Integrating emission indicators in investment decisions – An evaluation of OLS Regression, kNN and Gradient Boosting Classification approaches

Axel Fredric Larsson Bure Burman

Dissertation written under the supervision of Professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, January 2022.

Abstract

This dissertation studies the application of ordinary least squares regressions and supervised machine learning classification models on emission indicator integration on listed share investments. A large set of emission and financial variables are gathered from STOXX600 constituents stretching 2011 - 2020. Implementing a backward elimination feature selection narrow down 60 emission indicators to *Internal Carbon Pricing* and *NOx and SOx Emissions Reduction Initiatives* showing statistically significant relations with next quarter returns. The selected emission indicators are complemented by a set of control variables and used in three approaches to forming investment portfolios. A comparative analysis of the approaches through - a rolling window OLS regression, kNN classification and Gradient Boosting classification - show that a kNN approach to forming percentile portfolios outperform both the regression and Gradient Boosting approach. Both the kNN and Gradient Boosting approaches provide next quarter Up/Down return signal prediction higher than 50%. No approach outperforms a 1/N strategy composed of the source index constituents and only the best ranked percentile portfolio shows statistically significant 3 and 5 factor model alphas in all portfolio creation approaches.

Keywords: ESG indicators, emissions, environmental, STOXX600, Europe, machine learning, k-Nearest neighbours, gradient boosting classification, Internal Carbon Pricing

Title: Integrating emission indicators in investment decisions – An evaluation of Regression, kNN and Gradient Boosting Classification approaches

Author: Axel Fredric Larsson Bure Burman

Resumo

Esta dissertação estuda a aplicação de regressões de mínimos quadrados ordinários e modelos de classificação de aprendizado de máquina supervisionado na integração de indicadores de emissão em investimentos listados. Uma seleção de recursos de eliminação para trás restringe 60 indicadores de emissão para Preço interno de carbono e Iniciativas de redução de emissões de NOx e SOx, mostrando uma relação estatisticamente significativa com os retornos do próximo trimestre. Os indicadores de emissões significativas são complementados por um conjunto de variáveis de controle e implementados em três estratégias de investimento. Uma análise comparativa das estratégias de investimento formadas usando regressões OLS de período de tempo rolando, classificação kNN e classificação Gradient Boosting mostram que uma abordagem kNN para formar carteiras percentuais supera tanto a regressão quanto a abordagem Gradient Boosting. Ambas as abordagens kNN e Gradient Boosting fornecem previsão de sinal de retorno Up / Down para o próximo trimestre superior a 50%. Nenhuma abordagem supera uma estratégia 1 / N composta pelos constituintes do índice de origem e apenas o portfólio de percentil melhor classificado mostra alfas de modelo de 3 e 5 fatores estatisticamente significativos em todas as abordagens de criação de portfólio.

Palavras-Chave: Indicadores ESG, emissões, meio ambiente, STOXX600, Europa, aprendizado de máquina, k-vizinhos mais próximos, classificação de aumento de gradiente, preço interno de carbono

Titulo: Integrando indicadores de emissão em decisões de investimento - Uma avaliação das abordagens de Regressão, kNN e Classificação de Impulso de Gradiente

Autora: Axel Fredric Larsson Bure Burman

Acknowledgments

Starting the MSc in Finance in the middle of a pandemic had its challenges. However, the downs of seemingly endless coursework and uncertainty of the pandemic's development was undoubtedly made less difficult with the support of family, friends, and classmates. All of whom have helped make this dissertation possible.

I would like to specifically thank my supervisor and my family. My supervisor, Dan Tran, for challenging me in the creation of this dissertation and making machine learning approachable. My family for guidance and support, without which neither this degree nor my previous education would have been possible. Thank you.

Table of Contents

List of Tables

List of Figures

List of Appendices

1. Introduction

1.1. Motivation and Context

In recent years, there has been a growing interest among market participants to incorporate environmental, social and governance (ESG) factors into investment decisions. With this, numerous studies have been conducted on the relationship between the ESG score of a company and its financial returns. However, ESG scores are not created from a vacuum. Most rely on a large set of ESG indicators that show the company's previous, current, and planned engagements related to mitigating ESG risks. There have been few studies on these components' relationship with financial returns and even fewer studies have been conducted implementing machine learning models in an ESG indicator – financial return context. This paper aims to fill this gap by studying how well emission indicators perform when implemented in a basic investment strategy through three portfolio construction approaches.

