Norwegian School of Economics
Bergen, Spring 2022

Enhancing Fund Selection Using Supervised Machine Learning

Evidence From the Nordic Mutual Fund Market

Daniel André Voll Eriksen & Niklas Hagen

Supervisor: Andreas Ørpetveit

Master of Science in Economics and Business Administration

Major in Financial Economics and Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

Insightful and invaluable years at NHH soon come to an end, and both of us feel honored to attend such a great school with brilliant professors and associate students.

This paper has proven to be demanding, although very meaningful and rewarding. The immense scale of data to be pre-processed and cleaned has taken much time, but doing it thoroughly has been a priority and helped us optimize our thesis. We believe that this thesis contributes to broadening the literature in the Nordics regarding machine learning and its ability to predict alpha.

We want to express our most profound honor to our supervisor, Andreas Ørpetveit, for insightful discussions and counseling in this project. We acknowledge that his expertise has pushed us in the right direction. We would also like to thank André Sjuve for his profound and thoughtful comments on this thesis. Their competence in asset allocation and finance has proven invaluable in writing this thesis.

Finally, we would like to thank Eugene Fama for some insightful recommendations regarding our paper. The discussions have lifted the quality of our thesis to another level.

Norwegian School of Economics

Bergen, June 2022

Daniel André Voll Eriksen

Niklas Hagen

Abstract

In this research we aim to extend the literature on the performance predictability in actively managed mutual funds. We use the Nordic mutual fund market as our laboratory. We develop a performance-enhancing system to assist retail investors in selecting mutual funds by utilizing gradient boosting, random forest, and deep neural networks. Furthermore, we seek to obtain positive abnormal returns from our predicted quintile portfolios. We thus retrieve data free of survivorship bias for 2748 Nordic mutual funds from Morningstar Direct. First, we run the algorithms to test the possibility of classifying alphas. Secondly, we create a ranking system that categorizes funds based on predicted alpha, enabling us to separate the best from the worst-performing mutual funds. At last, we benchmark our findings against Morningstar's acknowledged rating platform to examine whether our top quintile portfolios manage to outperform Morningstar's top quintile portfolio. We find that our models can classify the sign of the alpha coefficient, whereas gradient boosting and random forest does this exceptionally well. Further, we manage to create a categorization system significantly outperforming both an equally weighted and asset weighted benchmark on risk-adjusted returns. Finally, our best performing portfolios generate risk-adjusted returns in excess of Morningstar, although only significantly for gradient boosting. Results are further robust to changes in risk-adjustment models for both equity funds and fixed income funds. The findings are consistent with the current machine learning literature and enable us to state that machine learning algorithms can be used to select successful mutual funds.

Keywords – Mutual Fund, Nordic Market, Machine Learning, Performance Predictability, Predictive Analytics, Alpha, Morningstar, Ranking, Fama French, Abnormal Returns

Contents

List of Figures

List of Tables

1 Introduction

Despite the growing popularity of passive investing, actively managed mutual funds still captures significant market shares. Investing in such a fund needs to provide benefits over passive benchmarks, which is why mutual funds often brag about their ability to deliver excess returns. However, the problem arises from the fact that the empirical data disfavors that statement to a large degree.[1] French (2008) concludes that retail investors would be better off by 0.7% p.a by changing their investment style from active to passive. A more recent study by Leippold and Rueegg (2020) investigate a large global sample of mutual funds and found that they cannot reject their null hypothesis that fund alphas are indistinguishably different from zero. If it then exists low-hanging fruits in the market that allow for $\alpha \neq 0$, we want to test whether we can exploit this caveat systematically using machine learning algorithms.

A popular paper by Ryll and Seidens (2019) reviewed more than 150 studies applying machine learning to financial market forecasting. Interestingly, the authors find that machine learning algorithms outperform most traditional stochastic forecasting methods in the financial markets. The authors further argue that robust forecasting models that try to find patterns using large amounts of data are becoming even more valuable to investors, motivating us to further investigate the topic. Given the findings of Ryll and Seidens (2019), we want to find out whether it is possible to exploit machine learning capabilities to increase investors' chance of choosing funds that possess positive alpha and perform in excess of a benchmark.[2] Hence, our goal is to predict next year's mutual fund alpha and produce a ranking system, enabling retail investors to select the most successful mutual funds. Thus, we front the following research question:

How is the applicability of machine learning to pick successful mutual funds?

To efficiently answer the research question, we examine whether it is possible to classify the sign of $alpha_{t+1}$ in the first hypothesis. This gives us an indication of whether it is possible to use machine learning to classify positive from negative alphas. The second hypothesis investigate whether we can create a successful classification system where we rank the

[1] See Carhart (1997), Fama and French (2010) and Ferreira, Keswani, Miguel, and Ramos (2013) for more evidence on the underperformance of actively managed mutual funds.

[2] To measure the outperformance, we use different benchmakrs, but more on this in chapter 4.

predicted mutual fund alphas into quintile portfolios.[3] The third hypothesis enables us to benchmark our categorization system against Morningstar's five-star portfolio to stress test our performance and legitimize the predictive power of our models.

Concerning the aforementioned hypotheses, we underline three main findings. First, we feel comfortable stating that we have managed to produce a classification system that enables us to separate negative from positive alphas. We produce a mean AUROC above 75%, supplemented with a sensitivity score above 69% across all models.[4] These scores confirm that we substantially surpass a threshold of 50%, considered to be a random guess (Mandrekar, 2010). However, we see that the tree-based methods perform particularly well compared to neural networks (Tabnet). Results from the second hypothesis finds that the top quintile portfolios produced by XGBoost and random forest yield an average annual alpha of 1.28% and 0.6%, indicating that investors can gain statistically significant risk-adjusted excess returns by selecting funds from our top quintile portfolios. Contextualizing an investment from the beginning of our hold-out set, an investor would earn a cumulative alpha from XGBoost and random forest of 7.72% and 1.44%. We also successfully and significantly outperform the asset weighted and equally weighted benchmark in the hold-out period. Investing in our best performing portfolio (XGBoost 5-star), would earn an excess cumulative alpha of 43.6% and 52.2% compared to the equally weighted and asset weighted portfolios.

Investigating the benchmarks isolated, we find that the asset and equally weighted benchmarks produce a cumulative alpha of -44.5% and -35.9%, while producing a mean annual alpha of -5.7% and -4.5%. Finally, we find evidence of significant outperformance on average and cumulative alpha to Morningstar's top quintile portfolio. When we reduce the number of annual positions in the top-quintile portfolio to match Morningstar, we further outperform on risk-adjusted measures. Knowing that the predictive model and our mutual fund ranking system performs excellent both economically and statistically, investors can buy our top quintile portfolio to gain excess alpha over the benchmarks and its lower quintile portfolios. We further perform a robustness check, where we change the

[3] A quintile is a statistical value of a data set that represents 20% of a given population, so the first quintile represents the lowest fifth of the data (1% to 20%); the second quintile represents the second fifth (21% to 40%) and so on.

[4] The area under the receiver operating characteristic (AUROC) is a performance metric which can be used for evaluating classification models.

risk-adjustment model used to predict and rank $alpha_{t+1}$. Our findings indicate that we successfully manage to reject the null hypotheses as our results stay robust.

Our research topic has been investigated to a certain extent in the hedge fund industry by J. Chen, Wu, and Tindall (2016) and Wu, Chen, Yang, and Tindall (2021), and to our knowledge, the research on this topic is not widespread in the mutual fund industry. Li and Rossi (2020) however use machine learning and detect significant predictability of mutual fund performance by utilizing 94 different predictors. However, our thesis is most related to the work of DeMiguel, Gil-Bazo, Nogales, and AP Santos (2021). The authors study whether machine learning and fund characteristics can help to select mutual funds with positive alpha in the American market. Interestingly, the authors find significant results and conclude that investors can benefit from machine learning and active mutual funds instead of holding the market portfolio.[5] The authors found evidence of their machine learning algorithms creating a monthly alpha of 0.4% in their top-decile portfolio.[6]

We have chosen 24 empirically backed mutual fund characteristics that our algorithms utilizes. Jones and Mo (2021) conducted an interesting study documenting the relationships between mutual fund characteristics and fund performance by studying the out-of-sample performance of variables to forecast mutual fund alphas. The authors find that the ability of fund characteristics to predict performance has declined over time and further argue that mutual fund competition and increased arbitrage activity might be the reason. DeMiguel et al. (2021) mention that a strong association between fund characteristics and performance not alone guarantees that mutual funds exploiting only that single characteristic would earn positive net alphas. This highlights the importance of choosing documented characteristics. Additionally, the model's weighting of characteristics can be hard to interpret in an economic context, although it is clear in a mathematical context. However, we are humble regarding the fact that our endogenous variables may have relationships that cannot be substantiated by financial theory. In this regard, our paper will not contribute to develop new characteristics nor use less popular risk-adjusting models than the ones developed by Fama and French (1992), Fama and French (1993), and Fama and French (2015), but rather aim to utilize them to answer the research question.

Our paper contributes to the literature as no paper has been published on the topic in the

[5] Holding the market portfolio refers to buying index funds that aim to replicate the market economy.
[6] Achieved by using the Random Forest to estimate the alpha, using a 24-month rebalancing technique.

Nordic markets, which is smaller and less dynamic than the American, where previous literature has been published. Furthermore, neither of the aforementioned studies have included fixed income funds as we do in this thesis.[7] The above factors is important because different factors such as investment styles, factor tilting, regulations, economic drivers, and overall geographies can be vital for the research outcome.[8] Moreover, financial institutions and practitioners often use data to supplement their decision making. We posit that only relying on foreign research papers, especially from the U.S, to consider Nordic investment decisions can lead to biases due to the different economies and dynamics between the two markets. Correspondingly, knowing that Nordic practitioners look for new ways to gain a comparative advantage, we believe that machine learning and its applicability is a trend that only becomes more important.

We present the thesis in a systematic approach. Chapter 2 presents the literature review to contextualize the goal of the thesis and to compare it against the consensus in the market. Chapter 3 presents the hypotheses development and our approach to testing the hypotheses. Chapter 4 specifies our data processing and cleaning steps, as well as important choices regarding our data. We also present descriptive statistics from our factor model regressions and the characteristics used as predictors in our machine learning models. Further, chapter 5 explains our performance-evaluation methodology regarding how we evaluate our classification and regression models. We then present the resampling techniques and an overview of the machine learning methods utilized. Chapter 6 presents findings and concludes on the three different hypotheses, enabling us to answer the research question. Chapter 7 presents a discussion on the results and a robustness check on the risk-adjustment model. We also elaborate on the weaknesses of the thesis and our view for further research. Lastly, in chapter 8, we give a conclusion to our thesis.

[7] We refer to the European Fund and Asset Management Association in https://bit.ly/3OsVpUk for proof.

[8] See Coval and Moskowitz (2001) for evidence on nearby investments and its possibility to generate substantial abnormal returns.

2 Literature Review

This chapter starts with section 2.1, where we present literature on how investors select mutual funds. Further, in section 2.2, we present literature on active management, discussing fund performance and manager persistence. At last, in section 2.3, we present evidence regarding machine learning prediction to give a thorough outline of our research topic.

2.1 Investor Fund Selection

Understanding how investors allocate capital is an essential question in the study of financial markets. Several studies have tried to explain what investors actually do when selecting funds. Berk and Van Binsbergen (2016) and Barber, Huang, and Odean (2016) find evidence that investors appear to discount performance attributable to exposure to the market factor when allocating capital according to past fund performance. However, a paper from Evans and Sun (2021) finds the argument surprising that retail investors are sophisticated enough to account for beta in their risk assessment, but not for other systematic factors.[9] The authors further suggest that if investors are unsophisticated, they will focus on total fund returns or active returns. Frazzini and Lamont (2008) documents that individual investors have a striking ability to do the wrong things, suggesting that they allocate their money to mutual funds that own stocks that perform poorly over the subsequent years. The findings of Evans and Sun (2021) suggest that investor heuristics, such as Morningstar ratings, have a significant and causal impact on investor decisions and that it drives fund flows. This suggestion is further supported by Ben-David, Li, Rossi, and Song (2019), who finds that investors appear to follow ratings blindly, not likely to understand how Morningstar constructs its ratings.

Despite the controversy of whether investors behave rationally and incorporate asset pricing models or if they follow easy-to-process signals in terms of ratings, we present a relatively new approach to selecting mutual funds. We internalize the abovementioned considerations, meaning that we predict $alpha_{t+1}$, substantiated from our risk-adjusted

[9] Such as the one found by Fama and French (1992), Fama and French (1993), and Fama and French (2015).

factor models, and additionally categorize the predicted alphas into a ranking system.[10] The aim is to present the best-performing mutual funds in our top quintile portfolio on a "silver platter" to the investor.

2.2 Active Management

On a general level, there should not exist any free lunches in the market. Researchers today mostly agree that the market is in semi-strong form, suggesting that the market price is reflected by past and all public information. Fama (1970) argues that investors in active mutual funds underperform the market and that mutual fund returns are unpredictable. Fama (1970) also argues that the market is fully efficient if all available information is embedded in the prices, including inside information.

Sharpe (1991) proves that active and passive investors must earn identical gross returns. As a result, investors have no incentive to choose active management, considering that active investors lose to passive investors after fees. In turn, this disfavors the choice of active management. However, this paper has been criticized by Berk and Van Binsbergen (2015) and Pedersen (2018), who argue that passive investors must trade to follow the market.[11] The authors suggest that passive investors can lose to active investors if they trade at prices systematically less favorable than active investors. However, Pedersen (2018) acknowledges that the although the history of active managers in general terms is not exceptional, he does not believe that passive funds will have a 100% allocation in the future.

Several papers have studied fund managers' abilities. Ferreira et al. (2013) find underperformance of equity mutual funds on average. Further, the paper finds no evidence of consistent stock-picking skill of fund managers and that only 43% of managers outperformed their benchmark, which again argues in favor of passive investing. This leads to asking why investors should invest their hard-earned capital in active funds rather than a safe and diversified index fund. Carhart (1997) endorsed the hypothesis of manager skill and found persistence only among the worst-performing funds. That led to the conclusion

[10] Knowing the different investors heuristics, we benchmark our top quintile portfolio against Morningstar's five-star portfolio in the third hypothesis, which helps to answer our research question.

[11] The authors argue that this is the case because the market change over time (IPO's, share-issuances, delistings and repurchases).

that the data did not support the existing evidence of skilled or informed mutual fund portfolio managers. Berk and Green (2004) argue that if a skill is in short supply, the net return is determined in equilibrium by competition between investors and not by managers' skills. Berk and Van Binsbergen (2015) find that the average mutual fund has used its skill to generate about $3.2 million per year, although not by measuring skill in terms of net alpha. The authors also find that, on average, active funds have a net alpha of 36 basis points p.a, when compared to index mutual funds with similar styles.[12] Their evidence can give hope to the survival of active management.

Active managers operate in a far different environment today than the managers in the older literature. Cremers, Fulkerson, and Riley (2019) reviewed the past 20 years of academic literature on the subject and found that the direct costs regarding expense ratios have decreased significantly over the past decades. He also mentions that the decline in indirect costs to investors of trading within actively managed funds has fallen. The authors conclude that the conventional literature is too pessimistic about the value of active management.

Kosowski, Timmermann, Wermers, and White (2006) examined the statistical significance of the performance and the persistence of the best and the worst mutual funds. The authors found that the performance of these managers were not solely due to luck, implying that they find that a sizeable minority of fund managers performs well enough to more than cover their costs. The authors conclude that superior alphas of these fund managers persist.[13] Another study by Fama and French (2010) finds that if there are funds that have enough skill to produce benchmark adjusted returns that cover their costs in an aggregated dataset, the evidence is hidden in the aggregate results of the bad managers with inadequate skill. In addition, if they add back the expense ratios and analyze on a gross level, there is evidence for both inferior and superior performance in the extreme tails of the cross-section of mutual fund alpha estimates. This finding gives hope for active management and mutual funds' ability to create alpha. Cremers et al. (2019) emphasize that recent research indicates that many active managers have significant observable skills, that those skills create real value for investors, and that those skills persist over time.

[12]The authors suggest that a better measure of skill is value added, and not the product of the funds abnormal return. The authors argue that mutual funds has diseconomies of scale, and that 1% gross return on a $10 billion fund adds more value than 10% on a $1 million fund.

[13]The findings are true among growth-oriented funds, but not within income-oriented funds.

Cremers et al. also reviewed fixed income funds explicitly. The authors mention that there are several studies providing evidence that active bond fund managers are skilled and that fixed income funds generate alpha before costs but provide underperformance after fees.[14] Gutierrez, Maxwell, and Xu (2009) also found fixed income funds displaying persistence in performance that is long-lived. On the other hand, Boney, Comer, and Kelly (2009) present evidence that managers are generally unsuccessful at timing the yield curve, questioning the value of bond management. Cremers et al. (2019) posit the statement that bond fund managers appear to make informed decisions on behalf of their investors, which is consistent with the findings for U.S. equity funds.

2.3 Machine Learning Prediction

If the asset owner chooses active management, he needs to find the most optimal tool for fund selection to maximize the probability of achieving positive net excess returns. Weigert (2021) studied the prediction of active mutual fund performance and concluded that it is challenging. He based his conclusion on the arithmetic of active management by Sharpe (1991), the efficient market hypothesis by Fama (1970), and the performance chasing of investors, found in Berk and Green (2004). Nevertheless, J. Chen et al. (2016) add to the existing literature on the value of active management and machine learning algorithms. The authors forecast hedge fund returns with different rebalancing frequencies, which differentiates from our paper trying to predict mutual fund alphas by defaulting a one-year rebalancing frequency. However, they find that when exposed to Carhart (1997) factor model, their machine learning portfolios generate large alphas compared to the traditional OLS and Lasso model.

Wu et al. (2021) further found evidence in favor of evaluating out-of-sample performance. The authors utilized returns-based and macro derivative features as predictors in their algorithms, specifying that their return-based characteristics lead to higher returns than the macro derivative features. The authors also mention that their forecast model yields the best performance when these two features are combined. At last, they find that deep neural networks appear to be the overall most effective out of four machine learning methods implemented.

[14] See Moneta (2015) for a comprehensive review.

3 Hypotheses Development

This chapter explains the development of our hypotheses, created to answer our research question. We base our first hypothesis on the possibility of distinguishing funds producing negative or positive alpha in the next 12 months. In *Hypothesis 2* we investigate if we can predict each funds $alpha_{t+1}$, to separate funds on the degree of alpha generated. We are interested in creating a classification system by separating funds on their alpha net of costs, aiming to replicate the selection problem that the investor face. It also allows us to differentiate a top quintile portfolio from the lower ones, giving us the preconditions to approve or disprove the hypothesis. Furthermore, in hypothesis three, we aim to stress test our best performing machine learning portfolios against Morningstar, as they are recognized as the market leader in terms of rating funds.[15] We hope the benchmarking can legitimize our model's predictive ability and visualize our relative performance to this industry benchmark.

3.1 Alpha Classification Hypothesis

As discussed in the literature review, the evidence posits that there might be a possibility for supreme managers to produce positive abnormal returns. The uplifting papers from Kosowski et al. (2006) and Cremers et al. (2019) give hope for active management, but no paper that we are aware of examines whether it is possible to classify the sign of alphas. Our first hypothesis is thus:

H0: It is not possible to classify positive alphas

H1: It is possible to classify positive alphas

To reject the null hypothesis, we have to be consistent in our prediction and produce AUROCs significantly above 0.5. Additionally, we require our results to be economically significant relative to an asset weighted and equally weighted benchmark.

[15] See Kamal et al. (2013) for insight in Morningstar's rating system.

3.2 Categorization Hypothesis

In this hypothesis, we train our data by using alpha as a target for the prediction. The machine learning algorithms supply us with predicted alpha coefficients that enable us to sort the funds on predicted alpha.[16] We systematize our predicted alphas by creating a ranking system based on quintile portfolios. To the best of our knowledge, we have not found any other papers doing this in the Nordic markets. We ask ourselves if this is due to its difficulty or due to us pioneering on this quest. Thus, our second hypothesis is the following:

H0: It is not possible to create a successful ranking system

H1: It is possible to create a successful ranking system

To reject the null hypothesis, we have to be consistent in our prediction and produce top quintile portfolios that outperform the benchmarks and their lower quintiles. We also require our results to be statistically and economically significant.

3.3 Endurance Testing Hypothesis

As argued in the literature review, we know that investor heuristics, such as Morningstar ratings, have a significant causal impact on investor decisions (Evans & Sun, 2021). This motivates us to compare our top quintile portfolios against Morningstar's top quintile portfolio. We hope that this benchmarking process substantiates our model's predictive ability, if not else, demonstrating better performance than Morningstar. Additionally, several studies investigate whether Morningstar's analyst ratings can predict future mutual fund performance. Blake and Morey (2000) found that Morningstar was able to predict low-performing funds and further found weak statistical evidence that the five-star funds outperform the four and three-star funds.

A study by Kräussl and Sandelowsky (2007) found that the rating system of Morningstar did not beat the hypothesis of the random walk Fama (1970). However, Kamal et al. (2013) find, contrary to Kräussl and Sandelowsky (2007), that Morningstar's analyst ratings are significantly positively related to the future performance of funds, measured

[16] An explanation of how we measure α is found in chapter 4.

by the 3-year alpha.[17] We find this last hypothesis interesting to test, as the evidence for Morningstar's predictable ability is disputable. We front the following hypothesis:

> ***H0: Our top quintile portfolio will not outperform Morningstar's five-star portfolio***
>
> ***H1: Our top quintile portfolio outperforms Morningstar five-star portfolio***

We use both statistical and economic measurements to test this hypothesis, similar to *Hypothesis 2*, where we test the statistical performance from the top-quintile portfolio to its lower quintiles. Furthermore, we investigate the different economic measurements, as shown in table 5.4, to have a consistent evaluation method. We can successfully accept the alternative hypothesis if we manage to outperform Morningstar's five-star portfolio both statistically and economically. This gives substance to our models and helps to legitimize both the predictive ability of the machine learning models and the economic rationale for investing in our best portfolios.

[17] We again emphasize that we benchmark against their star-rating, and not analyst-rating. We refer to the previous abbreviations for their differences.

4 Data

In this chapter, we describe the data used in our analysis. We also elaborate on our data sources, pre-processing steps, and data limitations. In section 4.1, we describe the data extracted from the Morningstar Direct database.[18] Further, in section 4.2, we describe the predictors implemented to capture information from both fund and market characteristics. At last, section 4.3 present the descriptive statistics for the dataset.

