Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Samuli Turunen

Feasibility of Nonlinear Multifactor Classifiers for Predicting Share Returns

Master's Thesis
Espoo, April 5, 2019

Supervisor: Professor Antti Punkka
Instructor: M.Sc. Antti Sivonen, Evli Fund Management Company Ltd

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author:	Samuli Turunen		
Title:	Feasibility of Nonlinear Multifactor Classifiers for Predicting Share Returns		
Date:	April 5, 2019	**Pages:**	vii + 79
Major:	Systems and Operations Research	**Code:**	SCI3055
Supervisor:	Professor Antti Punkka		
Instructor:	M.Sc. Antti Sivonen, Evli Fund Management Company Ltd		

This thesis determines the feasibility of nonlinear multifactor models for predictive classification of share returns. To accomplish this, a simple model fitting methodology is developed, applying machine learning methods to fit the factor models. Their performance is examined in terms of classification and portfolios created based on model predictions. Portfolio performance is furthermore compared to benchmarks, including a replication of the industry standard Fama-French Five-Factor Model.

Research applies data of US stocks covering a period from 2004 to 2016 for the predictions and portfolios. The data consists of return and accounting data. A three-month lag is imposed to accounting data to account for the reporting delay. Thus all the data applied in the predictions is available to the market at the time. The smallest companies are excluded from the data sample due to their disproportionate number and small market value. The factors calculated from the data are gathered from recent scientific literature.

The results provide weak support to the feasibility of nonlinear factor models to the task examined in this thesis. Additionally, the results are less than robust with respect to changing the model parameters. Moreover, options were identified to extend the model fitting methodology with possibly significant performance improvements.

The findings and the identified improvements are weakly in favor of the feasibility of nonlinear multifactor models for predictive classification of share returns. The conclusion inspires multiple directions for further research, which are presented along with the conclusion.

Keywords:	factor model, nonlinear, return prediction, classification, machine learning
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Master's Programme in Mathematics and Operations Research

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Samuli Turunen		
Työn nimi:	Epälineaaristen multifaktoriluokittimien soveltuvuus osakkeiden tuottojen ennustamiseen		
Päiväys:	5. huhtikuuta 2019	**Sivumäärä:**	vii + 79
Pääaine:	Systems and Operations Research	**Koodi:**	SCI3055
Valvoja:	Professori Antti Punkka		
Ohjaaja:	KTM Antti Sivonen, Evli-Rahastoyhtiö Oy		

Tässä diplomityössä määritetään epälineaaristen multifaktorimallien soveltuvuus osakkeiden tuottojen ennustavaan luokitteluun. Soveltuvuuden määrittämiseksi kehitetään yksinkertainen mallinsovitusmetodologia, jossa sovelletaan koneoppimista mallien sovittamiseen. Niiden toimintaa tarkastellaan luokittelun tarkkuuden ja luokittelun perusteella laadittavien osakeportfolioiden mukaan. Portfolioiden tuottoa ja ominaisuuksia vertaillaan yksinkertaisiin malleihin, joiden joukossa on alalla laajasti käytetty Fama-Frenchin viiden faktorin malli.

Tutkimus hyödyntää yhdysvaltalaisten pörssilistattujen yritysten tilinpäätös- ja osakedataa ja portfolioiden tarkastelu kattaa vuodet 2004-2016. Tilinpäätösdataan sovelletaan kolmen kuukauden viivettä raportointiviivästyksen huomioimiseksi. Täten kaikki ennustamiseen käytetty data on markinnoiden saatavissa ennusteen tekohetkellä. Pienimmät yritykset jätetään tarkastelun ulkopuolelle niiden suhteettoman suuren lukumäärän ja pienen markkina-arvon takia. Datasta laskettavat faktorit koostetaan viimeaikaisesta tieteellisestä kirjallisuudesta.

Tulokset tukevat heikosti epälineaaristen multifaktorimallien soveltuvuutta työssä tarkasteltavaan tehtävään. Tulokset eivät ole kuitenkaan robusteja mallien parametrien muutoksille. Työssä tunnistettiin kuitenkin mahdollisuuksia kehittää mallien sovitusmenetelmiä, joilla saattaa olla merkittäviä vaikutuksia suorituskykyyn.

Työn tulokset ja tunnistetut kehitysmahdollisuudet puoltavat heikosti epälineaaristen multifaktorimallien soveltuvuutta osakkeiden tuottojen ennustavaan luokitteluun. Yhteenvedon perusteella esitetään myös useita mahdollisuuksia jatkotutkimukselle.

Asiasanat:	faktorimalli, epälineaarinen, tuoton ennustaminen, luokittelu, koneoppiminen
Kieli:	englanti

Acknowledgements

It is easy to forget the extent to which we are products of our environment, most importantly affected by the people we surround ourselves with.

Therefore, I wish to express my gratitude to my mother, father and sister, as well as my partner for the guidance and support along my way.

Furthermore I would like to thank my advisor Antti, as well as Mattias and Peter at Evli for both exceptional advices and enjoyable discussions over the course of this project. Last, but not least, I thank Antti at Aalto for the efforts in his role as the supervisor including valuable comments and critique, which greatly helped writing this thesis.

Espoo, April 5, 2019

Samuli Turunen

Abbreviations and Acronyms

AMEX	American Stock Exchange
APT	Arbitrage Pricing Theory
BE	Book Equity
BM_{FF5}	Benchmark, FF5 based
BM_{LR}	Benchmark, Logistic Regression
BM_{simple}	Benchmark, simple
CAPM	Capital Asset Pricing Model
CART	Classification and Regression Tree
CCM	CRSP/Compustat Merged database
CMA	Conservative minus active, a factor in the FF5
CRSP	Center for Research of Security Prices
FF3	Fama-French Three-Factor Model
FF5	Fama-French Five-Factor Model
GB	Gradient Boosting
HML	High minus low, a factor in the FF3 and FF5
K-NN	K-Nearest Neighbors
LR	Logistic Regression
ME	Market Equity
NYSE	New York Stock Exchange
RF	Random Forest
RMW	Robust minus weak, a factor in the FF5
SMB	Small minus big, a factor in the FF3 and FF5
SR	Sharpe Ratio
SVM	Support Vector Machine

Contents

Chapter 1

Introduction

An investor attempting to earn higher than average long-term returns in the stock market faces a problem many consider impossible. In spite of the justified doubt, many still try to *beat the market* both in academia and in practice. Academic interest stems from achieving a greater understanding of the financial markets. The motivation for an individual investor is obvious, while a fund manager would be able to provide more attractive products to their customers. Intuitively, an interested party should develop a method for identifying the shares that are likely to yield high returns in the future and invest in them. One method to accomplish this could be to start analyzing individual companies in effort to identify the most likely high performing ones. An alternative approach would be a systematic method to search for and to exploit common factors that predict returns and differentiate the high and low performing shares. The systematic approach has the interesting advantage of providing a selection method that scales well with the number of shares considered. An army of professional analysts would be required in order to have the capacity to investigate the business and performance of individual companies, whereas a *factor model* could tackle the same problem using minimal labor and time once the model is developed. Regardless of the approach, the problem of making any sophisticated predictions about the stock market is extremely difficult.

Applying a factor model to the problem of explaining and predicting stock returns is among the major innovations of 20^{th} century financial economics (Cochrane, 1999a). In principle, the factor model assumes that the return of a share is a function of some systematic factors and the exposure of the share to each factor. The most prominent factor models are linear, as introduced by Pitsillis (2004). They are easy to interpret and have strong ties

to economic intuition, but are poorly suited to capture any nonlinearity in factor returns or interactions between the factors. This leads to the reasonable assumption that a method able to exploit the information in nonlinear returns or interactions could predict stock returns more accurately than a linear model. Motivated by these potential "blind spots" of the linear model, nonlinear and nonparametric approaches have been suggested. Early examples include applying neural networks to estimate a nonlinear factor model for stock selection by Levin (1995) and nonlinear pricing kernels by Dittmar (2002). Modern algorithms for fitting nonlinear models are included under the popular *machine learning* umbrella, including neural networks applied in Levin's work. A common advantage of such algorithms is that an explicit definition of the functional relationship between the input and output variables is not required. This makes them a flexible and efficient alternative for complex problems, such as return prediction in the factor model framework. However, literature on the feasibility of nonlinear factor models for the problem is scarce.

Thus, the objective of this thesis is to determine the feasibility of nonlinear factor models for classifying future stock returns. The objective is interesting to the asset management industry in particular because a nonlinear model could be implemented in investment products with reasonable effort in the existing factor model framework. The determination of the feasibility is accomplished by implementing promising methods for fitting nonlinear factor models to classify future share returns, comparing the models to a linear benchmark in terms of prediction accuracy, and by comparing the performance of portfolios created according to the predictions of the models. Nonlinear models are assumed to exploit the information in nonlinear factor returns and factor interactions, which are the two assumed sources of additional information a linear model captures poorly. Further motivation stems from the flexibility and efficiency of nonlinear models in complex prediction problems.

This thesis applies data on US stocks. Reflecting on the factor model framework, the scope excludes both the exploration of new factors or the assessment of the significance of documented factors. The factors applied in the models are gathered from recently published scientific literature. Furthermore, extending economic models in the direction of nonlinear models or to link the built models into economic intuition is excluded from the scope, relaxing the requirements for identifying the functional relationships between the factors and returns. Similarly, the objective excludes exhaustive research to the contribution of each phenomena to the relative performance of the nonlinear models. The returns and characteristics of portfolios constructed

based on the predictions of the models are compared, but the scope does not include developing or analyzing any detailed investment or trading strategy.

The rest of this thesis in organized as follows. In Chapter 2, the background of the subject is introduced, including the most prominent factor models in asset pricing, the current state of the factor research and the drawbacks of current models. Since this thesis applies machine learning methods to fit factor models, a brief introduction to machine learning is also provided, focusing on related applications and challenges in finance. Machine learning methods used to fit the models, along with comparison methods, are presented in detail in Chapter 3. Chapter 4 describes the examined dataset, including the source of the data, the selection of the shares included in the analysis, and the selection and construction of the factor variables from the data. The implementation of the experiments using the dataset is described in Chapter 5. Moreover, Chapter 5 presents the results, including prediction performance and portfolio performance metrics, characteristics and sensitivity analysis. Chapter 6 discusses the results reflecting on related research, and Chapter 7 presents the conclusions of the thesis along with propositions for further research.

Chapter 2

Background of Multifactor Models

This chapter presents the background of the subject of this thesis. The chapter begins by discussing the related terminology and outlining the terminology used in this thesis. The overview includes the most prominent factor models in asset pricing as well as the current status of related research. As machine learning methods are used to fit the nonlinear models, an introduction to machine learning with examples of factor model related literature is presented.

2.1 Factor Models in Asset Pricing

2.1.1 Terminology

An unambiguous definition for a *factor* is lacking in related literature. In this thesis, a factor is a systematic driver of asset returns, whereas a *factor model* refers to a model explaining asset returns by exploiting one or more factors. The ambiguousity could be viewed to arise from the lack of agreed criteria for calling a driver of asset returns a factor. Ross (2017) points out this concern by drawing attention to to the required statistical significance for a factor as well as noting "For me, perhaps even more troubling than the empirical evidence is the lack of a strong economic foundation for many of the factor candidates." The term *anomaly* also appears in related literature (Hou et al., 2017). Anomalies, when identified, can also be seen as drivers of

asset returns, but they may not meet the criteria of being accepted as factors. They are usually referred to as factors where risk-based explanations do not apply.

2.1.2 From Single to Multifactor Models

To focus on the most relevant theory, we skip the fundamental asset pricing theory of risk and investor preference by appealing to the intuition behind risk and return: higher risk must be compensated with higher expected returns, also known as *risk premium*. With this principle, we introduce some of the most prominent factor models in asset pricing. Ang (2014, Chapter 6) presents further details on factor models from a risk-premium perspective.

The Capital Asset Pricing Model (CAPM), based on the works of Sharpe (1964), Lintner (1965) and Mossin (1966), is regarded as the first factor model in asset pricing. Derived from market equilibrium conditions, the CAPM explains the returns of an asset with a market risk premium:

$$\begin{aligned} \mathbb{E}(r) - r_f &= \frac{\text{Cov}(r, r_m)}{\text{Var}(r_m)}(\mathbb{E}(r_m) - r_f) \\ &= \beta(\mathbb{E}(r_m) - r_f), \end{aligned} \tag{2.1}$$

where Var and Cov denote the variance and covariance operators, $\mathbb{E}$ is the expected value operator, r is the asset return, r_f is the risk-free rate and r_m is the market portfolio return. The risk-free rate is defined as the interest on a deposit with no counterparty default risk, a theoretical concept usually approximated with US Treasury bond yields. The model states that the expected return of an asset in excess of the risk-free rate depends on its tendency to move with the market portfolio with return r_m. Shares with $\beta = 1$ behave like the market portfolio, whereas $\beta > 1$ implies the returns tend to be greater in amplitude than those of the market portfolio, and the opposite for $\beta < 1$ (Ang, 2014, Chapter 6).

According to Ang (2014, p. 202), the general basis for multifactor models was set by the Arbitrage Pricing Theory (APT). The APT, published by Ross (1976), states that the returns of an asset can be explained as a linear combination of returns of several factors and the asset's exposure to these factors:

$$r = \alpha + \sum F_k x_k, \tag{2.2}$$

where α is a constant, F_k is a systematic factor explaining asset returns and x_k is the factor loading.

Building on previous results in asset pricing, Fama and French (1993) introduced their famous three-factor model (FF3). Referring to the introduction to the FF3 by Ang (2014, Chapter 7), FF3 extends CAPM by adding two factors capturing the *size effect* and the *value effect*, respectively the outperformance of small companies relative to big companies and the outperformance of high *value* companies relative to low *value* companies. In FF3, size is measured as the company market capitalization. Value, on the other hand, is measured as the market capitalization of the company divided by the book value of the equity of the company, also known as the *book-to-market* ratio. The FF3 model is

$$\mathbb{E}(r) = r_f + \alpha + \beta(r_m - r_f) + s\text{SMB} + h\text{HML}, \tag{2.3}$$

where α, β, s and h are regression coefficients. The first of the new factors, *small minus big* (SMB), captures the size effect as the return differential of small and big companies. Similarly, *high minus low* (HML) captures the value effect as the return differential of high and low value companies. *Factor mimicking portfolios* are used as proxies for these factors, which are designed to capture the effects by averaging over many stocks. The construction of these portfolios is described in detail in Section 3.2.1.

The FF3 model is the predecessor to two extensions. Carhart (1997) published a four-factor model in 1997, extending the FF3 model by adding a momentum factor. Momentum quantifies the tendency of outperforming shares to keep outperforming and underperforming to keep underperforming. It is defined as the average return over the past year. Fama and French (2015) extended their FF3 model to a five-factor model (FF5) by adding factors for profitability and investment activity. The FF5 model is

$$\mathbb{E}(r) = r_f + \alpha + \beta(r_m - r_f) + s\text{SMB} + h\text{HML} + r\text{RMW} + c\text{CMA}, \tag{2.4}$$

where α, β, s, h, r and c are regression coefficients. The first of the added factors, *robust minus weak* (RMW), refers to robust and weak profitability. The latter, *conservative minus active* (CMA) stands for conservative and active investments, that is, low and high investment activity companies. The two additional factors RMW and CMA are constructed by applying the same principle as with the SMB and HML factors.

The discussed factor models can be applied with monthly returns, following the example set by Fama and French (2015). Moreover, the discussed models do not explicitly impose a lookback on historical data, even though in practice they rely on it. This reliance arises from the estimation of the factor coefficients, which requires the historical returns of the asset, the market portfolio, risk-free rate and the returns of the considered factors. On the other hand, most company financials can be interpreted to contain an implicit lookback to historical data, as for example the assets or the profits of a company at a given time have accumulated over a historical period.

The most successful factor models and their performance has been put under scrutiny and research indicates that the models cannot perfectly describe asset returns, but also that the newer multifactor models outperform the CAPM. Indeed, Ang (2014, p. 197) states that "the CAPM is well known to be a spectacular failure."

2.1.3 The Zoo of Factors

The publication and success of the factor models has inspired research answering the logical question: what factors exist and how can they be exploited? Cochrane discusses the so called new multifactor world and its implications in two excellent companion articles "New Facts in Finance" (Cochrane, 1999a) and "Portfolio Advice in a Multifactor World" (Cochrane, 1999b). Over a decade later, Cochrane (2011) notes that the research for new factors has lead to a "zoo of factors" to where he demands discipline to map out the findings and examine the explanatory power of the found drivers of returns.

The current focus of the research in factor models is on exploring new factors, and, on the other hand, on assessing which of the documented factors provide additional information on asset prices. Traditionally the first problem has been approached by deriving the factors from empirical evidence with strong ties to economic theory, such as the FF3 model. During the last decades, the traditional approach has been challenged by data mining for new factors enabled by the increased availability of computational power and development in data mining algorithms. Harvey et al. (2016) conduct an exhaustive review on the existing literature on found factors and notice an exponential increase in the number of documented factors starting from 1962 until 2012 leading to at least 316 factors. According to their work, the rapid increase in the number of documented factors is at least partly caused by the advances in machine learning and data mining. This imposes new requirements for

a newly found driver of asset returns to be accepted as factor. Therefore, Harvey et al. (2016) suggest a new hurdle of 3 for the t-statistic of the excess returns of a factor to adjust for multiple testing usually occurring in data mining.

2.1.4 Motivation for Nonlinear Models

As applied in asset pricing for example in the FF3 model (2.3), the linear factor model presented in (2.2) does not account for nonlinear factor returns or factor interactions. It should be noted, that such effects could be described by the model by parametrization, but defining a functional form for all factors and interactions would be a daunting task.

The idea of applying nonlinear models in the factor model framework is not new. For example, McMillan (2001) tested the relationships of asset returns with macroeconomic and financial variables using non-parametric methods and found evidence supporting a nonlinear relationship between returns and interest rates. However, such evidence was not found with other tested variables and the increase in performance was marginal. Nonlinear return-variable relationships have motivated the use on nonlinear models already in the early neural network example by Levin (1995). Furthermore the monotonicity of asset returns under some conditioning variable has been under examination and statistical tests for monotonicity have been developed, mainly by Romano and Wolf (2013) and by Patton and Timmermann (2010). In summary, the nonlinearity of asset returns with respect to a conditioning variable can be viewed as a reasonable extension to the factor model framework.

