return and last return used to train the model can grow up to 12 months. Since we use lagged variables, this means that for prediction of stock return t we alway use stock characteristics from $t-1$, but some factors are only updated yearly. To mimic information set investor would have had available in historical periods we have to account for the delay in reporting balance sheet information. Therefore timeline of Fama and French (1993) is followed and models are trained each year at end of June.

4.2 Random forest

Decision trees are one example of nonparametric machine learning algorithms. 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. Starting point of the decision tree is called a root node. At each iteration of 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 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. Final prediction of the regression tree leaf is the average of the dependant variable values of training set observations inside it. Gu et al. (2020) formulate prediction of a regression tree with K leaves as

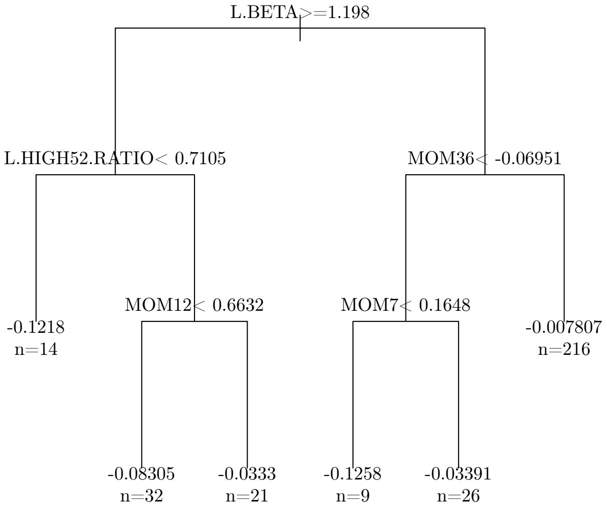
$$f(x_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k 1_{\{x_{i,t}\in C_K(L)\}} \tag{5}$$

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. Since observation can only belong to one leaf, partition $C_K(L)$ is the product of the above partitions.

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 aims to avoid this problem by deriving the predictions from ensemble of regression trees. As the name might suggest random forest consists of multiple decision trees.

Idea of the random forest is to randomly generate 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 the overfitting trees inside random forest should not be too correlated and this is ensured including randomness in the 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. Number of the variables model considers in each split as well as maximum depth of the decision tree and 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

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



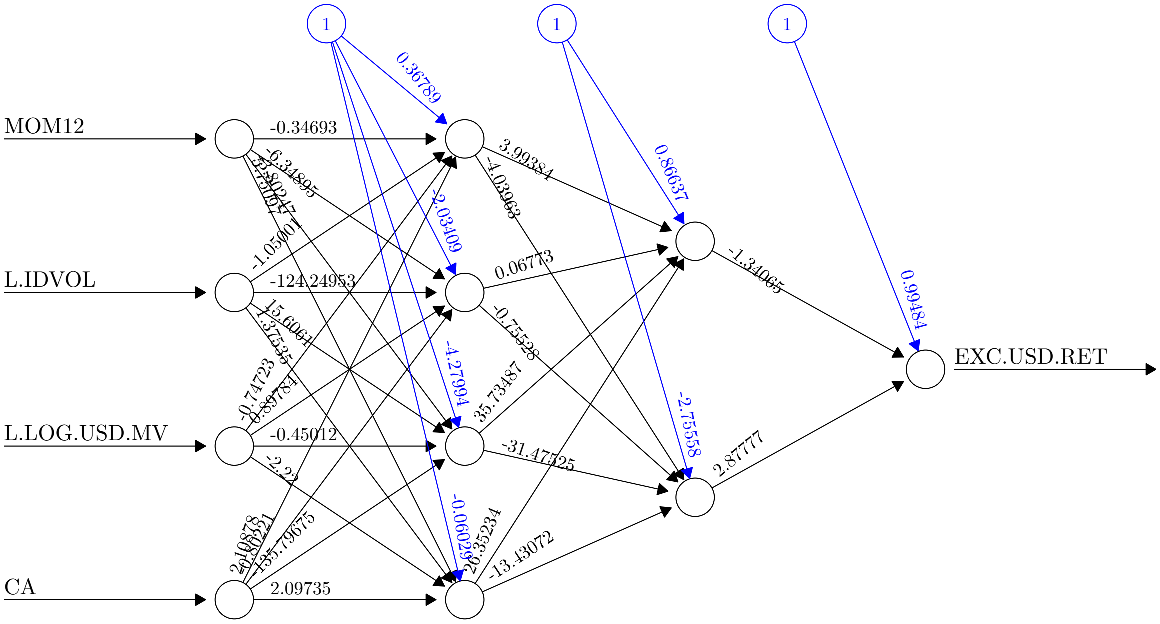
which require input from the user, but which also can be optimized for different tasks. Table 4 in Appendix shows which values were considered for each hyperparameter that were optimized for random forest.

4.3 Neural networks

Artificial neural networks are powerful machine learning method category. Currently neural networks are popular approach to many real world prediction issues. Due to their strong performance in multiple domains, neural networks are often considered as state of the art machine learning method. 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 parameter initialization. Some of the parameter, such as learning rate, are usually optimized during the training of the model while others such as architecture of the model are usually fixed.

One of the first things user has to decide while training a neural network is the architecture of the model. This study focuses on feedforward neural networks which consists of input layer, hidden layers and an output layer. Input layer consists of the predictive variables whereas output layer

Figure 3: Illustrative neural network
Figure serves only for illustrative purposes. Sole purpose of the figure is to visualize 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 narrower architecture was chosen for better visualization.



produces the final prediction. In between thee exist 1 to N hidden layers. Hidden layers again consists of so called neurons. Similar to number of hidden layers, user has to also decide 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 lot of previous literature simultaneously examine multiple different architectural forms, due to computing capacity in this study only one architecture will be examined (Gu et al., 2020; Hanauer & Kalsbach, 2023; Tobek & Hronec, 2021). Neural network of this study has two hidden layers. First hidden layer has 16 neurons and following 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). In order to improve and fasten the converging of the model batch normalization is implemented between all layers.

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

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

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

As mentioned neural networks include numerous hyperparameters that can be optimized during the training of the model. Training neural network is computationally demanding. Due to limited computing capacity hyperparameters are not optimized in this study, but predefined values are used. Hyperparameters and their values are presented in Table 4. 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 validation set. Instead of inserting whole dataset to the model at once data is inserted to 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 gradient. This is done in repetitive iterations. In each iteration size of the change is defined by hyperparameter called learning rate. Since learning rate is a hyperparameter it needs an input from the user. It can also be optimized. Setting correct learning rate is crucial, since too big learning rate might prevent algorithm from converging to optimal solution, but too small learning rate makes converging slow. For above described reasons learning rate is not optimized in this study, but using learning rate scheduler it is adjusted during the training of the model. In order to ensure efficient training learning rate is set 0.01 in the beginning of the training and after ten epochs learning rate will be multiplied by 0.9.

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. Ensemble is implemented by training five separate models with different initial weights. Final prediction will be then 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. 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 set of statistical tests which are described in more detail below. This allows us to evaluate the relation between predicted and realized excess returns.

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-Mariano 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 i. Nevertheless, Gu et al. (2020) argue that the historical mean return is so noisy estimator that it underperforms compared to static estimation of zero and therefore artificially improves the out-of-sample R^2. Instead they propose alternative out-of-sample R^2 measure where the squared sum of returns in the denominator is not demeaned.

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

R^2 presents the prediction accuracy as a single figure, whereas Diebold-Mariano allows for pairwise comparison of different models. Diebold-Mariano value is calculated as

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

where $\hat{r}_{i,t,1}$ is the return prediction of first model for company i at time t and $\hat{r}_{i,t,1}$ is the return prediction of the second model for company i at time t. N_t is number of observations in prediction period t. Therefore, $d_{12,t}$ is a time series of differences in average squared prediction errors between model 1 and model 2. $\overline{d}_{12}$ is the mean of $d_{12,t}$ and $\hat{\sigma}_{d_{12}}$ is the Newey and West (source) standard error of $d_{12,t}$. Diebold-Mariano test allows us to estimate statistical significance of the prediction accuracy of two models. Under assumption that the there does not exists difference in prediction accuracy between models Diebold-Mariano test statistic follows norma distribution with 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 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. 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. Magnitude of the beta coefficients can provide information of possible over or undershooting of the models.

Second perspective that is evaluated is the economic profitability of the methods. Profitability is estimated via portfolio construction following approach of Lewellen (2015). First expected returns are derived from each model. This process is introduced in more detail in following in 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 to the model. Therefore, also the portfolio allocation is repeated monthly. Each month all stocks are allocated to one of the ten expected return portfolio, but to avoid result to be mainly driven by small stocks approach from Hanauer and Kalsbach is applied (2023) 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 97% of the aggregated total market value of respective month.

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 return of the highest expected return portfolio and 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. 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

$$\text{Sharpe Ratio} = \frac{\overline{r_i}}{\sigma_i} \tag{11}$$

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

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

where Y_t stands for cumulative return from the beginning of the period until date t. In order to examine risk adjusted returns long-short portfolio returns will be regressed against Fama-French (2015) six factor model factors[9]. From this regression alphas, which can interpret 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

$$\begin{gathered} \hat{r}_{i,t} = \alpha + \beta_1\ RMRF_t + \beta_2 +\ SMB_t + \beta_3\ HML_t + \\ \beta_4\ CMA_t + \beta_5\ RMW_t + \beta_6\ MOM_t + \epsilon_t \end{gathered} \tag{13}$$

where $RMRF$ 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 between returns of stocks that had highest return in period $t-1$ and the stocks that had lowest return in period $t-1$. Factors are constructed from the same dataset as machine learning portfolios, except that the micro-cap stocks are not excluded. Construction of these factors is described in more detail in Section 5.1.

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 model generates excess returns but the monthly turnover remains large, it could lead to situation where after counting for transactions costs more passive strategy would be more profitable. Therefore, turnover provides valuable information of the practical implementability of the constructed machine learning models. For month t turnover is defined as

[9] Fama and French (2015) introduced the five factor model. Factors used to regress machine learning portfolio returns include five factor model factors and momentum factor from Carhart (1997).

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

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. 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 relative return in month t compared to the return of the corresponding portfolio.

4.5 Variable importance

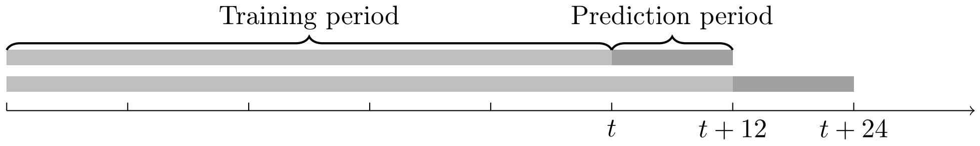
One challenge in dealing with various statistical methods is that they lack common metrics for explanatory inference. Many of the models have metrics for variable importance, but comparability of these metrics can be questioned. Therefore, approach of Gu et al. (2020) is implemented to define variable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

4.6 Sample splitting

It is common while training machine learning models to split the data to three sets. Training set will be used as the name suggests to train the model. In case machine learning model includes hyperparameters these can be optimized with validation set. Finally the true out-of-sample predictions can be performed for testing data. Because we want to mimic situation and information set of an investor we have to take into a consideration 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 rolling 10 year rolling scheme where they first train model using the first seven years of the data and then optimize hyperparameters using last three years of the rolling window. Finally they train the model with optimal hyperparameter initalization with whole ten year window to predict returns for the next year. Gu et al. (2020) use slightly different approach. Instead of using rolling window they increase the training window size after each training period by one year. Common for these two approaches is that they both train the model only once a year.

Figure 4: Sample splitting
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 next 12 months to predict the stock returns. Each year training period is extended by 12 months. Training period is further split into training data and validation data randomly allocating 80% of the data to training set and 20% of the data to the validation set. Validation set is used to optimize hyperparameters. Minimum length of the training period is 50 months.