4.1 Morningstar Direct Data

We extract both annual and monthly data of Nordic active mutual funds from the Morningstar Direct database.[19] Given our interest in selecting funds outperforming a factor benchmark, we exclude index funds from the analysis.[20] Additionally, we exclude fund-of-funds and closed-end funds from the analysis to avoid lock-up periods and restricted portfolio rebalancing. Consequently, the analysis comprises of open-end funds, which allow for ongoing new contributions and withdrawal from investors, supporting the yearly rebalancing used in this study. The returns retrieved from the database are net of expenses, but since we want to pragmatically proxy for the investment universe an investor faces, we include both true no-load and load funds.

On the contrary to DeMiguel et al. (2021), we aggregate fund share-classes to a single fund-level unit. All fund share-classes have the same holdings and fund-specific characteristics but different returns after expenses. Hence, we aggregate the share-level metrics by a weighted average on the proportion invested in each share-class. We also exclude funds with less than three years of existence to reduce the impact of volatile early cycle returns. Additionally, we make the dataset robust to the survivorship bias by including funds that have been liquidized, closed, merged, or acquired. Furthermore, all funds without ISIN are removed from the dataset, as this will be a unique identifier for each fund. Moreover, a handful of funds without management history is removed from the dataset. For *hypothesis 3*, we extract monthly data of Morningstar's overall star rating for the

[18] Morningstar Direct is an investment analysis platform that have specialized in fund data and analysis at a large scale.

[19] We define the Nordic market as Denmark, Sweden, Norway and Finland.

[20] In this thesis we use factor benchmarks to determine alpha. Our default benchmark is FF6F, but we additionally run FF3F as a robustness checks in section 7.

funds in our investment universe.[21] We aggregate this to an annual metric by averaging the rating for every fund in every year.

The final dataset consists of 25 839 unique annual and 270 049 monthly observations distributed across 2748 mutual funds. Of these funds, 1964 are diversified equity mutual funds, 702 are fixed income funds, and the residual 82 funds are characterized by having an alternative investment strategy.

4.2 Response and Predictors

A substantial part of this study is the collection, structuring, and preparation of the data used in the machine learning algorithms. The inclusion of fund specific characteristics with capabilities of explaining mutual fund excess returns is vital for the machine learning algorithms being able to predict fund alphas. Additionally, the exclusions of predictors with high correlation are essential for mitigating any multicollinearity bias. We refer to figure A3.1 for a visualization of this.

Research on the performance of mutual fund predictors is not the aim of this thesis, which is why the selection of predictors is based upon the extensive research of others. Several studies have investigated the association between various theoretically motivated variables to predict fund returns. The results are mixed, and from this, a moderate number of variables have shown to do so Jones and Mo (2021); Li and Rossi (2020); DeMiguel et al. (2021). This thesis utilizes acknowledged predictors that have proved to have good out-of-sample performance to reduce the risk of biases and data-snooping.[22]

In the following paragraphs we provide an explanation of the computation of main predictors among our 24 fund characteristics.[23] Table 4.1 provides an overview of the fund-level predictors included in the final machine learning models. We refer the interested reader to appendix part A2 for a short introduction on the rationale for including these specific predictors.

[21] For more information see: https://s21.q4cdn.com/198919461/files/doc_downloads/2019/07/MRQ_Ratings_Infographic_070219.pdf.

[22] Data snooping is a form of statistical bias where you manipulate data or analysis to artificially get statistically significant results.

[23] Note that we have cross-sectionally winsorized the two year cumulative return variable at the 1st and the 99th percentiles. This means that we replace extreme observations that are below the 1st percentile or above the 99th percentile with the value of those percentiles, due to outliers biasing our dataset.

Table 4.1: Fund characteristics

The table summarizes the 24 characteristics used as predictors (independent variables) in the machine learning models. We also denote and cite each predictor.

Predictor	Predictor description	Citation
Recent returns		
Ret2	Two year cumulative return	
Alpha	Realized alpha of the past 12 months'	Jensen (1968)
AlphaStat	Alpha (t-statistic)	Fama and French (2015)
MarkBeta	Market beta (t-statistic)	Fama and French (2015)
ProfBeta	Profitability beta (t-statistic)	Fama and French (2015)
InvBeta	Investment beta (t-statistic)	Fama and French (2015)
SizeBeta	Size beta (t-statistic)	Fama and French (2015)
ValBeta	Value beta (t-statistic)	Fama and French (2015)
MomBeta	Momentum beta (t-statistic)	Carhart (1997)
DefBeta	Default premium beta (t-statistic)	Fama and French (1993)
TermBeta	Term premium beta (t-statistic)	Fama and French (1993)
R2	R-squared of the past 12 months' returns	Amihud and Goyenko (2013)
Risks		
Sharpe	Sharpe ratio of the past 12 months' returns	Sharpe (1966)
Skew	Skewness of the past 12 months' returns	Wu et al. (2021)
Kurt	Kurtosis of the past 12 months' returns	Wu et al. (2021)
M2	M2 of the past 12 months' return	Modigliani and Modigliani (1997)
TE	Tracking Error	Gupta, Prajogi, and Stubbs (1999)
IdioRisk	Idiosyncratic risk of the past 12 months' returns	Gu, Kelly, and Xiu (2020)
VIX	Average of the VIX index in the past 12 months	Wu et al. (2021)
Fund management		
ManTen	Manager tenure	Weigert (2021)
ManSize	Size of fund management team	Weigert (2021)
Expense	Funds annual expenses	Nanigian (2012)
IR	Information Ratio	Gupta et al. (1999)
Fund characteristics		
TNA	Total net assets	DeMiguel et al. (2021)
FundType	Fixed Income/Equity Mutual fund (Factor)	

As stated in section 4.1, 26% of all mutual funds contained in the dataset are fixed income funds. These funds have a different risk exposure than the equity mutual funds. Instead of regressing the net excess returns on the FF5F augmented with momentum, which is our base case for the equity mutual funds,[24] we follow Bauer, Christiansen, and Døskeland (2022) and use credit factors as regressors for the net excess returns.[25] Conversely, alphas and betas for fixed income funds are computed on the credit premium factor and the term premium factor as:

$$\alpha_{i,m} = r_{i,m} - \hat{\beta}_{TERM_{i,m}} TERM_m - \hat{\beta}_{DEF_{i,m}} DEF_m \tag{4.1}$$

[24] Including sector funds, which is funds that invest in only one type of industry or sector.

[25] The credit factors are sourced from: https://www.nbim.no/en/publications/reports/2021/annual-report-2021/.

where $\alpha_{i,m}$ is the estimated alpha for fixed income fund i in months m. $r_{i,m}$ is the net excess return of fund i in month m, $TERM_m$ and DEF_m are the returns in the mth month of the two credit factors, and $\hat{\beta}_{TERM}$ and $\hat{\beta}_{DEF}$ are the factor loadings of the ith share class excess return with respect to the fixed income regressions. For the remaining equity mutual and sector funds, we compute the monthly realized alpha for the ith share class in the mth month ($\alpha_{i,m}$) as

$$\begin{aligned} \alpha_{i,m} = r_{i,m} - \hat{\beta}_{MKT-RF_{i,m}} MKT - RF_m - \hat{\beta}_{SMB_{i,m}} SMB_m - \hat{\beta}_{HML_{i,m}} HML_m \\ - \hat{\beta}_{RMW_{i,m}} RMW_m - \hat{\beta}_{CMW_{i,m}} CMW_m - \hat{\beta}_{MOM_{i,m}} MOM_m \end{aligned} \tag{4.2}$$

where the $MKT - RF_m$, HML_m, SMB_m, RMW_m, CMA_m and MOM_m are the returns in month m of the Fama-French and momentum factors. Further, $\hat{\beta}_{MKT-RF_{i,m}}$, $\hat{\beta}_{SMB_{i,m}}$, $\hat{\beta}_{HML_{i,m}}$, $\hat{\beta}_{RMW_{i,m}}$, $\hat{\beta}_{CMA_{i,m}}$ and $\hat{\beta}_{MOM_{i,m}}$ are the factor loading's of the ith share class with regards to the Fama French five factor model and momentum.

We compute alphas by performing a rolling window regression in the month m - 36. We follow DeMiguel et al. (2021) and calculate the annual realized alpha by adding the monthly realized alphas in each calendar year. From the rolling-window regressions of monthly excess returns on the equity and fixed income factors, we obtain a series of return-based characteristics. These include alpha, the *t-statistics* from the factor loadings of the ith share class, as proposed by Hunter, Kandel, Kandel, and Wermers (2014). We also retrieve the R-squared consistent with Amihud and Goyenko (2013) findings. Utilizing both alpha and beta *t-statistic* as our predictors, we better capture the level of a fund's exposure to the fund characteristic. Beyond the aforementioned return based characteristic, we compute a series of risk characteristics based on monthly frequencies. We compute a funds idiosyncratic risk as the residual sum of squares from our rolling-window regression of monthly excess returns against the factor models. Furthermore, we extract skewness, kurtosis, M2, and tracking error from the Morningstar Direct database.

After accounting for the return and risk characteristics, we compute a series of characteristics related to the fund and its management team, as Weigert (2021) suggest. Additionally, the research of Chevalier and Ellison (1999) shows that managerial characteristics are important when explaining fund performance, which is why our study examining the possibility of selecting high-performing mutual funds should account for

managerial factors. To capture fund-management characteristics, we include a series of management related characteristics proven to have a relationship with fund returns. For every fund, we extract detailed fund management data, including managers' names, date of designation, and departure from the fund. These details range from the fund's inception date to its conclusion date. By utilizing natural language processing, we are able to compute a time series variable on manager tenure as the years from which a fund manager was designated to the year of prediction for a given fund. For management teams, management tenure is calculated as the arithmetic mean of manager tenures of all current team members in the fund's management team at the point of prediction. Ideally, we could have created a detailed characteristic based on each team member's contribution, we do unfortunately not possess such data. Lastly, management team size is computed as the number of managers with the fund at the at period t, to create a time series variable.

The data availability of expense ratios from the Morningstar Database are sparse, which is why we compute a funds expense ratio as the difference between the Morningstar's net return and gross return. As the analysis is conducted on a fund-class level, differences in expense ratio within share-classes is aggregated to a single unit by a weighted averaged on the proportion invested in each share-class.[26]

4.3 Descriptive Statistics

This section presents descriptive statistics from our data. First, table 4.2 presents descriptives from our 24 predictors before we present alpha distributions from the rolling-window Fama French regression in figure 4.1.

[26]For more information see: https://awgmain.morningstar.com/webhelp/glossary_definitions/mutual_fund/Gross_Return.htm.

Table 4.2: Fund characteristics descriptives

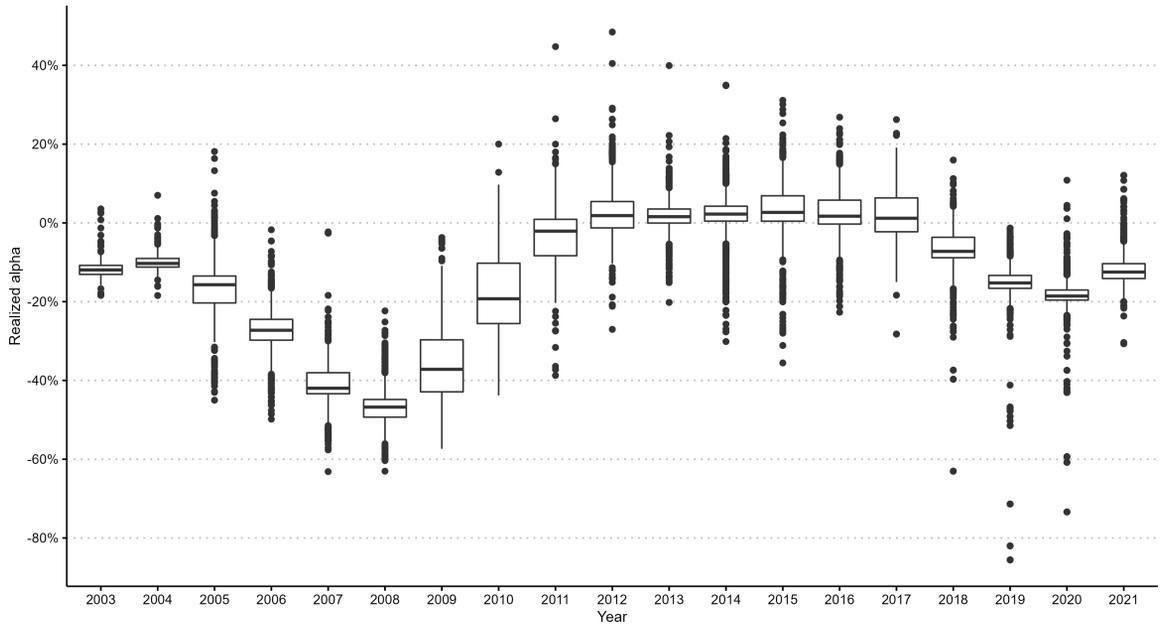
This table shows summary statistics on the predictors presented in table 4.1. Computations ranges the entire period, from 01/01/2002 through 12/31/2021. The table visualizes the mean value, the median value, standard deviation, 0 percentile, 25th percentile, 50th percentile, 75th percentile and 100th percentile.

Fund Characteristic	**Mean**	**Median**	**Std.Dev**	**P0**	**P25**	**P50**	**P75**	**P100**
Ret2	0.15	0.09	0.25	(0.46)	0.03	0.09	0.24	1.09
Alpha	(0.12)	(0.09)	0.16	(0.86)	(0.20)	(0.09)	0.01	0.48
AlphaStat	(3.02)	(1.98)	5.21	(65.57)	(5.02)	(1.98)	0.29	34.52
R2	0.49	0.53	0.23	0.00	0.32	0.53	0.68	0.98
Sharpe	(0.22)	0.83	24.84	(840.48)	(0.95)	0.83	2.20	703.88
Skew	(0.25)	(0.24)	0.82	(3.44)	(0.73)	(0.24)	0.26	3.46
Kurt	0.63	0.17	1.90	(2.25)	(0.69)	0.17	1.36	11.99
M2	0.06	0.05	0.13	(0.67)	(0.01)	0.05	0.13	0.82
TE	0.11	0.11	0.05	0.00	0.08	0.11	0.14	0.65
IdioRisk	0.08	0.08	0.05	0.00	0.04	0.08	0.11	0.49
VIX	19.18	16.67	6.34	11.09	14.23	16.67	22.55	32.70
ManSize	1.35	1.00	0.99	0	1.00	1.00	2.00	16.00
ManTen	5.25	4.00	4.30	0	2.00	4.00	7.00	37.00
Expense	0.011	0.011	0.008	0	0.005	0.011	0.015	0.185
IR	(0.73)	(0.63)	1.57	(7.43)	(1.53)	(0.63)	0.18	9.05
TNA (MNOK)	3 414.63	975.62	7 480.17	0	309.04	975.62	3 110.64	98 537.50
MarkBeta	4.09	3.86	2.41	(8.57)	2.63	3.86	5.12	29.53
ProfBeta	0.21	0.10	1.12	(4.53)	(0.53)	0.10	0.87	4.16
InvBeta	(0.34)	(0.41)	1.22	(5.57)	(1.16)	(0.41)	0.49	6.52
SizeBeta	0.25	0.27	1.30	(6.00)	(0.63)	0.27	1.17	5.37
ValBeta	0.03	(0.00)	1.02	(3.93)	(0.62)	(0.00)	0.65	4.57
MomBeta	(0.06)	0.04	1.18	(5.30)	(0.70)	0.04	0.74	5.98
DefBeta	1.45	1.02	2.32	(3.68)	0.00	1.02	2.45	28.17
TermBeta	2.24	1.78	2.49	(3.69)	0.39	1.78	3.68	21.56

Figure 4.1 displays the estimated alpha from the FF6F rolling-window regression across all years in our dataset:

Figure 4.1: Alpha distributions

The figure illustrates the distribution in the actual alpha of the funds contained in the dataset by regressing the default FF6F. The box plot show the negative outliers, minimum, the first quartile, median, the upper quartile, the maximum, and the positive outliers. The minimum is computed by $Q1 - 1.5 \times IQR$, and the maximum by $Q3 + 1.5 \times IQR$.



5 Methodology

This chapter is divided into two parts. The first part elaborates on how the machine learning algorithms are evaluated and cross-validated.[27] The second part provides an overview of the machine learning algorithms, where we provide a short outline on the machine learning methods utilized in this thesis. We want to emphasize that machine learning theory is not the prospect of this thesis. Hence, we will not elaborate on the topic in detail but rather briefly introduce our data management and the methods implemented.[28]

5.1 Performance Evaluation and Validation

This section describes the process employed for splitting the dataset with regards to the training of our machine learning models. We also present the process for evaluating and comparing our models on independent data. Furthermore, we describe the different performance evaluation metrics utilized across the three hypotheses.

5.1.1 Performance-Evaluation Methodology

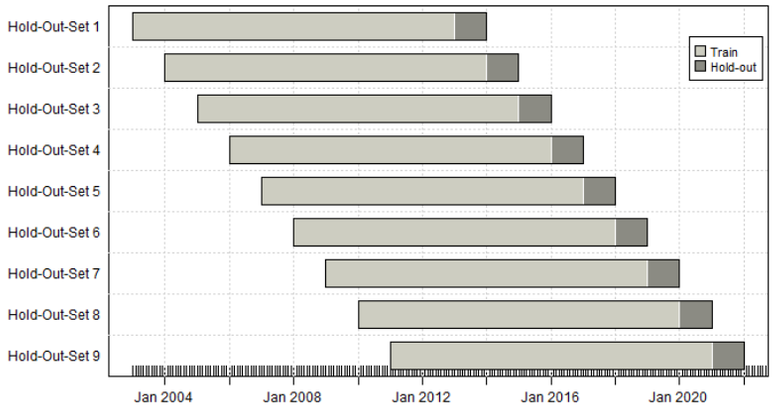
This section describes the procedure used to evaluate our model performance with respect to the different hypotheses. We use a well-known model building approach that encompasses splitting the data into two parts. The first part is used for training the model, and the latter is used to evaluate model performance out-of-sample (Kuhn, Johnson, et al., 2013). This structure simulates model deployment in real life, where the model will be tuned on the data readily available at the point of prediction, and evaluation will occur on data not yet seen by the model. We use a 11-year rolling window to train and evaluate the performance of our model, where the first 10-years will be training data, and the 11th year is validation data, characterized as a hold-out-sample. The rolling train-to-validation split is illustrated in figure 5.1:

[27] Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

[28] We use the well-known packages Caret and Tidymodels and refer the interested reader to the package documentation for more detailed descriptions.

Figure 5.1: Train and hold-out-set splits

The figure shows the rolling train-to-validation split utilized for simulating real-life implementation of the portfolios across the evaluation period of 9 years. The figure visualizes each hold-out-sets training and hold-out period. The training set is used to train and cross-validate the model, and the hold-out-set to evaluate the models performance out-of-sample.



From our dataset time frame, ranging from 2002 through 2021, we are able to produce a series of 9 validation sets used to test and compare the predictive ability over a period of time. Subsequently, we obtain 9 observations of each performance measure, for the years 2013 through 2021. The models will be evaluated with respect to single hold-out-set metrics and metrics across the entire hold-out-sample. From a financial perspective evaluating the model across different market conditions is essential to detect and mitigate biases. Financial market dynamics and performance vary across time, which is why we deem it crucial in the evaluation of our models. As visualized in figure 4.1, we observe that the distribution of our alpha coefficient greatly varies over time. This is not exclusively due to lower mutual fund returns, but the systematic risk exposure of the factor loadings and the price of risk have changed. This substantiates the fact that having a rolling train-to-validation split makes the modeling realistic regarding changes in market dynamics. In the following subsections, we aim to describe which metrics is used to evaluate the hold-out-sample results with respect to the three hypotheses developed in chapter 3.

Evaluating Classification Models

This subsection presents the measures applied to evaluate hypothesis 1, where the aim is to classify whether a fund will produce a negative or positive alpha in the next 12 months. For such a problem, it is important to choose an accuracy measure that measures

how well the classifier (predictive classification model) distinguishes between the classes, regardless of their relative proportion. In simple terms, we want an accuracy measure that is unbiased to imbalance in the number of observations in each class. For a binary classification problem such as the one in our thesis, the AUROC (Area Under the Receiver Operating Characteristic) is the most popular measure that corrects for this bias. The AUROC is a probability curve of the *true positive rate* (TPR) against the *false positive rate* (FPR) at various thresholds. The TPR and FPR are best explained by a confusion matrix, as illustrated in table 5.1:

Table 5.1: Confusion matrix

The figure illustrates a confusion matrix. The confusion matrix is further used for computing performance evaluation metrics for the classifier implemented in hypothesis 1.

		Actual	
		Positive Alpha	Negative Alpha
Predicted	Positive Alpha	True positive	False positive
	Negative Alpha	False negative	True negative

The confusion matrix divides the absolute prediction for each class, positive or negative alpha, into correct or false predictions. The rates are obtained by dividing the count of the predictions by the count within the actual class. Table 5.2 summarizes the key measures computed from the confusion matrix:

Table 5.2: Classifier evaluation metrics

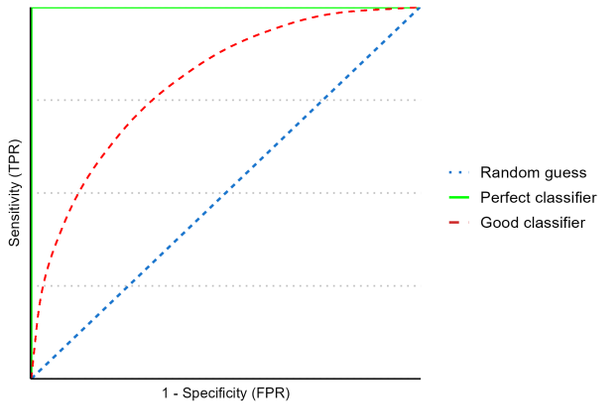
The table illustrates key classifier metrics and their formula. We refer the interested reader to the cited papers for a more in-depth explanation of the metrics than we provide in this thesis.

