CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Can Machines Explain Stock Returns?

Master's thesis

Author: Bc. Karolína Chalupová

Study program: Economics and Finance

Supervisor: doc. PhDr. Jozef Baruník, Ph.D.

Year of defense: 2020

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, June 30, 2020

Karolina Chalupova

Abstract

The abstract should concisely summarize the contents of a thesis. Since potential readers should be able to make their decision on the personal relevance based on the abstract, the abstract should clearly tell the reader what information he can expect to find in the thesis. The most essential issue is the problem statement and the actual contribution of described work. The authors should always keep in mind that the abstract is the most frequently read part of a thesis. It should contain at least 70 and at most 120 words (200 when you are writing a thesis). Do not cite anyone in the abstract.

JEL Classification	C45, G12
Keywords	(interpretable) machine learning, neural networks, stock returns

Title	Can Machines Explain Stock Returns?
Author's e-mail	`chalupova.karolina@gmail.com`
Supervisor's e-mail	`barunik@fsv.cuni.cz`

Abstrakt

TODO cesky preklad abstraktu.

Klasifikace JEL	C45, G12
Klíčová slova	(interpretovatelné) strojové učení, neuronové sítě, akciové výnosy

Název práce	Mohou stroje vysvětlit akciové výnosy?
E-mail autora	`chalupova.karolina@gmail.com`
E-mail vedoucího práce	`barunik@fsv.cuni.cz`

Acknowledgments

The author is grateful especially to doc. PhDr. Jozef Baruník, Ph.D., [TODO add]

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record
Chalupova, Karolina: *Can Machines Explain Stock Returns?*. Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2020, pages 42. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

ML Machine Learning

Master's Thesis Proposal

Author	Bc. Karolína Chalupová
Supervisor	doc. PhDr. Jozef Baruník, Ph.D.
Proposed topic	Can Machines Explain Stock Returns?

Motivation

Hypotheses

Methodology

Expected Contribution

Outline

Author

Supervisor

Core bibliography

Chapter 1

Introduction

Understanding of the drivers of stock prices is important for a number of reasons. First, fair stock valuation is vital for proper functioning of the stock market. The stock market, in turn, needs to function so as to maintain its roles: it enables firms to obtain financing for their investments, it allows investors to store their present wealth for the future and to share risk. [TODO elaborate] Moreover, as history has shown, incorrect stock price valuations can have severe ramifications. [TODO explain this more and give examples.]

Over the last 50 years, the academia has accumulated hundreds of variables that are proposed to explain stock returns. Harvey *et al.* (2016) and McLean & Pontiff (2016) have shown that most of these existing research findings are likely false and the field entered a deep crisis. The multidimensional challenge (Cochrane 2011) emerged: which of the published and unpublished determinants of stock returns are valid, and which are erroneous? The search for a reliable model is, essentially, same as in any other field. First, the explanatory variables should be based on theory. Second, rigorous statistical methods should be used to ensure robustness of the model. Specifically, if there are very many potential explanatory variables, it is necessary to consider all of them jointly, that is, to control for the rest of the variables. It is possible to include all the candidate variables in a single model, allowing the effects to crowd each other out. However, with hundreds of candidate variables the asset pricing field has accumulated over decades, traditional methods break down. R^2 goes very deep to the negative territory in unpenalized linear regressions (Gu *et al.* 2020) and portfolio-sorts become unusable as early as with 4 variables. This is where ML methods come in.

Machine learning is on the rise in the asset pricing field, both in academia

and industry. The academia is using machine learning (ML) methods to tackle the biggest challenges of the field, such as the problem of high dimensionality. Meanwhile, in the industry, finance practitioners use unparalleled predictive power of ML methods to forecast asset returns (Gu *et al.* 2020). [TODO add evidence of ML rise].

However, there is a trade-off between model performance and its intrinsic interpretability, which is why ML models are often perceived as black-box. [TODO develop further]

The solution to overcome this trade-off is to use ML interpretability methods to extract insights about the model ex-post. This can be readily done even for more complex models such as neural networks. [TODO develop further]

Interpretable ML can be used as the driving force of scientific discovery in many fields, including asset pricing. It can be used to propose plausible causes of phenomena, i.e., to help find the possible X in $y = f(X)$ problems. Typically, human researchers propose the X, and then devise a cause-effect mechanism and test it. The proposal of X can be done using interpretable ML. This is already done in practice [TODO examples from chemistry or others]. [TODO explain usefullness to asset pricing academia].

Interpretability of a ML model is also crucial for the asset-pricing industry. First, the typical approach to model discovery, that is, trying different models and backtesting them on the same data, leads to false positive discovery: it only takes 20 trials to find a false positive at the standard 5% significance, in other words, it is quite easy to overfit the backtest. Understading the sources of model performance can be used as an alternative research tool to guide strategy discovery (De Prado 2018): for example, using feature importance measures, one uncover which inputs to the model are important, and use this knowledge to limit the amount of noise or add more strong predictors. Second, knowing the sources of model performance can help with marketability of the model, as when ML interpretability methods are used, the model ceases to be a black box. From a practitioners perspective, my thesis can serve as a case study of some of the methods, issues, and results that occur when interpreting return-predicting ML models.

The objective of this thesis is .... [TODO]

The thesis is structured as follows: ... [TODO]

Chapter 2

Literature Review

The question of what drives the stock returns is a thrilling one to answer from several perspectives: that of an economist, of a statistician, of a machine learning engineer, and that of a finance practitioner. An economist sees that stock returns reflect human decisions: putting a price tag on uncertain future payoffs reflects the trade-off between current and future consumption, human impatience and attitudes to risk (Cochrane 2009) as well as human behavioral biases [TODO add reference]. She also sees that they speak of the complex web of relations between firms: as companies form relationships, they become exposed to similar risks, and their returns correlate. For the economist, studying the determinants of stock returns therefore is an opportunity to understand human behavior as well as the macro-phenomena that emerge in the network of firms' relationships. The statistician sees that the determinants of stock returns are notoriously over-studied, which brings many issues of publication bias and multiple hypothesis testing, leading to many false discoveries (Harvey *et al.* 2016). From this perspective, the question stands: how to separate the wheat from the chaff? The machine learning engineer sees the problem as a prediction task: there is a large amount of data, the variables likely interact in complex ways, and are highly correlated. Machines have proven to excel in the task of predicting stock returns. They learn from financial data, model the interactions, and produce unparalleled returns predictions (Gu *et al.* 2020). Finally, for the finance practitioner seeks to identify the future winners and losers, and sell and purchase them to make profit. As the machine-learning approaches are becoming the norm in the applied field, a need arises to understand the models on a deeper level to interpret them, which brings us back to the economic underpinnings of stock returns.

This thesis attempts to bring the four approaches together. I use machine learning models to predict stock returns and then interpret them to bring us closer to the answers to the economists' questions. This approach can also help to filter out false discoveries. Finally, the finance practitioner can employ ML interpretability methods to find an intelligible model, of which she understands the weaknesses and sources of performance.

The literature review proceeds as follows. First, I review the economic theory behind asset prices and common attitudes to modeling them. Second, I present the statistician's view: the multidimensional challenge faced in modeling stock returns (Cochrane 2011) and show what different approaches have been used to answer the challenge. Third, I present the machine learning approaches to asset pricing. Fourth, I review ML interpretability methods.

2.1 Economist's Perspective

Valuing a stock amounts to putting a price tag on a stream of future payoffs. The field concerned with doing just that is called asset pricing. Consider the following basic asset-pricing equation (Cochrane 2009):

$$p_t^i = E_t(m_{t+1}x_{t+1}^i) \tag{2.1}$$

In words, the price of an asset i at time t (p_t^i) is proportional to the expected asset's payoff at time $t+1$ (x_{t+1}^i). However, since the payoff does not come now (t), but in future ($t+1$), we discount it to present by the factor m_{t+1}, called the *stochastic discount factor*. The term stochastic stands to express the idea that m_{t+1} is not known with certainty at time t. There is only one assumption made in order to write the equation: it can be shown that m_{t+1} exists if and only if the *law of one price* holds, that is, if two assets generate the exact same payoff in all possible states of nature have the same price. If this is the case, there is a discount factor such that the equation holds for all assets i (Cochrane 2009).

In the stock market, equation 2.1 translates to:

$$1 = E_t(m_{t+1}R_{t+1}^i) \tag{2.2}$$

In words, an investor pays 1 dollar now to collect R_{t+1}^i dollars in the future. The payoff is called gross return and is the sum of future price and dividend

$R^i_{t+1} = p^i_{t+1} + d^i_{t+1}$. Using the definition of covariance, we can rearrange 2.2 (Cochrane 2009):

$$1 = E_t(m_{t+1})E_t(R^i_{t+1}) + Cov_t(m_{t+1}, R^i_{t+1}) \tag{2.3}$$

Defining the *risk-free gross return* as $R^f_{t+1} = \frac{1}{E_t(m_{t+1})}$ and using it to further rearrange (Cochrane 2009):

$$E_t(R^i_{t+1}) = \underbrace{R^f_{t+1}}_{\text{risk-free return}} - \underbrace{R^f_{t+1}Cov_t(m_{t+1}, R^i_{t+1})}_{\text{risk adjustment}} \tag{2.4}$$

This important result shows that expected return if stock i can be decomposed into risk-free return and risk adjustment. Note that the risk adjustment occurs if and only if returns are correlated to the discount factor, so *idiosyncratic* risk, that is, uncorrelated with the discount factor, is uncompensated. Returns positively correlated to the discount factor should be low, and vice versa. So to explain average returns, we "only" need to explain the returns' correlation to the discount factor m_{t+1}.