While the return-factor relationships have been explicitly addressed in previous research, the literature on factor interactions is scarce. Intuitively the task is more difficult than the simple return-variable relationship, because there are many variables instead of two. Since the most popular factor models, such as the FF3, implicitly assume that the factors are independent, the interactions are mainly ignored in the research considering these models. On the other hand, approaches using a model that can handle interactions barely characterize them, examples including the early work by Levin (1995). One can conclude, that no widely accepted methods exist for characterizing multiple factor interactions. However, interactions between specific factors have been examined and the evidence suggests that some interactions exist, as concluded for example by Fama and French (2008).

2.2 Nonlinear Multifactor Models via Machine Learning

The umbrella term *machine learning* (ML) can be defined with varying extent. Lopez de Prado (2018, p. 15) describes a machine learning algorithm via its function: "An ML algorithm learns patterns in a high-dimensional space without being specifically directed." In other words, a machine learning algorithm estimates a function to map a usually high-dimensional input data matrix to a target vector. Bishop (2006, p. 3) differentiates different types of machine learning applications to *supervised learning* in cases where the target vector is known, *unsupervised learning* when the target vector is unknown and the algorithm attempts to discover patterns within the data and *reinforcement learning* when the algorithm seeks a suitable action to maximize a reward. He also presents a division of applications based on the type of the target vector: in *classification* the target vector consists of categorical or discrete variables, whereas in *regression* the target is continuous. For a thorough overview of machine learning, we refer to Bishop (2006), where common algorithms are also presented making apparent that machine learning consists of both linear and nonlinear models. Combining the presence of nonlinear models within machine learning with their common ability to learn patterns from data motivate the selection of machine learning algorithms as methods for fitting the nonlinear multifactor models in this thesis.

In contrast, major challenges follow. Reflecting on the concern of the lacking economic foundation for factor candidates by Ross (2017), one could also argue that a factor model should have rigorous economic intuition behind it. This principle is violated by using machine learning to fit the nonlinear models, as the methods are characterized by their ability to discover patterns without directions. A major technical challenge is *overfitting* when seeking for a model with the smallest prediction error, or *loss*. Bishop (2006, pp. 6 – 7) explains overfitting intuitively with an example, but the concept can be described simply as a model being overfit, if it performs well on the input data sample, but generalizes poorly to new observations from the same data generating process. Overfitting is particularly difficult to avoid in financial applications of machine learning, mainly due to a low signal-to-noise ratio, as noted by Lopez de Prado (2018, p. 101).

2.2.1 Related Research

Related research applies various machine learning algorithms for stock return prediction with a factor model or a similar approach. Levin (1995) published an early example applying neural networks. Later examples include the work by Fan and Palaniswami (2001) applying Support Vector Machines (SVM) for predicting outperforming stocks in the Australian stock exchange, and the publication by Huerta et al. (2013) also applying SVMs for classifying stocks according to high and low future returns. Research comparing multiple algorithms includes the work by Sugitomo and Minami (2018) with SVMs, Gradient Boosting (GB) and neural networks as well as the comparison by Imandoust and Bolandraftar (2014) applying Decision Trees, Random Forests (RF) and Naïve Bayes classifiers. Moreover, Ballings et al. (2015) compares a number of different methods for predicting price movements of European shares. The working paper by Gu et al. (2018) includes one of the most exhaustive syntheses in related research. They formulate the prediction problem as regression and document support for the high comparative performance of nonlinear models. Furthermore they find that a small set of momentum, liquidity and volatility variables dominate the predictions of all models. In general, the related research reports positive findings in support of various nonlinear algorithms being able to improve prediction accuracy and portfolio performance with respect to the chosen benchmark.

However, the degree to which the related research seems to agree on the advantages from applying these algorithms to stock return prediction is rather striking. Specifically, Harvey et al. (2016) point out that replication studies are rarely published in finance, as opposed to other fields of science.

Chapter 3

Model Fitting and Comparison Methods

This chapter describes the methods applied to determine the feasibility of nonlinear multifactor classifiers for predicting share returns. This includes an introduction to model fitting and selection as well as a description of the benchmark models and the methods to fit the nonlinear models. Moreover, this chapter presents the model comparison methods, including the classification performance metrics, the construction of *model based portfolios*, and the metrics applied to compare the performance of those portfolios. A model based portfolio hereafter refers to a portfolio constructed according to a model introduced in this chapter.

The methods fall into two distinctive categories: classification and investment portfolios. Such a combination of methods is required, because we are interested in the classification performance of the different models, but also in the performance of the model based portfolios based on the classifier outputs. An outline of the analysis procedure is the following:

1. Fitting of and predicting with a classifier
2. Measuring classification performance
3. Constructing portfolios based on the predictions of a classifier
4. Measuring model based portfolio performance

3.1 Model Fitting and Selection

3.1.1 Model Fitting

The focus of this thesis is on classification. Therefore the following theoretical background of model fitting, as well as the models introduced thereafter, are applied in the context of classification, even though some parts of the theory and models generalize also to regression. For the theory presented in this section we refer to Bishop (2006).

Let $\mathbf{x} \in \mathbb{R}^{n \times m}$ be a matrix of model input data and $\tau \in L^n$ a vector of class labels associated with $\mathbf{x}$. Now there are n observations with input data over m variables. Furthermore L is the set of class labels indicating membership to a class $\mathcal{C}_k$, where $k = 1, \ldots, K$. Moreover, we denote the data by $\mathcal{D} = \{\mathbf{x}, \tau\}$. The aim of training a model is to find a function $y(\mathbf{x}, \mathbf{w}) : \mathbb{R}^{n \times m} \to L^n$, where $\mathbf{w}$ is model parameters, that maps $\mathbf{x}$ to τ as accurately as possible. The modeling task can be formulated as a either a *discriminative* or *probabilistic* problem.

In discriminative modeling we are simply interested in the agreement of the model output labels $\mathbf{y}$ with the true labels τ. Assuming a probabilistic model instead, we model the conditional probability of the class $\mathcal{C}_k$ given the input data $\mathbf{x}$:

$$y(\mathbf{x}, \mathbf{w}) = \{p(\mathcal{C}_1|\mathbf{x}), \ldots, p(\mathcal{C}_K|\mathbf{x})\}$$

In other words, we are now interested in modeling the probability distribution defined by $\mathcal{D}$. The output labels can also be obtained with a probabilistic model. For an observation $\mathbf{x}_i$ the output label corresponds to the class k that maximizes $p(\mathcal{C}_k|\mathbf{x}_i)$. *Bayes' theorem* is important for obtaining the probabilities $p(\mathcal{C}_k|\mathbf{x})$, which states, that

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}.$$

where the term $p(\mathcal{C}_k|\mathbf{x})$ is called the *posterior probability*, $p(\mathbf{x}|\mathcal{C}_k)$ the *likelihood* and $p(\mathcal{C}_k)$ the *prior*.

The procedure of fitting a classification model depends on the type of the selected model. Therefore the principles of fitting the function $y(\mathbf{x}, \mathbf{w})$ are introduced later along with presenting the models applied in this thesis.

However, one principle is applicable to both discriminative and probabilistic modeling. First assume that we have fitted a model with parameters $\mathbf{w}$ on a fixed data $\mathcal{D}$. Moreover we are interested in the quality of the model fit to the data, measured by a *total error*. With a fixed data, the error is dependent only on the model parameters $\mathbf{w}$ and the total error can be measured with a function $E(\mathbf{w})$. The principle decomposes the total error to a data dependent error $E_{\mathcal{D}}(\mathbf{w})$ and to a parameter dependent error $E_{\mathcal{W}}(\mathbf{w})$:

$$E(\mathbf{w}) = E_{\mathcal{D}}(\mathbf{w}) + \lambda E_W(\mathbf{w}), \tag{3.1}$$

where λ is a *regularization parameter*. The regularization parameter makes the principle useful, since it allows controlling model complexity. This may help in avoiding overfitting, as a less complex model tends to be less prone to it.

3.1.2 Model Selection

In model fitting the focus was on finding the best version of the selected model. Model selection, in contrast, addresses the question of finding the best alternative from several fitted models of different type. As overfitting is a major challenge both in model fitting and selection, the effect of overfitting can be quantified by decomposing a model expected error to *bias*, *variance* and *noise* as

$$\text{expected error} = \text{bias}^2 + \text{variance} + \text{noise},$$

where bias relates to the extent to which the average prediction deviates from the underlying patterns in the data, variance to the sensitivity of the predictions on the choice of the data sample and noise measures the intrinsic noise in the data that cannot be learned by a model. Moreover, a trade-off exists between bias and variance when minimizing error: models tend to have either high variance and low bias, or low variance and high bias. An overfit model would fall into the former group. We refer to Bishop (2006, Chapter 3.2) for further details.

The process of *model selection* tries to find the best model, that is, a model with a minimal error provided a suitable bias-variance balance. The outline for the process is presented in Bishop (2006, Chapter 1.3) with examples of typical methods which are summarized here.

For large quantities of data, a feasible approach is to partition the data to a *training set* and a *validation set*. The training set is used to fit the considered models, after which the independent validation set is used to evaluate the error of the fitted models on previously unseen data. Model selection can already be accomplished based on the validation set results, if we are only considering models of the same type with different parameters. However, if a selection has to be made between different types of models, an additional step to the process is required. In the third step, we apply an independent partition *test set* to evaluate the errors of the fitted models of different types and select the model with the smallest error. In addition to comparing different model types, the test set could be used to select between models from the same family with different *hyperparameters* (Bishop, 2006, p. 71). They are parameters of the model that are not learned from the data, such as the degree of a polynomial function.

For limited data, *cross-validation* is a common alternative to data partitioning. In cross-validation the data is divided into S groups, after which one group is reserved to be used as the validation set and the remaining $S - 1$ groups are used as the training set. The same procedure is repeated for all S groups. The special case of cross-validation, where all of the S groups consist of a single observation is known as *leave-one-out* validation. Lopez de Prado (2018, Chapter 7) discusses the particular challenges of applying cross-validation in finance and concludes that the standard cross-validation methods are poorly suited for financial applications.

3.2 Benchmark Models

We build three linear benchmark models to compare the nonlinear models to. Two of the benchmarks, an FF5 based benchmark described in Section 3.2.1 and a simple benchmark described in Section 3.2.2 are not predictive models, but rather follow deterministic rules to form portfolios. Therefore, they are not explicitly subject to the model fitting and selection methods presented in Sections 3.1.1 and 3.1.2. The third benchmark, logistic regression is a predictive model, as described in Section 3.2.3.

3.2.1 Fama-French Five-Factor Model

The widely used FF5 is estimated by *linear regression*. In linear regression, the problem is to find the parameters $\mathbf{w}$ that minimize the error for the

model

$$y = \sum_{i=1}^{m} \mathbf{w}_i \mathbf{x}_i + \epsilon_i,$$

where ϵ_i is an error term. Common methods to fit a linear regression include *least squares* and *maximum likelihood* Bishop (2006, pp. 4–6, 140–143). Linear regression is easy to interpret and computationally efficient, but by definition cannot capture any nonlinear effects between the input features and the inputs, unless explicit variable transformations are made, such as substituting $\hat{x} := x^2$. Furthermore it performs poorly with a high-dimensional data matrix, a phenomenon know as the *curse of dimensionality* (Bishop, 2006, Chapter 1.4).

The coefficients of FF5 in (2.4) can be solved as a linear regression problem, if the market portfolio return and the factors SMB, HML, RMW and CMA are known. The coefficient β is alike the CAPM β in (2.1), whereas the other four factors are formed via factor mimicking portfolios. The following method is based on so called 2×3 sorts, as defined by Fama and French (2015). Firms are sorted by their size, and the New York Stock Exchange (NYSE) companies median size is used to divide the firm into groups of small and big. Both groups are subsequently and independently sorted by the book-to-market ratios using the 30^{th} and the 70^{th} percentile of NYSE companies as the breakpoints into three groups within each size group, forming six portfolios used to construct the HML factor. A similar procedure is repeated for operating profitability and investment activity to construct the RMW and CMA factors, respectively. The resulting 18 portfolios as summarized in Table 3.1.

Table 3.1: The 2×3 sorts used to construct the FF5 factor portfolios. The shares are sorted by their size, book-to-market ratio, profitability and investment activity to form 18 subportfolios. The subportfolio accronyms are formed by the following keys: S, small size; B, big size; H, high book-to market; Nb, neutral book-to-market; L, low book-to-market; R, robust profitability; Np, neutral profitability; W, weak profitability; C, conservative investment activity; Ni, neutral investment activity; A, aggressive investment activity.

		Size	
		Small	Big
Book-to-market	High	SH	BH
	Neutral	SNb	BNb
	Low	SL	BL
Profitability	Robust	SR	BR
	Neutral	SNp	BNp
	Weak	SW	BW
Investment	Conservative	SC	BC
	Neutral	SNi	BNi
	Aggressive	SA	BA

The FF5 factors are calculated from the subportfolios presented in Table 3.1 as follows:

$$\begin{aligned}
\mathrm{SMB} = \Big(& \frac{(\mathrm{SH} + \mathrm{SNb} + \mathrm{SL}) - (\mathrm{BH} + \mathrm{BNb} + \mathrm{BL})}{3} \\
& + \frac{(\mathrm{SR} + \mathrm{SNp} + \mathrm{SW}) - (\mathrm{BR} + \mathrm{BNp} + \mathrm{BW})}{3} \\
& + \frac{(\mathrm{SC} + \mathrm{SNi} + \mathrm{SA}) - (\mathrm{BC} + \mathrm{BNi} + \mathrm{BA})}{3} \Big) / 3, \\
\mathrm{HML} = & \frac{(\mathrm{SH} + \mathrm{BH}) - (\mathrm{SL} + \mathrm{BL})}{2}, \\
\mathrm{RMW} = & \frac{(\mathrm{SR} + \mathrm{BR}) - (\mathrm{SW} + \mathrm{BW})}{2}, \\
\mathrm{CMA} = & \frac{(\mathrm{SC} + \mathrm{BC}) - (\mathrm{SA} + \mathrm{BA})}{2}.
\end{aligned}$$

While FF5 is usually applied as a regression model on asset returns, in this thesis we also construct a benchmark model based on the FF5 as an equal

weighted portfolio of replications of the factors SMB, HML, RMW and CMA. The resulting benchmark is denoted by $\mathrm{BM}_{\mathrm{FF5}}$.

3.2.2 A Simple Benchmark

By definition, the FF5 model is specified only for five factors, whereas a larger number factors is considered in this thesis. Comparing the performance of the nonlinear models only to FF5 would raise the question if the possible outperformance of the other models could simply be explained by the greater number of variables included in the models. Therefore, a simple linear benchmark containing all considered factors is included and constructed as follows.

A subportfolio is formed with respect to each considered factor by sorting the shares according to the variable and substracting the return of the lowest decile shares from the top decile shares. This corresponds to taking a *long* position in the top decile and a *short* position in the bottom decile. The top-bottom difference return is similar to the method used by Hou et al. (2017). The simple benchmark portfolio is formed by equal-weighting the individual factor subportfolios. The resulting benchmark is denoted by $\mathrm{BM}_{\mathrm{simple}}$.

3.2.3 Logistic Regression

Logistic regression (LR) is a probabilistic linear classification method, despite its name refers to regression. Bishop (2006, p. 198) writes logistic regression for as a model of the posterior probability of class $\mathcal{C}_k$ as

$$p(\mathcal{C}_k|\phi) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{3.2}$$

where $a_k = \mathbf{w}_k^T\phi$ and $\exp(\cdot)$ denotes $e^{(\cdot)}$. Here $\phi(\mathbf{x})$ is a vector of basis functions, which are nonlinear functions of the input data. Indeed, logistic regression can be constructed as a linear combination of nonlinear transformations of the data, but in this thesis we consider only the most simple case of $\phi(\mathbf{x}) = \mathbf{x}$. Now the weights $\mathbf{w}_k$ are the model parameters that need to be learned, which can be done by maximizing a log-likelihood function. We refer to Bishop (2006, pp. 196 – 210) for the derivations of (3.2) and for the method of learning the parameters $\mathbf{w}_k$, which are both omitted here.

Logistic regression, as described, is subject to similar limitations as linear regression. To address the performance of logistic regression especially with high-dimensional data, a *lasso* method introduced by Tibshirani (1996) can be applied, alike many other models. The lasso is a regularization method, that adds the following regularization term in the model error in (3.1):

$$\frac{1}{2}\sum |w_j|,$$

as described by Bishop (2006, p. 145). Thus the lasso includes a penalty term on the absolute value of the parameters to the model error, effectively forcing some of the model coefficients to zero and yielding a simpler model.

Since logistic regression is used as a linear predictive benchmark model, it is denoted by $\mathrm{BM}_{\mathrm{LR}}$.

3.3 Machine Learning Methods

3.3.1 K-Nearest Neighbors

K-NN (K-Nearest Neighbors) is among the simplest methods to fit a classification model. It is a nonparametric method based on distances between the observations in the feature space so that a new observation is classified according to the classes of the K nearest neighbors.

Suppose that the dataset $\mathcal{D}$ consist of N_k observations in class $\mathcal{C}_k$, so that $\sum_k N_k = N$, and that K_k points from class k are included in the K nearest neighbors of $\mathbf{x}$. Now the posterior probability of class k for $\mathbf{x}$ is

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{K_k}{K}.$$

Consequently, "fitting" a classifier via K-NN consists simply of storing the dataset. Calculations are made only when predicting with the classifier as the distances of the new observations to the stored data need to be calculated, in order to determine the K nearest neighbors. In practice, K-NN is quick to "fit" but requires substantial memory capacity for large data sets and can be slow to predict with.

While the K-NN has no parameters to fit, there are hyperparameters that can be optimized. Hyperparameters include the distance metric, the parameter

K and the method of obtaining the class of the new observation from the nearest neighbors' classes. Common distance metrics include the Euclidean distance and the Manhattan distance, whereas a usual choice to derive the class of a new observation is to choose the majority class of the K nearest neighbors.

The simple nonparametric specification of K-NN can be considered both an advantage as well as a drawback. It is very flexible, but computation is increasingly resource demanding as the dataset increases in size. In financial applications, however, data tends to be rather limited in quantity. Therefore, the K-NN is considered a suitable option for fitting nonlinear classifiers.

Bishop (2006, pp. 124 – 127) provides additional details on the K-NN, including the derivation of the class posterior probability for an observation $\mathbf{x}$.

