Norwegian School of Economics
Bergen, Spring 2021

ESG: All Bark and No Bite?

Exploring the utility of environmental, social and governance variables in empirical asset pricing via machine learning

Ola Silgjerd

Supervisor: Francisco Santos

Master's thesis, Economics and Business Administration

Major: Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

First I would like to thank my supervisor Francisco Santos for his invaluable guidance throughout the writing process. His knowledge, curiosity and enthusiasm are deeply inspiring, and his teachings will undoubtedly benefit me for many years to come. I would also like to extend my gratitude to my colleagues at Astrup Fearnley for kindly granting me access to some of their computing power. Finally, a sincere thank you to my friends and family for their unconditional support from beginning to end.

Norwegian School of Economics

Bergen, June 2021

Abstract

In this thesis I investigate the impact of including environmental, social and governance (ESG) variables in explaining the cross section of expected stock returns. Using three machine learning frameworks applied to a broad dataset of firm characteristics, macroeconomic predictors and ESG-related variables, I find that ESG contributes to a small but statistically significant increase in explanatory power. The governance category appears to be most important, followed by the environmental category. The social category is not found to contribute significant explanatory power, but does impact predicted excess returns comparably to the other categories. Governance variables contribute to a 4.54% increase in out-of-sample R^2 on average, whilst environmental variables contribute to a 1.44% increase. Including all ESG variables increases explanatory power by around 3.87% on average, but results are highly dependent on model selection, with some models yielding as much as 13.22%. Large firms experience the biggest increase in explanatory power from the inclusion of ESG variables. Finally, I expand on some recent findings in the literature such as the risk premium for CO2 emissions. Using neural network bivariate marginal effects, I find that premiums for younger firms are steeper and more sensitive to CO2 intensity.

Contents

List of Figures

List of Tables

1 Introduction

"The business of business is business" is the famous mantra commonly attributed to Milton Friedman's doctrine, in which he argues that the firm's only social responsibility is to its shareholders. For the same reason that firms should leave diversification and investment decisions to its investors (Modigliani and Miller, 1958), social causes are best addressed by individuals rather than corporations. How, then, is the pricing of firms affected when individuals and investors alike concern themselves more and more with environmental, social and governance issues?

Environmental, social and governance (ESG) aspects, and closely related topics such as sustainability, have garnered a lot of interest from researchers in the past few decades. It is considered a wide encompassing umbrella term that covers a range of interrelated topics. Perhaps the main common aspect among these is the effect that companies and organizations have on the environment, and society at large. Up until the 2000s and 2010s, the term corporate social responsibility (CSR) was primarily used in the context of sustainability, social and governance related matters. As the terms "global warming" and later "climate change" rose into public awareness, along with reigniting social issues, a broader term for these general activities became necessary. Today, the common terms used in finance to refer to such activities mainly include variations of "ESG", "sustainable", "responsible" and "green" investing.

Sustainable investing is defined as an investment approach that incorporates ESG factors in the portfolio construction and asset management process. In the U.S., the number of sustainable funds increased by over 30% from 2019 to 2020, and have seen a four times increase over the past decade (Hale, 2021). Globally there was a reported $30.7 trillion in capital allocated with associations to some form of sustainable investing at the start of 2018 (GSIA, 2018). Depending on how the figure is estimated, the total collective assets under management with any kind of responsible investment mandate might range from $86.3 to $103.4 trillion (UN, 2020). This encompasses a wide range of investing approaches, including exclusionary screening, best-in-class screening, ESG integration and sustainability themed investing, the most common of which being exclusionary or negative screening ($19.8 trillion). This is significant as it implies that a large proportion of global allocated capital is constrained by a selection of measurable variables, which in turn might systematically impact asset prices.

Since the single-factor capital asset pricing model (CAPM) of Sharpe (1964); Lintner (1965); Mossin (1966) and others, and the later widely adopted three factor model of Fama and French (1993), hundreds of factors claiming to contribute in explaining the cross section of expected stock returns have been published. Harvey et al. (2016) review a collection of 316 factors from the literature, and Feng et al. (2017) refer to it as a "zoo of factors". These and several other researchers caution against the mass discovery of risk factors, citing the incentives of journals and researchers to publish positive findings. With increasing data availability and computing power, they argue that the risk of data mining and "*p*-hacking" will inevitably lead to an array of non-replicable discoveries.

This is the motivation for the application of machine learning to the problem of explaining the cross section. As many recent papers, including Chen et al. (2020a) have noted, "It is a natural idea to use machine learning techniques like deep neural networks to deal with the high dimensionality and complex functional dependencies of the problem". Many frameworks and techniques in the field of machine learning have been developed specifically in order to address problems in which the main challenges are high dimensionality, multicollinearity and low signal-to-noise ratio. Therefore, such methods might be useful tools in determining the contribution, or lack thereof, of ESG variables in empirical asset pricing models.

The main research question of this thesis is investigating the impact and predictive power that ESG-related variables contribute in explaining the cross section of expected stock returns. Do non-financial variables within the categories of environmental-, social- and governance-related issues produce a measurable difference in explanatory power when included in asset pricing models? Furthermore, what is the impact of each category and how do these variables affect predicted returns?

To approach this problem, I construct a cross-sectional dataset consisting of firm characteristics and macroeconomic predictors from the literature, as well as ESG-related variables, using a selection of firms from the three major U.S. stock exchanges between 1993 and 2020. The goal is to construct a broad dataset that includes a large proportion of known and available anomalies, such that there is no confounding variable to disrupt measurement of the effect from ESG. Because of the use of regularized machine learning methods, it is less of a concern to include irrelevant information than it otherwise would be when using techniques such as linear regression. The hope is that the models are specified in such a way that the weights and coefficients of irrelevant variables are minimized.

Next, I estimate a selection of three types of machine learning models: elastic net (penalized regression), XGBoost (decision trees) and feedforward neural networks. They are selected to broadly but concisely cover various machine learning approaches of increasing complexity, all with the condition of incorporating regularization. The problem is formulated such that monthly individual excess stock returns are modelled using all available features in the training sample. To evaluate the performance of the models on different feature samples, Monte Carlo simulation is used in order to isolate the impact of ESG categories. Whilst being fitted on the total feature sample, categories of ESG variables are individually removed by imputing randomized values in the test set. For instance, to evaluate the effect of the environmental category of ESG, all non-environmental ESG variables are removed by replacing them with random noise. Because of the stochastic element introduced by randomization, the data generating process is simulated repeatedly to minimize bias in the estimate. The average performance of models for each feature sample is then compared to the average performance of models where all ESG variables have been randomized, controlling for model fixed effects. This allows for significance testing of the difference estimates for each feature sample.

The second part of the research question inquires about how ESG categories and individual variables impact predicted excess returns. However, due to the nature of the applied models, expectations should be moderated with regards to the precision and confidence with which marginal relationships and variable importance can be determined. Many machine learning models are highly complex, with nonlinear interdependencies and high parameterization. The same aspect that improves their explanatory performance unfortunately also obscures their interpretability. With this in mind, several approaches are applied in order to gain insight into modelled interactions and their impact on predicted returns. First, various approaches for determining feature importance are utilized, in order to gain a more detailed and granular perspective on the individual contribution of each variable by model. Second, Shapley values are estimated, which is a game theoretic approach in interpretable machine learning to explain individual predictions. Finally, univariate and bivariate marginal relationships are plotted in attempts to interpret the marginal effects of selected variables on model output values.

The primary empirical finding of this thesis is that small but statistically significant improvements in the explanatory power of asset pricing models are achieved by including ESG-related variables. The increase might be as much as 13%, but results differ across models. The governance category of ESG is found to be the selection of variables that contributes to the largest positive impact on explanatory power. Its inclusion increases out-of-sample R^2 of nonlinear models by an average of 4.54%, and the effect seems to be relatively stable over time. Following governance is the environmental category, which individually contributes to a 1.44% average increase. ESG variables in the social category are generally not found to contribute to a statistically significant positive impact on explanatory power, but do contribute comparably to the other categories in terms of variable importance. Including ESG variables appears to have the greatest impact on explanatory power for bigger firms. This might be due to some selection bias, as ESG reporting might be more extensive at larger companies. The ESG variables with highest importance within the models are executive compensation, employee turnover and CO2 intensity. Finally, bivariate marginal effects are examined using feedforward neural networks and contribute to some recent literature on the subject. One example of this is the risk premium of CO2 emissions documented by Bolton and Kacperczyk (2020). In addition to validating these findings, the risk premium is decomposed by age in order to gain a more detailed perspective. The positive marginal relationship appears to be more pronounced for younger firms, implying that these firms might face tighter constraints and higher average cost of capital due to e.g. exclusionary screening by investors compared to more established firms.

This thesis is largely motivated by the methodology, models and findings of Gu et al. (2018). This comprehensive comparative study applied a wide range of machine learning models to a broad, cross-sectional dataset of firm characteristics and macroeconomic variables. They use this to measure asset risk premia, and unify the empirical asset pricing literature with many widely used machine learning techniques and frameworks. The methodology for processing and modelling a large and high-dimensional dataset of macroeconomic data is inspired by Chen et al. (2020a), although the approach for incorporating macroeconomic information differs.

The main contribution of this thesis is expanding on the relatively recent practice in the empirical asset pricing literature of applying machine learning methods, to include ESG data. Using such methods allows for both more robust and more detailed analysis of the impact that ESG contributes to explaining the cross section of expected stock returns. Furthermore, it is demonstrated that linear model frameworks are unable to effectively incorporate marginal information such as that which might be contributed by ESG variables into asset pricing models. Complex, nonlinear and highly parameterized models however, are generally observed to benefit from marginal information. These findings are validated by using Monte Carlo simulation, estimating multiple models and controlling for model fixed effects, which is a further contribution. Moreover, this is also in accordance with the findings of Gu et al. (2018), who demonstrate large economic gains and increases in predictive accuracy from utilizing nonlinear models, with decision trees and neural networks being among the best performers. The consensus of these results serve to further motivate and justify the decision to utilize different models and methodologies than those conventional in the literature.

The thesis is structured into the following chapters: Chapter 2 provides a review of the literature, which is split into two subsections of ESG and machine learning in the field of empirical finance. Chapter 3 presents the dataset used. Chapter 4 describes methodology used for data preprocessing, estimating models and answering the research question. Chapter 5 presents the empirical results, and Chapter 6 contains discussion around the results and their implications.

2 Literature review

2.1 ESG in finance

Perhaps the main research topic relating to ESG in finance, which has received the most attention from researchers, is the relation between ESG and financial performance. The problem is usually formulated as investigating the relationship between a proxy for ESG performance such as ESG rating, and market or corporate financial performance. It has been a topic of interest since the beginning of the 1970s, and remains highly debated. The most comprehensive systematic review by Friede et al. (2015), with evidence from more than 2000 empirical studies on the subject, finds that around 90% of studies show a non-negative relation between ESG and financial performance, and a large majority find a positive relation.

Third-party ESG scores published by rating agencies have been adopted by academics and practitioners in recent years in order to facilitate decision making.[1] However, these ratings have received criticism for their inconsistency. Daines et al. (2010) find that corporate governance ratings do not provide useful information to investors. Berg et al. (2019) even find abnormal returns from divergence in ESG ratings, and identify three sources of divergence: scope, measurement and weight of categories. A clear limitation of many studies such as those reviewed in Friede et al. (2015) is the narrow focus placed on these unreliable and inconsistent ratings. This thesis seeks to address this issue by including a much broader selection of ESG measures, in order to gain more detailed insight into how these variables impact expected returns.

Hartzmark and Sussman (2019) find that investors value sustainability through higher fund inflows, but high-sustainability funds do not outperform low-sustainability funds. Krüger (2015) finds that investors react strongly negatively to negative events relating to a firm's corporate social responsibility, and this is especially pronounced when the information has strong legal or economic implications. A classic paper by Hong and Kacperczyk (2009) identifies "sin" stocks, which are companies involved in alcohol, tobacco and gambling, and argue that the effects of social norms lead to these firms being neglected

[1] Some examples of adoption include: Engle et al. (2020); Pedersen et al. (2020); Lins et al. (2017) and Dyck et al. (2019).

by analysts and norm-constrained investors. Lins et al. (2017) found that firms with high social capital were much better off than their low social capital counterparts during the 2008-2009 financial crisis. Firms with high CSR intensity had higher stock returns, profitability, growth and revenue, indicating that these firms were somewhat insulated from the broader effects of the recession.

Investors might impact the ESG behavior of firms. Dyck et al. (2019) find that institutional shareholders drive environmental and social performance of firms, and Chen et al. (2020b) use a quasi-natural experiment involving Russell Index reconstitutions to show the same effect for CSR performance. Noh and Oh (2020) use a demand-system approach to estimate a firm-level value of institutional pressure for greenness, and find that this relates to better future environmental performance.

Researchers have also examined certain ESG-related firm-level variables. Pedersen et al. (2020) compute an empirical ESG-efficient frontier using CO2 intensity, "sin" industries and accruals as measures of E, S and G. Similarly, Bolton and Kacperczyk (2020) use data on CO2 emissions to examine risk-adjusted returns and find that firms with higher total emissions earn higher returns, arguing the interpretation that investors demand compensation for their CO2 emission risk exposure. There is a broad literature within corporate governance on board composition affecting firm performance, looking at factors such as board member and executive compensation (e.g., Ryan Jr and Wiggins III (2004) and Chhaochharia and Grinstein (2009)), family involvement (e.g., Anderson and Reeb (2003)), size (e.g., Coles et al. (2008)) and proportion of women (e.g., Adams and Ferreira (2009)). Some of these factors, such as number of female directors, are not thought to have a direct effect on market returns (Post and Byron, 2015), but might affect firm performance in different ways, such as through board meeting attendance or pay-performance incentives (Adams and Ferreira, 2009). ESG variables examined in this thesis are selected based on findings from this literature, with the hope of being able to make inferences based on a more complete foundation that unifies some of this research.

2.2 Machine learning in finance

Machine learning models and techniques are increasingly being used in the financial domain, both in academia and by practitioners. According to Weigand (2019) this is due to lower storage costs, data availability, free open-source software and increasingly available and affordable computing capabilities. Given the increase in data available to researchers, as more and more factors explaining expected returns are published, it is becoming increasingly difficult to examine these datasets using traditional methodology. Feng et al. (2017) and Freyberger et al. (2020) approach this problem by applying variations of the LASSO method. Both papers conclude that many of the factors that are claimed to be predictive of expected returns do not provide incremental information.[2] There is an emerging literature applying machine learning to deal with the problem of dimensionality in asset pricing. Kelly et al. (2019) use dimensionality reduction and extend the technique to allow for time-varying factor loadings. Rapach et al. (2013) apply LASSO for predictor selection and Stambaugh and Yuan (2017) use cluster analysis to identify mispricing factors that explain anomalies better than competing models in the literature. Moritz and Zimmermann (2016) perform portfolio sorts and Bryzgalova et al. (2020) explain cross-sectional return predictors using tree-based methods.

Recently, Gu et al. (2018) applied and compared many different machine learning techniques using a high-dimensional, cross-sectional dataset of firm characteristics and stock returns. They found that tree-based models and feedforward neural networks performed best for the problem of measuring asset risk premia. Chen et al. (2020a) apply an even more complex generative adversarial network (GAN) model, which is a type of deep learning framework in which two neural networks compete with each other, to estimate the stochastic discount factor using a similar dataset. Furthermore, they also apply recurrent neural networks (RNN) with long short-term memory (LSTM) to estimate hidden macroeconomic state variables. They find that this approach outperforms all benchmark models out-of-sample in terms of Sharpe ratio, pricing errors and explained variation. Worth noting about the papers applying machine learning methods is that the variables used are generally based on known anomalies and risk factors at the firm level. For various reasons, few studies incorporate non-financial or alternative data. Moreover, none have yet included a broad selection of ESG-related variables beyond ratings using machine learning models, to my awareness.

[2] This might be interpreted as confirming the concerns of Harvey (2017) regarding the incentives to publish positive results and his prediction of “an embarrassing number of false positives—effects that will not be repeated in the future”.