Further rearranging 2.4 provides one more insight (Cochrane 2009). Multiply both sides by $Var_t(m_{t+1})/Var_t(m_{t+1})$ to obtain:

$$E_t(R^i_{t+1}) = R^f_{t+1} + \underbrace{\frac{Cov_t(m_{t+1}, R^i_{t+1})}{Var_t(m_{t+1})}}_{\text{denote } \beta_{i,t}} \cdot \underbrace{\left(-\frac{Var_t(m_{t+1})}{E_t(m_{t+1})}\right)}_{\text{denote } \lambda_t} \tag{2.5}$$

$$= \underbrace{R^f_{t+1}}_{\text{risk-free return}} + \underbrace{\beta_{i,t} \cdot \lambda_t}_{\text{risk adjustment}} \tag{2.6}$$

This equation is so-called *beta-representation* of 2.2. It shows that the risk-adjustment, the premium a stock pays for being correlated with the discount factor, can be decomposed into $\beta_{i,t}$ and λ_t. λ_t is the volatility of the discount factor and it is unrelated to properties of asset i. It can be interpreted as the price of risk. $\beta_{i,t}$ can be interpreted as the amount of the risk inherent in asset i. Empirically, it can be obtained as the coefficient from regressing asset i's returns on the discount factor [TODO does this estimation introduce the assumption that beta is time-invariant?]:

$$R^i_{t+1} = a_i + \beta_i m_{t+1} + \epsilon_{i,t+1} \tag{2.7}$$

Equations 2.4 and 2.6 show that the discount factor is the key to explaining

stock returns, as the latter is nothing but compensation for correlation with the former. Thus, specifying the discount factor is the only content of any asset-pricing model Cochrane (2009). I review the consumption-based model, the CAPM, and multiple-factor models as special cases of these equations providing more intuition about what drives the discount factor.

2.1.1 Consumption-Based Model

The consumption-based model takes the additional assumption that the investor's preferences can be captured by her utility from consumption. The investor needs to decide how much to consume today and how much to save for tomorrow. First order conditions for this problem lead to the following specification of the discount factor Cochrane (2009):

$$m_{t+1} = \kappa \frac{u'(c_{t+1})}{u'(c_t)} \tag{2.8}$$

where $u'(c_{t+1})$ denotes marginal utility from consumption. That is, the discount factor is the consumer's marginal rate of substitution between consumption at time t and $t+1$ and captures the willingness to trade consumption today for consumption tomorrow, where parameter κ is the weight placed on future utility. This means that such an investor would demand a high return for stock that perform badly at times when she is unwilling to give up today's consumption, and a low return for stock that performs badly at times she is willing to give up today's consumption.

For simplification, one can assume constant relative risk aversion:

$$u(c_t) = \frac{c^{1-\gamma} - 1}{1 - \gamma} \tag{2.9}$$

where γ is a parameter positive for risk-averse individuals, and plug in to 2.8:

$$m_{t+1} = \kappa \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} \tag{2.10}$$

That is, the discount factor is inversely proportionate to consumption growth. 2.4 says that assets with positive correlation with the discount factor should have lower returns, so it follows that assets highly correlated with consumption growth should earn high returns. This makes sense: the investor is assumed to care only about her marginal utility from consumption. Therefore, she de-

mands a higher return for holding a stock that performs badly at times when her consumption decreases, and only a low return on insurance-like stocks that perform well during bad times.

For the sake of completeness, we can rewrite this to the equivalent beta-representation using 2.6 [TODO odvodit, overit, zpresnit notaci]:

$$E_t(R^i_{t+1}) = R^f_{t+1} + \beta_{i,t} \cdot \lambda_t \tag{2.11}$$

$$\lambda_t = \gamma Var_t(\Delta c_{t+1}) \tag{2.12}$$

$$R^i_{t+1} = \alpha_{i,t} + \beta_{i,t}\Delta c_{t+1} + \epsilon_{i,t+1} \tag{2.13}$$

However, this model does not perform very well empirically Cochrane (1996). There are two possible reasons, either, the aggregate consumption data are imperfectly measured, or the assumption that the investor maximizes the utility from consumption is off. This motivates tying the discount factor to other variables.

2.1.2 Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM) is the possibly most widely-known asset pricing model. It is the special case of the consumption-based model, making the additional assumption that the utility from consumption is logarithmic (Rubinstein 1976):

$$u(c_t) = ln(c_t) \tag{2.14}$$

Plugging this to 2.8, one obtains

$$m_{t+1} = \kappa \frac{c_t}{c_{t+j}} \tag{2.15}$$

The CAPM operates with the concept of wealth portfolio, which comprises all world's wealth, inclusing real estate, metals, machinery or art. The price of this portfolio is (by 2.1 and plugging in for the discount factor):

$$p^W_t = E_t \sum_{j=1}^{\infty} m_{t+j} c_{t+1} = \frac{\kappa}{1-\kappa} c_t \tag{2.16}$$

and the return on wealth portfolio is therefore

$$R^W_{t+1} = \frac{p^W_{t+1} + c_{t+1}}{p^W_t} = \frac{1}{\kappa}\frac{c_{t+1}}{c_t} = \frac{1}{m_{t+1}} \tag{2.17}$$

Thus, according to CAPM, the discount factor is the inverse of the return on the wealth portfolio. As 2.4 says that assets with positive correlation with the discount factor should earn low returns, if follows that assets positively correlated with the wealth portfolio should earn higher returns. Again, this makes sense: an investor who only cares about consumption demands a higher return for stocks that perform badly when other sources of wealth perform badly, and vice versa, she is willing to forgo some average return in exchange for good performance at times when all else fails.

Again, for completeness, we can rewrite this to the equivalent beta-representation using 2.6 [TODO odvodit, overit, zpresnit notaci]:

$$E(R^i_{t+1}) = \gamma + \beta_{i,t} \cdot \lambda_t \tag{2.18}$$

$$\lambda_t = E_t(R^W_{t+1}) - \gamma \tag{2.19}$$

$$R^i_{t+1} = \alpha_{i,t} + \beta_{i,t} R^W_{t+1} + \epsilon_{i,t+1} \tag{2.20}$$

CAPM accomplishes the task of removing consumption data from empirical estimation of discount factor (Cochrane 2009), replacing them by return on wealth portfolio. However, this return is unobservable, so it is empirically often replaced by return on market portfolio of stocks. This may be problematic, as wealth comprises many more assets than just stocks, so the approximation is likely too crude (Roll 1977).

2.1.3 Multi-factor Models

Multi-factor models are motivated by the notions that consumption growth and return on wealth portfolio do not proxy the discount factor empirically very well (in the consumption-based model and CAPM respectively), and they try to find other proxies for the discount factor m_{t+1}, factors $\boldsymbol{f}_{t+1}$. If we retain the idea that investor maximizes her utility from consumption, then the factors should proxy the marginal utility growth (Cochrane 2009):

$$m_{t+1} = \kappa \frac{u'(c_{t+1})}{u'(c_t)} \approx a + \boldsymbol{b}'\boldsymbol{f}_{t+1} \tag{2.21}$$

The expression on the left-hand side can be interpreted as the rate at which the investor is willing to swap future for current consumption. [TODO verify that this understanding of mine is correct]. There are many variables that affect this rate. The current state of the economy, such as GDP growth, interest rate or investment, as well as new of the future states, forecasting either asset returns or macroeconomic variables. Many candidate factors can be defended on these grounds.

Multi-factor models explain returns as

$$E_t(R^i_{t+1}) = R^f_{t+1} + \boldsymbol{\beta}'_{i,t}\boldsymbol{\lambda}_t \tag{2.22}$$

$$R^i_{t+1} = \alpha_{i,t} + \boldsymbol{\beta}'_{i,t}\boldsymbol{f}_{t+1} + \epsilon_{i,t+1} \tag{2.23}$$

The *factor loadings* $\boldsymbol{\beta}'_{i,t}$ of size $1 \times K$ are exposures to risk factors $\boldsymbol{f}_{t+1}$ of size $K \times 1$, where K is the number of risk factors. They represent the amount of risk inherent in asset i due to its correlation to the corresponding risk factor. The $\boldsymbol{\lambda}_t$ of size $K \times 1$ are interpretable as the risk prices at time t (Kelly *et al.* 2019). (Note that the CAPM and consumption-based models are special cases of the multiple-factor models, where K, the number of factors, is 1.)

