GÖTEBORGS UNIVERSITET
HANDELSHÖGSKOLAN

Making Use Of The Factor Zoo

An unpretentious attempt to predict asset returns using machine learning methods.

Authors:
Line Clausen
Jesper Strömberg

Bachelor thesis (15hp)

Thesis advisor: Charles Nadeau
Date: February 11, 2022
School of Business, Economics and Law
Department of Economics and Statistics

Contents

Abstract

Factor modeling with the purpose of estimating assets returns is a dynamic and ever changing subject within finance. Recent literature has presented over 300 different factors that seem to have significance in predicting asset returns. This new phenomenon in factor modeling has been dubbed "The Factor Zoo". At first, the subject was marked by optimism, but soon researchers with a more pessimistic view entered the debate criticizing performed studies mainly concerning the choice of methodology. A presented solution to the issue of imperfect models is to utilize Machine Learning methods as they can handle the extensive nature of the "Factor Zoo". Feature selection is the idea of reducing the dimensionality in data sets to increase performance, interpretability and lower computational time. This thesis presents the evaluation of an ensemble of different feature selection models on a financial data set comprised of 30 firms from the S&P500 index and 123 firm characteristics during a ten-year period spanning from 2011-07 to 2021-07. Several established models handling feature selection were chosen. By comparing the models' performance the thesis reached confidence in the selection of final model. The proposed final model was the Lasso which outperformed both the other regressor models but also the benchmark as in Fama French Five Factor Model. By analysing the selection of firm characteristics by the Lasso, important features such as *'relative strength index'*, *'price/sales'* and *'price rel 52 week high'* showed great significance regardless of the chosen portfolio at hand.

Acknowledgements

We would like to thank our supervisor Charles Nadeau, senior lecturer in the department of Economics and Statistics, for his patience and great support during the journey of our thesis.

1 Introduction

1.1 Background

The subject of asset pricing is centred around apprehending the monetary value of uncertain forthcoming payments. The payment in question is an asset such as an equity. The value of an asset is composed of two parts; the value of time and the value of risk. The first component is straightforward as it represents the time value of money. In contrast, the second component is harder to evaluate, making it a more complex feature in the valuation and pricing of an asset[1].

Pricing of assets is done in multiple areas for many different purposes. It is especially common within the field of equity valuation, where asset pricing can be done with the use of factor models. To calculate the expected return and risk of an equity, factor models are utilized as it is assumed that they operate as a proxy for risk. As investors carry a certain individual appetite for risk and consequently strive to hold assets according to their level of risk, it is necessary to have well functioning models that can approximate an assets level of risk[2].

The assumption that factor models can proxy the risk of an asset is based on the *Capital Asset Pricing Model*[3]. The CAPM is a single factor linear regression model, where the single factor is the market risk of the asset. The simplicity of the CAPM makes it easy to interpret, thus it is usually taught in academics. Even though the CAPM is theoretically reasonable, when testing the model, anomalies has been discovered. This implies that the simplicity of the model may in fact be a downside.

In the light of these discoveries, models consisting of multiple other factors have been developed. For example the well known *Fama French Three Factor model*(Fama French 1993) [4]. The model includes the market factor just as the CAPM, but also includes a factor representing the size and a factor representing the value. Since the outcomes of Fama and French, the concept of using factor models for predicting asset return has evolved as the accessibility to large data sets has improved and as many new factors have shown to have predictive power for cross-sectional expected return. In recently publiched study, Harvey et al. (2016) identified over 300 significant factors with the purpose to predict asset returns [5].

The issue with finding a considerable amount of factors as relevant for estimating asset returns, is that some of the factors may be nested in each other. Furthermore, with the consequent discovery of factors that can potentially clarify certain market movements, hypotheses like the Efficient Market Hypothesis, are becoming more and more controvertible. Last but not least, in line with the reason for using factor models, it is highly relevant to stress the importance of the fact that any new factor added to a new model, need to contribute with the same predictive power in *out of sample data*, for the factor to actually be relevant. In other words, the models need to work on data that has not yet been generated. In his presidential address, John Cochrane examines the mentioned difficulties at the present state of the search for significant factors for asset pricing. He dubs the great access to factors as the "factor zoo" [6].

The issue presented by Cochrane has been challenged by multiple researches. In general, an emphasis has been made on the econometric challenges of the issue. Because of the nature of the problem, with large data sets and the challenge of extracting distinct significant features, machine learning approaches have shown to be the prominent choice of method for creating the models. Especially, the LASSO, developed by Robert Tibshirani, is a method that has been shown to tackle the issue adequately as it has been used for the purpose of both variable selection and prediction of asset returns [7].

By demonstrating an unconditional attempt to utilize the Factor Zoo with the purpose of predicting asset returns, this thesis is a contribution to the ongoing research on the development of factor models. The thesis evaluates how the LASSO and other methods perform and what factors they select, in a setting where the researches are limited by both time and knowledge.

1.2 Purpose and Research Question

The purpose of this thesis is to participate in the challenge presented by Cochrane and find factors that have power in predicting asset returns using Machine Learning methods. The assets are common stocks that currently lie within the Standard and Poor's stock market index.

In the light of Cochrane's stated question, and to support the purpose of the thesis, the research question is formulated as follows:

Which firm characteristics are most predictive for asset returns?

The question is although somewhat two-folded, since to reach the answer, one must first answer which models performs the best when extracting firm characteristics.

1.3 Delimitations

The data is constrained to only currently listed companies on the S&P500 index. This is due to the relatively high quality of the data and the relatively easy access to the data, which is necessary for the purpose of the thesis. Data for the period 2011-07-01 to 2021-07-01 is extracted for 30 firms from Capital IQ, composing 3 sample portfolios each consisting of 10 companies. All firms have been listen for the entire period. The specific firms are listed in the Methodology section.

The Methodology of the thesis is limited by time and knowledge. The time limit means that the essay only focuses on one main algorithm, namely LASSO. The other methods used are not presented in detail nor are the results in feature selection presented. The knowledge limitation means that some of the methods are not as rigorously theoretically derived as those presented in the literature review.

2 Literature Review

The search of the optimal model for estimating returns for common stocks has been going on for a long time, making asset pricing a well-established and well-known field within economics and finance. During past years, researchers have unveiled hundreds of different firm characteristics that seem to explain asset returns, which has led to the creation

of many new models. Many of which seem to deviate from the commonly known Capital Asset Pricing Model(Cochrane 2011)[6]. These findings have led to a rise in factor-based investing and asset-pricing theory revaluation (Ang 2014)[8].

In the paper "Presidential Address: Discount Rates" by John H. Cochrane, published in 2011, Cochrane presents the "The multidimensional challenge" [6]. The challenge describes the complexity of generating new models consisting of numerous factors. He acknowledges that, as the number of factors in a model is extensive, the conventional way of examining return via portfolio sorting, see Fama French (1993)[4], Carhart (1997) [9] and Fama French (2015) [10] performs inadequately. With this in mind, he presents the following questions
1) "First, which characteristics really provide independent information about average returns? Which are subsumed by others?",
2) "(...) does each new anomaly variable also correspond to a new factor formed on those same anomalies?",
3) "(...) how many of these new factors are really important?" and
4) "(...) why do prices move?".

Cochrane's questions form the basis for posterity literature on the subject. The following part of the literature review presents research based on Cochrane's question, focusing on the data and methodology used and the results produced. Also, recurring criticism on the subject is presented.

When Harvey et al. (2016) [5] enter the debate concerning the Factor Zoo, they hold a doubting attitude believing that most of the literature presented on the subject is likely false as the criteria used to select factors are insufficient. They argue that utilizing the usual statistical significance cutoffs in data mining contexts does not work. Their results show that as many as 316 factors demonstrate strength in estimating returns. However, the authors return to the thesis statement in the paper's conclusion, arguing that these are probably false discoveries.

Green et al.(2017) [11] accepted the challenge presented by Cochrane. In the light of the predecessors Fama and French (1992) [12] , they use a Fama - McBeth regression method where they account for 94 independent factors. To avoid overestimating the value of microcaps(firms with a market capitalization between 50 and 100 million U.S dollars), they conduct regression using market-value-weighted least square(VWLS). They also perform a regression using Ordinary Least Squares(OLS). The data for the analysis consists of all firms listed on either AMEX, NASDAQ or NYSE with a non-missing value for common equity in their annual financial statements and month-end market value on Center for Research in Security Prices(CRSP). They claim that most firms' characteristics became robust in 1980 and performed the analysis on data ranging from 1980 to 2014. The results they obtained from the OLS regression showed that only 12 of these factors show significance, namely: *"asset growth, growth in industry-adjusted sales, the percentage change in shares outstanding, growth in inventory, earnings announcement return, growth in book equity, growth in CAPEX, growth in long-term net operating assets, growth in PPE plus inventory, number of consecutive quarters with earnings higher than the same quarter a year ago, growth in sales less growth in inventory, and standardized unexpected quarterly earnings"* . For the VWLS regression, only the factor "growth in long-term net operating

assets" showed significance. Also, these factors only showed significance for the period before 2003. For the period after 2003, only two factors showed significance.