3 Data

The dataset is largely based on that used in the papers of Gu et al. (2018) and Chen et al. (2020a), due to the similarity of ambition to apply nonlinear machine learning models in empirical asset pricing. The variables are selected based on the feature importance documented in the literature, as well as data availability.[3] The primary aim in constructing the dataset is to include a broad and diverse selection of characteristics, in order to leverage the methods and techniques used as these are known to handle high dimensionality efficiently. This will be discussed further in the next section.

In total, the dataset consists of 65 firm characteristics, of which 32 are ESG-related and the remaining 33 belong to the following categories: intangibles, investment, past returns, profitability, trading frictions and value. Firm characteristics are documented in Table A.1 and ESG-related variables in Table A.2 of Appendix A. Fama and French 12 industry classifications are also included as binary predictors. In addition to firm-level characteristics, a comprehensive dataset consisting of 109 macroeconomic data series is constructed in order to capture systematic risk factors. This set of time series is similar to the one used in Chen et al. (2020a) to extract macroeconomic state processes. An exhaustive list of the variables can be found in Table A.3 of Appendix A.

ESG variables are divided into environmental, social and governance categories. They are selected based on the component variables in each pillar of ESG as defined by the data provider, as well as satisfying requirements of data availability and documentation in the literature. Additionally, there is a category for score variables that are meant as more general measures of ESG, which are constructed by data providers as a weighted average score of certain selected variables. The categorization will later be used for feature sampling, in order to investigate the effect and predictive power of different categories. An overview of the components in each category is provided in Table 3.1.

Many papers studying ESG in a financial context focus primarily on the ESG scores provided by rating agencies, and the findings are often inconsistent and inconclusive.[4]

[3] Examples of papers that document results using these variables include Gu et al. (2018); Chen et al. (2020a); Freyberger et al. (2020) and Feng et al. (2017).

[4] The dissensus among researchers on the topic of ESG and financial performance is documented by Friede et al. (2015). Moreover, Berg et al. (2019) trace divergence in ratings to different sources, including scope, measurement and weights of categories.

Table 3.1: ESG-related measures by category

Environmental	Social	Governance	Scores
CO2 intensity	Female managers	Independent board members	ESG score
Energy intensity	Female employees	Female board members	ESG combined score
Water intensity	Staff turnover	Board meeting attendance	ESG controversies
Waste generated	Working conditions	Board size	Environmental pillar
Resource use	Health and safety	Executive compensation	Social pillar
Emissions	Workforce	Non-executive board members	Governance pillar
Environmental innovation	Human rights	Board member term duration	
	Community	Board member compensation	
	Product responsibility	Management	
		CSR strategy	

Full description of each measure, variable names and source information are provided in Table A.2 of Appendix A.

For this reason, in addition to the application of machine learning models, the selected ESG-related variables have a broad scope and measure a wide range of issues and factors. A roughly equal number of variables are selected from each category of ESG, and are based on the constituents that make up the weighted average scores. This enables far more accurate measurement and attribution of each individual category and variable as it contributes to explaining the cross section. Furthermore, utilizing the components of scores instead of the aggregated ratings allows for the inclusion of variation from each contributing source to the resulting ratings, which might be orthogonal as it relates to expected returns.

The sample period of the dataset spans from January 1993 to December 2020. This is primarily limited due to the lack of historical ESG-related data. It consists of variables measured at a wide array of frequencies, including daily, monthly, quarterly and annually. All variables are aggregated such that the final dataset is measured at a monthly frequency. Monthly returns are known to exhibit the highest degree of normality (Richardson and Smith, 1993), as well as providing a reasonable compromise due to the large differences in measurement frequency.

Data is gathered from multiple different sources, which are indicated in Table A.1 of Appendix A. Market data as well as monthly, quarterly and annual firm characteristics are from the Center for Research in Security Prices (CRSP) monthly and daily stock files and the CRSP/Compustat merged database for quarterly and annual fundamentals, accessed via the Wharton Research Data Services (WRDS) platform. Some financial ratios at the firm level, as well as risk factor loadings are downloaded from WRDS Beta, which combines data from the previously mentioned sources. Fama and French factor data, as well as the Fama and French 12 industry classifications and the 1 month Treasury bill rate used to calculate excess returns are downloaded from Kenneth R. French's data library. Macroeconomic data series are all downloaded from the Federal Reserve Economic Data (FRED) website of the Federal Reserve Bank of St. Louis. Finally, ESG-related data is retrieved from the Refinitiv Datastream platform.

A very important point to note regarding the dataset is the frequencies at which the different variables are measured. Firm-level characteristics are constructed from income statement, balance sheet and cash flow statement data, which are available at quarterly and annual frequencies. In some cases, these data are combined with market data such as price or market capitalization, which are measured at a monthly frequency. Additionally, some variables are constructed from daily market data such as price, volume or bid-ask spread, which are aggregated from daily to monthly frequency. The macroeconomic dataset is entirely measured on a monthly basis, and the ESG-related variables are all annual. This is important to note in order to set general expectations with regards to which variables are going to exhibit the greatest explanatory power of the cross section. On a dataset measured at a monthly frequency, variables constructed from daily and monthly data have a clear advantage due to their variability, whilst variables incorporating financial statement data and ESG-related data stemming from quarterly or annual data are disadvantaged for the same reason.

4 Methodology

In this section, I will describe the methodology used to address the research question. The section begins with a description of the preprocessing steps and methods for the dataset. Next, the models are presented and the methods used to estimate and tune the models are described. Finally, measures used for evaluating model performance are presented, along with different techniques for interpreting and explaining model predictions, variable importance and marginal relationships.

4.1 Preprocessing

4.1.1 Sample selection

The first steps in preprocessing are filters which exclude certain observations from the dataset. Filters are applied at several stages throughout the preprocessing stage. First, following the convention of the literature, only equities listed on the NYSE, AMEX or NASDAQ with share codes of 10 or 11 (indicating common stocks) and listed in USD are included. However, it is further convention to exclude stocks with prices below \$5, micro-cap stocks and financial firms following e.g. Fama and French (1992), but this is not done here, heeding the cautions of Lo and MacKinlay (1990) against data-snooping.[5] Penny stocks (below \$5) and micro-caps (bottom 20% market capitalization of the NYSE sample) might cause problems in asset pricing and are often removed because their pricing might be driven by market microstructure issues. Financial firms are also often excluded, with the reasoning that high leverage might not have the same meaning as for other firms—a high debt ratio might indicate distress for nonfinancial firms but is normal for banks.

Later in the preprocessing stage, firms with less than 24 months of continuous historical pricing data are also excluded, as this is necessary to calculate some of the characteristics. Next, a series of data cleansing filters are used in order to remove unwanted or invalid observations, such as a negative observed value for total assets. Similar filters are applied

[5] This is following Gu et al. (2018), who point out that it might be problematic to use these common filters which exclude certain components of the S&P 500 index from an asset pricing analysis. Furthermore, the authors of this paper apply a similar set of models to their dataset, thereby I find it reasonable to follow a similar procedure.

both before and after feature construction to remove values which are not sensible in a financial context.

4.1.2 Missing values

Next, a crucial step in preprocessing is the handling of missing values. As the final dataset is a result of the merging of datasets measured at a wide range of frequencies, it is expected to generate a large proportion of missing values at different points in time and for different variables. Techniques to handle missing values are applied at almost every stage of preprocessing. Imputation might be required before the construction of variables requiring a constituent that is missing, but it is crucially important that the imputation is performed using an appropriate method and at an appropriate stage of the construction process. Mean, mode and zero imputation are used where each is appropriate, and, of equal importance, imputation is not used where it is not applicable or would otherwise risk introducing bias to the variables.

After merging the annual Compustat dataset containing financial statement information with the CRSP dataset containing market variables at a monthly frequency, the technique "last observation carried forward" (LOCF) is applied to be able to access the latest data point at each monthly period. This method involves simply filling in the last known observation of each feature until the next observation occurs, given that the dataset is sorted by date and the relevant firm-level grouping is applied. This allows us to access the latest available financial statement data at any given point in time in order to construct firm characteristics. This leads us to another important topic in the construction of the dataset, which is avoiding look-ahead bias.

4.1.3 Look-ahead bias

When combining data from multiple different sources measured at various frequencies, it is very important to ensure that all observations used in modelling are publicly available at the time provided. Many firm-level characteristics, especially those originating from the balance sheet or income statement, are not immediately available to the public. With regards to annual and quarterly financial statement variables, I follow the convention of Fama and French (1992) in introducing at least a 4 month lag for quarterly and 6 month lag for annual data. Market data originating from monthly stock files are lagged for 1 month, if constructed using daily market data it is lagged a minimum of 1 month from the following monthly period end.[6]

4.1.4 Transformation

Before modelling there are a number of transformation that can be applied to the dataset that might be beneficial for the analysis. A very widely used preprocessing transformation used to improve both convergence and generalization of machine learning models is normalization (Huang et al., 2020). To normalize the dataset, I use the rescaling approach min-max normalization, which is a simple method to scale the features to a range of two selected values. The transformation is defined as

$$x' = a + \frac{(x - min\,(x))\,(b - a)}{max\,(x) - min\,(x)}, \tag{4.1}$$

where x is the feature to be transformed and $[a, b]$ is the set of min-max values. Worth noting here is that the minimum and maximum values extracted from the features are gathered solely from the training set and applied blindly to the full dataset. This is to avoid introducing bias from the test set. Further discussion of the validation techniques can be found in Section 4.3. The selected min-max values for the transformation are $[-0.5, 0.5]$ and follow Chen et al. (2020a).

Next, categorical variables are transformed using one-hot encoding. This is a method that creates separate dummy variables for each category, which are simple binary variables indicating whether the category is present in the observation or not. One-one encoding is a very common technique in machine learning and is found to increase the performance of complex models (Seger, 2018). Furthermore, logarithmic transformations are applied to some ESG-related variables, as well as in the macroeconomic dataset. Details on the transformation used for each individual macroeconomic variable can be found in Table A.3 of Appendix A. For the variables scaled by revenue measured in millions such as

[6] Daily observations of market variables such as price, holding period return, bid, ask and trading volume are used to construct some characteristics such as Standardized Unexplained Volume (SUV) or bid-ask spread. These variables are constructed using contemporaneous or lagged components for each monthly period, and subsequently lagged an additional 1 month period after the end of the construction period to ensure that the information is publicly available.

carbon intensity (carbonint) or executive compensation (execcomp), the differences in scale between firms across the sample are impractical. As these variables are defined as positive values, a log transformation is applied which is given by $x' = (1 + ln(x))$.

4.1.5 Sample splitting

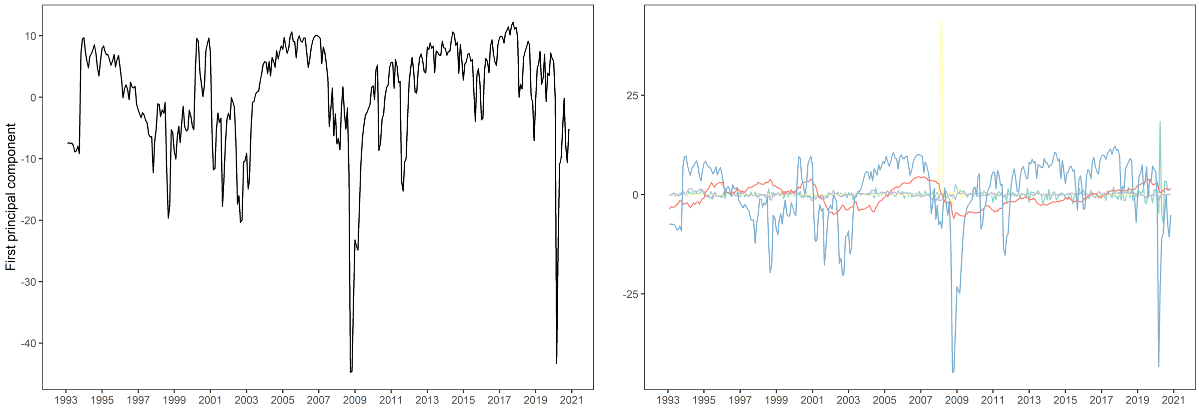
In order to evaluate the performance of the models out-of-sample it is necessary to split the data into a training sample and a testing sample. Due to the chronological nature of the data, the split cannot be randomized with regard to time. Instead, a fixed splitting scheme must be used. The dataset is split sequentially based on number of observations using a ratio of 75%. This corresponds to samples from January 1993 to March 2016 for the training sample and March 2016 to December 2020 for the testing sample. The training sample will be used to fit the models, and is further divided into an 80% training sample and 20% validation sample. This procedure is performed while training in order to estimate out-of-sample prediction error in-sample, as well as for model selection. The testing sample is used to estimate the performance of the models out-of-sample.

4.1.6 Macroeconomic variables

Separate from the firm-level characteristics, I construct a dataset of 109 macroeconomic variables. The selection of variables as well as their respective transformations follow Chen et al. (2020a), which in turn follow McCracken and Ng (2016). The variables are transformed in order to make them stationary, which is useful for dimensionality reduction. Full details on the different types of transformations, and which transformation is used for each variable is provided in Table A.3 of Appendix A. My approach differs from Chen et al. (2020a), whilst they apply a recurrent neural network (RNN) with long short-term memory (LSTM) in order to incorporate lagged values of the predictors in the estimation of the hidden macroeconomic state variables, I instead opt for using hierarchical clustering and subsequently the dimensionality reduction technique principal component analysis (PCA).

First, I use agglomerative or “bottom-up” hierarchical clustering, which is an algorithm that starts by assigning each variable to its own cluster, and merges the nearest clusters together based on gains in cohesion as it moves up the hierarchy. The stability of the

Figure 4.1: Expanding window first principal components from macroeconomic dataset



First principal component of total macroeconomic dataset (left) and first principal component decomposed into each cluster (right).

clusters is evaluated using a bootstrap approach which applies the clustering algorithm to B bootstrap samples of n observations. I select 5 as the number of clusters to be used, as this is the lowest stable number of clusters. The components of each cluster are indicated in Figure B.1 of Appendix B. Next, I apply an expanding window sampling approach to compute principal components for each cluster of variables. Starting with an initial window of observations from January 1985 to January 1993, I calculate the first principal component for each cluster. The final value of each first principal component is then extracted and used as the observation for that month.[7] Then, the window of observations is expanded by one period and the procedure is repeated until the end of the dataset. Time plots of the resulting series are presented in Figure 4.1. By using this method, I am able to perform dimensionality reduction that incorporates the entire dataset of macroeconomic variables, whilst also ensuring that only information available at the time is used.

There are two reasons why dimensionality reduction might be useful in this case, instead of simply passing the entire set of macroeconomic variables to the return models. First, there is a large proportion of time-dependent macroeconomic state information that could be incorporated into the models through the use of lagged values. However, it is simultaneously useful to transform the variables into stationary increments, effectively removing most of the time-dependent information. The way Chen et al. (2020a) solve this problem is by modelling a small number of hidden macroeconomic states using an

[7] The entire resulting dataset is lagged by 1 month at the end of construction to avoid look-ahead bias.

LSTM model, before passing these to a feedforward neural network. Here, a similar outcome is produced through the use of the expanding window approach, which enables me to incorporate all available lagged values for each data point. Second, with such a large dataset there is bound to be multicollinearity and redundant information. Although the models used in this analysis should in principle be able to deal with this through regularization, there is some literature that suggests that it might be beneficial to reduce the number of variables beforehand, especially when dealing with such a large set of predictors.[8]

4.2 Models

Here I present the models used for the analysis. I use a small selection of models motivated by the findings of Gu et al. (2018), with the goal of broadly but concisely covering the most useful modelling approaches.