First multi-factor model is Fama & French (1996) [TODO make sure the time indexation is right]:

$$E(R^i) = R^f + \boldsymbol{\beta}'_i\boldsymbol{\lambda} \tag{2.24}$$

$$R^i_{t+1} = \alpha_i + \boldsymbol{\beta}'_i\boldsymbol{f}_{t+1} + \epsilon_{i,t+1} \tag{2.25}$$

with factors

$$\boldsymbol{f}_{t+1} = \begin{pmatrix} R^M_{t+1} - R^f_{t+1} \\ SMB_{t+1} \\ HML_{t+1} \end{pmatrix} \tag{2.26}$$

and with factor prices

$$\boldsymbol{\lambda} = \begin{pmatrix} E(R^M) - R^f \\ E(SMB) \\ E(HML) \end{pmatrix} \tag{2.27}$$

The $R^M_{t+1} - R^f_{t+1}$ is the return on market portfolio above risk-free rate (as

in CAPM), SMB_{t+1} is the return on small-cap portfolio minus the return on big-cap portfolio, and HML_{t+1} is the return on portfolio of stocks with small market values relative to book value, minus the return on portfolio of stocks with high market values relative to book value. Note that the risk prices $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}_i'$ are time-invariant. Explanations why SMB and HML should be correlated to the discount factor include that they proxy for financial distress (Fama & French 1996), (Heaton & Lucas 2000) or that they forecast the future state of the economy (Liew & Vassalou 2000). Fama & French (2015) augment this three-factor model to five factors, adding RMW and CMA, which are, respectively, the difference in returns on portfolios of firms with robust and weak profitability, and difference in returns on portfolios of firms with conservative and aggressive investment.

[TODO add a systematic overview of existing anomalies.]

[TODO describe portfolio sorting metodology]

It is difficult to empirically estimate a multifactor model. The main challenge is that both $\boldsymbol{f}_{t+1}$ and $\boldsymbol{\beta}_{i,t}'$ are unobserved, so we have no directly observed variables at the right hand side of the estimated equation. There are two different approaches to tackle this issue (Kelly *et al.* 2019). The first is to use use prior knowledge of empirical behavior of average returns to pre-specify the factors, treat them as observable and then estimate β_{t-1}. An example of this approach is (Fama & French 1993) and most of the hundreds of published factors are too Cochrane (2009). But as noted by (Kelly *et al.* 2019, p. 3), this prior specification relies on "a partial understanding at best, and at worst is exactly the object of empirical interest." Even Fama & French (1993) note that without a clear economic theory for factors, the choice of the right-hands side variables is arbitrary. But as we have seen, the economic theory justifies inclusion of almost anything as the right-hands side variable: "One can appeal to the APT or ICAPM to justify the inclusion of just about any desirable factor" (Cochrane 2009, p. 124)), so we do not have a clear economic theory.

The other way to uncover the factor structure is, therefore, to instrument the unobserved factor loadings using firm-level variables, commonly called *characteristics*.

$$E_t(R^i_{t+1}) = R^f_{t+1} + \boldsymbol{\beta}'_{i,t}\boldsymbol{\lambda}_t \tag{2.28}$$

$$R^i_{t+1} = \alpha_{i,t} + \boldsymbol{\beta}'_{i,t}\boldsymbol{f}_{t+1} + \epsilon_{i,t+1} \tag{2.29}$$

$$\boldsymbol{\beta}'_{i,t} = g(\boldsymbol{z}'_{i,t}) \tag{2.30}$$

The first two equations are the same as in multiple-factor models, while the third shows instrumenting of the factor loadings with firm characteristics $\boldsymbol{z}'_{i,t}$. This approach is exemplified by Kelly *et al.* (2019). The key insight is that characteristics serve as proxies for *exposure* to different sources of systematic risk. Once risk loadings are instrumented, the authors use them to estimate the corresponding factors.

2.1.4 Arbitrage Pricing Theory

[TODO add economic motivation for drivers of expected returns from APT.]

2.2 Statistician's Perspective

> We argue that most claimed research findings in financial economics are likely false.
>
> Harvey *et al.* (2016)

There are hundreds of published factors suggested as explanation of average stock returns. Harvey *et al.* (2016) count 313 variables, considering the top journals only and calling the count "surely too low". Indeed, Cochrane (2011) refers to the state of asset pricing as a "zoo of factors". At the same time, the number of factors should be low in theory (Cochrane 2011) and empirically (Ahn *et al.* 2012) [TODO elaborate].

(Cochrane 2011, p. 1060) formulates the problem of separating the wheat from the chaff in asset pricing in his "multidimensional challenge": "We have a lot of questions to answer: First, which characteristics really provide independent information about average returns? Why are subsumed by others? Second, does each new anomaly variable also correspond to a new factor formed on those same anomalies? (...) Third, how many of these new factors are really important?"

The main issue with the hundrends of factor models is that they find the studied factor significant controlling only for the effects of market return or 3 or five other factors, typically factors from Fama & French (1996) or Fama & French (2015). To establish the significance of the variables, it is necessary to consider them jointly in a single model, allowing the factors to crowd each other out.

Moreover, there are numerous issues that severely diminish the reliability of these factors. First and most obviously, some results are bound to be just false positives due to the typical confidence levels (Harvey *et al.* 2016). Second, as it is difficult to publish a non-result, the insignificant results remain in the drawer and the significant ones get published, artificially driving up the significance in an instance of publication bias (Harvey *et al.* 2016). Third, the published studies are often biased in themselves: the specification search bias (selection of the model based on model's result), the sample selection bias (selection of the data based on model's results) and the multiple hypothesis testing bias (conducting multiple tests of the same hypothesis) all result in artificially high significance of returns predictors (McLean & Pontiff 2016). The situation is worsened by the fact that there is only a limited amount of data in finance: CRSP and Compustat are limited resources, and researches use the same data over and over, resulting in collective over-fitting (Harvey *et al.* 2016). Finally, unlike in other fields, it is difficult to publish a replication study in finance, which leaves the spurious factors unnoticed (Harvey *et al.* 2016). There are two ways to correct these biases ex post: out-of-sample tests and using a framework that accounts for the multiple testing and rises the usual significance levels. McLean & Pontiff (2016) perform the out-of-sample tests, studying 82 variables. As much as 10 of them could not even be replicated in-sample. The rest are biased on average by 10%. Harvey *et al.* (2016) take the multiple-testing approach. Out of 296 published significant factors, 158, 142, 132 and 80 are false discoveries, the precise number depending on statistical framework used Harvey *et al.* (2016).

2.3 Machine Learning Perspective

> To adress [the multidimensional challenge] in the zoo of new variables, I suspect we will have to use different *methods*.
>
> Cochrane (2011)

A possible definition of ML is from (Mitchell 1997, p. 2): "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In the case of predicting stock returns, the task T is a regression, performance measure can be for example the mean-squared-error, and the experience E is the historical data. Using this definition, ML is a broad collection of methods, including the very simplest, like linear regression or principal components analysis, that are well-known to any economist.

ML offers many varied methods that are well-suited to the task of explaining stock returns and that have an advantage over the more traditional methods (Gu *et al.* 2020). First, ML methods can handle data with high dimensionality and correlation in predictors. If the hundreds of candidate factors are used in linear regression, R^2 becomes negative and no predictors are significant (Gu *et al.* 2020) due to the high dimension and multicollinearity. Portfolio sorts, another traditional method in asset pricing, become unusable at around 5 predictors Cochrane (2011); Bryzgalova *et al.* (2019). ML offers dimension reduction and variable selection methods to overcome this challenge. Second, it seems that the linear functional form for predictors is too restrictive: Gu *et al.* (2020) statistically reject the traditional linear models in favor of the non-linear ones. ML offers methods of choosing the right functional form systematically. Third, the model selection is ML is based on validation data, as opposed to in-sample (training) data. In traditional econometrics approach to stock returns, model is tuned by researcher based on in-sample fit, which results in spurious findings that are very often not replicable out-of-sample (McLean & Pontiff 2016). On the other hand, ML puts an emphasis on out-of-sample performance, which allows to select a robust model.

Gu *et al.* (2020) apply a wide range of ML methods to the stock-returns prediction task. Their model is

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1} \tag{2.31}$$

where

$$E_t(r_{i,t+1}) = g(\boldsymbol{z}_{i,t}) = g(\boldsymbol{x}_t \otimes \boldsymbol{c}_{i,t}) \tag{2.32}$$

The function g stands for any ML model and the variables $\boldsymbol{z}_{i,t}$ are 94 characteristics of stocks $\boldsymbol{c}_{i,t}$ interacted with 8 macroeconomic predictors $\boldsymbol{x}_t$. The multiple-factor model 2.28 is a special case of this specification, with g linear:

$$E_t(r_{i,t+1}) = g(\boldsymbol{z}_{i,t}) = \boldsymbol{c}'_{i,t}\Theta'_1\Theta_2\boldsymbol{x}_t = \boldsymbol{\beta}'_{i,t}\boldsymbol{\lambda}_t \tag{2.33}$$

The authors make a horse race of several ML and traditional methods. First, they include methods to estimate linear g, both without regularization (linear regression with all variables (OLS), linear regression only with size, book-to market, and mometum (OLS-3)) and with regularization (partial least squares (PLS), principal components regression (PCR), elastic net (ENet)). Second, they include methods to estimate nonlinear g: generalized linear models (GLM), random forest, gradient-boosted regression trees, and neural networks of depths 1 to 5. They find that OLS performs badly, regularized linear models are an improvement, and, and nonlinear models are the best-performing, namely neural networks followed by random trees.