3.3.2 Support Vector Machine

SVM is a method for fitting a linear classifier in a feature space defined by a transformation $\phi(\mathbf{x})$ (Cortes and Vapnik, 1995). The transformation can be nonlinear, making the SVM suitable also for fitting nonlinear models.

To present the principle behind the SVM, let us first consider a linear binary classifier

$$y(\mathbf{x}) = \mathrm{w}^T \phi(\mathbf{x}) + b, \tag{3.3}$$

where b is a bias term and $\phi(\mathbf{x})$ is a fixed transformation of $\mathbf{x}$. Moreover, assume that the training data $\mathcal{D}$ consists of observations $\mathbf{x}$ and corresponding target values τ, where $\tau_i \in \{-1, 1\}$. The classification of a new observations $\mathbf{x}$ is done according to the sign of $y(\mathbf{x})$.

The SVM fits a *hyperplane* in the feature space that separates the observations in two classes with a *maximum margin*, that is, maximizes the distance to the closest points. For now, assume that there exists a hyperplane, that perfectly separates the two classes in the feature space $\phi(\mathbf{x})$. This means that all observations of a given class are located on one side of the hyperplane with no observations of another class on the same side. and Now for some $\mathbf{w}$ and b there exists a hyperplane of the form (3.3) that satisfies $y(\mathbf{x}_i) > 0$ for all $\tau_i = 1$ and $y(\mathbf{x}_i) < 0$ for all $\tau_i = -1$. The two inequalities impose a requirement that all observations must be correctly classified. The margin is defined

by the perpendicular distance from $y(\mathbf{x}) = 0$, where $y(\mathbf{x})$ is of the form in (3.3), to the closest data points. It can be shown that the distance of an observation $\mathbf{x}$ to $y(\mathbf{x})$ is given by $|y(\mathbf{x})|/||\mathbf{w}||$. Requiring that all obseravations are correctly classified, the distance of an observation $\mathbf{x}$ to the hyperplane is given by

$$\frac{\tau_i y(\mathbf{x}_i)}{||\mathbf{w}||} = \frac{\tau_i(\mathrm{w}^T\phi(\mathbf{x}) + b)}{||\mathbf{w}||}.$$

The desired hyperplane is obtained by maximizing the minimum distance to the closest data point:

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \left\{ \frac{1}{||\mathbf{w}||} \min_i [\tau_i(\mathrm{w}^T\phi(\mathbf{x}) + b)] \right\}. \tag{3.4}$$

The problem (3.4) is complex to solve, but can be manipulated to a form for which an easy solution exists. This is accomplished by scaling the parameters $\mathbf{w}$ and b so that the closest point to the hyperplane satisfies

$$\tau_i(\mathrm{w}^T\phi(\mathbf{x}_i) + b) = 1,$$

and consequently all points satisfy the condition

$$\tau_i(\mathrm{w}^T\phi(\mathbf{x}_i) + b) \geq 1. \tag{3.5}$$

Furthermore, it can be shown that minimizing

$$\frac{1}{2}||\mathbf{w}||^2 \tag{3.6}$$

with respect to the constraints (3.5) is equivalent to solving (3.4). Solving (3.6) is a *quadratic optimization* problem, for which exist efficient solution algorithms. The resulting SVM perfectly separates the two classes in the original input space $\mathbf{x}$.

For cases when the classes overlap, a *soft margin* approach has been developed. Overlapping classes cannot be separated by the hyperplane, that is, some observation are located on the wrong side of the hyperplane. The soft margin allows some points to be misclassified by the hyperplane, but penalizes any points located within the margin or on the wrong side of the

hyperplane. This is accomplished by adding a *slack variable* $\xi_i \geq 0$ for each observation such that $\xi_i = 0$ for points that are on or inside the correct margin boundary, and $\xi_i = |\tau_i - y(\mathbf{x}_i)|$ for other points. The classification constraints in (3.5) can then be replaced with

$$\tau_i(\mathrm{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$$

to formulate the soft margin optimization problem to minimize

$$C\sum_i \xi_i + \frac{1}{2}||\mathbf{w}||^2. \tag{3.7}$$

A computationally efficient solution algorithm exists also for (3.7), alike for (3.6).

The freedom to select the function ϕ along with the soft margin approach make SVM already a flexible tool for classification. As described, the SVM falls into the category of discriminative classifiers. Probabilistic outputs can be obtained by fitting a logistic regression to the classifier outputs. Moreover, a common method to extend the binary SVM to a multiclass problem is the *one-versus-one* approach, where binary classifiers are estimated on all $C(C-1)/2$ pairs of classes, assuming C classes, and the output label is decided by a voting scheme among the binary outputs.

Bishop (2006, pp. 326 – 339) presents further details on SVM, where the reader can refer to for details.

Findings in related scientific literature indicate the SVM is a powerful tool even in application related to this thesis. For example Huerta et al. (2013) apply SVMs with a radial basis function to stock return prediction. They note that the use of the nonlinear radial basis function increases model performance, but is computationally more demanding. These findings motivate the selection of the radial basis function for the SVM. Promising findings in literature, along with the detailed specification of the method motivate the inclusion of the method in this thesis.

3.3.3 Random Forest

RF is an ensemble method consisting of *decision trees* (Breiman, 2001). It applies *bagging* to form a *committee* of the individual decision trees, as introduced by Breiman (1996).

Bagging, (bootstrap aggregation), generates several samples of the training data and trains a copy of the same *base learner* on each generated data set. The committee is a combination of base learners that are used as an ensemble model to produce predictions often significantly more accurate than a single base learner. Under certain assumptions with M base learners, it can be shown that the committee reduces the error by a factor of M when compared to the average error of a single base learner. A simple average of base learner predictions is a common voting scheme to form the ensemble prediction, which the RF also applies. In a case of decision trees as base learners, bagging results in an RF.

Classification and Regression Trees algorithm (CART) is a common variant of decision trees (Breiman et al., 1984). A CART tree consists of a sequence of *binary splits* along some variable in the feature space. For selecting this variable, a metric is needed to measure the "purity" of output classes with respect to the real training classes. Typical metrics for classification include *cross entropy* and *Gini-index*. The binary split is made by selecting a variable and a threshold, that result in the greatest increase in the "purity" of the model output, when splitting the observations along the selected variable at the selected threshold. This sequence of binary splits can be visualized as a tree graph, bringing intuition to the model name. Bishop (2006, pp. 663 – 666) provides further details on CART models.

Tree-based methods, including RF, are also present in related literature. For example, Imandoust and Bolandraftar (2014) apply methods including RF to forecasting the direction of stock market index movements. On the other hand, Lopez de Prado (2018, pp. 100 – 101) discusses the preference of bagging over boosting in financial applications, concluding that bagging would likely provide better results in such applications. The inclusion of RF is motivated by published related applications, as well as rationale behind the model specification implying it would be a suitable method for this thesis.

3.3.4 Gradient Boosting

GB is a specific application of a general *boosting* ensemble method (Friedman, 2001). Both the GB and boosting are applicable to several types of base learners, including decision trees. In boosting, the base learners are trained in sequence, each time reducing the ensemble model total error. Finally, a voting scheme is applied to the predictions of the base learners to produce the committee prediction. Bishop (2006, pp. 657 – 663) introduces the general *boosting* framework with further details.

Friedman (2001) describes the specific GB algorithm in detail, including the sequential training, for different types of base learners such as decision trees. The GB exploits the negative gradient of the error function in the sequential training process. The algorithm begins with calculating the initial base learner, then performing iterations including the calculation of the negative gradient and generalizing it in order to perform and update step in the model. Because of the exploitation of the negative gradient when performing the updates, GB can be called a *gradient descent* algorithm. The update step in GB is analogous to adding new weak learners to the ensemble.

A boosting method is also included, even though the selection of RF is based on the rationale that bagging methods could be a better option for financial applications. The inclusion of GB as a boosting method is justified as a test to the hypothesized preference of bagging methods.

3.4 Comparison of Classifier Performance

The comparison of classifier performance is based on the output labels. The metrics applied are *precision* and *recall*, which are summarized to an F_1 score. Sokolova and Lapalme (2009) provide definitions for these metrics.

The metrics are calculated with per class *true positive*, *true negative*, *false positive* and *false negative* counts, denoted respectively by tp, tn, fp, and fn. For each class k, true positive counts the observations, for which the predicted label is k and agrees with the true label t. True negative, on the other hand, counts the cases where $y \neq k$ and the true label $t \neq k$. False positive counts the cases where $y = k$, but $t \neq k$, whereas false negative counts the cases where $y \neq k$, but $t = k$.

Precision measures the degree to which the output labels for class k agree with the true labels for those observations. Moreover, we measure precision as the simple average of per class precisions, given by

$$\text{Precision} = \left(\sum_{i=1}^{K} \frac{tp_i}{tp_i + fp_i} \right) / K. \tag{3.8}$$

Recall measures the proportions of observations in class k that were detected by the classifier. As precision in (3.8), recall is also measured as an average over the K classes, given by

$$\text{Recall} = \left(\sum_{i=1}^{K} \frac{tp_i}{tp_i + fn_i} \right) / K. \tag{3.9}$$

F_β score summarizes (3.8) and (3.9) as a harmonic average, given by

$$F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}, \tag{3.10}$$

where β is a chosen coefficient. We select $\beta = 1$ in order to give precision and recall an equal weight in the score, and denote the resulting measure as the F_1 score.

A beneficial property of the selected measures is that they are robust to an uneven class distribution. Moreover, precision and recall have a clear interpretation with relevance to the specific task. Precision could be viewed as the most important metric in terms of forming portfolios based on the classifier output, since a high precision implies that a high proportion of correctly classified shares are included in the portfolio. A high recall, on the other hand, implies that the classifier notices most of the shares in a given class. While a high recall is also a desirable property for a classifier, we can deduct that an investor would likely prefer a classifier with a high precision over one with a high recall. For example, assume a binary classification to high and low returns and an investor constructing a portfolio of the shares with high predicted returns. An investor would likely prefer the classifier that provides portfolios with a greater proportion of correctly classified shares, corresponding to a greater precision, while being relatively indifferent to the classifier failing to detect some shares with high returns.

3.5 Construction and Comparison of Model Based Portfolios

3.5.1 Portfolio Construction

While classification accuracy provides a measure of the classifier performance, it fails to capture the economic impact of the predictions, which is important when reflecting to the motivation of producing portfolios with higher returns. The impact can be assessed by constructing model based portfolios from

the classifier predictions and inspecting the performance of these portfolios. The focus is on *long-short* portfolios constructed according to the classifier predictions.

A long-short portfolio is constructed by combining *long* and *short* positions (Luenberger, 1998, pp. 137 – 141). A long position means simply buying and holding the share. If the share is bought at time 0 at price X_0 and held until time t, when it is sold at a price X_t, the return of the position would be

$$r = \frac{X_t - X_0}{X_0}.$$

Note, that in practice the prices must be adjusted for corporate events such as dividend payouts and changes in the number of shares outstanding. Now r is positive, in case the share price increases, and negative otherwise. A short position means borrowing, for example from a brokerage firm, and then selling the share at time 0 for a price x_0. At time t the share is bought back at a price X_t and returned to the lender. The investor now receives X_0 at time 0 and pays X_t at time t. Therefore, the return of the short position is

$$r = \frac{X_0 - X_t}{X_0},$$

which is positive in case the share price declines. In practice, short positions include an interest cost, because borrowing is involved.

Portfolio return is a function of the individual share *weights* and returns. The weight w_i of a share i in a portfolio is simply its fraction of the portfolio value, so that $\sum_i w_i = 1$. The realized portfolio return is given by

$$r = \sum_i w_i r_i,$$

where r_i is the realized return of share i. As shown by Luenberger (1998, pp. 14 – 15) for interest, returns alike can be compounded to obtain the return from time t to $t + k$:

$$1 + r_{t,t+k} = \prod_{i=t}^{t+k} (1 + r_i). \tag{3.11}$$

In this thesis, long-short portfolios are constructed based on classifier prediction according to the following principle. Long positions include only the shares with a predicted class corresponding to high returns, whereas the short positions include shares with a predicted class corresponding to low returns. In addition to selecting the shares, a weighting also needs to be determined. Typical weighting schemes include market equity weighting and equal weights. Market equity weighting has the advantage of limiting the disproportionate impact of shares with a very small market equity, but a similar effect can be achieved by excluding the smallest companies when using equal weights, as noted by Hou et al. (2017).

Controlling the number of shares in a portfolio is accomplished by exploiting the output class posterior probabilities. A predetermined number of shares with the highest posterior probability of class k is selected to the portfolio with a position corresponding to the class k. This approach, however, holds in a problem where it is possible for a share to be selected both to the long and short portfolio. This problem is mitigated by imposing a condition that the share can be selected only once according to the class with the greatest posterior probability.

3.5.2 Portfolio Performance Metrics

Portfolio performance is measured in terms of returns, risk, risk adjusted returns, and *turnover*. Return r is measured simply as the rate given by (3.11), and risk as the standard deviation of the returns, often referred to as *volatility* and denoted by σ. The interest in these quantities in finance stems from the portfolio theory by Markowitz (1952), summarized for example by Luenberger (1998, Chapter 6).

An investor's attitude towards the combination of risk and return is reflected by their individual risk preference (Luenberger, 1998). In general, an investor valuates a combination of risk and return when comparing different investments: an investment with a higher expected return may not be preferred in case it is too risky, whereas the less risky out of two alternatives with the same expected return would be chosen. This is the intuition behind measuring risk adjusted returns, which is measured by the *Sharpe Ratio*, denoted by SR. Luenberger (1998) defines the Sharpe Ratio of a portfolio as

$$SR = \frac{\mathbb{E}(r) - r_f}{\sigma}, \tag{3.12}$$

where $\mathbb{E}(r)$ estimates the true expected return as the simple mean of the individual share returns. In this thesis, the definition in (3.12) is modified to omit the risk-free rate. Consequently, σ is the volatility of the portfolio return. The SR applied in this thesis is given by

$$SR = \frac{\mathbb{E}(r)}{\sigma}.$$

In addition to the risk-return characteristics, the turnover is of interest since there are no constraints on how much the model based portfolio differ between consecutive months, likely leading to a high turnover. In practice, a portfolio with a high turnover is subject to high transaction costs, although the cost varies greatly by between investors and trades, see Carhart (1997) for discussion. In this thesis, turnover is measured as the sum of the absolute changes in the portfolio constituents not attributable to returns. Therefore turnover measures the implied trading activity of the portfolio over time. For a portfolio of N shares, the turnover at time t is given by

$$\mathrm{TO}_t = \sum_{i=1}^{N} |w_{i,t-1} r_{i,t-1} - w_{i,t}|.$$

Return, volatility, risk adjusted returns and turnover provide insight on the comparative performance of the model based portfolios. In addition to simply comparing these metrics, the model based portfolio returns are examined with the standard FF5 model. This is accomplished by fitting the linear regression model in (2.4) with standard FF5 factor returns on the model based portfolio returns.

The regression fit is evaluated by the following characteristics: the regression coefficients, their statistical significance and the *coefficient of multiple correlation*, denoted by R^2. Examining the coefficients dissects the model based portfolio returns in term of exposure to the standard FF5 factors. The significance is related to the probability of finding a coefficient as great or greater in the regression, in case the real coefficient is equal to 0. R^2, on the other hand, measures to the degree of variation explained in the model based portfolio returns. Allison (1999) provides details on the significance of the coefficients, and the R^2.

As one of the benchmark models specific to this thesis, $\mathrm{BM}_{\mathrm{FF5}}$, is based on a replication of the FF5 factors, the differences between the replication and the standard FF5 factors are discussed in Section 5.1.1.

Chapter 4

Data and the Selection of Factors

This chapter describes the following data specific matters: the source of the examined return and fundamental data, combining them, selecting the included shares, return labeling for classification as well as the selection and definitions of the applied factors.

4.1 Database

The data used in this thesis is obtained from the Center for Research in Security Prices (CRSP) from CRSP (2019) and Compustat by Standard and Poor's datasets via the CRSP/Compustat Merged (CCM) database from CRSP/Compustat (2018). The data applied covers the period starting from January 1988 to December 2016. The used CRSP dataset includes mainly the security monthly prices and returns, whereas the used Compustat data includes the quarterly company fundamental data. It should be noted, that the coverage of the datasets by the two providers differs and that they use different proprietary company and share identification regimes. The CCM database includes a linking table from CRSP identifiers to Compustat identifiers, making it convenient to combine data by the two providers.

To support the experiments made with the constructed database, the data available at the database of French (2019) is used to validate the constructed database. Moreover the database of French (2019) includes the standard FF5 factor returns.

4.2 Combining Price and Fundamental Data

To coherently combine return and fundamental data, different fundamental data reporting regimes and a reporting delay must be accounted for. US listed companies report their accounting data mostly on a quarterly frequency, but the reporting period is not necessarily aligned with the calendar year. The fiscal year can be defined, for example, to begin at July 1st of year t and end at June 30th of year $t + 1$. Therefore, for companies that report their accounting data with respect to a fiscal year ending in any other month than December, the fiscal year and fiscal year quarter end dates are converted to calendar year dates.

In addition to the different reporting regimes, a reporting delay must also be taken into account. Easton and Zmijewski (1993) discuss and examine the reporting delays for US companies. They usually do not report their accounting data for a given period at the end period end date, resulting in a reporting delay. Therefore, the information cannot be immediately reflected in the share price, since it is not yet public to the market. The length of the reporting delay can vary, but in this thesis a 3 month delay is applied to ensure that the fundamental data is reflected on the share prices. As discussed by Asness and Frazzini (2013) while, the literature standard delay of six months using annual data is a conservative and justified choice, it is not optimal. The findings of Easton and Zmijewski (1993) and the conclusion of Asness and Frazzini (2013) motivate the selection of a shorter than standard delay in this thesis.

4.3 Included Shares

The included shares are restricted only to common shares of US companies publicly traded in NYSE, American Stock Exchange (AMEX) and Nasdaq. Companies with less than two years, or eight quarters, of data in Compustat are excluded to mitigate a survivorship bias. Early examples discussing the bias include the work by Kothari et al. (1995). Furthermore, only the shares found in both CRSP and Compustat data are included.

The impact of microcaps, shares of companies with a very small relative market equity (ME), is mitigated by excluding them using monthly ME breakpoints. As Fama and French (2008) point out, microcaps constitute only approximately 3% of the combined NYSE-AMEX-Nasdaq ME, but make up

for 60% of the number of shares. Following their definition of microcaps as companies with ME below the 20th percentile NYSE ME, such companies are excluded from each monthly share universe. The universe for month t is constructed using the market equities and breakpoints from month $t-1$. ME is calculated as the number of shares outstanding times the share price. Figure 4.1 presents the 20th percentile NYSE breakpoint over time for the included shares in comparison to the breakpoints available at the database of French (2019).