The sample splitting scheme applied in this study is slightly different from above described ones. Above approaches use disjoint time periods to mimic the out-of-sample setting in the hyperparameter optimization. In this study training and validation set are separated from the testing data based on time. Approach is closer to approach of Gu et al. (2020) in a sense that training data window is increased each year. Nevertheless, the difference is that the data is distributed to training data and validation randomly instead of using the disjoint periods. Reason why this scheme is chosen is that we want to avoid the retraining of the model after the hyperparameter optimization which is necessary if the most recent data should be included to the model. Sample splitting scheme is illustrated in Figure **??**.

Table 4: Hyperparameters
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 stand 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 random forest model we actually optimize the hyperparameters. Therefore, training window is split to training and validation data so that 80% of the data is used in the training and 20% is assigned to the validation set. For 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. Approach of this study follows the common approach to only train the models once a year.

5 Benchmark factors

This sections serves as prerequisite for machine learning portfolios. In this section well established factors from Fama and French (2015) framework are constructed. These factors include market, size, value, investment, profitability and momentum factor. 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 risk adjusted performance of the different machine learning models. Factors also provide interesting insight to the existence of different traditional stock market factor anomalies in Nordic markets in a bit more traditional setting.

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

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. Since on compared to US markets Nordic markets have less companies with high market value, using NYSE breakpoints could lead to highly un-diversified 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 small part of the total market capitalization. Therefore approach of Fama and French (2012) is applied.

In the end of each June stocks are first distributed to two size portfolios. Companies with biggest market value that account 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 dataset used to construct benchmark factors is slightly different than the dataset used to train the machine learning models as it includes also the smallest stocks. Next stocks are allocated to three value, investment, profitability and momentum portfolios. For all of 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 indicator of value characteristic of a company. Book-to-market value is calculated as ratio between sum of common equity and deferred taxes and market capitalization on December $t-1$. Profitability is defined as net income before extra items/preferred dividends divided by the book equity of the company. Investment variable is calculated as annual change in total assets. Momentum is defined as cumulated return from $t-12$ to $t-2$. Returns are calculated using total return index that is converted to US dollars for comparability between different countries. Market value used in size allocation as well as to weight portfolio returns is also converted to US dollars.

Equation 15 shows formula for each factor. Abbreviation for each variable is derived from how they are calculated. Value factor is called high minus low (HML), profitability factors is called robust minus weak (RMW), investment factor is called conservative minus aggressive (CMA). Only momentum factor is exception to this rule and more intuitive naming is used. Portfolio allocation results six portfolios for value, investment, profitability and momentum factors and 24 two-fold size portfolios. After portfolio construction portfolio returns are calculated as difference on value weighted average returns on portfolios formed based on respective variable. E.g. value factor return is difference between average of value weighted returns of two high book-to-market portfolios and average of value weighted returns of two low book-to-market portfolios. Market factor is the average value weighted excess return of the whole market. 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 five factor (2015) model factors extended by the momentum factor in Nordic stock markets. 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 5: Benchmark factor summary statistics
Table presents the mean returns and standard deviations of the benchmark factors together with two-sided *t*-statistics and corresponding p-values. For each factor minimum and maximum monthly return is 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 average of value weighted return of two high value portfolios and average of value weighted return of two low value portfolios. RMW, CMA and MOM are calculated in similar manner, but portfolio sort are done based on investment, profitability 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. Risk free rate used to calculate excess returns is the US dollars one-month Treasury bill rate. Time period spans from January 1990 to December 2022.

	Mean	Std.	*t*-stat.	p-value	Min	Max
HML	0.0014	0.0022	0.6299	0.5291	-0.2662	0.2508
RMW	0.0013	0.0015	0.8711	0.3842	-0.1251	0.1640
CMA	0.0014	0.0015	0.9102	0.3633	-0.1077	0.1704
MOM	0.0090	0.0021	4.2990	0.0000	-0.1501	0.1828
SMB	-0.0001	0.0014	-0.1091	0.9132	-0.1204	0.1042
RMRF	0.0074	0.0032	2.3051	0.0217	-0.2576	0.2072

Table 5 provides the time series averages of the factor returns, standard deviation of the factor returns, corresponding t and p-values as well as monthly minimum and maximum returns for all six factors. From Table 5 it is clear that the momentum factor shows strongest performance measured both by the magnitude of the return as well as the statistical significance. Monthly momentum factor return is 0.9% with t-statistic of 4.3. Table B.1 in appendix shows the correlations between the factor returns. In Nordic markets correlation of momentum factor with other factors is only minor. Interestingly in Nordic markets momentum factor seems to negatively correlate with market factor.

Strong performance of the momentum factor is inline with previous literature. Many previous studies have documented excess momentum returns either in pooled or individual Nordic markets (e.g. Grobys and Huhta-Halkola (2019) and Leivo and Pätäri (2011)). Slightly more surprising is the poor performance of the value factor. Average return of the value factor is 0.14% and it is not statistically significant. Some of the previous studies document value premium 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.

Figure 5: Benchmark factor performance
Plot presents the cumulative return of the benchmark factors. RMRF is average value weighted excess return of pooled Nordic market. Portfolio returns are calculated based on 2 × 3 sorts on size and one other factor. HML is the difference in average of value weighted return of two high value portfolios and average of value weighted return of two low value portfolios. RMW, CMA and MOM are calculated in similar manner, but portfolio sort are done based on investment, profitability 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. Risk free rate used to calculate excess returns is the US dollars one-month Treasury bill rate.

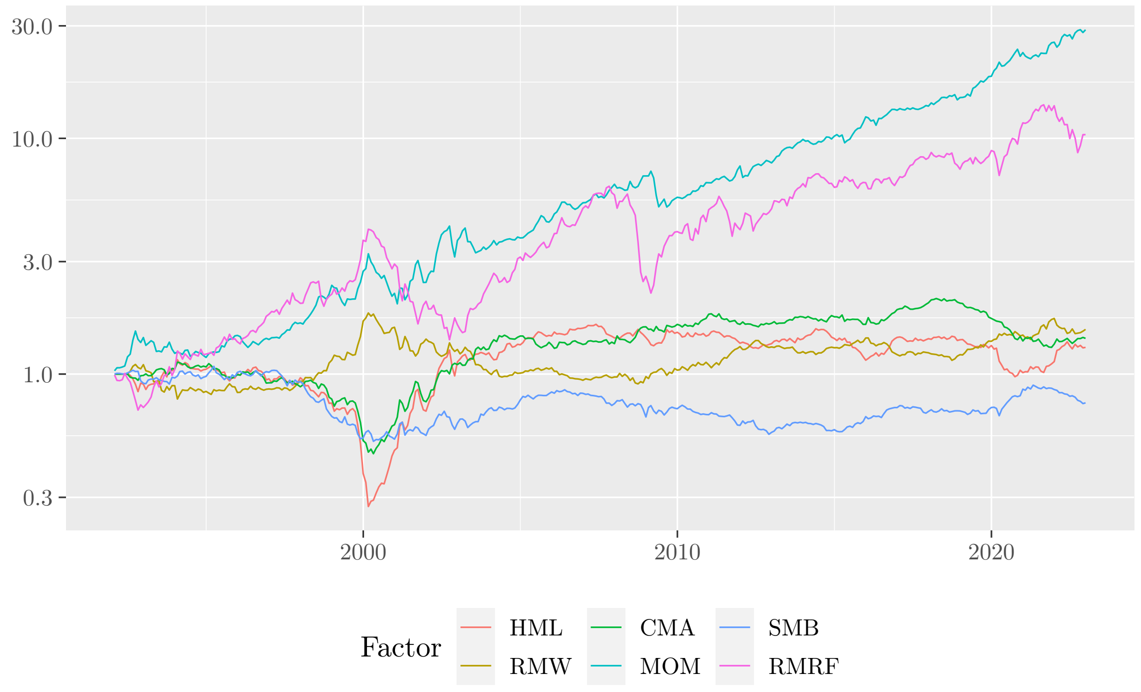


Figure 5 shows the historical performance of the benchmark factors. It shows that until 2008 even momentum factor could not exceed the market return, but still seemed to be less volatile. In post financial crisis period momentum factor has been clearly the strongest performing factor. Interestingly late nineties, which was strong period for market return, was extremely difficult period for value factor. Contrary in the early 2000's value factor performed strongly when market factor was decreasing steeply. Until 2003 performance of value factor is 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 through out 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 marker capitalization stocks. Nevertheless, negative return of size factor is small and not statistically significant. Performance of the benchmark factor indicate that momentum indicators could hold profitable information about Nordic stock market returns.

6 Empirical results on Nordic equities

Discuss optimal hyper parameters

Performance of the machine learning models will be evaluated from two aspects. Prediction accuracy is evaluated by the out-out sample R^2 and Diebold-Mariano tests. Additionally, in the spirit of Lewellen (2015) relationship between expected and realized excess returns are examined by regressing the realized excess returns with the individual stock level predicted returns. Then the economic profitability of the models will be evaluated by investigating performance of univariate expected return sort portfolios. Construction of these portfolios is explained in more detail in Section 4. Second aspect that is examined is the prediction accuracy of the machine learning models. Finally, the variable importance for different methods are calculated to see effect of each explanatory variable to the prediction accuracy of the model. Process to define variable importance is described in Section 4.5.

6.1 Prediction accuracy

Panel A of Table 6 presents the out-of-sample predictive accuracies for all models. Panel B presents the pairwise Diebold-Mariano statistics. It can clearly be seen from Table 6 that random forest model produces the most accurate out-of-sample predictions. Out of the three models it is the only one that produces positive out-of-sample R^2 value of 0.2% even with the more conservative definition where the benchmark model is prediction of zero. While linear and neural network models both produce negative out-of-sample R^2 of -1.95% and -0.05%, it can still be seen that neural network model performs better than the linear model in regards of the predictive accuracy.

Another interesting insight Table 6 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. 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 regards results are in line with findings of Fieberg et al. (2023).

Table 6: Prediction accuracy
Table presents the prediction accuracy metrics for different machine learning models. Panel A presents two out-of-sample R^2 values. First one uses zero prediction as a benchmark model. This means that the denominator in the calculation of the metric is squared excess return. 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 table presents the pairwise Diebold-Mariano statistics for all the methods. Bolded figure indicated significance at 5% level, whereas asterisk indicates significance at 5% level after three-way Bonferroni adjustment. FM stands for linear regression model, RF stand 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	0.0021	-0.0005
$R^2_{oos\ Trad.}$	0.0499	0.0492	0.0467

Panel B: Diebold-Mariano statistics

	FM	RF	NN
FM		**8.4462*** (0.0000)	**2.1495** (0.0316)
RF			**-10.0946*** (0.0000)

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

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

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

In this study machine learning models predict the stock returns between July 1994 and November 2022. Meaning that 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 strongest predictive accuracy with out-of-sample R^2 of 0.9%. During that time the out-of-sample R^2 of random forest model is 0.6% whereas predictive accuracy of the linear model remains negative with out-of-sample R^2 of -3.7%. 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 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 model measured by out-of-sample R^2 is partially inline with previous literature. Results are inline with findings of Drobetz and Otto's (2021) study in European stock markets in a sense that the linear model offers 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 for linear model. Results are also inline in a sense that random forest model shows strong predictive performance. Results of Fieberg et al. (2023) are slightly more contradictory, since they show positive out-of-sample R^2 also for linear model[10]. 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 of 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.

Where the results clearly deviate from previous literature is the predictive performance of the neural network model. In studies of Drobetz and Otto (2021) and Fieberg et al. (2023) neural network model produces highest, clearly positive, out-of-sample R^2 values. Naturally studies of Drobetz and Otto and Fieberg et al. are not directly comparable to this study since 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 also partially explain the relative poor performance of the neural network model, since usually neural network models require lot of data.

Panel B of Table 6 presents the pairwise Diebold-Mariano statistics for all the models. Calculation of the statistics is described in Section 4. Table 6 reports the Diebold-Mariano statistics

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

together with corresponding p-values. Bolding of the Diebold-Mariano figure imply significance in normal 5% level whereas asterisk implies more conservative 5% level which is Bonferroni adjusted for three-way comparisons. The three-way Bonferroni adjusted critical one-sided Diebold-Mariano value is 2.13.

