SHAPLEY VALUES, RANDOM FORESTS, AND LASSO REGRESSION FOR EXPLAINING FACTOR IMPORTANCE IN US CROSS-SECTIONAL RETURNS

Jared Cohen

An independent research project

The Kenan-Flagler Business School at

The University of North Carolina at Chapel Hill

Chapel Hill

2022

ABSTRACT

Jared Cohen

Shapley Values, Random Forests, and LASSO Regression for Explaining Factor Importance in US Cross-Sectional Returns

(Under the direction of Dr. Gill Segal)

An asset's expected return is based on a set of risk factors and the asset's exposure to each factor. A core pillar of factor investing research has been the identification of new factors that can best explain cross-sectional returns. Researchers have identified hundreds of factors, but many of these factors are redundant, containing similar information about risk. A key challenge is using this vast collection of discovered factors to determine which factors are actually the most important. I create LASSO and random forest machine learning models for this purpose due to these models' abilities to handle high-dimensional data. I use Shapley values, feature permutation, and mean decrease in impurity to evaluate feature importance for the random forest model, and I compare those results to the feature importance obtained through LASSO and OLS regression. From a set of 150 factors, I find that the momentum (UMD), earnings announcement return (ear), high-minus-low (HML), sales-to-cash ratio (salecash), and small-minus-big (SMB) factors are the most important overall. The different models produce moderate differences in factor importance rankings, and the different feature importance metrics for the random forest produce slight variations in feature importance rankings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION AND LITERATURE REVIEW

Introduction

The aim of factor investing is to model asset prices and returns, and the core idea of factor investing is that asset prices are determined by their exposures to various risk factors. The factor investing research dates back decades and includes hundreds of papers. In this literature review, I will discuss the foundational linear models of factor investing, explain the limitations of those linear factor models, survey various approaches in which machine learning is used to improve these models, discuss the use of Shapley values to increase the interpretability of machine learning models, and then explain how my research will fit into this body of literature.

Linear Models

Factor investing research relies heavily on linear models, and the majority of asset pricing research over the past decades employs linear models such as portfolio sorts. In this section, I will explain the capital asset pricing model and arbitrage pricing theory. I will then discuss the linear portfolio sort methodology and its use in the identification of factors.

The Capital Asset Pricing Model

The foundation of factor investing is the Capital Asset Pricing Model (CAPM). Developed by Sharpe (1964), CAPM is a one-factor model where an asset's returns are determined by its exposure to the market. The higher an asset's beta coefficient, the more exposure it has to this risk factor and the higher its expected return is. CAPM is a widely recognized and used model that is standard in both industry and academia to date (Jagannathan & McGrattan, 1995; Berk & van Binsbergen, 2017).

Arbitrage Pricing Theory

Arbitrage Pricing Theory (APT) was the next major development in the asset pricing literature. APT asserts that the returns of an asset can be modeled as a linear function of several factors with sensitivities to each factor being factor-specific beta coefficients (Ross, 1976). APT built on CAPM by allowing for the inclusion of several factors to better explain the sources of risk and how they affect asset prices.

The Identification of Factors Using Linear Models

Many researchers from the 1970s through today have focused on identifying factors that explain cross-sectional returns to create factor models in an APT framework. Hundreds of factors have been identified throughout the research using a portfolio sort methodology which was first introduced by Fama and French (1993). Portfolio sorting shows whether there is a relationship between a characteristic and expected returns by sorting returns by the characteristic value, dividing the assets into portfolios based on that characteristic, and then comparing differences in average returns across portfolios (Cattaneo et al., 2020). The following table displays the contributions of various researchers over time in seminal papers identifying factors and proposing factor models.

Table 1.

Contributions to Factor Model Research

Researchers	Year	Contribution
Fama & MacBeth	1973	Fama and MacBeth showed that pricing of common stocks reflected attempts of risk-averse investors to hold efficient portfolios based on expected value and dispersion of returns.
Basu	1977	Basu showed that portfolios with lower price-earnings ratios had higher risk-adjusted returns, providing evidence against the efficient market hypothesis.
Stattman	1980	Stattman demonstrated that portfolios with lower price-book values had higher returns.
Banz	1981	Banz identified the size premium whereby smaller firms have larger expected returns.
De Bondt & Thaler	1985	De Bondt and Thaler demonstrated empirically that stocks with low long-term past returns tend to have higher future returns.
Jegadeesh	1990	Jegadeesh showed empirical evidence for the predictability of individual stocks using previous returns. He showed a negative first-order serial correlation and a positive higher-order serial correlation.
Fama & French	1992	Fama and French showed that size and book-to-market equity capture the cross-sectional variation in average stock returns associated with market risk, size, leverage, book-to-market equity, and earnings-price ratios. They also showed that the relationship between market risk and average return in flat when tests allow for variation in market risk that is unrelated to size.
Fama & French	1993	Fama & French identified the value premium whereby firms with lower Price/Book values have higher expected returns. They also proposed a 3-factor model with market risk, size, and value.
Jegadeesh & Titman	1993	Jegadeesh and Titman identified a high momentum premium whereby firms with higher past returns have higher expected future returns.
Fama & French	1996	When looking at patterns in returns not explained by CAPM from firm characteristics such as size, earnings/price, cash flow/price, book-to-market equity, past sales growth, long-term past return, and short-term past return, Fama and French found that by using their 3-factor model instead of CAPM, these anomalies largely disappeared except for short-term momentum.

Carhart	1997	Carhart proposed a 4-factor model with market risk, size, value, and momentum.
Fama & French	2015	Fama & French updated their 3-factor model by proposing a 5-factor model with market risk, value, size, profitability, and firm investment.
Papenkov	2019	Papenkov demonstrated meaningfully heterogeneous risk across sectors for each factor in the Fama-French five-factor model. The sector-heterogenous model improved R-squared by 5% on average.

Much of the asset pricing literature from the 1970s through the 2010s focused on exploring the identification and testing of additional factors and attempting to create parsimonious models to explain as much variation as possible with as few factors as possible.

Limitations of Linear Models

The accuracy and understandability of linear models often deteriorate in high-dimensional space (Feng et al., 2020; Bryzgalova et al., 2020; Gu et al., 2020; Chen et al., 2021). In this section, I will explain the shortcomings of previous linear models related to the robustness and reproducibility of results, the construction of test assets, and the curse of dimensionality arising from the multitude of factors.

Robustness and Reproducibility of Results

Concerns over robustness and reproducibility of results arise from the testing of many factors and the variability in data and test asset construction. Harvey et al. (2016) argued the usual cutoff levels for statistical significance may not be appropriate given the known number of factors that have been tried and the reasonable assumption that many more factors have been tried but were not published. Significant results are likely found by testing many hypotheses without controlling the false discovery rate. With so many factors being tested, some factors appear significant due to chance (Harvey et al., 2016, p. 36).

Challenges to reproducibility arise from a lack of internal validity that is not robust to slightly different methodologies or data (Jensen et al., 2021). Additionally, models such as the Fama-French 3-factor model are used as benchmark models to evaluate whether new factors add explanatory power; however, this decision imposes the unrealistic assumption that the selected model is the true one and does not miss any additional factors (Feng et al., 2020, p. 1336). Selecting an incorrect model is problematic since it can lead to omitted variable bias when useful factors are not included or to efficiency loss when many useless or redundant factors are included.

Constructing Test Assets

Researchers have identified challenges with the portfolio sort method, especially when dealing with a large number of factors (Moritz & Zimmermann, 2016). In this method, the researcher sorts stocks into three to ten portfolios each month based on the value of a particular variable. In the next step, subsequent returns for each portfolio are calculated, checking whether there is a monotone relation between the sorting variable and these subsequent portfolio returns. The relevance of the sorting variable is then assessed by comparing the return to some equilibrium model of asset prices (such as the capital asset pricing model) or by assessing the monotonicity of the returns over deciles. To sort by a second characteristic, the stocks are further split into three to ten groups by the new variable. The importance of that second characteristic is determined by assessing the behavior of returns over the deciles for that second characteristic (Fama & French, 1993).

This portfolio sort methodology is a powerful, nonparametric tool that works best in low-dimensional cases; however, problems arise when sorting on more variables since the portfolios

will have fewer stocks the more variables are added. For example, triple-sorted 10x10x10 portfolios would lead to the creation of 1,000 test portfolios. Each portfolio would only have a few stocks in it which would be problematic due to an inability to diversify away the idiosyncratic risk. Considering the presence of hundreds of factors, this portfolio sort methodology is not feasible in the higher dimensional spaces of all factors (Moritz & Zimmermann, 2016).

Beyond the issues in higher dimensional space, Bryzgalova et al. (2020) have also shown that the conventional sorting-based portfolios fail to span the stochastic discount factor, thus leading to the wrong conclusions when used to evaluate or construct asset pricing models. These conventional cross-sections do not reflect the joint effect of multiple characteristics, neglecting their interactions. This problem arises even in low dimensional space, and stacking additional sorts against each other compounds this problem (Bryzgalova et al., 2020, p. 1). Similarly, Feng et al. (2020) have also demonstrated that these models have generally poor performance in explaining a large available cross-section of expected returns beyond 25 size and value-sorted portfolios, indicating omitted factors are likely to be present in the data.

The Curse of Dimensionality / Too Many Factors for Linear Models

Researchers have identified that standard linear models are not well-suited for high dimensionality (a large number of predictor variables) since linear models are unable to account for variable interactions and non-linear effects, which have been identified as important in asset pricing (Bryzgalova et al., 2020; Gu et al., 2020; Chen et al., 2021). Standard linear models also become inefficient or ineffective when the number of factors approaches the number of observations in the data (Wang et al., 2016; Feng et al., 2020, p. 1335; Gu et al., 2020, p. 2234).

Linear models cannot properly incorporate all identified factors, yet Kozak et al. (2018) demonstrate that linear four or five-factor models are also unable to adequately explain the cross-section of expected stock returns. Standard methodologies do not work well in a high-dimensional setting due to this curse of dimensionality, but including a large amount of information is important to produce more accurate models. Feng et al. (2020) thus argue that dimension-reduction and regularization techniques are needed for valid inference in this high-dimensional factor space.

Machine Learning

Machine learning models are well-suited for dealing with the high-dimensionality of factor data due to allowances for nonlinearity, regularization, and interaction effects. By moving beyond the traditional portfolio sorts and linear regression methodology, the existing factors can be tested in a new way, creating more accurate models (Gu et al. 2020; Feng et al., 2020). In this section, I will discuss machine learning approaches for constructing test assets, evaluating new and existing factors, and selecting models.

Methods for Constructing Test Assets

Machine learning approaches allow researchers to create test assets in a way that spans the stochastic discount factor and solve the problems of the standard portfolio sort methodology. Moritz and Zimmermann (2016) created a novel approach to grouping individual stocks into managed portfolios that reflect the information in a given set of characteristics. They used conditional portfolio sorts where the sorting variable and value need to be estimated. They made AP-Trees, which deliver a small cross-section of interpretable, well-diversified portfolios that provide a robust span of the SDF, conditional on many characteristics. The method allows for

flexible variable selection at each branch by using sorts deeper than two levels, which allows for the creation of better test assets. Bryzgalova et al. (2020) argued that the choice of test assets has a large influence on results since expected returns are explained by the returns of test assets and the model that succeeds in pricing them. While most of the literature has historically focused on the pricing aspect as discussed in the preceding sections, choosing the right test assets is equally crucial (Bryzgalova et al., 2020).

Methods for Evaluating Given Factors

Statistical techniques, machine learning regression models, and machine learning clustering algorithms can be used to evaluate the importance of given factors among the set of hundreds of factors. For example, Harvey et al. (2016) provide a multiple testing statistical framework for factor significance where the t-statistic must be greater than 3.0 to be significant (rather than the 2.0 typical cutoff). This statistical method evaluates previously discovered factors with stricter conditions, which leads to fewer factors being significant, condensing the factor set to those with higher significance.

In a machine learning regression model approach, Feng et al. (2020) created a LASSO regression-based methodology for estimating and testing the marginal importance of any factor in pricing the cross-section of expected returns beyond what the existing factors can explain. This methodology provides a framework for assessing the importance of a given factor after accounting for all other identified factors and can help filter redundant factors.