First, a linear regression technique must be included due to its popularity in finance. LASSO and ridge regression are popular penalized regression methods, and elastic net combines both types of penalties. Using repeated cross-validation allows me to let the data dictate which method is most effective. All the selected models incorporate regularization in some way, which is necessary due to the nature of the problem, as well as the dataset. Next is the gradient boosted decision tree framework XGBoost. Tree boosting techniques have been noted as particularly effective for financial applications by many practitioners, and was found by Gu et al. (2018) to be one of the best performing models. Using deep ensembles of decision trees allows for the modelling of highly complex nonlinearities. It might also help uncover how different variables, especially ESG-related variables interact with financial firm characteristics within a return model. Finally, artificial neural networks are employed as the most complex model. Feedforward neural networks are hypothesized to be "universal approximators" and might be helpful in uncovering complex nonlinear interdependencies. They take a vastly different approach to modelling nonlinearities compared to gradient boosted regression trees, and might therefore be valuable in providing a broader perspective and deeper insight into variable interactions.

[8] Chen et al. (2020a) found that using such a large number of predictor variables as in the approach of passing the entire set to the return models negatively impacted their performance.

4.2.1 Elastic net

The first model used is the penalized linear regression technique known as elastic net, first proposed by Zou and Hastie (2005). It incorporates a linear combination of L_1 (least absolute shrinkage and selection operator or LASSO) and L_2 (ridge) regularization, and is known to overcome some of the limitations of the LASSO method, such as variable selection when the number of predictors p is much larger than the number of observations n. Ordinary least squares (OLS) linear regression estimation is given by

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 , \tag{4.2}$$

where y_i is the observed value and $\hat{y}_i$ is the estimated value. The L_1 and L_2 regularization penalty terms are given by

$$L_1 = \lambda \sum_{j=1}^{p} \left| \beta_j \right| , \quad L_2 = \lambda \sum_{j=1}^{p} \beta_j^2 , \tag{4.3}$$

where λ is a weight parameter that adjusts the magnitude of the penalty. In the case of $\lambda = 0$, an OLS regression is returned. The elastic net penalty term combines both L_1 and L_2 regularization, and is given by

$$\lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{p} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{p} \left| \hat{\beta}_j \right| \right) , \tag{4.4}$$

where the additional α parameter allows for adjustment of the linear combination of the L_1 and L_2 penalty terms. In a case where $\alpha = 0$, the penalty term is equivalent to ridge regression, and likewise equivalent to LASSO where $\alpha = 1$. The estimation of elastic net regression can then be stated as

$$\arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{p} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{p} \left| \hat{\beta}_j \right| \right) . \tag{4.5}$$

4.2.2 XGBoost

The next model is the very widely used tree boosting model XGBoost, which started as a research project by Tianqi Chen and is described in Chen and Guestrin (2016). It is well known for producing winning results in many machine learning competitions, as well as its scalability. The system is a gradient boosted tree (GBT) algorithm and is built on a gradient boosting framework, which is an ensemble method that uses multiple decision trees together to generate predictions. It builds decision trees sequentially, such that trees are fitted on the residuals of previous ones. This way, even though each tree is a relatively weak learner with high bias, the resulting ensemble model can become a strong learner. Here, the method is applied to a regression problem, making it a type of gradient boosted regression tree (GBRT) model, which is the same type as is applied by Gu et al. (2018).

For a dataset of n observations and m variables, XGBoost uses K additive functions to predict the target variable, and the model is given by

$$\hat{y}_i = \phi\left(\mathbf{x}_i\right) = \sum_{k=1}^{K} f_k\left(\mathbf{x}_i\right), \qquad f_k \in \mathcal{F}, \tag{4.6}$$

where $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the space of all possible regression trees, T is the number of leaves in the tree and each f_k corresponds to a tree structure q and leaf weights w (Chen and Guestrin, 2016). To train the model, the regularized objective function

$$\mathcal{L}\left(\phi\right) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right), \quad where \quad \Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{4.7}$$

is minimized. l is a convex loss function measuring the residuals of the predicted values $\hat{y}_i$ and the actual target values y_i, whilst Ω is the regularization term.

The XGBoost system provides a range of hyperparameters that must be given as input in order to train the model. The performance is highly dependent on the selected hyperparameters, and they must be tuned individually for each dataset. Ranges of commonly recommended values were specified for each hyperparameter, and the optimization was performed using random search, which has been shown to be more efficient than grid search and manual search (Bergstra and Bengio, 2012). The ranges used for the hyperparameter optimization, selected values and brief descriptions of the tuned hyperparameters are provided in Table 4.1.

Table 4.1: XGBoost hyperparameters

Hyperparameter	Range	Selected	Function	Description
eta	{0.001, 0.1}	0.01	Learning rate	Step size of optimization for each iteration.
gamma	{0, 1.0}	0	Control overfitting	Minimum loss reduction required to make node split.
max_depth	{2, 25}	3	Control overfitting	Maximum depth of each tree.
min_child_weight	{1, 15}	5	Control overfitting	Minimum sum of instance weight required in child node.
subsample	{0.5, 1.0}	0.8	Add randomness	Fraction of observations to subsample for each tree.
colsample_bytree	{0.5, 1.0}	0.5	Add randomness	Fraction of features to subsample for each tree.
alpha	{0, 1.0}	0.1	Regularization	L_1 regularization.
lambda	{0.01, 1.0}	0.01	Regularization	L_2 regularization.
nrounds	{100, 2000}	800	Complexity	Number of trees.

4.2.3 Neural network

The final model used is an artificial neural network model, more specifically a deep feedforward neural network. They are a classic type of neural network, and are widely applied in many fields, both in academia and in practice. The objective of a feedforward neural network is to approximate a function f^*. For a regression model $y = f^*(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector of predictors and y is the output variable, a feedforward neural network defines a mapping as $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$ where the parameters $\boldsymbol{\theta}$ are optimized such that the resulting function is the best approximation (Goodfellow et al., 2016).

Neural networks consist of *units* which are loosely based on neurons found in biological brains. Units are connected to each other such that information passes through, and are typically grouped together in layers. The parameters $\boldsymbol{\theta}$ include weights $\boldsymbol{W}$ which specify a scaling factor for each connection between units. In feedforward neural networks, the flow of information is unidirectional. Deep feedforward neural networks consist of one or more *hidden layers*, which indicates that these are used for intermediary computations. Increasing the number of layers is known to substitute a large increase in number of units for equivalent performance at a lower computational cost due to the increase in complexity. The layers of units are fully connected and arranged in a chain structure, such that each layer is a function of the preceding layer. The first layer is defined as

$$\boldsymbol{h}^{(1)} = g^{(1)}\left(\boldsymbol{W}^{(1)\top}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right). \tag{4.8}$$

The second layer is a function of the first layer, and is defined as

$$\boldsymbol{h}^{(2)} = g^{(2)} \left(\boldsymbol{W}^{(2)\top} \boldsymbol{h}^{(1)} + \boldsymbol{b}^{(2)} \right), \tag{4.9}$$

and so on, where g is an activation function, $\boldsymbol{W}$ is a vector of weights, $\boldsymbol{x}$ is a vector of input values and $\boldsymbol{b}$ is a vector of biases (Goodfellow et al., 2016).

Decisions regarding the number of units and layers to provide the network with relate to the *architecture* of the network. They are all considered hyperparameters that need to be tuned in order to achieve optimal performance, as neural networks are both the most complex and highly parameterized of the models used. Exhaustive search optimization of the parameters $\boldsymbol{\theta}$ is not computationally feasible for this reason. Stochastic gradient descent (SGD) is a commonly used method for optimizing neural networks, which is an approximation approach to deal with the computational intensity of the problem. More specifically, the adaptive moment estimation (Adam) optimization algorithm of Kingma and Ba (2014) is used. It can be regarded as combining the RMSProp algorithm and the momentum method, and has stood out in the literature for generalizing well to a wide range of problems.

In selecting the architecture, I follow general recommendations from the literature, as well as Gu et al. (2018), and select an initial number of units as a power of two (2^n). Hidden layer units follow the geometric pyramid rule, wherein each subsequent hidden layer is given half the number of units as the previous (e.g. 32, 16, 8). Furthermore, in selecting the activation function g, I find that the rectified linear unit (ReLU) activation function is by far the most commonly used and extensively tested in the literature. It has been shown to be highly effective at training deep neural networks on complex, high-dimensional datasets, and is given by

$$g(x) = max(0, x). \tag{4.10}$$

Because of the complexity, nonlinearity and parameterization of deep neural network models, the risk of the resulting model overfitting the training sample is large, and it is therefore common to apply many different methods of regularization to avoid this. L_1

regularization is used, which has been discussed previously. However, L_2 regularization is omitted as it might cause "weight decay" which has been shown to be equivalent to early stopping (Bishop et al., 1995), which is used instead.

Early stopping monitors the loss on the validation set, and halts training when validation loss ceases to improve for a specified number of training iterations. It is often the case in training neural networks that, because of the high number of parameters, accuracy on the training sample continues to improve with more training iterations whilst validation accuracy reaches a peak. This is due to overfitting of the training sample, and a simple yet powerful remedy is using early stopping and restoring the weights at the iteration of highest validation accuracy.

Next, dropout is applied at each layer, which is a very commonly used regularization technique first proposed by Srivastava et al. (2014). A certain proportion of units in each layer is omitted from the training process, given by a specified dropout rate hyperparameter. This helps in diluting the weights as the network is unable to rely on certain units and complex co-adaptions in the training sample, which in turn helps the model generalize better to new data.

Finally, batch normalization of Ioffe and Szegedy (2015) is applied at each layer. It is known to stabilize the performance of neural networks, as well as make training more efficient, by normalizing the inputs of each layer. Batch normalization is applied after the nonlinearity of the layer (i.e. the ReLU activation function) and before the dropout, following what is recommended by the authors. Ranges and final selected hyperparameter values for the neural network model are provided in Table 4.2.

Table 4.2: Neural network hyperparameters

Hyperparameter	**Range**	**Selected**	**Function**	**Description**
Number of units	{8, 2048}	512	Complexity	Number of hidden layer units.
Number of hidden layers	{1, 4}	3	Complexity, depth	Depth of the network.
Learning rate	$\{10^{-6}, 10^{-2}\}$	10^{-3}	Learning rate	Step size of optimization for each iteration.
Epochs	{10, 500}	100	Optimization	Number of training epochs.
Dropout	{0, 0.6}	0.6	Regularization	Ratio of input values to drop for each layer.
Batch normalization	{Y/N}	Y	Regularization, stability	Normalization of input values for each layer.
L_1 regularization	$\{10^{-5}, 10^{-3}\}$	10^{-4}	Regularization	L_1 regularization.
Patience	{2, 50}	10	Regularization	Number of iterations for early stopping.

4.3 Estimation

The three types of models used take very different approaches in order to produce predictions and are different in many ways, but some general principles and methods can be applied. All models are trained and validated on the training set, and hyperparameters are selected entirely based on the predictive performance within the training set, in a manner which is entirely indifferent to the contents of, and the predictive performance on the test set. The objective when fitting models and selecting hyperparameters is minimizing the mean squared error (MSE) of predictions, which is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4.11}$$

The objective function is modified by introducing regularization, such as in the form of penalty terms for the estimated parameters.

Another common aspect for the training of all models is the use of a random search algorithm for hyperparameter optimization. This entails first specifying a range of values for each hyperparameter to tune, and then simply sampling values randomly from these ranges. This method has been shown by Bergstra and Bengio (2012) to be significantly more efficient than the popular alternative grid search, which would require the exhaustive search of every possible combination. With random search, larger ranges with higher granularity can be used in the search for optimal hyperparameters at a lower computational cost.

Each model selected for this analysis incorporates regularization. This is very important due to the nature of the dataset used as well as the nature of the problem. The large number of variables and low signal-to-noise ratio would be very disadvantageous to a model specification lacking any form of regularization of the coefficients, and would furthermore lead to more overfitting of the training data for more complex models. Elastic net uses a linear combination of L_1 and L_2 regularization, XGBoost and the neural network also have the potential for combining both. The latter two models also implement early stopping as an additional form of regularization, which is one of the most commonly used types of regularization in deep learning, due to its effectiveness and simplicity (Goodfellow et al., 2016). With early stopping, the validation loss of the model is monitored for each learning iteration. If the loss stops improving for a prespecified number of iterations, the training is stopped and the parameters with which validation loss is minimized are restored. This simultaneously saves computing power and helps prevent overfitting.

To measure the effect of different ESG variables, feature sampling is applied by fitting models on a complete dataset in the training sample, and subsequently removing selected features by imputing randomized values in the testing sample. The process is repeated for multiple simulations in order to reduce any idiosyncratic bias that might arise for individual models. Alternatively, selected variables could be imputed with zero values, as is done when estimating variable importance, but this is unlikely to produce comparably robust results due to the complexity and nonlinear nature of the models. Furthermore, due to the stochastic elements in initialization and estimation of machine learning models, there is high variability in the performance of identically parameterized models. Because of this, it is necessary to perform repeated simulations of randomized generated values in addition to estimating multiple models, in order to isolate the effect of certain predictors with reasonable confidence. Using this method and controlling for model fixed effects, it is possible to achieve a much more robust estimate of the impact from each feature sample.

4.3.1 Model evaluation

To evaluate the performance of models on different feature samples, out-of-sample R^2 is estimated as

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\tau_3}^{T} \left(y_{i,t+1} - \hat{y}_{i,t+1}\right)^2}{\sum_{(i,t)\in\tau_3}^{T} \left(y_{i,t+1} - \bar{y}\right)^2}, \tag{4.12}$$

where τ_3 indicates the testing sample, which is entirely independent from training sample, consist only of observations measured at a later point in time and not used for model estimation or selection. The use of R^2_{OOS} follows Gu et al. (2018) and Chen et al. (2020a).

To compare the predictive accuracy out-of-sample for each model and feature sample against each other, the Diebold-Mariano test of Diebold and Mariano (2002) is used. This allows for significance testing of forecast accuracy, with a null hypothesis of no difference in accuracy between two competing forecasts. Following Gu et al. (2018), the Diebold-Mariano test is adapted by comparing prediction errors calculated from cross-sectional average excess returns instead of individual predicted values. The test statistic is defined as $DM = \bar{d}/\hat{\sigma}_{\bar{d}}$, where

$$d_{t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_3} \left(\left(\hat{e}_{i,t+1}^{(1)} \right)^2 - \left(\hat{e}_{i,t+1}^{(2)} \right)^2 \right), \tag{4.13}$$

$\hat{e}_{i,t+1}^{(1)}$ and $\hat{e}_{i,t+1}^{(2)}$ are prediction errors for excess return i at time t for competing forecasts (1) and (2), and $n_{3,t+1}$ is the number of observations in each period of the test set.

4.3.2 Variable importance

As part of the research question asks not only if, but *how* individual and categories of ESG variables impact the return models, methods for measuring their effect on predicted values are needed. Using complex nonlinear machine learning models has both advantages and disadvantages. Amongst the disadvantages, it is often pointed out that many machine learning models lose interpretability and explainability in their complexity. This might be at least part of the reason why the academic literature in empirical finance has been so reluctant to adopt many of these useful techniques, in favor of simpler and more intuitive models.

A lot of effort has been put into explaining the complex behavior of machine learning models, and Molnar et al. (2020) provide a summary of the history and state-of-the-art in the field of interpretable machine learning. Here, I am primarily interested in analyzing and quantifying the effect certain variables have on model accuracy, as well as the general directional impact of certain variable categories. I employ several different approaches to estimate feature importance, univariate and bivariate marginal effects, which are detailed below.

Elastic net variable importance is estimated as the absolute value of the t-statistic. The elastic net model is trained using randomized, repeated cross-validation on the training set, and the t-statistics for each resulting model parameter is used.

XGBoost variable importance uses a gain value, which is the improvement in accuracy from each feature in the model based on the total gain of the feature's splits. For each split in a decision tree for a given variable, the difference in accuracy can be measured as a way to quantify the amount of improvement the variable contributes to the model.

An individual decision tree improves a certain amount from adding a split using a given variable, and the sum of total improvements over all decision trees in the ensemble is the gain of that variable, which is used as its measure of feature importance.

Neural network variable importance is difficult to estimate due to the complexity, non-linear relationships and parameterization. There is no definitive approach and any method will only be a rough approximation of the actual importance of a given variable to a feedforward neural network model. Following Gu et al. (2018) and in accordance with recent literature in interpretable machine learning, I apply a simple method of individually imputing each variable with zero whilst keeping the remaining testing sample fixed and measuring the reduction in R^2.