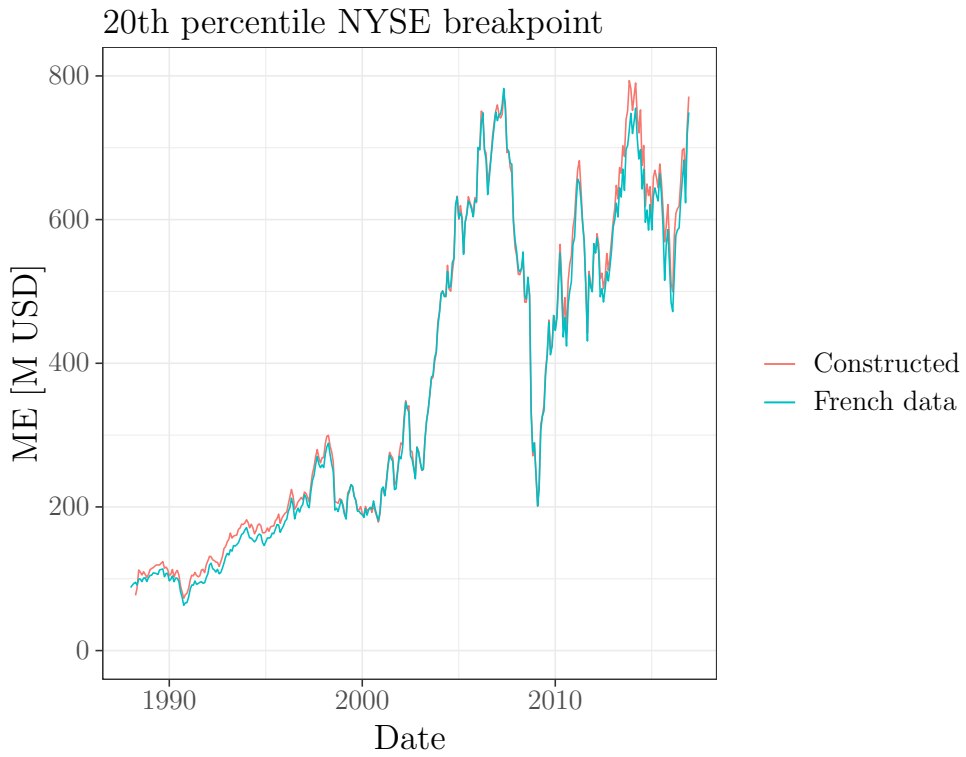

Figure 4.1: The breakpoints calculated from the constructed dataset are very close to the breakpoints available at the database of French (2019), indicating that similar sets of shares are included.

Furthermore, financial companies are excluded each month from the share universe, because of their differing nature of fundamental data. Moreover firms with a negative BE (book equity) value or missing ME are excluded. BE is measured as shareholders' equity minus the book value of preferred stock. The counts of total shares as well as the counts of shares included after excluding financial companies and applying the market value breakpoints are presented in Figure 4.2. The total count of shares exhibits a decreasing trend,

while the count of included shares is fairly stable, leading to an increasing trend in the fraction of shares included in the analysis. This is possible, because the ME breakpoint is calculated based on only NYSE shares.

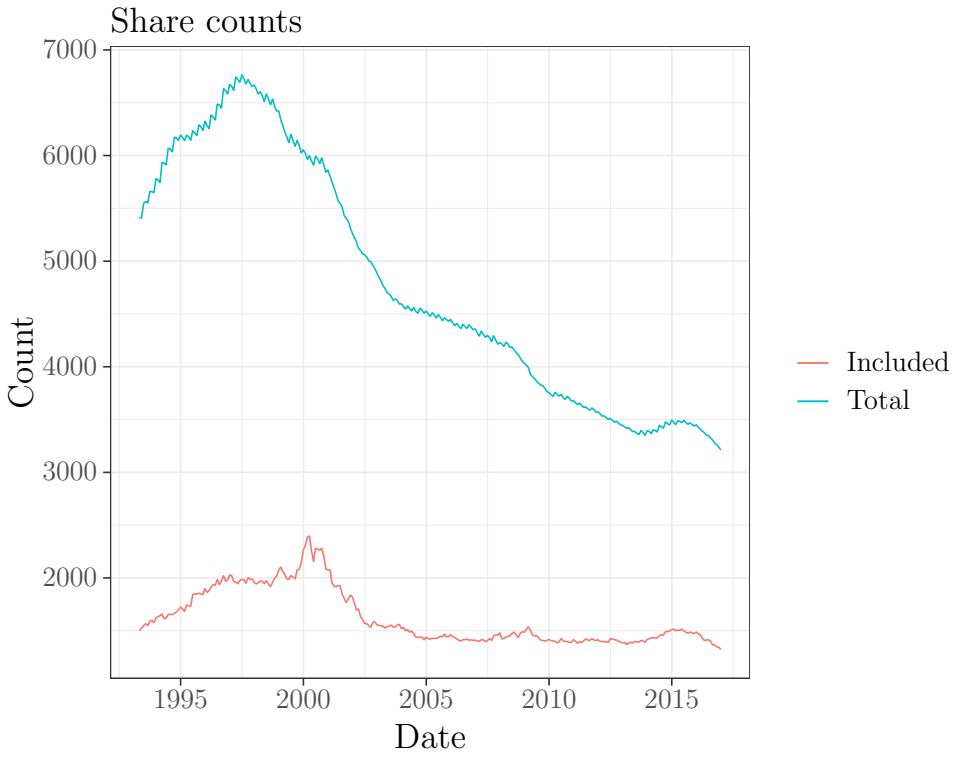

Figure 4.2: The total number of shares exhibits a decreasing trend, while the number of shares included after excluding financial companies and applying the market value breakpoints is relatively stable. This results in an increasing trend in the fraction of included shares.

4.4 Return Labeling

Since we are predicting future returns by classification, we transform the $t+1$ return of each observation to a class label using rank percentiles. Note that this implies the use of historical data, so that the realized data for t and $t+1$ are available. Let l and u be the percentile breakpoints for labeling. The $t+1$ future returns are labeled within monthly cross-sections as follows:

$$\begin{cases} 1, \text{ if rank percentile } \geq u \\ 0, \text{ if } l \geq \text{rank percentile} < u \\ -1, \text{ if rank percentile } < l \end{cases} \tag{4.1}$$

The simplest way to label the returns is binary labeling, for example using the median as the breakpoint. Then we would simply select $l = u$ in (4.1), leaving the middle class empty. On other hand, applying three-class labeling allows excluding the middle class from the portfolios. This focuses the long and short positions to, respectively, the shares with the highest and lowest expected returns, which could yield better portfolio performance.

4.5 Factor Selection and Scaling

The selection of the factors is an important step in constructing a multifactor model. When applying a factor model to the problem of stock return prediction, two possible approaches for selecting the factors can be considered. The first is to apply a data mining method to extract the factors from the data, and the other is to gather a set of documented factors based on empirical finance. While both can be justified, the factors are summarized from literature in this thesis to maintain a connection to economic and financial theories.

The selection is based on the replication work done by Hou et al. (2017). They replicate a large number of documented anomaly variables and test for the significance of the difference in returns between the highest and the lowest decile according to each variable. Their nonparametric approach is robust, but applying only the differential of the two extreme deciles fails to reveal any possible nonlinearity in the returns. This can, however, be interpreted as a choice of variables favorable to linear models, creating a rather pessimistic case in terms of the performance of the nonlinear models. Referring to the results of Hou et al. (2017), the variables are screened by excluding complex variables and variables related to analyst forecasts. Consequently variables with a reported t-statistic of at least three are chosen. However, in this thesis all selected variables are computed with quarterly data, differing from Hou et al. (2017) using both quarterly and annual variables. Moreover, we apply a one year sum of the quarterly income statement items, such as revenues. The sum mitigates the quarterly seasonal variation commonly found in company income items. This results in a set of 30 factors applied in this thesis. From

this set, we identify a subset of 5 factors that are applied in the FF5 model, as described in Section 3.2.1.

Table 4.1 presents the applied factors along with their accronyms. The factors are grouped into six categories: momentum, value-versus-growth, investment, profitability, intangibles and trading frictions factors.

Table 4.1: Factors used in this thesis. They are gathered from the replication work by (Hou et al., 2017). Factors marked with an asterisk (*) belong to the identified set of 5 factors applied in the FF5 model.

Factor	**Accronym**	**Category**
6-month prior return	R^6	Momentum
11-month prior return	R^{11}	Momentum
Book-to-market*	Bm	Value-vs-growth
Earnings-to-price	Ep	Value-vs-growth
Enterprise multiple	Em	Value-vs-growth
Cash flow-to-price	Cp	Value-vs-growth
Operating cash flow-to-price	Ocp	Value-vs-growth
Investment-to-assets*	Ia	Investment
Changes in gross property, plants and equipment and inventory-to-assets	dPia	Investment
Changes in net operating assets	dNoa	Investment
Net stock issues	Nsi	Investment
Composite equity issuance	Cei	Investment
Inventory changes	Ivc	Investment
Changes in net non-cash working capital	dWc	Investment
Change in net non-current operating assets	dNco	Investment
Change in non-current operating assets	dNca	Investment
Change in financial liabilities	dFnl	Investment
Return on equity	Roe	Profitability
Changes in return on equity	dRoe	Profitability
Changes in return on assets	dRoa	Profitability
Gross profits-to-lagged assets	Gla	Profitability
Operating profits-to-equity*	Ope	Profitability
Operating profits-to-lagged equity	Ole	Profitability
Operating profits-to-lagged assets	Ola	Profitability
Cash-based operating profits-to-lagged assets	Cla	Profitability
12 month lagged return	R_a^1	Intangibles
Years 2-5 lagged returns	$R_a^{[2,5]}$	Intangibles
Beta*	Beta	Intangibles
Asset liquidity	Alm	Intangibles
Market equity*	ME	Trading frictions

The factor values for each month t are normalized by a percentile transformation to restrict the values of each factor score to a fixed interval and to control outlier observations. One should notice, that the normalization is already a nonlinear transformation of the data, but it preserves the order of the observations.

The missing observation percentages for the calculated factors over all observations are presented in Table A.1, Appendix A. In general the factor coverage is satisfactory, ranging from no missing items for Market equity, to 47.76% missing for Changes in gross property, plants and equipment and inventory-to-assets.

Chapter 5

Experiments and Results

Experiments are conducted to determine the feasibility of nonlinear multifactor classifiers for predicting share returns. This chapter describes the experiments that apply the methods introduced in Chapter 3 on the data described in Chapter 4, as well as how the methods are compared to each other. The experiment results are presented and evaluated both in terms of classification performance and model based portfolio performance. Moreover, the experiments motivate a sensitivity analysis, which is described and for which the results are presented following the experiment results.

5.1 Implementation of the Experiments

5.1.1 Modeling and Portfolio Parameters

The experiments cover the period of 12 years from January 2004 to December 2016. While the data starts already from 1988, the first 16 year are only used to construct the factors, the reporting delay and the modeling lookback period. The classifiers are fit monthly in a *rolling window* approach. It means the classifiers are fitted at month t on normalized factor scores available at month $t-1$ to predict month t return labels. Returns for $t+1$ are predicted with the fitted model using the normalized scores from month t. The prediction is repeated for all months. The lookback period determines how many months of historical panel data is used per each month to fit the models. The experiments are made with loockback periods of 36 and 120 months. Any missing observations in factor scores in the constructed data

set are replaced by the median for fitting the predictive models. Classification performance measurement is based on the performance of each classifier to predict $t + 1$ return labels with t data. This means that the performance is always measured using data not used in model training. Moreover, the predictive classifiers are always fitted with the same hyperparameters, that is, no hyperparameter optimization is conducted.

The applied hyperparameter values are the default values in the computational implementation packages described in Section 5.1.2. Only the K for K-NN has to be determined, since it lacks a default value. An odd K is preferred to prevent even votes in binary classification, but otherwise the selection of K without model selection methods is rather arbitrary and $K = 49$ was selected. This selection reasonable as it considers a great number of neighboring observations, yet it is relatively small in comparison to the monthly counts of included shares, which are illustrated in Figure 4.2. This approach was selected due to practical limitations in thesis scoping and computational resources. For $\mathrm{BM_{LR}}$, the lasso approach is applied in experiments, where the whole set of 30 factors is considered. Furthermore, to make computation times feasible, random training data downsampling is applied to restrict the sample to a maximum of 5000 for SVM and RF and to a maximum of 50000 for GB.

The portfolios based on the predictive classifiers are formed by selecting shares with a high (low) predicted return, label 1 (-1), for long (short) positions. Class posterior probabilities provided by the predictive models are exploited in portfolio construction by selecting the top 200 shares to the position corresponding to the class. The shares are equal weighted within long (short) positions, meaning an equal amount is invested in the long and short positions. Moreover the total long and short positions are given equal weights in the long-short portfolio. Alike the deterministic benchmark portfolios, the classifier based portfolios are updated each month.

The deterministic benchmark models do not require fitting alike the predictive models and are not used for prediction, but only to construct benchmark portfolios. $\mathrm{BM_{FF5}}$ is constructed by replicating the methodology presented in Section (3.2.1) with two differences from the standard method: the replicated factor mimicking portfolios are updated monthly using the constructed data set, and the breakpoints are calculated from the whole monthly sets of included shares as opposed to updating the portfolios annually using NYSE breakpoints as in the standard methodology of Fama and French (2015). The original methodology does not apply the 20^{th} percentile NYSE ME breakpoints on the monthly sets of included shares, as done in this thesis. There-

fore, we opt to use the breakpoints of the whole universe in the replicated factor mimicking portfolio construction. This difference in the methodology is important, because the standard FF5 factors are applied in the analysis of model based portfolio returns with the FF5. This means the FF5 model in (2.4) is fitted to explain the model based portfolio returns with the standard FF5 factor returns, available at the database of French (2019). Since the modeling yields monthly predictions and monthly returns, the analysis of the model based portfolio returns with the FF5 model is also conducted with monthly returns. Figures B.1 and B.2 in Appendix B illustrate the monthly returns of the replicated FF5 factors and their standard counterparts.

Four experiments are conducted with the models and the resulting portfolios by changing one modeling parameter at a time: the lookback period; the number of factors; and the return labeling breakpoints. The experiments and the varied parameter values are presented in Table 5.1. They provide an overview of model performance in terms of accuracy and portfolio performance, as well as examining the sensitivity of the results with respect to different modeling parameters. However, the experiments are fairly different in terms of the parameters. The arguably scarce set of experiments was selected to compromise between computational requirements and coverage of different modeling parameters.

Table 5.1: Modeling parameters in each experiment. The number of factors refers to the use of either the whole set of 30 factors or the subset of 5 factors as described in Section 4.5. Return breakpoints state the percentiles applied in return labeling. If a single number is given, there are only two classes for return labeling. Panel data length describes the lookback period of data included in model training each month, while portfolio weighting states the method of weighting individual shares in a portfolio.

Experiment	1	2	3	4
Number of factors	30	5	30	30
Panel length [months]	120	120	36	120
Breakpoints [%]	30;70	30;70	30;70	50

Experiment 1 applies all 30 factors with a long training data lookback period and three-class labeling. The second experiment is identical to the first, except applying only the subset of five factors. The predictive models in Experiment 1 are expected to outperform their counterparts in Experiment 2 due to the greater number of factor applied in the first experiment. Exper-

iment 2 is included to test, how the predictive models perform applying the same factor information as the FF5 model. Experiments 3 and 4 continue with the whole set of 30 factors, where the third experiment applies a short training data lookback, and the fourth experiment tests a binary labeling. Experiment 1 is expected to yield better results than Experiment 3 because of the longer training data lookback in the first experiment. Similarly, the binary labeling in Experiment 4 is could yield weaker results to the otherwise identical Experiment 1, provided that the predictive models exploit the additional information of the three-class labeling in Experiment 1.

A baseline value for the classifier performance metrics precision, recall and F_1 can be deducted from the return labeling breakpoints in Table 5.1, since they determine the return class prior probabilities. A "predictive" model simply reflecting the class prior distribution can be defined as randomly assigning the class labels according to their prior distribution. The distributions of the true and "predicted" labels are independent. Applying this random model, a randomly selected observation belongs to class $\mathcal{C}_k$ with probability $p(\mathcal{C}_k)$. This observation is labeled with k with probability $p(\mathcal{C}_k)$, and in with some other label with probability $1 - p(\mathcal{C}_k)$, leading respectively either to a true positive result with probability $p(\mathcal{C}_k)p(\mathcal{C}_k)$, or a false negative result with probability $p(\mathcal{C}_k)(1 - p(\mathcal{C}_k))$. On the other hand, the randomly selected observation belongs to class $h \neq k$ with probability $1 - p(\mathcal{C}_k)$ and is labeled in class k with probability $p(\mathcal{C}_k)$, leading to a false positive result with probability $(1 - p(\mathcal{C}_k))p(\mathcal{C}_k)$. Moreover, we notice that a false positive has the same probability as a false negative. This deduction can be summarized to estimate the true positive, false positive and false negative rates in terms of the class prior probabilities as

$$\begin{aligned} tp_k &= p(\mathcal{C}_k)p(\mathcal{C}_k), \\ fp_k &= (1 - p(\mathcal{C}_k))p(\mathcal{C}_k) = fn_k. \end{aligned} \tag{5.1}$$

This can be applied to precision in (3.8), recall in (3.9) and the F_1 score in (3.10) in order to determine the threshold values for these metrics. Intuitively, any predictive model in required to exceed this value as a sign of providing predictive power.

In Experiments 1 to 3 the return labeling yields class prior probabilities of 0.3, 0.4 and 0.3 for classes 1, 0 and -1, respectively. The threshold values are obtained by placing the true positive, false positive and false negative rates in (5.1) in the metric definitions as follows:

$$\begin{aligned}
\text{Precision}_{BL} &= \left(\frac{0.3^2}{0.3^2 + 0.3 \cdot 0.7} + \frac{0.4^2}{0.4^2 + 0.4 \cdot 0.6} + \frac{0.3^2}{0.3^2 + 0.3 \cdot 0.7} \right) /3 \\
&= (0.3 + 0.4 + 0.3)/3 = \frac{1}{3}, \\
\text{Recall}_{BL} &= \text{Precision}_{BL}, \\
F_{1,BL} &= \frac{(1+1)(1/3)(1/3)}{1/3 + 1/3} = \frac{2/9}{2/3} = \frac{1}{3},
\end{aligned} \tag{5.2}$$

that is, a value of 1/3 for all of the three metrics. In Experiment 4, the prior probabilities are $(0.5, 0.5)$, and all metrics are 1/2 as a result of a similar calculation as in (5.2).