Investments related to ESG is often categorized into three different groups. First, the morally conscious investors could seek to only invest in companies that are aligned with their moral compass, thereby pursuing value-based investing. This category of investors can be considered a passive voice in companies' ownership structure as opposed to impact investing. By pursuing impact investing one could seek to influence and drive change among portfolio constituents through active stock ownership and influencing the companies' board, strategy, or overall operations. Lastly, one could consider ESG-integration, in which investors seek to minimize and handle environmental, social and governance related risks through incorporating ESG factors in an investment strategy. Through studying the impact of emission indicators on financial returns, this paper mainly lies within the ESG-integration field of research.

ESG scores are naturally biased from the model used by the rating agency. Being in a rapid growth and young stage, with the acronym ESG first being coined in 2004, the reliability of scores provided across different agencies and firms differ heavily. There is no clear standardization of what variables are relevant for scoring and how those variables should be defined. Furthermore, the lack of standardization has contributed to a heterogeneity between the scores provided by different rating agencies. In the case of ESG-integration, investors are therefore at the mercy of rating agencies. This paper could prove helpful in identifying which indicators are most material for asset managers to implement into their investment decisions and to what extent. Additionally, by building a model that relies solely on ESG indicators as opposed to ESG scores, investors could lessen their exposure to the heterogeneity and biases of ESG scores.

A commonly used method in portfolio creation uses a regression model to build portfolio. In such a strategy, the stocks are ranked according to desired metrics and formed into portfolios based on their ranking. However, linear regression models simplify what is oftentimes a much more complex data structure. By utilizing machine learning techniques that have become increasingly more accessible, a OLS regression approach to creating portfolios could be supplemented to better take complex data structures into account. This would allow the portfolio creation process to better understand the underlying tendencies of the dataset and minimize the risk of underfitting the data.

This paper studies emission indicators through three approaches. First, an ordinary least squared (OLS) regression approach to creating percentiles portfolios is implemented using the most material emission indicators: Internal Carbon Pricing, NOx and Sox Emissions Reduction Initiatives, Emission Scores and CO2 Equivalents Emission Total. These variables are complemented with financial indicators. Second, the percentile portfolios are created using a k-Nearest Neighbors classification algorithm, using the probability of a positive return as the determining factor of the portfolios. Lastly, a similar approach implementing a gradient boosting classification algorithm is studied. The three approaches are then analyzed separately and jointly based on their financial and model performance.

Through data exploration and a Backward Elimination feature selection, the paper identifies Internal Carbon Pricing and NOx and SOx Emission Reduction Initiatives as indicators which are more important in next quarter returns. Integrating these indicators, along with a set of financial indicators and control variables, the percentile portfolios created using the Regression, kNN and Gradient Boosting approaches all fail to outperform a 1/N portfolio of the index and provide limited economic and statistical significance measured against 3 and 5 factor models. A comparison of the approaches indicates that a k-nearest neighbors' approach was marginally more successful than the regression and gradient boosting approaches in incorporating the emission and financial indicators into the portfolio creation process.

1.2. Objective and Structure

The objective of this paper is to determine to what extent an investment strategy based on emission indicators perform when implemented in a well-established subset of firms. More specifically, it aims to answer three specific questions:

(1) Which emission indicators are especially relevant in ESG-integration?
(2) Are emission indicators viable to use in investment decisions from a risk-return perspective?
(3) Can a simple machine learning approach better the results of the investment strategy as compared to a traditional regression approach?

The paper is structured as follows: A brief description and discussion of the previous literature and relevant concepts related to the research topic. Thereafter a description of the steps to which the data was gathered and treated and the methodology of the investment strategy. The paper is concluded with a discussion of the results and its relation to the previous research on the topic.

2. Literature review

2.1. ESG Indicators in a Financial Context

The concept of integrating other factors than fundamental value in investment decisions is not new. However, for the better part of the late 1900s the Friedman Doctrine of firms' responsibility towards shareholders heavily influenced research and asset managers alike. Research sentiments saw a large shift as James S. Coleman introduced social capital as a resource in investment decisions. As argued by Hart & Zingales, the sentiment that ethical behavior and money making is often interlinked has become more prevalent among asset managers. (Hart & Zingales, 2017) ESG as a concept is even more novel, first being introduces in 2004 as part of by the UN Global Compact article 'Who Cares Wins'. Since, further research has been conducted to showcase how and to what extent corporate social responsibility influence prices and a multitude of ESG scores and measurements have been established. Much of this research has found a nonnegative relation between ESG and returns. (Friede, G., Busch, T., & Bassen, A. 2015). However, the relative novelty of the concept has led to several issues in much of ESG research. Most notably, ESG scores are undermined by the inconsistencies in data and benchmarks. How these inconsistencies are handled by researchers can lead to disagreements and discrepancies. (G. Serafeim & S. Kotsantonis, 2019) Furthermore, a lack of transparency in ESG rating agencies can further reduce their reliability. (Stubbs, W & Rogers. P, 2013)