Interestingly, the authors find that all methods (with the exception of OLS) agree in the variables that have an important first-order impact on returns. These are, ordered by importance: recent price trends (such as momentum and long- and short-term reversal), liquidity measures (such as turnover), risk measures (such as return volatility and market beta), valuation ratios (such as earnings-to-price), and fundamental signals (such as asset growth). On annual frequency, recent price trends become less important, and industry emerges as an important predictor, but otherwise the results are similar to the monthly frequency.

Bryzgalova *et al.* (2019) use random trees to build the discount factor. They generalize portfolio-sorts to accommodate a large number of predictors. Similarly to Kelly *et al.* (2019), they consider factor loadings $\boldsymbol{\beta}_{i,t}$ as function of characteristics. Specifically, the characteristics include recent price trends

(momentum and long- and short-term reversal), liquidity measures (such as turnover), investment, profitability, accruals and book-to-market ratio. Their approach leads to selection of a very sparse and readily interpretable set of basis assets that are used to construct the discount factor. This is done using a global approach to pruning a tree. Global pruning is necessary as the task is to use the tree's nodes to span the stochastic discount factor, which cannot be done using local decision criteria. The pruning method proceeds in the following steps.

The intermediate and final nodes of the tree then serve as basis assets to construct the discount factors. This is done by optimally weighting the basis assets. The weights are found by finding the minimum variance weights for each expected return, solving the Markowitz problem with additional shrinkage of the weights using elastic net penalty, choosing the the optimum expected return and shrinkage parameters to maximize the Sharpe ratio on validation data. Specifically, the approach takes three steps:

First, for each node of the tree (including the first and intermediate nodes, i.e., not just the leaves), calculate the sample estimates of mean and covariance matrix of the excess returns, denote them $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ respectively.

Second, find weights $\boldsymbol{w}$ of the nodes such that the resulting portfolio will have minimum variance, at least a given mean return, and such that the weights are small (elastic net shrinkage). That is, for given μ_0, and shrinkage parameters λ_1, λ_2, find weights $\boldsymbol{w}$ that solve

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2}\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{w} + \lambda_1 \|\boldsymbol{w}\|_1 + \frac{1}{2}\lambda_2 \|\boldsymbol{w}\|_2 \\ \text{subject to} \quad & \boldsymbol{w}^T\mathbb{1} = \mathbb{1} \\ & \boldsymbol{w}^T\hat{\boldsymbol{\mu}} \geq \mu_0 \end{aligned} \tag{2.34}$$

Third, the parameters μ_0, and shrinkage parameters λ_1, λ_2 are selected on validation set such that the Sharpe ratio is maximized.

Bryzgalova *et al.* (2019) and Gu *et al.* (2020) agree that the ability to uncover non-linearities and interactions is the crucial driving force of the superior ML performance. If a variable has a non-linear relationship with returns, the linear models can conclude there is no association, while in reality there is an important non-linear relationship. For example, in Gu *et al.* (2020), linear models deem size and volatility unimportant predictors, while non-linear models find the contrary. A similar problem may arise if predictors are have zero

association with return themselves, but not in interaction with other predictors. [TODO include examples].

2.4 Interpretable Machine Learning

ML interpretability methods can be grouped into into several categories (Molnar 2020). A first dictinction can be made between *intrinsically* interpretable models, that are intrepretable per se (linear and logistic regression, random tree), and *post-hoc methods*, that can be applied to extract interpretable knowledge from the model ex-post. A further distinction can be made as to the applicability of the interpretability methods: some methods are *model-specific*, that is, can only be applied to some ML models others are *model-agnostic*, applicable to all models from linear regressions to random trees and neural networks. Yet other distinction can be made between *global* and *local* ML interpretability methods: while the former assess sources of model performance across all observations, the latter can be used to interpret the drivers within a single observation.

2.4.1 Global Feature Importance Measures

Feature importance focuses on how much the model relies on an input feature to produce a prediction. There are several different notions of what qualifies a feature as important:

Mean Decreased Performance proposed by (Fisher *et al.* 2019), also known as **Permutation Feature Importance** is a model-agnostic measure of feature importance, which measures how much of model's performance is lost when permuting a given feature. It compares the performance of a model fit on a set of features with the performance of the same model after permuting a given feature (randomly shuffling its features across examples, without refitting the model). If the particular feature is important, permuting it will decrease model's performance. There are several special cases depending on the measure the model's performance, e.g., **Model Reliance** (Fisher *et al.* 2019) (mean squared error), or **Mean Decreased Accuracy** [TODO add citation] (accuracy). Since the measure depends on random permutation, its variance can be decreased by computing it several times and averaging the result. [TODO add alogorhitmic description.]

Mean Decreased Explained Variance is also a model-agnostic measure of feature importance, and measures how much of model's explained variance is lost when replacing the feature's values. It is different to Mean Increased Loss in that rather than measuring changes in model's *error*, it measures the decrease in explained variance of the target, or R^2. It is used for example by Gu *et al.* (2020) or Kelly *et al.* (2019). [TODO find original methodological paper introducting the measure]. Gu *et al.* (2020) use replacement of values by 0, as their features are scaled from -1 to 1, but generally a different replacement is needed [TODO could be permutation? I think so. In that case average needs to be taken.]

Sum of Squared Derivatives proposed by Dimopoulos *et al.* (1995) is the sum of squared partial derivatives of the model's prediction with respect to the given feature. The disadvantage of this measure that is its not model-agnostic, specifically it is unusable if derivatives cannot be taken such as in random trees.

Mean Decreased Impurity proposed by [TODO find out] can be used for tree-based methods, unlike Sum of Squared Derivatives.

2.4.2 Shapley Values

Shapley values are a local measure of feature importance. They are due to Shapley (1953), (originating in coalitional game theory) and in the ML context are well-described in Molnar (2020). They can be used to decompose the prediction for given observation into effects of individual features. For given feature and observation, it is the contribution of the feature to the difference between the actual prediction and the mean.

In case of linear model

$$\hat{f(x)} = \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_K x_K \tag{2.35}$$

where $\boldsymbol{x}$ are the values for particular observation, the Shapley value of feature j is exactly

$$\psi_j(\hat{f}) = \beta_j x_j - \beta_j E(X_j) \tag{2.36}$$

Thanks to Shapley values, we can arrive at the same level of interpretability also for complex models. In case of linear models, the order in which the

features "vote" for the prediction is irrelevant (addition is commutative). This is not so in nonlinear models, and this is why the Shapley value takes account for all possible orders in which features can participate in the prediction: in a model with K features, there are 2^K possible coalitions between them. Intuitively, imagine that all feature values of a particular observation enter a room in random order. At each point, all feature values make a prediction. Shapley value for feature j is the average change in prediction that happens when feature j enters the room (Molnar 2020).

Exact Shaplay value for given observation and feature can be calculated as follows: [TODO add]

Since the computation time of exact Shapley value grows exponentially with number of features, it may be infeasible to compute the exact value. Instead, the Shapley value can be estimated by only computing some of all the possible configurations (Štrumbelj & Kononenko 2014).

[TODO include algorithm]

2.4.3 Accumulated Local Effects

Accumulated Local Effects (ALE) are due to [TODO] described in Molnar (2020). [TODO describe in detail.]

2.5 Unused text

Of course, studying only the first-order impact of a characteristic provides only a very limited picture. For example, Bryzgalova *et al.* (2019) find that the first-order impact of accruals on returns is flat (zero), but when controlling for size, the relationship turns out to be U-shaped.

Chapter 3

Data and Methodology

3.1 Raw Data

The data originates from two sources. The first and most important source is Thomson Reuters Datastream. It covers publicly traded shares from developed countries of Europe, Japan and Asia-Pacific (Australia, New Zealand, Hong Kong, and Singapore) and contains the firms' yearly accounting figures and daily market information (e.g., price and volume). The second data source is Institutional Brokers Estimate System (I/B/E/S) from Wharton Research Data Services, which provides analysts' forecasts.

The firms are filtered based on [TODO add source of filtering methods]. The final sample includes 8350 companies.

3.2 Anomalies Data

The anomaly dataset consists of 153 anomalies published in leading financial and accounting journals. The data is monthly and spans from January 1990 to December 2018. It covers the same 8350 firms as the raw data, totaling 1,607,117 observations.