4.3.3 Marginal effects

Beyond the measure of variable importance, it is also of interest to interpret the behavior of different models through marginal effects and relationships. However, it is important to moderate our expectations of meaningful and significant insight from analyzing univariate or multivariate marginal effects. Some machine learning models are considered by many to be highly opaque and difficult to interpret due to their complexity. Whilst it is true that large numbers of estimated parameters and nonlinear relationships make some models' predictions difficult to explain, it might still be possible to make some useful inferences using univariate or multivariate imputation.

Univariate and bivariate marginal effects on expected returns are estimated by allowing a primary active variable to vary by imputing the full normalized range $[-0.5, 0.5]$ for the given variable whilst keeping the remaining variables fixed at their respective mean values over a constructed sample. For bivariate effects the same method is applied repeatedly for the active variable while a secondary passive variable is held constant at a fixed level over the range and varies stepwise for each iteration.

SHAP (SHapley Additive exPlanations) is a method proposed by Lundberg and Lee (2017) to explain the individual contribution of each variable to model predictions. It is a method for estimating Shapley values for predictive models, which were originally applied as a solution concept in cooperative game theory. They are used to attribute the proportional contribution of each predictor to the predicted value. A feature value z_i

has a Shapley value $\phi_i \in \mathbb{R}$ which is its weighted contribution to the output value. The explanation model g is defined as

$$g\left(z'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{4.14}$$

where $z' \in \{0,1\}^M$ is the vector of simplified input features. More details regarding how SHAP is used to estimate Shapley values can be found in Lundberg and Lee (2017).

5 Results

This section presents the empirical results. First, results relating to the main research question of investigating the usefulness of ESG variables in explaining the cross section of expected returns are presented in Table 5.1. Here, out-of-sample R^2 is compared between feature samples of no ESG, each ESG category individually, and all ESG using two different approaches.

In order to compare the usefulness of certain variables, Monte Carlo simulation of imputed randomized values in the testing sample is applied. The first approach (individual simulations) estimates out-of-sample predictive accuracy for each model and for each simulation. This provides a large number of estimates for the R^2_{OOS} in each feature sample, which in turn allows for the comparison and significance testing of different categories. In the second approach (aggregate simulations), each set of predicted excess return values is aggregated for each model and feature sample. This generally improves predictive accuracy in all models, which is likely due to the diversifying effect of aggregation, which minimizes idiosyncratic predictive error. The table should be interpreted as examining the effect on explanatory power resulting from including each category of ESG. For comparison and to put the results into perspective, estimates from removing more established risk factors such as size, value and momentum are also included.

Before interpreting the results it should be noted that the impact of ESG is very small compared to the established risk factors. This is likely both due to underlying associated risk, as well as the measurement frequency of the variables. Furthermore, the effects measured here are as precise of a measurement as was possible given the frequency at which ESG data are measured. Comparing variables measured at monthly, quarterly and annual frequencies is disadvantageous to the low-variance variables, and might understate their impact.

The primary findings from Table 5.1 are the statistically significant positive impacts of the environmental and governance categories, and the negative impact of the social category. The negative impact finding in particular validates the randomized simulation approach as an effective tool to identify not only positive but also negative contributors to return models. The governance category provides the largest positive contribution, with a 0.29%

Table 5.1: Percentage out-of-sample R^2 by model for different feature samples

	Individual simulations				Aggregate simulations			
	EN	XGB	NN	All	EN	XGB	NN	Mean
No ESG	0.033	0.927	0.595	0.648	0.042	1.007	0.637	0.562
Environmental	0.033	0.927	0.643	0.650	0.038	1.016	0.624	0.560
Δ	0.000	0.000	0.049	0.001	-0.004	0.009	-0.013	-0.003
t	6.199	1.959	9.834	9.450				
Social	0.033	0.924	0.583	0.646	0.037	1.005	0.748	0.597
Δ	0.000	-0.003	-0.012	-0.002	-0.005	-0.003	0.110	0.034
t	-15.659	-23.311	-2.553	-17.117				
Governance	0.033	0.931	0.621	0.650	0.029	1.016	0.744	0.596
Δ	0.000	0.004	0.026	0.002	-0.013	0.009	0.106	0.034
t	-39.373	30.041	7.092	17.135				
All ESG	0.033	0.927	0.626	0.648	0.022	1.015	0.722	0.587
Δ	0.000	0.000	0.032	0.000	-0.020	0.008	0.084	0.024
t	-44.136	2.698	5.170	-0.123				
Excluding size	0.009	0.313	0.452	0.202	0.015	0.409	0.518	0.314
Δ	-0.024	-0.614	-0.174	-0.446	-0.007	-0.606	-0.204	-0.272
t	-77.654	-573.160	-33.367	-230.260				
Excluding value	0.002	0.472	0.293	0.287	0.004	0.521	0.379	0.302
Δ	-0.031	-0.455	-0.334	-0.362	-0.018	-0.494	-0.342	-0.285
t	-81.836	-271.640	-24.962	-206.370				
Excluding momentum	0.026	0.161	0.195	0.124	0.016	0.220	0.214	0.150
Δ	-0.007	-0.766	-0.432	-0.524	-0.006	-0.795	-0.508	-0.437
t	-66.928	-743.698	-39.712	-174.770				
Number of models	100	100	10	210	1	1	1	3
Number of simulations	100	100	100	300	1 000	1 000	1 000	3000
Number of observations	80 000	80 000	8 000	168 000	8	8	8	24
Model fixed effects	Yes	Yes	Yes	Yes	No	No	No	No

This table presents out-of-sample R^2 of predicted excess returns using different models and feature samples, as well as difference Δ and t-statistic. Each model is trained on the full dataset, and variables are subsequently removed by replacing their respective test set values with randomized values. The first row (No ESG) contains results where all ESG variables have been randomized. The following rows contain results where each category of ESG is added individually, and (All ESG) contains results using the full dataset including all ESG variables. Estimates excluding known risk factors such as size, value and momentum are also included for comparison. Due to the stochastic element introduced by randomizing variables, the procedure is repeated for n number of simulations per model in order to minimize idiosyncratic effects. The left section contains results in which multiple models are trained and values are predicted using multiple simulations of noise which are subsequently aggregated individually, controlling for model fixed effects. The right section utilizes a different approach, aggregating all predicted values using noise simulations from one fitted model into a combination forecast. Note: The "all" and "mean" columns summarizing all models in each approach should not be directly compared, as they utilize different methodologies. The "all" column incorporates observations from all models to arrive at a novel estimate, whilst the "mean" column is a simple arithmetic mean.

increase in out-of-sample R^2 when utilizing results from all models (the "all" column of individual simulations). The environmental category is also measured to provide a positive contribution of around 0.20%, whilst the social category contributes negatively at approximately -0.33%. Including all ESG variables yields no significant difference in explanatory power when incorporating all model estimates, but the effect varies across different models.

Elastic net (EN) stands out as the worst performing model, both in absolute terms and in terms of incorporating ESG variables. Due to its nature as a penalized linear regression method, it should not be expected to efficiently incorporate marginal information, and is therefore considered a benchmark model. XGBoost (XGB) and neural network (NN) generally perform better when including ESG, and the neural network model displays higher granularity which is likely due to its level of parameterization.

Calculating the average difference of including each ESG category from only the nonlinear models, including both individual and aggregate simulation approaches, produces the following estimates: environmental (1.44%); social (2.86%); governance (4.54%); all ESG (3.87%). Combining these estimates with the previous significance testing, it can inferred that the previous observations hold. Governance remains the strongest category, followed by environmental. The social category negatively impacts all models except one, but this positive observation is a large outlier and thereby biases the average estimate. Due to the findings from significance testing of all individual simulations estimates which also control for model fixed effects, the impact of the social category should be regarded as statistically insignificant.

The "all" column of individual simulations contains coefficients calculated from the largest number of estimates. Therefore, these results are likely to be the most robust. Here the magnitudes and directional impacts implied by the means are confirmed, with environmental and governance contributing positively and social contributing negatively. Utilizing all estimates, it is not observed that including all ESG-related variables causes a statistically significant difference in model performance. However, this appears to be biased due to the effect of including elastic net estimates. XGBoost and the neural network exhibit increases in explanatory performance of 0.04% and 5.30% respectively. From this it can be inferred that adequate model complexity and nonlinear variable interactions are not only beneficial but necessary in order to effectively incorporate the available ESG information.

The aggregate simulations results serve to further confirm most of the findings drawn from using individual simulations. Due to the smaller number of estimated coefficients, less emphasis should be placed on these results. However, using a different approach might serve to confirm or refute previous findings, as well as increase the robustness with which conclusions are drawn. The XGBoost and neural network models confirm the increase in performance from including all ESG variables, providing 0.80% and 13.22% increase in performance respectively. Furthermore, using aggregate simulations produces a positive contribution also from the social category, which is likely caused by the diversifying effects of aggregation, as discussed previously. Here, the environmental category is the weakest in terms of contribution, conflicting with the comparably stronger results from the individual simulations.

Comparing the effect sizes between ESG variables and established risk factors such as size, value and momentum, the impact of including ESG is very small. Using the same average estimate from nonlinear models as above, the following values are differences when excluding known risk factors: size (-45.49%); value (-49.61%); momentum (-75.06%). However, as has been mentioned previously, in addition to the primary effect from the underlying risk-based explanation, it should be noted that the measurement frequency of each variable introduces a strong bias in favor of variables such as momentum, which are measured at monthly rather than annual frequency. This might cause the underestimation of actual effect sizes for variables measured at lower frequencies.

Next, a Diebold-Mariano test is used in order to compare the forecast accuracy of each model and feature sample. This differs from the approach of comparing out-of-sample R^2 as it examines forecast accuracy measured in mean squared error, as opposed to explanatory power. The Diebold-Mariano test is a significance test of predictive accuracy between two competing forecasts. The predicted excess returns originate from the aggregate simulations method, as it is only possible to use one set of values for each model and feature sample to compare accuracy. Results from the test are presented in Table 5.2.

The results of the Diebold-Mariano test primarily demonstrate that the differences in predictive accuracy both between models and across feature samples are quite small.

Table 5.2: Diebold-Mariano test results using different models and feature samples

		EN					XGB					NN				
		None	E	S	G	ESG	None	E	S	G	ESG	None	E	S	G	ESG
EN	None		-1.01	-1.41	-0.47	1.31	-0.53	-0.99	-1.15	-1.07	0.68	-1.09	-0.84	**2.00**	**2.59**	-1.01
	E	1.01		-1.37	0.23	1.51	0.29	-0.91	-0.99	-0.04	1.35	-0.85	0.40	1.66	1.81	-0.72
	S	1.41	1.37		1.39	1.43	1.39	0.82	1.10	1.39	1.42	1.21	1.39	1.43	1.44	1.25
	G	0.47	-0.23	-1.39		0.83	0.03	-0.93	-1.03	-0.26	0.67	-0.92	0.06	0.92	1.03	-0.79
	ESG	-1.31	-1.51	-1.43	-0.83		-0.96	-1.02	-1.22	-1.57	-0.53	-1.20	-1.55	0.33	0.92	-1.11
XGB	None	0.53	-0.29	-1.39	-0.03	0.96		-0.94	-1.05	-0.33	0.77	-0.94	0.03	1.18	1.21	-0.81
	E	0.99	0.91	-0.82	0.93	1.02	0.94		0.40	0.91	1.00	0.59	0.94	1.02	1.04	0.67
	S	1.15	0.99	-1.10	1.03	1.22	1.05	-0.40		0.99	1.19	0.32	1.06	1.24	1.26	0.47
	G	1.07	0.04	-1.39	0.26	1.57	0.33	-0.91	-0.99		1.47	-0.84	0.44	1.71	1.88	-0.71
	ESG	-0.68	-1.35	-1.42	-0.67	0.53	-0.77	-1.00	-1.19	-1.47		-1.15	-1.20	0.92	1.41	-1.06
NN	None	1.09	0.85	-1.21	0.92	1.20	0.94	-0.59	-0.32	0.84	1.15		0.96	1.22	1.25	0.18
	E	0.84	-0.40	-1.39	-0.06	1.55	-0.03	-0.94	-1.06	-0.44	1.20	-0.96		1.78	**2.17**	-0.84
	S	**-2.00**	-1.66	-1.43	-0.92	-0.33	-1.18	-1.02	-1.24	-1.71	-0.92	-1.22	-1.78		0.79	-1.13
	G	**-2.59**	-1.81	-1.44	-1.03	-0.92	-1.21	-1.04	-1.26	-1.88	-1.41	-1.25	**-2.17**	-0.79		-1.17
	ESG	1.01	0.72	-1.25	0.79	1.11	0.81	-0.67	-0.47	0.71	1.06	-0.18	0.84	1.13	1.17	

This table presents Diebold-Mariano test statistics comparing predicted values of excess returns from each model applied on different feature samples. The test statistic compares forecast accuracies of two given forecasts. A positive value indicates that the column model outperforms the row model and vice versa. It is adapted to the problem of cross-sectional return prediction by estimating prediction errors on average stock returns in each given period, as opposed to on individual return predictions. Each model section contains results from one fitted model and each column or row indicates if and which ESG variables are included. Predicted values are estimated as combination forecasts where omitted ESG variables are randomized and used to form aggregated predictions over 1 000 simulations. (None) indicates that all ESG-related variables are randomized, (E, S, G) indicates that the category is included but the others are randomized, and (ESG) indicates that all ESG variables are included. Bold and outlined indicates statistical significance at the 5% level.

Important to note here however, is that the test is adapted to the cross-sectional return prediction problem by aggregating forecasts at each period. This provides more level grounds for competing between models, but also decreases the significance where differences in individual return predictions might have been large. As was noted previously in the aggregate simulations results, the social category seems to in fact contribute to increased forecast accuracy, and this is most pronounced when incorporated in the neural network model. However, it is again outperformed by the governance category, which is consistently the most significant positive contributor to predictive accuracy. The final conclusion to draw here is the outperformance of the governance category over the environmental category. As has been previously implied through mean and total differences in R^2_{OOS} is confirmed here through a 2.17 test statistic for the neural network model, showing that governance variables contribute more than environmental variables.

To gain a more detailed perspective of the differences caused by including ESG, I perform double-sorting of the results, denoted both in R^2_{OOS} and mean squared error. Individual predicted excess return values for the testing sample are sorted by size (lme) and book-to-market equity ratio (bm) for each monthly period and assigned into quintiles. The quintile breakpoints are determined ex-post for each period. The results are presented in Table 5.3 and indicate estimated differences between each category added and a benchmark measurement using no ESG variables. Colors are applied as a gradient scale from white to green, where darker color indicates better model performance in both tables (higher R^2_{OOS} and lower MSE). Furthermore, color gradients are applied within each model to facilitate comparison between feature samples for each model, but colors should not be compared between models.

In total, the double-sorted estimates of R^2_{OOS} reach the same conclusion as the aggregated results. Mean squared errors however, indicate far more positive effects from including ESG variables. It is important to note the difference in purpose of the two measures. R^2 was selected due to its non-demeaning nature as an estimate of explained variation, which allows it to capture a more detailed estimate of both the positive and negative effects introduced by additional variables. Mean squared error on the other hand, is a more simplistic direct measure of deviation between values. It is therefore expected to display more favorable results when including additional predictors, which is what is observed here. Results from adding variables are mostly non-negative measured in MSE, and larger number of variables seem to systemically lead to higher improvement in predictive accuracy. Both tables reiterate the order of explanatory power for the ESG categories, with governance on top, followed by environmental.

Considering both R^2_{OOS} and MSE, improvements from incorporating ESG variables seem to be generally most pronounced in larger companies. Measuring using MSE also exhibits some large outlier improvements amongst small companies, driven by all models, but primarily by elastic net. It should be noted that there is likely some bias present causing these results. ESG reporting is a highly specialized activity, driven by regulatory but also pecuniary incentives, both of which might be naturally biased towards larger and more resourceful companies. With regards to book-to-market equity, there is mostly no relationship, but might be weakly favoring value firms.