Using a different approach to answer Cochrane's question, Freyberger et al. (2017) [13] use the adaptive group LASSO(least absolute shrinkage and selection operator), a non-parametric method, to identify significant factors. The method does not impose a strong functional-form, it is less sensitive to outliers and can handle a large number of input factors. Furthermore, it is shown that compared to linear panel regression, the method has a higher explanatory power in the out of sample test. The data consists of monthly return data retrieved from The Center For Research In Security Prices(CRSP). Just like Green et al.(2017), the data is compromised of stock prices for firms listed on either AMEX, NASDAQ or NYSET. However, the amount of data is more extensive, ranging from 1963 to 2015, since Freyberger et al. (2017) also consider the impact of time on estimating asset return. The researches estimate their model using 36 firm characteristics. These are *"size, book-to-market, beta, and other prominent variables and anomalies"*. The result they obtained showed that 15 factors showed significance. These include *"size, idiosyncratic volatility, and past return-based predictors"*. Only seven factors, including *"size, past returns, and standardized unexplained volume"*, continue to be identified as significant factors, as the model takes into account only stocks with a market capitalization above 20 per cent of the NYSE (size percentile).

Feng et al. (2017) [14] discuss the suitable method for handling the "Factor zoo". They propose a model selection method that systematically evaluates the impact of each new-found factor on the estimation of asset pricing. They point out that traditional methods tend to assume perfectly variable selection making the final models ambiguous due to potential variable bias from potentially omitted variables. Their method takes these mistakes into account, making their model more reliable. They argue that the single use of LASSO to produce the relevant variables for estimating average returns is not a method that produces credible results, as a significant coefficient for a variable does not guarantee that it is precisely this variable that is one of the true factors. The method they propose can evaluate the contribution of a factor in a high-dimensional setting by combining the traditional econometric methods with the newer, Fama-MacBeth regression and double selection- LASSO (DS LASSO) developed by Belloni, Chernozhukov, and Hansen (2014b) [15].

Kelly et al. (2019) [16] attempt to answer Cochrane's (2011) question differs from the previously presented articles concerning model choice as they focus on feature extraction rather then feature selection. They develop their own instrumental Principal Component Analysis (IPCA). When comparing their method with other established methods, they find that IPCA delivers higher out-of-sample mean-variance efficiency. Furthermore, in the light of both the out-of-sample test and the in-sample test, and in comparison with, among others, the benchmark model Fama French Five-Factor Model (Fama & French 2015) [10], they find that their model performs much better in terms of getting accurate predictions. 12-month momentum and size are the factors that demonstrate the highest predictive power.

In a recently published paper, Peng et al. (2021) [17] examine feature selection for asset pricing by utilizing multiple machine learning methods, such as deep neural networks,

for feature selection. The data was daily observations from 2008 and 2019 for stocks listed on several market indexes. The researchers accounted for 124 features and applied the different methods to achieve some form of consensus estimation. The results show that the feature selection algorithms did not choose the same variables, which indicated that the algorithms did not perform as they would have hoped.

Bartram et al. (2021) [18] comment on Crochane's Factor Zoo by taking the standpoint of an institutional investor [18]. The researchers concentrate on the usefulness of the evidence found concerning factor-based investment strategies. An emphasis is on the need to examine the recent achievements within financial literature and the many factors documented, as they may be more apparent than actual from a practical perspective. The researchers also point out that most of the results presented in the literature rely upon the inclusion of firms with small- and micro market capitalization. The authors argue that this is not the case in practice as they are generally not a part of the investment universe because of the relatively high transaction costs and the issue of illiquidity. The aim is to design a theoretically valid rationale upon which investment decisions can lie. In conclusion, the authors somewhat criticize the frequently cited "Factor Zoo" and the possibilities it entails. They mean that it is essential that the developments are implementable asset pricing models, arguing that almost no models pass the "real-world test".

The presented literature is only a compilation of the most relevant concerning the purpose of the thesis. The subject is complex, as reflected by the different objectives presented in the papers.

3 Theory

This part of the thesis will introduce the fundamentals of factor modelling in the context of asset pricing. It will begin with a brief overview of the risk-return relationship represented by the Capital Asset Pricing Model, followed by a recap of the development of factor models through time. Further on, a walk-through of the relevant machine learning approaches for factor modelling will be provided to accompany the financial-theoretical reasoning. This part will cover the main types of filter and embedded feature selection algorithms that can be considered plausible when creating factor models for asset pricing. Note however that the theory of each introduced model will not be covered.

3.1 Asset Pricing

The purpose of asset pricing can be generalized into two typical objectives. Within academia, the subsequent development of the factor model is an essential part of the ongoing research on comprehending market movements. Within investing, factor models for asset pricing bear the function of being both a predictor and an evaluation method. The model is used to determine the level of risk an asset possesses and to reckon whether the asset should be considered cheap or expensive. With these purposes in mind, a generalization may be made about the objective of asset pricing. The objective of asset pricing is to predict the value of an uncertain stream of cash flows(Cochrane 2005) [1]. The intent is thus to compute the value of an asset at a specific time, with a specific payoff structure. The future payoff is a random variable, but in the light of probability

theory, estimation of different scenarios and outcomes is possible. Note that the payoff structure differs for dissimilar types of assets. In the case of common stocks, the payoff is a function of the stock price in the future and possible dividends

3.1.1 Capital Asset Pricing Model

As mentioned, pricing of assets is done for several reasons, one of the main reasons being to estimate risk. When estimating risk, a commonly instructed and used model is the CAPM. The CAPM was initially developed from modern portfolio theory by Markowitz in 1952 [19]. The model estimates the return of an asset by its sensitivity to the market factor, and the proxy for risk is, therefore, the market risk premium. The market risk premium is the difference between the expected return on the market portfolio and the risk-free rate, often proxied by the one-month US Treasury bill rate(Fama & French 2004) [20]. The value of beta, called the factor loading, in the expression displays the asset's sensitivity to market movements. The formula for the CAPM is presented below.

$$E(R^i) = \gamma + \beta_i(E(R^m) - \gamma), \quad i = 1, 2, ..., N \tag{1}$$

Summarizing the CAPM, the model captures the securities exposure to the systematic risk. The model is still today often used as a benchmark-model and the theory behind the model is considered fundamental in the field of asset pricing. Despite this, empirical evidence reveals that the model is flawed. To emphasize is that it is regarded fair rather than distrustful that model does not operate flawlessly. Had the model performed perfectly, the research on the subject of asset pricing would likely not have been continued. One approach to handle the issue of anomalies is to add other factors to the model that may proxy for risk. The aim is thus to create a model based on firm characteristics that seem to function as a proxy for exposure to systematic risk in light of empirical evidence. The factor loadings represent the assets sensitivity to the factor in question. The *factor-model* would be expressed as follows

$$E(R^i) = \gamma + \beta_{i,a}\,\lambda_a + \beta_{i,b}\,\lambda_b + ..., \quad i = 1, 2, ..., N \tag{2}$$

Where, the betas are defined as the coefficients from the following time series regression

$$R^i_t = \alpha_i + \beta_{i,a}\,f^a_t + \beta_{i,b}\,f^b_t + ... + e^i_t, \quad t = 1, 2, ..., N \tag{3}$$

The factor model consists of multiple explanatory factors and is, therefore, directly a transformation of the single-beta Capital Asset Pricing Model. It assumes a linear relationship between the factors and the expected return. The factors chosen are variables that have predicted average returns well on past evidence, and are therefore assumed to capture risk premiums adequately. As the model includes several explanatory variables, the model may provide investors with more specific descriptions of the security's risk and return relationship. With the aim of asset pricing being to detect and evaluate risk, this attribute thus is valued highly.