Jensen et al. (2021) created a machine learning classification model to algorithmically classify factors into one of 13 themes. These themes possess a high degree of within-theme correlation and conceptual economic similarity while inter-theme correlation is low. This factor taxonomy allows for viewing factors as 13 highly correlated clusters rather than hundreds of

individual factors. Factors that are economically similar likely contain the same information about risk, so the methodology of Jensen et al. (2021) provides a systematic framework for condensing the factor set in a mathematically sound and conceptually intuitive way.

Methods for Selecting Models

Another branch of research focuses on using machine learning to estimate a factor model rather than creating frameworks for evaluating given factors. As shown in Table 2, researchers have employed various machine learning methods such as neural networks, principal component analysis, LASSO and other generalized linear models, and tree-based models in order to process the multitude of factors to find key drivers and propose parsimonious models (Messmer, 2017; Kozak et al., 2018; Wolff & Neugebauer, 2019; Feng et al., 2020; Gu et al., 2020; Chen et al., 2021).

Table 2.

Machine Learning Methods for Selecting Factor Models

Researchers	Year	Approach & Contribution
Messmer	2017	Messmer used a deep feedforward neural network based on 68 firm characteristics to predict the cross-section of US stock returns. Messmer found that long-short portfolios can generate attractive risk-adjusted returns compared to linear benchmarks, and the results are robust to size, weighting schemes, and portfolio cutoff points. Messmer identified that price-related characteristics, like short-term reversal and 12-month momentum, are the main drivers of the return prediction while the majority of firm characteristics are of little importance.
Kozak et al.	2018	Kozak et al. showed that a small number of principal components of the universe of potential characteristics-based factors can approximate the stochastic discount factor well.
Wolff & Neugebauer	2019	Wolff and Neugebauer used general linear models and tree-based machine learning models to predict equity premiums based on fundamental, macroeconomic, sentiment, and risk data. They found that linear models such as penalized least squares and principal component regression outperformed the benchmark while other ML models used failed to outperform the benchmark. An investment strategy using machine learning prediction in a market timing strategy, however, outperformed a buy-and-hold investment.
Feng et al.	2020	Feng et al. imposed lower dimensionality on a model using LASSO regression.
Gu et al.	2020	Gu et al. compared the performance of various machine learning methods for asset pricing and showed large economic gains emerged from using ML-based forecasts. Neural networks and trees were the best performing methods due to the allowance for nonlinear predictor interactions that other methods miss, and all methods identified variations in momentum, liquidity, and volatility as the dominant predictive signals.
Chen et al.	2021	Chen et al. used deep neural networks to estimate an asset pricing model for individual stock returns.

Neural networks and tree-based models were most often shown to be the best models, likely due to an allowance for nonlinear effects; however, these models suffer from lower interpretability (Messmer, 2017; Gu et al. 2020; Chen et al., 2021). Additionally, some papers show that generalized linear models such as penalized least squares, principal components, and LASSO regression actually outperformed other machine learning models like trees (Wolff & Neugebauer, 2019; Feng et al., 2020). Across different models, momentum-related factors were often the largest drivers of returns while short-term reversal, liquidity, and volatility were also dominant return drivers (Messmer, 2017; Gu et al., 2020).

Shapley Values for Increasing Model Interpretability

Shapley values can explain how much each feature contributes to the value of a model's prediction using a game-theoretical method for assigning payouts to players depending on their contribution to the total payout of a game (Molnar, 2022, p. 215). In this section, I will explain the game-theoretical basis of Shapley values, the application of Shapley values to model interpretability, and the previous application of Shapley values in the finance literature.

The Game Theoretical Basis of Shapley Values

Shapley values are the payouts that each player in a game receives for their contribution to the total payout of the game, and these values rely on certain assumptions and axioms. This method assumes that utility is objective and transferable and that games are cooperative affairs and adequately represented by their characteristic functions (Shapley, 1952, p. 1). Another assumption defines a game as a set of rules with specified players in the playing positions, and rules describe an abstract game (Shapley, 1952, p. 2). The value of a game is a function that associates each player with a real number and satisfies the following axioms:

1. Symmetry: The value is a property of an abstract game.
2. Efficiency: The value represents a distribution of the full yield of the game.
3. Law of Aggregation: When two independent games are combined, their values must be added player by player (Shapley, 1952, p. 4).

Players cooperate in a coalition and receive a certain profit from this cooperation. First, the players agree to play a game in a grand coalition. Starting with a single player, the coalition randomly adds one member at a time until all N players have been admitted. Each player demands and is promised the amount which their adherence adds to the value of the coalition as determined by the function ∇. The grand coalition then plays the game efficiently, achieving the amount ∇(N), which is just enough to meet the total amount promised (Shapley, 1952, p. 13).

Application of Shapley Values for Model Interpretability

Shapley values give a measure of how much each feature contributes to the difference between an observation's predicted value and the average prediction for all observations. Each feature value of an observation is a player in a game where the prediction is the payout. The Shapley value is the average marginal contribution of a feature value across all possible combinations (coalitions) of features (players). Shapley values thus determine how to fairly distribute the payout among features based on how much each feature contributes to explaining the difference between the predicted value of a given observation and the average prediction across all observations. That difference equals the sum of these contributions across each feature, so the results are easily comparable and interpretable. The larger the payout a given feature receives, the more important that feature is in driving predictions. (Molnar, 2022, p. 215).

Shapley Values in Finance Literature

Shapley values have been used in finance research to provide alternate measures of risk and to explain price predictions. Mussard and Terraza (2008) used Shapley values to decompose portfolio risk as measured by sample covariance. Those values allowed for classifying the securities in portfolios according to risk scales. The Shapley values expressed how much each security in the portfolio contributed to overall portfolio risk, including both systematic and idiosyncratic risk. Ortmann (2016) similarly demonstrated that Shapley values can be used to decompose market risk, specifically the beta factor in CAPM. Ortmann explained that a given asset's beta factor can be interpreted as that asset's share of market risk or as that asset's average marginal contribution to market risk. Through this linking of Shapley values and the beta factor, Ortmann claimed that Shapley values could lead to a deeper understanding of systematic risk and the beta factor. Shalit (2020) further explored portfolio risk decomposition by using Shapley values to quantity relative risk of securities in optimal portfolios, comparing the risk ranking derived from Shapley values to that derived from betas. Systematic risk measured as the relative covariance of stock returns plays a large role in pricing securities, but estimating this beta value is difficult. By viewing portfolios as cooperative games where players (assets) are playing to minimize risk, investors can calculate the exact amount that each asset contributes to portfolio risk. This decomposition of risk using Shapley values for mean-variance and mean-Gini efficient portfolios provides a better ranking of assets by their total contribution to the risk of an optimal portfolio (Shalit, 2020; Shalit, 2021). In another application of risk decomposition, Tarashev et al. (2016) used Shapley values to assess the allocation of system-wide risk to individual banking institutions to provide a measure of their systemic importance, and they found that size is the main determinant of systemic importance for banks. Lastly, Giudici and Raffinetti (2021) used

Shapley values to derive variable importance in a machine learning model for predicting the price of Bitcoin.

The Need for My Research

My research will be the first application of Shapley values to interpret feature importance in machine learning factor models. My research will address the limitations of standard linear models, the conflicting evidence about the best machine learning models, and the lack of interpretability for machine learning models.

Addressing Limitations of Linear Models

I will address the high-dimensionality problems associated with standard linear models by using the LASSO regression and random forest regression machine learning models within the Fama and MacBeth (1973) double-pass regression framework. I will use standard sorted portfolios and use ordinary least squares first-pass regressions to obtain each asset's exposure to each factor. However, instead of using another linear second-pass regression to determine which factors are priced, I will substitute LASSO and random forest regression models. Both models should offer improvements due to regularization (LASSO) and nonlinearity (random forest) while acting as a sound comparable due to using traditional sorted portfolios as test assets.

Machine Learning Model Comparative Efficacy

Previous literature disagrees on whether generalized linear models or nonlinear models outperform, so I will use both a LASSO (generalized linear) and random forest (nonlinear) model to provide additional evidence. For example, Wolff and Neugebauer (2019) and Feng et al. (2020) supports the use of general linear models like LASSO. Conversely, Messmer (2017), Gu et al. (2020), and Chen et al. (2021) show outperformance for neural networks and tree-based

models. By comparing the accuracy and factor importances resulting from each model, I will contribute valuable information about how these models behave on this data.

Shapley Values for Increasing Model Interpretability

I will employ a novel use of Shapley values to determine feature importance for machine learning factor models to address the lack of interpretability common in higher-performing machine learning models. By using Shapley values, I will determine how much each factor contributes to the difference between an asset's return and the average return, resulting in a ranking of factor importance. Shapley values have previously been applied to decomposing risk within a portfolio to determine how much each asset contributes to portfolio risk (Mussard & Terraza, 2008; Ortmann, 2016; Shalit, 2020; Shalit, 2021); however, no one has used Shapley values to decompose stock returns to determine the contributions of various factors in driving returns. My research will not necessarily infer a parsimonious model; however, it will determine which factors tend to have the strongest effect on returns and provide a new method of testing factor importance that can be used in conjunction with any model.

METHODOLOGY

Introduction

My research aims to explain variations in stock returns using LASSO and random forest machine learning models that address many limitations of traditional factor investing research. I will first download all necessary data and import it into Python to prepare it for analysis. Next, I will construct test assets, build models, and then analyze the output data. In the following sections, I will discuss my plan for data collection and cleaning, provide background information on factor investing, detail my research design, and explain the rationale of my design.

Data Collection and Cleaning

My research involves building models to explain variations in asset returns based on factors constructed from fundamental stock data. The primary types of data I need are monthly returns of test assets and monthly values of factors. Empirical finance researchers prefer factors to be robust over time, so I will use data as far back as available. For monthly test asset return data, I will use the following datasets from Kenneth French's website: "10 Portfolios formed on momentum," "25 Portfolios Formed on Size and Book-to-Market (5 x 5)," and "48 Industry Portfolios." I will also use the risk-free rate from the dataset "Fama/French 3 Factors" from French's website. These datasets include data from January 1927 to February 2022. For factor data, I will use a dataset provided by Feng et al. (2020) containing monthly values for 150 different factors from July 1976 to December 2017. A summary of this dataset is provided in the Appendix. This dataset is built using data from the CRSP and Compustat databases. I will use data from July 1976 through December 2017 for my analysis because that is the time period covered by all of my datasets.

Background on Factor Investing Research

The core idea of factor investing is that asset prices are determined by their exposures to various risk factors. Researchers in asset pricing and factor investing have thus pursued the identification of such factors in order to create models that better estimate asset prices and returns. I will explain the construction of test assets and factors, traditional Fama-MacBeth regressions, and the issue of factor proliferation.

Test Asset and Factor Construction

Test assets' returns are used to construct factors, and researchers typically use characteristic-sorted portfolios rather than individual stocks in order to diversify out the idiosyncratic risk (Fama & French, 1993; Cattaneo et al., 2020). For example, a 3x2 double sorted portfolio on size and value would involve splitting all stocks into three buckets based on market capitalization and then further separating those buckets based on price to book ratio. The differences in monthly returns of these portfolios are used to construct the factors.

Fama-MacBeth Regressions

Traditional factor investing research uses a two-pass linear regression methodology employed by Fama and MacBeth (1973). First, monthly returns for each asset are regressed on various factor data. Then, average test asset monthly returns are regressed on the beta coefficients of each factor determined in the first set of regressions. The coefficients from the second regression are the risk premia for each factor, and if the coefficient is statistically significant, the factor is a priced factor (Fama & MacBeth, 1973).

Factor Proliferation

Asset pricing research has led to hundreds of potential factors for explaining cross-sectional stock returns, and more advanced models are needed to sort through these factors (Feng et al., 2020; Bryzgalova et al., 2020; Gu et al., 2020; Chen et al., 2021). Linear models are poorly suited for high-dimensional space, and this high-dimensionality makes statistical inference difficult (Wang et al., 2016; Gu et al., 2020, p. 2234). Variable selection techniques can be useful in reducing dimensionality but produce poor estimates unless appropriate econometric methods are used to account for model selection mistakes. For example, Feng et al. (2020) combine a double-selection LASSO econometric method with the Fama-MacBeth two-pass regressions to evaluate the marginal contribution of a factor to explaining asset prices. Determining whether new factors can add explanatory power beyond the previously discovered factors is important for creating accurate models without redundancy or unnecessary complexity.