5.1.2 R Implementation

R is used to carry out the experiments mainly due to its ease of use and availability of high-level packages for statistical analysis and modeling. Table 5.2 lists the most important R packages used for data handling and implementing the methods.

Table 5.2: The R packages used in the experiments containing the predictive model implementations.

Package	Description
`glmnet`	LASSO Logistic Regression (Friedman et al., 2018)
`caret`	K-NN with posterior probabilities (Kuhn, 2018)
`e1071`	Implementation of SVMs (Meyer et al., 2019)
`gbm`	Implementaion of GB (Greenwell et al., 2019)
`randomForest`	Implementation of RF (Breiman et al., 2018)

The computations for the experiments are performed with a standard laptop computer and all methods are required to be computed within reasonable time. While this imposes restrictions on the modeling methods, the requirements are justified not only from a practical point of view for conducting the research, but also because computationally extremely heavy methods would be more demanding to apply in the industry. In this sense, methods feasible with standard equipment and reasonable time are in the focus.

5.2 Results

The results of the four experiments, including the deterministic benchmarks, are presented in Table 5.3. The results are expressed in terms of the F_1 score, precision and recall for classifier performance. Note that the classifier performance metrics are not applicable to the deterministic benchmarks, which are not used for classification. Classifier based portfolios are inspected in terms of accumulated returns, volatility, Sharpe Ratio and turnover.

Table 5.3: Experiment results. A bold typeface indicates the preferred value for the column metric per experiment. The F_1 score, precision and recall are all medians of monthly values over the time period. For interpretation, note that Experiment 4 uses a binary return labeling with a median breakpoint, as opposed to the other experiments applying a three-label regime with 30^{th} and 70^{th} percentile breakpoints. SVM F_1 score and precision are not defined in Experiment 1 for two occurrences, in Experiment 2 for two occurrences, in Experiment 3 for one occurrence, and in Experiment 4 for 13 occurrences due to SVM predicting no observations to some of the classes. These occurrences are omitted from the SVM precision and F_1 score. r is the cumulative return, σ the sample volatility and TO the mean of monthly turnovers.

Model	F_1	Precision	Recall	r	σ	SR	TO
Benchmark							
$\text{BM}_{\text{simple}}$				-0.29	**1.98**	0.00	25.49
BM_{FF5}				**33.10**	4.87	**0.48**	**14.83**
Experiment 1							
BM_{LR}	0.3709	0.3746	0.3677	1.77	4.58	0.05	**36.69**
K-NN	0.3704	0.3745	0.3684	3.04	**2.99**	0.09	59.81
SVM	0.3316	0.3321	0.3301	-25.76	5.22	-0.42	87.47
GB	**0.3815**	**0.3868**	**0.3780**	6.56	5.31	0.12	51.99
RF	0.3770	0.3789	0.3757	**21.50**	3.09	**0.50**	72.68
Experiment 2							
BM_{LR}	0.3647	0.3707	0.3599	**12.79**	5.77	0.19	**28.29**
K-NN	0.3677	0.3677	**0.3671**	7.99	2.87	**0.22**	52.44
SVM	0.3339	0.3383	0.3337	-9.30	5.72	-0.10	87.42
GB	**0.3699**	**0.3756**	0.3650	0.20	4.18	0.02	47.58
RF	0.3587	0.3589	0.3588	-1.53	**2.37**	-0.04	80.64
Experiment 3							
BM_{LR}	0.3722	0.3765	0.3696	1.93	4.06	0.06	**37.02**
K-NN	0.3706	0.3733	0.3671	-8.81	3.36	-0.20	58.49
SVM	0.3402	0.3432	0.3400	-12.09	6.44	-0.12	87.54
GB	**0.3856**	**0.3906**	**0.3785**	**24.89**	4.58	**0.40**	44.72
RF	0.3751	0.3757	0.3735	12.12	**3.36**	0.28	74.01
Experiment 4							
BM_{LR}	0.5021	0.5021	0.5021	**35.23**	5.44	**0.46**	**43.69**
K-NN	0.5013	0.5013	0.5012	3.22	**2.77**	0.10	69.42
SVM	0.4998	0.4997	0.5000	8.57	4.44	0.17	91.01
GB	0.5047	0.5048	0.5046	27.77	5.20	0.39	58.67
RF	**0.5053**	**0.5053**	**0.5053**	-4.32	3.33	-0.09	81.60

Section 5.2.1 examines the results in Table 5.3 in terms of classifier performance, and Section 5.2.2 examines them in terms of model based portfolio performance.

5.2.1 Classifier Performance

A few main findings and general patterns are found in the classifier performance metrics in Table 5.3. The differences between respective precisions and recalls are small. Therefore we focus the analysis of classification performance on the F_1 score. All classifiers except the SVM constantly provide a mean F_1 metric above the class prior probability based threshold of $1/3 \approx 0.3333$, as determined in Section 5.1.1, in Experiments 1 to 3. In Experiment 4 the corresponding threshold is $1/2 = 0.5000$, which all classifiers except the SVM achieve. However, the margins to the thresholds are small. The highest F_1 score is 0.3856 for GB in Experiment 3 out of Experiments 1 to 3 with comparable labeling classification metrics. In Experiment 4 the highest score is 0.5053 for RF. The lowest F_1 scores are produced by SVM: 0.3316 in Experiment 1 and 0.4998 in Experiment 4, the only instances where the thresholds are not met. In general the predictive classifiers seem to have at least some predictive power. Evidence is stronger in Experiments 1 to 3, where the margins to the corresponding threshold are greater than in Experiment 4.

A general pattern over all experiments appears that SVM has the lowest classifying performance on all metrics without exception. In addition the SVM classification metrics are very close and even below to the prior probability based thresholds. These findings indicate that the SVM provides little, if any, predictive power. This is likely to be related to the heavy downsampling of the SVM training data sample, or the implementation default hyperparameters poorly suited to this specific task. GB, on the other hand, is the strongest classifier on all metrics in Experiments 1 and 3, and the strongest in Experiment 2 based on the F_1 score and precision. This indicates that GB is the most powerful predictive classifier in this set of experiments. RF is the strongest classifier in Experiment 4 on all three metrics. In fact, RF provides fairly high classification performance, especially considering it is subject to the same heavily downsampled training data sample as SVM. This could be explained with the ensemble structure of RF, where the bagging procedure mitigates the impact of the training data downsampling on the classification performance.

Comparing the Experiments 1 to 3 to one another, Experiment 3 yielded

overall the highest classifier performance metrics, whereas Experiment 2 in general the lowest. Recall that Experiments 1 and 3 are identical, except for the training data lookback period. The longer lookback in Experiment 1 surprisingly produces weaker classification metrics than the shorter period in Experiment 3. Possible causes are pure coincidence and that the longer lookback contains historical events not reflected in the returns anymore. A suspect cause for the relatively poor performance in Experiment 2 is the smaller set of factors used, implying that the classifiers are able to exploit additional information from the more extensive set of factors applied in the other experiments. A direct comparison cannot be made between Experiments 1 to 3 with Experiment 4 because of the differing return labeling. However, it seems that the labeling in Experiment 4 is more difficult for the classifiers, because the classifier metric margins to the thresholds are relatively small in Experiment 4, as noted earlier.

Fitting and predicting with the methods is fairly quick. This is affected by the downsampling of SVM and RF training data. A monthly cycle of fitting the classifiers and predicting with them takes under two minutes with the specified implementation.

5.2.2 Portfolio Performance

Portfolio performance measures return, volatility and Sharpe Ratio seem to display less obvious patterns than the classification performance measures. The overall highest return is 35.23% generated by $\mathrm{BM}_{\mathrm{LR}}$ in Experiment 4 and the lowest -25.76% by SVM in Experiment 1. The benchmark $\mathrm{BM}_{\mathrm{simple}}$ yields poor returns of -0.29%, while $\mathrm{BM}_{\mathrm{FF5}}$ provides high returns of 33.10%. Within the experiments, the highest return is attributed to $\mathrm{BM}_{\mathrm{LR}}$ in two experiments, and to GB and RF in one pear each model. SVM provides the lowest return in all experiments, except Experiment 4.

Interestingly, Experiment 4 produces relatively high returns, even though the classification performance is barely above the threshold of predictive power. The median breakpoint labeling unique to Experiment 4 is likely the cause, even though it was expected to be more difficult to the classifiers than the three-class labeling in the other experiments. A possible explanation for the high returns in Experiment 4 lies in the model fitting methodology, which considers all misclassifications equally bad. In the binary classification of Experiment 4 it makes no difference, but likely has an effect on the other experiments. Intuitively, it is a worse error to misclassify a true 1 as a -1, than it is to misclassify it as a 0. The model fitting does not account this

"order" of the classes, which could explain the higher returns in Experiment 4. Missing the "order" is a characteristic of simple classification methods applied in this thesis, where the classifier output is of categorical rather than of ordinal type.

Out of the predictive models, RF produced the globally lowest volatility of 2.37% in Experiment 2, while the highest volatility of 6.44% is measured for SVM in Experiment 3. The least volatile portfolios are provided by RF in three experiments, while K-NN provides it in one. On the other hand, BM_{LR} produces the most volatile portfolios in two experiments.

The Sharpe Ratios are in general low and in some cases even negative. The BM_{simple} SR is zero, whereas BM_{FF5} yields a relatively high SR of 0.48. The highest SR in all experiments is 0.50 produced by RF in Experiment 4, while the lowest attributes to SVM at -0.42 in Experiment 1. Within experiments, the highest SR is produced along with the highest return in three experiments. BM_{LR}, K-NN, GB and RF all provide it in one of the experiments. Comparing all predictive models to BM_{FF5} in terms of SR, only RF in Experiment 1 manages to beat the it, while BM_{LR} in Experiment 4 and GB in Experiments 3 and 4 come close.

Differing from the other metrics, turnover exhibits fairly strong patterns. First, however, we notice that BM_{FF5} has the lowest average monthly turnover of 14.83% with a great margin. All other portfolios have considerably greater turnovers, ranging from 25.49% for BM_{simple} to 91.01% for SVM in Experiment 4. The highest turnovers for all models are produced in Experiment 4. Out of the models, SVM generates the highest turnover in all experiments, while BM_{LR} the lowest respectively. The significant differences in turnover have severe implications in practical applications. A high turnover generates high transaction costs, which are covered from portfolio returns. Taking the portfolio based on RF predictions in Experiment 1, Figure B.3 in Appendix B illustrates the effects of various levels of constant transaction costs. The presence of costs dramatically lowers the returns of a high turnover portfolio.

While this thesis focuses on equal weight portfolios, market weighted portfolio performance metrics respective to the four experiments are presented in Table A.2, Appendix A. Volatility and turnover are similar to the equal weight results, but return and therefore the Sharpe Ratios differ significantly. In general, the market weight portfolio performance is weaker than their equal weight counterparts, even though some exceptions exist. A possible cause is that the classifier fitting equally weights the observations making the predictions a poor basis for market weighted portfolios. The discussed

market weighted portfolio performance serves as a robustness check to the main equal weight results.

5.2.3 Accumulation and Correlations of Returns

In addition to the summary performance metrics, the over time accumulated returns of the equal weight portfolios are examined. The over time development of returns is fairly flat for BM_{simple}, whereas BM_{FF5} shows an upwards trend.

The first part of Figure 5.1 presents the cumulative return series of the model based portfolios of Experiment 1. The poor performance of SVM is clear, since the returns are nearly constantly the lowest. GB generates relatively very high returns until 2008, when the portfolio suffers dramatic losses. In fact, most model portfolios suffer from the 2008 financial crisis, except the SVM and BM_{FF5}. However, it could be argued that the predictive power of SVM is so weak, that the portfolio is comparable to random selection. Opposite to most of the portfolios, BM_{FF5} gains substantial returns during the financial crisis. After the crisis, RF accumulated returns are constantly high compared to those of other portfolios and it is the only portfolio that experiences a clear upwards trend in returns. Furthermore, the predictive classifier based returns seem fairly correlated, especially GB and BM_{LR}.

Experiment 2 returns, presented in the second part of Figure 5.1, are fairly different than those of Experiment 1. A peak in returns alike GB in Experiment 1 does not appear. On the other hand, the crash in returns during the 2008 crisis is far less severe, but all portfolios exhibit a mostly flat trend in returns after the ciris, except BM_{FF5}. K-NN is constantly among the highest in terms of returns, unlike in the other experiments. Moreover, GB and BM_{LR} appear again fairly correlated, except for the period starting from 2015. Also in common to the former experiment, SVM underperforms constantly. Moreover, we notice that BM_{FF5} dominates the cumulative returns after the 2008 crisis and that the other model based portfolios seem to display fairly different patterns in returns. Since the predictive models are trained using only the subset of five factor containing the same information that is applied in the FF5 model, it can be concluded that the predictive models do not learn a similar model to the FF5 from the same information.

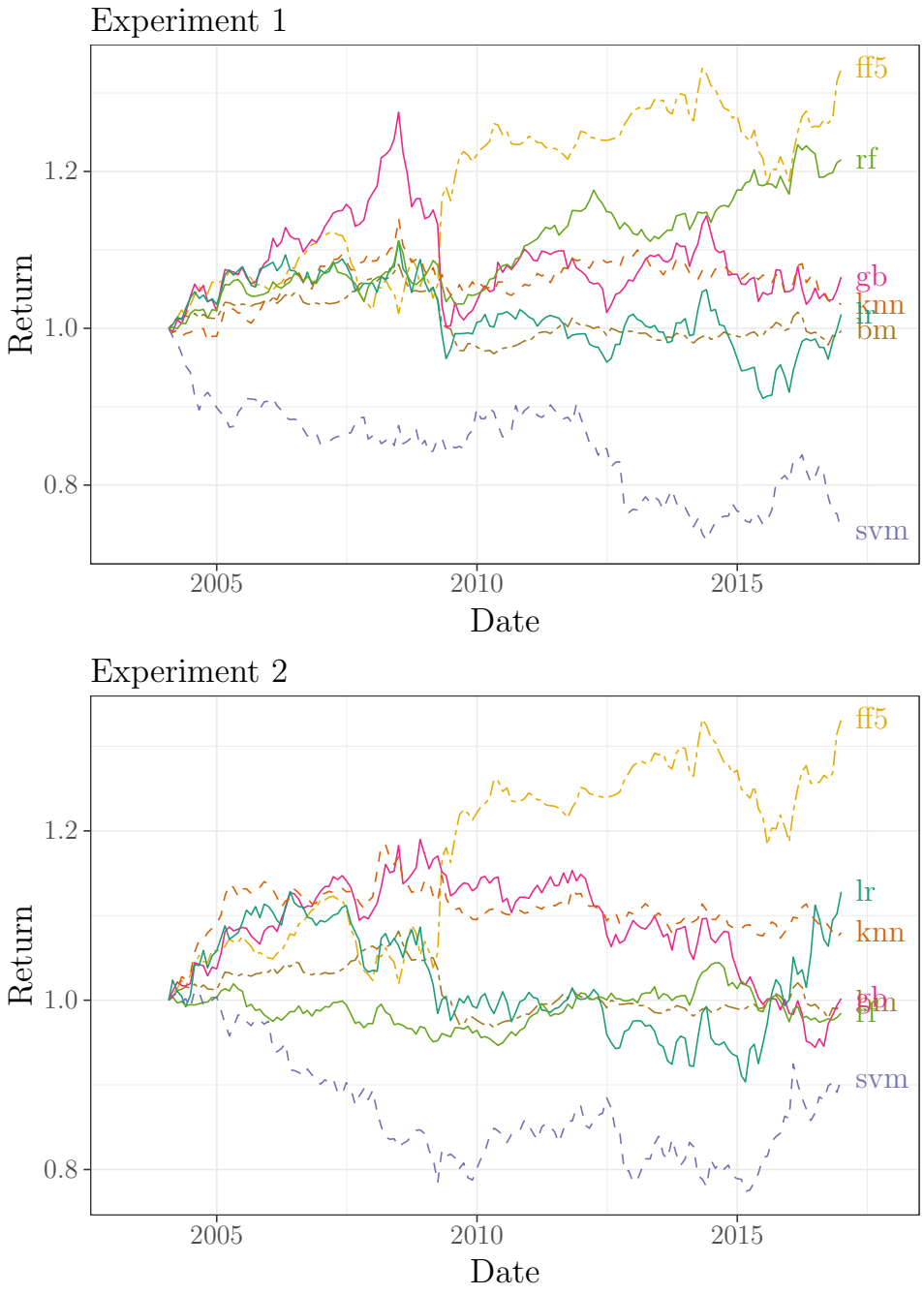

Figure 5.1: Cumulative returns for the model based portfolios in Experiments 1 and 2. BM_{simple} is denoted by bm, BM_{FF5} by ff5, BM_{LR} by lr, K-NN by knn, SVM by svm, GB by gb and RF by rf in the figure legend.

Experiment 3 cumulative returns for the model based portfolios, illustrated in

the first part of Figure 5.2, exhibit a 2008 crash less severe than Experiment 1, and the portfolios recover from the crash faster. GB exhibits very high returns for most of the period, with cumulative returns even higher than BM_{FF5} until the last years of the period. Once again, SVM has almost constantly the lowest returns. Alike in Experiment 2, the portfolios exhibit a rather flat trend starting from 2010. Interestingly, the cumulative returns in Experiment 3 are in general higher than in Experiment 1, even though the first experiment uses a longer training data lookback expected to yield better results. This assumption is not supported by the evidence from comparing Experiments 1 and 3.

The results from Experiment 4 show, in turn, fairly similar patterns to Experiment 1, as presented in the second part of Figure 5.2. GB and BM_{LR} seem again highly correlated and both exhibit the peak followed by a dramatic crash in returns before and during the 2008 crisis. Furthermore, both follow the BM_{FF5} cumulative returns closely. RF exhibits an upwards trend prior to the 2008 crisis similar to GB and BM_{LR}, but weaker. Moreover, RF cumulative returns exhibit a downwards trend after the crash during the crisis. The comparison of the returns in Experiment 4 to those of Experiment 1 indicates that the binary labeling of the fourth experiment yields higher returns than the three-class labeling in the otherwise identical first experiment.