Measure	Definition		Citation
Sensitivity/TPR	$\frac{True\ Positive}{True\ Positive + False\ Negative}$	(5.1)	Altman and Bland (1994) Fawcett (2006)
FPR	$\frac{False\ Positive}{False\ Positive + True\ Negative}$	(5.2)	Altman and Bland (1994)
Specificity/TNR	$\frac{True\ Negative}{True\ Negative + False\ Positive}$	(5.3)	Altman and Bland (1994) Fawcett (2006)
FNR	$\frac{False\ Positive}{False\ Positive + True\ Negative}$	(5.4)	Altman and Bland (1994)
Prevalence	$\frac{TP + FP}{TP + FP + TN + FN}$	(5.5)	Kuhn et al. (2013)

The Receiver Operating Characteristic Curve (ROC) (Fawcett, 2006) is an extension of the confusion matrix and the measures presented in table 5.2, used to compute the AUROC measure. This measure presents the classifier's overall performance summarized over all possible thresholds. For this reason, AUROC is the favored classification metric that offers benefits of independence of class frequency or specific false negative/positive costs (Moro, Cortez, & Rita, 2014); (Martens & Provost, 2011). A classifier able to surpass a random guess, meaning an AUROC of above 0.5, is considered informative, and an AUROC of above 0.7 is considered a good model (Lingo & Winkler, 2008).

Figure 5.2: ROC curve example

The figure illustrates three different ROC curves, at three different levels. The Area Under the ROC (AUROC) measures the accuracy of the classifiers implemented in *hypothesis 1*. The perfect classifier, illustrated by the green line, returns an AUROC of a 100% as all predictions are classified correctly. The red curve, shows a good performing model, which is able to surpass an naïve approach, involving random guessing, which could yield an AUROC of 50%.



To answer *hypothesis 1*, the main source of evaluation will be the AUROC since the use of financial evaluation measures can be misleading for a classification problem. This is due to the evaluation measures not correcting for the class imbalance biases (Japkowicz & Stephen, 2002). Because the alpha distribution is imbalanced in most years, an evaluation measure biased towards class imbalances will likely not give the best tuned model nor give a fair representation of the prediction results in the hold-out-set (James, Witten, Hastie, & Tibshirani, 2013).

In addition to the AUROC, sensitivity, and specificity, the prevalence will be used to quantify the proportion of true positive alphas relative to true negative alphas in the hold-out-set. The prevalence aims to exhibit the prerequisite for model performance in the year by quantifying the class imbalance. Furthermore, the PPV (Predicted Positive

Value) is computed to quantify the model's ability to identify true positive alphas in the hold-out-set. The PPV defines the proportion of predicted positive alphas that actually was positive and reflect the post-prediction probability of a positive alpha, given a positive alpha prediction (Altman & Bland, 1994).

$$PPV = \frac{Sensitivity * Prevalence}{Sensitivity * Prevalence \ + \ (1 - Specificity) * (1 - Prevalence)} \tag{5.6}$$

Evaluating numeric prediction models

This section presents the measures applied to evaluate the second and third hypotheses that encompass a numeric prediction outcome.

To evaluate the machine learner's ability to predict mutual fund alphas, we start by evaluating the models in a statistical context. As illustrated in table 5.3, we utilize two measures: the root mean squared error (RMSE) measure the predictive accuracy, and the Spearman's rank correlation coefficient (r_s/Spearman's Rho) measures the model's ability to rank new observations (Kuhn et al., 2013).

Table 5.3: Evaluation metrics for numerical predictors

The table visualizes the metrics used to evaluate the predictive ability of the numeric prediction models implemented in hypotheses 2 and 3. We refer the interested reader to cited papers for a more in-depth explanation of the metrics, than provided in this thesis.

Measure	Definition		Citation
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$	(5.7)	Chai and Draxler (2014)
Spearman's Rho	$1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n(n^2-1)}$	(5.8)	Spearman (1961)

The RMSE is a function of the model's residuals, signifying the observed minus the predicted values. The formula for RMSE is presented in equation 5.7, where y_i is the observed value, and $\hat{y}_i$ represents the predicted value.[29]

In contrast to the RMSE, Spearman's Rho does not measure accuracy, but the degree of correlation between hierarchically ranked variables. The properties of the measure

[29] We refer the reader to Chai and Draxler (2014) for a more detailed description of the RMSE.

have values between -1 and 1, where 0 indicate no correlation. In simple terms, strongly positive r_s indicate that high ranks of actual alpha coincide with high ranks of predicted alpha. On the contrary, a strongly negative r_s indicate that high realized alpha frequently occurs with low ranks in predicted alpha. Equation 5.8 presents the formula for r_s, where R_i represents the rank of y_i, Q_i represents the rank of $\hat{y}_i$. Notations y_i and $\hat{y}_i$ represent the actual alpha value and the value predicted by the model for the i-th sample. With respect to the second hypothesis, highly positive Spearman's rank correlations would showcase the model's ability to rank fund alpha.

Besides the statistical measures, we evaluate the performance of the algorithms in an economic context. Hence, we compute out-of-sample net-alphas of the constructed quintile portfolios and benchmark the performance across the quintile portfolios, and against two constructed benchmarks. Additionally, we compute the return-based characteristics presented in table 5.4:

Table 5.4: Performance measurement analysis

The table present economical measures used as a supplement to alpha when evaluating the performance of the constructed machine learning and benchmark, mutual fund portfolios. Metrics are computed on monthly values across the entire hold-out-sample and annualized. For the mean and geometric returns, R_t is the return of period t, and n the number of periods. Further, the R_p is the expected portfolio return, R_f the risk-free rate, σ_p the standard deviations of the portfolio, and σ_{p_D} the downside standard deviation of the portfolio.

Measure	Definition		Citation
Mean return	$\sum_{t=1}^{n} \frac{R_t}{n}$	(5.9)	
Geometric return	$\sqrt[t]{R_1 R_2 \cdots R_t}$	(5.10)	
Sharpe ratio	$\frac{R_p - R_f}{\sigma_p}$	(5.11)	Sharpe (1966)
Sortino ratio	$\frac{R_p - R_f}{\sigma_{p_D}}$	(5.12)	Rollinger and Hoffman (2013)
Cumulative return	$\frac{P_t - P_{t=0}}{P_{t=0}}$	(5.13)	

We emphasize that our target variable is alpha, which does not necessarily coincide with the financial performance metrics proposed in table 5.4. However, it is interesting to evaluate the funds on return based characteristics as these measures supplement alpha.

By contextualizing these performance measures with the alpha, we get more financially confident that our top quintile portfolios are overall well-performing.

5.1.2 Resampling Techniques

This subsection introduces the data resampling techniques utilized in the thesis. We provide a short overview of key terminology such as training, testing, and validation set. Further, we introduce the cross-validation techniques used for tuning the machine learning models.

The machine learning models utilized in this thesis are highly adaptable and capable of modeling complex relationships, which mean that they can very easily overemphasize patterns which are not reproducible (Kuhn et al., 2013). This problem is generally characterized as overfitting, which if not corrected for will reduce the usability of the model. This is especially relevant for our thesis as the stock market is highly dynamic and overfitting would result in an overstatement of the true predictive power of the model. To control this flexibility, the models implemented use a series of tuning parameters that govern the model's complexity and where poor choices may cause overfitting (Kuhn et al., 2013). To find the optimal values of each tuning parameter, we perform a grid search which involves searching through a range of candidate tuning parameters.[30] This is executed by applying the tuning parameters to the training set and evaluating the model's predictive performance with this set of parameters. However, when evaluating the model on the same data it was trained upon, we could risk overfitting the data to a relationship only present in the training data. Consequently, our model performance would likely be poor in hold-out-samples and in real-life implementations.

In order to correct for the overfitting risk, we use a well-known model-building approach that encompasses tuning of model parameters and evaluation. The overall goal is to find a reproducible structure in the data. This method involves repeatedly dividing the training data into two sections; a training set and a test set.[31] The training set has the distinct purpose of tuning model parameters, and the test set has the purpose of evaluating predictive performance (Kuhn et al., 2013). The predictive performance on the test set is then aggregated into a performance profile, where key accuracy measures are

[30] A Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters.
[31] Be aware that the train and test set are both components of the training set presented in figure 5.1.

extracted and compared to other tuning parameter combinations. The combination of tuning parameters producing the best performance in the test set are then chosen and applied to the entire training set to produce the final model.

The target of this thesis involves a time-series component, which is why we are careful in disregarding the time factor by randomly sampling data into folds, as done by most cross-validation techniques. When using random samples, we would take the risk of future-looking when training our model, meaning that we would have used values from the future to predict values from the past. In a time series such as fund returns and the characteristics computed from these returns, there might be a temporal dependency between observations, which is why we must account for those relations when evaluating model performance in the test set (Bergmeir & Benítez, 2012). To preserve the time series relationships, our models are trained using time series cross-validation, which separates itself from regular cross-validation by the fact that the test set is always ahead of the training set. In this thesis, we use a look-back period of 4 years and a test period of 1 year, which replicate the real-life implementation of predictions, as explained in section 5.1.1, but on a smaller window:

Figure 5.3: Time series cross-validation

The figure illustrates a time series cross-validation as implemented in training of our machine learning models. *Train data* denotes the closest temporal 10-years of data readily available at point of prediction, as illustrated in figure 5.1

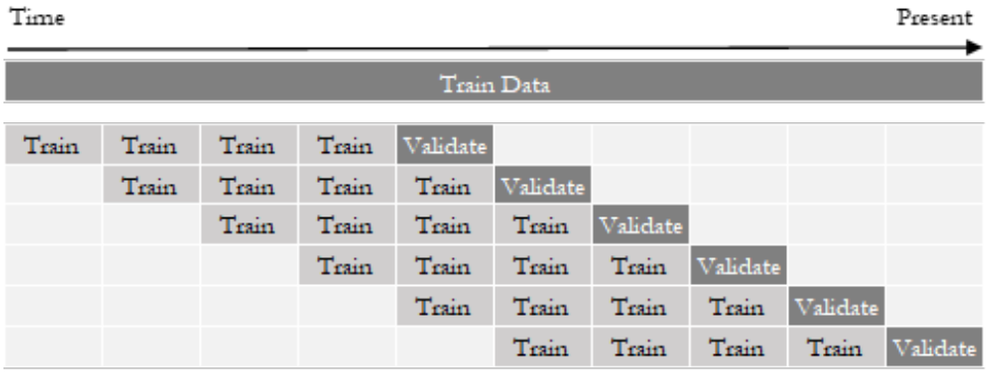


Figure 5.3 illustrates time series cross-validation implemented in the training of our model. The method involves a rolling origin forecast, where the first 4 time series components is used as the training set, and the subsequent component is used as the test set. Every set of parameter combinations is applied to all combinations of the train and test sets, before the performance measures are aggregated.

5.2 Machine Learning Algorithms

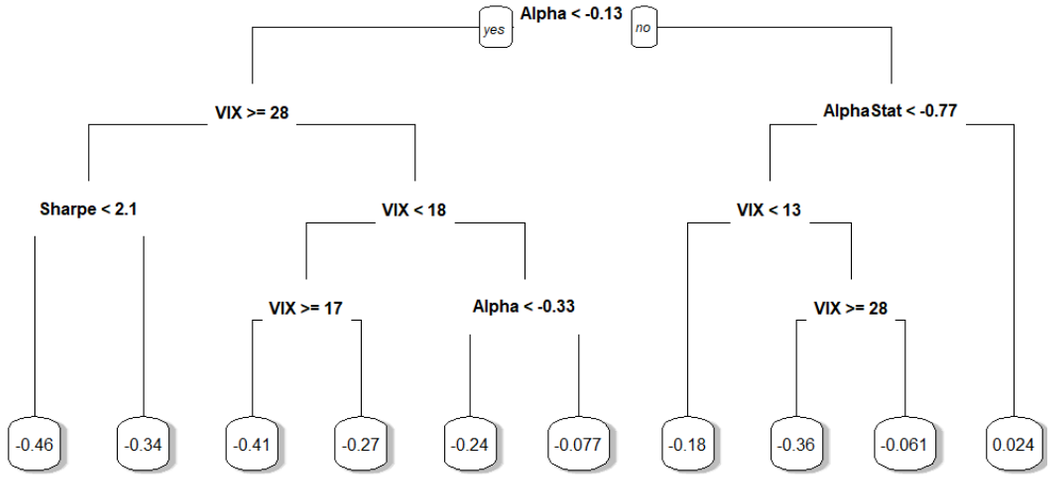
We utilize the three supervised machine learning methods, random forest, gradient boosting, and neural networks, which are selected for their capability of capturing complex non-linear relationships (Ryll & Seidens, 2019). Random forest and gradient boosting are both decision trees, which often perform well on structured (tabular) data, as the dataset constructed for this thesis. We also implement neural networks, which tend to perform well on non-structured or highly non-linear data. We move away from the previous literature on this particular subject and implement a recently developed neural network, tabnet, which is a neural network specialized for structured data (Arık & Pfister, 2021). Neural networks differ from decision trees as they apply many tuning parameters to capture non-linearities, which is why they require a large number of observations to deliver accurate estimates. As a result, neural networks are not as well suited to our dataset as decisions-trees. However, we want to test the tabnet algorithm, as its authors prove outperformance on tabular data in comparison compared to decision-trees such as random forest and XGBoost (Arık & Pfister, 2021).

5.2.1 Random Forest

Random forest is a decision tree model, which is a supervised substrata of machine learning algorithms that involve stratifying, or segmenting, the predictor space into several simple regressions (James et al., 2013). The procedure for generating these regions is often illustrated in a tree, where a sample is split at each node based on the characteristic that is most important at the specific node. The tree expands from the root-node to the leaf-nodes, where predictions are the average value of the target variable, of the observations at each leaf node. A simple decision tree is presented in figure 5.4:

Figure 5.4: Decision tree example

The figure presents an example of a singular decision tree, which is utilized by both random forest and XGBoost.



Random forest is an extension of decision trees that make ensembles of decision trees formed by bootstrap aggregation (Breiman, 2001). Decision trees are praised for being highly interpretable, but their out-of-sample performance can be poor due to the high variance of their predictions. By utilizing a bootstrap aggregation, random forest can improve its prediction accuracy drastically. In simple terms, the algorithm involves applying multiple decision trees trying to predict alpha, where the independent variables used for each model are randomly selected and where the predictions made by each model are averaged to produce the final predictions. Because the trees comprise of only a subset of the characteristics, predictions from the different trees will be less correlated than regular bagging trees, reducing the variance of predictions.

In this thesis, we train 1000 decision trees for each hold-out-set and use bootstrap with resampling to select the observations included in each tree. We use time-series cross-validation as explained in section 5.1.2 to tune the number of characteristics chosen and the required minimum number of data points in a node for the node to be split further.

5.2.2 Extreme Gradient Boosting Machines

Extreme Gradient Boosting Machines (XGBoost) is another extension on decision trees, equivalently to random forest. However, in contrast to random forest, XGBoost works by aggregating trees sequentially in order to give more influence to observations poorly

explained by its previous trees. In simple terms, all trees in random forest try to explain the same variation in the target variable, whereas in XGBoost, the next three try to explain what could not be explained by the previous trees (Mayr, Binder, Gefeller, & Schmid, 2014). Consequently, gradient boosting trees should achieve improved predictions by reducing prediction variance and prediction bias (Schapire & Freund, 2012).

For gradient boosting we apply the XGBoost-package developed by T. Chen et al. (2015). We train 1000 decision trees for each hold-out-set and tune all parameters with time series cross-validation.

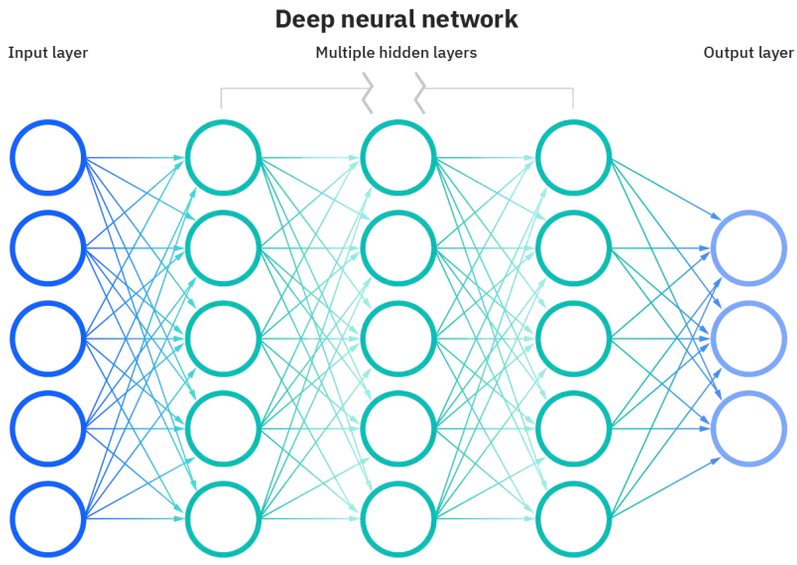
5.2.3 Deep Neural Networks

Deep learning is substrata of machine learning which applies multi-layered neural networks to extract useful features from raw data. Deep learning mimics the thinking patterns of humans to learn the patterns of the data without pre-programming of the rules (IBM, 2020).

Deep neural networks are based on multi-layered interconnected nodes (neurons) and comprises of three main layers, the input layer, the hidden layer, and the output layer. The input and output are considered the visible layers. The input is where the network takes data in for processing, and the output layer is the results of the classification or regression. The hidden layers are where the data is transformed to make predictions through forward and backwards propagation (Kuhn et al., 2013). We illustrate a simple structure of a deep neural network in figure 5.5:

Figure 5.5: Structure of Deep Neural Network

The figure presents a simple example of the structure of a deep neural network (IBM, 2020).



For deep learning we utilize the newly developed torch library, and the relatively new tabnet deep learning algorithm developed by Arık and Pfister (2021). This algorithm takes a new approach to deep neural networks with an algorithm optimized for tabular data, as will be used in this thesis.

6 Results

In this chapter we presents results from our hypotheses sequentially. In section 6.1 , we present the *Alpha Classification Hypothesis*. In section 6.2, we explore the results of the *Categorization Hypothesis*. At last, section 6.3 investigate the results of the *Endurance Testing Hypothesis*. In each subsection, we draft the premise that we base our hypothesis on and conclude on the results in regards to the respective hypothesis. This allows us to reject or accept the hypothesis, enabling us to answer the research question.

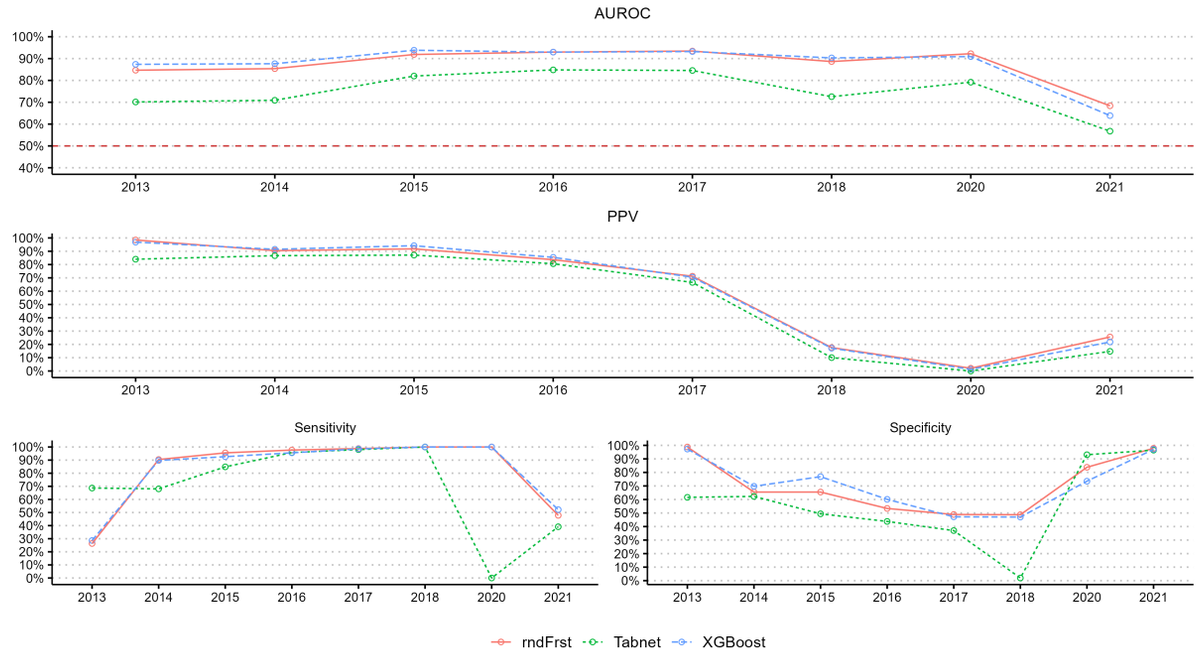
6.1 Alpha Classification

This section presents results of the *Alpha Classification Hypothesis* using the classifiers: XGBoost, random forest, and tabnet (neural networks). Further, we discuss our statistical and financial results to conclude the first hypothesis. Table A4.1 presents results for the three classifiers distributed across the key metrics described in section 5.1.1. Moreover, prevalence, sensitivity, specificity, and PPV are computed with regards to positive alpha. We want to notify the reader that there were no observable positive alphas in 2019 when FF6F was used as a risk-adjustment model. Consequently, the metrics subject to the True Positive rate would not be representative, and they will thus not be displayed in table A4.1, nor the figures presented in this chapter.

Figure 6.1 summarizes results from the evaluation metrics across the three different machine learning algorithms and the 9 hold-out sets:

Figure 6.1: Classifier metrics

The figure presents results for the three classifiers implemented in hypothesis 1. All measures are computed with regards to positive alphas, e.g. a PPV of 0.968 for XGBoost in 2013 show that 96.8% of all predicted positive alphas was correct for that model, in the year. In 2019, no observations of positive alpha exists, making certain measures non-representative, as a consequence the year is not presented in the figure. The red horizontal line in the AUROC figure represents the 50% threshold from Mandrekar (2010).



Focusing on the Prevalence in figure A4.1, we find that the distribution between positive and negative alphas fluctuates across the 9 hold-out-years. For the earlier hold-out-sets, ranging from 2013 until 2017, the distribution is tilted towards positive alphas. However, the distribution shifts in the later periods and becomes exceedingly tilted towards negative alphas, affecting the models' prerequisites for performance and their ability to classify positive alphas. However, the models maintain a high AUROC in 8 out of 10 years, signifying good out-of-sample performance. A good classifier will have an AUROC of above 0.7, which for all classifiers is achieved in all years except 2021. However, the AUROC in 2021 still outperforms a random guess with AUROC greater than 0.5. Further, following, Skalská and Freylich (2006), we compute bootstrapped AUROCs for the three classifiers. Results, presented in table A4.2 shows that we produce an AUROC statistically greater than 0.8 for both XGBoost and random forest. Tabnet produces bootstrapped AUROCs statistically greater than 0.7, hence underperforming the tree-based models, but still sufficient for a good classifier.