Table 5.3: Difference in accuracy measures by category, sorted by size and BE/ME

Out-of-sample R^2

Size quintile	EN					XGB					NN					Mean				
	Book-to-market quintile					Book-to-market quintile					Book-to-market quintile					Book-to-market quintile				
	Low	2	3	4	High	Low	2	3	4	High	Low	2	3	4	High	Low	2	3	4	High
			Environmental					Environmental					Environmental					Environmental		
Small	-0.05	0.04	-0.46	0.26	-0.04	0.28	0.42	0.02	-0.08	0.22	-0.41	-1.77	-0.03	0.40	-0.65	-0.06	-0.44	-0.16	0.19	-0.16
2	0.03	0.21	1.16	1.57	1.21	-0.01	-0.01	-0.01	0.00	0.03	-0.26	-0.06	-0.23	0.01	-0.47	-0.08	0.05	0.31	0.53	0.26
3	0.09	1.02	0.71	-2.16	0.77	-0.02	-0.01	0.00	-0.01	0.02	0.38	0.17	0.62	0.17	2.40	0.15	0.39	0.44	-0.67	1.06
4	-0.14	-0.70	-0.31	-3.68	0.83	0.00	-0.02	0.01	0.01	-0.02	0.02	-0.66	-0.55	0.61	0.59	-0.04	-0.46	-0.29	-1.02	0.47
Big	1.56	-1.33	1.72	-1.49	0.49	0.00	0.01	0.01	0.03	0.03	0.99	-2.79	0.84	-1.52	4.06	0.85	-1.37	0.86	-1.00	1.53
			Social					Social					Social					Social		
Small	-0.09	0.15	0.59	-0.28	0.04	0.30	0.17	0.10	-0.10	0.00	0.31	0.55	0.97	-0.25	1.20	0.17	0.29	0.55	-0.21	0.41
2	-0.39	-0.05	1.97	0.51	-0.21	-0.07	-0.02	0.00	0.00	-0.21	-0.20	-0.05	0.11	-0.38	0.63	-0.22	-0.04	0.69	0.04	0.07
3	-1.00	1.49	-2.89	-2.29	0.61	-0.05	0.00	0.00	-0.07	-0.20	-1.81	0.27	0.73	1.41	2.60	-0.95	0.59	-0.72	-0.31	1.01
4	1.13	-1.04	0.22	-2.27	-0.63	0.00	0.00	0.00	-0.05	-0.32	2.31	1.70	2.06	0.30	0.28	1.15	0.22	0.76	-0.68	-0.23
Big	2.13	0.20	1.87	-0.64	-0.74	0.00	0.00	0.01	0.01	0.01	2.73	2.34	0.92	0.20	2.91	1.62	0.85	0.93	-0.14	0.73
			Governance					Governance					Governance					Governance		
Small	-0.14	-0.08	0.77	0.49	-0.14	0.40	-1.01	0.14	0.22	0.29	0.21	-1.06	0.29	0.55	1.11	0.16	-0.71	0.40	0.42	0.42
2	2.88	0.04	0.27	0.32	0.28	-0.05	0.07	0.02	0.03	-0.13	-0.26	-0.05	0.19	-0.37	1.54	0.85	0.02	0.16	-0.01	0.56
3	1.16	0.74	2.38	0.46	0.73	0.02	0.06	0.02	0.10	-0.01	0.70	0.27	0.89	2.29	2.37	0.63	0.36	1.10	0.95	1.03
4	-0.57	0.53	0.95	-0.62	-0.69	0.01	0.04	0.00	0.28	0.12	-0.80	1.26	2.26	0.85	0.42	-0.45	0.61	1.07	0.17	-0.05
Big	2.50	-0.08	2.17	0.63	-0.07	-0.02	0.01	0.07	0.02	0.07	4.24	1.70	3.03	2.50	7.54	2.24	0.54	1.76	1.05	2.51
			All ESG					All ESG					All ESG					All ESG		
Small	-0.08	0.03	0.23	0.29	-0.19	1.04	-1.20	0.19	-0.09	0.36	0.95	-1.16	0.10	0.92	0.67	0.64	-0.78	0.17	0.37	0.28
2	2.09	-0.03	0.39	0.33	0.42	-0.13	0.10	0.02	0.02	-0.29	0.26	0.64	0.35	-0.38	1.49	0.74	0.24	0.25	-0.01	0.54
3	-0.71	1.13	-0.54	-2.71	1.13	-0.04	0.08	0.02	0.07	-0.23	-2.68	-0.01	2.54	2.90	3.96	-1.14	0.40	0.67	0.09	1.62
4	1.45	-0.33	-0.47	-4.22	-0.08	0.01	0.04	0.00	0.16	-0.11	-3.68	0.61	1.23	0.98	0.60	-0.74	0.10	0.25	-1.02	0.14
Big	2.49	0.41	4.70	-0.32	1.25	-0.02	0.01	0.07	0.02	0.05	1.81	0.31	4.27	-0.43	-0.27	1.43	0.24	3.01	-0.25	0.34

Mean squared error

Size quintile	EN					XGB					NN					Mean				
	Book-to-market quintile					Book-to-market quintile					Book-to-market quintile					Book-to-market quintile				
	Low	2	3	4	High	Low	2	3	4	High	Low	2	3	4	High	Low	2	3	4	High
			Environmental					Environmental					Environmental					Environmental		
Small	-0.41	-0.33	-0.29	-0.26	-0.25	-1.09	-0.67	-0.04	0.21	-1.81	0.13	0.03	0.18	-0.07	-0.15	-0.14	-0.11	-0.09	-0.09	-0.09
2	-0.28	-0.20	-0.16	-0.16	-0.01	0.04	0.05	0.02	0.00	-0.05	0.32	0.09	0.10	0.24	-0.25	-0.08	-0.06	-0.05	-0.05	-0.01
3	-0.23	-0.20	-0.18	-0.11	-0.06	0.07	-0.02	0.00	0.02	-0.03	-0.01	0.17	-0.11	-0.33	-0.47	-0.08	-0.06	-0.06	-0.05	-0.03
4	-0.28	-0.29	-0.19	-0.18	0.00	0.03	0.03	0.04	-0.01	0.02	0.14	0.30	-0.12	-0.17	-0.70	-0.09	-0.09	-0.07	-0.07	-0.02
Big	-0.34	-0.30	-0.24	-0.14	-0.14	0.01	-0.02	-0.03	-0.03	-0.10	0.19	0.16	-0.08	-0.29	-0.23	-0.11	-0.09	-0.08	-0.06	-0.05
			Social					Social					Social					Social		
Small	0.02	0.00	0.03	0.00	0.00	-0.76	-0.01	-0.13	0.27	0.40	0.46	0.17	-0.22	0.26	-0.07	0.02	0.01	0.00	0.01	0.00
2	0.00	0.00	0.01	0.00	0.00	0.18	0.06	0.01	0.03	0.34	0.15	-0.16	-0.08	-0.10	-0.21	0.00	0.00	0.00	0.00	-0.01
3	-0.01	0.00	-0.01	-0.01	0.00	0.17	0.00	0.02	0.10	0.24	0.14	-0.18	-0.11	-0.31	-0.34	0.00	-0.01	-0.01	-0.01	-0.01
4	-0.02	-0.03	-0.01	-0.03	-0.01	0.00	0.00	0.00	0.04	0.26	-0.06	-0.02	-0.19	-0.11	-0.38	-0.01	-0.01	-0.01	-0.01	-0.02
Big	-0.05	-0.05	-0.04	-0.03	-0.03	0.00	0.00	-0.02	-0.02	-0.08	0.05	-0.04	-0.09	-0.25	-0.19	-0.02	-0.02	-0.02	-0.02	-0.02
			Governance					Governance					Governance					Governance		
Small	-0.49	-0.44	-0.38	-0.29	-0.34	-1.16	2.11	-0.19	-0.58	-1.95	-0.71	-0.28	-0.30	-0.52	-1.99	-0.19	-0.15	-0.14	-0.12	-0.19
2	-0.33	-0.29	-0.22	-0.23	-0.01	0.13	-0.16	-0.05	-0.17	0.68	-0.03	-0.07	0.07	0.08	-0.20	-0.11	-0.10	-0.07	-0.07	-0.01
3	-0.32	-0.28	-0.24	-0.13	-0.07	-0.03	0.19	-0.08	0.08	0.41	-0.11	-0.12	-0.07	-0.22	-0.33	-0.11	-0.10	-0.08	-0.05	-0.03
4	-0.39	-0.39	-0.25	-0.25	0.00	0.12	-0.05	0.00	-0.02	0.38	-0.01	-0.09	-0.16	-0.12	-0.30	-0.13	-0.13	-0.09	-0.09	-0.01
Big	-0.47	-0.39	-0.32	-0.17	-0.19	0.04	0.02	-0.16	0.42	0.15	-0.13	-0.10	-0.17	-0.21	-0.35	-0.16	-0.13	-0.11	-0.06	-0.07
			All ESG					All ESG					All ESG					All ESG		
Small	-0.77	-0.64	-0.53	-0.43	-0.46	-2.90	2.63	-0.25	0.21	-2.01	-0.48	-0.15	0.30	-0.30	-1.69	-0.28	-0.21	-0.17	-0.15	-0.22
2	-0.49	-0.38	-0.28	-0.29	0.05	0.35	-0.28	-0.06	-0.13	0.93	0.59	-0.07	0.07	0.45	-0.66	-0.14	-0.13	-0.09	-0.08	0.00
3	-0.45	-0.36	-0.31	-0.14	-0.03	0.14	0.20	-0.08	0.12	0.66	0.47	0.23	-0.26	-0.73	-0.90	-0.13	-0.11	-0.11	-0.07	-0.04
4	-0.53	-0.55	-0.32	-0.32	0.09	0.14	-0.05	0.00	0.06	0.57	0.58	0.36	-0.27	-0.17	-0.82	-0.16	-0.17	-0.12	-0.11	0.01
Big	-0.71	-0.57	-0.46	-0.20	-0.21	0.04	0.01	-0.16	0.42	0.18	0.38	0.09	-0.23	-0.50	-0.30	-0.22	-0.19	-0.16	-0.08	-0.08

These tables present difference in accuracy measures, R^2_{OOS} and MSE respectively, by adding each category of ESG across size and book-to-market sorted quintiles. Market equity (lme) and book-to-market ratio (bm) variables are used to form five quintiles for each variable, resulting in 25 size-BE/ME portfolios per model per feature sample. Quintile breakpoints are determined ex-post for each period. Measures are estimated as aggregate predictions consisting of predicted values from 1 000 simulations of randomized noise replacing excluded variables. Due to the small size of differences in estimates, reported differences are scaled by 10^3. XGB is scaled by 10^4. In both tables, darker color indicates better model performance and coloring is separate for each model.

In the next section of results, the focus moves beyond the categories and onto the individual variables. First I examine feature importance implied by each model. The measures of feature importance differ for each model, but are normalized to be comparable. Results can be found in Table 5.4.

The first point to note from the feature importance table is again the difference in effects, with ESG variables contributing only very slightly in each model. Top features are biased with regards to measurement frequency as expected, with established risk factors from the literature such as size, value and momentum among the most important. Furthermore, it is apparent from the table which models incorporate a larger amount of marginal information, and which rely more on shrinkage of the coefficients, with XGBoost and the neural network model being more diversified. Among the ESG variables, it might be observed that governance variables appear to lead in terms of importance, however the selection is quite diversified across the different categories. Feature importance by ESG category over time can be found in Figure C.1 of Appendix C. In this figure, it is apparent that the governance category tends to be most important throughout the sample, occasionally being surpassed by the others. Furthermore, the importance of ESG categories seem to correlate somewhat, with periods of high and low general importance placed on ESG variables.

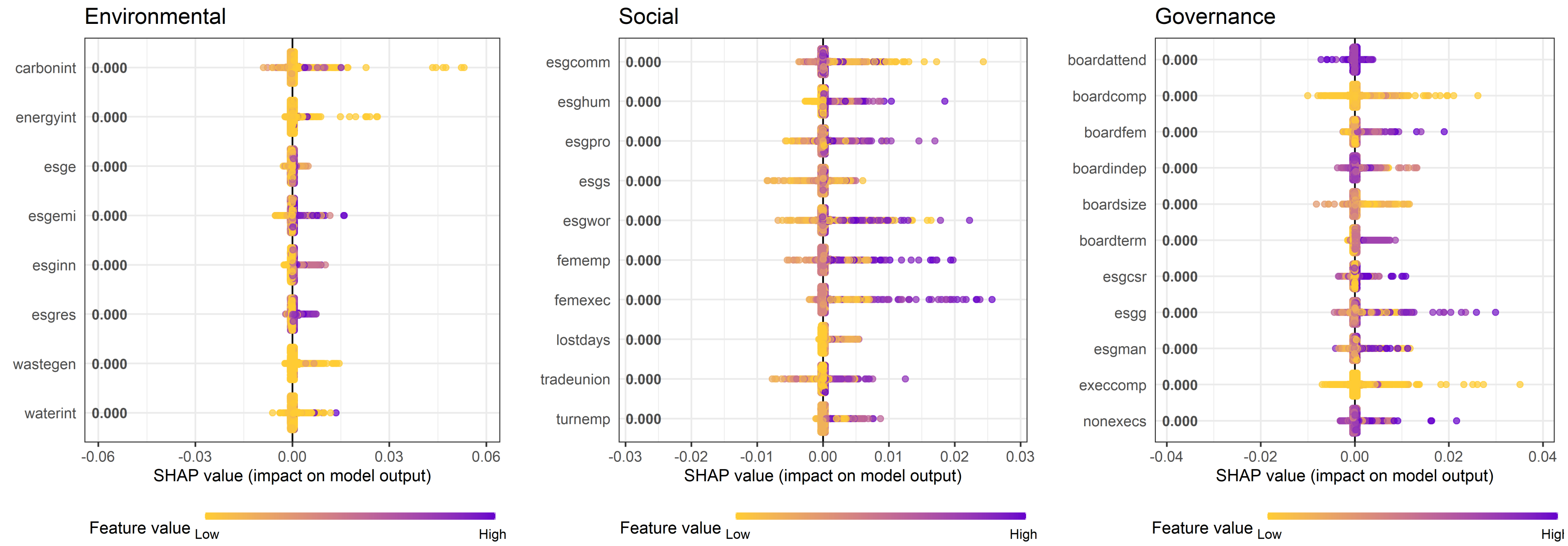
Next, SHAP values for each ESG variable by category can be found in Figure 5.1. I calculate SHAP values using the XGBoost model in order to gain insight into the directional effect and magnitude of impact for each individual ESG variable on predicted returns. The XGBoost model is chosen due to its superior explanatory performance. SHAP is a method used to explain individual output values from a model. The SHAP value can be thought of as the magnitude with which the underlying variable affects the predicted return value. If the SHAP value is positive, the effect is positive with regards to predicted returns, and vice versa. Each estimated SHAP value has a corresponding observation of the underlying variable, which is also included and illustrated using colors. This way, the effect of an observation can be linked to its corresponding actual value. A simple example to demonstrate would be the size effect. Market capitalization of firms are known to exhibit a negative relationship with expected returns. One would expect to observe large absolute SHAP values for the size variable, with positive SHAP values corresponding