Research Design

My research will extend the two-pass regression methodology of Fama and MacBeth (1973) by substituting machine learning models for the second regression. My research also draws from the methodology of Feng et al. (2020) by using LASSO regression in conjunction with Fama-MacBeth regressions to determine the marginal contribution of a factor for explaining the cross-section of returns. My research design will involve several phases: test asset selection, first-pass linear regression, second-pass linear regression, second-pass LASSO regression, second-pass random forest regression, calculation of mean decrease in impurity and feature permutation feature importance, and computation of Shapley values.

Selecting Test Assets

I will use pre-constructed test assets from Kenneth French's website. These test assets include 25 5x5 double-sorted portfolios on size and book-to-market, 10 portfolios sorted on momentum, and 48 portfolios sorted by industry. These three datasets contain portfolios constructed in various ways, so insights derived from these test assets should be generalizable.

First-Pass Linear Regression

I will obtain every test asset's exposure to each factor by regressing the monthly returns of test assets on the values of each factor. I will run these regressions using the LinearRegression module from Python's scikit-learn machine learning library. This process will entail running 83 regressions—one regression for each test asset—giving me the factor exposures for each asset.

Second-Pass Linear Regression

A cross-sectional regression of test asset average returns on the assets' factor exposures will then determine the coefficients and significance of each factor. Python's scikit-learn will again be used for this regression. The coefficients in this regression correspond to risk premia (the market prices of risk), so larger values indicate more important factors. This ordinary least square regression model, however, will be limited since the number of observations (83) is less than the number of independent variables (150).

Second-Pass LASSO Regression

LASSO regression produces a parsimonious model by selecting the few factors that best explain the cross-section of returns, and I will employ a LASSO for the second regression to determine which of the 150 factors are most important. LASSO regression is a machine learning model that extends multilinear regression by penalizing the inclusion of more predictor variables.

The LASSO regression will drop less important variables by scaling them to zero while the more important variables remain in the model. A regularization parameter (lambda) is used to indicate how strict the penalty will be: a stricter penalty means more variables will be dropped. A lambda of 0 is equivalent to an ordinary least square regression model, while the strictest lambda, 1, scales all coefficients to zero. Using the Lasso module in scikit-learn, I will construct LASSO models with lambda values ranging from 0 to 1 in increments of 0.001 in order to observe which factors are selected as the penalty becomes stricter. The most important factors will be selected at the highest lambda level, and the next most important factors will be revealed as the lambda value is progressively lowered.

Second-Pass Random Forest Regression

Random forest regressors combine many independent, nonlinear tree models into a single prediction to reduce variance and increase accuracy. Each tree within the model samples features and observations randomly from the dataset, resulting in different trees. An individual tree is constructed by splitting the data at nodes in a way that produces the most separation between observations in the left node and the right node. These splitting conditions are determined from the features in the dataset (e.g., the left node has data where feature "a" is less than "x" and the right node would include data where feature "a" is greater than or equal to "x"). This splitting process repeats until a specified depth level or until the tree can no longer produce heterogeneous child nodes. Large variations occur between the trees since the individual trees can only choose from a random subset of features, resulting in lower correlation across trees. A random forest will average the predictions that each tree independently generates to produce an aggregate output.

I will create a random forest regression model to predict test asset returns using factor exposures. Using the RandomForestRegressor module in scikit-learn, I will construct a random forest model containing 1,000 individual decision trees. The maximum depth will not be specified, so the trees will continue splitting until the random sample of data used by a given tree has no more variation to split on.

Mean Impurity Decrease and Feature Permutation Factor Importance

I will calculate measures of feature importance for the random forest using the mean decrease in impurity and feature permutation. To find the mean decrease in impurity, I will use the built-in "feature_importances_" attribute of a model made using the RandomForestRegressor module in Python's scikit-learn library. For feature permutation, I will utilize Python's permutation_importance module in scikit-learn. These methods derive feature importance based on two different approaches for evaluating a feature's contribution to the model: one based on how much including a factor improves a model (mean decrease in impurity) and one based on how much excluding a factor hurts a model (feature permutation).

Computation of Shapley Values

Shapley values represent how much each factor contributes to the difference between the predicted return of a given asset and the average return across all test assets. I will compute the mean absolute Shapley value for each factor in my random forest model to determine the overall importance that each factor has in the model. Using the TreeExplainer method in the SHAP Python package, I will compute every factor's Shapley value for each observation. The Shapley value for each feature acts as a force to increase or decrease the prediction. The prediction starts at a baseline, which will be the average monthly return across my test assets. For a given

observation, each feature's Shapley value will drive the prediction away from baseline such that the cumulative effect of all features leads to the actual prediction for that observation. The size of a factor's Shapley value indicates how much of the variation between an observation's predicted return and the average return can be attributed to that factor. The mean absolute value of Shapley values for a factor can measure the overall importance of that factor in the model. The higher the mean absolute Shapley value, the more important that factor is in explaining cross-sectional returns (Molnar, 2022).

A factor's Shapley value is its contribution to the difference between the average test asset return and an individual observation's predicted test asset return, weighted and summed over every possible combination of feature values. This computation is performed by evaluating all possible sets of feature values with and without a given marginal feature. To calculate the Shapley values, the first step is to select a feature, j, an instance, x, and a number of iterations, M. In each iteration, a random observation, z, is selected from the data, and a random order of features is chosen. Two new instances are then created by combining values from the instances z and x: X+j and X-j. X+j is similar to x, but all feature values occurring after feature j based on the random order are replaced by the corresponding feature value in z. X-j is the same as X+j, except the value for feature j from z is used instead of the value for feature j from x. The difference between the predictions of these instances, X+j and X-j, is then computed, and this difference corresponds to the marginal contribution of feature j. The marginal contributions from each iteration are then averaged together to produce the Shapely value for feature j and observation x, and the number of iterations should be sufficiently large such that every permutation of features and every instance are randomly selected at least once. This procedure must be repeated for every feature and for every observation in the data in order to obtain each

feature's Shapley value for every observation, and the SHAP package takes care of all of these calculations. Computing the mean absolute value of Shapley values for each feature across all observations can then provide a measure of overall feature importance (Molnar, 2022).

Rationale

This research design extends the widely used factor investing methods of Fama and MacBeth (1973) by incorporating machine learning methods to address the challenges presented by the vast array of previously identified factors. Below, I will explain the benefits of the LASSO and random forest models, compare approaches to feature importance, and then discuss the limitations of my methodology.

Benefits of LASSO and Random Forest Models

Machine learning models, like LASSO and random forests, are well-suited for dealing with the high-dimensionality of factor data due to allowances for nonlinearity, regularization, and interaction effects (Gu et al., 2020; Feng et al., 2020). LASSO is effective for handling a large number of predictor variables to determine a smaller number of driving factors. The model is also easily interpretable since it is a type of linear model (Feng et al., 2020). Random forests are an effective model for factor data due to their nonlinearity and allowance for complex interaction effects between features (Gu et al., 2020). Random forests are not easily interpretable, however, so methods, such as mean impurity decrease, feature permutation, and Shapley values, are needed to understand feature importance within the model. Using LASSO and random forests can lead to better models that uncover insights into which factors drive the most variation in cross-sectional returns.

Comparative Approaches to Feature Importance

I will compare the feature importance obtained through OLS coefficients and LASSO selection to the feature importance of random forests obtained through the mean decrease in impurity, feature permutation, and Shapley values in order to analyze the persistence or variation of feature importance across models and feature importance metrics.

I will evaluate feature importance for the ordinary least squares model by comparing the regression coefficients. These coefficients are the risk premia for each factor. Comparing coefficient size is acceptable since the units of all independent data are the same. However, looking at coefficients is imperfect since they represent the price per unit risk, and the amount of risk may vary between factors. The expected asset return is the sum of all factors times their respective risk premiums, so the product of the factor value and factor risk premium is what better represents the importance of a factor in explaining returns. I will thus use coefficient size as an imperfect proxy for factor importance.

LASSO determines feature importance by eliminating less important features from the model and selecting only the most important factors. The stricter the lambda penalty, the fewer features remain and the more important those features are. I will rank factors by importance based on their order of addition to the model as the penalty term shrinks.

Impurity-based feature importance is based on the decrease of Gini impurity when a feature is used to split a node. Gini impurity, ranging from 0 to 0.5, is a measurement of the likelihood of misclassification of a new instance of a random variable. Gini impurity is used when constructing decision trees, as minimizing Gini impurity leads to the choice of which feature splits a node. Feature importance is thus measured by the magnitude of the decrease in Gini impurity when a given feature is chosen to split a node. The decreases in Gini impurity

whenever a given feature is chosen to split a node across every tree are summed, and this sum is divided by the number of trees in the forest. This average decrease in Gini impurity gives a measure of relative feature importance.

Feature permutation determines feature importance by measuring the decrease in model performance when a given feature is randomly shuffled. This random shuffling breaks the relationship between that feature and the dependent variable. A decrease in model score thus indicates how much the model depends on that feature, and the more model performance drops, the higher the importance is of that feature.

Shapley values measure feature importance by attributing a portion of variation to each feature. The mean absolute value of Shapley values for a factor can measure the overall importance of that factor in the model. The higher the mean absolute Shapley value, the more important that factor is in explaining cross-sectional returns. The benefit of Shapley values is their greater interpretability. The Shapley value of a feature for a single observation gives the amount by which that feature moves the average prediction towards the expected prediction. The mean absolute Shapley value for a feature shows how much that feature contributes to the actual predictions of each observation on average. Shapley values are thus easily interpretable because they are units of the target variable: monthly returns of test assets.

Addressing Limitations

The primary limitations of this methodology center around lack of inference for feature importance, lack of robustness for LASSO, and competition between correlated features for Shapley values. First, none of the measures of feature importance are technically statistical inferences. Shapley values, feature permutation feature importance, and mean decrease in impurity are not supposed to be methods of inference; however, recognizing this fact is

important. Also, parsimonious models selected through LASSO may not be robust without econometric modifications, which I do not use. As explained by Feng et al. (2020), changing the LASSO penalty value would likely change the output of which factors were most important. Finally, the Shapley values obtained may not be ideal because attribution must be split between economically similar factors. If the data contains more factors within one economic theme than another, then the total Shapley values attributed to factors within the former theme will be divided by a larger number, resulting in any given factor having a lower Shapley value. One potential method to address this issue would be to incorporate clustering of economically similar factors like in the methodology of Jensen et al. (2020).

RESULTS

Introduction

In this section, I explore the performance of my second-stage regression models and describe the features selected as important by each model. I will first describe the output of my ordinary linear model. Next, I will discuss the performance and results of my LASSO model with varying penalty parameter values. I will then analyze the output of my random forest regression model. Lastly, I will compare feature importances obtained through OLS, LASSO, mean decrease in impurity for random forest, feature permutation for random forests, and Shapley values for random forests.

Ordinary Least Squares Regression Model

The second-stage ordinary least squares multilinear regression with all 150 factors had a perfect R-squared of 1 but was plagued by statistical concerns. First, while a higher R-squared typically signals that a model is accurate, in this case, the R-squared is 1 by virtue of the fact that the model contains 150 independent variables. Adding additional independent variables to an OLS regression can only increase R-squared. Even a random variable with no correlation to the output variable will still slightly increase R-squared. Due to this effect, the 150 factors in this model drove R-squared up to 1 despite any lack of actual prediction power. The model is also unable to conduct hypothesis tests due to the lack of any degrees of freedom. Since the number of features (150) exceeds the number of observations (83), the measures of statistical inference break down. As **Figure 1** illustrates, the standard errors are infinite, and t-statistics are zero while p-values and confidence intervals are undefined. The only information obtainable from this

regression is the coefficient values; however, the statistical significance of any of these coefficients is indeterminate, so the trustworthiness of this data is low.

Figure 1.