3.1.2 Multiple Factor Models

Fama and French exemplifies the approach of composing factor models based on the CAPM(Fama & French 1993) by constructing a multi-factor model such as the three-factor model

$$E(R^{ei}) = \alpha + \beta_{1,i}(E(R^m - R^f) + \beta_{2,i}E(R^{SMB}) + \beta_{3,i}E(R^{HML}) + e_i, \tag{4}$$

The first model, developed in 1992, is a three-factor model that consist of the same market factor as in the CAPM, a factor representing the return on a portfolio of stocks in excess of the return on a portfolio of large stocks(frequently mentioned as Small Minus Big, thus *SMB*), and a factor representing the return of a portfolio of stocks with high book to market ratio in excess of the return on a portfolio of stocks with a low book to market ratio(frequently mentioned as High Minus Low, thus *HML*). The five-factor model, developed as recently as in 2015, expands the three-factor model and consists of the same factors as well as the difference in expected return between robust operating profitability and weak operating profitability firm portfolios(Robust Minus Weak, thus *RMW*). It also presents the difference in expected return between conservative investment and aggressive investment firm portfolios(Conservative Minus Agressive, this *CMA*) [2]. The Fama and French five factor model is expressed as follows

$$E(R^{ei}) = \alpha + E(R^{ei}) = \alpha + \beta_{1,i}(E(R^m - R^f) + \beta_{2,i}E(R^{SMB}) + \beta_{3,i}E(R^{HML}) \\ +\beta_{4,i}E(R^{RMW}) + \beta_{5,i}E(R^{CMA}e_i),$$

It is essential to underline that Fama and French's modeling captures an implied aspect of factor modelling in general, namely the use of factors as proxies of risks, rather than viewing the factors as explanatory of risk. The view is that *SMB* , *HML*, *RMW* and *CMS* may proxy for hard-to-measure, more fundamental risk factors, and that they are not themselves necessarily relevant risk factors. This assumption is based on historical correlation between the variables and return on assets. The Fama and French's models and their variants are a big part of the past and the ongoing research within asset pricing. The models are considered to be the two most famous developments of the CAPM. Furthermore, the models are frequently used as benchmarks in research when evaluating new models(Fama & French 2015) [10].

With extensive data sets and rapid computers, new possibilities for developing factor models exist. In 2016, Harvey et al. identified more than 300 factors that showed power in predicting expected asset return [**?**] . These factors are referred to as the "Factor Zoo", dubed in 2011 by John Cochrane. In his paper, Cochrane challenges researchers to find which factor independently explains return instead of researching new potential factors.

3.1.3 Factor models and "big data"

Recall the discussion of the future price of an asset being a random variable. This fact has implications for the use of factor models for asset pricing. The performance of selecting variables that will proxy for a risk factor is not an easy performance. A central problem with statistical modelling in general, is that when creating models in the

form of regression, a prominent issue is that correlation does not always imply causality. When using empirical approaches for composing multi-factor models such as the Fama and French's, the unique factors representing a risk-return relationship may be patterns that have happened merely by chance. This aspect is very inconvenient when using the models for asset pricing since the factor-loadings resulting from the regressions may be directly misleading and perhaps even a result from a change in share price rather than the opposite. This problem is, as mentioned, an overall difficulty for all statistical methods and is therefore thought of as an aspect that should be taken into account when analyzing newly developed models.

When overseeing this general problem with factor models, the issue concerning which factors that should be chosen can be somewhat solved with the help of new statistical methods. This notion invites the discussion on data mining and feature selection methods in the field of asset pricing. Today, there is almost an abundance of available data for nearly every publicly listed security. In the light of the prominent development of new techniques that can handle extensive data settings, researchers suggest that the use of Machine learning methods is reasonable for screening for factors that have power in predicting asset returns.

Machine learning methods are getting more and more popular as time goes on in economic and econometric analysis. During the last few decades, the amount of collected data from enterprises has increased drastically, both in terms of the number of observations that are collected but also the number of variables that are treated. These large sets of data can be useful in various amount of fields, including medicine, health-care, industrial commerce, advertising and more. However there are many poised challenges in order to gain valuable information from them. The arisen problems often revolves around the computational time which can be immense for large data sets with great amount of features. Another challenging problem to be aware of is the curse of dimensionality, which encompasses all kinds of bad effects when working with data in high-dimensional space due to the exponential increase in data size which creates sparsity in the data and difficult handling of Machine Learning algorithms [21].

By scientific advancements, researchers have tackled these issues with big data through reducing the dimensionality. This basically refers to the process of reducing the number of features in the data while keeping as much of the variation in the original data set as possible. Dimension reduction techniques can be divided into two categories. *Feature extraction* and *feature selection* [22]. Feature extraction is the idea of projecting the data onto a lower-dimensional space, and thus creating new kinds of features. Feature selection, on the other hand, aims to preserve the features as is. Rather its focus lies on identifying irrelevant or excess features from the data set. It can also be used to find information regarding the relation between the output values and the original features [23].

A typical data set that consists of p features, there are a maximum of $2^p - 1$ possible feature subsets, with the exception of really small data sets. Thus when working with high dimensional data, an exploratory search over all possible subsets is not feasible. This creates need for machine learning algorithms to handle these issues. Feature selection algorithms is a compilation of three different categories of methods: *Filter methods*, *wrapper methods* and *embedding methods*.

3.2 Filter methods

Filtering techniques are generally used as a preprocessing step. The selection of features is thereby independent of any machine learning algorithms. Instead its makes use of various statistical tests for their correlation with the output variable.

3.2.1 F-Score

By running an univariate linear regression test, one can find the cross correlation between each regressor and the target using the following equation

$$\frac{(X - \mu_X)(y - \mu_y)}{\sigma_X \sigma_y} \tag{5}$$

where X, y is the value data for each feature and response, μ_X, μ_y is the corresponding mean value for feature values and responses. σ_X, σ_y represents the corresponding standard deviations. The outputs later is converted to an F-score and its corresponding p-values. Based on the F-test, the model can estimate the degree of linear dependency between two random variables. The method for this is called *fregression* and is wholly extracted from the library in Python called *Scikit-learn* [24].

3.3 Embedding methods

Embedded methods select features in the process of another task. The methods that will be covered are of penalized regression type. The following section will explain the special case of penalized regression which is called LASSO.

3.3.1 LASSO

In papers covering the topic of the factor zoo, both Freyberger et al. (2016) and Feng et al. (2017) utilise LASSO for variable selection and prediction. LASSO, *Least Absolute Shrinkage and Selection Operator* is a type of embedded method used in feature selection, fabricated by Tibshirani [7]. The LASSO regression problem is stated as

$$\hat{\beta}_{lasso}(\lambda) = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2 + \lambda ||\beta||_1. \tag{6}$$

The first term of the optimization problem estimates the linear regression coefficients of ordinary least square (OLS). The second term explains the LASSO penalty factor which constrains β through L1 penalization. λ is the parameter that controls the penalization of β through cross-validation. Beginning with the simple linear regression model in the cross-section setting with p independent variables

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \epsilon_i. \tag{7}$$

In the traditional least squares regression, estimated parameters are chosen to minimize the residual sum of squares(RSS), where RSS is

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{8}$$

In cross-validation, the validation data is chosen at random and the model will be bootstrapped several times for the same λ, bringing forth a new validation set for each iteration. However, the problem arises when p is large, which makes regularization a good use to shrink parameter estimates to zero. The Figure 1 below visualizes how the OLS coefficients are shrunk to fit the constraint of the LASSO, causing one of the coefficient to be pushed to zero.

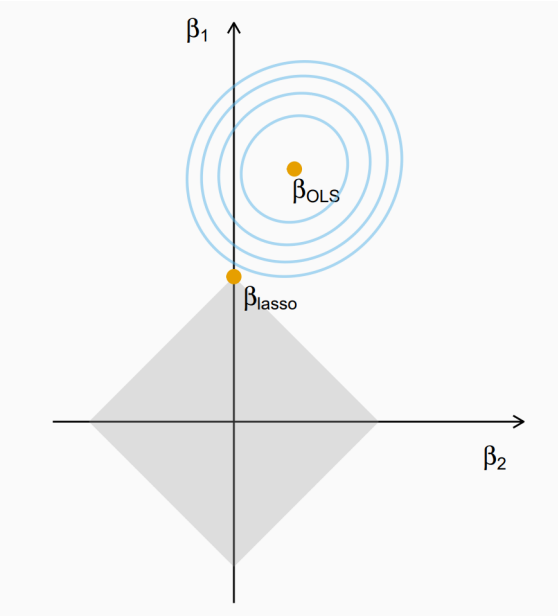

Figure 1: The least squares RSS(*residual sum of squares*) is minimized for β_{OLS}, if a constraint is added, then the RSS is minimzed by the closes β possible that fulfills that constraint. The blue lines are the contour lines for the RSS.

For feature selection, LASSO can be seen as accessing larger values of the $\hat{\beta}$ coefficients. Larger values of a certain coefficient implies that the corresponding feature is important.

Further improvements can be done to sharpen the algorithm further. By choosing a λ a bit larger, more coefficients could be pushed to zero. Assume that cross validation is done with C folds and tested for M choices for the hyperparameters. Let $e \in \mathbb{R}^M$ be the average MSE across all C folds and $s \in \mathbb{R}^M$ be the corresponding standard deviations. $\lambda_{\min}$ corresponds to the minimal average MSE at index $i_{\min}$ which leads to a better choice of hyperparameter is found by solving the problem

$$i_{1se} = \arg\max_j \{ \boldsymbol{e}^{(j)} | \boldsymbol{e}^{(i_{min})} \leq \boldsymbol{e}^{(j)} \leq \boldsymbol{e}^{(i_{min})} + \boldsymbol{s}^{(i_{min})} / \sqrt{C} \text{ and } \lambda^{(j)} \geq \lambda_{min} \}. \tag{9}$$

Here $\lambda^{(j)}$ represents the j-th hyperparameter tested during cross-validation. The λ_{1se} refers then to the i_{1se}-th hyperparameter among the choice of M hyperparameters. The idea is to choose the final model with the least non-zero parameter.