Table 5.4: Feature importance by model

Firm characteristics					ESG variables					
Variable	EN	XGB	NN	Mean	Variable	ESG	EN	XGB	NN	Mean
r12_2	0.080	0.073	0.032	0.062	execcomp	G	0.008	0.009	0.013	0.010
lme	0.074	0.041	0.053	0.056	turnemp	S	0.002	0.008	0.016	0.009
strev	0.069	0.038	0.051	0.053	carbonint	E	0.008	0.007	0.011	0.008
r2_1	0.080	0.042	0.014	0.045	esgwor	S	0.002	0.007	0.013	0.007
bm	0.070	0.025	0.040	0.045	esgg	G	0.000	0.009	0.010	0.006
variance	0.050	0.044	0.030	0.042	boardindep	G	0.000	0.006	0.014	0.006
a2me	0.058	0.034	0.029	0.040	boardcomp	G	0.000	0.005	0.013	0.006
spread	0.054	0.033	0.015	0.034	boardfem	G	0.000	0.006	0.013	0.006
chmom	0.047	0.036	0.015	0.032	nonexecs	G	0.004	0.007	0.008	0.006
roa	0.059	0.015	0.013	0.029	esgpro	S	0.000	0.006	0.012	0.006
ep	0.025	0.039	0.021	0.028	esgs	S	0.000	0.006	0.010	0.006
pm	0.046	0.014	0.022	0.027	boardsize	G	0.001	0.004	0.012	0.005
maxret	0.015	0.030	0.034	0.026	esgcomm	S	0.001	0.007	0.008	0.005
rel2high	0.010	0.037	0.030	0.026	esgscore	ESG	0.000	0.006	0.010	0.005
r36_13	0.030	0.037	0.008	0.025	esgman	G	0.000	0.007	0.009	0.005
rna	0.042	0.016	0.015	0.024	lostdays	S	0.000	0.001	0.013	0.005
r6_2	0.020	0.036	0.015	0.024	esgcomb	ESG	0.000	0.002	0.012	0.005
divyield	0.036	0.010	0.023	0.023	boardterm	G	0.000	0.001	0.011	0.004
accrual	0.016	0.020	0.028	0.021	waterint	E	0.000	0.003	0.009	0.004
ivol	0.013	0.029	0.021	0.021	fememp	S	0.000	0.004	0.008	0.004
chsho	0.032	0.016	0.011	0.020	femexec	S	0.000	0.003	0.009	0.004
noa	0.011	0.013	0.030	0.018	boardattend	G	0.000	0.002	0.009	0.004
turnover	0.010	0.026	0.011	0.016	energyint	E	0.000	0.004	0.007	0.004
r12_7	0.000	0.031	0.015	0.015	esgres	E	0.000	0.002	0.008	0.004
zerotrade	0.016	0.001	0.027	0.015	esghum	S	0.000	0.002	0.007	0.003
beta	0.000	0.028	0.015	0.014	wastegen	E	0.000	0.003	0.007	0.003
cf2p	0.000	0.018	0.019	0.013	esgcontr	ESG	0.002	0.001	0.006	0.003
suv	0.003	0.025	0.007	0.012	esge	E	0.000	0.003	0.006	0.003
leverage	0.004	0.012	0.019	0.012	esginn	E	0.000	0.002	0.006	0.003
lat	0.000	0.015	0.011	0.009	esgemi	E	0.000	0.006	0.001	0.002
sga2s	0.000	0.012	0.013	0.008	tradeunion	S	0.000	0.003	0.003	0.002
age	0.000	0.008	0.011	0.006	esgcsr	G	0.000	0.001	0.003	0.001
tobinsq	0.001	0.005	0.006	0.004						

This table provides feature importance for each variable by model, separated into firm characteristics and ESG-related variables. The measured feature importance values are normalized for each model and sorted by the mean value for all models. Values indicate the loss in predictive performance associated with the exclusion of a given variable, or the statistical significance of the coefficient for a given variable. Elastic net (EN) uses absolute value of t-statistic, XGBoost (XGB) and neural network (NN) use loss in accuracy. Colors indicate variable importance as a gradient with darker green color for higher importance.

Figure 5.1: SHAP values for ESG variables by category



This figure contains SHAP values for ESG variables estimated on the training sample using the XGBoost model. Color indicates the value of the variable and the position on the x-axis indicates the magnitude and direction of impact by the observation on the predicted excess return output value. Variables with many large absolute SHAP values (variation on the x-axis) could generally be interpreted as having greater impact on predicted values. As an example, market equity or size (lme) would be expected to display large absolute SHAP values, with purple (high feature value) for negative SHAP values and yellow (low feature value) for positive SHAP values. This would indicate a negative relationship between the variable and predicted excess return, i.e. higher predicted excess return for smaller companies.

to low observed values (indicated by yellow color) and negative SHAP values with high observed values (purple color).

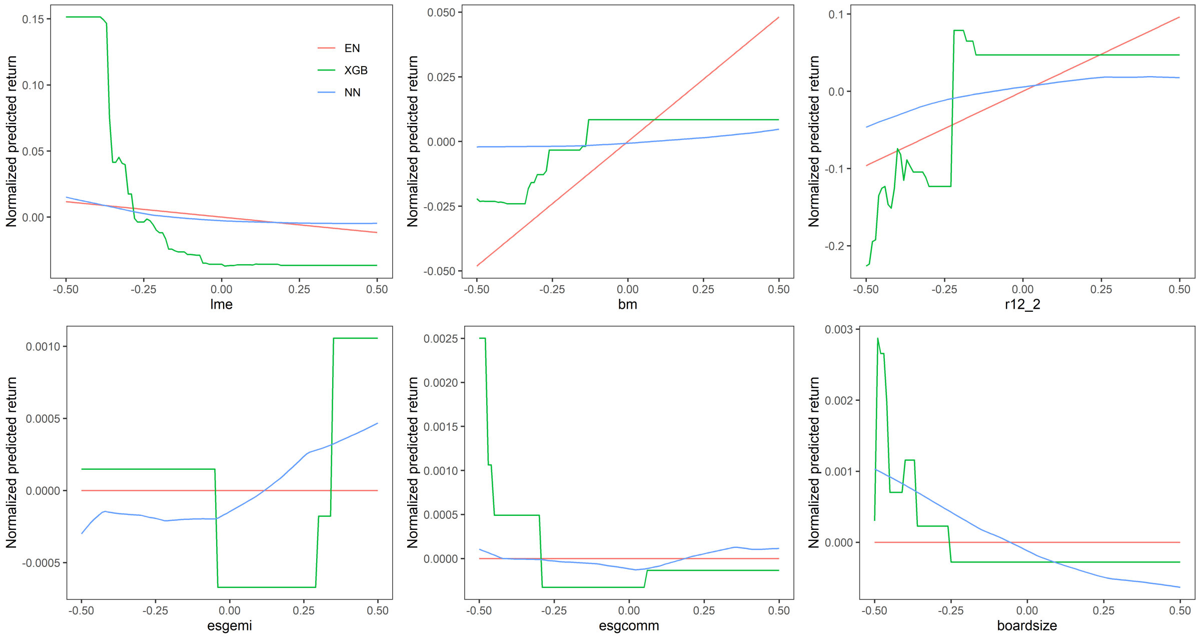
In the SHAP figure (Figure 5.1), a detailed visualization of predicted excess return values is presented. It should be noted that the scale of the x-axes are dynamic and that outlier values might have large impacts. This should be taken into account particularly with regard to the environmental category, which is comparable in impact and magnitude of SHAP values to the other two categories. Variables with mostly small absolute SHAP values can be thought to have relatively small impact on predicted values. The first observation from the figure is the unexpectedly large impact measured from the social category. Given that its impact on explanatory power was generally not found to be significant, it appears to impact the model output disproportionately.

Of the environmental variables, CO2 intensity (carbonint) and energy intensity (energyint) stand out with the largest impact. The effect of environmental measures conform with expectations, wherein higher resource use and total emissions should earn a risk premium. Social variables with high impact seem to be scores relating to the community (esgcomm) and workplace (esgwor), as well as proportional measures of female employees (fememp) and executives (femexec). Governance variables that stand out correspond with those previously identified: compensation-related variables such as board member compensation (boardcomp) and executive compensation (execcomp), governance score (esgg), and board composition factors such as independent (boardindep), female (boardfem) and non-executive (nonexecs) board members.

In the final section of results I explore univariate and bivariate marginal effects and relationships. First, Figure 5.2 examines the univariate effect of a selection of established risk factors and one feature from each category of ESG using each model. Next, Figure 5.3 presents a selection of four ESG variables, each with an additional covariate in order to investigate the development of marginal relationships between two variables with regards to predicted returns.

Figure 5.2 presents six subfigures which are divided into two rows. The top row presents the estimated relationship with predicted excess returns for three established risk factors: size, value and momentum. This is to introduce the method using a familiar set of features, as well as to gain some insight into each model. The figures present univariate effects for

Figure 5.2: Univariate marginal effects by model



The figure presents univariate marginal effects of selected variables over their total range on predicted excess returns for each model. Predicted excess returns have been normalized by subtracting the mean from each model prediction to facilitate comparison. The top row displays the modelled effect of commonly used variables from the literature: size (lme), value (bm) and momentum (r12_2). The bottom row displays ESG variables: emissions score (esgemi), community score (esgcomm) and board size (boardsize). All other features are kept constant at their respective mean values. Note: The scales of the y-axes are not constant, therefore the magnitude of relationships should not be directly compared between subfigures.

each model on normalized predicted excess returns, holding all other features constant. The marginal univariate effects for each of the established risk factors appear to be modelled as expected, with some idiosyncratic differences between modelling approaches. The relationship of size to predicted returns is negative, whilst value and momentum are positive. Notably, the XGBoost model selectively and distinctively emphasizes the effect of some characteristics asymmetrically at the extremes. The size effect is estimated to be very strong for the smallest firms, and likewise momentum causes a large negative impact given highly negative observations. It is clear from the figures that elastic net is limited to linear marginal relationships, whilst XGBoost and the neural network model have nonlinear capabilities.

The bottom row of the figure presents a selection of ESG variables, one for each category of ESG. It is important to note here that when comparing the top and bottom row, the scales of the y-axes are not held constant, which implies that the magnitude of marginal relationships are not directly comparable between subfigures. Modelled relationships for

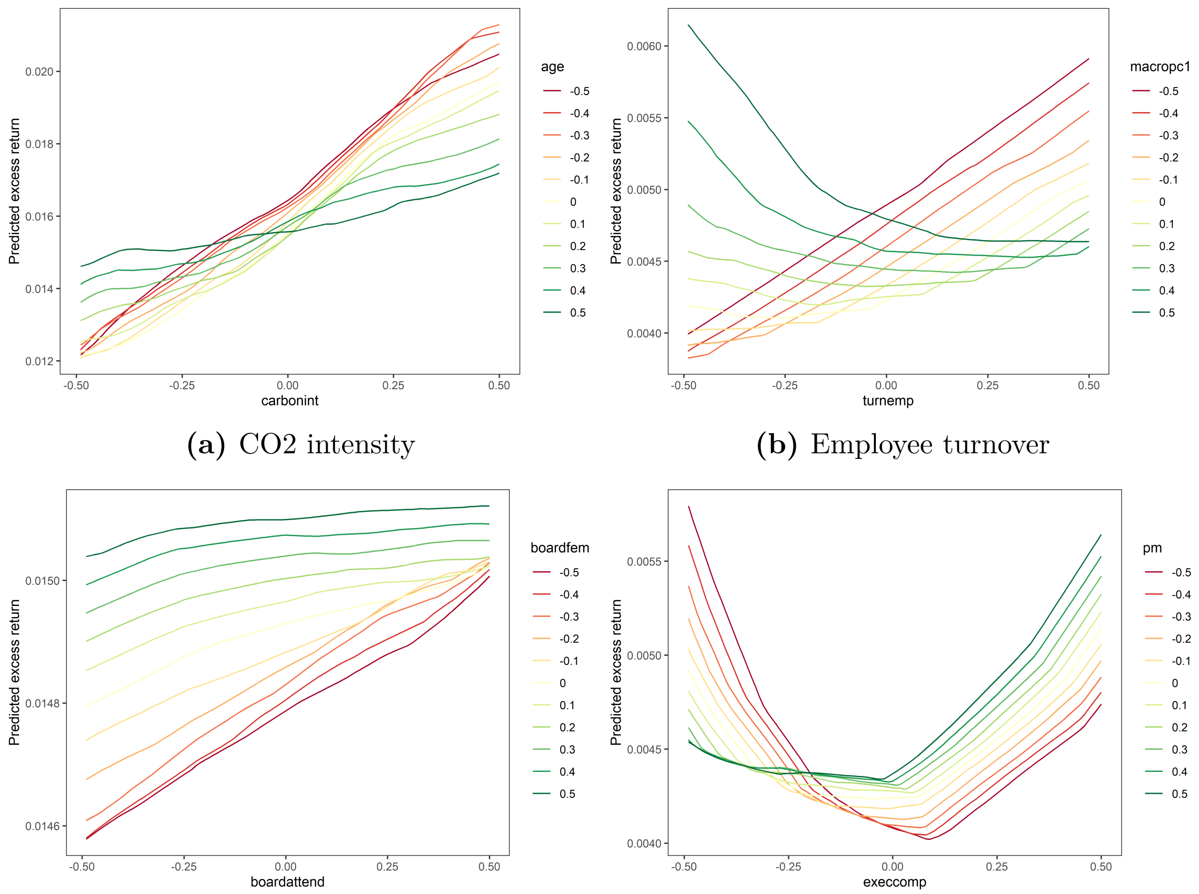
ESG variables are as expected considerably weaker than those of the established risk factors. The models generally tend to be in agreement with regards to the modelled directional relationship. High levels of emissions (esgemi), low community score (esgcomm) and board size (boardsize) all appear to cause higher predicted excess returns.

The final figure (Figure 5.3) provides a more detailed exploration of marginal relationships between selected ESG variables and predicted excess returns, by adding a secondary variable and studying how the primary marginal relationship develops. An important point to note before discussing the results is that all variables were selected ex-post, which poses the risk of selection bias. Discussion of this issue can be found in the next chapter.

The first bivariate marginal relationship (a) examines CO2 intensity (carbonint) as the primary variable, with age of the firm (age) as the secondary variable. CO2 intensity is measured as total CO2 equivalent emissions scaled by revenue. This pair is selected as it provides detail into the CO2 emissions risk premium documented by Bolton and Kacperczyk (2020). Controlling for known risk factors, they find higher returns for firms with higher total CO2 emissions. This effect is also displayed here, with the added contribution of exploring how the relationship differs across firms based on age. It appears that the modelled positive marginal relationship between CO2 intensity and predicted excess returns is steeper for younger firms and flatter for older firms. Speculatively, a portion of this effect might be caused by the types of investors in different firms. With the recent rising prominence of sustainable investing, it is conceivable that investors in younger firms might face different sets of constraints compared to investors in more established corporations, increasing risk premiums. Bolton and Kacperczyk (2020) point to negative or exclusionary screening on the basis of direct emissions intensity as an explanation for the aggregate effect.

Next, subfigure (b) looks at employee turnover (turnemp) as its primary variable and the first macroeconomic principal component (macropc1) as its secondary variable. The macroeconomic variable is intended to serve as a proxy of systematic risk, and should be interpreted as an indicator of the general state of the economy. High values should indicate “good times” or the boom part of a business cycle, and low values should indicate “bad times” or recessions. From this figure we can infer how the relationship between employee turnover and predicted excess returns changes based on the state of the economy.

Figure 5.3: Bivariate marginal effects of selected variables



(a) CO2 intensity

(b) Employee turnover

(c) Board attendance

(d) Executive compensation

The figures show the marginal effect of a primary active variable (x-axis) over the full range $[-0.5, 0.5]$ on predicted excess return (y-axis) using the neural network model. A secondary passive variable is given 10 constant input values whilst the active variable varies over the full range to show how the marginal relationship changes at different levels of the passive variable. All other features are kept constant at their respective mean values. Note: The scales of the y-axes are not constant, therefore the magnitude of relationships should not be directly compared between subfigures.

In times of economic upturn, employee retention earns higher returns and high employee turnover reduces predicted returns. The relationship is most pronounced for very low levels of turnover. In recessions, the relationship is inverted, with higher turnover earning a positive risk premium. Again, interpretation of these relationships is very speculative, but it might be conceivable that firms with large layoffs during recessions are riskier for a variety of reasons, and vice versa for firms with low employee turover.