Second-Stage OLS Regression Output

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     nan
Dependent Variable: Asset Returns    AIC:                -5253.8663
Date:               2022-05-04 19:50 BIC:                -5053.1026
No. Observations:   83               Log-Likelihood:     2709.9
Df Model:           82               F-statistic:        0.000
Df Residuals:       0                Prob (F-statistic): nan
R-squared:          1.000            Scale:              inf
------------------------------------------------------------------
                   Coef.   Std.Err.     t     P>|t|  [0.025  0.975]
------------------------------------------------------------------
const               0.4807       inf   0.0000    nan     nan     nan
MktRf               0.2325       inf   0.0000    nan     nan     nan
beta                0.0154       inf   0.0000    nan     nan     nan
ep                  0.0864       inf   0.0000    nan     nan     nan
dy                  0.0453       inf   0.0000    nan     nan     nan
sue                 0.1041       inf   0.0000    nan     nan     nan
pps                 0.0276       inf   0.0000    nan     nan     nan
LTR                 0.0318       inf   0.0000    nan     nan     nan
lev                -0.0399       inf  -0.0000    nan     nan     nan
```

*Note***.** The output for the coefficients is limited to the first nine rows in order to demonstrate the issues with the output without using unnecessary space.

Factors cannot be compared based on significance levels or marginal contribution to R-squared in this model, so the only way to evaluate factor importance is by comparing the magnitude of the coefficients. As displayed in **Table 3**, the corporate investment (cinvest) factor has the largest coefficient by a wide margin. The return on invested capital (roic), convertible debt indicator (convind), organizational capital (orgcap), earnings announcement return (ear), and momentum (UMD) factors have the next largest coefficients by absolute value. **Table A1** contains a key for all factor abbreviations.

Table 3.

Factors with Highest Absolute Value OLS Coefficients

Factor	**Coefficient**
cinvest	0.416292
roic	0.246895
convind	0.238540
orgcap	0.234098
ear	-0.216295
UMD	0.212912
herf	0.207966
mom36m	0.202833
ol	-0.195675
absacc	-0.195352
quick	-0.192430
HML	0.191448
SMB	0.188681
ala	0.185452
poa	-0.171021
pm	-0.170793
dcoa	-0.147339
std_dolvol	-0.146275
etr	-0.145250
zerotrade	-0.144881

Note. Negative coefficients are red, and positive coefficients are black for readability

LASSO Regression Model

I ran many LASSO regression models to analyze the behavior of model performance and feature selection with varying regularization penalty levels. I first tested LASSO models with penalty values ranging from 0 to 1 at increments of 0.01. A penalty of 1 always shrinks all coefficients to zero, and I found that in this model, any penalty value greater than or equal to 0.066 was sufficient to shrink all coefficients to zero. Conversely, a penalty value of 0 does not shrink any coefficients and is equivalent to an OLS regression with all 150 factors. **Table 4** displays the R-squared values, mean absolute errors, and the number of features selected in each LASSO model with different penalty values. As the penalty value increases, the number of coefficients selected decreases, and the R-squared also decreases while the mean absolute error increases.

Table 4.

LASSO Results with Varying Regularization Penalty

Penalty Value	R-Squared (%)	Mean Absolute Error	# Features Selected
0.000	100.0	0.0002	150
0.001	90.6	0.0449	56
0.002	81.9	0.0637	45
0.003	72.0	0.0808	38
0.004	65.3	0.0908	28
0.005	59.9	0.0977	24
0.006	54.6	0.1034	20
0.007	49.7	0.1086	18
0.008	45.0	0.1135	16
0.009	40.6	0.1182	14
0.010	36.4	0.1223	12
0.012	28.6	0.1294	11
0.013	24.5	0.1329	10
0.014	20.5	0.1361	9
0.015	17.0	0.1387	6
0.017	11.1	0.1432	6
0.018	8.0	0.1456	5
0.019	6.3	0.1470	3
0.023	3.6	0.1492	2
0.030	1.7	0.1510	1
0.066	0.0	0.1534	0

My next step was determining which coefficients were selected at each level of penalty strictness. I ran another series of LASSO models with penalties ranging from 0 to 0.065, but I now used increments of 0.0005. The goal of this increased granularity was to ensure only one coefficient was cut after each increment, so I could rank the importance of each factor. I determined the most important factor to be the factor that remained in a 1-factor model. Similarly, the n-th most important factor would be the n-th last factor to be dropped by the LASSO as the penalty became more stringent. The n-th most important factor would also be the n-th factor to be selected as the penalty becomes more lenient. **Table 5** displays the ranking for the 20 most important factors and the highest penalty value at which each factor is included. The most important factors in the LASSO model were the invest (capital expenditures and inventory), salecash (sales to cash), pchsaleinv (% change sales-to-inventory), gma (gross profitability), and UMD (momentum) factors.

Table 5.

Ranking of Factor Importance in LASSO Model

Factor	**Importance Rank**	**Highest Penalty Value to Select**
invest	1	0.0630
salecash	2	0.0290
pchsaleinv	3	0.0225
gma	4	0.0185
UMD	5	0.0180
quick	6	0.0175
dnca	7	0.0145
HML	8	0.0145
std_turn	9	0.0140
pctacc	10	0.0135
nef	11	0.0125
ear	12	0.0110
sp	13	0.0095
roic	14	0.0090
dsti	15	0.0085
SMB	16	0.0080
ol	17	0.0075
orgcap	18	0.0070
aeavol	19	0.0065
QMJ	20	0.0060

Random Forest Regression Model

In this section, I will discuss the factor importance determined from using mean impurity decrease, feature permutation, and Shapley values on my random forest model.

Mean Decrease in Impurity Feature Importance

The UMD (momentum), ear (earnings announcement return), currant (current ratio), dsti (change in short-term investments), salescash (sales to cash), and HML (high minus low) factors were the most important based on mean decrease in impurity. **Table 6** displays the mean decrease in impurity feature importance score for each factor. As **Figure 2** illustrates, relative feature importance scores drop off rapidly. When looking at the feature importance for all 150 factors, the distribution of scores exhibits an exponential decay. When resticing the distribution to the 50 most important factors (as **Figure 3** shows), the drop-off in values is less severe; however, the handful of most important features still seems to dominate. Six factors emerge as the most important by a decent margin. This difference can be seen more easily in **Figure 4**, as the view is restricted to only the 20 most important factors.

Table 6.

Mean Decrease in Impurity Feature Importance

Factor	**Importance**
UMD	0.044
ear	0.038
currat	0.034
dsti	0.030
salecash	0.028
HML	0.026
dy	0.021
nef	0.021
stdacc	0.020
LIQ_PS	0.019
SMB	0.018
stdcf	0.018
maxret	0.018
sue	0.015
QMJ	0.015
grltnoa	0.013
ta	0.013
divo	0.012
age	0.012
sp	0.011
retvol	0.011

Note. All other values are 0.010 or below

Figure 2.

All 150 Factors by Mean Decrease in Impurity Feature Importance

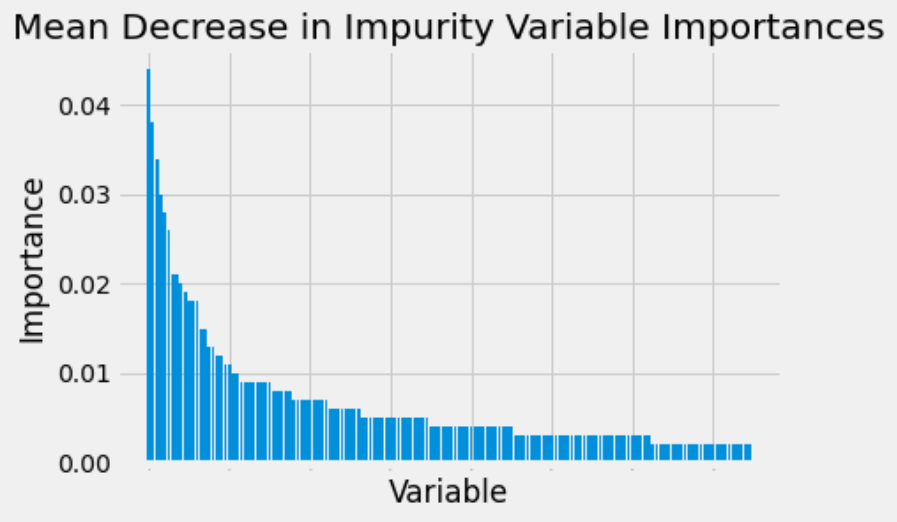


Figure 3.

Top 50 Factors by Mean Decrease in Impurity Feature Importance

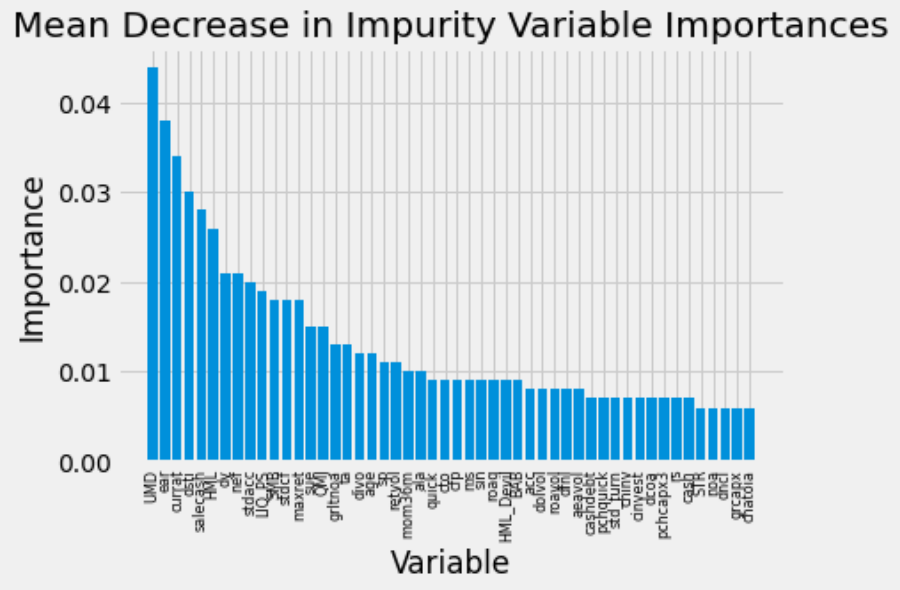
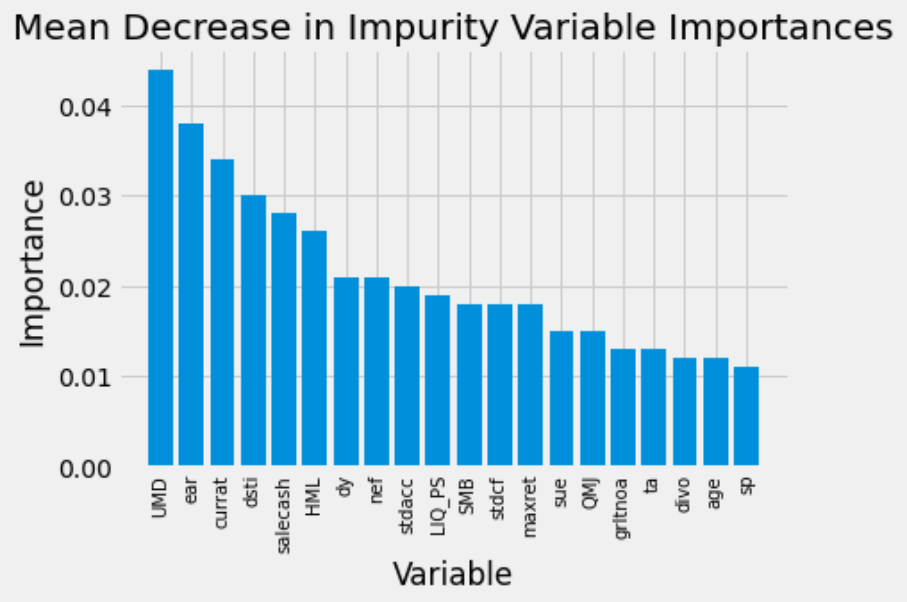


Figure 4.

Top 20 Factors by Mean Decrease in Impurity Feature Importance



Feature Permutation Feature Importance

Feature permutation factor importance can be interpreted as the average amount by which the score of the model (i.e., R-squared) drops when a given factor is randomly shuffled. Similar to mean impurity decrease, the UMD, ear, currant, dsti, and HML factors were the most important factors based on feature permutation-based factor importance too. **Table 7** displays the feature permutation feature importance score for each factor. As **Figure 5** illustrates, relative feature importance scores drop off rapidly, even more so than the mean decrease in impurity importance. When limiting the distribution to the 50 most important factors (as **Figure 6** shows), the drop-off in values more less severe for the first half of features while feature importances begin dropping more slowly for the less important features. **Figure 7** shows that the 5 most important factors domine in relative feature importance.