Discuss the results.

Next the prediction accuracy is examined by following approach of Lewellen (2015) by regressing the realized excess returns by the return predictions from different models. Table 7 presents the summary statistics for these regressions. 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 7 to descriptive statistics in Table 3 shows that the mean expected return is really close to actual realized mean excess return for neural network and random forest model. Both mean expected and realized return are calculated as time series average 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 7 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 7: Expected return regression summaries
Table provides univariate properties of the return predictions for all models and summary statistics for regression where realized excess returns are regressed with expected returns. Mean and standard deviation are reported for expected returns. 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. 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 stand for random forest model and NN stands for neural networks model. 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	6.584	0.0003
RF	0.0074	0.0095	0.4055	0.0300	13.500	0.0015
NN	0.0068	0.0123	0.3719	0.0224	16.612	0.0023

Results from right side of Table 7 support the remarks from out-of-sample predictive performance and univariate properties. For all of the models there exists 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 7 should be one. Random forest model has the highest predictive slope of 0.41, which means that 1% change in expected return respond to 0.41% change in realized return. Neural network and linear models have slightly smaller slopes of 0.37 and 0.15 correspondingly. R^2 values from the regressions are also presented as 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 first two as the neural network and random forest perform better than linear model. With regression based R^2 neural network model is able to explain more variation realized excess returns than the random forest model.

Given that the mean predicted return matches quite well 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 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 6. Further insight for this will be provided in next section where 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. Approach attempts to mimic of information set of a historical investor. Section describes the historical realized returns for an investment strategies build based on the machine learning models. First part of the section mainly focuses on evaluating performance of the expected return portfolios whereas second par discusses 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 8 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 is reported for each portfolio. Left side of the table presents the values for equal weighted portfolios, whereas right side of the table presents the values for the value weighted portfolios. Panel A of the table shows the results for linear regression model, panel B shows the results for random forest model and panel C shows the results for 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 short position on lowest expected return portfolio and long position in highest expected return portfolio.

Looking at the equal weighted part of Table 8 provides is quite clear message. Even though out-of-sample R^2 remained rather low, especially for linear and neural network model, examined variable set seems to contain information about cross section of future stock returns. For all models there is clear rising trend of realized excess returns across expected return portfolios.

Table 8: Machine learning portfolio performance

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 current month account for 97 percent of cumulative market value of all stocks in the dataset. H-L is zero investment portfolio which consist of short position in portfolio formed from stocks with lowest expected return and long position in portfolio formed from stocks with highest expected return. Time series average of predicted return and realized excess return of each portfolio is reported for each model together with standard error of realized excess return. Additionally, Sharpe ratios are reported. Left side of the table reports result for equally weighted portfolios and right side reports results for portfolios where each stock in portfolio is weighted by its lagged market value. 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	-0.0060	0.0007	0.0753	0.1707	0.0093	-0.0048	0.0043	0.0777	1.0141	0.0550
2	0.0010	0.0042	0.0647	1.2114	0.0657	0.0010	0.0047	0.0663	1.3177	0.0715
3	0.0035	0.0075	0.0612	2.2721	0.1232	0.0036	0.0090	0.0656	2.5278	0.1371
4	0.0057	0.0087	0.0628	2.5530	0.1385	0.0058	0.0086	0.0652	2.4351	0.1321
5	0.0077	0.0072	0.0598	2.2355	0.1212	0.0078	0.0079	0.0628	2.3171	0.1257
6	0.0097	0.0092	0.0611	2.7685	0.1501	0.0097	0.0054	0.0627	1.5987	0.0867
7	0.0115	0.0113	0.0634	3.2810	0.1779	0.0115	0.0080	0.0670	2.1914	0.1188
8	0.0132	0.0123	0.0620	3.6509	0.1980	0.0132	0.0105	0.0657	2.9379	0.1593
9	0.0153	0.0139	0.0631	4.0488	0.2196	0.0153	0.0114	0.0656	3.2039	0.1738
High	0.0224	0.0165	0.0705	4.3041	0.2334	0.0210	0.0135	0.0742	3.3591	0.1822
H-L	0.0284	0.0158	0.0422	6.8846	0.3734	0.0259	0.0092	0.0574	2.9678	0.1610

Panel C: Neural network

	Equal weighted					Value weighted				
	Pred.	Avg.	Std.	*t*-stat	SR	Pred.	Avg.	Std.	*t*-stat	SR
Low	-0.0143	0.0014	0.0800	0.3156	0.0171	-0.0121	0.0055	0.0836	1.2160	0.0659
2	-0.0015	0.0054	0.0674	1.4704	0.0797	-0.0015	0.0066	0.0694	1.7639	0.0957
3	0.0021	0.0074	0.0646	2.1098	0.1144	0.0021	0.0055	0.0685	1.4862	0.0806
4	0.0046	0.0087	0.0601	2.6784	0.1453	0.0046	0.0095	0.0633	2.7524	0.1493
5	0.0066	0.0090	0.0595	2.7819	0.1509	0.0067	0.0090	0.0631	2.6242	0.1423
6	0.0085	0.0100	0.0595	3.0946	0.1678	0.0085	0.0085	0.0619	2.5283	0.1371
7	0.0104	0.0096	0.0589	2.9958	0.1625	0.0104	0.0075	0.0628	2.1906	0.1188
8	0.0126	0.0111	0.0617	3.3128	0.1797	0.0127	0.0086	0.0632	2.5208	0.1367
9	0.0156	0.0111	0.0618	3.3103	0.1795	0.0156	0.0114	0.0647	3.2419	0.1758
High	0.0245	0.0137	0.0674	3.7442	0.2031	0.0241	0.0127	0.0744	3.1384	0.1702
H-L	0.0388	0.0123	0.0439	5.1716	0.2805	0.0362	0.0072	0.0548	2.4058	0.1305

Linear model has couple of outliers where the return of lower expected return portfolio actually exceeds the return of 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 minimum expected return portfolio and maximum expected return portfolio is more than one percent for all models.

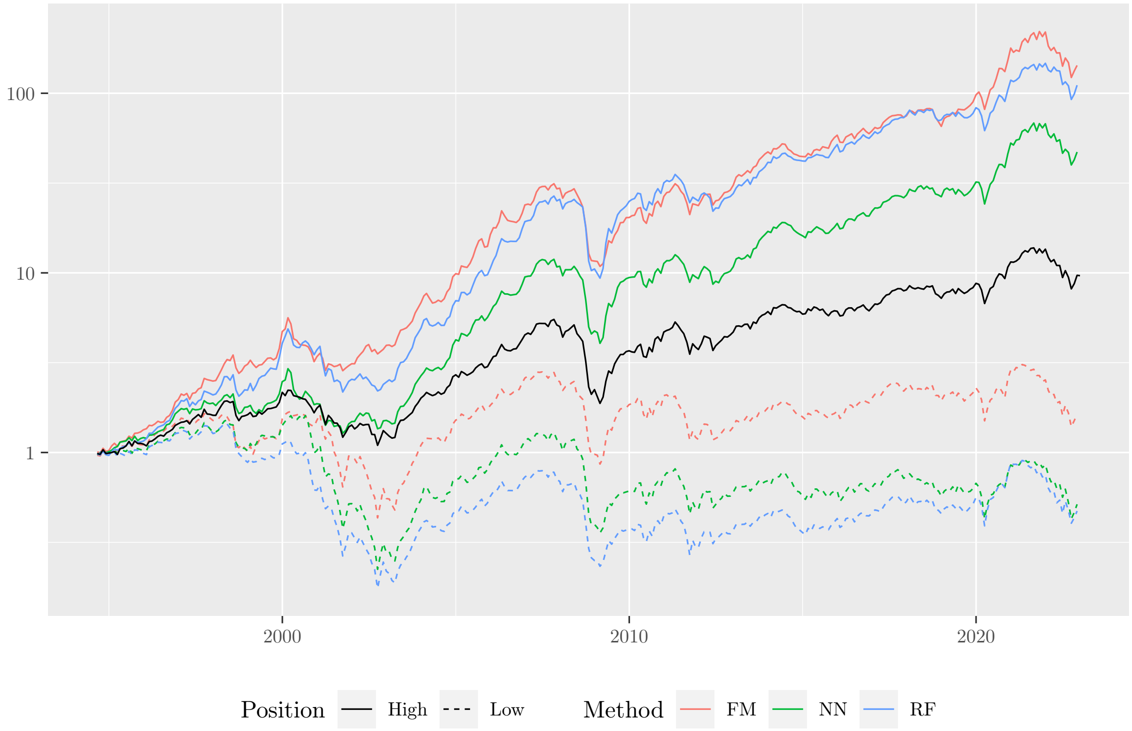
Models struggle more with value weighted portfolios. Increasing trend among value weighted is not as smooth and more outliers exist than among equal weighted portfolios. Nevertheless, for all models on average five smallest expected return portfolios generate lower realized returns than five largest expected return portfolios. For all models portfolio with highest expected return also has the highest realized excess return. Not surprisingly even among the value weighted portfolios random forest and neural network models show stronger performance. For these two models value weighted portfolio with highest expected return has the highest realized excess return, but also value weighted portfolio of lowest expected return has the lowest realized return. Additionally, it can be seen from Table 8 that realized return of the two lowest expected return portfolios is clearly below average market return from Table 3. Simultaneously, for these two models return of two highest expected return portfolios is above average market return with clear premium.

It is quite expectable that predictability of the value weighted portfolios is lower than the equal weighted portfolios. One reason for this is that is that stocks are divided to ten expected return portfolios. From Table 1 it can be seen that on average month dataset contains approximately 340 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 equal weighted portfolio. Nevertheless, typical to stock markets also Nordic stock markets have few extremely large market capitalization companies. Performance of these companies can even after winsorizing the market value drive the performance of the whole portfolio if number of stock inside the portfolio is limited.

Additionally, it is not guaranteed that each expected return portfolio would consists of same amount of stocks. This is because breakpoint expected returns for the portfolio allocation are calculated from expected returns of large market capitalization stocks. 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 8 is that standard deviation of the realized excess portfolio returns does not increase together average realized returns. For this reason Sharpe ratios

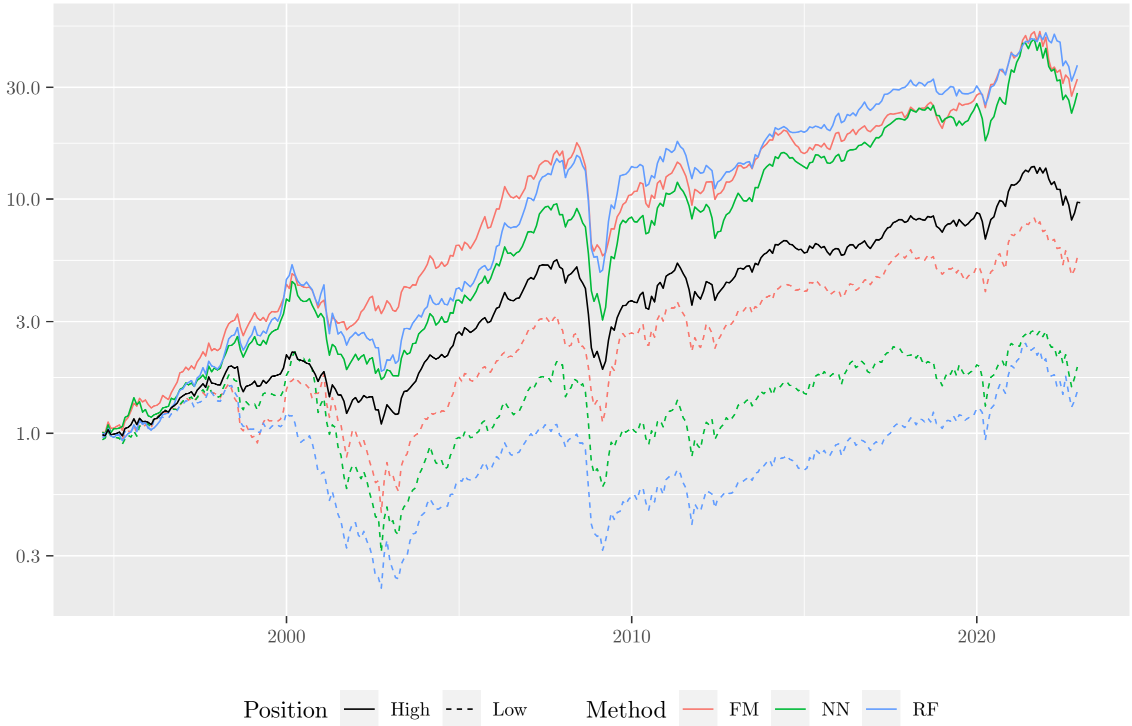
Figure 6: Cumulative return of equal weighted machine learning portfolios
Figure plots the realized historical cumulative excess return of the out-of-sample predictions. Figure shows performance of portfolios that are formed allocating all except micro-cap stocks to ten portfolios based on their expected 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. Solid line plots the cumulative performance of the highest expected return portfolio whereas dashed line plots the cumulative excess return for the lowest expected return portfolio. All portfolios are equal weighted. Solid black line shows the value weighted marked return. All returns are converted to US dollars. Prediction period spans from July 1994 to November 2022.