The third subfigure (c) displays average board meeting attendance (boardattend) as the primary variable, with proportion of female board members (boardfem) as the secondary variable. This selection was made on the basis of the paper by Adams and Ferreira (2009), which posits that female board representation might have a disciplining effect on attendance. The results here do not consider the correlation but rather board meeting attendance as it relates to predicted excess returns. It seems that there might be some evidence in accordance with the study, as low proportion of female board representation appears to make predicted returns more sensitive to attendance. The largest risk premia are found in firms where proportion of female board members are highest, which is also confirmed by the observed SHAP values for these variables. The risk-based explanation for this, as discussed by Adams and Ferreira (2009), might relate to shareholder rights as well as the general governance of the firms.

Finally, subfigure (d) examines executive compensation (execcomp) as its primary variable and profit margin (pm) as its secondary variable. Executive compensation is measured as total senior executive compensation scaled by revenue. This relationship is mostly negative for unprofitable firms and mostly positive for profitable firms. Furthermore, in all cases there appears to be a dip around the median value, indicating that the largest effects are found at the extremes, something which confirms the previous findings from SHAP values. Unprofitable firms granting executives relatively low compensation appear to carry higher risk. There might be a confounding effect relating to which industry the firm operates in causing these results, which is not examined here. Furthermore, profitable firms granting high executive compensation might indicate that these firms operate in more complex, dynamic and competitive industries. As compensation is driven by supply and demand, higher levels might imply higher demand for the scarce resource of competent and experienced leadership. This might be more sought after in more competitive industries, which in turn might carry higher risk.

As a closing note, I would like to caution against expedient extrapolation and inference on the basis of these findings. Marginal effects estimated from a highly complex and parameterized model are rough approximations, and do not account for nonlinear predictor interactions as would be if applied to observed data. As the entire remaining set of features are held constant, a very narrow set of assumptions would have to hold in order to be able to argue that these results will generalize to new data, or even hold in sample. Therefore, reason and rigour should be applied when interpreting the implications of these findings.

6 Discussion

The differences in measurement frequency across the dataset is a crucial aspect in interpreting the empirical findings. It is also one of the primary limitations of the analysis. Measuring, storing, analyzing and incorporating ESG-related factors in the investment process remains a relatively recent practice in finance. Because of this, despite the sometimes ostentatious claims of data providers, such data remain narrow, lack detail and history, are inconsistent across different data providers and are measured at low frequencies. This might often be due to the nature of the variables themselves, but it nonetheless complicates the empirical analysis. A broader dataset might contribute to more detailed and robust findings, but this might also be at the cost of quality and reliability in the ESG data.

As the dataset is constructed using variables measured at a range of different frequencies, the resulting findings will naturally be biased towards certain types of features. This tendency can also be found in papers utilizing similar datasets, such as Gu et al. (2018) and Chen et al. (2020a). When comparing variable importance in models that incorporate both market data with high variance and financial statement data with relatively low variance, variables constructed from market data such as momentum and volume tend to be deemed more important in explaining the cross section. It then follows that if variables measured at even lower frequencies with lower variance than financial statement data are included, these will be considered even less important. For this reason, variable importance measured from variables constructed at different frequencies should not be directly compared. Furthermore, these variables should not be expected to outperform variables of higher measurement frequencies in explanatory power, but should instead be considered for their combined additive contribution. The findings are in accordance with these expectations; including ESG variables demonstrates a significantly lower contribution of explanatory power compared to established, high-variance risk factors.

A wide array of regularization techniques are applied throughout the modelling process. This is common when using highly complex and parameterized models, in order to avoid overfitting and deal with the problem of dimensionality. Moreover, financial data are known for their high level of noise, as such it is crucial to apply regularization in order for the models to generalize to novel data. A problem then arises when trying to compare various feature samples to analyze the effect of certain variables on the explanatory power of the models. If regularization is applied effectively, adding information that is not beneficial will at worst have no effect on the explanatory power of the model. Even entirely irrelevant information might either have no effect or a slightly positive effect, but never a negative effect, which would complicate the analysis and make drawing conclusions difficult. This would be true if models were trained on different feature samples and compared directly against each other, as the models would find an optimal solution with or without irrelevant variables by shrinking their coefficients or otherwise excluding them. This is solved by instead training models using all features and subsequently repeatedly imputing randomized values.

By performing repeated simulations of randomized noise generated in the normalized range and imputing the excluded variables in the testing sample, the explanatory power of included variables can be measured more accurately. Because of the complex nonlinearities of the models, imputing constant values would be insufficient and because of regularization, training on different feature samples would be ineffective. Instead, multiple models are trained on the full feature sample using identical hyperparameters. Next, Monte Carlo simulation is applied by generating predictions using imputed randomized values for certain predictors in order to isolate the effect of these predictors on the explanatory power of the models. Because of the stochastic elements introduced in weight initialization and optimization, it is also important to repeat the simulations across multiple instances of estimated models in order to avoid idiosyncratic effects of individual models.

When examining the bivariate marginal relationships using the neural network model, the results appear interesting, relevant and in accordance with what one might expect ex-ante. However, it is important to note that the selection of variables is performed ex-post through deliberate selection of relationships that yielded the most interesting results. In other words, there is a very strong selection bias present in this sample. Risk of data mining is an entirely valid concern, and with such a high-dimensional dataset it is likely that many relationships are simply present by coincidence. Therefore, it is very important to apply reasonable judgement when interpreting these marginal relationships, and appropriately scrutinize the economic implications.

7 Conclusion

In this thesis I have applied machine learning models to a cross-sectional dataset of excess returns, macroeconomic data, firm characteristics, and ESG-related variables from the three major U.S. stock exchanges. The research question was to investigate whether ESG-related variables contribute to explaining the cross section of expected stock returns, and furthermore how the various variables and categories impact predicted returns.

Using Monte Carlo simulation to isolate category and variable effects, I found that ESG-related variables can contribute to increasing the explanatory power of asset pricing models. However, the impact of ESG is very small compared to more established risk factors such as size, value and momentum. Some categories of ESG contribute more than others. The governance category provided the largest contribution in explanatory power, followed by the environmental category. The social category did generally not contribute to a significant increase in explanatory power, but did impact output values similarly to the other categories when included.

Among the main limitations discussed were the measurement frequency, quality and scope of ESG data. For a variety of reasons, measuring different aspects of ESG at the firm level with the quality, consistency and level of detail that would allow for direct comparison to known anomalies and risk factors from the literature is difficult. Furthermore, the sample is limited by historical availability, as well as to firms meeting coverage criteria, and will for these reasons likely continue to be fragmented for many years to come. In the future, it will probably be possible to examine a far broader dataset, as ESG reporting becomes increasingly standardized and regulated. More comprehensive samples of ESG measures will allow researchers to draw more robust conclusions, and explore the relationship with risk premia in greater detail and with a long-term perspective. Although the effect measured on monthly excess returns was found to be relatively small, the core of sustainable investing primarily relates to the long-term, both with regards to financial- and value-based motives. As the prominence of ESG investing continues to rise, so too will its empirical basis likely continue to develop.

References

Adams, R. B. and Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics*, 94(2):291–309.

Anderson, R. C. and Reeb, D. M. (2003). Founding-family ownership and firm performance: evidence from the s&p 500. *The Journal of Finance*, 58(3):1301–1328.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1):259–299.

Balakrishnan, K., Bartov, E., and Faurel, L. (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics*, 50(1):20–41.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, 99(2):427–446.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18.

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics*, 12(1):129–156.

Berg, F., Koelbel, J. F., and Rigobon, R. (2019). *Aggregate confusion: The divergence of ESG ratings*. MIT Sloan School of Management.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2).

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43(2):507–528.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Bolton, P. and Kacperczyk, M. (2020). Do investors care about carbon risk? Technical report, National Bureau of Economic Research.

Bryzgalova, S., Pelger, M., and Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Chen, L., Pelger, M., and Zhu, J. (2020a). Deep learning in asset pricing. *Available at SSRN 3350138*.

Chen, T., Dong, H., and Lin, C. (2020b). Institutional shareholders and corporate social responsibility. *Journal of Financial Economics*, 135(2):483–504.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chhaochharia, V. and Grinstein, Y. (2009). Ceo compensation and board structure. *The Journal of Finance*, 64(1):231–261.

Chung, K. H. and Zhang, H. (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets*, 17:94–120.

Coles, J. L., Daniel, N. D., and Naveen, L. (2008). Boards: Does one size fit all? *Journal of Financial Economics*, 87(2):329–356.

Daines, R. M., Gow, I. D., and Larcker, D. F. (2010). Rating the ratings: How good are commercial governance ratings? *Journal of Financial Economics*, 98(3):439–461.

Datar, V. T., Naik, N. Y., and Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, 1(2):203–219.

Desai, H., Rajgopal, S., and Venkatachalam, M. (2004). Value-glamour and accruals mispricing: One anomaly or two? *The Accounting Review*, 79(2):355–385.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Dyck, A., Lins, K. V., Roth, L., and Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? international evidence. *Journal of Financial Economics*, 131(3):693–714.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebel, J. (2020). Hedging climate change news. *The Review of Financial Studies*, 33(3):1184–1216.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

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

Feng, G., Giglio, S., and Xiu, D. (2017). Taming the factor zoo. *Chicago Booth research paper*, (17-04).

Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1):1–25.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Friede, G., Busch, T., and Bassen, A. (2015). Esg and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4):210–233.

Gandhi, P. and Lustig, H. (2015). Size anomalies in us bank stock returns. *The Journal of Finance*, 70(2):733–768.

Garfinkel, J. A. (2009). Measuring investors' opinion divergence. *Journal of Accounting Research*, 47(5):1317–1348.

George, T. J. and Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, 59(5):2145–2176.

Gettleman, E. and Marks, J. M. (2006). Acceleration strategies. *SSRN Electronic Journal*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

GSIA (2018). Global sustainable investment review 2018, global sustainable investment alliance. *Available at https://www.gsi-alliance.org/wp-content/uploads/2019/03/GSIR_Review2018*, 3.

Gu, S., Kelly, B., and Xiu, D. (2018). Empirical asset pricing via machine learning. Technical report, National Bureau of Economic Research.

Hale, J. (2021). Sustainable funds us landscape report 2021, morningstar. *Available at https://www.morningstar.com/lp/sustainable-funds-landscape-report.*

Hartzmark, S. M. and Sussman, A. B. (2019). Do investors value sustainability? a natural experiment examining ranking and fund flows. *The Journal of Finance*, 74(6):2789–2837.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, 72(4):1399–1440.

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

Hirshleifer, D., Hou, K., Teoh, S. H., and Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, 38:297–331.

Hong, H. and Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, 93(1):15–36.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., and Shao, L. (2020). Normalization techniques in training dnns: Methodology, analysis and application. *arXiv preprint arXiv:2009.12836.*

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Jiang, G., Lee, C. M., and Zhang, Y. (2005). Information uncertainty and expected returns. *Review of Accounting Studies*, 10(2-3):185–221.

Kaldor, N. (1966). Marginal productivity and the macro-economic theories of distribution: Comment on samuelson and modigliani. *The Review of Economic Studies*, 33(4):309–319.

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

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, 115(2):304–329.

Lins, K. V., Servaes, H., and Tamayo, A. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance*, 72(4):1785–1824.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4):587–615.

Litzenberger, R. H. and Ramaswamy, K. (1979). The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics*, 7(2):163–195.

Liu, W. (2006). A liquidity-augmented capital asset pricing model. *Journal of Financial Economics*, 82(3):631–671.

Lo, A. W. and MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, 3(3):431–467.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review*, 48(3):261–297.

Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. *arXiv preprint arXiv:2010.09337*.

Moritz, B. and Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. *Available at SSRN 2740751*.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the Econometric Society*, pages 768–783.

Noh, D. and Oh, S. (2020). Are green investors green-inducing? a demand system approach. *A Demand System Approach (June 30, 2020)*.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, 103(3):429–453.

Pedersen, L. H., Fitzgibbons, S., and Pomorski, L. (2020). Responsible investing: The esg-efficient frontier. *Journal of Financial Economics*.

Pontiff, J. and Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, 63(2):921–945.

Post, C. and Byron, K. (2015). Women on boards and firm financial performance: A meta-analysis. *Academy of Management Journal*, 58(5):1546–1571.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2013). International stock return predictability: What is the role of the united states? *The Journal of Finance*, 68(4):1633–1662.

Richardson, M. and Smith, T. (1993). A test for multivariate normality in stock returns. *Journal of Business*, pages 295–321.

Ryan Jr, H. E. and Wiggins III, R. A. (2004). Who is in whose pocket? director compensation, board independence, and barriers to effective monitoring. *Journal of Financial Economics*, 73(3):497–524.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review*, pages 289–315.

Soliman, M. T. (2008). The use of dupont analysis by market participants. *The Accounting Review*, 83(3):823–853.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, 30(4):1270–1315.

UN (2020). Annual report 2020, united nations principles for responsible investment. *Available at https://www.unpri.org/annual-report-2020.*

Weigand, A. (2019). Machine learning in empirical asset pricing. *Financial Markets and Portfolio Management*, 33(1):93–104.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320.

Appendix

A Data

Table A.1: List of firm characteristics with database and source information

Variable	Measure	Category	Compustat variable(s)	CRSP variable(s)	Other	Reference
a2me	Assets to market equity	Value	AT	PRC, SHROUT		Bhandari (1988)
accrual	Accruals	Intangibles	ACT, CHE, LCT, DLC, TXP			Sloan (1996)
age	Age of firm	Intangibles				Jiang et al. (2005)
beta	CAPM beta	Trading Frictions			WRDS Beta	Frazzini and Pedersen (2014)
bm	Book to market ratio	Value	TXDITC, SEQ, CEQ, PSTKRV, PSTKL, PSTK	PRC, SHROUT		Fama and French (1992)
cf2p	Cash flow to price	Value	IB, DP, TXDB	PRC, SHROUT		Desai et al. (2004)
chmom	Change in momentum	Past Returns		RET		Gettleman and Marks (2006)
chsho	Change in shares outstanding	Investment		SHROUT		Pontiff and Woodgate (2008)
divyield	Dividend yield	Value	DVT			Litzenberger and Ramaswamy (1979)
ep	Earnings to price	Value	EPSFX			Basu (1983)
ivol	Idiosyncratic volatility	Trading Frictions			WRDS Beta	Ang et al. (2006)
lat	Total assets	Trading Frictions	AT			Gandhi and Lustig (2015)
leverage	Leverage	Value	DLTT, DLC, SEQ			Bhandari (1988)
lme	Size	Trading Frictions		PRC, SHROUT		Banz (1981)
maxret	Maximum daily return	Past Returns		RET	Daily returns	Bali et al. (2011)
noa	Net operating assets	Investment	AT, CHE, IVAO, DLTT, MIB, PSTK, CEQ			Hirshleifer et al. (2004)
pm	Profit margin	Profitability	OIADP, SALE			Soliman (2008)
r12_2	12-2 momentum	Past Returns		RET		Jegadeesh and Titman (1993)
r12_7	12-7 momentum	Past Returns		RET		Novy-Marx (2012)
r2_1	Short-term momentum	Past Returns		RET		Jegadeesh and Titman (1993)
r36_13	Long-term momentum	Past Returns		RET		Jegadeesh and Titman (1993)
r6_2	6-2 momentum	Past Returns		RET		Jegadeesh and Titman (1993)
rel2high	Price relative to yearly high	Trading Frictions		PRC	Daily price	George and Hwang (2004)
rna	Return on net operating assets	Profitability	OIADP			Soliman (2008)
roa	Return on assets	Profitability	IB, AT			Balakrishnan et al. (2010)
sga2s	SG&A to sales	Profitability	XSGA, SALE			Freyberger et al. (2020)
spread	Bid-ask spread	Trading Frictions		BID, ASK	Daily bid-ask spread	Chung and Zhang (2014)
strev	Short-term reversal	Past Returns		RET		Jegadeesh and Titman (1993)
suv	Standardized unexplained volume	Trading Frictions		RET, VOL	Daily returns and volume	Garfinkel (2009)
tobinsq	Tobin's Q	Value	AT, CEQ, TXDB	PRC, SHROUT		Kaldor (1966)
turnover	Turnover	Trading Frictions		VOL, SHROUT		Datar et al. (1998)
variance	Daily variance	Trading Frictions		RET	Daily returns	Ang et al. (2006)
zerotrade	Zero trading days	Trading Frictions		VOL	Daily volume	Liu (2006)

List of all non-ESG firm characteristics used, category, variable names from CRSP and Compustat and reference to the original author(s).