4 Data

This part will clarify what input data the thesis lies upon and the sources of data.

The data is comprised of many different firm characteristics. The raw data set is collected from the database Capital IQ and consists of 108 observations and 123 features [25]. The data consists of securities that are listed on NYSE(*New York Stock Exchange*) or NASDAQ and that currently are constituents of the S&P500 index. The reason for picking the S&P500 index is because of its function as a representative for the U.S stock market which is because of its size. Also, the data for the firm characteristics are easily accessible and there are only few missing values, making the index an informative source for the aim of the thesis. For some of the existing features, percentage change is added over the previous period. That is due to the anticipation that in some cases the percentage change from the previous quarter could become more significant than the quantity for a single quarter. In addition to the data extracted from Capital IQ, we added the factors for the Fama and French Five factor model to the data set [26].

The data took the form of $\{X^i, y^i\}$ where X^i is the feature matrix for company i and the rows of X^i represented time while the columns represented the features. Rows corresponded to time in $\{2011Q2, ..., 2021Q2\}$. Together one can average each of these data matrices over a set of companies to make a portfolio.

The response variable or output variable is based upon the same logreturn variable $(log(p_t^i) - log(p_{t-1}^i))$ that Green et al. (2017) and other researchers after him have used in his similar analysis about asset returns.

In order to be able to fully work with the data, pre-processing needs to be done. This includes basic exploratory data analysis through finding missing values, or transforming feature values based on convenience. The following chapter will explain these problems and why it's necessary to take them into account.

4.1 Exploratory data analysis and data preprocessing

Exploratory data analysis, or data mining, is a tool to find patterns, relationships and trends in larger data sets [27]. Before applying relevant machine learning algorithms and techniques on the data, it's necessary to inspect the the current scaling of features, the correlation between them and missing values.

The scaling of features can be determined through analyzing the means and variances of the features. Depending on the results it could be necessary to either keep the data as it is, or standardize/normalize it in order for the algorithm to not favour features with large scale variations.

In the data set, problems could arise due to multicollinearity between features. By analyzing this, some features can be kept in mind during the actual regression analysis. Correlations were mentioned earlier as being a filter method, but it's generally always necessary to include when you perform data pre-processing. The correlation values is

calculated through Pearson's correlation criteria. Features that are highly correlated, $\sim 90\%$ or more means that no additional information about the data is gained. Thus these features will only add noise to the model which could make it harder for machine learning algorithms to classify the data.

5 Methodology

This section presents the operating procedure used to promote the purpose of the thesis. An emphasis is made on clarifying the usage of the data and a presentation of different models will be demonstrated and evaluated for finding the most important features that explains asset returns. All analysis will be programmed in Python, and the relevant models are part of the large statistical library called *Scikit-learn* [24].

5.1 Portfolio Arrangement

The data will be divided upon three different portfolios. Each based on relative market cap in the S&P500-index. The "small" portfolio(Portfolio 3) consists of data for ten firms in the lowest segment sorted by market cap in the S&P 500-index. The "medium" portfolio (Portfolio 2) consists of data for ten firms with medium relative market cap in the middle segment of all firms sorted by market cap. The "largest" portfolio(Portfolio 1) consists of the top ten firms with highest market cap.

The constituents of the different portfolios can be found in the table 1 below.

Portfolio 1 (Large relative market-cap)	Portfolio 2 (Medium)	Portfolio 3 (Small)
Johnson & Johnson (NYSE:JNJ)	AMETEK, Inc. (NYSE:AME)	Citrix Systems, Inc. (NasdaqGS:CTXS)
JPMorgan Chase & Co (NYSE:JPM)	Nucor Corporation (NYSE:NUE)	Loews Corporation (NYSE:L)
The Home Depot Inc. (NYSE:HD)	Old Dominion Freight Line, Inc. (NasdaqGS:ODFL)	WestRock Company (NYSE:WRK)
Berkshire Hathaway Inc. (NYSE:BRK.A)	The Allstate Corporation (NYSE:ALL)	Comerica Incorporated (NYSE:CMA)
Nvidia Corp (NasdaqGS:NVDA)	American Water Works Company, Inc. (NYSE:AWK)	Tapestry, Inc. (NYSE:TPR)
Alphabet Inc. (NasdaqGS:GOOG.L)	Public Service Enterprise Group Incorporated (NYSE:PEG)	Lumen Technologies, Inc. (NYSE:LUMN)
Amazon.com, Inc. (NasdaqGS:AMZN)	TransDigm Group Incorporated (NYSE:TDG)	Juniper Networks, Inc. (NYSE:JNPR)
Apple Inc. (NasdaqGS:AAPL)	Zebra Technologies Corporation (NasdaqGS:ZBRA)	W. R. Berkley Corporation (NYSE:WRB)
Microsoft Corporation (NasdaqGS:MSFT)	The Williams Companies, Inc. (NYSE:WMB)	The Williams Companies, Inc. (NYSE:WMB)
Tesla, Inc. (NasdaqGS:TSLA)	Copart, Inc. (NasdaqGS:CPRT)	A. O. Smith Corporation (NYSE:AOS)

Table 1: The constituents of each of the three portfolios.

5.2 Designing the algorithm to predict asset returns

As the data section presents, a data set is used comprised of 30 companies with 123 features, which are listed in the Appendix. The idea is that each feature should be handled and assumed as a random variable. Thus, instead of performing each method on a single asset, companies will be assorted into portfolios depending on their size which in this case is determined by market cap. The portfolios are made by averaging each component through all constituents to the portfolio, i.e the return on portfolio 3 for a specific date is the average return of the constitutions of the portfolio.

An ensemble of different methods will be used in order to evaluate feature selection. The models are presented in the table 2 below. The results of these will only be made to emphasize the confidence of the lasso selection stability. Thus, we do not demonstrate these methods further.

Models	Linear Regression	Decision Tree	Random Forest	XGBoost	KNearest-Neighbors	Lasso

Table 2: The ensemble of models used to gain confidence in feature selection.

The overall pipeline showcasing the order of methodology used is presented in the Figure 2 below. The pipeline encompasses all models used to arrive to the solution of the research question.

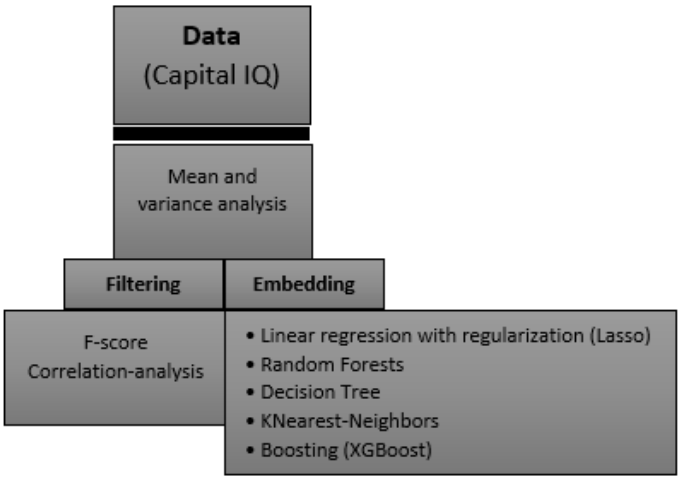

Figure 2: Pipeline containing all models used.

5.3 Model training and evaluation

The evaluation on all models except for the Lasso is based on train-test splitting where each data sample is divided into 80% training and 20% test sets. This involves splitting the data set into two subsets, where the first subset is used to fit the model and the other is used to evaluate the fitted model.

Evaluation for the Lasso is based on k-folded cross-validation. This means that the training set is split into k smaller sets. The model is then trained using $k - 1$ of the folds as training data and the remaining part of the data acts as a validation set for the resulting model. A basic sketch to how this works is showed in Figure 3 below.