For the years 2018 through 2021 the models strive to classify positive alphas caused by the

class imbalance in the period.[32] The average PPV is 0.134 for XGBoost, 0.151 for random forest, and 0.083 for tabnet. This means that for every positive alpha prediction made in this period, less than 15% are correct. Further, the mean specificity score for the same years is 0.755 for XGBoost, 0.805 for random forest, and 0.620 for tabnet. This implies some success in filtering out true negatives, and we argue that the results makes sense due to the class imbalance in these years. Still, the sensitivity score in 2018 is relatively high for XGBoost and random forest due to the model successfully classifying all positive alphas in this period. In context with the low PPV, this signifies that the two models make several false positive predictions of alpha in these years.

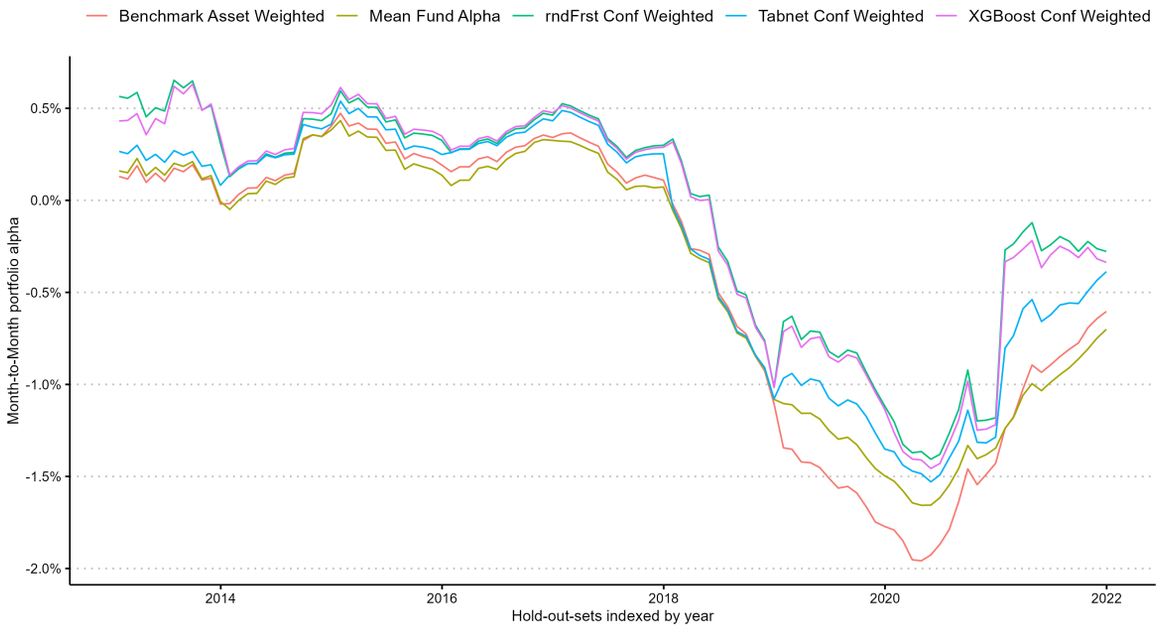
The overall classification performance is affirmative in the hold-out-sets of years 2013 to 2017. The mean sensitivity of all models is above 0.8, which signifies that the models are able to correctly predict most true positive alphas. A mean PPV of above 0.9 supports the model's classification performance for both random forest and XGBoost, demonstrating the classifiers' excellent ability to make correct positive alpha predictions. Tabnet performs slightly worse, with an average PPV of 0.81, which signify that out of all predicted positive alphas, 81% are truly positive. The specificity score is somewhat lower, which indicates that the models make some false positive predictions during this period. However, XGBoost makes the least false positive predictions with a mean specificity of 0.7 compared to random forest with 0.66 and tabnet with 0.51.

Besides the statistical measures, we also consider the classifiers' ability to identify alpha in a financial aspect, as the overall goal of our research question is to help investors select successful mutual funds. Figure 6.2 visualizes how the monthly net-alphas of the three machine learning portfolios develop over the hold-out-sample.

[32] Figure A4.1 visualizes the class imbalance.

Figure 6.2: Monthly realized portfolio alphas

The plot exhibits the monthly out-of-sample net-alphas of 5 different portfolios. The two benchmarks, equally and asset weighted comprises of all funds contained in the dataset. Similarly, the *mean fund alpha* is the average net-alpha of every fund in the dataset at month m. The portfolios, *rndFrst conf weighted, XGBoost conf weighted, and Tabnet conf weighted* are portfolios of all funds predicted to have a positive alpha in the year, weighted by the machine learners estimated probability of a fund being a positive alpha.



When investigating the alphas from the figure, we observe that nearly all of our specified portfolios achieved an abnormal net-alpha in the period 2013 to 2018. The subsequent period yields diminishing alphas due to the true alphas in this period being negative on average.[33] When comparing the confidence-weighted portfolio alphas in figure 6.2 to the equally weighted portfolio alphas in figure A4.2, there is evidence that the tree-based methods achieve the best alphas, consistent with the statistical measures.[34] Consistently, for both the confidence-weighted portfolios in figure 6.2, and the equally weighted portfolios in figure A4.2, the tree-based methods outperform the mean fund alpha portfolio and the asset weighted and equally weighted benchmark portfolios, which we find to be interesting. The results indicate that the machine learning portfolios perform in excess of their accompanied benchmarks. When interpreting the figure in a broader context, we realize that it is not only important to pick the winners, but also the best losers when the

[33] The alpha distribution is visualized in figure 4.1 in chapter 4.

[34] A confidence-weighted portfolio is a portfolio in which the machine learning algorithms estimates a probability for an observation turning positive or negative. We use this probability to make a weighted portfolio of all alphas in the respective algorithm which is predicted to be positive.

average fund achieves negative abnormal returns.

To conclude on *Hypothesis 1*, we contextualize both the statistical measures and the economic interpretations. The null hypothesis is:

It is not possible to classify positive alphas.

Given the premise that it is possible to successfully classify positive alphas as long as we exceed the threshold of 50% for AUROC (random guess), we can state that we have done this significantly and successfully. With a mean AUROC above 75% on all models, we show that we are able to classify alphas. The overall performance of the machine learning models proves that XGBoost has the best precision, followed by random forest and tabnet. The average PPV is above 0.5 for all three models, which supplements our decision regarding the hypothesis. However, we want to emphasize that our scores are not optimal in the last third of the hold-out-sample. We believe this to derive from relationships in the hold-out-set set that are not present in the training set. It might also be the case that our chosen predictors does not capture changes in market conditions in a optimal way in this period. The financial considerations further indicate that we are able to outperform the mean fund alpha and its equally weighted and asset weighted benchmarks. Finally, both XGBoost and random forest produce AUROCs statistically greater than 0.8 and tabnet greater than 0.7. The models greatly outperforms a random guess of 0.5, signifying that the models are good classifiers of alpha. Conclusively, we reject the null hypothesis and accept the fact that we manage to classify alphas in the Nordic mutual fund market.

6.2 Fund Categorization

This section presents results of the *Categorization Hypothesis*. This hypothesis aims to test whether we can utilize machine learning algorithms to successfully rank mutual funds on their alpha in the next 12 months. We start by analyzing the predictive algorithms on a series of statistical metrics to measure our ability to predict alpha and our success in the ranking of mutual funds based on predicted alpha. Subsequently, we construct 5 portfolios by descending ranking on predicted alpha and divide predictions into quintile portfolios.[35]

[35]For presentations in figures and tables we replicate Morningstar abbreviations, meaning the top quintile portfolio will be categorized as 5 Star, and bottom quintile as 1 Star.

Thereafter, we compute the portfolios actual obtained alpha in the hold-out-set, as well as other financial measures as explained in chapter 5.1.1.

Figure 6.3: Statistical evaluation of predictive models

The figure presents development in the RMSE and Spearman's rho of the three machine learning algorithms across the hold-out-sample of 2013 through 2021. The RMSE measures the predictive accuracy and Spearman's rho measures the models ability to rank observations.

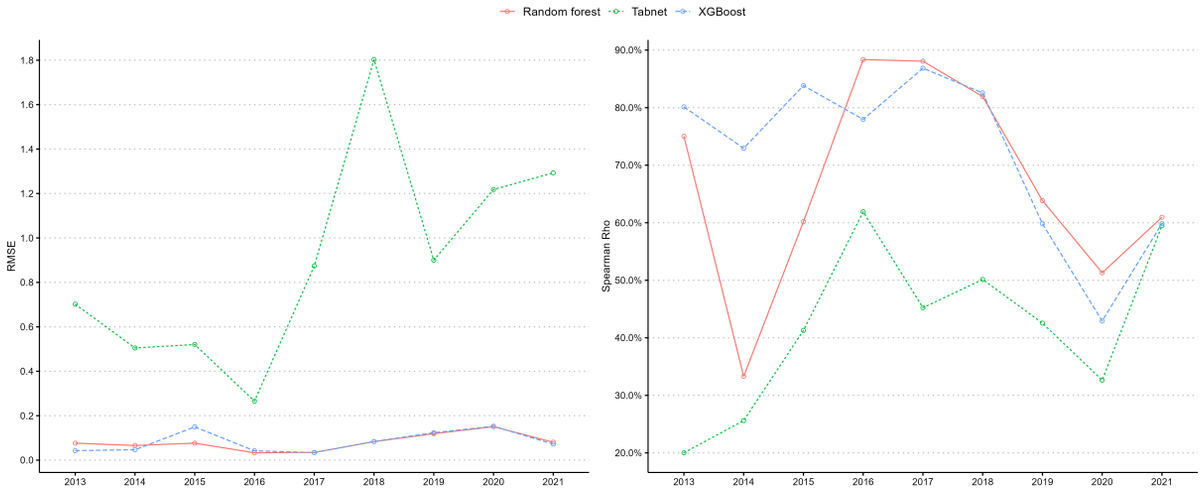


Figure 6.3 and table A5.1 report out-of-sample statistical metrics on the three machine learning algorithms; XGBoost, random forest, and tabnet. Our first finding is that the two tree-based machine learning algorithms, XGBoost, and random forest, outperform tabnet on both RMSE and Spearman's rho. This indicates that the tree-based methods do not only produce more accurate predictions of alpha, but are also better at ranking funds in terms of alpha. This is captured by a mean Spearman's rho of 24 percentage points higher than tabnet. These results are consistent with findings in *Hypothesis 1*, where the tree-based models illustrate superior classifier performance on both negative and positive alphas, when compared to tabnet. Interestingly, XGBoost underperforms on RMSE, but outperforms on Spearman's rho, when benchmarked against random forest. This implies that the overall predictive accuracy of random forest is better, but the XGBoost model is better at predicting a funds alpha relative to other funds. In simple terms, this translates to XGBoost being a better algorithm for ranking mutual funds on next year's alpha, which is reflected in the confusion matrices presented in the appendix part A5. This outperformance on ranking largely coincides with the hold-out-set of 2014, where XGBoost exceeds random forest with 39.6 percentage points on Spearman's rho. Therefore, it is reasonable to conclude that XGBoost is the best model due to more persistent results than

random forest, which yield poor ranking in the hold-out-set of 2014. Also, with an average Spearman's rho of 0.72, the XGBoost model shows good overall fund ranking abilities out-of-sample. Further, we investigate hold-out-sample net-alphas on the constructed quintile portfolios, in table 6.1:

Table 6.1: Cumulative portfolio alphas

The table reports hold-out-set annual cumulative alphas of the top quintile portfolios, across all algorithms. In addition, the table presents two benchmark portfolios comprising of all funds in the dataset, one equally weighted and one asset weighted. The stars signify results from a Welch t-test, testing whether difference in monthly net-alphas are statistically greater than zero. The average presents the mean annual cumulative alpha of the 9 hold-out-sets, and Cumulative exhibits the cumulative alpha of the entire hold-out-sample, ranging from 2013 through 2021. The asterisks denote statistical significance: $***$ p<0.01, $**$ p<0.05, $*$ p<0.1

	Benchmarks		**XGBoost**			**Random Forest**			**Neural Networks**		
	Equally	Asset	5 Star	Diff to Equally	Diff to Asset	5 Star	Diff to Equally	Diff to Asset	5 Star	Diff to Equally	Diff to Asset
2013	.0183	.0152	.0715	.0531***	.0562***	.0685	.0501***	.0532***	.0368	.0184***	.0215***
2014	.0192	.0216	.0745	.0554***	.0529***	.037	.0178***	.0154***	.0447	.0255***	.0230***
2015	.0332	.0394	.119	.0855***	.0793***	.0703	.0371***	.0309***	.0778	.0445***	.0383***
2016	.0257	.0312	.0882	.0625***	.0570***	.106	.0802***	.0746***	.0849	.0592***	.0537***
2017	.0207	.0251	.0981	.0774***	.0730***	.0973	.0767***	.0723***	.0648	.0442***	.0398***
2018	-.0694	-.0720	-.0045	.0649***	.0675***	-.0048	.0647***	.0672***	-.0486	.0208***	.0234***
2019	-.158	-.234	-.107	.0504***	.126***	-.107	.0506***	.126***	-.116	.0421***	.118***
2020	-.180	-.240	-.148	.0327***	.0926***	-.141	.0396***	.0995***	-.157	.0234***	.0834***
2021	-.111	-.0991	-.0760	.0353***	.0231***	-.077	.0345***	.0222***	-.0786	.0327***	.0205***
Average	-.0446	-.0569	.0128	.0575	.0698	.0055	.0501	.0624	-.0101	.0345	.0468
Cumulative	-.359	-.445	.0772	.4362	.5222	.0144	.3734	.4594	-.119	.240	.326

Our main finding is that the two tree-based methods, XGBoost, and random forest, select long-only portfolios (5-stars) which deliver statistically significant cumulative net alphas of 7.72% and 1.44% with respect to the FF6F. In contrast, the benchmarks portfolios, equally weighted and asset weighted, yield cumulative net alphas of -35.9% and -44.5%, respectively. Equivalently, all three models, XGBoost, random forest, and tabnet, yield statistically significant outperformance of month-to-month alphas across all hold-out-sets (2013 through 2021) when compared to the benchmarks.[36] The average monthly alpha of the XGBoost, random forest and tabnet portfolios are 7.2 bps (0.87% p.a.), 1.6 bps (0.19% p.a.) and -11.4 bps (-1.37% p.a), while the equally and asset weighted portfolios yields averages of -40.8 bps (-4.89% p.a) and -53.9 bps (-6.47% p.a). Interestingly the equally weighted portfolios outperform the asset weighted portfolios, which is consistent with the findings of DeMiguel et al. (2021).

Investigating further, we observe that XGBoost and random forest are the only top quintile portfolios able to return positive average and cumulative alphas across the 9

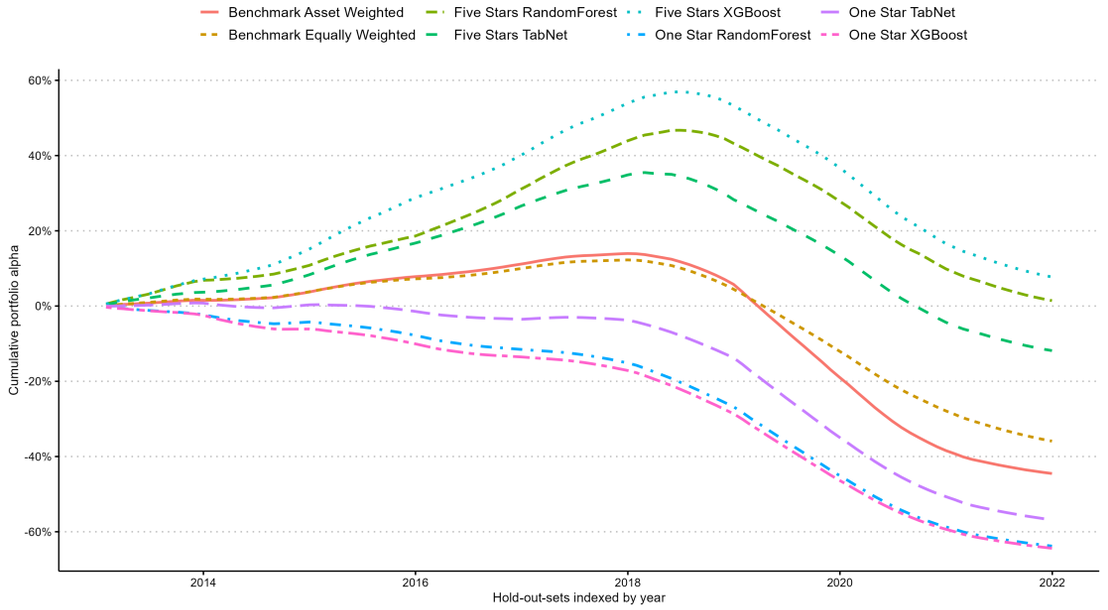
[36]For more information, see table A5.3.

hold-out-sets. In context with the fact that no other portfolio outperforms the top quintile portfolios in any of the hold-out-sets, this indicates some success in helping investors avoid underperforming funds. However, the tree-based methods XGBoost and random forest illustrate superior performance when compared to tabnet.

The hypothesis aims to measure our success in creating an effective ranking system for $alpha_{t+1}$. Accordingly, it is interesting to investigate how our top quintile portfolios perform relative to their lower quintiles and to assess the development in comparison to the two benchmarks. Following this, a cumulative representation of the portfolios is illustrated in figure 6.4:

Figure 6.4: Portfolio cumulative alphas

The figure illustrates the development in net-alphas across the hold-out-set, ranging from 2013 through 2021. The time series is the cumulative of monthly net-alphas for the top and bottom quintile portfolios of the machine learners. Included are also the cumulative of monthly net-alphas for two benchmark portfolios comprising of all funds in the dataset, one asset weighted and one equally weighted.



The illustration show that the outperformance of the top quintile are economically significant. Investing in our best performing portfolio (XGBoost 5-star), would earn an excess cumulative alpha of 43.6 percentage points compared to the equally weighted portfolio, and 52.2 percentage points compared to the asset weighted portfolio. Further, the top quintiles of XGBoost, random forest, and tabnet ultimately manage to differentiate from both the bottom quintile portfolios (one-star) and the asset weighted and equally weighted portfolios. We interpret this as a success in regards to the model's target, which

is to optimize alpha in the top-quintile portfolio. The counterargument to the performance of the top quintile portfolios would be to state that the rising tide lifts all the boats, due to better market conditions subsequent to the financial crisis of 2008. Our findings defy this, as results indicate that we successfully manage to separate funds producing higher alphas from those producing lesser or negative alphas. This suggests that our capability to filter out mutual funds producing lower alphas is materialized.

To validate the machine learners ability to rank observations, we evaluate whether the top quintile portfolios can outperform its lower quintile portfolios. Consequently, we perform a series of Welch t-tests (Welch, 1947), testing whether out-of-sample net-alphas of the top quintile are greater than the net-alpha of its lower quintile portfolios.[37] Table 6.2 reports the annualized mean net-alphas, and that the difference between the net-alphas of the top and bottom quintile portfolios is significant for all models. These findings are consistent with the Spearman's rho presented in table A5.1 and confirm our ability to rank mutual funds on net-alpha. Table A5.4 shows that the difference is still significant for the higher quintiles, substantiating the ability to rank funds based on predicted net-alpha. We emphasize that the difference in net-alpha is apparent and financially relevant to an investor.

Table 6.2: Mean portfolio alphas

The table reports mean hold-out-set alphas (annualized), of the top and bottom quintile portfolios across all algorithms. The asterisks signify a Welch t-test, testing whether the difference in monthly net-alphas of the top quintile portfolios and bottom quintile portfolios are statistically greater than zero. The t-test samples annualized monthly net-alpha in the hold-out-set. The final row exhibit annualized monthly alphas of the entire hold-out-sample, ranging from 2013 through 2021. The asterisks denote statistical significance: $***$ p<0.01, $**$ p<0.05, $*$ p<0.1

	XGBoost			**Random Forest**			**Neural Networks**		
	5 Star	1 Star	Difference	5 Star	1 Star	Difference	5 Star	1 Star	Difference
2013	.0692	-.0246	.0939***	.0664	-.0232	.0897***	.0362	.0078	.0283***
2014	.0721	-.0382	.110***	.0364	-.0201	.0565***	.0438	-.0046	.0484***
2015	.113	-.0428	.156***	.0682	-.0375	.106***	.0751	-.0175	.0926***
2016	.0848	-.0395	.124***	.101	-.0415	.143***	.0818	-.0214	.103***
2017	.0939	-.0424	.136***	.0933	-.0412	.134***	.0630	-.0022	.0652***
2018	-.0044	-.148	.144***	-.0047	-.147	.142***	-.0496	-.109	.0599***
2019	-.113	-.282	.169***	-.113	-.284	.171***	-.122	-.278	.156***
2020	-.159	-.274	.115***	-.151	-.282	.132***	-.169	-.274	.105***
2021	-.0788	-.132	.0532***	-.0797	-.131	.0511***	-.0816	-.133	.0510***
2013 - 2021	.0087	-.114	.1227	.0019	-.112	.1139	-.0137	-.0925	.0788

[37] Presented in table A5.4.

The main objective of this thesis is to rank funds in terms of alpha. However, it is also interesting to study how the constructed portfolios perform in terms of other performance metrics. Table 6.3 presents 6 different performance metrics relevant to an investor in our constructed portfolios:

Table 6.3: Monthly portfolio performance metrics

The table reports annualized return-based performance metrics of the top and bottom quintile portfolios, including the equally weighted and asset weighted benchmarks. Metrics are computed on monthly portfolio returns across the entire hold-out-period, ranging from 2013 through 2021.

	Benchmarks		**Top quintile / 5-star**			**Bottom quintile / 1-Star**		
	Equally	Asset	XGBoost	RandomForest	TabNet	XGBoost	RandomForest	TabNet
Mean return	.0856	.0886	.162	.142	.119	.0467	.0476	.0676
Std.Dev.	.0753	.0852	.110	.109	.0997	.0607	.0633	.0681
Sharpe ratio	1.01	0.930	1.32	1.17	1.07	.654	.641	.875
Sortino ratio	1.01	0.825	1.50	1.27	1.13	.689	.651	.810
Geometric return	.0825	.0846	.155	.135	.113	.0447	.0455	.0651
Cumulative return	1.04	1.08	2.67	2.13	1.63	.483	.492	.765

The ranking of arithmetic and geometric returns closely resembles the ranking in alphas. All machine learners outperform the benchmarks in their top quintile portfolio on both arithmetic and geometric means. This visualizes the dominant performance of the top quintile and adds to the evidence that our top quintile manages to select and differentiate funds performing better than others. The relative outperformance in out-of-sample return can partially be explained by higher volatility, consistent with the findings of Wu et al. (2021). Further, the risk-return relationship illustrated by the Sharpe and Sortino ratio is consistent with ranking in terms of alpha, as the machine learners top quintile portfolio outperforms the benchmarks, while the bottom quintile underperforms. Importantly, this also proves that we do not achieve return only at the expense of volatility. As a result, we can more confidently state that our best quintile portfolio performs better than the lower quintiles and the benchmarks. Across the models, we observe that XGBoost excels on both random forest and tabnet on all measures in the top quintile, especially for cumulative returns.