increase together with expected returns. For all models equal weighted portfolios portfolio with highest expected return has also the highest Sharpe ratio if high-low portfolios are not considered. Among value weighted portfolios this is true for linear regression and random forest model. Since the volatility of the returns does not increase together with magnitude of the returns, it means that machine learning models are able to generate excess returns without simply investing in more volatile stocks. Naturally, 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 risk adjusted performance of the machine learning portfolios. Risk adjusted performance is discussed in more detail for high-low portfolios later.

Results from Table 8 further support the findings of Section 6.1 that models overshoot in their predictions. Table shows that average predicted returns for middle expected returns portfolios are close to mean return from Table 3 especially for random forest and neural network models.

Figure 7: Cumulative return of value weighted machine learning portfolios
Figure plots the realized historical cumulative excess return of the out-of-sample predictions. Figure shows performance of portfolios that are formed allocating all except micro-cap 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. Solid line plots the cumulative performance of the highest expected return portfolio whereas dashed line plots the cumulative excess return for the lowest expected return portfolio. All portfolios are value weighted. Solid black line shows the value weighted marked return. All returns are converted to US dollars. Prediction period spans from July 1994 to November 2022.



On the other hand realized excess returns of the predicted return portfolios between third and fifth decile land closest to the markets mean return. On the other hand expected returns for minimum and maximum expected return portfolios are rather extreme. For example for neural network model spread between average expected returns of the two extreme portfolios is almost four percent for both equal and value weighted portfolios.

Given that there is 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 mean return of the market, it seems that the models do 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 similar situation with lowest expected return portfolios as well. Realized average excess return of the lowest expected return portfolios from neural network model are clearly below average market return, but still positive. As the average predicted return for these portfolios is between -1.4% and -0.12% percent, the spread between realized and expected return is rather large. This phenomena could at least partially explain the low out-of-sample R^2 values seen in Section 6.1.

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

Figures 6 and 7 show the historical cumulative return of the highest and lowest expected return portfolios for all models for equal value weighted portfolios correspondingly. Solid line shows the cumulative excess return for the highest expected return portfolio, whereas dashed line indicates the cumulative excess return of the lowest expected return portfolio for each model. Figures are in line with the results from the Table 8. Overall market trends can be seen from all portfolios, but there is clear spread between low and high expected return portfolios. Compared to Table 8 Figures 6 and 7 provide the time series dimension of the returns. For equal weighted portfolios Figure 6 reveals rather constant spread between high and low expected return portfolio, which results divergent cumulative returns. Despite the average realized return being slightly positive for lowest expected return portfolios for random forest and neural network models, the cumulative return of these portfolios is negative.

Same overall market trends can be seen from Figure 7 for value weighted portfolios. Compared to equal weighted portfolios, value weighted portfolios show more seasonality. Especially value weighted high expected return portfolios from neural network model seems to be sensitive to overall market distress. Both in early 2000's and during the financial crisis value of this portfolio decreases close to the market return. Among market value weighted portfolios both performance of the highest expected return portfolio as well as performance of the lowest expected return portfolio are more modest than among equal weighted portfolios. Even for neural network and random forest model cumulative return of the lowest expected return portfolio is positive.

Next part of the paper focuses on evaluating the performance of the long-short portfolios. Table 9 reports set of performance metrics for both equal and value weighted portfolios. Riskiness of the portfolios are evaluated by maximum drawdown and maximum one-month loss. Additionally, risk adjusted performance is evaluated by regressing 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, table reports the turnover for long side of the long-short portfolios.

Message from Table 9 is in line with previous part. Overall performance of equal weighted long-

Table 9: Zero investment portfolio performance metrics
Table shows different performance metrics for the spread in realized excess return of highest and lowest expected return portfolios for each model. Left side of the table shows the results for equal weighted portfolios and right side for market value weighted. Loss metrics reported in the table include maximum drawdown and maximum one-month loss. 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. Table also shows Sharpe ratios for each of the long-short portfolios. Last row of the table shows the turnovers of long side of the long-short portfolios. FM stands for linear regression model, RF stand for random forest model and NN stands for neural networks model. Prediction period spans from July 1994 to November 2022.

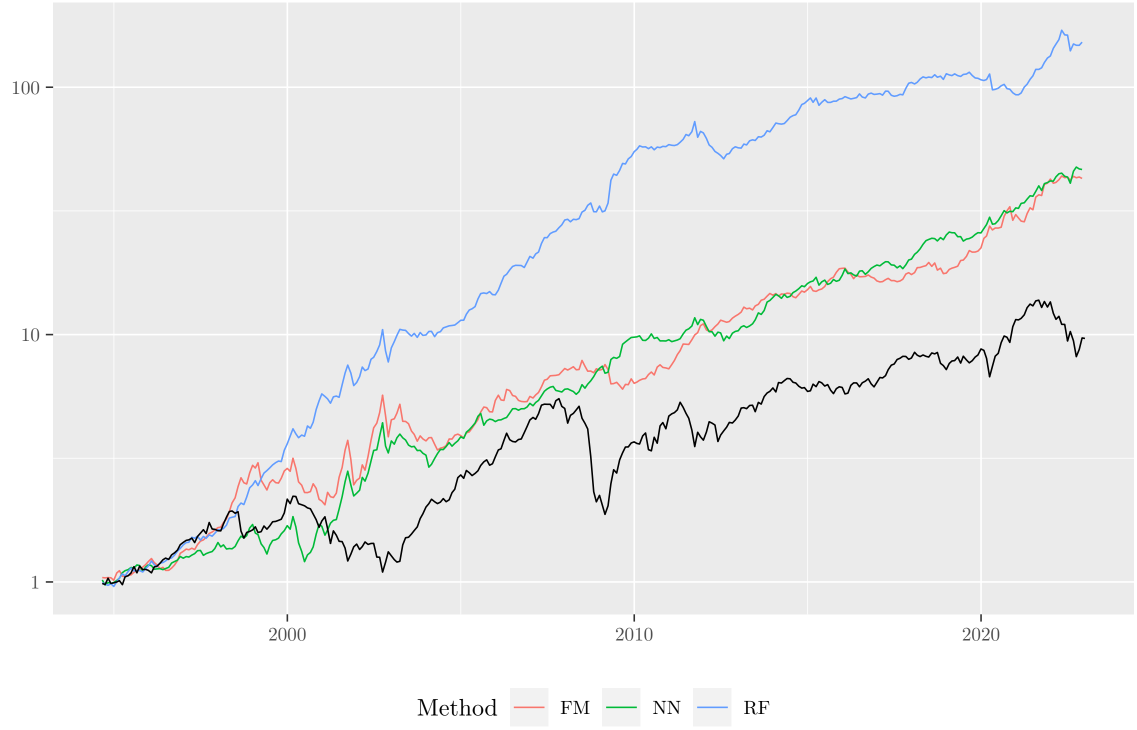
	Equal weighted			Value weighted		
	FM	RF	NN	FM	RF	NN
Max DD(%)	-0.4001	-0.2931	-0.3427	-0.6360	-0.4782	-0.4583
Max 1month Loss(%)	-0.2096	-0.1802	-0.1947	-0.3102	-0.3848	-0.3615
FF Alpha	0.0028	0.0118	0.0069	-0.0049	0.0037	0.0007
t-stats	1.4479	5.2782	3.2584	-1.9984	1.2254	0.2580
R^2	0.5853	0.1732	0.3180	0.4650	0.1747	0.2500
Sharpe ratio	0.2425	0.3734	0.2805	0.0777	0.1610	0.1305
Turnover (%)	0.4456	0.4458	0.4728	0.48842	0.6071	0.5875

short portfolios is stringer 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. Random forest model shows the strongest performance among the equal weighted portfolios with maximum drawdown of -29% and maximum one-month loss of -18%. Among the value weighted portfolios two risk measurements give slightly inconsistent message linear regression model has the highest maximum drawdown of -64%, but simultaneously lowest maximum one-month loss of -31%.

Table 8 showed the strong positive return long-short portfolios, which was statistically significantly different from zero for all portfolios except value weighted version from linear model. 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 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 9 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 linear model is also positive, but statistically not significant. Additionally, benchmark factors are able to explain 58% of the variation in returns of equal weighted long-short portfolio from 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

Figure 8: Cumulative return of equal weighted zero investment portfolios
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. Solid black line shows the value weighted marked return. All returns are converted to US dollars. Prediction period spans from July 1994 to November 2022.

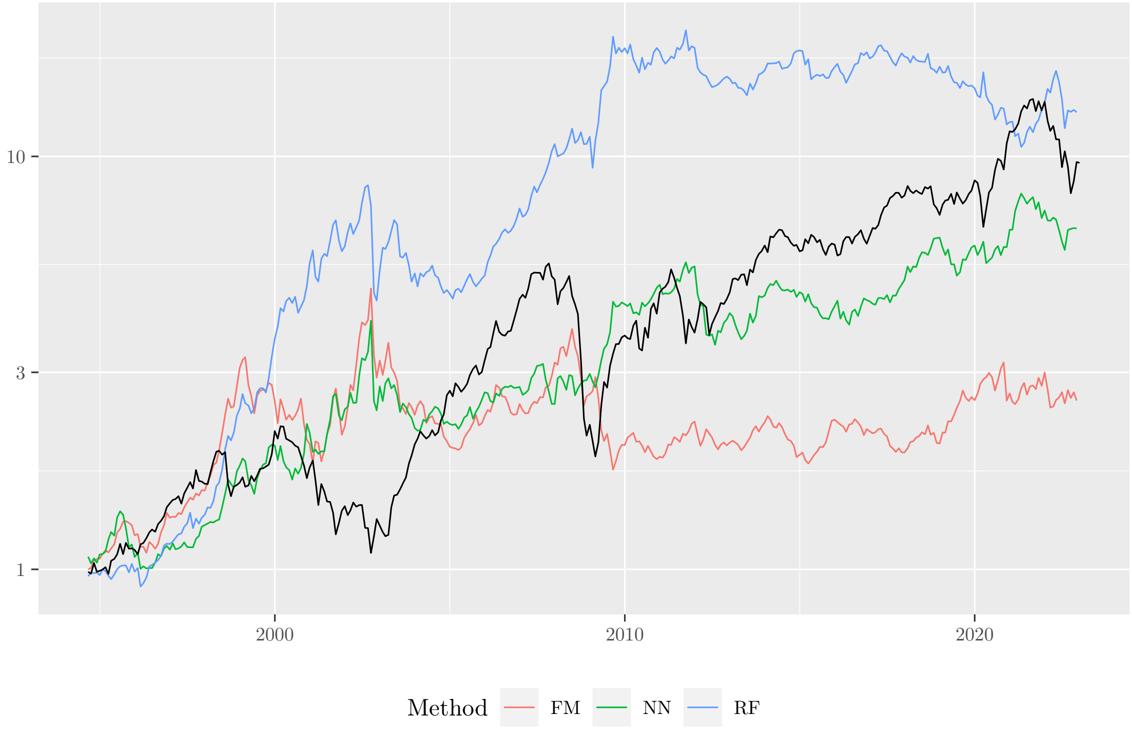


statistically significant. Only statistically significant alpha is the negative alpha of the linear regression model portfolio.