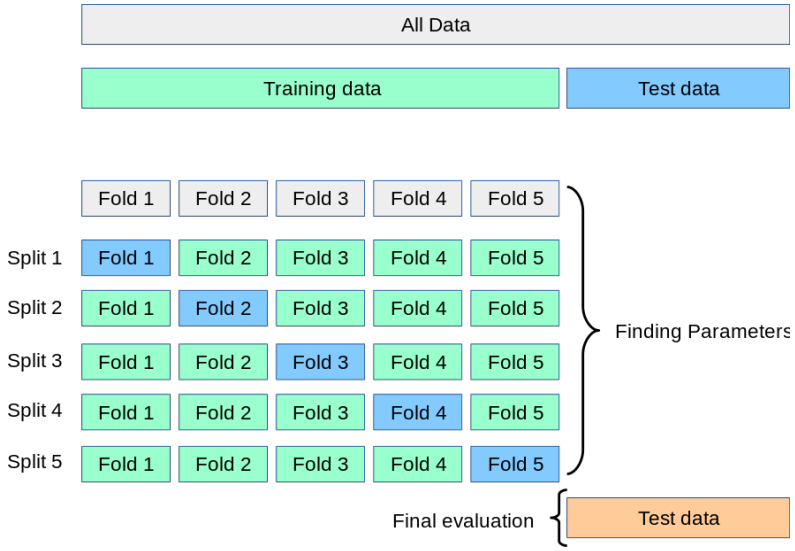

Figure 3: Design of 5-fold cross-validation. Figure extracted from the *Scikit-learn*-website. [28]

5.4 Extensions to the Lasso algorithm

At the moment, there is a single selection of features, stemming from one run of sparse multi-class logistic regression. Bootstrapping is a method that can be used to further explain how stable the selection is. Recall the improvements of the lasso algorithm stated in the theory section. By creating M bootstrap samples from the filtered data set, one can perform feature selection on each of these samples. By recording how often each feature got selected for each class, one can obtain a histogram where larger peaks indicate that a feature was chosen more often. How often a feature gets selected relative to M can then be perceived as selection stability.

Recall the LASSO regression problem from equation 6, where we stated that finding a certain λ will give the optimal feature selection. Also recall that we stated that we decided to pick the λ that had mean squared error within one standard deviation of the minimal one $\lambda_{\min}$. After picking $\lambda^{(i_{1se})}$, the data set is bootstrapped with 70% of the data without replacement and a lasso model was fitted with using $\lambda^{i_{1se}}$. For each fitted model, the selected features were recorded, i.e. all features i with $\beta_i > 0$, whilst also all β_i values were recorded. This was done for 1000 iterations.

These extensions but also the improvements regarding the theory of the Lasso stated in the problem 9 makes this Lasso model unique compared to the predecessors of former research stated in the literature review.

After a linear regression model is fitted onto the data, the most probable features will be added. Consensus is made by measuring how well the model fits the output data through calculating the R^2-score and the RMSE values. This was done through using

the following equations

$$R^2 = 1 - \frac{u}{v} \tag{10}$$

$$u = \sum^{n} (y_{pred} - y_{true})^2 \tag{11}$$

$$v = \sum^{n} (y_{true} - \bar{y}_{true})^2 \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)} \tag{13}$$

By those means, this serves as an indication of how important the feature is in predicting the response variable, i.e. average asset returns. Since if a feature is always is chosen, it must have some bearing upon the prediction, and vice versa, allowing the model to justify the resulting feature selection with confidence.

6 Results

In this section, at first, results will be provided regarding how the data was pre-processed through scaling of features. This follows by presenting the filter-methods being used in order to gain information on the feasible factors towards explaining asset returns. Extensive results are showed of the Lasso alongside the results from the other models. The basic idea is that an evaluation takes place which brings forth which model performed the best. The chosen model is the one the outperforms the others'. A final presentation of selected features but also the corresponding fitting results by this model is then showed. A quick comparison between the other models leads to confidence in the picking state, i.e. selection stability. Lastly a quick comparison between the benchmark of Fama French Five Factor Model and the proposed final model is presented.

6.1 Scaling of features

Before applying the respective models, it is important to inspect the scaling of features, since otherwise large scale features need smaller coefficients, which for example is favoured by the lasso since less regularization is applied. Both the means and variances differs wildly, and the features had to be standardized so that features doesn't favour these features.

Standardization occurs through using the following equation:

$$z_{\text{standard}} = \frac{X - \mu}{\sigma} \tag{14}$$

where X is the observed value, μ is the mean of the specific columns and σ is the standard deviation of the specific columns. Together the formula produces standardized values for each data point in the feature matrix.

6.2 F-Scores

The F-scores for portfolio 1 can be seen ordered from highest to lowest in Figure 5.

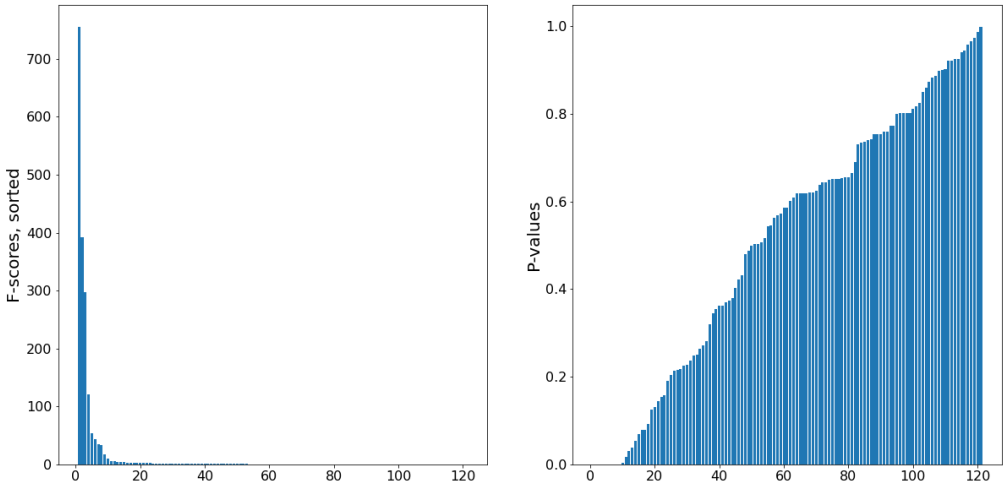

Figure 4: To the left: F-scores for the features ordered from highest to lowest. To the right: p-values in the same order as the left figure.

It's important to note that the numbering is not consistent with the the factor numbering in the data. However, the p-values indicate that about 15 features seems relevant for feature selection. It's also of essence to keep at hand when performing the embedding, and if these features should match the embedding methods it would account for better prediction. The results for the two other portfolios also seems consistent when addressing the number of relevant factors. These can be found in the appendix.

6.3 Lasso Regression

At first, Lasso regression on Portfolio 1 will be addressed. Lasso seeks to select features that have a high potential for omitted variable-bias. The number of factors selected by the Lasso can be interpreted as a measure for the dimensionality on the asset pricing model.

When testing cross validation with default parameters, one can see the number of selected features using both λ_{min} and λ_{1se} presented in the table 3 below.

Number of selected features	Portfolio 1	Portfolio 2	Portfolio 3
λ_{min}	41	38	48
λ_{1se}	38	37	44

Table 3: Number of selected features using both α_{min} and α_{1se}.

This reveals that this model propose that there are around 40 relevant factors feasible for explaining asset returns. Further on, we made extensions to the lasso algorithm. By bootstrapping the data, aside from the current single selection off features stemming from one run of sparse logistic regression, one can find a more stable feature selection.

Using the bootstrapping algorithm presented in the methodology section we can arrive at a histogram showing recording how often each feature is chosen across all bootstrap samples. Larger peaks indicate that a feature was chosen more often.

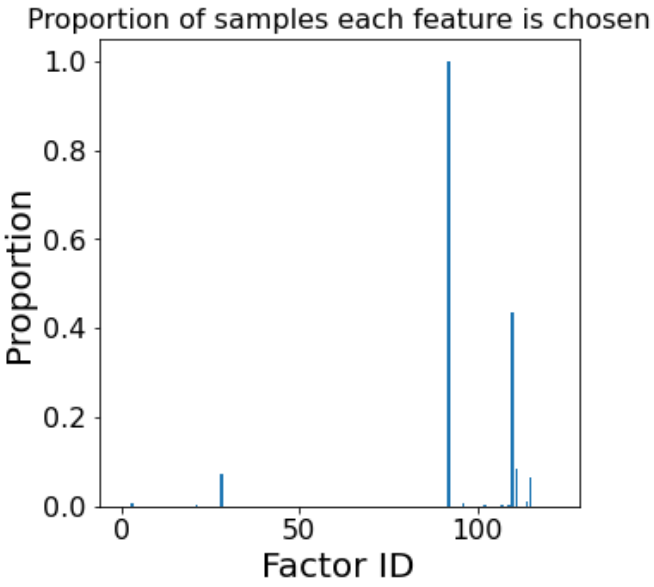

Figure 5: Proportion of samples each features is chosen by using the Lasso on portfolio 1. Factor ID stands for the corresponding label in the data set. The important thing is to note the peaks since they convey how many and how frequently the features are chosen.

Here we can see that among the 40 factors firstly picked through one run of sparse logistic regression, we instead regularized the model further and showed that only a handful of those picked seems relevant enough. The addressing of the selected features will be covered after model evaluation has been done. Similar histograms for Portfolio 2 and 3 can be found in the Appendix.