The anomalies can be classified into three groups: Fundamentals, Frictions, and I/B/E/S anomalies. Fundamental anomalies are primarily calculated from yearly accounting statements and are exemplified by Total Accruals, Asset Growth, or Leverage. Fundamental anomalies fall into 5 broad categories: Accruals (e.g., Total Accruals, Change in Common Equity, Inventory Growth), Investment (e.g., Asset Growth, Debt Issuance, Net Operating Assets), Intangibles (e.g., Asset Liquidity, Earnings Persistence, Herfindahl Index), Prof-

itability (e.g., Profit Margin, Return-on-Equity, Capital Turnover) and Value (e.g., Enterprise Multiple, Assets-to-Market, or Net Payout Yield). Friction anomalies are primarily calculated from daily market data. For example, they include Bid-Ask Spread, 11-Month Residual Momentum, 52-Week-High, and Long-Term Reversal. I/B/E/S anomalies are calculated from the I/B/E/S dataset and cover analysts' forecasts and recommendations, such as Analyst Value, Change in Recommendation, or Forecast Dispersion. There are 93 Fundamentals, 48 Frictions and 12 I/B/E/S anomalies. Within the Fundamentals category, there are 21 Accruals, 15 Investment, 25 Intangibles, 18 Profitability, and 14 Value anomalies.

A complete list of the anomalies can be found in the Appendix. [TODO].

Tables 3.1, 3.2, and 3.3 show descriptive statistics of the anomalies data. [TODO fit too long fund table to page or split across multiple pages.]

Figures 3.1, 3.2, and 3.3 show plots of correlation matrices of Fundamental, Frictions, and I/B/E/S anomalies respectively. Correlations between variables from different categories are not shown, as they are all very low but for negligible exceptions.

Figure 3.4 shows histogram of monthly returns.

	count	mean	std	min	25%	50%	75%	max
Accr	1326726.0	-3.130000e-02	3.010000e-01	-5.376980e+01	-0.0668	-0.0299	0.0045	2.323860e+01
AGr	1426822.0	inf	NaN	-1.000000e+00	-0.0557	0.0595	0.1852	inf
AL	1185337.0	-inf	NaN	-inf	-0.6710	-0.5098	-0.3670	1.196750e+01
AL2	1135791.0	-1.078751e+03	1.791780e+05	-4.988183e+07	-1.5728	-0.7994	-0.3784	4.935200e+00
ATurn	1174246.0	3.595700e+00	1.478946e+02	-4.632576e+03	0.8766	1.6008	2.7225	2.580516e+04
AtM	1463723.0	4.043164e+03	8.896111e+05	0.000000e+00	0.8506	1.6844	3.5502	2.202320e+08
BM	1443483.0	-inf	NaN	-inf	-1.0495	-0.4264	0.1646	1.494580e+01
CT	1424872.0	inf	NaN	-9.138000e-01	0.3921	0.8189	1.2702	inf
CFoMV	1358488.0	3.053290e+01	3.733290e+03	-3.856522e+04	0.0401	0.0865	0.1531	8.549874e+05
CBOP	1044783.0	-inf	NaN	-inf	0.0348	0.0902	0.1541	5.517662e+03
CtA	1356655.0	1.735000e-01	1.610000e-01	0.000000e+00	0.0611	0.1275	0.2321	2.461500e+00
dATurn	1011919.0	-2.033000e-01	2.720040e+01	-2.582384e+03	-0.2043	-0.0190	0.1081	3.323493e+03
dCE	1426774.0	inf	NaN	-8.540260e+03	-0.0156	0.0209	0.0796	inf
dCOA	1219678.0	inf	NaN	-3.769400e+00	-0.0216	0.0128	0.0631	inf
dCOL	1220041.0	-inf	NaN	-inf	-0.0191	0.0091	0.0475	2.031640e+07
dFL	1425235.0	-inf	NaN	-inf	-0.0248	0.0000	0.0472	2.488448e+07
dLTI	852533.0	1.010000e-01	2.134350e+01	-9.762000e-01	-0.0033	0.0000	0.0071	6.969071e+03
dNFA	1136823.0	inf	NaN	-2.488448e+07	-0.0587	-0.0006	0.0436	inf
dNNCWC	1219392.0	inf	NaN	-1.933512e+07	-0.0247	0.0031	0.0350	inf
dNNCOA	1219475.0	inf	NaN	-1.142870e+03	-0.0260	0.0161	0.0750	inf
dNCOA	1219945.0	inf	NaN	-1.496010e+01	-0.0258	0.0179	0.0795	inf
dNCOL	1219851.0	inf	NaN	-2.223446e+05	-0.0048	0.0005	0.0090	inf
dPM	1265175.0	NaN	NaN	-inf	-0.0133	0.0015	0.0178	inf
dSTI	970776.0	2.060000e-02	2.416700e+00	-9.642000e-01	-0.0066	0.0000	0.0084	6.835770e+02
ChNOA	1284565.0	inf	NaN	-3.521843e+03	-0.0400	0.0312	0.1140	inf
ChPPEIA	1227511.0	inf	NaN	-1.437410e+01	-0.0261	0.0329	0.1204	inf
CDI	667698.0	inf	NaN	-9.899100e+00	-1.0154	0.0821	1.3740	inf
CEI5Y	1288644.0	1.125000e-01	6.540000e-01	-1.139050e+01	-0.1206	-0.0373	0.1241	1.987440e+01
DI	1468025.0	5.174000e-01	4.997000e-01	0.000000e+00	0.0000	1.0000	1.0000	1.000000e+00
DA	1223877.0	-5.800000e-03	7.217000e-01	-1.100946e+02	-0.0409	-0.0004	0.0386	6.328230e+01
DurE	1400768.0	NaN	NaN	-inf	-4.9750	5.5412	11.6959	inf
ES	560569.0	2.560000e-02	1.424900e+00	0.000000e+00	0.0002	0.0005	0.0017	1.145981e+02
EoP	1251646.0	3.200700e+01	3.668273e+03	0.000000e+00	0.0295	0.0524	0.0878	6.797906e+05
ECon	802280.0	9.806000e-01	8.442385e+02	-3.277065e+05	0.7819	1.1617	2.0798	5.903915e+05
EConsit	655688.0	6.360000e-02	2.342000e-01	-8.782000e-01	-0.0736	0.0535	0.1844	2.187700e+00
EPer	763925.0	3.694000e-01	5.484000e-01	-3.102810e+01	0.0873	0.3641	0.6177	4.175850e+01
EPred	763925.0	1.211706e+13	2.038141e+15	0.000000e+00	0.0715	0.6041	5.9455	3.436214e+17
ETime	806492.0	3.496000e-01	2.300000e-01	0.000000e+00	0.1638	0.3068	0.5000	1.000000e+00
ECoBP	1335533.0	9.183000e-01	2.593297e+02	-1.982584e+05	0.3517	0.6901	1.0767	1.239914e+05
EM	1262616.0	NaN	NaN	-inf	4.5453	8.2825	14.0408	inf
FSc	249606.0	4.344500e+00	1.863700e+00	0.000000e+00	3.0000	4.0000	6.0000	9.000000e+00
GP	1277559.0	inf	NaN	-6.544031e+02	0.1272	0.2328	0.3823	inf
dINVoAvgA	1290712.0	1.700000e-03	1.540000e-02	-3.494000e-01	-0.0015	0.0001	0.0039	4.335000e-01
dLTNOA	1283630.0	4.350000e-02	1.396100e+00	-3.083692e+02	0.0055	0.0384	0.0748	3.263627e+02
HI	1468025.0	6.570000e-02	6.120000e-02	1.270000e-02	0.0301	0.0498	0.0800	1.000000e+00
HR	1212491.0	4.210000e-02	2.687000e-01	-2.000000e+00	-0.0317	0.0152	0.0855	2.000000e+00
HIAT	1468025.0	7.050000e-02	6.590000e-02	1.300000e-02	0.0305	0.0494	0.0899	1.000000e+00
HIBE	1468025.0	6.470000e-02	7.370000e-02	1.230000e-02	0.0296	0.0467	0.0794	2.980300e+00
IAOrgCap	1143847.0	-6.750000e-02	7.883000e-01	-2.337200e+00	-0.4339	-0.1511	0.0159	2.327240e+01
IARER	908308.0	-inf	NaN	-inf	-0.3188	-0.1041	0.1280	2.347817e+02
IntanRet	1042857.0	1.599000e-01	8.838000e-01	-8.913700e+00	-0.2582	0.1769	0.6343	9.485900e+00
dINVoLagA	1210533.0	inf	NaN	-2.366000e+00	-0.0072	0.0014	0.0188	inf
InventGr	1210551.0	inf	NaN	-1.600690e+01	-0.1102	0.0413	0.2268	inf
Invest	1082922.0	inf	NaN	-5.338500e+00	0.5948	0.9072	1.2720	inf
LFE	1260151.0	inf	NaN	-4.939930e+01	-0.0916	0.0235	0.1477	inf
Lvrg	1436299.0	6.382130e+01	9.761273e+03	0.000000e+00	0.0683	0.3001	0.8815	2.663432e+06
LCoBP	1335533.0	1.903845e+02	2.075323e+04	-1.239853e+05	-0.1000	-0.0057	0.0813	3.096410e+06
MBaAC	1468025.0	-1.560000e-02	2.397000e-01	-1.000000e+00	0.0000	0.0000	0.0000	1.000000e+00
NDF	797361.0	7.900000e-03	1.182000e-01	-4.455200e+00	-0.0222	0.0000	0.0239	9.869000e+00
NEF	1416103.0	8.600000e-03	1.999000e-01	-1.138550e+01	-0.0178	-0.0063	0.0000	3.741130e+01
NOA	1302719.0	inf	NaN	-5.263510e+03	0.4422	0.6048	0.7698	inf
NPY	1175147.0	-4.216800e+00	1.590547e+03	-5.510098e+05	0.0000	0.0136	0.0318	3.376611e+04
dNWC	988815.0	-9.693501e+02	2.783091e+05	-7.989080e+07	-0.0212	0.0045	0.0372	8.715248e+03
dNOA	1095129.0	6.043616e+02	1.821825e+05	-1.696300e+00	0.2774	0.4194	0.5986	5.503650e+07
OperLvrg	1035457.0	inf	NaN	-4.448300e+00	0.4465	0.7684	1.1327	inf
OPoA	1035175.0	-inf	NaN	-inf	0.0483	0.0941	0.1519	1.742333e+02
OPoE	1444335.0	NaN	NaN	-inf	0.0910	0.2066	0.3949	inf
OrgCap	1145437.0	6.288100e+00	1.110744e+03	-1.570000e-02	0.1025	0.2595	0.5009	2.363308e+05
OSc	1137808.0	inf	NaN	-6.694534e+03	-4.1795	-3.0869	-1.8525	inf
PY	1134797.0	1.483300e+00	2.205329e+02	-6.671690e+01	0.0048	0.0175	0.0366	3.571822e+04
prcOA	1138403.0	NaN	NaN	-inf	-1.8700	-0.6978	0.0373	inf
prcTA	1136281.0	NaN	NaN	-inf	-0.5959	0.4110	1.1743	inf
PM	1453792.0	NaN	NaN	-inf	0.0201	0.0600	0.1290	inf
RDoMV	562509.0	5.885000e-01	2.361709e+02	0.000000e+00	0.0067	0.0206	0.0518	1.248199e+05
RDtA	519178.0	inf	NaN	0.000000e+00	0.0116	0.0380	0.0909	inf
RDtS	562823.0	inf	NaN	-7.099000e-01	0.0051	0.0164	0.0400	inf
RNOA	1168531.0	inf	NaN	-4.443354e+13	0.0511	0.1070	0.2108	inf
RoE	1462826.0	NaN	NaN	-inf	0.0218	0.0727	0.1423	inf
SalesGr	1179126.0	1.823378e+03	8.422184e+02	7.466700e+00	1222.8667	1774.1333	2361.6667	5.306800e+03
SaleToMV	1456454.0	1.085103e+03	1.988591e+05	0.000000e+00	0.4752	1.1149	2.4347	4.810779e+07
SR	1468025.0	2.447000e-01	4.299000e-01	0.000000e+00	0.0000	0.0000	0.0000	1.000000e+00
SuGr	1427050.0	NaN	NaN	-inf	-0.0527	0.0684	0.1961	inf
TAN	1273947.0	6.914000e-01	1.384700e+00	1.000000e-04	0.5717	0.6883	0.8004	3.640305e+02
TA	1061566.0	inf	NaN	-8.588103e+03	0.0286	0.0816	0.0826	inf