Figures 8 and 9 show the cumulative return of the equal and value weighted long-short portfolios 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 lowest and highest expected return portfolios. For linear regression and neural network models the spread is almost identical. Figure 6 shows that high expected return portfolio of the linear model performs better than the high expected return portfolio of the neural network, but neural network model does better job picking low expected return companies.

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 highest and lowest expected return machine learning portfolios is not affected by the overall stock

Figure 9: Cumulative return of value weighted zero investment portfolios
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. Solid black line shows the value weighted marked return. All returns are converted to US dollars. Prediction period spans from July 1994 to November 2022.



market distress. For example effect of financial crisis around 2008 can be clearly seen from the cumulative market return, but cumulative return of any long-short machine learning portfolio is not remarkably changed. Figure 6 shows that the equal weighted long-short machine learning portfolios provide larger and more smooth returns than the market return.

Looking at Figure 9 shows that none of the value weighted long-short portfolios are able to remarkably exceed the market return. Cumulative return of the random forest model ends up slightly above market return whereas cumulative return of neural network model ends up slightly below it. Even though cumulative return of value weighted long-short neural network portfolio does not exceed market return it is the only model that show increasing trend in the cumulative return across the prediction period. Performance of the linear model is poor throughout the period whereas random forest performs strongly in the first 16 years of the period. In last twelve years of the prediction period cumulative return of the value weighted long-short portfolio from random forest model is actually negative.

Finally, the turnover of the long-short portfolios is surprisingly low. Lot of previous studies report turnover exceeding 100% for long-short portfolios (Gu et al., 2020; Tobek & Hronec, 2021). 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 9 shows that average monthly turnover of the long-short portfolios is around 50%. Average monthly turnover for value weighted long-short portfolios is varying between 49% for linear model and 61% for random forest model. This is slightly more than monthly turnover of equal weighted long-short portfolios which are 45% for linear random forest model and 47% for neural network model.

Partially low turnover could be result of the frequency of the predictor variables. This study includes lot covariates on 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 next 12 months. In a situation where set of companies would be constant through out the whole period and only annual predictors would be included, 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 the transaction costs, which naturally reduce the return of the investment. Investment strategy that requires monthly changing position in majority of invested capital is often not implementable in real world setting due to increased transactions costs compared to more passive investment strategy.

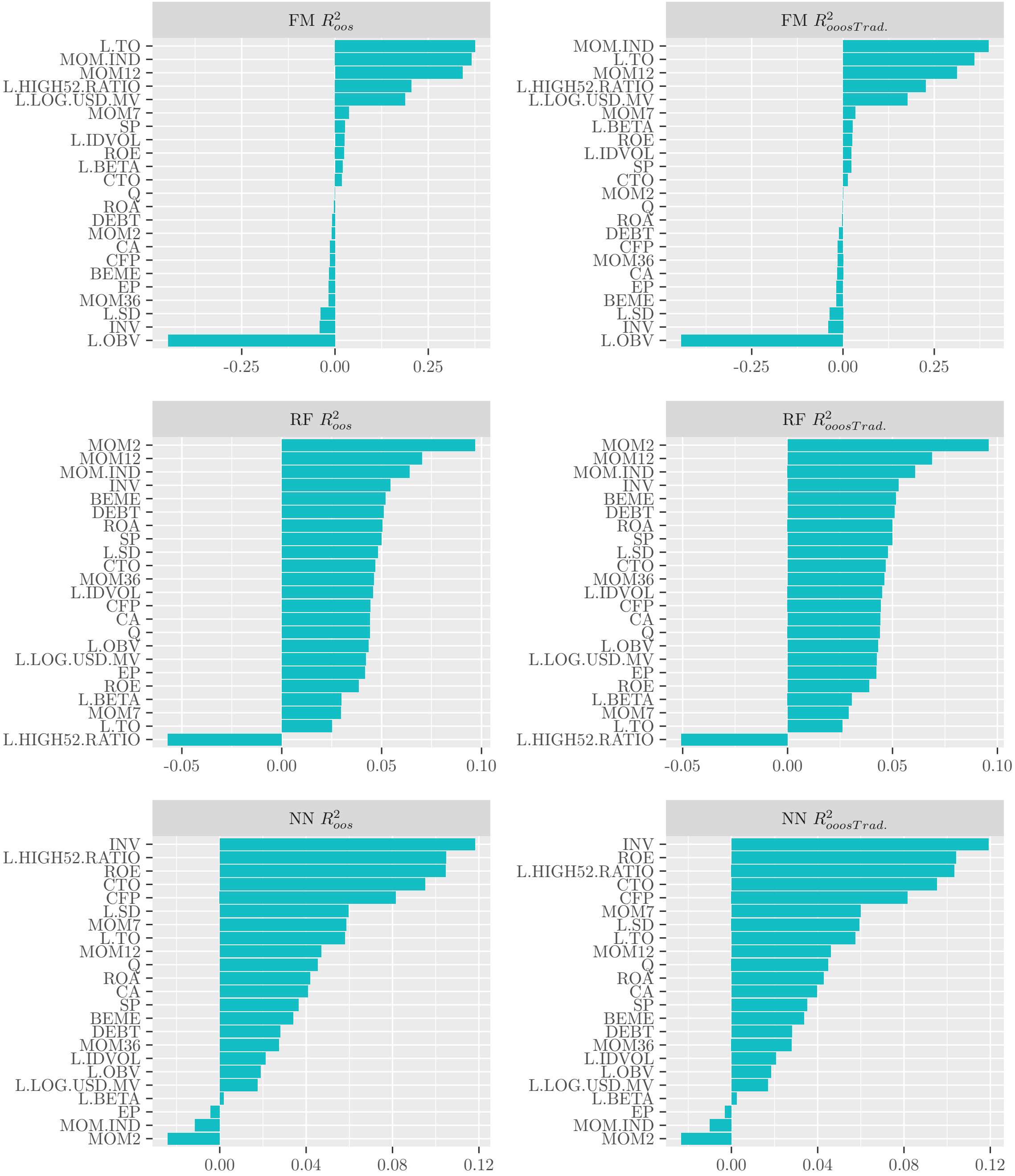
6.3 Predictive characteristic importance

This section sheds light to which covariates contribute the most to the predictive accuracy of the machine learning models. Figure 10 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. Final variable importance figure is then the time series average reduction in out-of-sample R^2. Therefore, variable importance measured using different definition of out-of-sample R^2 can result slightly different results, but as can be seen from Figure 10 differences are minor.

Figure 10 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 52 week high price (L.HIHG52.RATIO) and log market value (L.LOG.USD.MV) results clear reduction in prediction accuracy of the model. Rest of the variables seem to have only minor effect on the prediction accuracy of the model or the effect is even negative. Remarkable result

Figure 10: Relative variable importance

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



from Figure 10 is that on-balance volume (L.OBV) has clear negative effect on prediction performance of the linear model and predictions of the model are more precise when the variable is excluded.

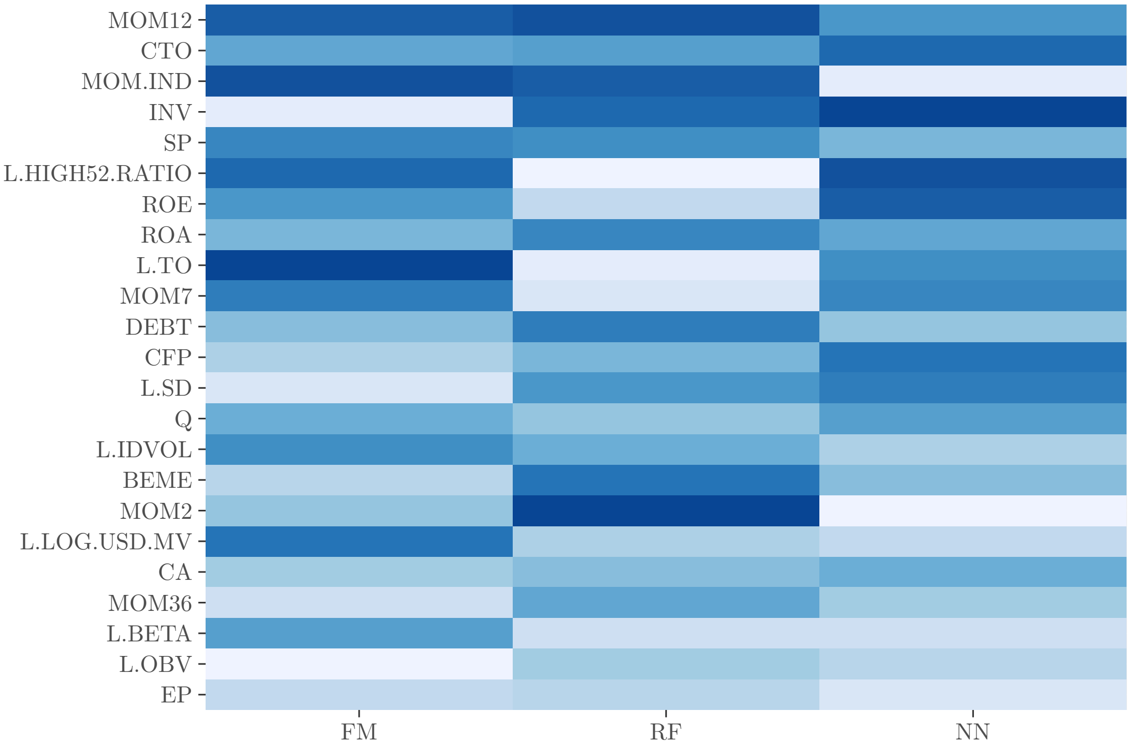
Variable importance of random forest model is not as skewed as for linear model and lot of variables show similar importance. Random forest model puts lot of weight on momentum variables. Most influential variables for random forest model are short-term momentum (MOM2), 12-month momentum (MOM12), industry momentum (MOM.IND) and investment (INV). Only 52 week high price (L.HIHG52.RATIO) worsens the predictive accuracy of the random forest model.

Variable importance for neural network model shown in Figure 10 is quite different from the two other models. Linear model had few important variables, whereas for random forest model majority of the models played similar importance. On the other hand for neural network variable importance is increasing almost linearly starting from least important variable. Variables that play the biggest role in prediction accuracy of the neural network model are investment (INV), 52 week high price (L.HIHG52.RATIO), return on equity (ROE) and capital turnover (CTO). Short-term momentum (MOM2) and industry momentum (MOM.IND) are the least important variables for neural network model, which actually slightly reduce the prediction accuracy of the model.

Figure 11 provides insight how aligned the variables importances of different firm characteristics are between different models as it shows the variable importance ranks of the models. For each model variable with highest variable importance gets variable importance rank of one. Therefore, darker colour in Figure 11 indicates lower variable importance rank and lighter colour higher variable importance rank. Figure shows some dispersion on the variable importances of the different models. For example short-term momentum is the most important variable for random forest model and 52 week high price is the least important. For neural network model the results are almost opposite as the short-term momentum is the least important variable and the 52 week high price is among the most important ones. Interestingly for linear regression model and random forest model most important variables have monthly frequency where neural network weights annual characteristics.