6.4 Model comparison & evaluation

In the table 4 below, fitting results are shown on portfolio 1. These are fitting results based on the models proposed through bootstrapping and ordinary train-test-splitting.

Model	Lasso	Linear Regression	Decision Tree	Random Forest	XGBoost	KNearestNeighbors
RMSE	0.0097	0.1422	0.0177	0.0213	0.0213	0.0213
R^2	0.8090	0.9961	1.0	0.9519	0.9996	0.5476

Table 4: RMSE and R^2 values for the corresponding models used on portfolio 1.

Model	Lasso	Linear Regression	Decision Tree	Random Forest	XGBoost	KNearestNeighbors
RMSE	0.0109	0.0261	0.0180	0.0072	0.0071	0.0226
R^2	0.7815	0.4203	0.7515	0.8239	0.8909	0.0985

Table 5: RMSE and R^2 values for the corresponding models used on portfolio 2.

Model	Lasso	Linear Regression	Decision Tree	Random Forest	XGBoost	KNearestNeighbors
RMSE	0.0081	2.8862	0.0115	0.02333	0.0215	0.0281
R^2	0.8624	-0.0262	0.7672	0.5349	0.6337	0.2762

Table 6: RMSE and R^2 values for the corresponding models used on portfolio 3.

Based on RMSE measurements, the proposed Lasso model performed the best. The R^2-values seems rather contradictory for which some models can reach both high RMSE-values and high R^2. Since the main model, the Lasso, seems to outperform the others' we are presenting the selected features by this model.

6.5 Presentation of selected features

Table 7: Top 10 chosen features through using the Lasso model on Portfolio 1.

Selected features	[% change in return on total assets, change in capital turnover, change in rsi, change in price target, change in total assets, rsi, change in P/sales, change in sales to price, price rel 52 week high, change in market cap]
Number of selected features	Top 10 selected
R^2 score	0.8090
RMSE score	0.0081

Comparing the 10 most chosen features from the ensemble of models used we gain confidence upon which factors seems to be the most relevant. The matches of these are presented in the table below.

Number of matches	Random Forest	Decision Tree	XGBoost	F-Score	KNearestNeighbors	Linear Regression
Lasso	4	4	3	6	5	0

Table 8: Matches between selected top 10 features of the lasso model and the other models on portfolio 1.

Results for Portfolio 2 and 3 can be found below.

Table 9: Top 10 chosen features through using the Lasso model on Portfolio 2.

Selected features	[change in price target, change in 4m lagged pp& e, price rel 52 week high, change in TEV/total revenues, rsi, change in P/sales, change in TEV/EBIT, change in sales to price, change in market cap, change in price target]
Number of selected features	Top 10 selected
R^2 score	0.7815
RMSE score	0.0109

Number of matches	Random Forest	Decision Tree	XGBoost	F-Score	KNearestNeighbors	Linear Regression
Lasso	5	3	4	4	3	0

Table 10: Matches between selected top 10 features of the lasso model and the other models on portfolio 2.

Table 11: Top 10 chosen features through using the Lasso model on Portfolio 3.

Selected features	[P/sales, change in sales, change in 4m lagged pp&e, rsi, change in P/sales, change in market cap, change in 52 week high, change in return on total assets, operations accurals, Mkt-RF]
Number of selected features	Top 10 selected
R^2 score	0.8624
RMSE score	0.0097

Number of matches	Random Forest	Decision Tree	XGBoost	F-Score	KNearestNeighbors	Linear Regression
Lasso	4	4	6	6	3	0

Table 12: Matches between selected top 10 features of the lasso model and the other models on portfolio 3.

6.6 Benchmark results

Lastly the benchmark-model containing only the Fama-French 5 factors used in regression for each Portfolio. The regression outputs for each model is seen in the table 13 below.

FF5	Portfolio 1	Portfolio 2	Portfolio 3
RMSE	0.0215	0.0204	0.0206
R^2	0.1638	0.1747	0.2491

Table 13: RMSE and R^2 values for the FF.

It can be seen clearly that linear regression on the benchmark clearly doesn't perform well compared to the proposed models.

7 Discussion

The section begins with an interpretation of how well the models perform with respect to the criteria presented in the Methodology section. Thereafter, a brief discussion about the composition of the models in the context of the results presented in the literature review is presented. Lastly, the arrangement of the portfolios is examined, followed by a discussion regarding the opportunities of alternative analyzes concerning the bias-variance trade-off. This is accompanied by suggestions regarding future work on the subject.

7.1 Model Evaluation

As presented in the section explaining the methodology, the models' fit is assessed according to R^2 and RMSE. These measures reflect the algorithm's ability to produce models that perform, thus reflecting if the algorithm is appropriate or not. A high R^2 value and a low RMSE value are desired when building a prediction model.

Portfolio 1 gets a R^2 of 0.8090 and a RMSE score of 0.0097. Portfolio 2 gets a R^2 of 0.7815 and and a RMSE score of 0.0109. Portfolio 3 gets a R^2 of 0.8624 and a RMSE score of 0.0097. This implies that the algorithm produces working models for all portfolios based on these measures, but that the best model is the one fitted for portfolio three consisting of companies with a lower relative market cap. It should be noted, however, that all models perform noticeably well. The results demonstrate that all three models share between 3 and 6 factors with the other machine learning models. Just as in the paper of Peng et al.(2021), these results serve as a measure of consensus between the methods and thus demonstrate the strength or weakness of the model generated by the algorithm. It is desirable to get as many matches as possible, and we see that the algorithm for portfolio 3 was/is best concerning the consensus estimate.

Altogether, the algorithm selects 18 unique factors that predict asset returns. 14 of these represent a value change during a period and not a unique value for a specific date. The result confirms the previously proposed anticipation that in some cases, the percentage change from the previous quarter could become more significant than the quantity for a single quarter.

There are only three factors that all three portfolio-models share. These are *"change in P / sales"*, *"rsi"* , and *"market cap changes"*. The models for portfolio 1 and 2 share three factors, namely *"change in price target"*, *"change in sales to price"*, and *"price rel 52 week high"*. One factor is shared for portfolios 2 and 3, namely *"change in 4m lagged pp&e"*. This result is consistent with that of Green et al.(2017). The fact that portfolios 1 and 3 do not share any factor can be considered strange or reasonable depending on the reasoning. On the one hand, the ideal is to find factors that can estimate returns on a universal level, at least predict asset returns for common stocks within the S&P500 index. On the other hand, firms with different market capitalization may be sensitive to dissimilar changes in firm characteristics because of the varied implications of the characteristics.

In addition to whether the models are good or bad predictors, it is relevant to examine whether they are reasonable or not. In this context, reasonable is defined as if the factors can be considered valid with respect to both financial and accounting theory, as well as the psychology of investing concerning if the factors can impact how investors value the asset's trading price. In the light of the discussion about reasonableness, it is essential to revisit the Efficient Market Hypothesis briefly mentioned in the introduction of the thesis. Theoretically the stock market is efficient. The price of a security on the market must reflect supply and demand, which is determined according to the consumer's access to information. Furthermore, all information must be equally accessible to all and all available information is thought to be accounted for in the price at which an asset is valued. Thus, predicting asset returns using factor models is not possible in theory, making the evaluation of the models somewhat complex.

If we look past the argumentation about the efficient market and instead shed light on the factors that the models present, it can be stated that they are reasonable. However, the remaining, and always present, problem with correlation not always being causality, and the risk of observing a relationship that is only due to chance, is vital to recognise.

Given the criticism highlighted in literature reviews on the Factor zoo, it is appropriate to point out that the results we have obtained through our method can not or should not be assumed to be perfect or even close to perfect. As mentioned in the introduction, the data, methodology, and results are developed under limited circumstances, especially regarding time and knowledge in relation to other researchers on the subject. Thus, it should be stated that we contribute to the research that takes place in the subject, but only to a certain extent

7.2 Portfolio Assessment and Bias-Variance trade-off

The issue behind focusing on portfolios rather than individual assets lies in the existence of bias-variance trade-off. However selecting portfolios rewards the analysis with being less prone to missing data problems, and having more stable betas. To reduce the bias one would rather select many more portfolios, which will minimize favouring of factors.

The idea of sorting the portfolios by the companies' market cap size, could not also be the most favorable. Instead, which implies robustness, one can make different sets of portfolios that depends on e.g. size, book-to-market ratio, net issuance, size and accruals or size and momentum. One could also include separability by sectors, of which some assets form a part of the health-care sector and some belong to the IT-sector. There are many ways to go about making a justifiable portfolio that reflects both the market and individual assets as a whole.