Table 7.

Feature Permutation Importance

Factor	**Importance**
UMD	0.043
ear	0.039
dsti	0.031
currat	0.029
HML	0.029
dy	0.021
sue	0.018
salecash	0.018
SMB	0.016
ta	0.014
nef	0.014
LIQ_PS	0.013
stdcf	0.013
stdacc	0.013
maxret	0.012
retvol	0.010
QMJ	0.010

Note. All other values are 0.008 or less

Figure 5.

All 150 Factors by Feature Permutation Feature Importance

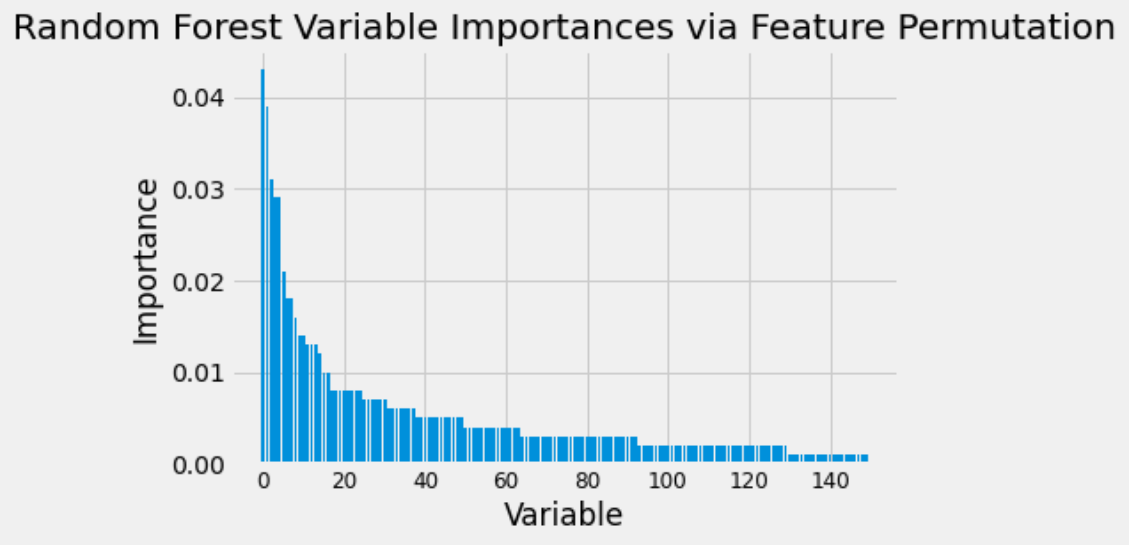


Figure 6.

Top 50 Factors by Feature Permutation Feature Importance

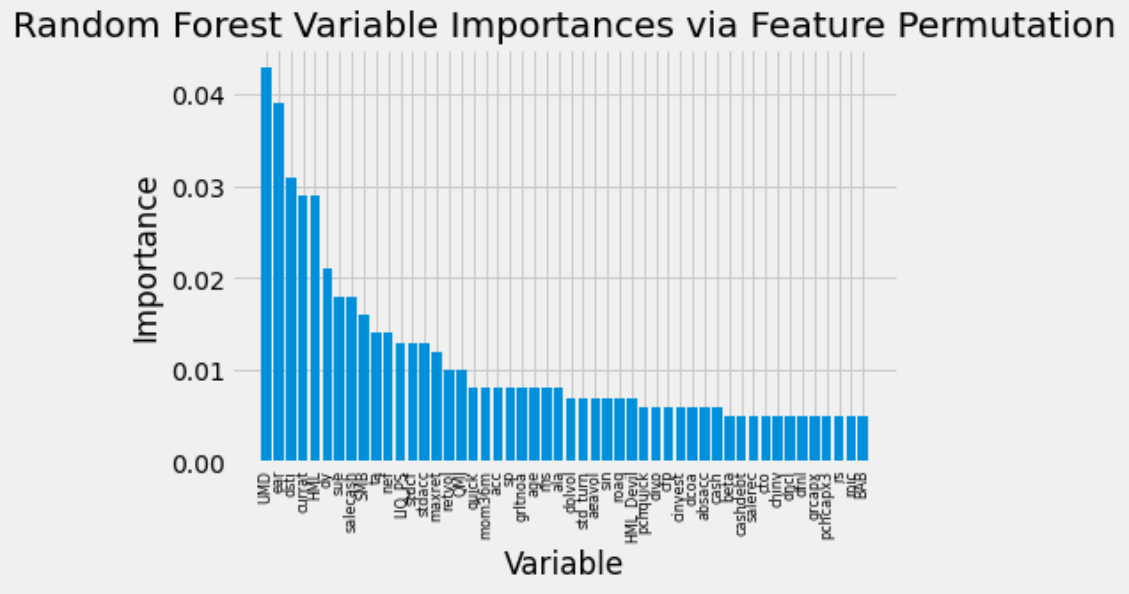
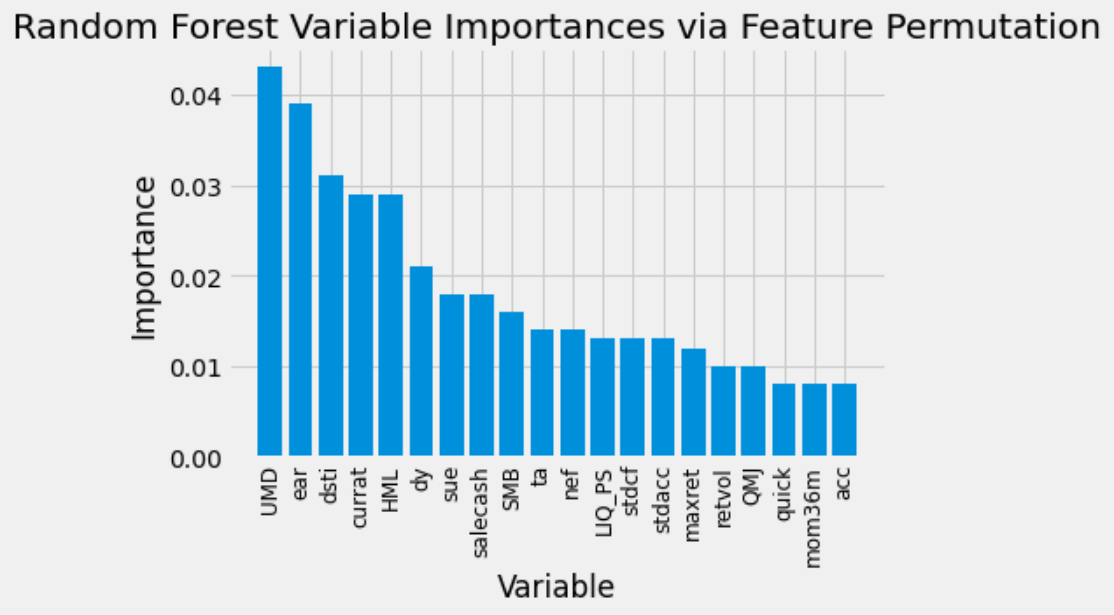


Figure 7.

Top 20 Factors by Feature Permutation Feature Importance



Shapley Value Feature Importance

The ear, UMD, and HML factors are most important based on Shapely values; however, the distribution of Shapley values decreases more slowly than the feature permutation feature importance and mean decrease in impurity feature importances. **Table 8** displays the mean absolute Shapley values for each of the 20 most important factors. Shapley values decrease more smoothly and slowly than the feature permutation and mean impurity decrease feature importance scores (demonstrated in **Figure 8** and **Figure 9**). Based on Shapley values, ear, UMD, and HML are the three most factors by a decent margin (see **Figure 10**).

Looking beyond the distribution of mean absolute Shapley values for each factor, I also looked at each factor's distribution of individual Shapley values among observations. All of the 20 most important factors have a large proportion of observations clustered at similar Shapley values; however, the more important factors have more individual points with much larger

absolute Shapley values (see **Figure 11**). This characteristic suggests that the ranking of mean absolute Shapley values is partly driven by large effects that factors have on a subset of test assets in addition to moderate effects that factors have on all test assets. I then assessed Shapley values even more granularity by looking at individual Shapley values without taking absolute values (see **Figure 12**). This measure allows me to determine which factors have either positive or negative effects on predictions. Interestingly, every factor had many positive values and negative values, suggesting that every factor can either increase or decrease the predicted amount. Some factors, such as ear and dsti, have a cluster of positive Shapley values and only a handful of negative Shapley values. Conversely, some factors, such as HML and dy, have a cluster of negative Shapley values and only a handful of positive Shapley values. Additionally, some factors, including UMD and sue, have both positive and negative Shapley value clusters. This variation of the sign of a feature's Shapley values between observations allows for a deeper understanding of the model's inner workings; however, focusing on the mean absolute Shapley value for each factor allows for a better understanding of the overall importance of a factor based on how much that factor affects predicted returns.

Table 8.

Mean Absolute Shapley Values of Features in Random Forest

Factor	**Mean Absolute Shapley Value**
ear	0.008717
UMD	0.008508
HML	0.007774
dsti	0.006407
sue	0.006281
dy	0.005701
SMB	0.005380
salecash	0.005100
currat	0.004663
stdcf	0.004600
ta	0.004251
nef	0.003727
maxret	0.003234
retvol	0.002994
stdacc	0.002743
LIQ_PS	0.002693
sp	0.002595
QMJ	0.002332
dolvol	0.002323
grltnoa	0.002305

Figure 8.

All 150 Factors by Shapley Value Feature Importance

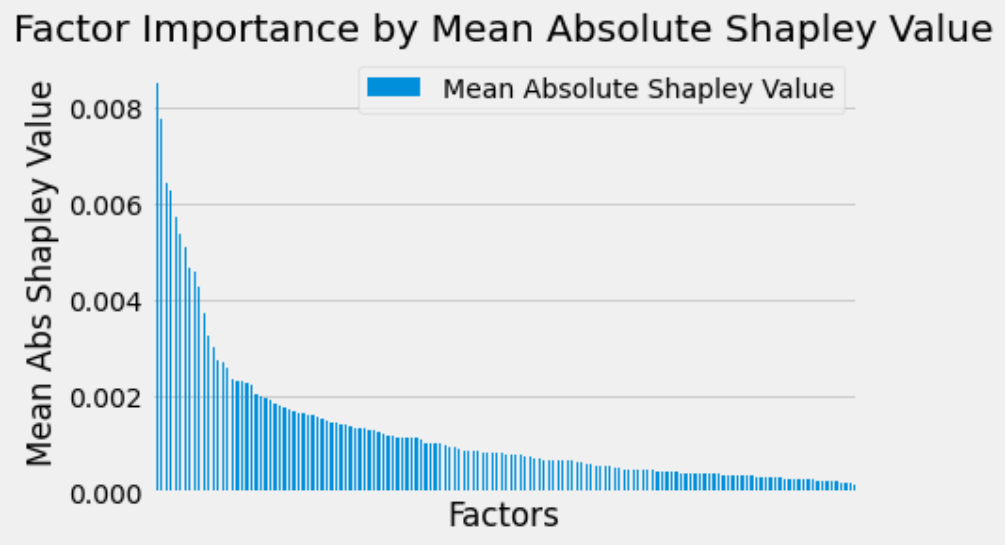


Figure 9.

Top 50 Factors by Shapley Value Feature Importance

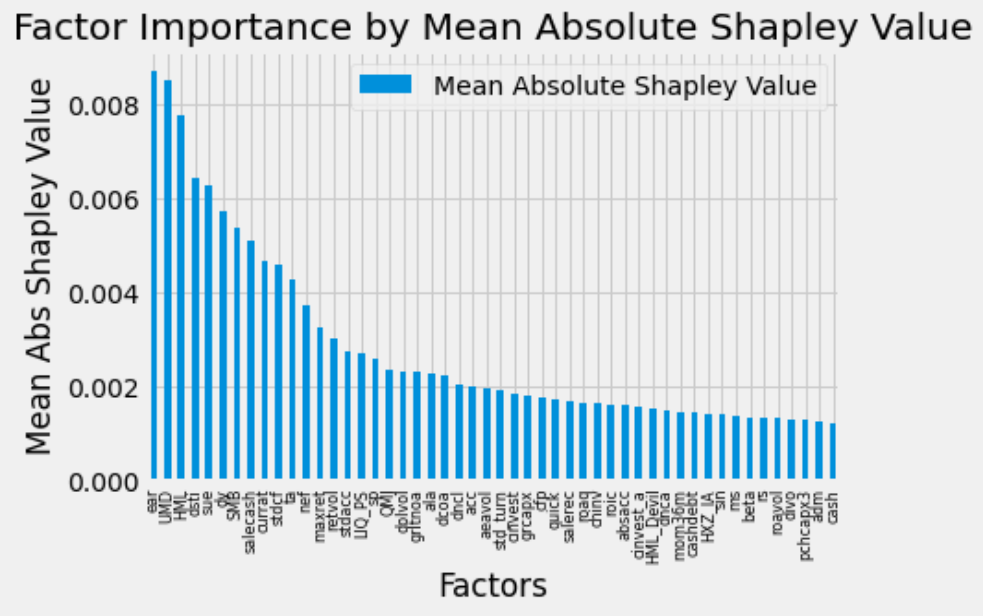


Figure 10.

Top 20 Factors by Shapley Value Feature Importance

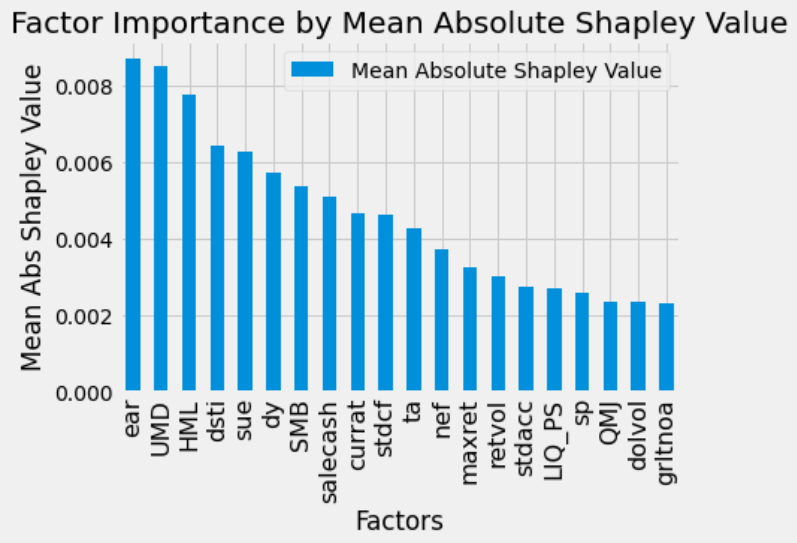


Figure 11.

Individual Absolute Shapley Values for Top 20 Factors by Absolute Mean Value

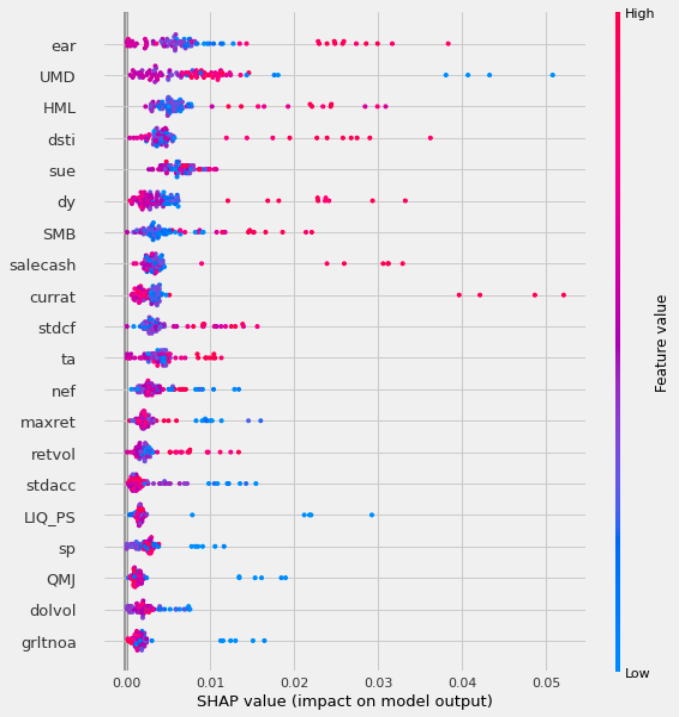
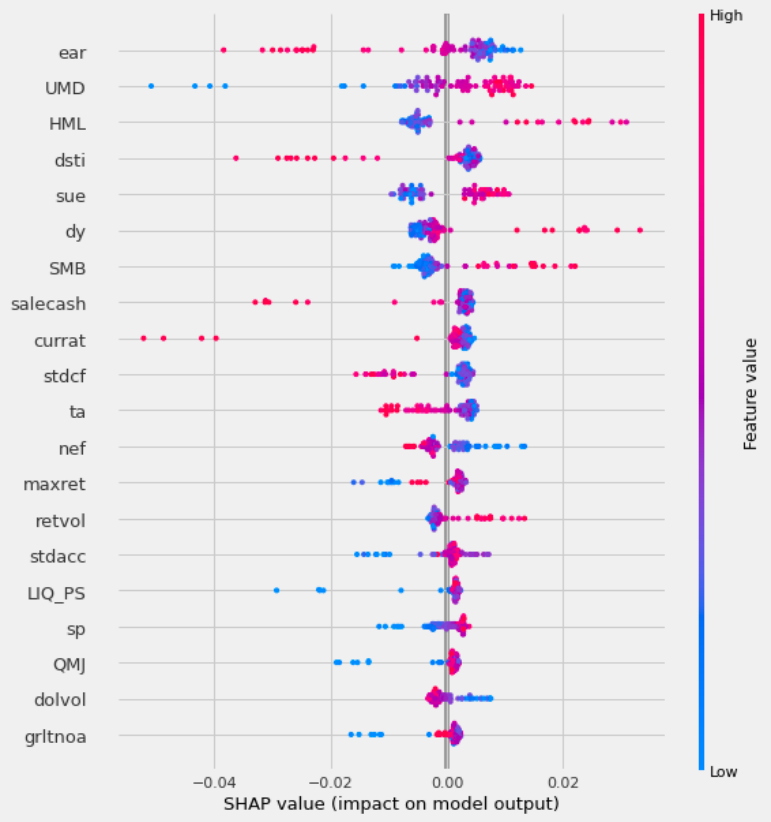


Figure 12.

Individual Shapley Values for Top 20 Factors by Absolute Mean Shapley Value



Comparing Factor Importance Across Models

The OLS, LASSO, and random forest models led to quite different factor importance rankings while the factor importance rankings for random forests were very similar whether calculated through mean decrease in impurity, feature permutation, or Shapley values. In this section, I will first discuss which factors were most important the most often in my models, and I will then discuss the overlap between models of which factors were most important.

Evaluating the Most Important Factors Overall

I evaluated factors based on how often they were in the top 1, 3, 5, 10, or 20 most important factors across the five approaches I used. Of the 150 factors I analyzed, 46 factors were in the top 20 at least once. 29 factors were in top 10 at least once. 14 were in the top 5 at least once. 11 were in the top 3 at least once, and 4 were the most important factor at least once (see **Table 9**). Based on these frequencies, I split the factors into 5 groups. Group 1 included the factors that were highly important in most or all of the models. Group 1 contained 5 factors: UMD, ear, HML, salecash, and dsti. Group 2 was comprised of factors that were highly important in some models but not important in others. This group included 7 factors: roic, orgcap, invest, cinvest, covind, pchsaleinv, and gma. Group 3 consisted of factors that were slightly important in most or all models, and this group included 14 factors: SMB, nef, QMJ, currant, dy, sue, quick, LIQ_PS, ta, stdacc, stdcf, sp, retvol, and maxret. Group 4 was the factors that were slightly important in one or two models but unimportant in the others. This group was comprised of 20 factors: ol, mom36m, grltnoa, dnca, herf, std_turn, pctacc, absacc, ala, poa, pm, dcoa, std_dolvol, divo, exr, aeavol, age, dolvol, zerotrade, and acc. Group 5 was comprised of the other 104 factors that were not among the 20 most important factors in any of my models. This grouping of factors suggests a five-factor model using UMD, ear, HML, salcash, and dsti since those factors were the most important the most often.

Table 9.

Number of Times Factors Were Selected at Various Relative Importance Levels

Factor	**# Top 20**	**# Top 10**	**# Top 5**	**# Top 3**	**# Top 1**
UMD	5	5	4	3	2
ear	5	4	4	3	1
HML	5	4	2	2	0
SMB	5	2	0	0	0
salecash	4	4	2	1	0
dsti	4	3	3	1	0
nef	4	1	0	0	0
QMJ	4	0	0	0	0
currat	3	3	2	1	0
dy	3	3	0	0	0
sue	3	2	1	0	0
quick	3	1	0	0	0
LIQ_PS	3	1	0	0	0
ta	3	1	0	0	0
stdacc	3	1	0	0	0
stdcf	3	1	0	0	0
sp	3	0	0	0	0
retvol	3	0	0	0	0
maxret	3	0	0	0	0
roic	2	1	1	1	0
orgcap	2	1	1	0	0
ol	2	1	0	0	0

mom36m	2	1	0	0	0
grltnoa	2	0	0	0	0
invest	1	1	1	1	1
cinvest	1	1	1	1	1
covind	1	1	1	1	0
pchsaleinv	1	1	1	1	0
gma	1	1	1	0	0
dnca	1	1	0	0	0
herf	1	1	0	0	0
std_turn	1	1	0	0	0
pctacc	1	1	0	0	0
absacc	1	1	0	0	0
ala	1	0	0	0	0
poa	1	0	0	0	0
pm	1	0	0	0	0
dcoa	1	0	0	0	0
std_dolvol	1	0	0	0	0
divo	1	0	0	0	0
exr	1	0	0	0	0
aeavol	1	0	0	0	0
age	1	0	0	0	0
dolvol	1	0	0	0	0
zerotrade	1	0	0	0	0
acc	1	0	0	0	0

Note. Looking at OLS, LASSO, random forest MDI, random forest FP, & random forest Shapley

Similarity of Factor Importance Between Models

The factor importance derived from the mean decrease in impurity, feature permutation, and Shapley values had a large overlap with each other while the OLS and LASSO models had little overlap with each other or the random forest feature importance methods. When looking at the 20 most important factors in each model, 4 factors are shared among all methods: UMD, ear, HML, and SMB. The 3 feature importance measures for random forests all share 9 or 10 of their top 20 factors with 2 other methods, meaning that those methods exclusively share about half of their top 20 factor importance. The OLS and LASSO models had the largest number of unique factors in the top 20 importance, demonstrating that these two models had less overlap (see **Table 10**). In comparing the features among the 5 most important factors for each selection method, less overlap is present. As shown in **Table 11**, no factor is in the top 5 importance values for all 5 selection methods; however, UMD and ear are among the 5 most important factors in 4 of the selection methods. All methods but LASSO include ear in the top 5, and all methods but OLS include UMD in the top 5. UMD was the sixth most important factor in the OLS model, however, so UMD seems to be the most important single factor overall, followed by ear as the second most important factor overall.