Most important variable among all models is the 12-month momentum. 12-month momentum shows relative large importance in all models, which is not big surprise given the strong performance of momentum benchmark factor in Section 5.2. Momentum effect is also well documented in previous Nordic stock market anomaly literature (e.g. Grobys and Huhta-Halkola (2019) and Leivo and Pätäri (2011)). Another variable that shows rather strong influence across the models is the capital turnover. On the other hand models agree on the low importance of on-balance volume and earnings-to-price ratio. Including earnings-to-price ratio even reduces the predic-

Figure 11: Variable importance
Figure plots the importance of the explanatory variables to the predictive performance of the three machine learning models. Variable importance is defined as reduction in out-of-sample R^2 when corresponding variable is replaced by zero before each training process. Definition of the out-of-sample R^2 is described in Section 4. Darker colour indicates higher variable importance. FM stands for linear regression model, RF stand for random forest model and NN stands for neural networks model. Prediction period spans from July 1994 to November 2022.



tion accuracy for two out of three models as can be seen from Figure 10. One objective of this study was to examine whether on-balance volume would contain information about 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 important factor for variable importance of linear regression model with clearly negative out-of-sample R^2. Hypothetically if exclusion of a characteristics would demean all the predictions of linear model to 0, variable importance of this variable would be clearly positive. This is because then in 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 positive change in out-of-sample R^2. Nevertheless, after excluding this hypothetical variable model would be useless for an investors, since constant prediction would not provide any information about the cross section of the future stock returns. Therefore, investor would not have any indicator to for the portfolios construction from Section 6.2.

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 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 exclusion of a variable. This way it could be seen if the including a variable brings the realized excess return of an expected return portfolio closer to its expected return.

7 Conclusion

More trading variables.

Even though it is argued that the linear models do not require hyperparameter optimization, as the Fama MacBeth variation is chosen one variable could be treated as a hyperparameter. Implemented linear model requires input from the user for the rolling window used to calculate mean factor loadings. One option for further research could be to treat also linear model more like the other machine learning models and optimize the rolling window.

Discuss other hyperparameters as well.

Low variable because lot of yearly variables and few stocks allocated to each portfolio.

Rolling time window, because r2 in the end volatile. If there is indicators that predictors relation to excess return is time varying shorter window could be justified.

A Data collection

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 dynamic screen, corresponding security is not necessary completely excluded from the dataset. Tables A.3 and A.4 shows the static screen. 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 key words that are searched among Datastream attributes NAME, ENAME and ECNAME. In case requirement from panel A of Table A.3 is not met or if maleficent keyword is found in the name of the security, security is excluded from the dataset completely.

Table A.1: Constituent lists and keywords
Table provides the constituent lists used in data collection.

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

Table A.2: Dynamic screens
Table provides the dynamics screens used in the data cleaning.

Affected attribute	Applied screen
RI	Observations where 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 last available value. Therefore, by removing all consecutive zero returns at the end of the dataset for all securities.

Table A.3: Static screens

It also provides the country specific keywords that are used to deleted entries from the dataset. Panel B provides keywords that were used to delete entries from each market separately. Keyword deletion follows Ince and Porter (2006) and Hanauer and Windmüller (2023). Same logic is applied to remove both country specific and generic keywords. Keyword is searched from Datastream attributes NAME, ENAME and ECNAME. In case if 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 keyword occurs at the beginning of the name, at the end of the name or as 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				CONVERTED INTO, USE,
ENAME	\\)CSE \\	USE		CONVERTED-,
ECNAME				CONVERTED - SEE

Table A.4: Common keywords

Table shows the general keywords that were used to delete entries from all markets. Keyword deletion follows Ince and Porter (2006) and Hanauer and Windmüller (2023). Same logic is applied to remove both country specific and generic keywords. Keyword is searched from Datastream attributes NAME, ENAME and ECNAME. In case if 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 keyword occurs at the beginning of the name, at the end of the name or as 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.

B Benchmark factor properties

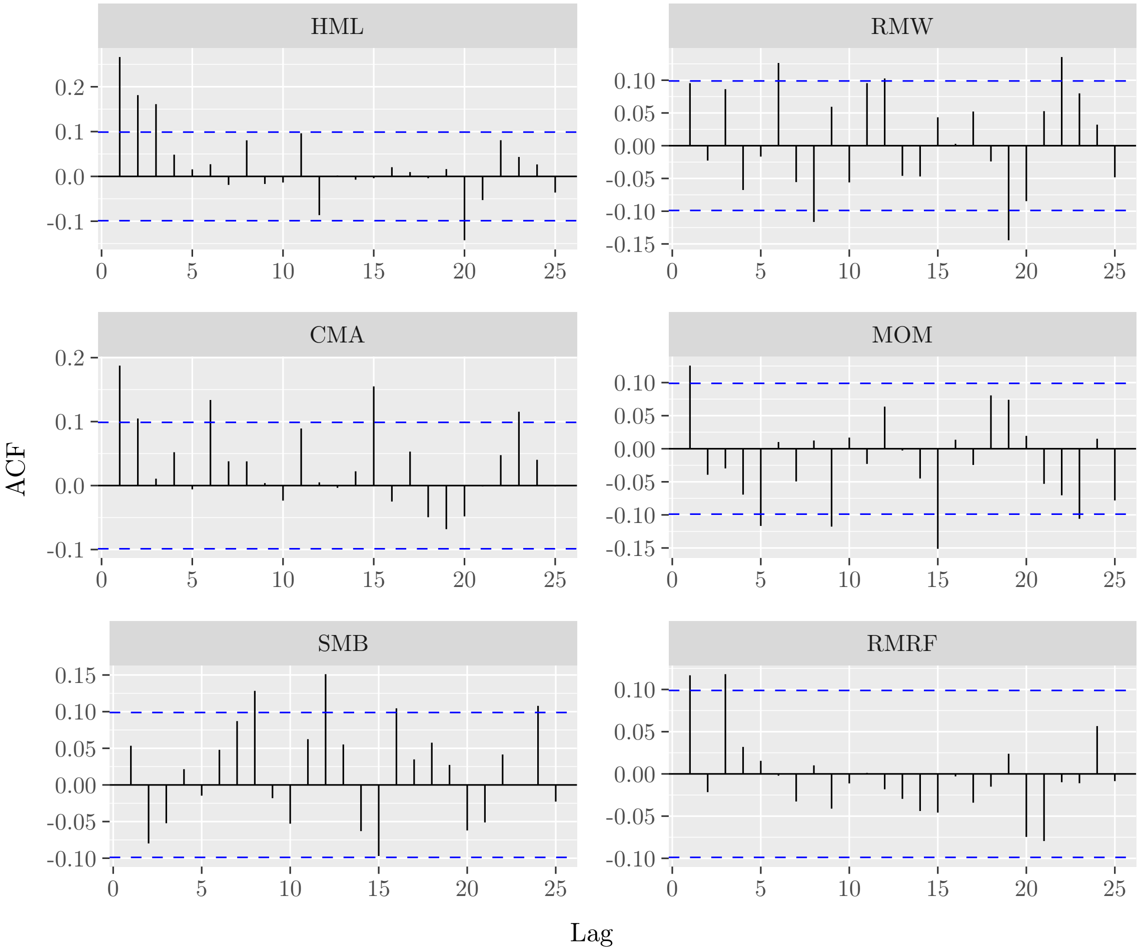
Supplementary materials for the benchmark factors.

Table B.1: Benchmark factor correlation matrix
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 average of value weighted return of two high value portfolios and average of value weighted return of two low value portfolios. RMW, CMA and MOM are calculated in similar manner, but portfolio sorts are done based on investment, profitability 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	-0.5707	0.5542	0.1122	0.2998	-0.2740
RMW	-0.5707	1	-0.5899	0.0857	-0.2568	0.0639
CMA	0.5542	-0.5899	1	-0.0703	0.1777	-0.2078
MOM	0.1122	0.0857	-0.0703	1	0.1544	-0.2040
SMB	0.2998	-0.2568	0.1777	0.1544	1	-0.2695
RMRF	-0.2740	0.0639	-0.2078	-0.2040	-0.2695	1

Figure B.1: Factor autocorrelation

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 average of value weighted return of two high value portfolios and average of value weighted return of two low value portfolios. RMW, CMA and MOM are calculated in similar manner, but portfolio sorts are done based on investment, profitability 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.



C Additional information

Figure C.1: Number of companies

Figure shows the development of total number of securities considered in the dataset from 1990 to 2022 for each Nordic country. Figures counts all securities that passed the static screens.

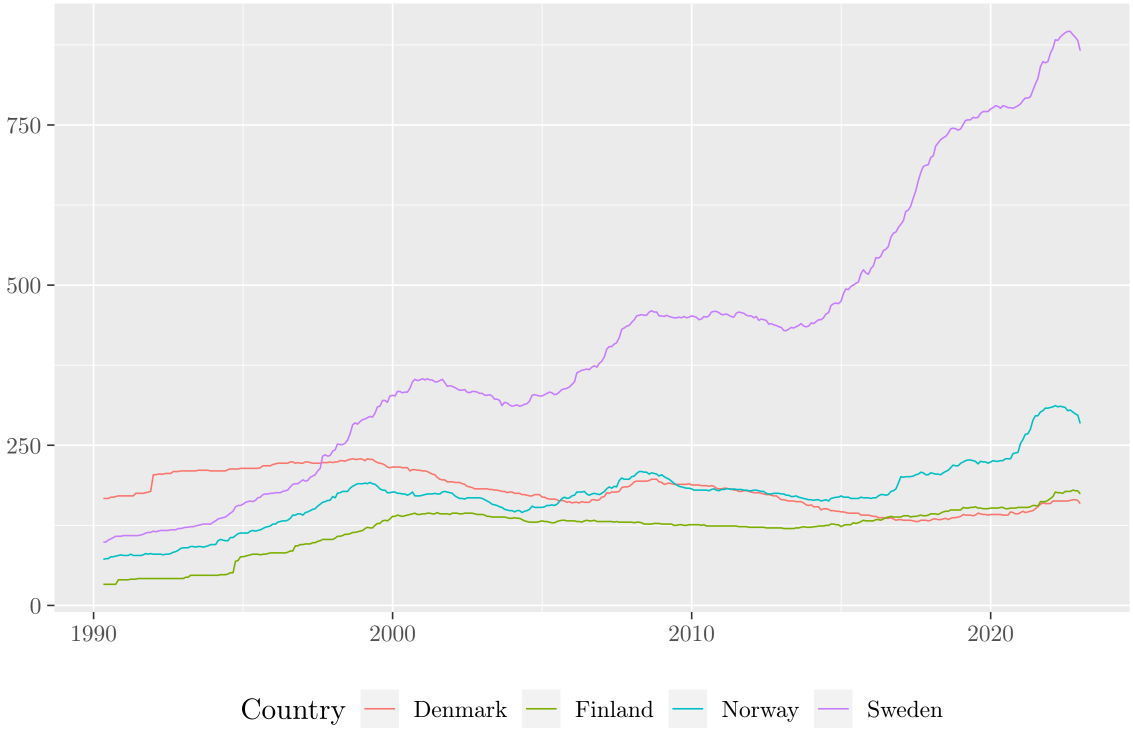


Figure C.2: Exchange rates
Figure shows the development of currency rates compared to US dollars. DK stands for Danish krone, E stands for Euro, MK stands for Finnish markka, NK stands for Norwegian krone and SK stands for Swedish krona.