	count	mean	std	min	25%	50%	75%	max
ResidMom	1418029.0	2.490040e+01	8.381478e+03	-2.101682e+06	7.9873	1.253450e+01	1.825540e+01	8.551546e+06
52WH	1607117.0	7.751000e-01	1.985000e-01	0.000000e+00	0.6694	8.286000e-01	9.312000e-01	1.000000e+00
Amihud	1490638.0	1.000000e-04	8.000000e-03	0.000000e+00	0.0000	0.000000e+00	0.000000e+00	2.486900e+00
Beta	1418468.0	1.069500e+00	9.988000e-01	-1.386059e+02	0.6827	1.010000e+00	1.371900e+00	1.392699e+02
BAB	1547675.0	9.475000e-01	4.613000e-01	-4.960500e+00	0.6348	9.041000e-01	1.197600e+00	1.130220e+01
BidAsk	1529557.0	5.500000e-03	1.870000e-02	0.000000e+00	0.0013	2.500000e-03	5.000000e-03	1.045870e+01
CFV	1172643.0	1.820336e+02	2.133074e+04	0.000000e+00	0.0249	4.960000e-02	1.157000e-01	3.362591e+06
CVoST	1505422.0	1.344600e+00	1.096500e+00	0.000000e+00	0.6920	1.034500e+00	1.587800e+00	1.148910e+01
Coskew	1418284.0	-7.030000e-02	3.206000e-01	-2.553600e+00	-0.2242	-5.530000e-02	1.117000e-01	2.142200e+00
DownBeta	1583536.0	9.467000e-01	6.030000e-01	-1.979120e+01	0.5925	9.221000e-01	1.263300e+00	1.462620e+01
Age	1607117.0	1.292560e+01	6.364700e+00	2.700000e-03	7.4192	1.360000e+01	2.000820e+01	2.001370e+01
MomAge	121042.0	8.970000e-02	5.920000e-01	-9.998000e-01	-0.1233	2.940000e-02	2.097000e-01	8.126400e+01
IdioRisk	1603598.0	2.220000e-02	2.040000e-02	0.000000e+00	0.0119	1.730000e-02	2.600000e-02	1.138000e+00
IndMom	1449788.0	8.650000e-02	2.137000e-01	-6.946000e-01	-0.0284	7.170000e-02	1.738000e-01	1.194260e+01
MomLag	1547992.0	7.660000e-02	6.699000e-01	-1.000000e+00	-0.1373	1.960000e-02	1.957000e-01	1.845991e+02
LB1	1344959.0	1.031200e+00	6.629000e-01	-1.278818e+02	0.6481	9.769000e-01	1.334300e+00	2.730200e+01
LB2	1344959.0	0.000000e+00	3.000000e-04	-3.200000e-02	0.0000	0.000000e+00	0.000000e+00	3.120000e-02
LB3	1344959.0	-2.200000e-03	4.100000e-03	-2.816000e-01	-0.0031	-1.700000e-03	-8.000000e-04	6.272000e-01
LB4	1344959.0	-4.500000e-03	2.880000e-02	-5.294500e+00	-0.0059	-2.200000e-03	-6.000000e-04	2.163700e+00
LB5	1344959.0	1.037900e+00	6.667000e-01	-1.285383e+02	0.6535	9.830000e-01	1.341800e+00	2.764130e+01
LiqShck	1456330.0	0.000000e+00	1.330000e-02	-2.108200e+00	-0.0000	-0.000000e+00	0.000000e+00	1.051950e+01
LTR	1288656.0	7.388000e-01	4.021400e+00	-1.000000e+00	-0.3022	1.807000e-01	9.401000e-01	1.301714e+03
Max	1603606.0	5.960000e-02	7.160000e-02	-2.400000e-03	0.0280	4.320000e-02	6.900000e-02	4.974300e+00
Mom	1573112.0	7.580000e-02	1.159300e+00	-1.000000e+00	-0.1395	1.770000e-02	1.943000e-01	1.036467e+03
MomLTRev	1602283.0	-4.400000e-03	2.433000e-01	-1.000000e+00	0.0000	0.000000e+00	0.000000e+00	1.000000e+00
MomRev	1517272.0	7.750000e-02	6.731000e-01	-1.000000e+00	-0.1371	2.030000e-02	1.970000e-01	1.845991e+02
MomVol	739681.0	1.137000e-01	7.108000e-01	-9.977000e-01	-0.1108	5.020000e-02	2.302000e-01	1.845991e+02
PRC	1607049.0	1.269857e+02	1.330390e+03	1.000000e-04	1.3621	5.701000e+00	1.905730e+01	9.965309e+04
Seas	1548041.0	1.380000e-02	1.497000e-01	-9.912000e-01	-0.0149	1.030000e-02	3.680000e-02	9.539920e+01
Seas1A	1548039.0	1.170000e-02	2.277000e-01	-1.000000e+00	-0.0567	1.000000e-03	6.360000e-02	1.151463e+02
Seas1N	1602475.0	1.150000e-02	1.181000e-01	-9.540000e-01	-0.0138	7.700000e-03	3.070000e-02	9.539920e+01
Seas11t15A	929489.0	1.480000e-02	1.316000e-01	-9.912000e-01	-0.0255	1.030000e-02	4.790000e-02	7.923940e+01
Seas11t15N	995223.0	1.490000e-02	4.370000e-02	-7.137000e-01	0.0010	1.240000e-02	2.550000e-02	7.842100e+00
Seas16t20A	619801.0	1.310000e-02	1.475000e-01	-9.912000e-01	-0.0273	8.700000e-03	4.690000e-02	7.923940e+01
Seas16t20N	670739.0	1.370000e-02	4.550000e-02	-7.016000e-01	-0.0004	1.100000e-02	2.440000e-02	7.842100e+00
Seas2t5A	1484904.0	1.280000e-02	1.599000e-01	-9.912000e-01	-0.0299	7.600000e-03	4.660000e-02	9.539920e+01
Seas2t5N	1542966.0	1.300000e-02	1.070000e-01	-9.532000e-01	-0.0012	1.000000e-02	2.280000e-02	9.539920e+01
Seas6t10A	1227461.0	1.350000e-02	1.552000e-01	-9.912000e-01	-0.0265	9.000000e-03	4.610000e-02	9.539920e+01
Seas6t10N	1283463.0	1.420000e-02	1.142000e-01	-7.137000e-01	0.0004	1.130000e-02	2.420000e-02	9.539920e+01
SI1Y	1547714.0	-inf	NaN	-inf	0.0000	0.000000e+00	6.500000e-03	1.122490e+01
ST	1519360.0	2.016000e-01	9.431000e-01	0.000000e+00	0.0381	1.032000e-01	2.234000e-01	4.427891e+02
STR	1602426.0	1.020000e-02	4.242000e-01	-1.000000e+00	-0.0568	1.000000e-04	6.060000e-02	4.555599e+02
Size	1607111.0	3.185566e+03	1.172882e+04	1.000000e-04	163.7863	5.161238e+02	1.788754e+03	5.630556e+05
TailRisk	1418117.0	5.288000e-01	5.414000e-01	-5.339830e+01	0.2381	4.731000e-01	7.582000e-01	2.653000e+01
TV	1603606.0	2.590000e-02	2.190000e-02	0.000000e+00	0.0145	2.080000e-02	3.070000e-02	1.171000e+00
VolMV	1498504.0	9.248000e-01	7.946800e+00	0.000000e+00	0.1833	4.461000e-01	9.262000e-01	4.808495e+03
VolTrend	1279306.0	2.500000e-03	3.030000e-02	-1.622000e-01	-0.0151	3.300000e-03	2.130000e-02	1.528000e-01
VarVol	1447254.0	7.529860e+06	2.450747e+07	0.000000e+00	445071.9655	1.584363e+06	5.310762e+06	9.164343e+08