Table 10.

Number of Factors in Top 20 Shared with Other Selection Methods

Method	**Unique**	**1 Shared**	**2 Shared**	**3 Shared**	**4 Shared**
OLS	11	4	1	0	4
LASSO	7	3	1	4	4
Mean Decrease in Impurity	2	2	10	4	4
Feature Permutation	1	1	9	4	4
Shapley	1	1	10	4	4

Table 11.

Number of Factors in Top 5 Shared with Other Selection Methods

Method	**Unique**	**1 Shared**	**2 Shared**	**3 Shared**	**4 Shared**
OLS	4	0	0	1	0
LASSO	3	1	0	1	0
Mean Decrease in Impurity	0	2	1	2	0
Feature Permutation	0	2	1	2	0
Shapley	1	1	1	2	0

DISCUSSION

Introduction

Given the proliferation of factors identified by various researchers, this paper uses the LASSO and random forest machine learning models (along with mean decrease in impurity, feature permutation, and Shapley values) to determine which factors best explain cross-sectional returns. My research has three main findings. First, UMD (momentum), ear (earnings announcement return), HML (high minus low), salecash (sales-to-cash), and SMB (small minus big) are the most important factors due to their persistent high importance across models. Second, factor importance varies moderately when using different models. Third, factor importance varies slightly when using different metrics of factor importance for the same model. In this section, I will discuss using machine learning models, comparing feature importance between models, and calculating feature importance through different methods. I will then discuss the limitations of my research and areas for further research.

Implications for Using Machine Learning Models

Machine learning models offer a great approach for dealing with the high-dimensionality of factor data. While my OLS model could not properly handle 150 independent variables, the LASSO and random forest were perfectly able to handle the number of variables and data used. The LASSO model used regularization to penalize the inclusion of too many terms in a linear model while the random forest used an ensemble of independent, nonlinear decision trees to produce a more powerful model. While both models are effective compared to OLS regression in hig-dimensionality, LASSO and random forest offer different benefits.

The current research has varying views on the most effective machine learning model for factor investing, and my results did not prove either LASSO or random forest to be definitivlety better

than the other. Considering I obtained different information about factor importance from each, I believe future research should continue exploring the use of several different machine learning models.

Implications for Comparing Feature Importance Between Models

My random forest, LASSO, and OLS models gave differing results on factor importance, so comparing several models is important for producing more robust results. In my research, some factors were important for all the models while some were only important for one or two of the models. For example, UMD, ear, and HML were very important according to the OLS, LASSO, and random forest models. Other factors, such as currat, dy, stdacc, LIQ_PS, and sue, were only important for the random forest while some factors, like invest, pchsaleinv, gma, dnca, and pctacc, were only important for the LASSO model. Further, the OLS model uniquely returned cinvest, convind, and herf as important factors. In order to be maximally confident in the final model and output, a framework is needed to use several highly accurate models in combination.

Implications for Calculating Feature Importance Using Different Methods

I calculated feature importance scores using mean decrease of Gini impurity, feature permutation, and Shapley values, and each method has advantages and disadvantages.

Mean Decrease of Gini Impurity

Mean decrease of Gini impurity measures the average amount by which each factor improves a decision tree by splitting a node on that factor, but the method has two problems: it is not easily interpretable, and it only works for tree-based models. The values given by mean

decrease of impurity are standardized in order to allow for direct comparison and give relative feature important; however, that standardization makes directly interpreting the value more difficult, as the value is essentially a type of test statistic. Also, since mean decrease of impurity can only be used for tree-based models, the method cannot be used for comparisons to other types of models.

Feature Permutation

Feature permutation measures feature importance based on how much the model suffers without a given feature, and the main advantages are that the technique is generalizable and is more interpretable. Feature permutation involves randomly shuffling the order of one of the features in order to break any link to the target variable. The feature importance is then measured as the drop in model score. Since that random shuffling process for any model, and every model has some type of scoring value, feature permutation works for any model. The interpretation of this value is the marginal contribution of a feature to model performance, so feature permutation is a great method for understanding factor importance in terms of how much they contribute to predictive power.

Shapley Values to Measure Feature Importance

Shapley values also work for any model and are easily interpretable because they involve assigning each individual feature a portion of the difference between a single observation's predicted value and the average predicted value. For example, while feature permutation illustrates that factor "x" contributes a marginal R-squared of 2%, Shapley values say that factor "x" contributes 0.3% worth of the 3% difference between the 8% predicted return of asset "A" and the 5% average predicted return across all assets. In other words, that Shapley value of 0.3,

means that factor "x" explains 10% of the difference between the average predicted value and the predicted value of observation "A." The mean absolute Shapley value is then interpreted as how much a given factor, on average, explains the returns of individual assets relative to the average return of all assets. This interpretation is more insightful into the impact of a factor on the target variable directly rather than on R-squared (or the accuracy of predicting the target variable). Shapley values and feature permutation provide alternate interpretations that give similar, but slightly different, factor importance, so using both seems like a strong approach.

Limitations and Further Research

This paper utilized Fama-MacBeth regressions modified with LASSO and random forest models to perform an in-sample analysis and determine factor importance, and the main limitations arise from this narrow scope. In this section, I will discuss the limitations of my research and potential future research directions centered on the following ideas: performing an out of sample analysis and grouping correlated factors.

Performing Out of Sample Analysis

My research was performed using only in-sample analysis, in order to more closely match the methodology of Fama and MacBeth (1973). In-sample analysis means that I used all available data to train the models. This approach provides more data to feed to model to explain past returns but is less likely to perform as well on new, future data. I would extend this research by holding back some testing data, and incorporating out of sample analysis.

Grouping Correlated Factors

One potential issue in my research is that correlated factors were competing with each other for feature importance scores, so my current ranking of feature importance may be skewed. An approach to improve this issue would be to group correlated factors together before running the model and assigning feature importance scores. In this approach, fewer factors would be given feature importance scores, but those scores would contain the value that was previously attributed to a multitude of correlated factors. This consolidation may help to increase economic interpretability by combining factors that contain similar economic information.