Table A.2: List of ESG variables with database and source information

Variable	Measure	Refinitiv description	Refinitiv variable	ESG
carbonint	CO2 intensity	Total CO2 Equivalent Emissions To Revenues USD in million	TR.AnalyticCO2	E
energyint	Energy intensity	Total Energy Use To Revenues USD in million	TR.AnalyticEnergyUse	E
waterint	Water intensity	Water Use To Revenues USD in million	TR.AnalyticWaterUse	E
wastegen	Waste generated	Total Waste To Revenues USD in million	TR.AnalyticTotalWaste	E
femexec	Female managers	Women Managers	TR.WomenManagers	S
fememp	Female employees	Women Employees	TR.WomenEmployees	S
turnemp	Staff turnover	Turnover of Employees	TR.TurnoverEmployees	S
tradeunion	Working conditions	Trade Union Representation	TR.TradeUnionRep	S
lostdays	Health and safety	Lost Days To Total Days	TR.AnalyticLostDays	S
boardindep	Independent board members	Independent Board Members	TR.AnalyticIndepBoard	G
boardfem	Female board members	Board Gender Diversity, Percent	TR.AnalyticBoardFemale	G
boardattend	Board meeting attendance	Board Meeting Attendance Average	TR.BoardMeetingAttendanceAvg	G
boardsize	Board size	Board Size	TR.BoardSize	G
execcomp	Executive compensation	Total Senior Executives Compensation To Revenues in million	TR.AnalyticSeniorExecsTotalComp	G
nonexecs	Non-executive board members	Non-Executive Board Members	TR.AnalyticNonExecBoard	G
boardterm	Board member term duration	Board Member Term Duration	TR.BoardTermDuration	G
boardcomp	Board member compensation	Board Member Compensation	TR.AnalyticBoardMemberComp	G
esgscore	ESG score	ESG Score	TR.TRESGScore	ESG
esgcomb	ESG combined score	ESG Combined Score	TR.TRESGCScore	ESG
esgcontr	ESG controversies	ESG Controversies Score	TR.TRESGCControversiesScore	ESG
esge	Environmental pillar	Environmental Pillar Score	TR.EnvironmentPillarScore	E
esgs	Social pillar	Social Pillar Score	TR.SocialPillarScore	S
esgg	Governance pillar	Governance Pillar Score	TR.GovernancePillarScore	G
esgres	Resource use	Resource Use Score	TR.TRESGResourceUseScore	E
esgemi	Emissions	Emissions Score	TR.TRESGEmissionsScore	E
esginn	Environmental innovation	Environmental Innovation Score	TR.TRESGInnovationScore	E
esgwor	Workforce	Workforce Score	TR.TRESGWorkforceScore	S
esghum	Human rights	Human Rights Score	TR.TRESGHumanRightsScore	S
esgcomm	Community	Community Score	TR.TRESGCommunityScore	S
esgpro	Product responsibility	Product Responsibility Score	TR.TRESGProductResponsibilityScore	S
esgman	Management	Management Score	TR.TRESGManagementScore	G
esgcsr	CSR strategy	CSR Strategy Score	TR.TRESGCSRStrategyScore	G

List of all ESG variables used, their descriptions and variable names as retrieved from Refinitiv Datastream.

Table A.3: List of macroeconomic variables with transformations

Variable name	Description	Transformation
AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
AWHMAN	Avg Weekly Hours : Manufacturing	1
AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	2
BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	1
BUSLOANS	Commercial and Industrial Loans	6
CE16OV	Civilian Employment	5
CES0600000007	Avg Weekly Hours : Goods-Producing	1
CES0600000008	Avg Hourly Earnings : Goods-Producing	6
CES1021000001	All Employees: Mining and Logging: Mining	5
CES2000000008	Avg Hourly Earnings : Construction	6
CES3000000008	Avg Hourly Earnings : Manufacturing	6
CLF16OV	Civilian Labor Force	5
CMRMTSPL	Real Manu. and Trade Industries Sales	5
CPIAPPSL	CPI : Apparel	6
CPIAUCSL	CPI : All Items	6
CPIMEDSL	CPI : Medical Care	6
CPITRNSL	CPI : Transportation	6
CPIULFSL	CPI : All Items Less Food	6
CUMFNS	Capacity Utilization: Manufacturing	2
CUSR0000SA0L2	CPI : All items less shelter	6
CUSR0000SA0L5	CPI : All items less medical care	6
CUSR0000SAC	CPI : Commodities	6
CUSR0000SAD	CPI : Durables	6
CUSR0000SAS	CPI : Services	6
DAAA	Moody's Seasoned Aaa Corporate Bond Yield	2
DBAA	Moody's Seasoned Baa Corporate Bond Yield	2
DCOILWTICO	Crude Oil, spliced WTI and Cushing	6
DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	6
DMANEMP	All Employees: Durable goods	5
DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	6
DPCERA3M086SBEA	Real personal consumption expenditures	5
DSERRG3M086SBEA	Personal Cons. Exp: Services	6
DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	6
DTCTHFNM	Total Consumer Loans and Leases Outstanding	6
EXCAUS	Canada / U.S. Foreign Exchange Rate	5
EXJPUS	Japan / U.S. Foreign Exchange Rate	5
EXSZUS	Switzerland / U.S. Foreign Exchange Rate	5
EXUSUK	U.S. / U.K. Foreign Exchange Rate	5
FEDFUNDS	Effective Federal Funds Rate	2
GS1	1-Year Treasury Rate	2
GS10	10-Year Treasury Rate	2
GS5	5-Year Treasury Rate	2
HOUST	Housing Starts: Total New Privately Owned	4
HOUSTMW	Housing Starts, Midwest	4
HOUSTNE	Housing Starts, Northeast	4
HOUSTS	Housing Starts, South	4
HOUSTW	Housing Starts, West	4
ICSA	Initial Claims	5
INDPRO	IP Index	5
INVEST	Securities in Bank Credit at All Commercial Banks	6
IPB51222S	IP: Residential Utilities	5
IPBUSEQ	IP: Business Equipment	5
IPCONGD	IP: Consumer Goods	5
IPDCONGD	IP: Durable Consumer Goods	5
IPDMAT	IP: Durable Materials	5
IPFINAL	IP: Final Products (Market Group)	5
IPFPNSS	IP: Final Products and Nonindustrial Supplies	5

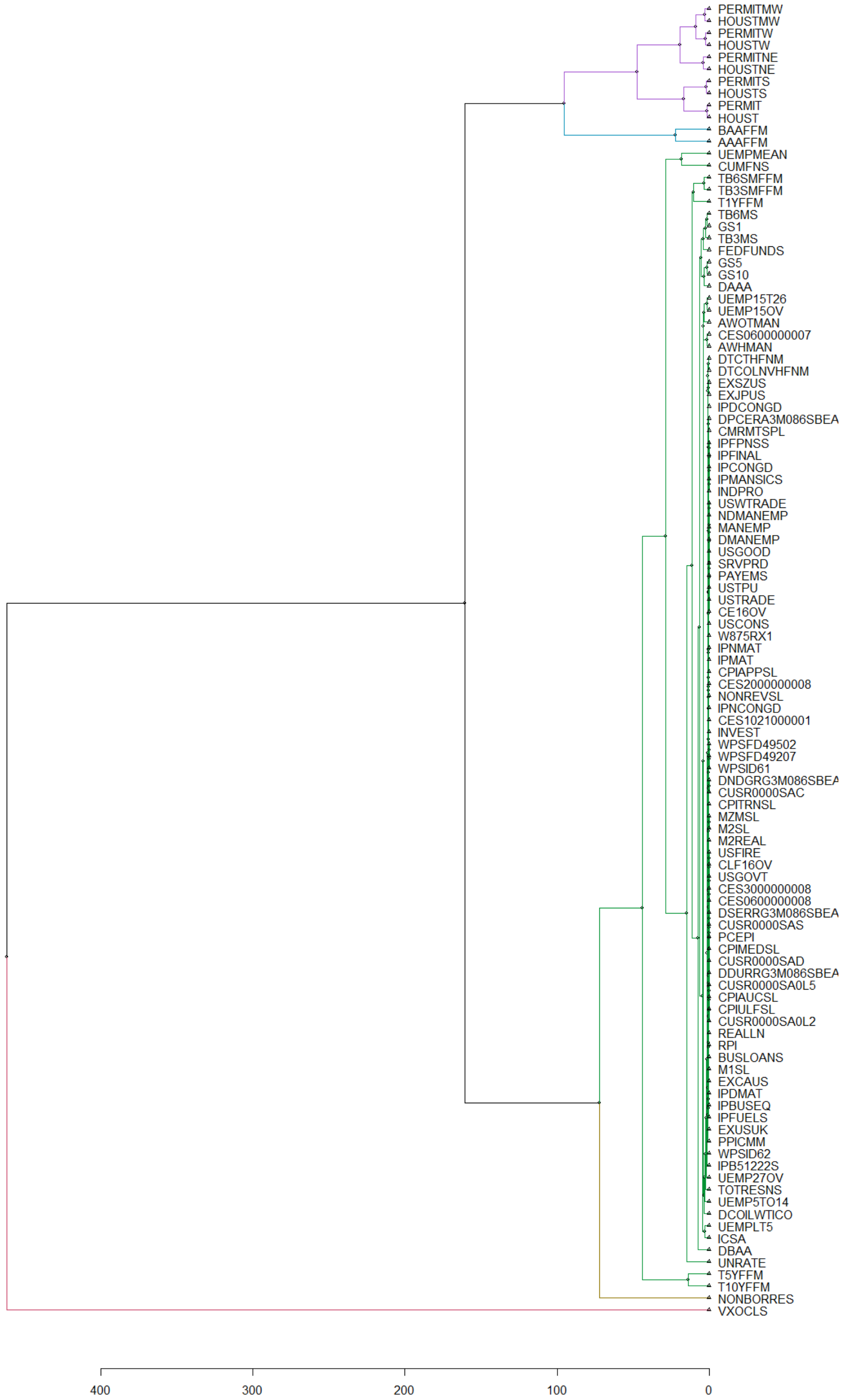
Table A.3 –

Variable Name	Description	Transformation
IPFUELS	IP: Fuels	5
IPMANSICS	IP: Manufacturing (SIC)	5
IPMAT	IP: Materials	5
IPNCONGD	IP: Nondurable Consumer Goods	5
IPNMAT	IP: Nondurable Materials	5
M1SL	M1 Money Stock	6
M2REAL	Real M2 Money Stock	5
M2SL	M2 Money Stock	6
MANEMP	All Employees: Manufacturing	5
MZMSL	MZM Money Stock	6
NDMANEMP	All Employees: Nondurable goods	5
NONBORRES	Reserves Of Depository Institutions	7
NONREVSL	Total Nonrevolving Credit	6
PAYEMS	All Employees: Total nonfarm	5
PCEPI	Personal Cons. Expend.: Chain Index	6
PERMIT	New Private Housing Permits (SAAR)	4
PERMITMW	New Private Housing Permits, Midwest (SAAR)	4
PERMITNE	New Private Housing Permits, Northeast (SAAR)	4
PERMITS	New Private Housing Permits, South (SAAR)	4
PERMITW	New Private Housing Permits, West (SAAR)	4
PPICMM	PPI: Metals and metal products	6
REALLN	Real Estate Loans at All Commercial Banks	6
RPI	Real Personal Income	5
SRVPRD	All Employees: Service-Providing Industries	5
T10YFFM	10-Year Treasury C Minus FEDFUNDS	1
T1YFFM	1-Year Treasury C Minus FEDFUNDS	1
T5YFFM	5-Year Treasury C Minus FEDFUNDS	1
TB3MS	3-Month Treasury Bill	2
TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	1
TB6MS	6-Month Treasury Bill	2
TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	1
TOTRESNS	Total Reserves of Depository Institutions	6
UEMP15OV	Civilians Unemployed - 15 Weeks & Over	5
UEMP15T26	Civilians Unemployed for 15-26 Weeks	5
UEMP27OV	Civilians Unemployed for 27 Weeks and Over	5
UEMP5TO14	Civilians Unemployed for 5-14 Weeks	5
UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	5
UEMPMEAN	Average Duration of Unemployment (Weeks)	2
UNRATE	Civilian Unemployment Rate	2
USCONS	All Employees: Construction	5
USFIRE	All Employees: Financial Activities	5
USGOOD	All Employees: Goods-Producing Industries	5
USGOVT	All Employees: Government	5
USTPU	All Employees: Trade, Transportation & Utilities	5
USTRADE	All Employees: Retail Trade	5
USWTRADE	All Employees: Wholesale Trade	5
VXOCLS	CBOE S&P 100 Volatility Index: VXO	1
W875RX1	Real personal income ex transfer receipts	5
WPSFD49207	PPI: Finished Goods	6
WPSFD49502	PPI: Finished Consumer Goods	6
WPSID61	PPI: Intermediate Materials	6
WPSID62	PPI: Crude Materials	6

List of all macroeconomic variables retrieved from FRED. Transformation codes indicate as follows: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta^2 log(x_t)$

B Methodology

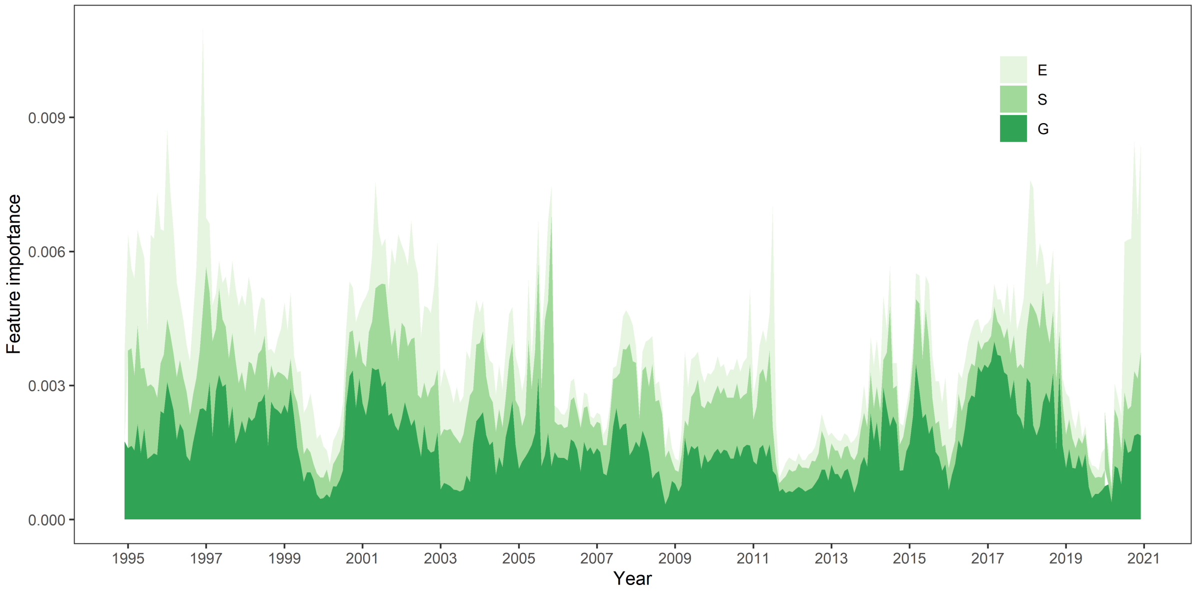
Figure B.1: Hierarchical clustering dendrogram



This figure illustrates the result of hierarchical clustering performed on the macroeconomic dataset. Colors indicate the selected clusters.

C Results

Figure C.1: Importance of ESG categories over time



The figure shows in-sample variable importance by ESG category over time measured using the XGBoost model. Feature importance is normalized per model and the values indicate average aggregated feature importance per category for each period. Models are estimated on a 24 month rolling window and each period represents $n = 100$ estimated models.