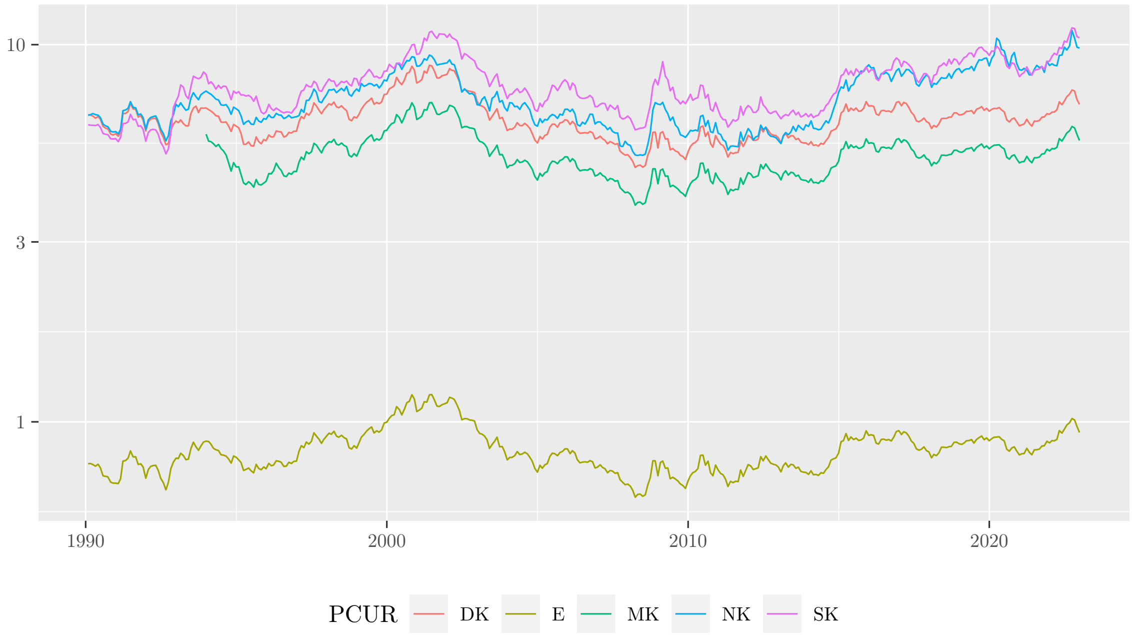


Figure C.3: US dollars one-month Treasury bill rate
Figures shows the development of US dollars one-month Treasury bill rate, which is used as risk free rate in this study.

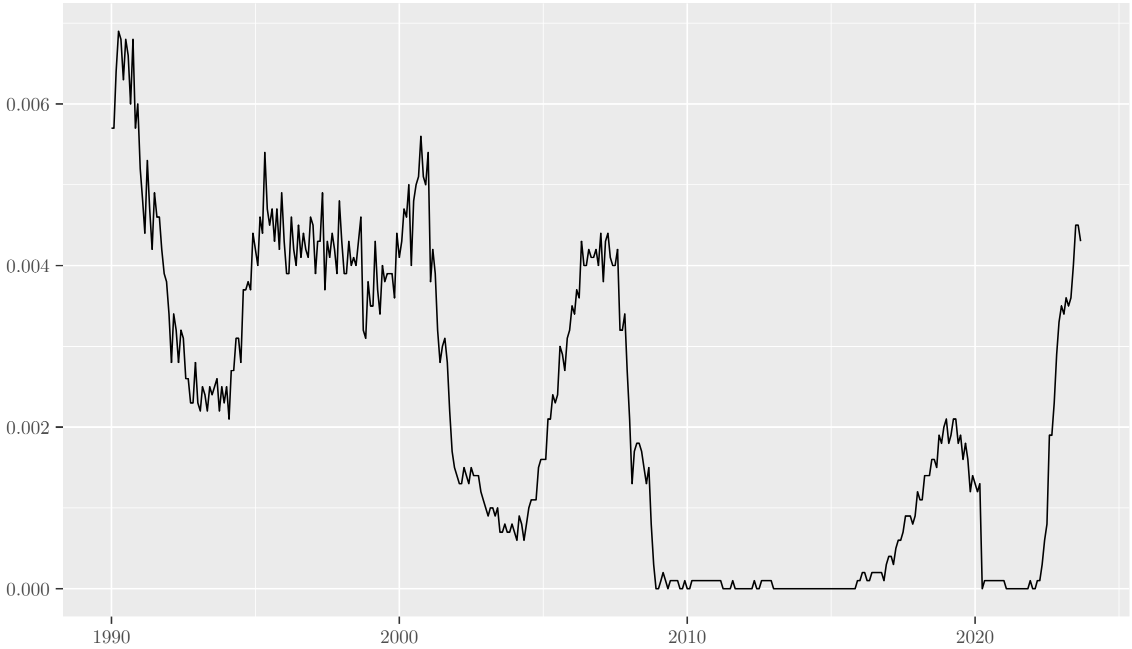


Figure C.4: Time series of the mean cross-sectional properties
Figure plots development of the firm characteristics across time. Values shown are the monthly cross-sectional averages. Construction of each variable is explained in detail in Section 3.

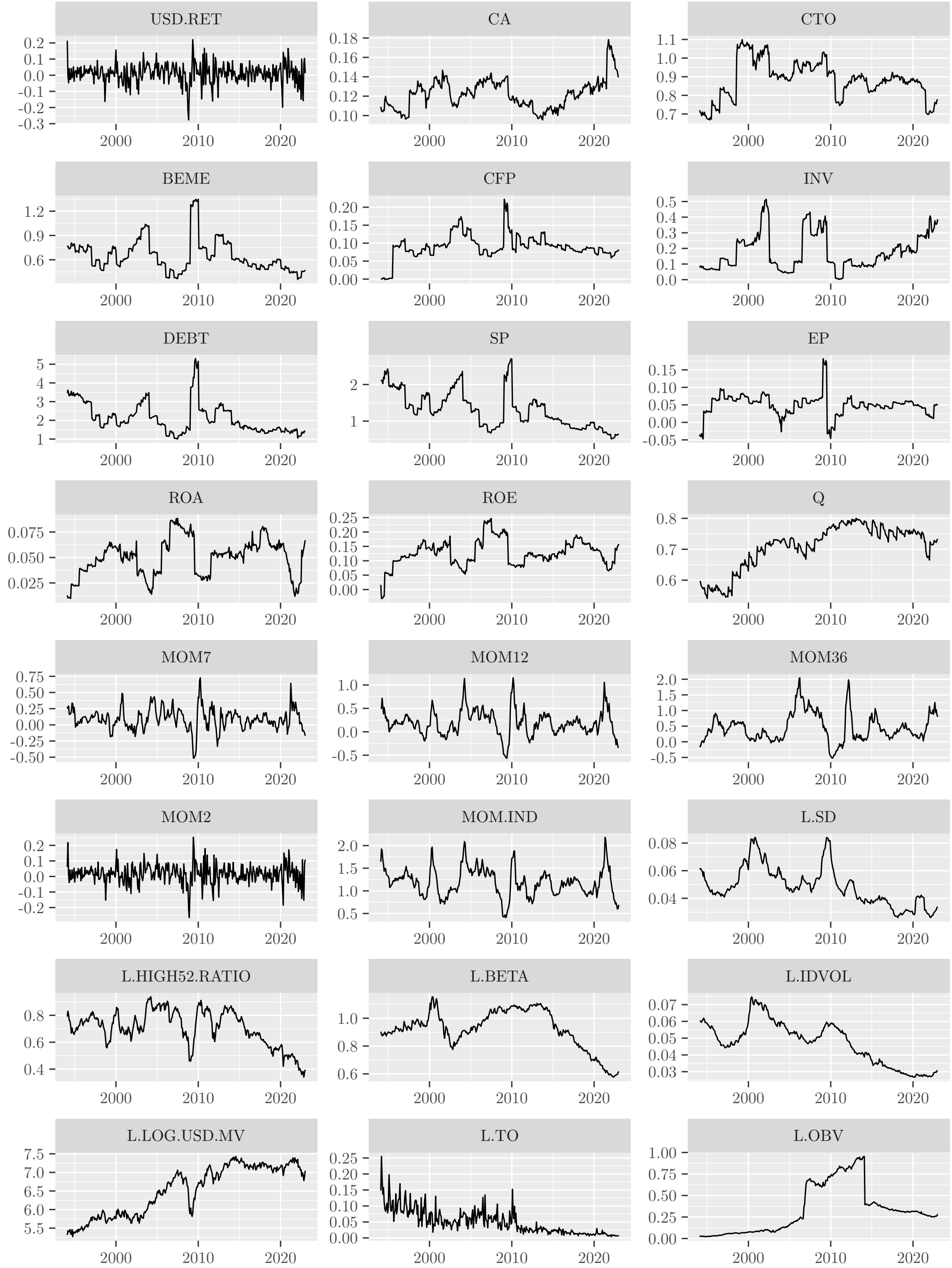


Figure C.5: Time series of out-of-sample R^2s

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 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.

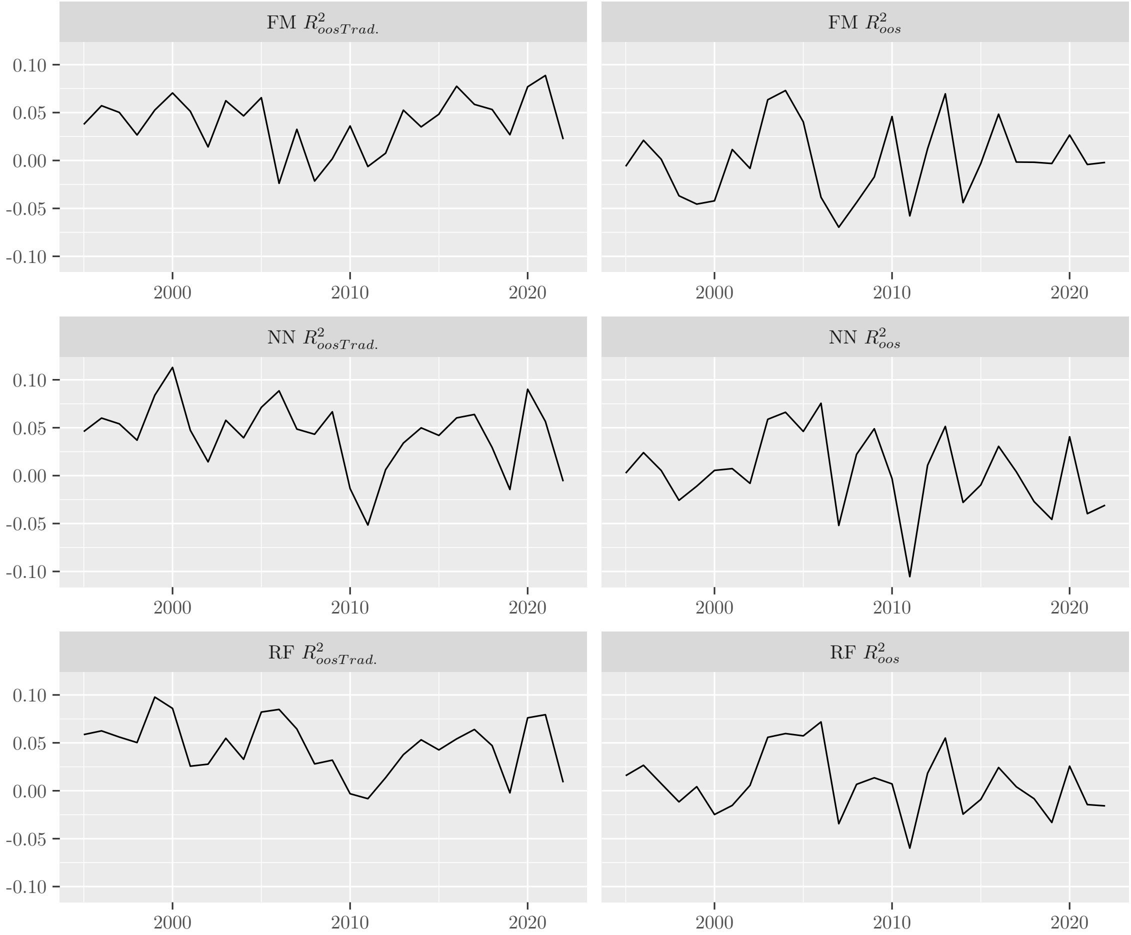


Figure C.6: Turnover of long-short machine learning portfolios
Figure plots time series of turnover of the long-short portfolios. Values show are mean turnovers of prediction periods. FM stands for linear regression model, RF stand for random forest model and NN stands for neural networks model.