To conclude on *Hypothesis 2*, we contextualize both the statistical measures and the economic interpretations. The null hypothesis is:

It is not possible to create a successful ranking system.

First, we test the predictive ranking ability of our models using Spearman's rho.

XGBoost outperforms random forest when interpreting Spearman's rho, but interestingly underperforms on the accuracy measure RMSE. Second, we compute a ranking system that categorizes the predicted alpha coefficient of the respective mutual fund, diverting mutual funds with the highest predicted alpha in the top quintile and mutual funds with the lowest predicted alpha in the bottom quintile. This way, we systematize the selection problem that retail investors face when picking funds. To test whether we have made a successful ranking system, we prove that our top quintile portfolio is statistically significant from the lower quintiles. Third, we compute the cumulative alpha from XGBoost and random forest in the hold-out-sample, showing cumulative alphas after adjusting for FF6F on 7.72% and 1.44%. Tabnet deliver -11.9%, indicating the model's underachievement in ranking of $alpha_{t+1}$ and further confirms the results from *hypothesis 1*. We further test our fund picking abilities by testing whether monthly out-of-sample net-alphas of our top quintile portfolios are statistically greater than two benchmark portfolios. Results where significant for all models. Lastly, we supplemented our alpha measurements with the performance measurements described in table 5.4. We find that the ratios from the top quintile portfolios outperform the bottom quintile portfolios and the benchmarks, consistent with results on alphas. Accordingly, we are confident in rejecting the second null hypothesis and accept the alternative hypothesis stating that we manage to create a successful ranking system.

6.3 Endurance Test

This section presents results of the *Endurance Testing Hypothesis*. We use the Morningstar rating system to create five equally weighted portfolios. This enables us to track the portfolio's return characteristics and net alpha through our 9 hold-out sets, ranging from 2013 through 2021. Thereby we construct an appropriate time-series benchmark for our portfolios, where the machine learning portfolios and the Morningstar portfolios are compared on the same hold-out-set alphas and return metrics. Subsequently, we compare the performance characteristics of Morningstar top quintile portfolio to our top quintile portfolios.

Table 6.4 visualizes the annual out-of-sample net-alphas of our quintile portfolios, as presented in table 6.1, along with the portfolios constructed from the Morningstar star

ratings.[38]

Table 6.4: Cumulative portfolio alphas

The table reports annual cumulative hold-out-set net-alphas, of the top and bottom quintile portfolios. The cumulative alphas are benchmarked against the cumulative alpha of the Morningstar star-rating system. The asterisks signify a Welch t-test, testing whether difference in monthly net-alphas of the top quintile portfolios are statistically greater than zero. The last row presents cumulative alphas of the entire hold-out-sample, while the second to last present annualized mean net-alpha of the entire hold-out-period. The asterisks of the second to last row presents results from table A6.2. The asterisks denote: $***$ p<0.01, $**$ p<0.05, $*$ p<0.1

	Morningstar	**XGBoost**		**Random Forest**		**Neural Networks**	
	5 Star	5 Star	Difference	5 Star	Difference	5 Star	Difference
2013	.0259	.0715	.0455***	.0685	.0426***	.0368	.0108***
2014	.0274	.0745	.0471***	.0370	.0096***	.0447	.0172***
2015	.0384	.119	.0803***	.0703	.0319***	.0778	.0394***
2016	.0317	.0882	.0565***	.106	.0741***	.0849	.0532***
2017	.0323	.0981	.0658***	.0973	.0650***	.0648	.0325***
2018	-.0283	-.0045	.0239***	-.0048	.0236***	-.0486	-.0203
2019	-.0901	-.107	-.0173	-.107	-.0172	-.116	-.0256
2020	-.0924	-.148	-.0551	-.141	-.0482	-.157	-.0644
2021	-.0548	-.076	-.0212	-.0768	-.0220	-.0786	-.0238
Average	-.0135	.0087	.0222**	.0019	.0154*	-.0137	-.0002
2013 - 2021	-.116	.0772	.1932	.0144	.1304	-.119	-.003

Of the three top quintile portfolios, only the XGBoost and random forest portfolios are able to outperform Morningstar 5-star portfolios in terms of average out-of-sample net-alphas. The tabnet portfolio underperform the Morningstar rating system when considering average annualized alphas. Surprisingly, Morningstar's 5-star portfolio outperforms all our top-quintile portfolios on annual out-of-sample net-alphas from 2019 through 2021. In these years, very few and sometimes no positive alphas exist, indicating that Morningstar may be better at filtering out funds producing highly negative alphas from their 5-star portfolio. However, when examining monthly out-of-sample net-alphas of the top quintile portfolios, we observe that the lowermost monthly net-alpha of all mutual funds included in the Morningstar portfolio is -2.51% (pre-weighting) with a mean of -28.35 bps. In contrast, the lowermost included fund net-alpha of the XGBoost portfolio is -2.38% (pre-weighting) with a mean of 3.66 bps.[39] Accordingly, we find no evidence for Morningstar to be better at filtering out funds producing highly negative alphas. Hence, the outperformance in

[38]Table A6.1 reports results for all quintile portfolios.

[39]This refers to the monthly net-alpha of individual funds included in the top quintile portfolios. Where the mean is the average net-alpha of all funds included in the portfolio.

annual-out-of sample alphas in these years may be caused by our portfolios including more funds than that of Morningstar. We further investigate this later in the chapter, where we recompute our portfolios to hold a similar number of annual positions to that of Morningstar 5-star portfolios.

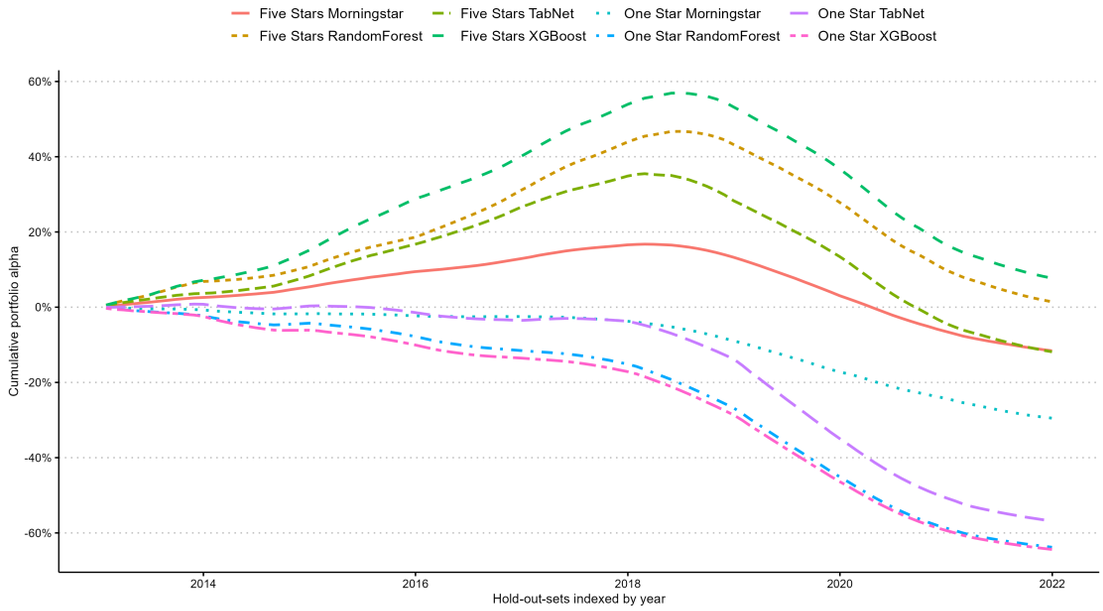
Our main finding is that we surpass Morningstar on out-of-sample cumulative net-alphas with our best performing portfolios, the XGBoost, and random forest top quintile portfolios. However, Morningstar outperform in the years 2018 through 2021, where the prevalence is tilted towards negative alphas. Still, the XGBoost and random forest top quintile portfolios exceed the mean annual alpha of Morningstar with 2.22 percentage points and 0.19 percentage points, respectively. However, only the XGBoost algorithm produces alphas significantly greater than that of Morningstar when we consider the hold-out-sample in its entirety. The difference in net-alpha for random forest is still apparent and financially relevant to an investor as he would earn an annual excess net-alpha of 1.54% when investing in the random forest portfolio as opposed to the Morningstar. Although the outperformance of random forest is not statistically significant at a 95% confidence-level, this illustrates that there is a possibility to make money from the five-star portfolio of random forest regardless of the statistical measure rejecting the significant difference from Morningstar. Contrarily, the tabnet portfolios slightly underperform Morningstar portfolios, with an average annual net-alpha of 2 bps lower than that of Morningstar. In summary, results show that we are able to select portfolios producing significantly higher out-of-sample monthly net-alpha when benchmarked against Morningstar portfolios. However, this is true only for our best performing model, XGBoost.

Further, figure 6.5 shows that the outperformance of both the XGBoost and random forest are economically significant, with a cumulative net alpha of 19.32% and 13.04%, in excess of Morningstar. Interestingly, Morningstar's bottom quintile portfolio (1-star) obtains higher cumulative net-alphas than that of our machine learners' bottom quintile portfolios. This may imply that Morningstar is not as successful at ranking alpha in such a way that the funds producing the lowermost alphas end up in the bottom quintile portfolio. Furthermore, the distance in obtained annual and cumulative net-alphas between the Morningstar top quintile and bottom quintile portfolios is less than that of the machine learner portfolios, substantiating that Morningstar star ratings are less successful than

the machine learners in the ranking of funds in terms of net-alpha.

Figure 6.5: Portfolio cumulative alphas

The figure illustrates the development in net-alphas across the hold-out-set, ranging from 2013 through 2021. The time series is the cumulative of monthly net-alphas for the top and bottom quintile portfolios of Morningstar and the machine learners.



The main goal of the thesis is to apply machine learning as a way of selecting mutual fund portfolios able to return the highest alpha net of costs. However, it is also of interest to evaluate how well we compare with Morningstar's well-known fund ranking system on return and risk based characteristics. Henceforth, we compute the same return-based measures as for *hypothesis 2* on both the Morningstar and the machine learning portfolios. Table 6.5 reports the results:

Table 6.5: Monthly portfolio performance metrics

The table reports annualized return based performance metrics of the top and bottom quintile portfolios of Morningstar and our machine learners. Metrics' are computed on monthly portfolio returns across the entire hold-out-sample, ranging from 2013 through 2021.

	Top quintile / 5-star				**Bottom quintile / 1-Star**			
	Morningstar	XGBoost	RandomForest	TabNet	Morningstar	XGBoost	RandomForest	TabNet
Mean return	.066	.162	.142	.119	.022	.047	.048	.068
Std.Dev.	.042	.110	.109	.100	.040	.061	.063	.068
Sharpe ratio	1.39	1.32	1.17	1.07	.410	.654	.641	.875
Sortino ratio	1.62	1.50	1.27	1.13	.403	.689	.651	.810
Geometric return	.065	.155	.135	.113	.022	.045	.046	.065
Cumulative return	.760	2.67	2.13	1.63	.212	.483	.492	.765

All three machine learners outperform Morningstar on arithmetic and geometric returns

in the top quintile, which is interesting given that the prediction target is fund alpha and not fund returns. As for *hypothesis 2* higher standard deviations partially explain the higher mean returns for the machine learning portfolios. This denotes that investors take on more risk to earn more returns. Conversely, when examining the Sharpe ratio, we observe that the Morningstar 5-star portfolio outperforms our best portfolio in terms of alpha (XGBoost 5-star) with 7 units. This is also the case when comparing the portfolios across the Sortino ratio, which only accounts for downside risk. The Morningstar top quintile portfolio exceeds the XGBoost top quintile portfolio by 12 units on the Sortino ratio. All top quintile machine learning portfolios surpass the Morningstar top quintile portfolio in terms of cumulative returns. The highest cumulative return is obtained when investing in the XGBoost top quintile portfolio, while the least is obtained from the Morningstar 5-star portfolio. Across the bottom quintile portfolios, the returns-based metrics reflect net-alpha observations, where the difference across the quintiles is the greatest for XGBoost, demonstrating the model's ability to rank funds.

The results stated above show that we outperform the Morningstar 5-star portfolio with regards to net-alpha on both XGBoost and random forest, though only significantly with XGBoost. However, in our base case, we compute the quintile portfolios based on equal distributions between all quintiles, resulting in a mean number of fund positions of 272 in our top quintile, 5-star portfolios. Contrarily, Morningstar has no such constraint, resulting in fewer funds included in the 5-star portfolios. Morningstar's top quintile portfolios hold between 103 and 200 funds p.a across the hold-out-sets, with a mean of 148 funds. Conversely, it is interesting to test whether our outperformance manifests due to our top quintile portfolios holding more positions than the Morningstar top quintile portfolio. As a result, we recompute the top quintile portfolios to hold positions in only the top 148 alpha predictions. Results, presented in table A6.3, show that we extend our outperformance on both cumulative alphas and cumulative returns. Interestingly our best portfolio, the XGBoost 5-star, converges from underperforming on Sharpe and Sortino ratio to outperforming when we reduce the number of holdings in the top quintile portfolio. Moreover, in addition to the increase in cumulative net-alphas being economically relevant, the increase in monthly net-alphas is also statistically significant.[40] As a result, we conclude that outperformance relative to Morningstar does not occur due to a higher

[40]See table A6.4 for t-test results.

number of positions, as all our performance metrics improve when reducing the number of positions p.a. to match Morningstar.

To conclude on *Hypothesis 3*, where we aim to check the endurance of our machine learning algorithms when compared to Morningstar, we refer to tables 6.4, 6.5, and A6.3. Our findings show that we deliver better cumulative alphas than Morningstar in the top quintile portfolio. However, this is not true when we consider the years 2019 through 2021 as standalone periods, where Morningstar exceeds all machine learning portfolios on annual cumulative net-alphas. This implies a weakness in our results, as our conclusion may be sensitive to an extended hold-out-sample. Nonetheless, table 6.5 shows that our top quintile portfolio surpasses Morningstar's top-quintile portfolio on all metrics besides Sharpe and Sortino ratio. However, when we reduce the number of annual positions in the top-quintile portfolio to match Morningstar, we further outperform on the Sharpe and Sortino ratio. Lastly, we supplemented our findings with a Welch t-test testing whether monthly out-of-samples net-alphas of our 5-star portfolios were significantly greater than that of the Morningstar 5-star portfolio. Results were significant only for XGBoost, although also economically relevant for random forest.

Based on the results presented above, we are confident both statistically and economically to reject the null hypothesis:

Our top quintile portfolio will not outperform Morningstar's five-star portfolio.

Henceforth, we claim that we are able to create a ranking system that surpasses the Morningstar star rating system on out-of-sample net-alphas.

7 Discussion and Robustness Checks

In this chapter, we first present robustness tests on our methodological choices of risk-adjustment model. Further, in section 7.2, we discuss the weaknesses of the thesis and how they may have affected our conclusion. Finally, in section 7.3, we elaborate on ideas for further research.

7.1 Robustness to Risk-Adjustment Model

Factor models have many potential difficulties, as they are estimated ex-post, meaning that they are based on historical data. Implementing a strategy to capture factor risk premiums *ex-ante* is thus considered challenging. Additionally, some factor analysis estimates are criticized for only estimating static exposures, assuming that betas are constant over time. We try to overcome this by running a rolling-window regression to better capture the dynamic changes in the factor exposures. For the net excess return of equity funds, we follow DeMiguel et al. (2021) and use the popular FF6F model, which is Fama and French (2015) 5-factor model augmented with the momentum factor from Carhart (1997). For the fixed income funds, we follow Bauer et al. (2022) and regress the net excess fund returns on the factors suggested by Fama and French (1993); term and default premium.

We posit that the theory of Fama (1991) on his joint hypothesis problem might be evidential: Measured abnormal returns can result from market inefficiency, a bad model of market equilibrium, or problems in the way the model is implemented. In simple terms, no one knows the true estimation parameters. For the fixed income funds, we emphasize the findings of Cremers et al. (2019):

> No model is generally accepted for controlling for bond portfolio risks. As a result, a wide variety of models have been used.

Moreover, Dahlquist, Polk, Priestley, and Ødegaard (2015) recommend the term and default factors and further elaborate that it is unnecessary to include a market factor in the isolated analysis of the fixed-income funds.[41] We believe that our model hence is

[41] We emphasize that regressions are net of the t-bill, while Bauer et al. (2022) are in net of their benchmark portfolio. This is due to the Norwegian Bank Investment Management having its own designated benchmark, while we do not.

estimated on the best practice.

However, knowing that our estimates are prone to measurement errors, we change the risk-adjustment model for the equity mutual funds to test if our results are robust to alternate factor models in the risk-adjustment process.[42] Conversely, we perform new rolling-window regressions with regards to FF3F on the second and third hypotheses (Fama & French, 1993). We omit to test the first hypothesis on FF3F due to our significant results and the time constraint. Furthermore, we do not control for a new risk-adjustment model regarding the fixed income funds, but inspect the potential impact of an inaccurate risk-adjustment model for fixed income funds on our conclusions.

7.1.1 Robustness Test of Fixed Income Funds

Robustness checking the risk-adjustment model for fixed income funds proves to be difficult, as there is no consensus on the correct model, and we believe that our default risk-adjustment is the best practice. However, to examine the impact of a potentially inappropriate risk-adjustment of fixed income funds, we inspect the population of the top quintiles, and redo our analysis with only the equity mutual funds. As a result, we consider if our conclusions would have been affected by an inappropriate risk-adjustment of fixed income funds.

We thus investigate the top quintile portfolio of XGBoost (our best performing model) and find that 28 fixed income and 798 equity funds are included across the hold-out-sample. Correspondingly, we find it interesting to study how the out-of-sample alphas materialize across the two asset classes. We filter our findings in the top quintile portfolio on equity and fixed income funds and construct two separate portfolios on the fund classes. Subsequently, we find that the mean monthly alpha for equity mutual funds and fixed-income funds is 3.33 bps (0.4% p.a.) and 22.2 bps (2.66% p.a.). We find it interesting that the fixed income funds outperform the equity funds on mean alpha, and believe this could be caused by our factor model not capturing the real systematic risk exposure of the fixed income funds, in line with Cremers et al. (2019). This could overestimate the alpha, thus making it larger than it actually is. We further investigate the results by controlling

[42] Akey, Robertson, and Simutin (2021) show that factor returns differ substantially depending on when the data were downloaded. They show that annual alphas of almost half of individual funds and even portfolios of funds change by more than 1% (in each direction).

for the fixed income funds. If the risk-adjustment model is truly correct, Cremers et al. (2019) statement that fixed income fund managers appear to make informed decisions on behalf of the investors is manifested. Table 7.1 presents our results when excluding fixed income from the portfolio:

Table 7.1: Metrics top quintile portfolios

The table presents the annualized performance metrics of the top quintile portfolios across all machine learning algorithms. The metrics are not adjusted for the number of positions, as we did in A6.3. All metrics are calculated for the hold-out period from 2013 through 2021, and visualize the top quintile portfolio of the machine learners, Morningstar, and the equally weighted and asset weighted benchmarks. The table includes results for the base case, and the robustness checks where we exclude all fixed income funds from the portfolios.

	Equity Funds						**Equity & Fixed Income Funds (Base case)**					
	Morningstar	Equally	Asset	XGBoost	RandomForest	Neural Network	Morningstar	Equally	Asset	XGBoost	RandomForest	Neural Network
Mean return	.083	.114	.124	.176	.151	.135	.066	.086	.089	.162	.142	.119
Std.Dev.	.052	.098	.104	.112	.113	.108	.042	.075	.085	.110	.109	.100
Sharpe ratio	1.41	1.04	1.07	1.41	1.20	1.12	1.39	1.01	.930	1.32	1.17	1.07
Sortino ratio	1.73	1.13	1.23	1.64	1.35	1.23	1.62	1.01	.825	1.50	1.27	1.13
Geometric return	.081	.109	.118	.169	.143	.129	.065	.083	.085	.155	.135	.113
Cumulative return	1.02	.025	.027	3.08	2.34	1.97	.760	1.04	1.08	2.67	2.13	1.63
Cumulative alpha	-.046	-.299	-.257	.159	.085	-.044	-.116	-.359	-.445	.077	.014	-.119

The table shows that our results are robust to a potentially inaccurate risk-adjustment model for the fixed income funds. The machine learning models outperform the benchmarks and Morningstar on all return and alpha metrics, although Morningstar outperforms on Sharpe and Sortino. Further, we find that monthly net-alphas of all our top quintile portfolios are significantly greater than the lower quintiles, consistent with our initial findings.[43] Additionally, table A7.1 proves that the net monthly alphas of our top quintile portfolios are significantly greater than the equally and asset weighted benchmarks and that we outperform the top quintile portfolio of Morningstar significantly with XGBoost. Overall, these results are in line with previous findings and illustrate that our conclusions are robust to a potential overstatement of alpha for fixed income funds.

7.1.2 Robustness Test of Equity Funds

As explained at the beginning of the chapter, we use the FF3F model to robustness check the risk-adjustment for equity funds, in accordance with Fama and French (1993). We thus perform new rolling-window regressions and retrain our models on the new dataset to predict the fund alpha coefficients. The results are presented in table 7.2:

[43] For results we refer to table A7.2.

Table 7.2: Metrics top quintile portfolios, FF3F

The table presents the annualized performance metrics of the top quintile portfolios across all machine learning algorithms. The metrics are not adjusted for the number of positions, as we did in A6.3. All metrics are calculated for the hold-out-sample from 2013 through 2021, and visualize the top quintile portfolio of the machine learners, Morningstar, and the equally weighted and asset weighted benchmarks.