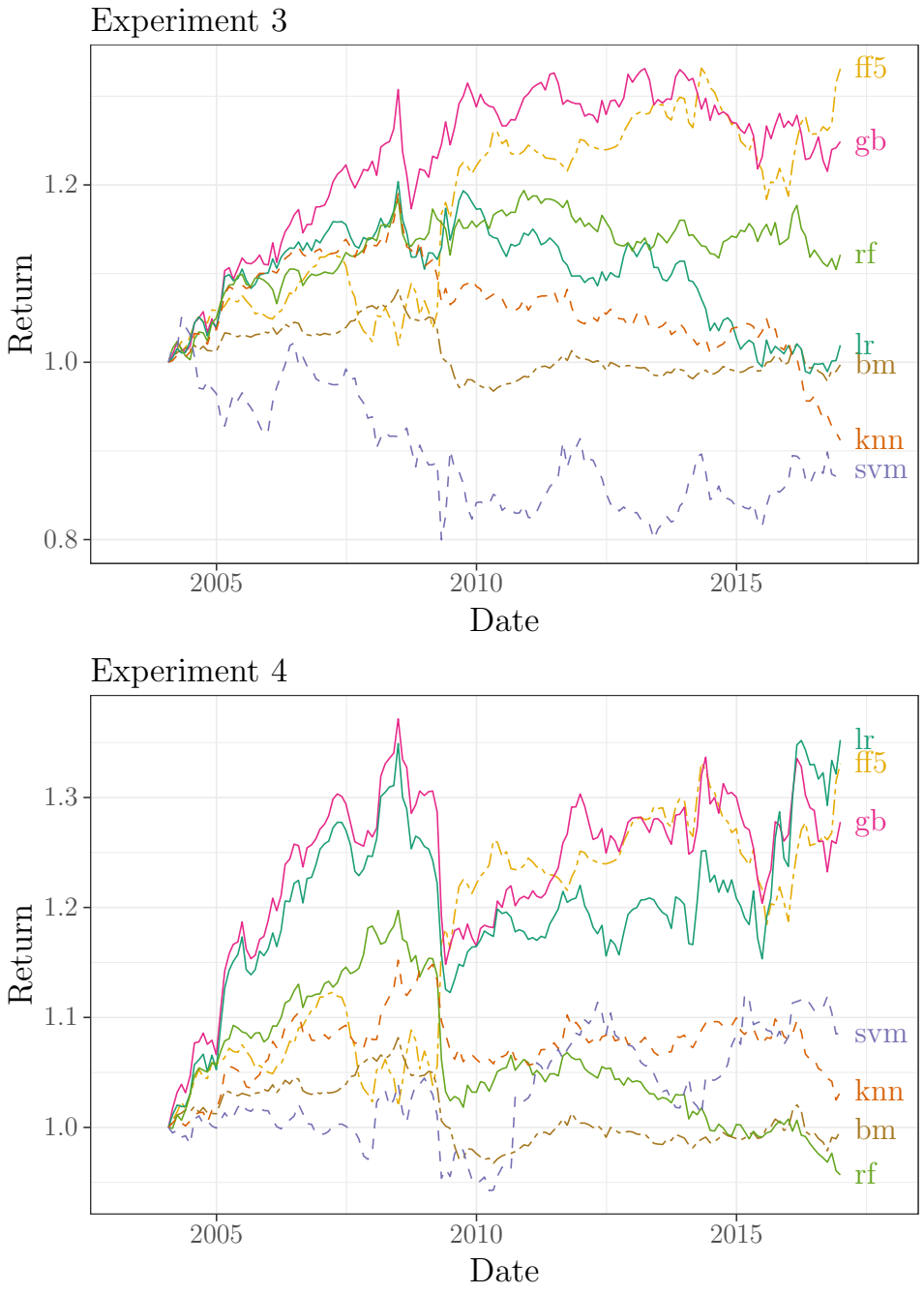

Figure 5.2: Cumulative returns for the model based portfolios in Experiments 3 and 4. BM_{simple} is denoted by bm, BM_{FF5} by ff5, BM_{LR} by lr, K-NN by knn, SVM by svm, GB by gb and RF by rf in the figure legend.

In general, BM_{FF5} dominates the other model based portfolios in terms of

portfolio performance, especially when judged by the results in Table 5.3. The examination of the cumulative return series in Figures 5.1 and 5.2 reveals, that other model based portfolios generate higher cumulative returns during some subperiods of the whole examined period. Especially GB clearly beats BM_{FF5} prior to the 2008 financial crisis in Experiments 1, 3 and 4. Furthermore, the inspection of the cumulative returns reveals that BM_{FF5} tends to generate the gap to the other model based portfolios during the financial crisis, where it generates outstanding returns while others suffer losses.

Since the return graphs indicate substantial correlations between portfolio returns, the correlations of the portfolio returns within experiments calculated and are presented in Table A.3, Appendix A. Indeed the correlations between BM_{LR} and GB are high in Experiments 1, 3, and 4: 0.70, 0.79, and 0.87, respectively. The correlations confirm, that GB and BM_{LR} produce similar returns in these three experiments. The other correlations are in general substantially lower than the few discussed exceptions.

5.2.4 Analysis of the Model Based Portfolio Returns with the Fama-French Five-Factor Model

The FF5 model in (2.4) is estimated for the monthly model based portfolio returns to examine how well the industry standard model explains the returns. The regressions describe how exposed the model based portfolios are to the standard FF5 factors, if there are any excess returns generated and how well the FF5 factors explain the portfolio returns. In summary, the objective of the regressions is to present the model based portfolio returns in terms of the industry standard FF5 model. The regressions conducted with the standard FF5 factors, available at the database of French (2019), as opposed to the replicated FF5 factors. Figures B.1 and B.2 in Appendix B illustrate, respectively, the replicated factors and their standard counterparts.

Table 5.4 presents the regression coefficients obtained by fitting the model in (2.4) with monthly returns of the standard FF5 factors, available at the database of French (2019), to the monthly returns of each model based portfolio. In addition to the coefficients, the table presents their statistical significance and the coefficient of multiple correlation R^2. In general, Table 5.4 exhibits low coefficients and low values of R^2. A notable exception is BM_{FF5}, for which the regression R^2 is 57%. This is naturally explained by the similar construction of the model to the standard FF5. Relatively high values of R^2 are also obtained for Experiment 4, except for the SVM, which indicates that

the return labeling in that experiment leads to portfolios more alike the FF5 factors in returns.

Table 5.4: Regression coefficients and R^2 from regressing the returns of each portfolio against the FF5 factor returns obtained from the database of French (2019) over the whole sample period. The significance of the coefficient is denoted with asterisks as follows: '**' for significance at 99%, and '*' for significance at 95%.

Model	α	$r_m - r_f$	SMB	HML	RMW	CMA	R^2
Benchmark							
BM_{simple}	-0.00**	0.01	-0.04**	-0.03*	0.11**	-0.06*	0.23
BM_{FF5}	-0.00	0.07**	0.12**	0.31**	0.31*	0.15**	0.57
Experiment 1							
BM_{LR}	-0.00*	0.03	-0.05	0.04	0.30**	0.14*	0.16
K-NN	-0.00*	0.01	-0.03	-0.00	0.16**	0.03	0.10
SVM	-0.00**	0.05	0.12**	-0.10*	0.17*	0.02	0.06
GB	-0.00	0.01	-0.01	-0.05	0.26**	0.07	0.09
RF	-0.00	0.01	-0.01	-0.03	0.14**	0.10*	0.08
Experiment 2							
BM_{LR}	-0.00	0.04	-0.08	0.17**	0.32**	0.06	0.16
K-NN	-0.00	0.01	-0.00	0.07	0.16**	-0.03**	0.11
SVM	-0.00	-0.03	0.15**	-0.03	0.05	-0.01	0.03
GB	-0.00**	0.02	0.05	0.06	0.32**	0.05	0.17
RF	-0.00**	0.02	-0.06**	0.07**	0.11**	0.08*	0.18
Experiment 3							
BM_{LR}	-0.00*	0.05*	-0.06	0.02	0.21**	-0.04	0.09
K-NN	-0.00**	0.01	-0.06	-0.05	0.13**	-0.02	0.12
SVM	-0.00	-0.01	0.03	-0.20**	-0.04	0.20	0.06
GB	-0.00	0.04	-0.01	-0.01	0.28**	-0.06	0.10
RF	-0.00	-0.02	0.00	-0.01	0.18**	0.03	0.13
Experiment 4							
BM_{LR}	-0.00	0.02	-0.20**	0.13**	0.42**	0.07	0.34
K-NN	-0.00	-0.03*	-0.05*	0.00	0.15**	0.01	0.23
SVM	-0.00	-0.02	-0.04	-0.05	0.01	0.13	0.04
GB	-0.00	-0.01	-0.12*	0.06	0.36**	0.06	0.26
RF	-0.00*	-0.02	-0.05	-0.03	0.11**	-0.26	0.14

A key interest in Table 5.4 are the coefficients α. A statistically significant $\alpha > 0$ for a portfolio would indicate that the portfolio earns excess returns

the FF5 is unable to capture. However, no values of α different from zero when rounded to two decimals are found. This implies that none of the portfolios is able to generate excess returns compared to the FF5. On the other hand, the overall low values of R^2 indicate that the FF5 factors cannot explain the returns of the classifier based portfolios to a high degree. These two implications could be synthesized as the classifier based portfolios not generating excess returns with respect to the FF5, but that any return generated is based on fairly different factors. The synthesis is logical, since the classifiers exploit more factors than the FF5, except in Experiment 2. In common to the values of α, the market factor coefficients in column $r_m - r_f$, which are the FF5 β coefficients, are very low. This is likely explained by the long-short portfolio construction, that mitigates the effects of market movements on the portfolio returns.

SMB coefficients, in turn, exhibit some relatively great absolute values that often are statistically significant. A positive SMB coefficient can be interpreted as portfolio weighting small companies, whereas a negative coefficient relates a greater weight in big companies. The greatest absolute significant coefficient value is -0.20 for BM_{LR} in Experiment 4. Interestingly, SVM has significant SMB coefficients of 0.12 and 0.15 in Experiments 1 and 2, respectively. However, no apparent reason is identified for these findings.

HML coeffients have an interpretation similar to the SMB: a positive HML coefficient indicates a value company portfolio, while a negative coefficient relates to a growth company portfolio. Most portfolios have very small HML coefficients. In contrast, BM_{FF5} yields the greatest significant HML coefficient of 0.31, and BM_{LR} based portfolios in Experiments 2 and 4 have relatively great significant coefficients of 0.17 and 0.13, respectively. SVM has the lowest significant coefficient of -0.20 in Experiment 3.

The RMW coefficients can be interpreted as a positive value indicating the portfolio weights high profitability companies, while a negative value indicates weight on weak profitability companies. Intuitively, high profitability should reflect as decent returns. Agreeing with this intuition, the RMW coefficients are positive, except for one exception. Moreover, the coefficients are significant, except for three exceptions and in general relatively great in value. It appears that the RMW factor has a large impact in the share returns during the examined period. In fact, the RMW factor has generated the greatest returns of the standard FF5 factors during the sample period, which can be seen in Figure B.2 in Appendix B. Furthermore, GB and BM_{LR} have the highest RMW coefficients in all experiments, implying that they have exploited RMW returns the most. The BM_{LR} yields the greatest significant

coefficient of 0.42 in Experiment 4. The last of the factors, CMA, exhibits no clear patterns. The coefficients are in general small in absolute value, save for a few exceptions.

In summary, the FF5 factor regressions imply that the predictions do not generate significant excess returns not captured by the FF5 model. On the other hand, the returns generated by the predictions are poorly explained by the FF5. Out of the FF5 factors, RMW seems to have had the greatest impact, although it is also the factor with the greatest returns and fluctuations during the sample period, which Figure B.2 in Appendix B exhibits. This may have an inflating effect on the RMW coefficients.

5.2.5 Sensitivity to the Training Data Lookback Length

Since the results are less than robust from the four experiments, we conduct sensitivity analysis of the results with respect to the training data lookback period length. The sensitivity analysis is specifically motivated by the findings from Experiments 1 and 3, indicating that the shorter training data lookback period in Experiment 3 (36 months) could yield higher performance than the otherwise identical Experiment 1 (120 months).

Therefore, we conduct sensitivity analysis by experimenting with training data lookback periods that fall in between those applied in Experiments 1 to 4, covering lookbacks from 36 to 120 months with 12 month intervals. Moreover, the whole set of 30 factors is applied. Differing from the four experiments, SVM and K-NN are excluded from the sensitivity analysis, since their respective results indicate little to no evidence of satisfactory performance. Consequently, the sensitivity analysis is applied with $\mathrm{BM}_{\mathrm{LR}}$, GB and RF. The sensitivity analysis differs also in terms of performance metrics. Since the respective precisions and recalls in Experiments 1 to 4 are generally very close to each other, we examine only the F_1 score in the sensitivity analysis. Moreover, turnover is also excluded from the analysis since it provides little patterns of interest based on the results form Experiments 1 to 4. Sensitivity analysis results are presented in terms of the F_1 score, cumulative returns, volatility and Sharpe Ratio.

Sensitivity is analyzed both under the three-class labeling applied in Experiments 1 to 3, as well as with the binary labeling applied in Experiment 4. First we examine the results obtained by applying the three-class labeling, followed by the respective results obtained by applying the binary labeling. Table 5.5 presents the results of the sensitivity analysis applying three-class

labeling in terms of the F_1 score, return, volatility and Sharpe Ratio.

Table 5.5: Sensitivity analysis results applying three-class labeling. Bold typeface denotes the preferred value for the column metric per experiment The F_1 score is the median of monthly values over the time period. r is the cumulative return and σ the sample volatility. Note, that the results for lookbacks of 36 and 120 months are obtained from Experiments 3 and 1, respectively.

Panel data length [months]	F_1	r	σ	SR
36 (Experiment 3)				
BM_{LR}	0.3722	1.93	4.06	0.06
GB	**0.3856**	**24.89**	4.58	**0.40**
RF	0.3751	12.12	**3.36**	0.28
48				
BM_{LR}	0.3746	-2.87	4.59	-0.03
GB	**0.3798**	10.17	4.96	0.18
RF	0.3752	**12.04**	**3.63**	**0.26**
60				
BM_{LR}	0.3713	-8.50	4.30	-0.14
GB	**0.3784**	-9.49	5.23	-0.12
RF	0.3747	**18.51**	**3.24**	**0.42**
72				
BM_{LR}	0.3712	-2.71	4.09	-0.03
GB	**0.3834**	1.90	4.70	0.05
RF	0.3770	**9.87**	**3.55**	**0.22**
84				
BM_{LR}	0.3719	-10.00	4.03	-0.18
GB	**0.3826**	**6.25**	4.77	**0.12**
RF	0.3757	-0.04	**2.94**	0.01
96				
BM_{LR}	0.3701	-0.29	4.01	-0.01
GB	**0.3820**	**1.77**	4.88	**0.05**
RF	0.3763	-1.75	**3.05**	-0.03
108				
BM_{LR}	0.3719	2.25	4.38	0.06
GB	**0.3831**	-1.86	5.15	0.00
RF	0.3763	**8.63**	**2.94**	**0.23**
120 (Experiment 1)				
BM_{LR}	0.3709	1.77	4.58	0.05
GB	**0.3815**	6.56	5.31	0.12
RF	0.3770	**21.50**	**3.09**	**0.50**

The F_1 scores indicates that the training data lookback has little effect on classification accuracy. However, GB produces the highest score for all lookback periods. The stability of the $\mathrm{BM_{LR}}$ F_1 scores implies that the model fails to exploit the incremental information provided by a longer lookback. The scores for RF are similarly stable, which could be caused by the heavy training data downsampling. Since the downsampling was done by limiting the number of observations, it excludes proportionally more data as the lookback period increases. GB is also subject to training data downsampling, but it is allowed to apply substantially more observations than RF, which could also allow the increasing trend in the GB F_1 scores.

Returns, alike the Sharpe Ratios, display less apparent patterns than classification performance. Perhaps the clearest are, that $\mathrm{BM_{LR}}$ never produces the highest portfolio performance, and that RF always yield the lowest volatility. The highest performing portfolios are obtained with the extreme values of the lookback period, while the performance obtained with the intermediate values tends to be less than satisfactory.

Figure 5.3 illustrates the F_1 scores and returns obtained from the sensitivity analysis in terms of the training data lookback period length. The F_1 scores do not display clear trends. However, it clearly presents the order of the three models in terms of classification performance, as GB has constantly the highest F_1 score, RF constantly the second highest, leaving $\mathrm{BM_{LR}}$ the weakest without exception. Such a pattern is not visible in the model based portfolio returns in Figure 5.3, and, in fact, a prominent relationship between the F_1 score and returns is not apparent.