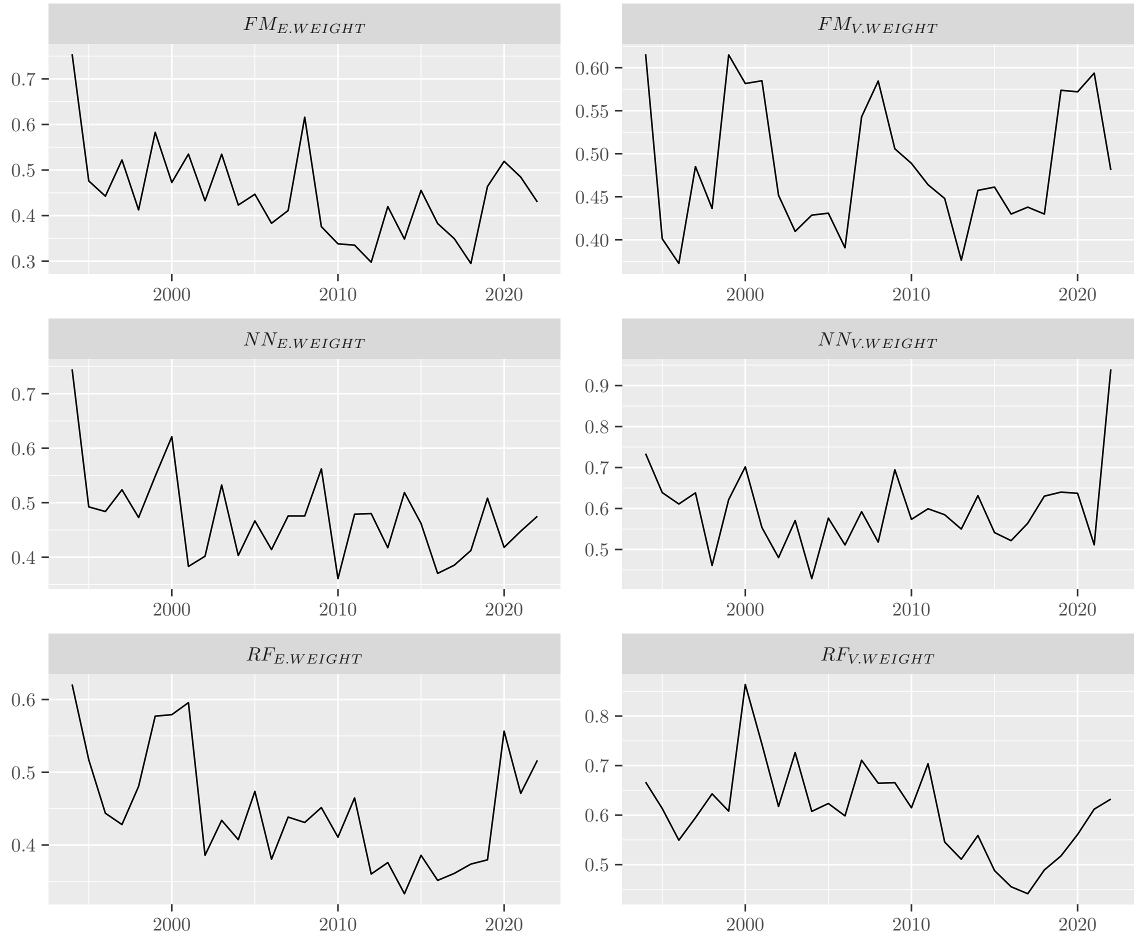
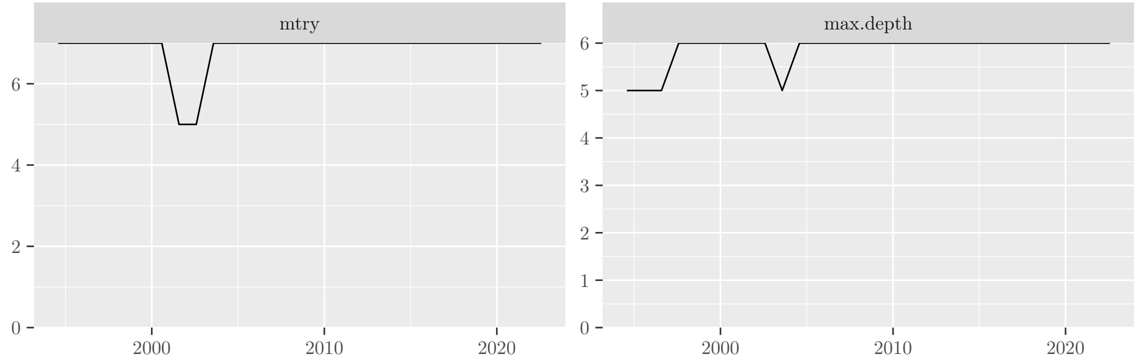


Figure C.7: Random forest optimized hyper parameters
Time series of optimal hyper parameters for random forest model. Mtry stands for number of features to possibly split at in each node and max.depth stands for maximum depth of the regression trees in random forest model.



References

Ali, A., Hwang, L.-S., & Trombley, M. A. (2003). Arbitrage risk and the book-to-market anomaly. *Journal of Financial Economics*, *69*(2), 355-373. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X03001168 doi: https://doi.org/10.1016/S0304-405X(03)00116-8

ANG, A., HODRICK, R. J., XING, Y., & ZHANG, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259-299. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2006.00836.x doi: https://doi.org/10.1111/j.1540-6261.2006.00836.x

Anwar, B. H., & Hogholm, K. (2020). The impact of illiquidity risk for the nordic markets. *Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad*, *49*(1), 28-47. Retrieved from https://doi.org/10.1080/02102412.2018.1555348 doi: 10.1080/02102412.2018.1555348

Balakrishnan, K., Bartov, E., & Faurel, L. (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics*, *50*(1), 20-41. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165410109000810 doi: https://doi.org/10.1016/j.jacceco.2009.12.002

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3-18. Retrieved from https://www.sciencedirect.com/science/article/pii/0304405X81900180 doi: https://doi.org/10.1016/0304-405X(81)90018-0

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, *32*(3), 663–682. Retrieved 2024-03-09, from http://www.jstor.org/stable/2326304

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, *43*(2), 507–528. Retrieved 2024-03-12, from http://www.jstor.org/stable/2328473

Bondt, W. F. M. D., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793–805. Retrieved 2024-03-09, from http://www.jstor.org/stable/2327804

Breiman, L. (2001). Random forests [Article]. *Machine Learning*, *45*(1), 5 – 32. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035478854&doi=10.1023%2fA%3a1010933404324&partnerID=40&md5=4b9f43897146098c0df3a2af232cf2f4 (Cited by: 78308) doi: 10.1023/A:1010933404324

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57–82. Retrieved 2023-11-24, from http://www.jstor.org/stable/2329556

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance*, *63*(4), 1609–1651. Retrieved 2024-03-09, from http://www.jstor.org/stable/25094485

Datar, V. T., Y. Naik, N., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, *1*(2), 203-219. Retrieved from https://www.sciencedirect.com/science/article/pii/S1386418197000049 doi: https://doi.org/10.1016/S1386-4181(97)00004-9

Davis, J. L., Fama, E. F., & French, K. R. (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance*, *55*(1), 389–406. Retrieved 2024-03-09, from http://www.jstor.org/stable/222559

Davydov, D., Tikkanen, J., & Äijö, J. (2017). Magic formula vs. traditional value investment strategies in the finnish stock market.. Retrieved from https://api.semanticscholar.org/CorpusID:220593553

Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: evidence from the european stock market. *Journal of Asset Management*, *22*(7), 507–538. Retrieved from https://doi.org/10.1057/s41260-021-00237-x doi: 10.1057/s41260-021-00237-x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3-56. Retrieved from https://www.sciencedirect.com/science/article/pii/0304405X93900235 doi: https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, *105*(3), 457-472. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X12000931 doi: https://doi.org/10.1016/j.jfineco.2012.05.011

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X14002323 doi: https://doi.org/10.1016/j.jfineco.2014.10.010

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636. Retrieved 2023-10-29, from http://www.jstor.org/stable/1831028

Fieberg, C., Metko, D., Poddig, T., & Loy, T. (2023). Machine learning techniques for cross-sectional equity returns'prediction. *OR Spectrum*, *45*(1), 289–323. Retrieved from https://doi.org/10.1007/s00291-022-00693-w doi: 10.1007/s00291-022-00693-w

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), pp. 2326–2377. Retrieved 2024-03-09, from https://www.jstor.org/stable/48574462

George, T. J., & Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, *59*(5), 2145–2176. Retrieved 2024-03-03, from http://www.jstor.org/stable/3694820

Green, J., Hand, J. R. M., & Zhang, X. F. (2017, 03). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies*, *30*(12), 4389-4436. Retrieved from `https://doi.org/10.1093/rfs/hhx019` doi: 10.1093/rfs/hhx019

Grobys, K., & Huhta-Halkola, T. (2019). Combining value and momentum: evidence from the nordic equity market. *Applied Economics*, *51*(26), 2872-2884. Retrieved from `https://doi.org/10.1080/00036846.2018.1558364` doi: 10.1080/00036846.2018.1558364

Gu, S., Kelly, B., & Xiu, D. (2020, 02). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223-2273. Retrieved from `https://doi.org/10.1093/rfs/hhaa009` doi: 10.1093/rfs/hhaa009

Hanauer, M. X., & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, *55*, 101022. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1566014123000274` doi: https://doi.org/10.1016/j.ememar.2023.101022

Hanauer, M. X., & Windmüller, S. (2023). Enhanced momentum strategies. *Journal of Banking & Finance*, *148*, 106712. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0378426622002928` doi: https://doi.org/10.1016/j.jbankfin.2022.106712

Haugen, R. A., & Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, *41*(3), 401-439. Retrieved from `https://www.sciencedirect.com/science/article/pii/0304405X9500868F` doi: https://doi.org/10.1016/0304-405X(95)00868-F

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, p. 278-282 vol.1). doi: 10.1109/ICDAR.1995.598994

Ince, O. S., & Porter, R. B. (2006). Individual equity return data from thomson datastream: Handle with care! *Journal of Financial Research*, *29*(4), 463-479. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6803.2006.00189.x` doi: https://doi.org/10.1111/j.1475-6803.2006.00189.x

Jacobs, H., & Müller, S. (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics*, *135*(1), 213-230. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0304405X19301618` doi: https://doi.org/10.1016/j.jfineco.2019.06.004

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, *45*(3), 881–898. Retrieved 2024-03-09, from `http://www.jstor.org/stable/2328797`

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91. Retrieved 2024-03-09, from `http://www.jstor.org/stable/2328882`

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

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance*, *49*(5), 1541–1578. Retrieved 2024-03-09, from http://www.jstor.org/stable/2329262

Leivo, T. H., & Pätäri, E. J. (2011). Enhancement of value portfolio performance using momentum and the long-short strategy: The finnish evidence. *Journal of Asset Management*, *11*(6), 401–416. Retrieved from https://doi.org/10.1057/jam.2009.38 doi: 10.1057/jam.2009.38

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review*, *4*(1), 1-44. Retrieved from http://dx.doi.org/10.1561/104.00000024 doi: 10.1561/104.00000024

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, *47*(1), 13–37. Retrieved 2024-03-12, from http://www.jstor.org/stable/1924119

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, *54*(4), 1249–1290. Retrieved 2024-03-03, from http://www.jstor.org/stable/798005

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, *103*(3), 429-453. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X11001152 doi: https://doi.org/10.1016/j.jfineco.2011.05.003

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics*, *104*(1), 162-185. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X11002856 doi: https://doi.org/10.1016/j.jfineco.2011.12.009

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442. Retrieved 2024-03-12, from http://www.jstor.org/stable/2977928

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, *56*, 100588. Retrieved from https://www.sciencedirect.com/science/article/pii/S1386418120300574 doi: https://doi.org/10.1016/j.finmar.2020.100588

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