	FF3F						FF6F					
	Morningstar	Equally	Asset	XGBoost	RandomForest	TabNet	Morningstar	Equally	Asset	XGBoost	RandomForest	TabNet
Mean return	.066	.085	.089	.156	.132	.107	.066	.086	.089	.162	.142	.119
Std.Dev.	.042	.075	.085	.109	.109	.093	.042	.075	.085	.110	.109	.100
Sharpe ratio	1.39	1.01	.930	1.28	1.09	1.04	1.39	1.01	.930	1.32	1.17	1.07
Sortino ratio	1.62	1.01	.825	1.46	1.18	1.04	1.62	1.01	.825	1.50	1.27	1.13
Geometric return	.065	.083	.085	.149	.125	.103	.065	.083	.085	.155	.135	.113
Cumulative return	.760	1.04	1.08	2.49	1.89	1.41	.760	1.04	1.08	2.67	2.13	1.63
Cumulative alpha	-.131	-.379	-.466	.032	-.019	-.203	-.116	-.359	-.445	.077	.014	-.119

The table visualizes that the results are robust to an alternative factor model to measure risk-adjusted performance. The top quintile portfolios outperform Morningstar and the equally and asset weighted benchmarks on all performance metrics except the Sharpe and Sortino ratios. We still argue that this is due to our portfolio taking more positions in the market, reducing the portfolio's expected return.

Further, we test whether the FF3F-models top quintile portfolios are statistically significant from its lower quintile portfolios as a comparison to *hypothesis 2*. The results presented in table A7.4 illustrate that both XGBoost and random forest successfully create a ranking system. This is consistent with the evidence from the FF6F default portfolio in table A5.4.

Finally, table A7.3 proves that the net monthly alphas of our top quintile portfolios are statistically greater than the equally and asset weighted benchmarks, and that we outperform the top quintile portfolio of Morningstar significantly with XGBoost. Overall, these results are consistent with previous findings and illustrate that our conclusions are robust to the FF3F.

7.2 Weaknesses

As in any other dissertation, our thesis is prone to weaknesses. First, note that we test the models over a relatively short period (2013 through 2021) and may risk not covering a long enough period to be persistent. Financial markets are dynamic, and models may have interchangeable periods of relevance, which is why testing the models over a substantive period is essential for robust results. As presented in tables 6.1 and 6.3, the performance

of the models fluctuates across the hold-out-sample. It especially varies in the last third of the hold-out-sample, where classifier performance and the significance in outperformance of Morningstar are depreciating. Consequently, results may have been affected if testing over a more extended period. However, we note that we observe solid results for all hypotheses for the majority of the data, which we believe substantiate our overall conclusions.

The aforementioned attributes of the financial markets underline another challenge when aiming to predict mutual fund alphas. Machine learning techniques try to predict the future by discovering and exploiting regularities in the training data. However, when the testing data contain regularities and relationships different from that of the training data, machine learning models may supply poor predictions. Although we significantly outperform Morningstar across the entire hold-out-sample, table 6.4 highlights that our ability to outperform Morningstar is insignificant in the last third of the hold-out-sample. The abrupt changes in market conditions in 2019, as illustrated by figures 4.1 and A7.1, combined with poor predictive performance, may indicate discrepancies in regularities between training and testing data for that year.[44] This is further substantiated by the model's improved performance in the subsequent years, as these new regularities enter the training data improving predictions.[45] Overall, this presents a general weakness in the implementation of machine learning in mutual fund selection.

Table 4.1 visualize that we use empirically backed and well-performing predictors in our thesis. However, basing our predictive machine learning algorithms on the work of others might lead to biases. We risk omitting other unknown relevant predictors, or including predictors which are no longer relevant. Market conditions may no longer be applicable, or fund managers might learn about the research on the predictors and apply it. The predictive ability can be reduced, and the predictive advantage could be reduced to an equilibrium position.

Finally, in the last hypothesis, we benchmark our quintile portfolios to Morningstars against the star rating system.[46] Ideally, we would have used Morningstar's Quantitative rating, which is created by statistical machine-learning models. Hence, it could be a more representative peer to the methods implemented in this thesis. Unfortunately, the data

[44] Appendix part A1 illustrate changes in the FF6F factors across the hold-out-sample.

[45] This is due to the rolling-window train-to-validation split as explained in figure 5.1.

[46] The star rating is a purely quantitative, backward-looking measure of our fund's past performance, measured from one to five stars.

coverage in the Nordic market on the quantitative rating is sparse.

7.3 Further Research

Earlier, we emphasized that the choice of factor models represents an uncertainty, which is why we have used the most recent factor model of Fama and French (2015) with momentum. However, another way of calculating the alpha would be to subtract a given benchmark return from the net fund return, which would yield the active return, often called raw alpha.[47] Thus, we do not risk-adjust more than the benchmark risk. The advantage of doing it this way is that we have an investable alternative, as factor models are often criticized as hard to interpret and require simultaneous long-short positions (Ang, 2014). Since our mutual funds can go long-only, we earn static equity and bond risk premiums by taking only long positions. Ang (2014) further describes that the dynamic factors require constant dynamic trading in the long-short positions, while our machine learning algorithms only rebalance once a year. The funds might also have restrictions such that they cannot replicate the factor portfolios due to risk-budget constraints, maximum holdings in individual companies, and liquidity (Bauer et al., 2022). We henceforth posit that it might be interesting to try the *raw alpha* due to its primitive calculation.

Further, it would be interesting to investigate the applicability of Berk and Van Binsbergen (2015) value added measure as a target variable for selecting funds with machine learning. The authors argue that this factor can be a better measure to consider how much the funds extract from the capital market. Unfortunately, due to time constraints, we were unable to supplement the proposed measures. Additionally, one could investigate different rebalancing periods, as Wu et al. (2021) proposed. Different frequencies for rebalancing could be a factor that either reduces or increases the net risk-adjusted alpha. However, we argue that a one-year rebalancing period is realistic for retail investors since they do not change mutual funds too often.

[47] Active return is denoted as R_A, calculated as $R_P - R_B$ and represent the non risk-adjusted excess return measure. The formula subtracts the return of a benchmark from the mutual fund return, meaning that it only considers benchmark risk.

8 Conclusion

Despite the growing popularity of passive investing, actively managed mutual funds still captures significant market shares. However, the evidence shows that these funds, on average, fail to produce significant fund alphas net of costs, raising the question of whether active investing is worth it when the available benchmark is cheaper, more diversified, and produces greater net excess returns on average. As an investor selecting mutual funds that show persistent performance is challenging. However, the recent evolution in data processing capabilities has boosted the use of machine learning in asset pricing, whereas some are able to generate positive abnormal returns. To test whether we can utilize machine learning, we pose the research question; How is the applicability of machine learning to pick successful mutual funds in the Nordic market?

To answer the research question, we have developed a performance-enhancing system to assist retail investors in selecting mutual funds. As the evidence shows that significant alphas are rare, we first create a classification system for separating funds based on their alpha. To systematize the testing, we predict the degree of alpha and create a ranking system that enables the investor to select a bundle of the best-performing mutual funds in the investment universe. Further, to legitimize our models, we benchmark our results against the Morningstar star rating system.

We underline three main findings. First, we manage to produce a classification system that separates negative from positive alphas. We produce an AUROC significantly above 80% for XGBoost and random forest, and significantly above 70% for tabnet, substantially outperforming a random walk. Secondly, the top quintile portfolios of XGBoost and random forest produce an average annual alpha of 1.28% and 0.6%, indicating that investors can gain statistically significant abnormal returns by investing in our top quintile portfolios. Contextualizing an investment from the beginning of our hold-out-sample, an investor would harvest a cumulative alpha from XGBoost and random forest of 7.72% and 1.44%. Our top quintile portfolios significantly outperforms the benchmarks on mean annual alphas, whereas our best portfolio exceeds the asset and equally weighted benchmarks by 7% and 6%, respectively. Third, we find evidence of significant outperformance on net alphas to Morningstar's top quintile portfolio, legitimizing our models. In addition, we show that our findings are robust to changes of risk-adjustment models. Based on the above, we argue that the applicability of machine learning to enhance fund selection and pick successful mutual funds is materialized. Our findings posit that investors can use machine learning to select mutual funds, generating a positive alpha net of all costs. In addition, we show that our findings are robust to changes of risk-adjustment models.

References

Akey, P., Robertson, A., & Simutin, M. (2021). Noisy factors. *Available at SSRN 3930228*. Retrieved from http://dx.doi.org/10.2139/ssrn.3930228

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552. Retrieved from https://doi.org/10.1111/j.1651-2227.2006.00180.x

Amihud, Y., & Goyenko, R. (2013). Mutual fund's r2 as predictor of performance. *The Review of Financial Studies*, *26*(3), 667–694. Retrieved from https://doi.org/10.1093/rfs/hhs182

Ang, A. (2014). *Asset management: A systematic approach to factor investing*. Oxford University Press.

Arık, S. O., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Aaai* (Vol. 35, pp. 6679–6687).

Barber, B. M., Huang, X., & Odean, T. (2016). Which factors matter to investors? evidence from mutual fund flows. *The Review of financial studies*, *29*(10), 2600–2642. Retrieved from https://doi.org/10.1093/rfs/hhw054

Barry, C. B., & Starks, L. T. (1984). Investment management and risk sharing with multiple managers. *The Journal of Finance*, *39*(2), 477–491. Retrieved from https://doi.org/10.1111/j.1540-6261.1984.tb02321.x

Bauer, R., Christiansen, C., & Døskeland, T. (2022). A review of the active management of norway's government pension fund global. *Available at SSRN 4003433*. Retrieved from http://dx.doi.org/10.2139/ssrn.4003433

Ben-David, I., Li, J., Rossi, A., & Song, Y. (2019). What do mutual fund investors really care about? *The Review of Financial Studies*, *35*(4), 1723–1774. Retrieved from https://doi.org/10.1093/rfs/hhab081

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213.

Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of political economy*, *112*(6), 1269–1295. Retrieved from https://doi.org/10.1086/424739

Berk, J. B., & Van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of financial economics*, *118*(1), 1–20. Retrieved from https://doi.org/10.1016/j.jfineco.2015.05.002

Berk, J. B., & Van Binsbergen, J. H. (2016). Assessing asset pricing models using revealed preference. *Journal of Financial Economics*, *119*(1), 1–23. Retrieved from https://doi.org/10.1016/j.jfineco.2015.08.010

Blake, C. R., & Morey, M. R. (2000). Morningstar ratings and mutual fund performance. *Journal of financial and Quantitative Analysis*, *35*(3), 451–483. Retrieved from https://doi.org/10.2307/2676213

Boney, V., Comer, G., & Kelly, L. (2009). Timing the investment grade securities market: Evidence from high quality bond funds. *Journal of Empirical Finance*, *16*(1), 55–69. Retrieved from https://doi.org/10.1016/j.jempfin.2008.06.005

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. Retrieved from https://doi.org/10.1023/A:1010933404324

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57–82. Retrieved from https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, *7*(3), 1247–1250. Retrieved from https://doi.org/10.1016/j.ins.2021.11.036

Chen, J., Wu, W., & Tindall, M. L. (2016). *Hedge fund return prediction and fund selection: A machine-learning approach* (Tech. Rep.). Federal Reserve Bank of Dallas. Retrieved from https://EconPapers.repec.org/RePEc:fip:feddop:2016_004

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... others (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.

Chevalier, J., & Ellison, G. (1999). Are some mutual fund managers better than others? cross-sectional patterns in behavior and performance. *The journal of finance*, *54*(3), 875–899. Retrieved from https://doi.org/10.1111/0022-1082.00130

Coval, J. D., & Moskowitz, T. J. (2001). The geography of investment: Informed trading and asset prices. *Journal of political Economy*, *109*(4), 811–841.

Cremers, K. M., Fulkerson, J. A., & Riley, T. B. (2019). Challenging the conventional wisdom on active management: A review of the past 20 years of academic literature on actively managed mutual funds. *Financial Analysts Journal*, *75*(4), 8–35. Retrieved from https://doi.org/10.1080/0015198X.2019.1628555

Dahlquist, M., Polk, C., Priestley, R., & Ødegaard, B. A. (2015). Norges bank's expert group on principles for risk adjustment of performance figures–final report. *Norges Bank report*. Retrieved from https://www.nbim.no/contentassets/f04b97db0e704572bfbda10525a6a3fc/expert_group_final_report_nov_2015.pdf

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & AP Santos, A. (2021). Machine learning and fund characteristics help to select portfolios of mutual funds. *Can Machine Learning Help to Select Portfolios of Mutual Funds*. Retrieved from https://dx.doi.org/10.2139/ssrn.3768753

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Evans, R. B., & Sun, Y. (2021). Models or stars: The role of asset pricing models and heuristics in investor risk adjustment. *The Review of Financial Studies*, *34*(1), 67–107. Retrieved from https://doi.org/10.1093/rfs/hhaa043

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383–417. Retrieved from https://doi.org/10.2307/2325486

Fama, E. F. (1991). Efficient capital markets: Ii. *The journal of finance*, *46*(5), 1575–1617. Retrieved from https://doi.org/10.1111/j.1540-6261.1991.tb04636.x

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, *47*(2), 427–465. Retrieved from https://doi.org/10.1111/j.1540-6261.1992.tb04398.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://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance*, *65*(5), 1915–1947. Retrieved from https://doi.org/10.1111/j.1540-6261.2010.01598.x

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22. Retrieved from https://doi.org/10.1016/j.jfineco.2014.10.010

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874. Retrieved from https://doi.org/10.1016/j.patrec.2005.10.010

Ferreira, M. A., Keswani, A., Miguel, A. F., & Ramos, S. B. (2013). The determinants of mutual fund performance: A cross-country study. *Review of Finance*, *17*(2), 483–525. Retrieved from https://doi.org/10.1093/rof/rfs013

Frazzini, A., & Lamont, O. A. (2008). Dumb money: Mutual fund flows and the cross-section of stock returns. *Journal of financial economics*, *88*(2), 299–322. Retrieved from https://doi.org/10.1016/j.jfineco.2007.07.001

French, K. R. (2008). Presidential address: The cost of active investing. *The Journal of Finance*, *63*(4), 1537–1573. Retrieved from https://doi.org/10.1111/j.1540-6261.2008.01368.x

Grinblatt, M., & Titman, S. (1992). The persistence of mutual fund performance. *The Journal of finance*, *47*(5), 1977–1984. Retrieved from https://doi.org/10.1111/j.1540-6261.1992.tb04692.x

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273. Retrieved from https://doi.org/10.1093/rfs/hhaa009

Gupta, F., Prajogi, R., & Stubbs, E. (1999). The information ratio and performance. *Journal of Portfolio Management*, *26*(1), 33. Retrieved from https://www.proquest.com/scholarly-journals/information-ratio-performance/docview/195576286/se-2?accountid=37265

Gutierrez, R. C., Maxwell, W. F., & Xu, D. (2009). On economies of scale and persistent performance in corporate-bond mutual funds. *Available at SSRN 1133959*. Retrieved from http://dx.doi.org/10.2139/ssrn.1133959

Hunter, D., Kandel, E., Kandel, S., & Wermers, R. (2014). Mutual fund performance evaluation with active peer benchmarks. *Journal of Financial economics*, *112*(1), 1–29. Retrieved from https://doi.org/10.1016/j.jfineco.2013.12.006

IBM. (2020, August). *What are neural networks?* Retrieved from https://www.ibm.com/cloud/learn/neural-networks

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429–449.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *The Journal of finance*, *23*(2), 389–416. Retrieved from https://doi.org/10.2307/2325404

Jones, C. S., & Mo, H. (2021). Out-of-sample performance of mutual fund predictors. *The Review of Financial Studies*, *34*(1), 149–193. Retrieved from https://doi.org/10.1093/rfs/hhaa026

Kamal, R., et al. (2013). Can morningstar analyst ratings predict fund performance? *Journal of Applied Business Research (JABR)*, *29*(6), 1665–1672. Retrieved from https://doi.org/10.19030/jabr.v29i6.8205

Kosowski, R., Timmermann, A., Wermers, R., & White, H. (2006). Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of finance*, *61*(6), 2551–2595. Retrieved from https://doi.org/10.1111/j.1540-6261.2006.01015.x

Kräussl, R., & Sandelowsky, R. M. (2007). The predictive performance of morningstar's mutual fund ratings. In *Se puede descargar en: http://ssrn. com/abstract* (Vol. 963489). Retrieved from https://icmaif.soc.uoc.gr/~icmaif/Year/11conf/docs/Crete2007.pdf

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Leippold, M., & Rueegg, R. (2020). How rational and competitive is the market for mutual funds? *Review of Finance*, *24*(3), 579–613. Retrieved from https://doi.org/10.1093/rof/rfz011

Li, B., & Rossi, A. G. (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*. Retrieved from http://dx.doi.org/10.2139/ssrn.3737667

Lingo, M., & Winkler, G. (2008). Discriminatory power - an obsolete validation criterion? *Banking & Financial Institutions eJournal*.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. Retrieved from https://doi.org/10.1097/JTO.0b013e3181ec173d

Martens, D., & Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1934670

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms. *Methods of information in medicine*, *53*(06), 419–427. Retrieved from https://doi.org/10.3414/ME13-01-0122

Modigliani, F., & Modigliani, L. (1997). Risk-adjusted performance. *Journal of portfolio management*, *23*(2), 45–54. Retrieved from https://doi.org/10.3905/jpm.23.2.45

Moneta, F. (2015). Measuring bond mutual fund performance with portfolio characteristics. *Journal of Empirical Finance*, *33*, 223–242. Retrieved from https://doi.org/10.1016/j.jempfin.2015.03.012

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. Retrieved from https://doi.org/10.1016/j.dss.2014.03.001

Nanigian, D. (2012). Why do mutual fund expenses matter? *Financial Services Review*, *21*(3). Retrieved from http://dx.doi.org/10.2139/ssrn.1977655

Patel, S., & Sarkissian, S. (2017). To group or not to group? evidence from mutual fund databases. *Journal of Financial and Quantitative Analysis*, *52*(5), 1989–2021. Retrieved from https://doi.org/10.1017/S0022109017000655

Pedersen, L. H. (2018). Sharpening the arithmetic of active management. *Financial Analysts Journal*, *74*(1), 21–36. Retrieved from https://doi.org/10.2469/faj.v74.n1.4

Rollinger, T. N., & Hoffman, S. T. (2013). Sortino: a 'sharper'ratio. *Chicago, Illinois: Red Rock Capital*. Retrieved from http://ea.kitgain.com/content/uploadfile/202102/9e631613532179.pdf

Ryll, L., & Seidens, S. (2019). Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey. *arXiv preprint arXiv:1906.07786*. Retrieved from https://doi.org/10.48550/arXiv.1906.07786

Schapire, R. E., & Freund, Y. (2012). Foundations of machine learning. *MIT Press*, 23–52. Retrieved from https://ieeexplore.ieee.org/document/6282245?reload=true&t=

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, *39*(1), 119–138. Retrieved from http://www.jstor.org/stable/2351741

Sharpe, W. F. (1981). Decentralized investment management. *The Journal of Finance*, *36*(2), 217–234. Retrieved from https://doi.org/10.1111/jofi.12024

Sharpe, W. F. (1991). The arithmetic of active management. *Financial Analysts Journal*, *47*(1), 7–9. Retrieved from https://doi.org/10.2469/faj.v47.n1.7

Skalská, H., & Freylich, V. (2006). Web-bootstrap estimate of area under roc curve. *Austrian journal of statistics*, *35*(2&3), 325–330.

Spearman, C. (1961). The proof and measurement of association between two things. *American Journal of Psychology, 1904*, *15*, 45–58. Retrieved from https://doi.org/10.1037/11491-005

Weigert, F. (2021). Which variables predict future active mutual fund performance? new insights from academic research.

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, *34*(1-2), 28–35. Retrieved from https://doi.org/10.1093/biomet/34.1-2.28

Wu, W., Chen, J., Yang, Z., & Tindall, M. L. (2021). A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science*, *67*(7), 4577–4601. Retrieved from https://doi.org/10.1287/mnsc.2020.3696

Appendix

A1 Abbreviations

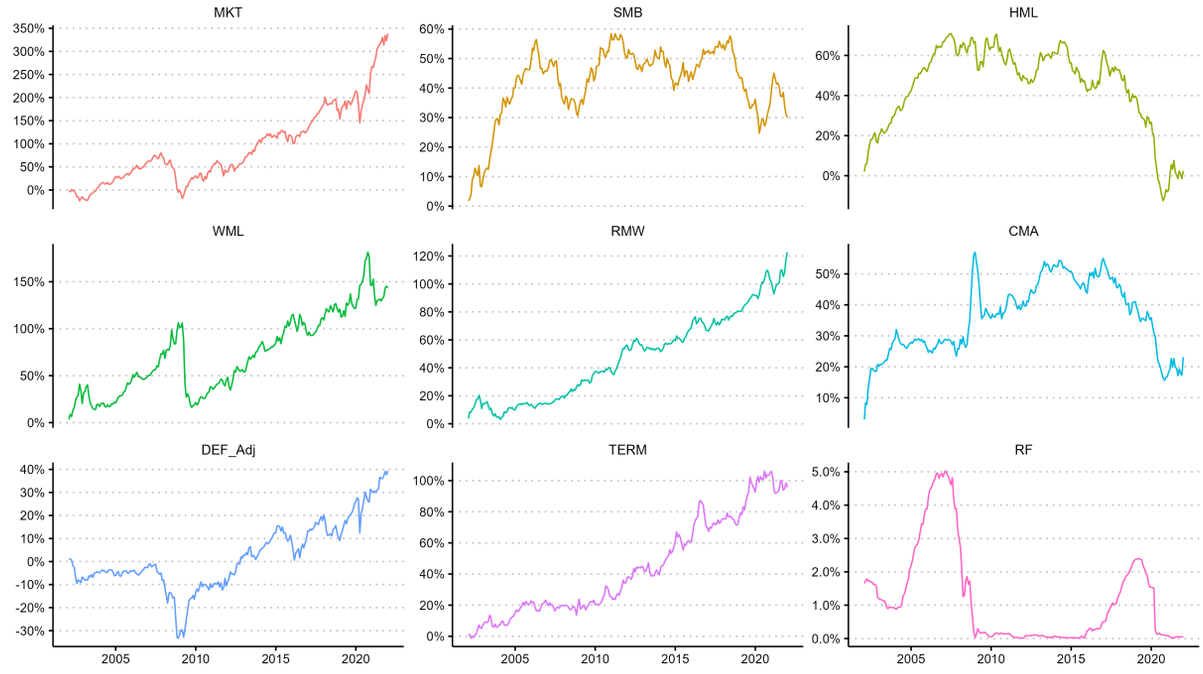
Table A1.1: Factor descriptions

This table introduces the risk-adjustment factors implemented for deriving alphas in this thesis.