Table 3.2: Descriptive Statistics of the Frictions Anomalies

	count	mean	std	min	25%	50%	75%	max
AV	855739.0	0.3203	26.4437	-11091.1307	0.3432	0.5792	0.8357	773.8677
AC	972979.0	8.9920	7.8251	1.0000	3.0000	7.0000	13.0000	55.0000
ChiFA	948693.0	0.0009	0.5431	-1.0000	0.0000	0.0000	0.0000	1.0000
ChR	1073823.0	-0.0139	0.0902	-1.0000	0.0000	0.0000	0.0000	1.0000
dAEF	957159.0	NaN	NaN	-inf	-0.0729	0.0000	0.0431	inf
DispLTST	436189.0	NaN	NaN	-inf	-116.1868	-18.8408	14.3489	inf
DispLT	122120.0	4.7469	9.9981	0.0000	0.4314	1.9835	5.2100	419.5918
DF	986222.0	0.3414	0.4742	0.0000	0.0000	0.0000	1.0000	1.0000
EFoP	972866.0	0.0348	2.6321	-1109.1131	0.0350	0.0584	0.0844	494.3105
FD	739277.0	inf	NaN	-183.0000	0.0007	0.0081	0.0602	inf
LTGrF	447385.0	5.8889	20.0011	-979.5057	0.1296	1.7640	7.9416	5725.2200
UF	986222.0	0.3102	0.4626	0.0000	0.0000	0.0000	1.0000	1.0000