7.3 Comparison against Benchmark

The Fama and French five-factor regression show that this model, concerning the calculated statistical criteria, is not suitable for predicting returns these portfolios. However, it should be said that this model should primarily function as a proxy for risk where the beta values should define the sensitivity to the particular factor. Thus, the factor-loadings are relevant when evaluating the Fama and French model results. Of course, the model's fit is also essential, and it can be stated that the results obtained in our thesis imply that the model is not suitable for neither of the three portfolios. This is also consistent with the fact that the algorithm did not include any of the FF5 factors for any model for any portfolio. Therefore, our result means that these factors are not relevant for estimating asset returns.

7.4 Future work

With respect to the available research on creating sparse factor models for asset pricing, Cochrane (2011)'s question regarding which firm characteristics provide independent information about asset returns remains unanswered. In our thesis, we have attempted to contribute to the research by simplifying the methods. Like Cochrane and his followers, we have used machine learning approaches to extract factors from the factor zoo.

The fact of that the factors retrieved by the predecessors of the literature review were not possible to obtain through our data extraction methods makes it hard to compare between the different works. A more extensive research could be done in retrieving these missing factors and add them to our model to see how the results would change.

In the light of good research practice, it is not possible to draw any conclusions from our results as they still need to be verified in the context of new observations. One can yield more interesting results by including additional factors and analysing a larger random sample and over different markets.

References

[1] J. H. Cochrane, *Asset Pricing: Revised Edition*, 2005.

[2] A. M. Zvi Bodie, Alex Kane, *Investments*, 2020, vol. 12.

[3] W. F. Sharpe, "Capital Asset Prices: A Theory Of Market Equilibrium Under Conditions Of Risk," *Journal of Finance*, vol. 19, no. 3, pp. 425–442, September 1964. [Online]. Available: https://ideas.repec.org/a/bla/jfinan/v19y1964i3p425-442.html

[4] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304405X93900235

[5] C. R. Harvey, Y. Liu, and H. Zhu, "... and the Cross-Section of Expected Returns," *The Review of Financial Studies*, vol. 29, no. 1, pp. 5–68, 10 2015. [Online]. Available: https://doi.org/10.1093/rfs/hhv059

[6] J. H. COCHRANE, "Presidential address: Discount rates," *The Journal of Finance*, vol. 66, no. 4, pp. 1047–1108, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x

[8] A. Ang, *Asset Management: A Systematic Approach to Factor Investing*, ser. ACS Symposium Series. Oxford University Press, 2014. [Online]. Available: https://books.google.se/books?id=e5yzAwAAQBAJ

[9] M. M. Carhart, "On persistence in mutual fund performance," *The Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997. [Online]. Available: http://www.jstor.org/stable/2329556

[10] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304405X14002323

[11] J. Green, J. R. M. Hand, and X. F. Zhang, "The characteristics that provide independent information about average u.s. monthly stock returns," *The Review of Financial Studies*, vol. 30, no. 12, pp. 4389–4436, 2017. [Online]. Available: https://www.jstor.org/stable/48616726

[12] E. F. Fama and K. R. French, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, vol. 47, no. 2, pp. 427–465, June 1992. [Online]. Available: https://ideas.repec.org/a/bla/jfinan/v47y1992i2p427-65.html

[13] J. Freyberger, A. Neuhierl, and M. Weber, "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2326–2377, 04 2020. [Online]. Available: https://doi.org/10.1093/rfs/hhz123

[14] G. Feng, S. Giglio, and D. Xiu, "Taming the factor zoo," 2017.

[15] A. Belloni, V. Chernozhukov, and C. Hansen, "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 29–50, May 2014. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29

[16] B. Kelly, S. Pruitt, and Y. Su, "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, vol. 134, no. 3, pp. 501–524, Dec. 2019, publisher Copyright: © 2019 Elsevier B.V. Copyright: Copyright 2019 Elsevier B.V., All rights reserved.

[17] Y. Peng, P. H. M. Albuquerque, H. Kimura, and C. A. P. B. Saavedra, "Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators," *Machine Learning with Applications*, vol. 5, p. 100060, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266682702100030X

[18] P. F. P. . A. R. Söhnke M. Bartram, Harald Lohre, "Scikit-learn: Machine learning in Python," *Journal of Business Economics*, vol. 91, pp. 655–703, 2021.

[19] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952. [Online]. Available: http://www.jstor.org/stable/2975974

[20] E. F. Fama and K. R. French, "The capital asset pricing model: Theory and evidence," *Journal of Economic Perspectives*, vol. 18, no. 3, pp. 25–46, September 2004. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/0895330042162430

[21] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *Computational Intelligence and Bioinspired Systems*, J. Cabestany, A. Prieto, and F. Sandoval, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 758–770.

[22] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen, "Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 320–333, 2006.

[23] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1, pp. 131–156, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1088467X97000085

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 2825–2830, nov 2011.

[25] "Sp global: Accelerating progress." McGraw Hill Financial, 2022. [Online]. Available: https://www.spglobal.com/marketintelligence/en/

[26] "Description of fama/french 5 factors (2x3)." Kenneth R. French, 2021. [Online]. Available: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5_factors_2x3.html

[27] J. W. Tukey *et al.*, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A F-scores for Portfolio 2 and 3

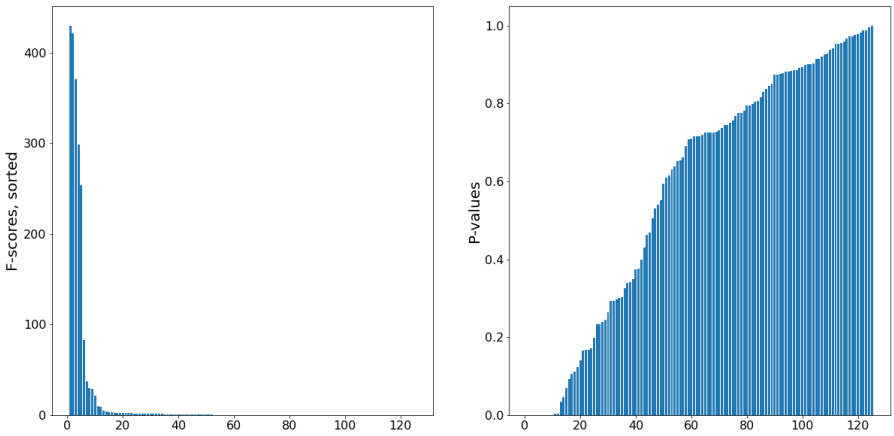

Figure 6: To the left: F-scores for the features ordered from highest to lowest. To the right: p-values in the same order as the left figure.

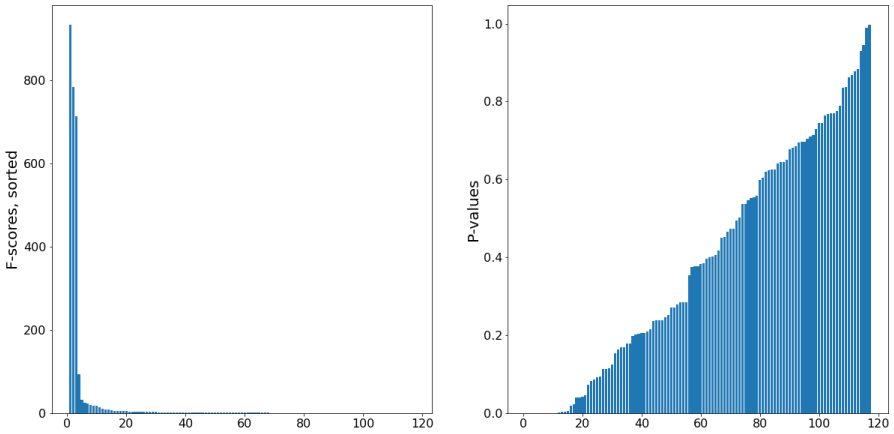

Figure 7: To the left: F-scores for the features ordered from highest to lowest. To the right: p-values in the same order as the left figure.

B Proportion of chosen features histograms

Histograms showcasing how often each feature is recorded in the bootstrapping algorithm.

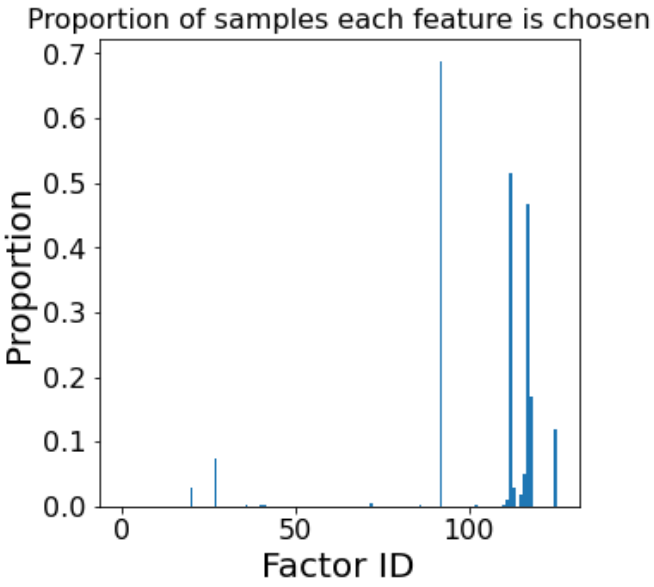

Figure 8: Proportion of samples each features is chosen by using the Lasso on portfolio 2.