APPENDIX: KEY FOR INCLUDED FACTORS

Row	Description	Mean	tstat	Authors	Year
MktRf	Excess Market Return	0.0064	3.2580	Black and Jensen and Scholes	1972
beta	Market Beta	-0.0008	-0.3468	Fama and Macbeth	1973
ep	Earnings to price	0.0028	1.9134	Basu	1977
dy	Dividend to price	0.0001	0.0384	Litzenberger and Ramaswamy	1979
sue	Unexpected quarterly earnings	0.0012	1.6969	Rendelman and Jones and Latane	1982
pps	Share price	0.0002	0.1393	Miller and Scholes	1982
LTR	Long-Term Reversal	0.0034	2.3402	De Bondt and Thaler	1985
lev	Leverage	0.0021	1.5649	Bhandari	1988
cashdebt	Cash flow to debt	-0.0009	-1.0963	Ou and Penman	1989
currat	Current ratio	0.0006	0.4957	Ou and Penman	1989
pchcurrat	% change in current ratio	0.0000	0.0306	Ou and Penman	1989
pchquick	% change in quick ratio	-0.0004	-0.7661	Ou and Penman	1989
pchsaleinv	% change sales-to-inventory	0.0017	2.9788	Ou and Penman	1989
quick	Quick ratio	-0.0002	-0.1849	Ou and Penman	1989
salecash	Sales to cash	0.0001	0.0994	Ou and Penman	1989
saleinv	Sales to inventory	0.0009	1.0383	Ou and Penman	1989
salerec	Sales to receivables	0.0014	1.4678	Ou and Penman	1989
baspread	Bid-ask spread	-0.0004	-0.2094	Amihud and Mendelson	1989
depr	Depreciation / PP&E	0.0011	0.7772	Holthausen and Larcker	1992
pchdepr	% change in depreciation	0.0008	1.4859	Holthausen and Larcker	1992

SMB	Small Minus Big	0.0021	1.5766	Fama and French	1993
HML	High Minus Low	0.0028	2.2125	Fama and French	1993
STR	1-month momentum	0.0015	1.3948	Jegadeesh and Titman	1993
mom6m	6-month momentum	0.0021	1.7922	Jegadeesh and Titman	1993
mom36m	36-month momentum	0.0009	0.8634	Jegadeesh and Titman	1993
sgr	Sales growth	0.0004	0.3723	Lakonishok and Shleifer and Vishny	1994
cp	Cash flow-to-price	0.0031	2.0921	Lakonishok and Shleifer and Vishny	1994
IPO	New equity issue	0.0010	0.5587	Loughran and Ritter	1995
divi	Dividend initiation	-0.0003	-0.2205	Michaely and Thaler and Womack	1995
divo	Dividend omission	-0.0018	-1.1620	Michaely and Thaler and Womack	1995
acc	Working capital accruals	0.0022	2.9644	Sloan	1996
sp	Sales to price	0.0035	2.6956	Barbee and Mukherji and Raines	1996
cto	Capital turnover	-0.0011	-1.0722	Haugen and Baker	1996
UMD	Momentum	0.0063	3.2345	Carhart	1997
turn	Share turnover	-0.0002	-0.1379	Datar and Naik and Radcliffe	1998
pchgm_pchsale	% change in gross margin - % change in sales	-0.0005	-0.8010	Abarbanell and Bushee	1998
pchsale_pchinvt	% change in sales - % change in inventory	0.0014	2.7127	Abarbanell and Bushee	1998
pchsale_pchrect	% change in sales - % change in A/R	0.0014	2.7997	Abarbanell and Bushee	1998

pchsale_pchxsga	% change in sales - % change in SG&A	0.0009	1.2649	Abarbanell and Bushee	1998
etr	Effective Tax Rate	-0.0004	-0.5832	Abarbanell and Bushee	1998
lfe	Labor Force Efficiency	-0.0003	-0.5485	Abarbanell and Bushee	1998
os	Ohlson's O-score	0.0005	0.6015	Dichev	1998
zs	Altman's Z-score	0.0020	1.4243	Dichev	1998
pchcapx_ia	Industry adjusted % change in capital expenditures	0.0010	1.3219	Abarbanell and Bushee	1998
nincr	Number of earnings increases	0.0001	0.1789	Barth and Elliott and Finn	1999
indmom	Industry momentum	0.0001	0.0921	Moskowitz and Grinblatt	1999
ps	Financial statements score	0.0008	1.1872	Piotroski	2000
bm_ia	Industry-adjusted book to market	0.0022	2.4475	Asness and Porter and Stevens	2000
cfp_ia	Industry-adjusted cash flow to price ratio	0.0026	3.3535	Asness and Porter and Stevens	2000
chempia	Industry-adjusted change in employees	-0.0001	-0.0968	Asness and Porter and Stevens	2000
mve_ia	Industry-adjusted size	0.0036	2.3359	Asness and Porter and Stevens	2000
dolvol	Dollar trading volume	0.0038	2.3052	Chordia and Subrahmanyam and Anshuman	2001
std_dolvol	Volatility of liquidity (dollar trading volume)	0.0020	2.4978	Chordia and Subrahmanyam and Anshuman	2001

std_turn	Volatility of liquidity (share turnover)	0.0002	0.1338	Chordia and Subrahmanyam andAnshuman	2001
adm	Advertising Expense-to-market	-0.0013	-1.0030	Chan and Lakonishok and Sougiannis	2001
rdm	R&D Expense-to-market	0.0034	2.3314	Chan and Lakonishok and Sougiannis	2001
rds	R&D-to-sales	0.0006	0.3544	Chan and Lakonishok and Sougiannis	2001
kz	Kaplan-Zingales Index	0.0022	1.6270	Lamont and Polk and Saa-Requejo	2001
chinv	Change in inventory	0.0018	2.6232	Thomas and Zhang	2002
chtx	Change in tax expense	0.0009	1.1614	Thomas and Zhang	2002
ill	Illiquidity	0.0034	1.8448	Amihud	2002
LIQ_PS	Liquidity	0.0038	2.4843	Pastor and Stambaugh	2003
idiovol	Idiosyncratic return volatility	0.0007	0.3312	Ali and Hwang and Trombley	2003
grltnoa	Growth in long term net operating assets	0.0022	3.3386	Fairfield and Whisenant and Yohn	2003
ob_a	Order backlog	0.0005	0.3703	Rajgopal and Shevlin and Venkatachalam	2003
grltnoa_hxz	Changes in Long-term Net Operating Assets	0.0024	3.6100	Fairfield and Whisenant and Yohn	2003
cfp	Cash flow to price ratio	0.0027	2.0438	Desai and Rajgopal and Venkatachalam	2004
rd	R&D increase	0.0006	0.7156	Eberhart and Maxwell and Siddique	2004
cinvest	Corporate investment	0.0013	2.3477	Titman and Wei and Xie	2004
roavol	Earnings volatility	0.0010	0.6890	Francis and LaFond and Olsson and Schipper	2004

cinvest_a	Abnormal Corporate Investment	0.0013	2.0126	Titman and Wei and Xie	2004
noa	Net Operating Assets	0.0031	4.2926	Hirshleifer and Hou and Teoh and Zhang	2004
dnoa	Changes in Net Operating Assets	0.0014	2.6791	Hirshleifer and Hou and Teoh and Zhang	2004
tb	Tax income to book income	0.0014	1.8241	Lev and Nissim	2004
pricedelay	Price delay	0.0007	1.0807	Hou and Moskowitz	2005
age	years since first Compustat coverage	0.0001	0.0718	Jiang and Lee and Zhang	2005
egr	Growth in common shareholder equity	0.0015	1.7763	Richardson and Sloan and Soliman and Tuna	2005
lgr	Growth in long-term debt	0.0006	0.8541	Richardson and Sloan and Soliman and Tuna	2005
dcoa	Change in Current Operating Assets	0.0019	2.2317	Richardson and Sloan and Soliman and Tuna	2005
dcol	Change in Current Operating Liabilities	0.0003	0.4083	Richardson and Sloan and Soliman and Tuna	2005
dwc	Changes in Net Non-cash Working Capital	0.0011	1.6224	Richardson and Sloan and Soliman and Tuna	2005
dnca	Change in Non-current Operating Assets	0.0021	2.8677	Richardson and Sloan and Soliman and Tuna	2005
dncl	Change in Non-current Operating Liabilities	0.0004	0.6216	Richardson and Sloan and Soliman and Tuna	2005
dnco	Change in Net Non-current Operating Assets	0.0023	2.2782	Richardson and Sloan and Soliman and Tuna	2005

dfin	Change in Net Financial Assets	0.0023	3.8008	Richardson and Sloan and Soliman and Tuna	2005
ta	Total accruals	0.0019	2.8855	Richardson and Sloan and Soliman and Tuna	2005
dsti	Change in Short-term Investments	-0.0003	-0.5371	Richardson and Sloan and Soliman and Tuna	2005
dfnl	Change in Financial Liabilities	0.0018	3.6149	Richardson and Sloan and Soliman and Tuna	2005
egr_hxz	Change in Book Equity	0.0017	1.9309	Richardson and Sloan and Soliman and Tuna	2005
ms	Financial statements score	0.0017	2.3900	Mohanram	2005
chmom	Change in 6-month momentum	0.0021	1.9213	Gettleman and Marks	2006
grcapx	Growth in capital expenditures	0.0014	1.9610	Anderson and Garcia-Feijoo	2006
retvol	Return volatility	-0.0002	-0.1077	Ang and Hodrick and Xing and Zhang	2006
zerotrade	Zero trading days	-0.0005	-0.2816	Liu	2006
pchcapx3	Three-year Investment Growth	0.0011	1.5172	Anderson and Garcia-Feijoo	2006
cei	Composite Equity Issuance	-0.0001	-0.1393	Daniel and Titman	2006
nef	Net equity finance	0.0008	0.6273	Bradshaw and Richardson and Sloan	2006
ndf	Net debt finance	0.0017	3.1129	Bradshaw and Richardson and Sloan	2006
nxf	Net external finance	0.0022	2.4879	Bradshaw and Richardson and Sloan	2006
rs	Revenue Surprises	0.0005	0.5816	Jegadeesh and Livnat	2006
herf	Industry Concentration	0.0003	0.2471	Hou and Robinson	2006

ww	Whited-Wu Index	-0.0002	-0.1700	Whited and Wu	2006
roic	Return on invested capital	0.0018	1.8907	Brown and Rowe	2007
tang	Debt capacity/firm tangibility	0.0005	0.4555	Almeida and Campello	2007