Table 3.3: Descriptive Statistics of the I/B/E/S Anomalies

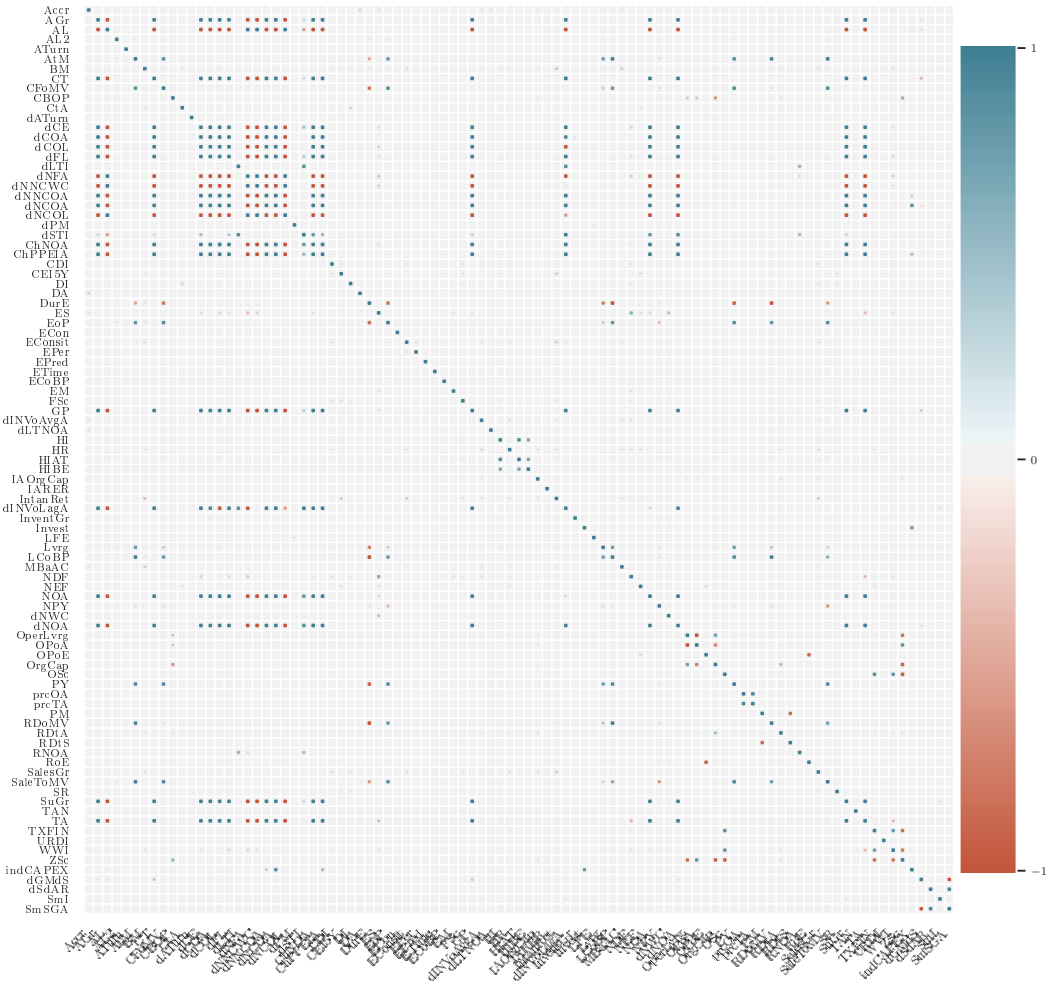

Figure 3.1: Correlation Matrix of Fundamental Anomalies

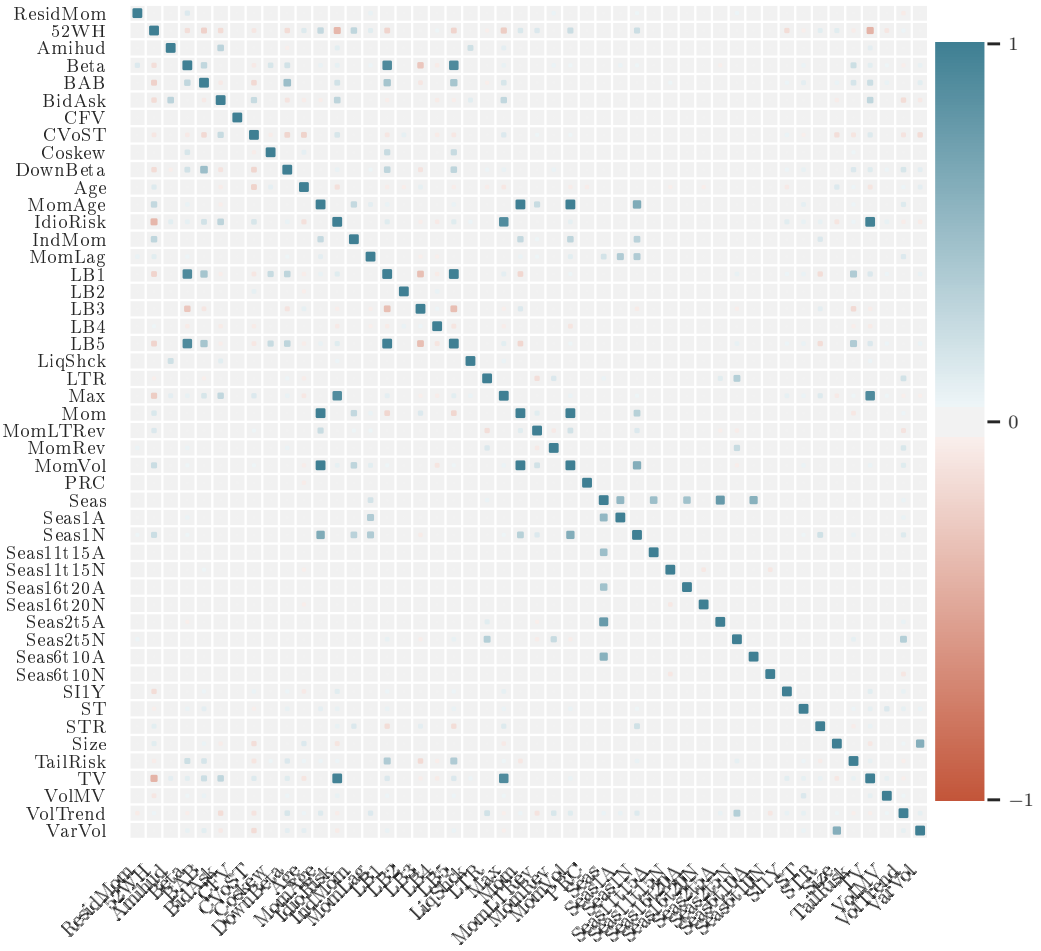

Figure 3.2: Correlation Matrix of Frictions Anomalies

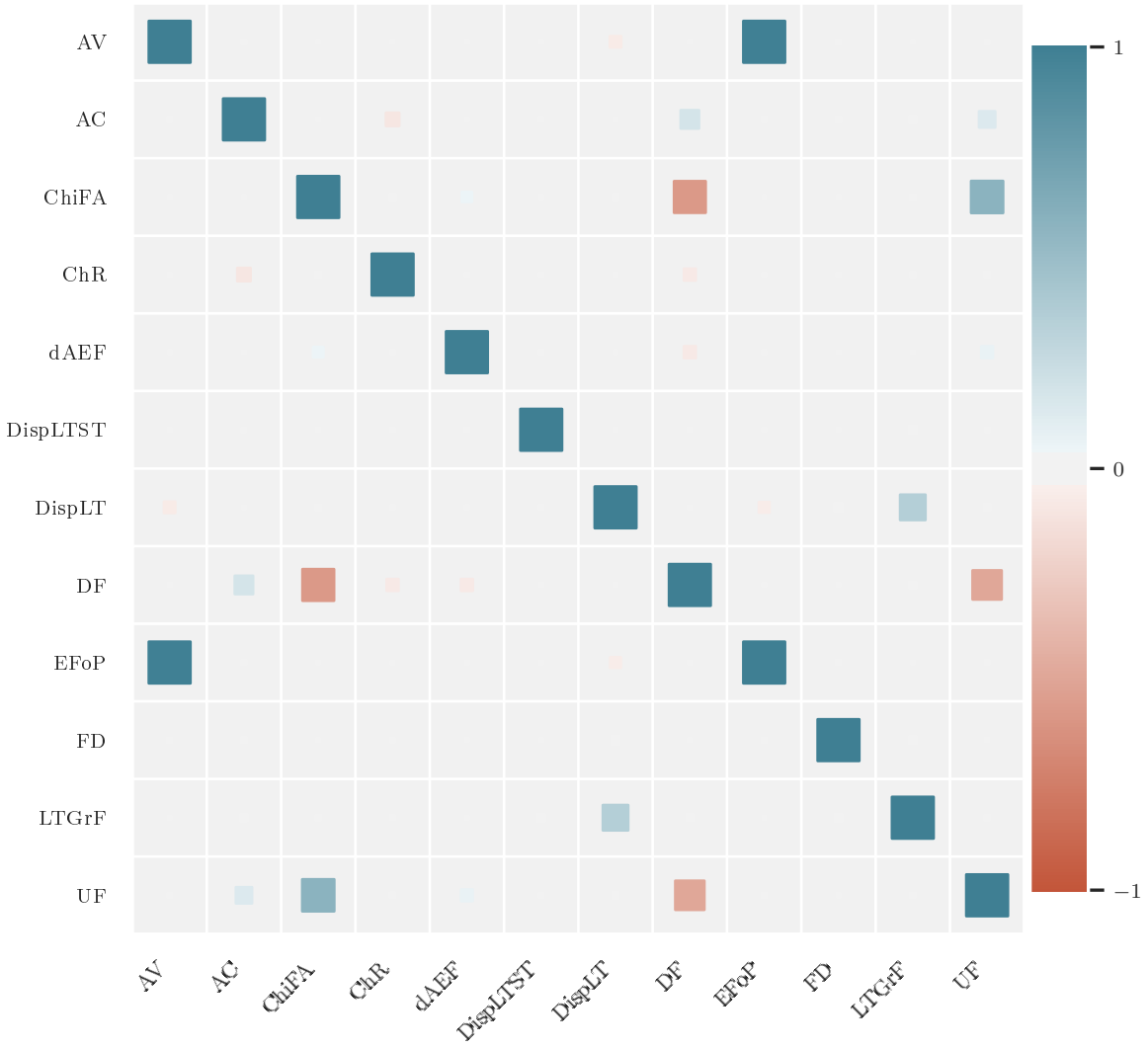

Figure 3.3: Correlation Matrix of I/B/E/S Anomalies

Figure 3.4: Histogram of Monthly Returns

Chapter 4

Results

Chapter 5

Conclusion

Bibliography

Ahn, S. C., A. R. Horenstein, & N. Wang (2012): "Determining rank of the beta matrix of a linear asset pricing model." *Technical report*, Working Paper, Arizona State University and Sogang University.

Bryzgalova, S., M. Pelger, & J. Zhu (2019): "Forest through the trees: Building cross-sections of stock returns." *Available at SSRN 3493458* .

Cochrane, J. H. (1996): "A cross-sectional test of an investment-based asset pricing model." *Journal of Political Economy* **104(3)**: pp. 572–621.

Cochrane, J. H. (2009): *Asset pricing: Revised edition*. Princeton university press.

Cochrane, J. H. (2011): "Presidential address: Discount rates." *The Journal of finance* **66(4)**: pp. 1047–1108.

De Prado, M. L. (2018): *Advances in financial machine learning*. John Wiley & Sons.

Dimopoulos, Y., P. Bourret, & S. Lek (1995): "Use of some sensitivity criteria for choosing networks with good generalization ability." *Neural Processing Letters* **2(6)**: pp. 1–4.

Fama, E. F. & K. R. French (1993): "Common risk factors in the returns on stocks and bonds." *Journal of* .

Fama, E. F. & K. R. French (1996): "Multifactor explanations of asset pricing anomalies." *The journal of finance* **51(1)**: pp. 55–84.

Fama, E. F. & K. R. French (2015): "A five-factor asset pricing model." *Journal of financial economics* **116(1)**: pp. 1–22.

Fisher, A., C. Rudin, & F. Dominici (2019): "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." *Journal of Machine Learning Research* **20(177)**: pp. 1–81.

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

Harvey, C. R., Y. Liu, & H. Zhu (2016): " and the cross-section of expected returns." *The Review of Financial Studies* **29(1)**: pp. 5–68.

Heaton, J. & D. Lucas (2000): "Portfolio choice and asset prices: The importance of entrepreneurial risk." *The journal of finance* **55(3)**: pp. 1163–1198.

Kelly, B. T., S. Pruitt, & Y. Su (2019): "Characteristics are covariances: A unified model of risk and return." *Journal of Financial Economics* **134(3)**: pp. 501–524.

Liew, J. & M. Vassalou (2000): "Can book-to-market, size and momentum be risk factors that predict economic growth?" *Journal of Financial Economics* **57(2)**: pp. 221–245.

McLean, R. D. & J. Pontiff (2016): "Does academic research destroy stock return predictability?" *The Journal of Finance* **71(1)**: pp. 5–32.

Mitchell, T. (1997): *Machine Learning*. McGraw Hill.

Molnar, C. (2020): *Interpretable Machine Learning*. Lulu. com.

Roll, R. (1977): "A critique of the asset pricing theory's tests part i: On past and potential testability of the theory." *Journal of financial economics* **4(2)**: pp. 129–176.

Rubinstein, M. (1976): "The valuation of uncertain income streams and the pricing of options." *The Bell Journal of Economics* pp. 407–425.

Shapley, L. S. (1953): "A value for n-person games." *Contributions to the Theory of Games* **2(28)**: pp. 307–317.

Štrumbelj, E. & I. Kononenko (2014): "Explaining prediction models and individual predictions with feature contributions." *Knowledge and information systems* **41(3)**: pp. 647–665.

Appendix A

Title of Appendix A

Appendix B

Internet Appendix

[TODO add github reference]