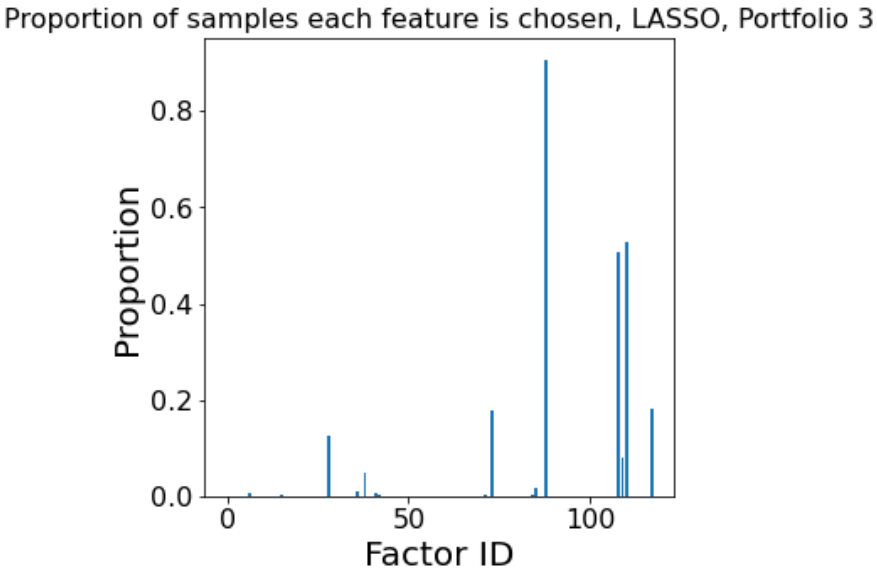

Figure 9: Proportion of samples each features is chosen by using the Lasso on portfolio 3.

C Data Dictionary

Name Of Feature	Definition Of Characteristic
earnings per share	Earnings Per Share
Market Cap	Market Capitalisation
Beta 5 years	5 Year Beta
Total assets	Total Assets
Change in total assets	Change in total assets
Sales to assets	Total Sales / Total Assets
cash	Cash And Equivalents
change in cash	Change in cash
short term investments	Short-term investments
change in total equity	Change in total equity
total equity	Total Equity
capital turnover	Total Revenue / Total Equity
capital intensity	Capital intensity
cash flow	Cash flow
Investments	Long-term Investments
total debt	Total Debt
leverage	Total Debt / Total Equity
gross profit	Gross Profit
sales	Total Revenues
gross profitability	Gross Profit / Total Revenues
year high(52 week high)	52 Week High
close price	Close Price
price rel 52 week high	Close Price / 52 Week High
total liabilities	Total Liabilities
financial liabilities	Financial Liabilities Fair Value
financial asssets	Financial Assets Carried At Fair Value Through Thrift
net operating assets	Total Assets - Total Liabilities - Financial Assets Carried At Fair Value Through Thrift +Financial Liabilities
net income	Net Income
Return on operating assets	Net Income / Net Operating Assets
Return on total assets	Net Income / total Assets
return on equity	Net Income / Total Equity
rsi	Relative Strength Index
rate of change 1 month	Close Price
sales to price	Price/Sales
sales and admin costs to sales	Selling General & Admin Exp.
bid-ask	Difference Between Bid Price and Ask Price
cash flow per share	Cash Flow Per Share
book value per share	Book Value / Share
cash/book	Cash Flow Per Share / (Book Value / Share)
operations accruals	Revenues earned or expenses incurred which affect net income
change in operating income	Change In Operating Income
change in sales	Change in Total Revenue
operating leverage	Change In Operating Income / Change in Total Revenue
shareholder equity	Total Assets - Total Liabilities
lagged shareholders equity	Lagged Total Assets - Total Liabilities
growth in common shareholder equity	((Total Assets - Total Liabilities) - (Lagged Total Assets - Total Liabilities)) / (Lagged Total Assets - Total Liabilities)
Capital expenditure	Capital Expenditure
Lagged capital expenditure	12 months lagged capital expenditure
growth in capital expenditures	(Capital Expenditure - Lagged Capital Expenditure) / Lagged Capital Expenditure
Total employees	Total Employees
lagged total employees	Lagged Total Employees
Capital expenditures and inventory	Inventory
operating income	Operating Income
operating profitability	Operating Income / Net Income
1y lagged total assets	1 Year Lagged Total Assets

Name Of Feature	Definition Of Characteristic
1y lagged debt	1 Year Lagged Total Debt
shares outstanding	Shares Outstanding on Balance Sheet Date
1y lagged shares outstanding	1 Year Lagged Shares Outstanding on Balance Sheet Date
inventory	Inventory
1y lagged inventory	1 Year Lagged Inventory
cash paid in taxes	Cash Taxes Paid
4m lagged cash paid in taxes	4 Month Lagged Cash Taxes Paid
PP&E	Property, Plant and Equipment
4m lagged pp&e	4 months lagged
8m lagged pp&e	8 months lagged
12m lagged pp&e	12 months lagged
4m lagged sales	4 months lagged
8m lagged sales	8 months lagged
12m lagged sales	12 months lagged
short term investments	Short Term Investments
total current assets	Total Current Assets
long term investments	Long-term Investments
short term borrowings	Short-term Borrowings
Current portion of long term debt	Current Portion of Long Term Debt
Long-term debt	Long-Term Debt
Short term debt issued	Short Term Debt Issued
total revenues	Total Revenues
Gross profit	Gross Profit
R&D EXP	Research and Development Expenditures
Depreciation & amort	Depreciation & Amortisation
Amort. Of goodwill and intangibles	Amortisation Of Goodwill And Intangibles
Total debt issued	Total Debt Issued
Income Tax Expense	Income Tax Expense
EBITDA	EBITDA
EBITA	**EBITA**
EBIT	EBIT
Total dividends paid	Total Dividends Paid
EBITDA-CAPEX	EBITDA - CAPEX
EBITDA margin%	EBITDA Margin
Total debt/equity	Total Debt / Equity
Capex as % of revenues	CAPEX In % Of Revenues
total short-term debt	Total Short-Term Debt
AVG broker recommendation	Average Broker Recommendation # 1 = Strong Buy 3 = Hold 5 = Strong Sell
Price target	Price Target
EST. Annual eps growth 5 yr	5 Year Expected Annual Earnings Per Share Growth
TEV/total revenues	Total Enterprise Value / Total Revenues
TEV/EBIT	Total Enterprise Value / EBIT
P/sales	Price / Sales
PEG ratio	Ratio Price / Earnings To Growth
Enterprise value	Enterprise Value
dividend yield	Dividend Yeild
Institutional holder total shares	Institutional Holder Total Shares
change eps	Change Earnings Per Share
change marketcap	Change In Market Capitalisation
change beta 5 years	Change 5 Year Beta
change total assets	Change Total Assets
change sales and admins costs to sales	Change Sales To Assets
change in total equity	Change In Total Equity
change in capital turnover	Change In Turnover
change in capital intensity	Change In Capital Intensity
change in cashflow	Change In Cashflow

Name Of Feature	Definition Of Characteristic
change in investments	Change In Investments
change total debt	Change In Total Debt
change in leverage	Change In Leverage
change in grossprofit	Change In Grossprofit
change in sales	Changes In Revenue
change in grossprofitability	Change In Gross Profability
change in 52weekhigh	Change 52 Week High
change in total liabilities	Change In Total Liabilities
change in operating assets	Change In Operating Assets
change in net income	Change In Net Income
change in return on operating assets	Change Return On Operating Asset
change in return on total assets	Change In return On Total Assets
change in return on equity	Change In return On Equity
change in rsi	Change Relative Strength Index
change sales to price	Change Sales to Price
change in sales and admin costs to sales	Change Sales And Admin Costs To Sales
change in 4m lagged pp&e	change in 4m lagged pp&e
change in avg broker recommendations	Change In Average Broker Recommendation
change in price target	Change In price Target
change in est annual eps growth 5yr	Change In 5 Year Estimated Annual Earnings Per Share Growth
change in TEV/total revenues	Change Total Enterprise Value / Total Revenues
change in TEV/EBIT	Change Total Enterprise Value / EBIT
change in P/sales	Change Price / Sales
change in peg ratio	Change In PEG Ratio
Mkt-RF	Fama & French Market Factor
SMB	Fama & French Small Minus Big
HML	Fama & French High Minus Low
RMW	Difference returns of firms operating profitability
CMA	Investment factor
RF	Fama & French Risk Free Rate