op	Payout yield	0.0016	1.1247	Boudoukh and Michaely and Richardson and Roberts	2007
nop	Net payout yield	0.0016	1.1098	Boudoukh and Michaely and Richardson and Roberts	2007
ndp	Net debt-to-price	0.0002	0.1631	Penman and Richardson and Tuna	2007
ebp	Enterprise book-to-price	0.0014	0.9454	Penman and Richardson and Tuna	2007
chcsho	Change in shares outstanding	0.0024	2.3268	Pontiff and Woodgate	2008
aeavol	Abnormal earnings announcement volume	-0.0008	-1.0963	Lerman and Livnat and Mendenhall	2008
ear	Earnings announcement return	0.0002	0.4409	Kishore and Brandt and Santa-Clara and Venkatachalam	2008
moms12m	Seasonality	0.0016	1.1130	Heston and Sadka	2008
dpia	Changes in PPE and Inventory-to-assets	0.0019	2.7073	Lyandres and Sun and Zhang	2008
pchcapx	Investment Growth	0.0017	2.5468	Xing	2008
cdi	Composite Debt Issuance	0.0008	1.3935	Lyandres and Sun and Zhang	2008
rna	Return on net operating assets	0.0009	0.5534	Soliman	2008
pm	Profit margin	0.0002	0.2861	Soliman	2008

ato	Asset turnover	0.0006	0.4344	Soliman	2008
chatoia	Industry-adjusted change in asset turnover	0.0014	2.6463	Soliman	2008
chpmia	Industry-adjusted change in profit margin	-0.0001	-0.2037	Soliman	2008
cashpr	Cash productivity	0.0027	2.4207	Chandrashekar and Rao	2009
sin	Sin stocks	0.0044	2.6829	Hong and Kacperczyk	2009
rsup	Revenue surprise	0.0012	1.2404	Kama	2009
stdcf	Cash flow volatility	0.0020	1.7155	Huang	2009
absacc	Absolute accruals	-0.0005	-0.5511	Bandyopadhyay and Huang and Wirjanto	2010
invest	Capital expenditures and inventory	0.0019	2.7572	Chen and Zhang	2010
roaq	Return on assets	-0.0009	-0.8961	Balakrishnan and Bartov and Faurel	2010
stdacc	Accrual volatility	0.0019	1.7132	Bandyopadhyay and Huang and Wirjanto	2010
realestate_hxz	Industry-adjusted Real Estate Ratio	0.0011	1.1166	Tuzel	2010
pctacc	Percent accruals	0.0016	2.2563	Hafzalla and Lundholm and Van Winkle	2011
maxret	Maximum daily return	0.0000	-0.0192	Bali and Cakici and Whitelaw	2011
ol	Operating Leverage	0.0020	2.1127	Novy-Marx	2011
ivg	Inventory Growth	0.0013	1.9368	Belo and Lin	2011
poa	Percent Operating Accruals	0.0015	1.8611	Hafzalla and Lundholm and Van Winkle	2011
em	Enterprise multiple	0.0011	1.1348	Loughran and Wellman	2011

cash	Cash holdings	0.0013	0.9838	Palazzo	2012
HML_Devil	HML Devil	0.0023	1.4569	Asness and Frazzini	2013
gma	Gross profitability	0.0015	1.4466	Novy-Marx	2013
orgcap	Organizational Capital	0.0021	2.0536	Eisfeldt and Papanikolaou	2013
BAB	Betting Against Beta	0.0091	5.9774	Frazzini and Pedersen	2014
QMJ	Quality Minus Junk	0.0043	3.8690	Asness and Frazzini and Pedersen	2014
hire	Employee growth rate	0.0008	0.8289	Bazdresch and Belo and Lin	2014
gad	Growth in advertising expense	0.0007	0.8395	Lou	2014
ala	Book Asset Liquidity	0.0009	0.7937	Ortiz-Molina and Phillips	2014
RMW	Robust Minus Weak	0.0034	3.2074	Fama and French	2015
CMA	Conservative Minus Aggressive	0.0026	3.0172	Fama and French	2015
HXZ_IA	HXZ Investment	0.0034	4.1706	Hou and Xue and Zhang	2015
HXZ_ROE	HXZ Profitability	0.0057	4.9901	Hou and Xue and Zhang	2015
Intermediary	Intermediary Risk Factor	0.0015	0.5118	He and Kelly and Manela	2016
convind	Convertible debt indicator	0.0011	1.6983	Valta	2016

Note. A version of this table was originally produced by Feng et al. (2020)

REFERENCES

Banz, R. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3-18. https://doi.org/10.1016/0304-405X(81)90018-0.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *Journal of Finance (Wiley-Blackwell)*, *32*(3), 663–682. https://doi.org/10.1111/j.1540-6261.1977.tb01979.x

Berk, J. & van Binsbergen, J. (2017). How do investors compute the discount rate? They use CAPM. *Financial Analysts Journal, 73*(2), 25-32. https://doi.org/10.2469/faj.v73.n2.6.

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. https://dx.doi.org/10.2139/ssrn.3493458

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

Cattaneo, M., Crump, R., Farrell, M., & Schaumburg, E. (2020). Characteristic-sorted portfolios: Estimation and inference. Review of Economics & Statistics, 102(3), 531-551. https://doi.org/10.1162/rest_a_00883

Chen, L., Pelger, M., & Zhu, J. (2021). Deep learning in asset pricing. http://dx.doi.org/10.2139/ssrn.3350138

De Bondt, W. & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793–805. https://doi.org/10.2307/2327804

Fama, E.F. & French, K.R. (1992). The cross-section of expected stock returns. *The Journal of Finance, 47*(2), 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

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

Fama, E.F. and French, K.R. (1996), Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55-84. https://doi.org/10.1111/j.1540-6261.1996.tb05202.x

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

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607. https://doi.org/10.1086/260061

Feng, G., Giglio, S, & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance, 75*(3), 1327-1370. https://doi.org/10.1111/jofi.12883

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz explainable artificial intelligence. *Expert Systems with Applications*, *167*, N.PAG. https://doi.org/10.1016/j.eswa.2020.114104

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

Harvey, C., Liu, Y., & Zhu, H. (2016). ...and the cross-section of returns. *The Review of Financial Studies*, *29*(1). https://doi.org/10.1093/rfs/hhv059

Jagannathan, R., & McGrattan, E. R. (1995). The CAPM debate. *Quarterly Review (02715287)*, *19*(4), 2.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, *45*(3), 881–898. https://doi.org/10.2307/2328797

Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65-91. https://doi.org/10.2307/2328882

Jensen, T.I., Kelly, B.T., & Pedersen, L.H. (2021). Is there a replication crisis in finance? NYU Stern School of Business (forthcoming). Available at SSRN: http://dx.doi.org/10.2139/ssrn .3774514

Kozak, S., Nagel, S., & Santosh, S. (2018). Shrinking the cross section. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2945663

Messmer, M. (2017). Deep learning and the cross-section of expected returns. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3081555

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Moritz, B. & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2740751

Mussard, S., & Terraza, V. (2008). The shapley decomposition for portfolio risk. *Applied Economics Letters*, *15*(9), 713–715. https://doi.org/10.1080/13504850600748968

Ortmann, K. (2016). The link between the shapley value and the beta factor. *Decisions in economics and finance, 39*(2), 311-325 https://doi.org/10.1007/s10203-016-0178-0

Papenkov, M. (2019). An empirical asset pricing model accommodating the sector-heterogeneity of risk. *Atlantic Economic Journal, 47*, 499–520. https://doi.org/10.1007/s11293-019-09637-2

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

Shalit, H. (2020). Using the shapley value of stocks as systematic risk. *Journal of Risk Finance (Emerald Group Publishing Limited)*, *21*(4), 459–468. https://doi.org/10.1108/JRF-08-2019-0149

Shalit, H. (2021). The shapley value decomposition of optimal portfolios. *Annals of Finance*, *17*(1), 1–25. https://doi.org/10.1007/s10436-020-00380-2

Shapley, L. S. (1952). A value for n-person games. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/papers/P295.html. Also available in print form.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditionals of risk. The *Journal of Finance*, *19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb 02865.x

Stattman, D. (1980). Book values and stock returns. *The Chicago MBA: A Journal of Selected Papers 4*, 25–45.

Tarashev, N., Tsatsaronis, K., & Borio, C. (2016). Risk attribution using the shapley value: Methodology and policy applications. *Review of Finance*, *20*(3), 1189–1213. https://doi.org/10.1093/rof/rfv028

Wang, X., Dunson, D., & Leng, C. (2016). No penalty no tears: Least squares in high-dimensional linear models. *Proceedings of The 33rd International Conference on Machine Learning, 48*, 1814-1822. https://doi.org/10.48550/arXiv.1506.02222

Wolff, D. & Neugebauer, U. (2019). Tree-based machine learning approaches for equity market predictions. *Journal of Asset Management, 20*, 273–288. https://doi.org/10.1057/s41260-019-00125-5