Factor	Premium	Factor description
Equity factors		
MKT	Market	Equity market return in excess of the risk free rate
SMB	Size	Return spread between small cap and large cap stocks
HML	Value	Return spread between high book-to-market and low book-to-market stocks
WML	Momentum	Return spread between winner stocks and loser stocks
RMW	Profitability	Return spread between high and low profitability stocks
CMA	Investment	Return spread between stocks low and high investment ratios
Fixed Income factors		
DEF Adj	Default	Excess returns from long-term corporate bonds to long-term government bonds, adjusted for differences in duration between corporates and treasuries
TERM	Term	Return spread between long and short term government bonds
Risk-free rate		
RF	Risk-free rate	Three month US Treasury bill

Figure A1.1: Factor descriptions

The figure presents development in monthly return on the factors across our data set, ranging from 2002 through 2021. The time series is the cumulative of monthly factor premiums, except for the RF (risk-free rate). The RF presents the month-to-month rate of the 3-month US Treasury bill.



A2 Predictors

In this subsection we briefly explain the rationale of our predictors. We again emphasize that selecting the right predictors comes in form of a large risk of omitting other relevant characteristics.

A2.1 Recent Returns

We have included return based characteristics on the argument of past performance predicting future performance. Grinblatt and Titman (1992) found evidence that differences in performance between funds persist over time and that this persistent is consistent with the ability of fund managers to earn abnormal returns. On the other side, Carhart (1997) found that the evidence did not support the existence of skilled or informed mutual fund portfolio managers. We also follow the famous Fama and French (2015) factor model, plus the momentum factor from Carhart (1997). Additionally, we follow Hunter et al. (2014) to retrieve the t-stats instead of the raw alphas and betas to account for the estimation error. At last we include the R^2 as proposed by Amihud and Goyenko (2013), as it materializes as a good predictor of future performance.

A2.2 Risks

The goal with our predictors is to help to differentiate bad from well-performing funds. This is why we utilize Sharpe ratio (Sharpe, 1966), which has become a common risk-adjused measure. Furthermore, we use tracking-error and the information ratio, as proposed by Gupta et al. (1999). They argue that it allows investors to form reasonable expectations of the performance of their money managers and evaluate them appropriately. We also use the M2 measure of risk adjusted performance to supplement the Sharpe Ratio (Modigliani & Modigliani, 1997). In terms of idiosyncratic risk, we follow Gu et al. (2020), as the rationale is that the fund needs to expose themselves for idiosyncratic risk in order to produce alphas. Lastly, we follow Wu et al. (2021) when we account for Skewness, Kurtosis and the VIX.

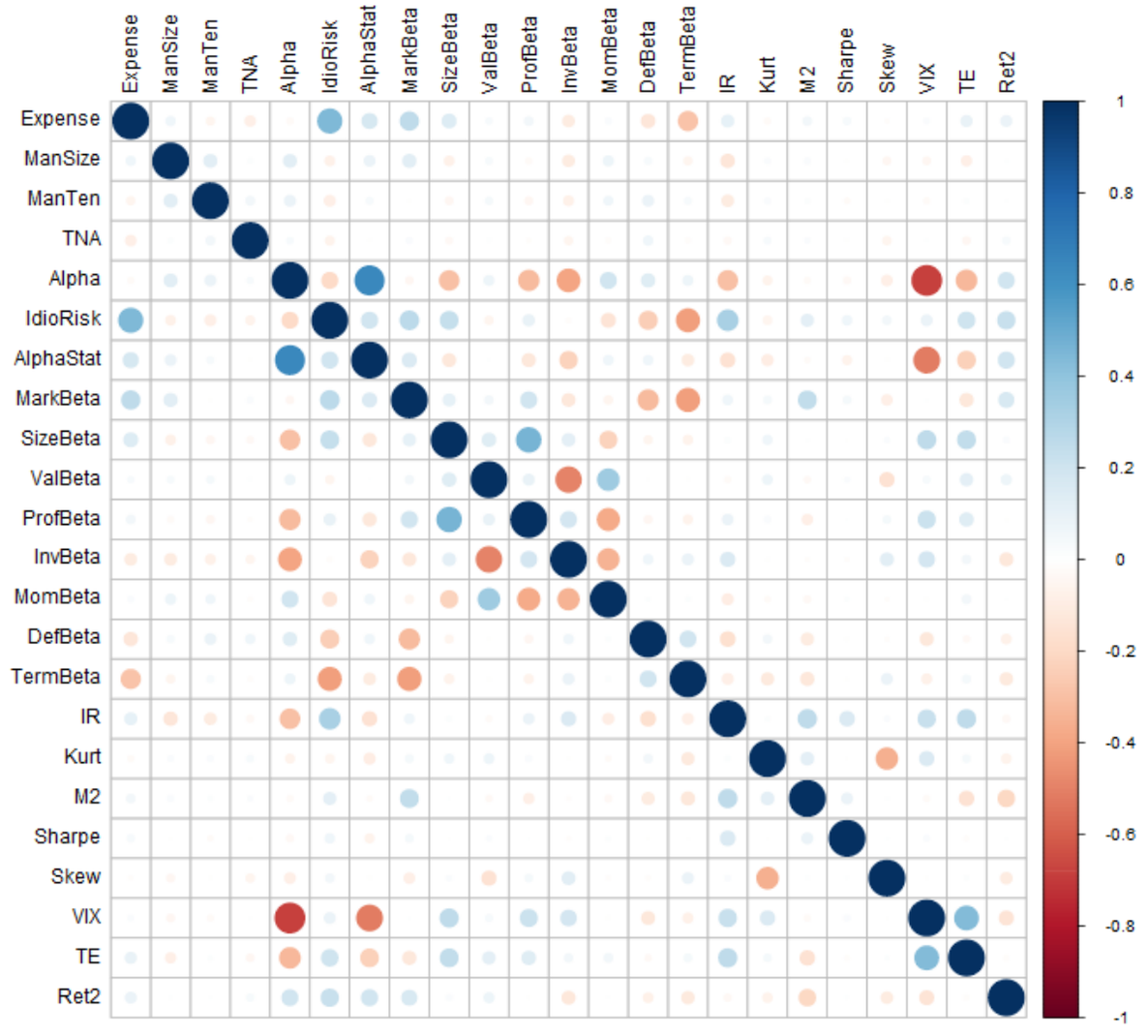
A2.3 Fund Management

To supplement the predictors based off of recent returns and risk factors we include a series of managerial characteristics. The research of Chevalier and Ellison (1999) show that managerial characteristics are important when explaining fund performance, hence a study examining the possibility of selecting high performing funds should account for these factors. Over the recent years, team-based mutual fund management has become the norm, explained by industry professionals as a trend occurring from a performance viewpoint. Sharpe (1981) and Barry and Starks (1984) argue that funds with team-management achieve a diversification of investment style and decision making that reduces portfolio risk, hence resulting in better performance. In a quantitative study, Patel and Sarkissian (2017) finds that on average, team-based funds have higher risk adjusted returns than single managed peers, hence considering managerial fund properties could be important in a study examining the possibility of selecting high performing funds. We also include certain fund characteristics, e.g. the expense ratio. Nanigian (2012) finds that fund expenses have a statistically significant negative impact on the performance of American funds.

A3 Correlations

Figure A3.1: Correlations plot

The figure illustrates a correlations plot of all numeric predictors implemented in the machine learning models. The figure show that no predictor set has a correlation greater than the threshold of 75% resulting in exclusion.



A4 Alpha Classification

Table A4.1: Classifier metrics

This table presents yearly classifier metrics for the three classifiers implemented in hypothesis 1. All measures are computed with regards to positive alphas, e.g. a prevalence of 0.746 in 2013 show that 74.6% of the mutual funds in the data set had a positive alpha in the year. In 2019, no observations of positive alpha exists, making certain measures nonrepresentative, and will not be presented in the table.

		XGBoost				**Random Forest**				**Neural Networks**			
	Prevalence	AUROC	Sensitivity	Specificity	PPV	AUROC	Sensitivity	Specificity	PPV	AUROC	Sensitivity	Specificity	PPV
2013	0.746	0.874	0.286	0.973	0.968	0.847	0.263	0.988	0.984	0.701	0.687	0.616	0.840
2014	0.783	0.876	0.897	0.698	0.915	0.854	0.904	0.655	0.904	0.709	0.680	0.623	0.867
2015	0.801	0.938	0.926	0.769	0.942	0.919	0.955	0.655	0.918	0.820	0.848	0.494	0.871
2016	0.710	0.929	0.956	0.601	0.854	0.930	0.977	0.534	0.837	0.848	0.958	0.438	0.806
2017	0.561	0.932	0.985	0.472	0.705	0.934	0.988	0.489	0.712	0.846	0.980	0.370	0.666
2018	0.099	0.903	1	0.470	0.171	0.886	1	0.487	0.176	0.726	1	0.0196	0.101
2019	0	-	-	0.844	-	-	-	0.867	-	-	-	0.565	-
2020	0.003	0.909	1	0.735	0.013	0.922	1	0.838	0.021	0.792	0	0.931	0
2021	0.0154	0.639	0.522	0.971	0.218	0.684	0.478	0.978	0.256	0.568	0.391	0.965	0.148
Average	0.413	0.875	0.822	0.726	0.598	0.872	0.821	0.721	0.601	0.751	0.693	0.558	0.537

Figure A4.1: Classifier metrics, prevalence

The figure presents yearly prevalence of the hold-out-sample. The measure is computed with regards to positive alphas, e.g. a prevalence of 0.746 for 2013 show that 74.6% of the mutual funds in the dataset had a positive alpha in the year.

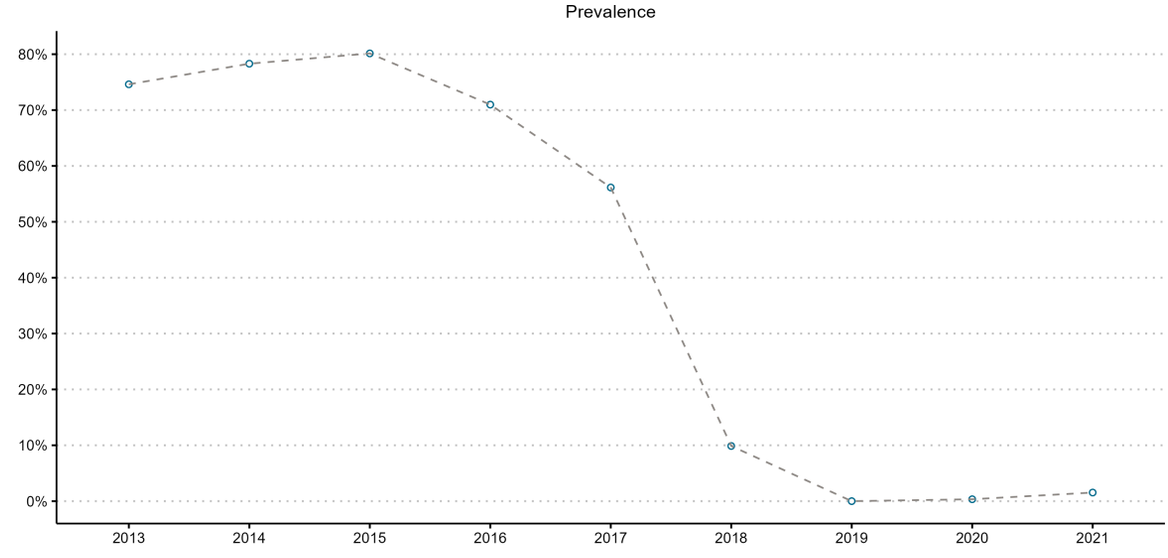


Table A4.2: Significance test of AUROC

The table presents results from bootstrapped t-tests, testing whether the hold-out-sample of AUROC are statistically greater than a given threshold. The t-tests are performed on 5 different AUROC thresholds, where 0.5 is considered to be a random guess and 1 a perfect model (Mandrekar, 2010). A classifier able to surpass a random guess is considered informative (Lingo & Winkler, 2008). The t-test samples a population of 100 AUROCs created from a bootstrapped (Efron & Tibshirani, 1994) sample off all predictions made in the hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
0.5	1206.40	$p < .001$	1232.70	$p < .001$	540.39	$p < .001$
0.6	882.92	$p < .001$	906.35	$p < .001$	317.38	$p < .001$
0.7	559.41	$p < .001$	579.96	$p < .001$	94.36	$p < .001$
0.8	235.90	$p < .001$	253.57	$p < .001$	-128.65	$p = 1$
0.9	-87.60	$p = 1$	-72.82	$p = 1$	-351.66	$p = 1$

Figure A4.2: Monthly realized portfolio alphas

The plot exhibits the monthly out-of-sample net-alphas of 6 different portfolios. The two benchmarks, equally and asset weighted comprises of all funds contained in the dataset. Similarly, the *mean fund alpha* is the average net-alpha of every fund in the dataset at month m. The portfolios, *rndFrst equally weighted, XGBoost equally Weighted, and Tabnet equally Weighted* are portfolios of all funds predicted to have a positive alpha in the year, equally weighted.

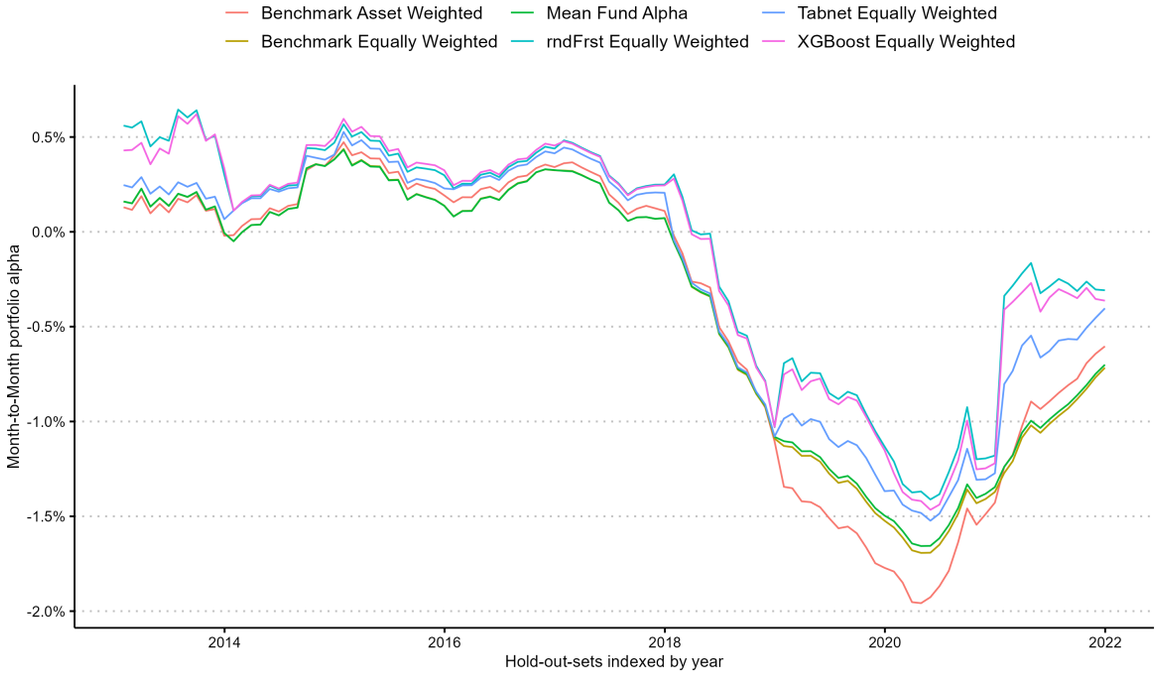


Figure A4.3: XGBoost confusion matrix

The figure illustrates the confusion matrix of the XGBoost classifier. It visualizes all classifications made in the entire hold-out-sample, ranging from 2013 through 2021.

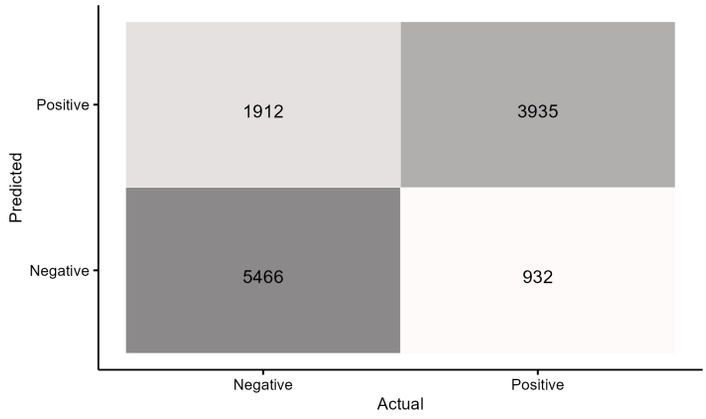


Figure A4.4: Random forest confusion matrix

The figure illustrates the confusion matrix of the random forest classifier. It visualizes all classifications made in the entire hold-out-sample, ranging from 2013 through 2021.

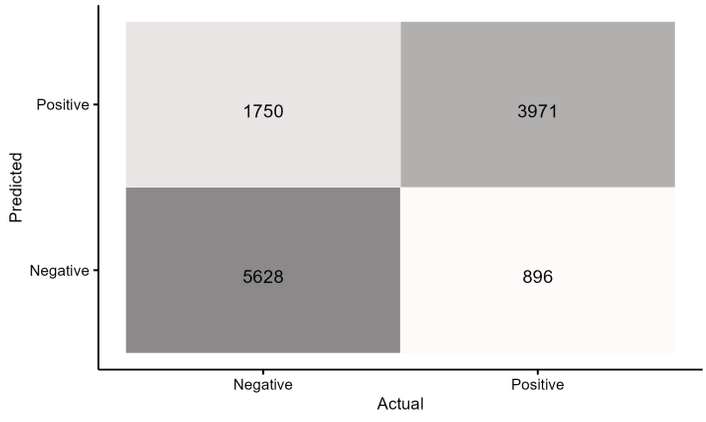
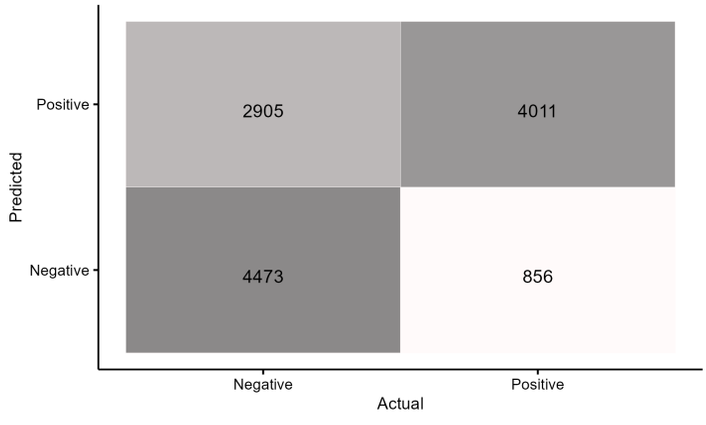


Figure A4.5: Tabnet confusion matrix

The figure illustrates the confusion matrix of the tabnet (Neural Network) classifier. It visualizes all classifications made in the entire hold-out-sample, ranging from 2013 through 2021.



A5 Fund Categorization

Table A5.1: Statistical evaluation of the predictive models

The table presents the RMSE and Spearmans Rho of the three different machine learners across the hold-out-sample of 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	RMSE	Spearmans Rho	RMSE	Spearmans Rho	RMSE	Spearmans Rho
2013	0.0428	0.801	0.0765	0.750	0.702	0.200
2014	0.0470	0.729	0.0659	0.333	0.505	0.256
2015	0.150	0.838	0.0765	0.602	0.520	0.413
2016	0.0421	0.780	0.0331	0.884	0.266	0.619
2017	0.0338	0.869	0.0349	0.881	0.874	0.452
2018	0.0838	0.826	0.0833	0.819	1.80	0.501
2019	0.123	0.598	0.119	0.638	0.900	0.426
2020	0.153	0.429	0.151	0.513	1.22	0.326
2021	0.0722	0.599	0.0806	0.609	1.29	0.595
Average	0.0831	0.7188	0.0801	0.6699	0.898	0.421

Table A5.2: Cumulative alpha

The table reports the hold-out-set annual cumulative net-alphas, of all quintile portfolios, across all algorithms. Included are also the equally weighted and asset weighted benchmark portfolios. The second to last row exhibits the annual cumulative mean alpha, and the last row presents the cumulative alpha of the entire hold-out-set, ranging from 2013 through 2021.

	Benchmarks		**XGBoost**					**Random Forest**					**Neural Networks**				
	Equally	Asset	5 Star	4 Star	3 Star	2 Star	1 Star	5 Star	4 Star	3 Star	2 Star	1 Star	5 Star	4 Star	3 Star	2 Star	1 Star
2013	.0183	.0152	.0715	.0275	.0152	.0045	-.0244	.0685	.0261	.0167	.0057	-.0230	.0368	.0186	.0129	.0159	.0079
2014	.0192	.0216	.0745	.0379	.0178	.0062	-.0376	.0370	.0330	.0280	.0189	-.0200	.0447	.0235	.0205	.0124	-.0046
2015	.0332	.0394	.1190	.0555	.0294	.0103	-.0419	.0703	.0641	.0475	.0242	-.0369	.0778	.0463	.0325	.0289	-.0174
2016	.0257	.0312	.0882	.0514	.0225	.0092	-.0388	.1060	.0467	.0189	.0029	-.0407	.0849	.0390	.0221	.0063	-.0212
2017	.0207	.0251	.0981	.0558	.0118	-.0156	-.0416	.0973	.0578	.0115	-.0176	-.0405	.0648	.0363	.0160	-.0101	-.0022
2018	-.0694	-.0720	-.0045	-.0434	-.0724	-.0827	-.139	-.0048	-.0423	-.0739	-.0836	-.1370	-.0486	-.0493	-.0708	-.0730	-.1040
2019	-.1580	-.2340	-.107	-.136	-.143	-.148	-.248	-.1070	-.1330	-.1460	-.1480	-.2500	-.1160	-.1380	-.1390	-.1450	-.2450
2020	-.1800	-.2400	-.148	-.160	-.173	-.177	-.242	-.1410	-.1630	-.1720	-.1740	-.2480	-.1570	-.1590	-.1690	-.1720	-.2420
2021	-.1110	-.0991	-.076	-.106	-.121	-.127	-.124	-.0768	-.1040	-.1210	-.1300	-.1230	-.0786	-.1050	-.1180	-.1290	-.1250
Average	-.0446	-.0569	.0128	-.0241	-.0459	-.0578	-.1042	.0055	-.0238	-.0433	-.0557	-.1022	-.0101	-.0320	-.0436	-.0518	-.0838
Cumulative	-.3590	-.4450	.0772	-.2240	-.3640	-.4300	-.6440	.0144	-.2220	-.3500	-.4200	-.6380	-.1190	-.2760	-.3500	-.3980	-.5680

Table A5.3: Significance test of 5-Star portfolios against benchmarks

The table presents results from Welch t-tests, testing whether monthly net-alphas of the top quintile machine learning portfolios are statistically greater than the asset and equally weighted portfolios. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
Equally Weighted	4.4978	$p < .001$	4.1239	$p < .001$	2.8645	$p < .001$
Asset Weighted	4.8758	$p < .001$	4.5489	$p < .001$	3.487	$p < .001$

Table A5.4: Significance test of ranking system, FF6F

The table presents results from a series of Welch t-test, testing whether monthly net-alphas of the top-quintile portfolio are statistically greater than that of lower quintiles. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
4 Star	2.836	$p = .0025$	2.3718	$p = .0093$	1.8166	$p = .0353$
3 Star	4.692	$p < .001$	4.0482	$p < .001$	2.8396	$p = .0025$
2 Star	5.8122	$p < .001$	5.2453	$p < .001$	3.5733	$p < .001$
1 Star	9.022	$p < .001$	8.5036	$p < .001$	5.6727	$p < .001$

Figure A5.1: XGBoost confusion matrix quintile ranking

The figure illustrate the ranking accuracy of the XGBoost machine learning model. The y-axis show the predicted quintile rankings, and the x-axis show the actual quintile ranking, both based on net-alphas.