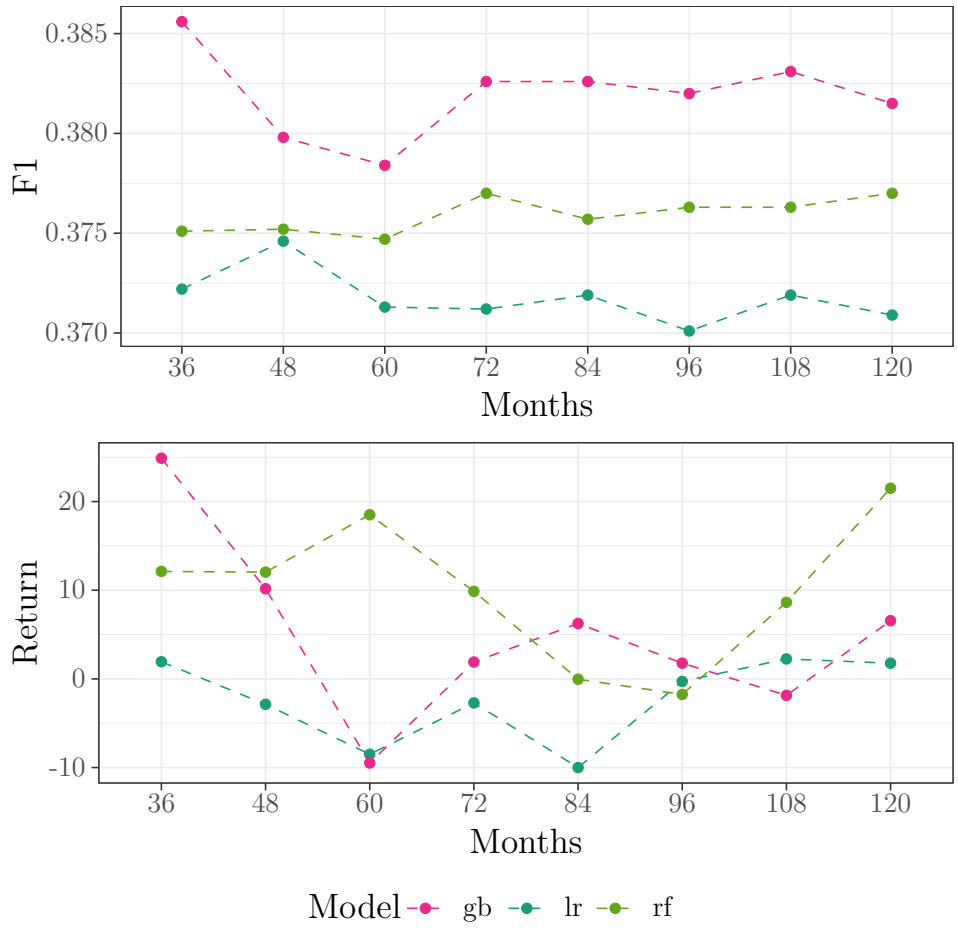

Figure 5.3: The F_1 scores and model based portfolio returns as a function of the training data lookback period. The figure is based on the results of sensitivity analysis applying three-class labeling presented in Table 5.5. BM_{LR} is denoted by lr, GB by gb and RF by rf in the figure legend.

The results from sensitivity analysis applying the binary labeling are presented in Table 5.6. GB dominates the F_1 scores by yielding the highest score for all but the longest lookback, where RF provides the highest scores. Overall, the F_1 scores are barely above the threshold of predictive power of 0.5, but returns are substantially higher than in the three-class regime, where the scores clear the respective threshold with a greater margin. The highest returns, on the other hand, are provided by BM_{LR} for four lookbacks, by GB for three and by RF for one. The highest Sharpe Ratios are also distributed among the models. RF constantly provides the lowest volatility, which in some cases explains the high Sharpe Ratios even with intermediate returns.

Overall, the binary labeling yields substantially higher portfolio performance than the three-class labeling. Especially, GB provides the overall highest return of 36.61% with a lookback of 96 months, and RF yields the maximum obtained Sharpe Ratio of 0.71 with a lookback of 108 months. Alike the results obtained with the three-class labeling in Table 5.5, the binary labeling results do not indicate a clear relationship between classification and portfolio performance.

Table 5.6: Sensitivity analysis results applying binary labeling. Bold typeface denotes the preferred value for the column metric per experiment The F_1 score is the median of monthly values over the time period. r is the cumulative return and σ the sample volatility. Note, that the results for the lookback of 120 months are obtained from Experiment 4.

Panel data length [months]	F_1	r	σ	SR
36				
BM_{LR}	0.5060	**25.47**	5.52	**0.35**
GB	**0.5092**	17.60	6.15	0.23
RF	0.5040	3.41	**3.36**	0.25
48				
BM_{LR}	0.5064	**16.97**	5.72	0.24
GB	**0.5096**	13.44	6.45	0.18
RF	0.5031	15.43	**3.36**	**0.35**
60				
BM_{LR}	0.5063	8.87	5.60	0.15
GB	**0.5083**	**15.14**	6.13	**0.21**
RF	0.5035	1.48	**2.93**	0.05
72				
BM_{LR}	0.5042	21.83	5.34	0.31
GB	**0.5061**	8.70	5.39	0.15
RF	0.5042	**25.88**	**3.16**	**0.58**
84				
BM_{LR}	0.5059	10.38	5.30	0.17
GB	**0.5073**	**28.80**	5.51	**0.38**
RF	0.5045	6.29	**3.08**	0.17
96				
BM_{LR}	0.5031	27.39	5.14	0.39
GB	**0.5078**	**36.61**	4.83	**0.52**
RF	0.5032	19.53	**2.87**	0.50
108				
BM_{LR}	0.5036	**30.84**	5.18	0.43
GB	**0.5057**	28.02	5.22	0.39
RF	0.5039	26.84	**2.65**	**0.71**
120 (Experiment 4)				
BM_{LR}	0.5021	**35.23**	5.44	**0.46**
GB	0.5047	27.77	5.20	0.39
RF	**0.5053**	-4.32	**3.33**	-0.09

Figure 5.4 illustrates the results presented in Table 5.6 in terms of the F_1 score and returns as a function of the training data lookback period length. The F_1 scores for BM_{LR} and GB displays slightly decreasing trend, whereas RF displays a weakly increasing trend. Alike the scores obtained with the three-class labeling, GB dominates the classification accuracy. However, the order of the three models is not as strict like with the three-class labeling. Fluctuation is substantial in returns, however a weakly increasing trend can be seen. Moreover, Figure 5.4 fails to illustrate a clear relationship between classifier and portfolio performance.

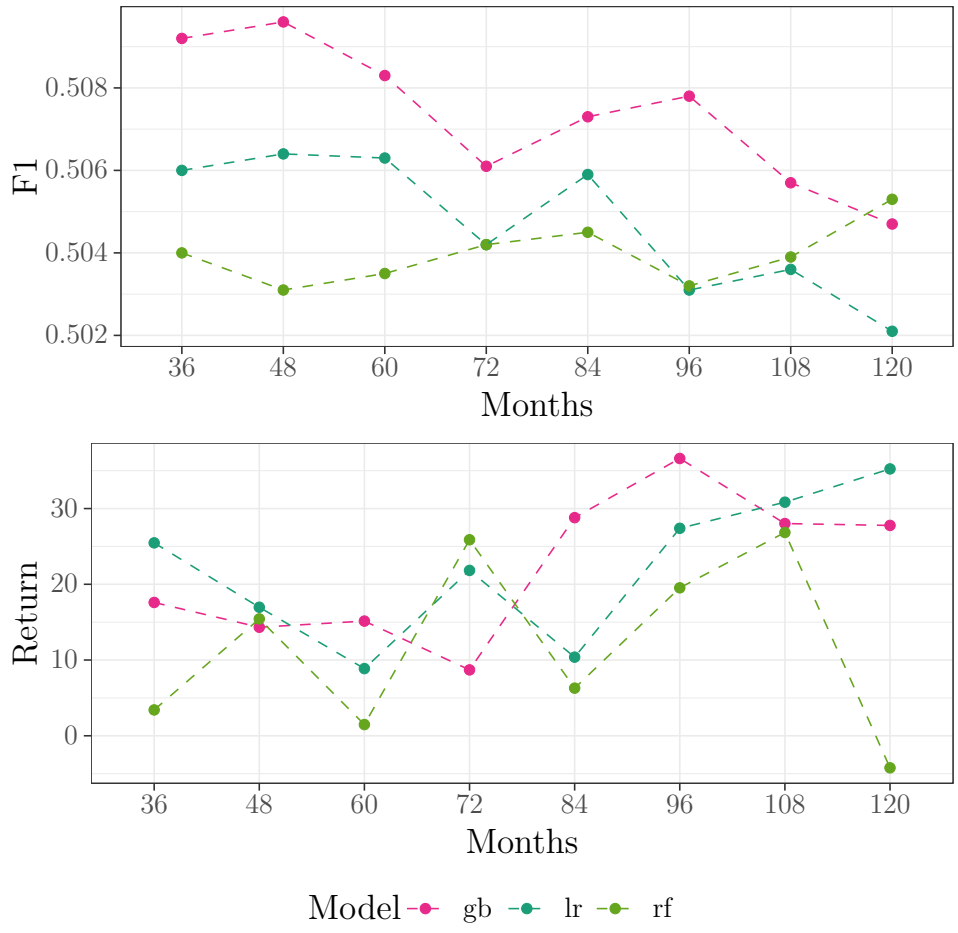

Figure 5.4: The F_1 scores and model based portfolio returns as a function of the training data lookback period. The figure is based on the results of sensitivity analysis applying binary labeling presented in Table 5.6. BM_{LR} is denoted by lr, GB by gb and RF by rf in the figure legend.

The sensitivity analysis for the two labeling regimes can be summarized as follows. Classification performance does not seem to improve as the training data lookback period increases, since the little evidence was found supporting positive relationship. On the other hand, portfolio performance exhibits some evidence of improvement as the lookback period increases. These two implications are intuitively contradictory, as one could reasonably expect returns to increase along with classification performance. Intuitively this means, that the applied metrics fail to accurately capture how the classification performance translates to portfolio performance. As a byproduct of the sensitivity analysis, the additional data obtained on model performance supports the findings of the four initial experiments: the predictive models have predictive power, but only occasionally beat the $\mathrm{BM}_{\mathrm{FF5}}$ presented in Table 5.3. Furthermore, the binary labeling yields substantially higher portfolio performance with than the three-class labeling.

Therefore we take the high performing GB with binary labeling as an example, and examine the effect of the training data lookback length on accumulation of returns in the sensitivity analysis. Specifically, we are interested in the return accumulation during the 2008 crisis, where the $\mathrm{BM}_{\mathrm{FF5}}$ was identified to generate high returns, unlike the predictive models. Figure 5.5 illustrates the return accumulation of GB with different lookbacks in comparison to $\mathrm{BM}_{\mathrm{FF5}}$.

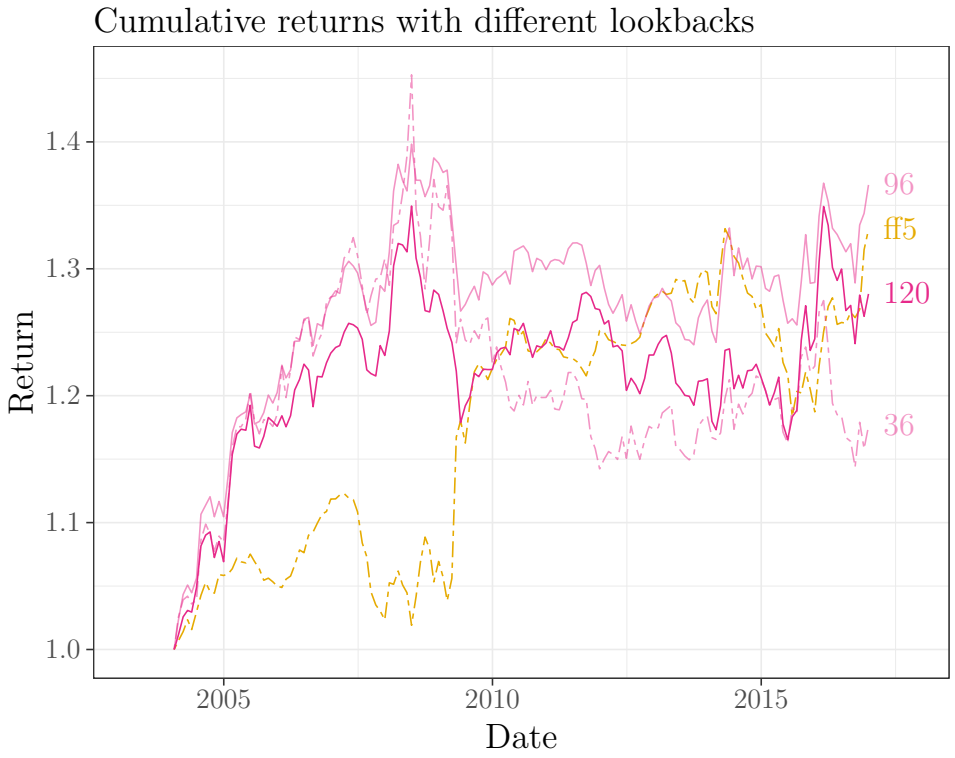

Figure 5.5: The return series of GB with binary labeling from the sensitivity analysis with different lookbacks. In the figure legend, BM_{FF5} is denoted by ff5, whereas 36, 96, and 120 denote GB with respective lookbacks.

The accumulation of returns for GB in Figure 5.5 is similar regardless of the lookback length. However, we notice that the shortest lookback of 36 months generates a very high peak before the crisis, but results with the smallest returns at the end of the period. In general, Figure 5.5 supports the evidence obtained from the four experiments, that the predictive models outperform the BM_{FF5} prior to the 2008 crisis, lose dramatically to the benchmark during the crisis and generate similar returns after the crisis.

Chapter 6

Discussion

Summarizing the four experiments, the obtained classification and portfolio performance measures are rather weak with positive exceptions. The results indicate at least some predictive power for most models, but the industry standard FF5 based benchmark is difficult to beat in terms of portfolio performance. The benchmark generates outstanding returns over the course of the 2008 financial crisis, while other models tend to generate dramatic losses, partly explaining the dominance of the benchmark. The results regarding classification performance are more harmonious than the portfolio performance results, where deviance is substantial. It can be concluded that the results are not very robust with respect to the parameters experimented with. Out of the nonlinear classifiers, GB and RF are arguably the highest performing methods, while SVM fails on most metrics and the performance K-NN is unsatisfactory. The comparison of LR, the linear classifier benchmark, to the nonlinear models provides less than strong evidence in support of superior performance of the nonlinear models.

The sensitivity analysis of the results with respect to the training data lookback period length supports the findings of the four experiments and exhibits little evidence of a clear pattern between performance and the lookback length. Moreover, the additional measurements of performance display less than clear patterns between classification and portfolio performance, while a reasonable assumption would be that higher classification accuracy translates to higher portfolio performance. In addition to coincidence, we identify two possible causes for this, which may coexist. The first is the calculation of the classification performance metrics over the whole set of shares per month, while only a subset of shares is selected in the portfolios. Now the selected shares may be accurately classified, while the excluded ones are not, which

can lead to low overall classification performance metrics together with satisfactory portfolio performance. The opposite effect is equally possible. To examine the effect of this possible phenomenon, the classification performance should be measured separately for the subset of selected shares. The other possible cause is the measurement of classification performance with respect to the class labels. The labels lose information about the magnitude of the share returns. The lost information affects the portfolio performance, but is not reflected on the classification performance metrics. However, this is rather a characteristic of the selected classification approach.

Moreover, the results suggest, that the binary labeling yields higher portfolio performance, although the three-class labeling produces stronger evidence of predictive power. A suspect cause for this phenomenon is the calculation of the classification performance metrics as an average over the classes. If the the middle class in the three-class labeling is the easiest for the classifiers, it could inflate the class-average classification performance metrics when applying the three-class labeling. However, it would not translate to portfolio performance, since the middle class is excluded from the portfolios. A detailed analysis of the per-class classification metrics would be required in order to capture and examine this suspected phenomenon.

The classification performance results indicating predictive power in future return classification agree in general with the related scientific literature, which mainly includes positive findings on performance. However, methodologies, data and the specification of the classification problem differ substantially. For example, Huerta et al. (2013) consider only the extreme tails of the returns in the classification and even drop out the middle section of the data from model training. Moreover they examine daily data, as opposed to monthly data used in this thesis, but similarly fit the models each month. Ballings et al. (2015), on the other hand, apply return thresholds on labeling of future returns to conduct binary classification to high and low return stocks using European data, and predicting returns one year ahead. Differences between approaches exist, in addition to differing problem specification, in the applied classification and portfolio performance measures. One logical explanation lies in the classification type, whether it is binary or multi-class problem. Common metrics such as *receiver operating characteristics* curve (ROC) and *area under the curve* (AUC) applied by Ballings et al. (2015), apply better to binary problems rather than multi-class problems. The classification measures applied in this thesis are equally suited to both types of classification tasks, even though the metric values may be incomparable directly. In fact, Huerta et al. (2013) provide no detailed analysis on classification performance, but present only portfolio performance

metrics.

Metrics of portfolio performance are, on the other hand, more similar between publications: return, volatility and the Sharpe Ratio are widely applied in the literature, with additional metrics varying between publications. Alike classification performance metrics, the specification of the classification problem affects the criteria for assessing the portfolio performance metrics. Further impact stems from the portfolio construction methodology. Values of portfolio performance measures for long-short portfolios constructed in this thesis can be assumed to be quite different from long-only portfolios. The choice of the shares included, as well as the time period examined, also affect the metrics, making direct comparisons between the results of the publications challenging. Further examining the aforementioned articles by Huerta et al. (2013) and Ballings et al. (2015) reveals that the former reports an annual α of 15% indicating substantially higher performance than the α values obtained in this thesis. Fan and Palaniswami (2001), in turn, apply SVMs to detect shares with exceptional returns using data from the Australian stock exchange, which are used to construct long-only portfolios. They compare the SVM produced portfolios to a "market" benchmark constructed as an equal weight portfolio of all examined shares, and report SVM providing approximately three-fold returns over the five-year period from 1995 to 1999. Compared to the results of this thesis, the related literature documents findings considerably stronger in favor of superior performance of nonlinear multifactor models.

A common property to related research, that is different to this thesis, is the inclusion of a cross-validation step in model fitting. The step dismissed in this thesis is usually performed in related research. It is likely to improve the performance metrics reported in the publications, and it can be assumed that the results obtained in this thesis would improve by implementing cross-validation. In addition, the selection of the factors could be done systematically by including a *feature selection* step, where only the factors with the most predictive power are selected for predicting. Moreover, a possible extension to the model fitting methodology with possibly significant effects on performance is to account the different cost of misclassification in multi-class problems. The methods in this thesis consider all misclassification equal, while it is obviously more dangerous to mistake a high-return share for a low-return share, instead of a neutral-return share. The implementation of this extension is not considered in this thesis, but most machine learning methods applied are likely extendable in this direction.

Notably, GB is relatively a well performing method, even though it is subject

to downsampling of the training data done in order to decrease computation time. SVM and RF are subject to heavier training data downsampling than GB, which likely the cause for the unsatisfactory performance of SVM. However, RF performs substantially better than the SVM, which could be due to the ensemble construction of the RF, mitigating the effect of the downsampling. The results obtained would likely be positively affected by omitting the training data downsampling.

The set of selected methods as well as the highly simplified fitting procedure can be viewed to as a base case regarding the methods. Moreover, we identified multiple approaches to enhance the performance of the methods. In conclusion, the results of this thesis combined with the improvement potential regarding the nonlinear methods weakly support the feasibility of the nonlinear multifactor models in classification of future share returns.

Chapter 7

Conclusion

The objective of this thesis was to determine the feasibility of nonlinear multifactor models for predictive classification of future share returns. The findings provide weak support to the feasibility. However, the model fitting methodology of this thesis is highly simplified. Extending the methodology with cross-validation or feature selection is expected to improve model performance. The findings and the identified extensions to the methodology indicate that the nonlinear multifactor models are feasible to the specific problem, but careful development and further research is required to possibly outperform the industry standard factor model FF5.

The feasibility was determined examining models fitted with four nonlinear machine learning methods: K-Nearest Neighbors, nonlinear Support Vector Machine, Gradient Boosting Machine and Random Forest. The methods were applied to fit classifiers for predicting future share returns using past data, and the predictions were in turn applied to construct long-short portfolios. The examination included a comparison of classification performance to thresholds indicating predictive power and to Logistic Regression, a linear predictive method. Moreover, the classifier based portfolios were compared in terms of performance to assess the economic impact of the classifier predictions. The comparison of portfolio performance was based on benchmarks including a replication of the industry standard FF5 factors. Moreover, the portfolio returns were analyzed with the widely applied standard FF5 factors by regression. The regressions indicate no excess returns were generated with respect to the FF5. On the other hand, the regressions are unable to explain the portfolio returns to a high degree.

The methodology was applied to monthly return and quarterly fundamental

data from 1988 to 2016 on US stocks to construct portfolios from 2004 to 2016. A lag of three months was imposed on the fundamental data to account for the reporting delay. This ensures that all data used to fit the classifiers was available to the market at the time. The effect of small stocks on the results was mitigated by excluding the smallest stocks from the analysis. The resulting dataset was used to construct 30 factors, on which the factor models were based. The factors were gathered from recent scientific literature.

Feasibility is further supported by readily available implementations of the machine learning methods. The implementations proved efficient enough in terms of computational requirements, even though some methods required considerable data downsampling. The machine learning methods applied in this thesis vary in complexity, but in general a machine learning model is difficult to interpret. This makes the interpretability of models alike FF5 hard to beat, which may discourage some practical applications of nonlinear multifactor models. Another practical limitation considering the prediction based portfolios is their high turnover, which generates high transaction costs, that eats away returns.

The findings of this thesis inspire a number of directions for further research. Most importantly the effects of including cross-validation and feature selection to the classifier fitting procedure should be examined, since it is expected to improve performance. Moreover, the methodology could be extended to account for different costs for misclassifications in multiclass model training: intuitively mistaking a high-return share for a low-return share, instead of a neutral-return share should have a high cost. On the other hand, the portfolio construction methodology of this thesis includes some fairly arbitrary decisions, such as the return labeling breakpoints and the number of shares in the portfolio, or how often and for what period the returns should be predicted. Moreover, the feasibility should be inspected with different data, for example with European or Asian shares, in addition to including other factors to the models. Moreover, as there is an identified need to map out the numerous findings of factors, the results of this thesis imply a similar need to map out the true potential of nonlinear multifactor models. Since the current related literature is scarce, most new findings contribute to the knowledge. Simultaneously, a question of common comparison methodology arises, not forgetting the impact the so far rare replication studies may have.

Bibliography

Paul D. Allison. *Multiple Regression: A Primer*. Pine Forge Press, 1999.

Andrew Ang. *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press, New York, 2014.

Clifford Asness and Andrea Frazzini. The Devil in HML's Details. *The Journal of Portfolio Management*, 39(4):49–68, 2013.

Michel Ballings, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20):7046–7056, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Ohlsen. *Classification and Regression Trees*. Taylor Francis, 1984.

Leo Breiman, Adele Cutler, Andy Liaw, and Matthew Wiener. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*, 2018.

Mark M. Carhart. On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1):57–82, 1997.

John H. Cochrane. New Facts in Finance. *Economic Perspectives*, 23(3): 36–58, 1999a.

John H. Cochrane. Portfolio Advice for a Multifactor World. *Economic Perspectives*, 23(3):59–78, 1999b.

John H. Cochrane. Presidential Address: Discount Rates. *The Journal of Finance*, 66(4):1047–1108, 2011.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

CRSP. Center for Research in Security Prices (CRSP): Monthly Stock File. Available at Wharton Research Data Sevices (WRDS), 2019. URL `https://wrds-web.wharton.upenn.edu/wrds/`.

CRSP/Compustat. Center for Research in Security Prices (CRSP): CRSP/Compustat Merged database - Fundamentals Quarterly. Available at Wharton Research Data Sevices (WRDS), 2018. URL `https://wrds-web.wharton.upenn.edu/wrds/`.

Robert F. Dittmar. Nonlinear Pricing Kernels, Kurtosis Preference, and Evidence from the Cross-Section of Equity Returns. *The Journal of Finance*, 57(1):369–404, 2002.

Peter D. Easton and Mark E. Zmijewski. SEC Form 10K/10Q Reports and Annual Reports to Shareholders: Reporting Lags and Squared Market Model Prediction Errors. *Journal of Accounting Research*, 31(1):113–129, 1993.

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

Eugene F. Fama and Kenneth R. French. Disecting Anomalies. *The Journal of Finance*, 63(4):1653–1678, 2008.

Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Alan Fan and Marimuthu Palaniswami. Stock Selection using Support Vector Machines. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, volume 3, pages 1794–1798, 2001.

Kenneth R. French. Data Library. Available at the Kenneth R. French website., 2019. URL `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

Jerome Friedman, Trevor Hastie, and Noah Simon. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2018.

Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Brandon Greenwell, Bradley Boehmke, and Jay Cunningham. *gbm: Generalized Boosted Regression Models*, 2019.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Chicago Booth Research Paper No. 18-04*, 2018.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ...and the Cross-Section of Expected Returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Kewei Hou, Cheng Xue, and Lu Zhang. Replicating Anomalies. *Fischer College of Business Working Paper Series 2017-03-010*, 2017.

Ramon Huerta, Fernando Corbacho, and Charles Elkan. Nonlinear support vector machines can systematically identify stocks with high and low future returns. *Algorithmic Finance*, 2(1):45–58, 2013.

Sadegh Bafandeh Imandoust and Mohammad Bolandraftar. Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange. *International Journal of Engineering Research and Applications*, 4(6):106–117, 2014.

S. P. Kothari, Jay Shanken, and Richard G. Sloan. Another Look at the Cross-Section of Expected Stock Returns. *The Journal of Finance*, 50(1): 185–224, 1995.

Max Kuhn. *caret: Classification and Regression Training*, 2018.

Asriel U. Levin. Stock Selection via Nonliear Multi-Factor Models. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, volume 8, pages 966–972. MIT Press, 1995.

John Lintner. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 47(1):13–37, 1965.

Marcos Lopez de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.

David G. Luenberger. *Investment Science*. Oxford University Press, 1998.

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

David G. McMillan. Nonlinear predictability of stock market returns: Evidence from nonparametric and threshold models. *International Review of Economics and Finance*, 10(4):353–368, 2001.

David Meyer, Evgenia Dimitiadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2019.

Jan Mossin. Equilibrium in a Capital Asset Market. *Econometrica*, 34(4): 768–783, 1966.

Andrew J. Patton and Allan Timmermann. Monotonicity in asset returns: New tests with applications to the term structure of the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98(3):605–625, 2010.

Mario Pitsillis. Review of literature on multifactor asset pricing models. In John L. Knight and Stephen Satchell, editors, *Linear Factor Models in Finance*, chapter 1, pages 1–11. Elsevier Science Technology, Oxford, 2004.

Joseph P. Romano and Michael Wolf. Testing for monotonicity in expected asset returns. *Journal of Empirical Finance*, 23:93–116, 2013.

Stephen A. Ross. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

Stephen A. Ross. Invited Editorial Comment. *The Journal of Portfolio Management*, 43(5):1–5, 2017.

William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 19(3):425–442, 1964.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

Seisuke Sugitomo and Shotaro Minami. Fundamental Factor Models Using Machine Learning. *Journal of Mathematic Finance*, 8(1):111–118, 2018.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

Appendix A

Tables

Table A.1: Missing observation percentages for the factors. The percentages are calculated over the whole data.

Factor	**Accronym**	**Missing**
6-month prior return	R^6	5.07
11-month prior return	R^{11}	8.26
Book-to-market*	Bm	1.03
Earnings-to-price	Ep	0.79
Enterprise multiple	Em	12.06
Cash flow-to-price	Cp	9.31
Operating cash flow-to-price	Ocp	10.87
Investment-to-assets*	Ia	9.29
Changes in gross property, plants and equipment and inventory-to-assets	dPia	47.76
Changes in net operating assets	dNoa	9.69
Net stock issues	Nsi	8.02
Composite equity issuance	Cei	32.83
Inventory changes	Ivc	11.85
Changes in net non-cash working capital	dWc	28.94
Change in net non-current operating assets	dNco	26.88
Change in non-current operating assets	dNca	26.84
Change in financial liabilities	dFnl	9.16
Return on equity	Roe	3.55
Changes in return on equity	dRoe	12.40
Changes in return on assets	dRoa	12.14
Gross profits-to-lagged assets	Gla	15.62
Operating profits-to-equity*	Ope	19.10
Operating profits-to-lagged equity	Ole	15.37
Operating profits-to-lagged assets	Ola	18.94
Cash-based operating profits-to-lagged assets	Cla	31.28
12 month lagged return	R_a^1	8.02
Years 2-5 lagged returns	$R_a^{[2,5]}$	34.66
Beta*	Beta	19.64
Asset liquidity	Alm	22.22
Market equity*	Me	0

Table A.2: Performance metrics for market weighted portfolios. They are based on the same predictions as the examined equal weight portfolios.

Model	r	σ	SR	TO
Benchmark				
BM_{simple}	5.83	1.89	0.24	29.07
BM_{FF5}	28.64	4.82	0.43	14.18
Experiment 1				
BM_{LR}	-14.27	5.77	-0.18	39.88
K-NN	1.85	5.56	0.05	67.23
SVM	-5.19	6.46	-0.03	89.74
GB	-4.81	6.51	-0.03	58.47
RF	16.11	4.30	0.29	81.81
Experiment 2				
BM_{LR}	17.31	6.62	0.22	27.88
K-NN	-7.90	4.72	-0.11	57.80
SVM	-12.71	6.74	-0.12	90.36
GB	-11.47	6.03	-0.13	51.58
RF	6.57	3.71	0.15	88.19
Experiment 3				
BM_{LR}	-19.18	5.45	-0.27	40.74
K-NN	-24.87	4.96	-0.42	48.09
SVM	10.35	6.82	0.15	89.71
GB	-2.13	6.62	0.01	48.09
RF	10.35	5.01	0.14	48.09
Experiment 4				
BM_{LR}	27.67	6.63	0.32	48.77
K-NN	3.24	3.36	0.09	71.38
SVM	27.48	4.62	0.43	93.58
GB	0.23	5.70	0.03	69.06
RF	-5.53	3.59	-0.10	85.38

Table A.3: Return series correlations.

	BM_{simple}	BM_{FF5}	BM_{LR}	K-NN	SVM	GB	RF
Benchmark							
BM_{simple}	1						
BM_{FF5}	-0.41	1					
Experiment 1							
BM_{LR}	0.47	0.02	1				
K-NN	0.46	-0.01	0.45	1			
SVM	0.05	-0.06	-0.01	-0.03	1		
GB	0.63	-0.25	0.70	0.45	-0.05	1	
RF	0.50	-0.04	0.45	0.45	0.00	0.51	1
Experiment 2							
BM_{LR}	0.27	0.27	1				
K-NN	0.17	0.31	0.28	1			
SVM	-0.12	0.06	-0.01	-0.05	1		
GB	0.33	0.11	0.40	0.34	-0.17	1	
RF	0.05	0.28	0.23	0.21	0.02	0.15	1
Experiment 3							
BM_{LR}	0.29	0.05	1				
K-NN	0.39	-0.34	0.56	1			
SVM	0.24	-0.26	-0.09	0.04	1		
GB	0.39	-0.05	0.79	0.56	-0.07	1	
RF	0.33	-0.05	0.52	0.54	-0.13	0.58	1
Experiment 4							
BM_{LR}	0.62	0.05	1				
K-NN	0.57	-0.26	0.62	1			
SVM	0.23	-0.17	0.15	0.21	1		
GB	0.66	-0.10	0.87	0.66	0.19	1	
RF	0.52	-0.38	0.54	0.57	0.13	0.61	1

Appendix B

Figures

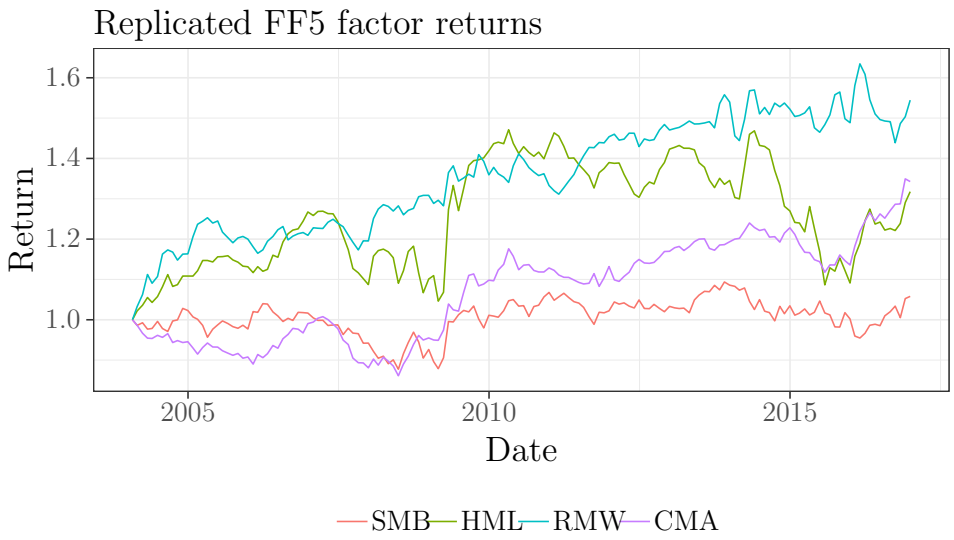

Figure B.1: The returns of the replicated FF5 factors SMB, HML, RMW and CMA. BM_{FF5} is the equal weighted mean of these four factors. Note, that the replicated factors are based on equal weighted factor mimicking portfolios.

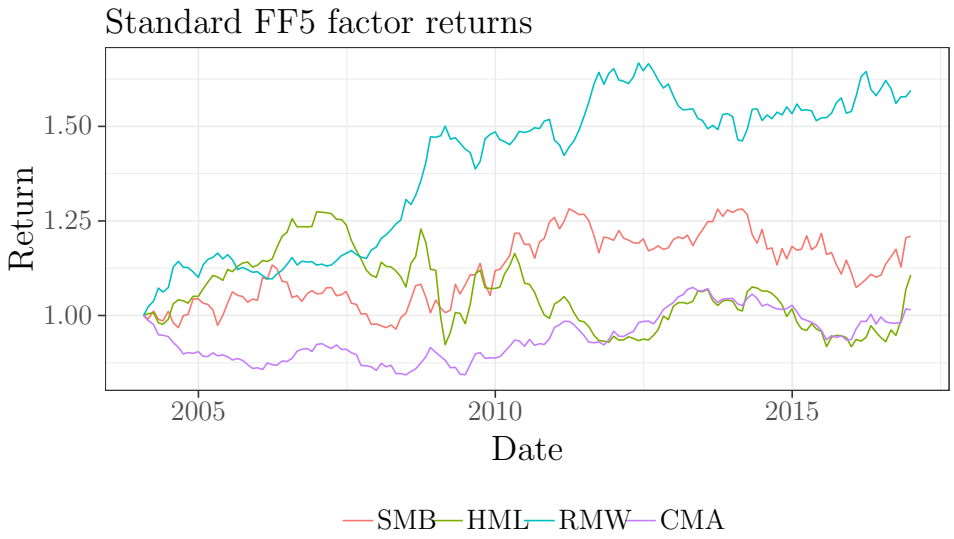

Figure B.2: The returns of the standard FF5 factors SMB, HML, RMW and CMA from French (2019) used in the regression analysis of the model based portfolio returns. The factors are based on market weighted factor mimicking portfolios.

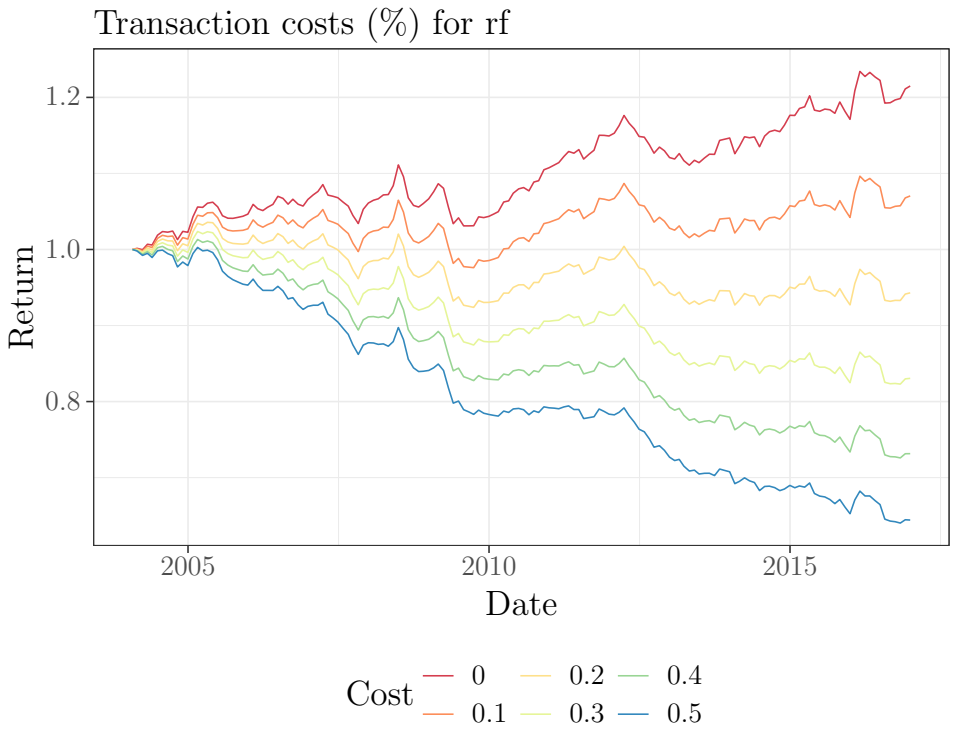

Figure B.3: The effect of various levels of constant transaction costs on the cumulative returns of RF based portfolio in Experiment 1.