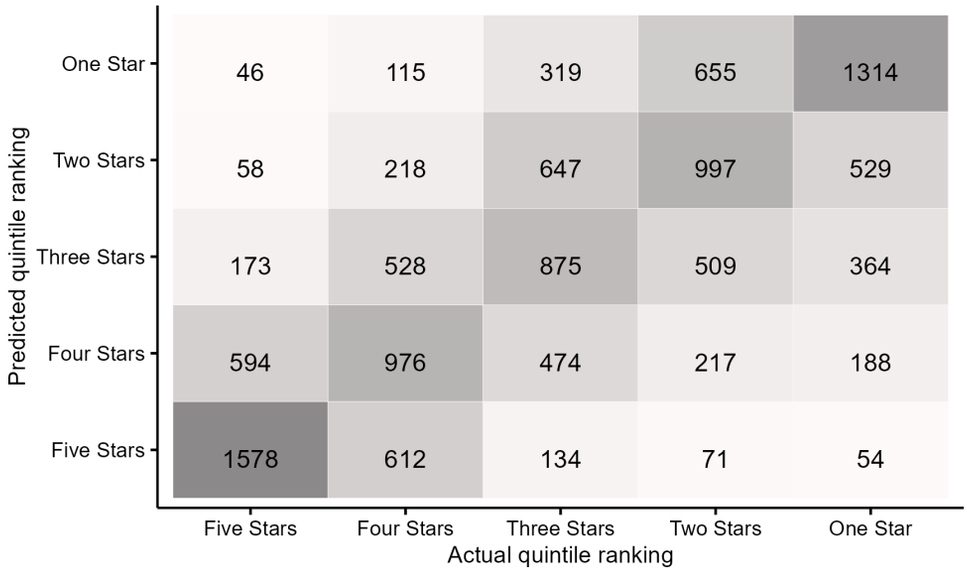


Figure A5.2: Random forest confusion matrix quintile ranking

The figure illustrate the ranking accuracy of the random forest machine learning model. The y-axis show the predicted quintile rankings, and the x-axis show the actual quintile ranking, both based on net-alphas.

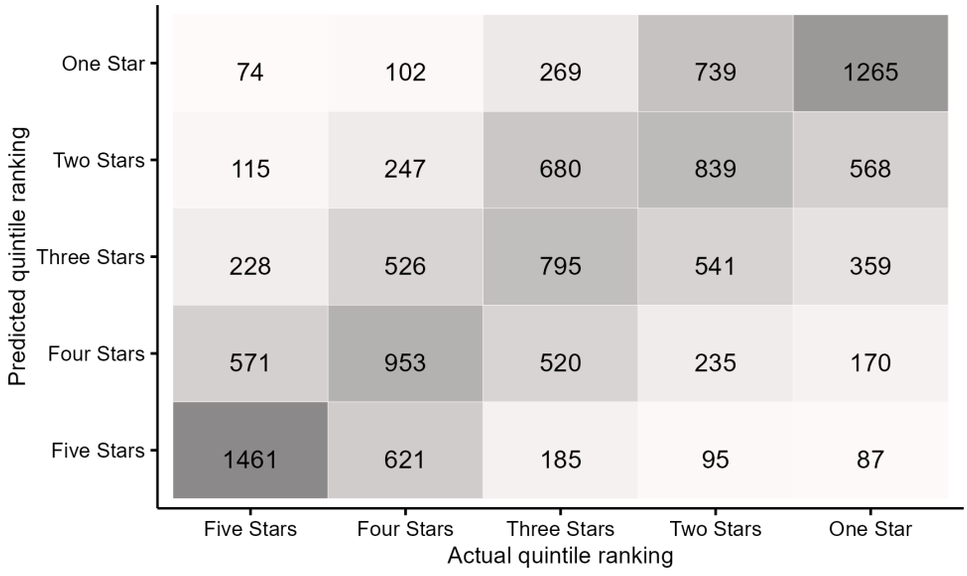
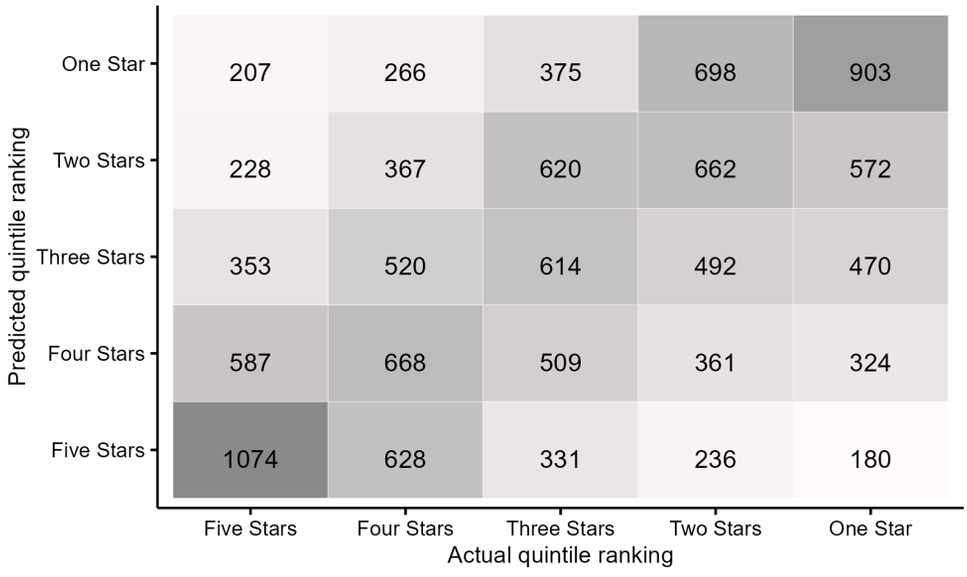


Figure A5.3: Tabnet confusion matrix quintile ranking

The figure illustrate the ranking accuracy of the tabnet machine learning model. The y-axis show the predicted quintile rankings, and the x-axis show the actual quintile ranking, both based on net-alphas.



A6 Endurance Test

Table A6.1: Cumulative alpha, Morningstar

The table reports the hold-out-set annual cumulative net-alphas, of all quintile portfolios, across all algorithms. Included are also the Morningstar star rating portfolios. The second to last row exhibits the mean annual cumulative net-alpha of the 9 hold-out-sets, and the last row presents the cumulative net-alpha of the entire hold-out-sample, ranging from 2013 through 2021.

	Morningstar					**XGBoost**					**Random Forest**					**Neural Networks**				
	5 Star	4 Star	3 Star	2 Star	1 Star	5 Star	4 Star	3 Star	2 Star	1 Star	5 Star	4 Star	3 Star	2 Star	1 Star	5 Star	4 Star	3 Star	2 Star	1 Star
2013	.0259	.0148	.0075	-.0004	-.0075	.0715	.0275	.0152	.0045	-.0244	.0685	.0261	.0167	.0057	-.0230	.0368	.0186	.0129	.0159	.0079
2014	.0274	.0186	.0085	.0002	-.0102	.0745	.0379	.0178	.0062	-.0376	.0370	.0330	.0280	.0189	-.0200	.0447	.0235	.0205	.0124	-.0046
2015	.0384	.0324	.0191	.0106	-.0055	.1190	.0555	.0294	.0103	-.0419	.0703	.0641	.0475	.0242	-.0369	.0778	.0463	.0325	.0289	-.0174
2016	.0317	.0247	.0143	.0086	-.0018	.0882	.0514	.0225	.0092	-.0388	.1060	.0467	.0189	.0029	-.0407	.0849	.0390	.0221	.0063	-.0212
2017	.0323	.0225	.0156	.0076	-.0129	.0981	.0558	.0118	-.0156	-.0416	.0973	.0578	.0115	-.0176	-.0405	.0648	.0363	.0160	-.0101	-.0022
2018	-.0283	-.0317	-.0372	-.0353	-.0543	-.0045	-.0434	-.0724	-.0827	-.139	-.0048	-.0423	-.0739	-.0836	-.1370	-.0486	-.0493	-.0708	-.0730	-.1040
2019	-.0901	-.0884	-.0921	-.0832	-.0906	-.107	-.136	-.143	-.148	-.248	-.1070	-.1330	-.1460	-.1480	-.2500	-.1160	-.1380	-.1390	-.1450	-.2450
2020	-.0924	-.0916	-.0977	-.0888	-.0851	-.148	-.160	-.173	-.177	-.242	-.1410	-.1630	-.1720	-.1740	-.2480	-.1570	-.1590	-.1690	-.1720	-.2420
2021	-.0548	-.0654	-.0718	-.0705	-.0691	-.076	-.106	-.121	-.127	-.124	-.0768	-.1040	-.1210	-.1300	-.1230	-.0786	-.1050	-.1180	-.1290	-.1250
Average	-.0122	-.0182	-.0260	-.0279	-.0374	.0128	-.0241	-.0459	-.0578	-.1042	.0055	-.0238	-.0433	-.0557	-.1022	-.0101	-.0320	-.0436	-.0518	-.0838
Cumulative	-.1160	-.2190	-.1620	-.2310	-.2950	.0772	-.2240	-.3640	-.4300	-.6440	.0144	-.2220	-.3500	-.4200	-.6380	-.1190	-.2760	-.3500	-.3980	-.5680

Table A6.2: Significance test of top quintile portfolios

The table presents results from a series of Welch t-test, testing whether monthly net-alphas of our top-quintile portfolios are statistically greater than that of the top quintile portfolio from Morningstar's Star rating system. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
Morningstar	2.046	.02116	1.5036	.06724	-0.010575	.5042

Table A6.3: Metrics of top quintile portfolios with reduced positions

The table reports hold-out-sample performance metrics of the top quintile portfolios, and where the number of annual positions in the machine learner portfolios are reduced to 148 p.a. to match Morningstar's top quintile portfolio. The table also reports the base case with equal distributions of funds across the quintiles. The Morningstar top quintile portfolio takes an average of 148 positions p.a. across the hold-out-sample of 2013 through 2021.

		Top quintile reduced positions			**Top quintile base case**		
	Morningstar	XGBoost	RandomForest	TabNet	XGBoost	RandomForest	TabNet
Mean return	.066	.180	.152	.124	.162	.142	.119
Std.Dev.	.042	.113	.113	.100	.110	.109	.100
Sharpe ratio	1.39	1.42	1.21	1.12	1.32	1.17	1.07
Sortino ratio	1.62	1.69	1.36	1.17	1.50	1.27	1.13
Geometric return	.065	.172	.145	.118	.155	.135	.113
Cumulative return	.760	3.19	2.38	1.73	2.67	2.13	1.63
Cumulative alpha	-.116	.221	.138	-.006	.0772	.0144	-.119

Table A6.4: Significance test of top quintile portfolios, 148 positions p.a

The table presents results from a series of Welch t-test, testing whether monthly net-alphas of the top-quintile portfolios with 148 postions p.a. are statistically greater than that of the top quintile base portfolios. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

XGBoost		**Random Forest**		**Neural Networks**	
T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
3.3412	$p < .001$	3.1212	$p < .05$	3.1164	$p < .05$

A7 Robustness Checks of Risk-Adjustment Model

A7.1 Robustness of Fixed Income Risk-Adjustment Methodology

Table A7.1: Significance test of top quintile portfolios, robustness check

The table presents results from Welch t-tests, testing whether monthly net-alphas of the top quintile machine learning portfolios (Without Fixed Income funds) are statistically greater than the asset weighted portfolio, equally weighted portfolio, and the Morningstar top quintile portfolio. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
Equally Weighted	4.3629	$p < .001$	3.9384	$p < .001$	2.8082	$p = .0027$
Asset Weighted	3.8177	$p < .001$	3.3756	$p < .001$	2.2566	$p = .0125$
Morningstar	2.0135	$p = .0229$	1.4149	$p < .0795$	0.0498	$p = .4802$

Table A7.2: Significance test of ranking system, robustness check

The table presents results from a series of Welch t-test, testing whether monthly net-alphas of the top-quintile portfolio are statistically greater than that of lower quintiles. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
4 Star	2.496	$p = .0067$	2.224	$p = .0136$	2.056	$p = .0205$
3 Star	4.086	$p < .001$	3.295	$p < .001$	2.565	$p = .0055$
2 Star	5.634	$p < .001$	4.822	$p < .001$	3.497	$p < .001$
1 Star	10.11	$p < .001$	9.569	$p < .001$	5.930	$p < .001$

A7.2 Change of Risk-Adjustment Model, FF3F

Table A7.3: Significance test of top quintile portfolios, FF3F

The table presents results from Welch t-tests, testing whether monthly net-alphas of the top quintile machine learning portfolios (FF3F risk-adjustment) are statistically greater than the asset weighted portfolio, equally weighted portfolio, and the Morningstar top quintile portfolio. The test samples monthly net-alphas of the entire hold-out-period, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
Equally Weighted	4.5941	$p < .001$	4.2200	$p < .001$	2.3862	$p = .0090$
Asset Weighted	5.0139	$p < .001$	4.6927	$p < .001$	3.1565	$p < .001$
Morningstar	1.8573	$p = .0325$	1.3522	$p < .0890$	-.97505	$p = .8346$

Table A7.4: Significance test of ranking system, FF3F

The table present results from the Welch t-test, testing whether monthly net-alphas of the top quintile portfolio is statistically greater than that of the lower quintiles in the hold-out-period. The test samples monthly net-alphas of the entire hold-out-sample, ranging from 2013 through 2021.

	XGBoost		**Random Forest**		**Neural Networks**	
	T-statistic	P-value	T-statistic	P-value	T-statistic	P-value
4 Star	3.0483	$p = .0013$	2.5603	$p = .0056$	.87823	$p = .1904$
3 Star	4.6518	$p < .001$	3.9898	$p < .001$	2.0358	$p = .0215$
2 Star	5.8042	$p < .001$	5.1714	$p < .001$	2.8531	$p < .001$
1 Star	9.2111	$p < .001$	8.986	$p < .001$	5.7163	$p < .001$

Figure A7.1: Alpha distribution of FF3F

The figure illustrates the distribution in actual alpha of the funds contained in the dataset, with regards to the FF3F risk-adjustment model. The box plot show the negative outliers, minimum, the first quartile, median, the upper quartile, the maximum, and the positive outliers. The minimum is computed by $Q1 - 1.5 \times IQR$, and the maximum by $Q3 + 1.5 \times IQR$.

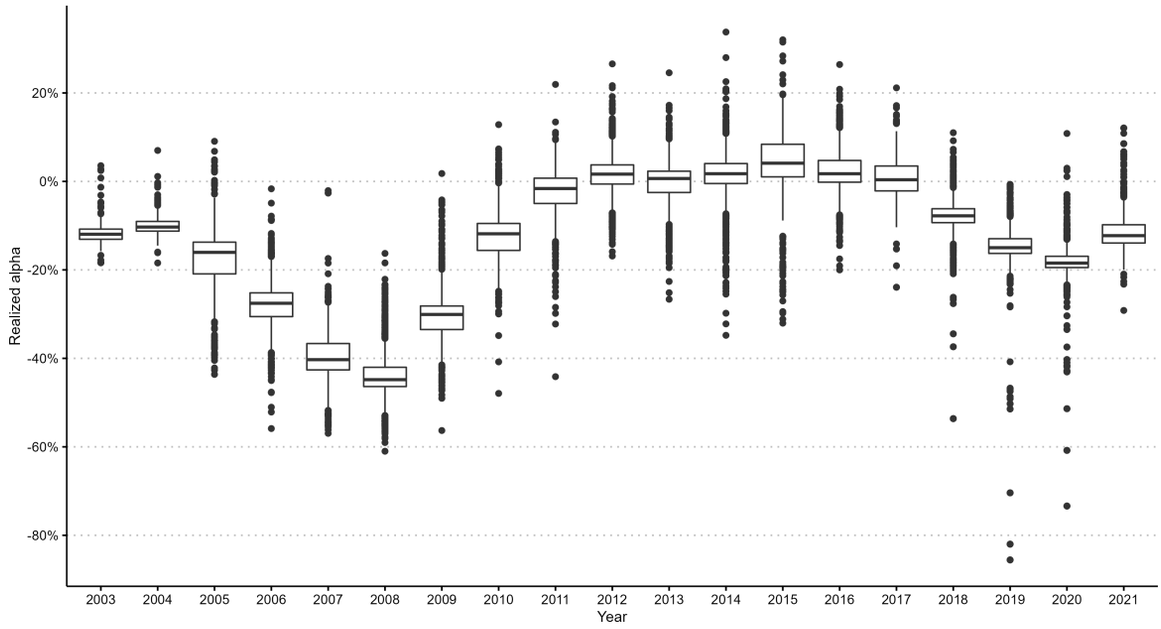


Table A7.5: Top-, median-, and bottom quintiles, FF3F & FF6F

This table present annual cumulative net-alphas of the top-, median-, and bottom quintile portfolios for all machine learners and for Morningstars star rating system. Included are also, the equally weighted and asset weighted benchmarks. The second to last row show the average annual cumulative alpha, and the final row show the cumulative net-alpha when investing in the portfolios at the start of 2013 and holding through 2021.

	Fama & French 3 Factor														**Fama & French 6 Factor**											
	Morningstar			**XGboost**			**RandomForest**			**Neural Networks**			**Benchmarks**		**Morningstar**			**XGBoost**			**RandomForest**			**Neural Networks**			**Benchmarks**	
	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	Equally	Asset	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	5 Star	3 Star	1 Star	Equally	Asset
2013	.0160	-.0056	-.0209	.0484	.0077	-.0720	.0298	.0092	-.0662	.0411	-.0037	-.0479	-.0029	-.0023	.0259	.0075	-.0075	.0715	.0152	-.0244	.0685	.0167	-.0230	.0368	.0129	.0079	.0183	.0152
2014	.0274	.0039	-.0138	.0824	.0178	-.0582	.0570	.0311	-.0595	.0474	.0260	-.0472	.0142	.0162	.0274	.0085	-.0102	.0745	.0178	-.0376	.0370	.0280	-.0200	.0447	.0205	-.0046	.0192	.0216
2015	.0470	.0259	.0031	.1360	.0308	-.0221	.1300	.0388	-.0296	.0703	.0440	.0017	.0463	.0507	.0384	.0191	-.0055	.1190	.0294	-.0419	.0703	.0475	-.0369	.0778	.0325	-.0174	.0332	.0394
2016	.0312	.0128	-.0015	.0893	.0193	-.0304	.0882	.0197	-.0286	.0299	.0280	.0086	.0236	.0279	.0317	.0143	-.0018	.0882	.0225	-.0388	.1060	.0189	-.0407	.0849	.0221	-.0212	.0257	.0312
2017	.0223	.0060	-.0148	.0567	.0064	-.0316	.0588	.0037	-.0324	.0344	.0073	-.0225	.0083	.0093	.0323	.0156	-.0129	.0981	.0118	-.0416	.0973	.0115	-.0405	.0648	.0160	-.0022	.0207	.0251
2018	-.0368	-.0452	-.0592	-.0356	-.0811	-.1330	-.0364	-.0816	-.1320	-.0554	-.0777	-.1090	-.0808	-.0861	-.0283	-.0372	-.0543	-.0045	-.0724	-.1390	-.0048	-.0739	-.1370	-.0486	-.0708	-.1040	-.0694	-.0720
2019	-.0890	-.0901	-.0894	-.0990	-.1430	-.2520	-.0991	-.1400	-.2500	-.1320	-.1360	-.2310	-.1550	-.2310	-.0901	-.0921	-.0906	-.1070	-.1430	-.2480	-.1070	-.1460	-.2500	-.1160	-.1390	-.2450	-.1580	-.2340
2020	-.0928	-.0971	-.0831	-.1400	-.1740	-.2430	-.1380	-.1720	-.2460	-.1500	-.1690	-.2410	-.1790	-.2390	-.0924	-.0977	-.0851	-.1480	-.1730	-.2420	-.1410	-.1720	-.2480	-.1570	-.1690	-.2420	-.1800	-.2400
2021	-.0525	-.0701	-.0680	-.0693	-.1160	-.1250	-.0745	-.1150	-.1270	-.0798	-.1160	-.1180	-.1090	-.0963	-.0548	-.0718	-.0691	-.0760	-.1210	-.1240	-.0768	-.1210	-.1230	-.0786	-.1180	-.1250	-.1110	-.0991
Average	-.0141	-.0288	-.0386	.0076	-.0480	-.1074	.0017	-.0451	-.1079	-.0216	-.0440	-.0895	-.0482	-.0612	-.0122	-.0260	-.0374	.0128	-.0459	-.1042	.0055	-.0433	-.1022	-.0101	-.0436	-.0838	-.0446	-.0569
Cumulative	-.1310	-.2390	-.3020	.0316	-.3760	-.6550	-.0193	-.3590	-.6570	-.2030	-.3530	-.5890	-.3790	-.4660	-.1160	-.1620	-.2950	.0772	-.3640	-.6440	.0144	-.3500	-.6380	-.1190	-.3500	-.5680	-.3590	-.4450