Three notable organs have been established to combat the issue of inconsistent disclosures. The Global Reporting Initiative (GRI) and the Principles for Responsible Investment (PRI). Furthermore, the Financial Stability Board has launched the Taskforce on Climate related Financial Disclosure (TCFD), looking to establish a framework for consistent disclosures of climate-related financial risks. However, these organs and concepts are still relatively novel and cannot be considered to fully mitigate the concerns addressed by G. Serafeim & S. Kotsantonis.

As to emission indicators potential of providing abnormal returns, most literature focus on the concept from a Corporate Social Responsibility (CSR) perspective and not emissions specifically. As argued by Cheng, Ioannou & Serafeim (2013), better CSR performance can help maintain better access to capital due to reduced agency costs from stakeholder engagement with the firm and increased transparency reducing informational asymmetry. This can be said for much of the literature on the subject. While ESG scores have largely been studied, the literature focusing more specifically on using emission indicators as an investment strategy is however more lacking. Some studies have been conducted on the carbon pricing effects on the performance, risk, and valuation on European firms, most from an economic policy perspective studying the European Union's Emission Trading Scheme (ETS). These have shown the presence of a carbon premium due to higher cash flows among firms allocated a free carbon allowance from the ETS (A.M. Oestreich & I. Tsiakas., 2015). Similarly, the prices set from the ETS positively correlate with energy stock prices (A. Dutta, E. Bouri & H. Noor., 2018), which indicate that investors at the very least factor in direct costs related to emissions.

2.2. Approaches Through Machine Learning

Machine learning techniques are normally split into two categories: supervised and unsupervised learning. The two methodologies differ in what type of data is used and how that data is represented. Supervised learning uses labeled data both on the input and output side to regress or classify the stated problem as opposed to unsupervised learning, which is characterized by unlabeled data, thus requiring no human intervention. Using machine learning to research risk-return questions in finance has been approached using both supervised and unsupervised methodologies. Similarly, unsupervised and supervised methodologies alike have successfully been implemented into return signaling classification. On the unsupervised side Artificial Neural Networks (ANNs) and Deep Neural Networks (DNNs) could help predict return signaling. (X. Zhong & D. Enke., 2019). On the supervised side, decision trees and random forests can be implemented to improve understanding of expected asset returns. (Gu. S, Kelly. B & Xiu. D., 2018).

While machine learning in asset pricing and return is not a novel subject, the implementation of machine learning models in ESG research is less explored. Specifically, two instances of machine learning related to ESG-integration have been found. Lanza, Bernardini & Faiella study to what extent ESG indicators manage climate risk through a large subset of ESG indicators and a decision tree model approach to creating optimized portfolios. (Lanza. A, Bernardini. E & Faiella. I., 2020) Engle, Giglio, Kelly & Lee study how climate risks can be hedged through the usage term frequency–inverse document frequency to interpret climate news. (Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. 2020).

3. Data

This section describes the process, assumptions made, and treatment of retrieving the necessary data used in conducting the research.

3.1. Data Retrieval

When studying ESG issues, the inconsistency in how companies report on ESG factors, and the heterogeneity among existing ESG rating agencies must be noted and treated to minimize bias. This research will rely on data from only one ESG agency, thus avoiding biases from inconsistency in data, but instead being limited in the results interpretability. Emission variables reported by Reuters Eikon as part of their ESG indicators are gathered from Datastream. The first selection of emission variables shows 60 potential variables, shown in Appendix 1. Once selected, the variables are then sorted based on if they are categorical or continuous. 40 continuous variables and 20 categorical variables are identified. The list of variables is then filtered based on the total list's characteristics. Variables are deleted based on if they are only relevant for a specific industry or type of company and if there are homogeneity among the existing variables. For example, the variables Total Waste and Waste Recycled Total are already covered by the provided Waste Recycling Ratio variable, as a result they are not used. Furthermore, some variables taken as is would likely yield misleading results. In the case of 'CO2 Equivalent Emissions Indirect Scope 3 To Revenues USD in million' for example, leaving the variable unadjusted would not consider the different emissions required by different industries. To minimize such effects, industry dummy variables based on Thomson Reuters Business Classification (TRBC) and country dummy variables are formed.

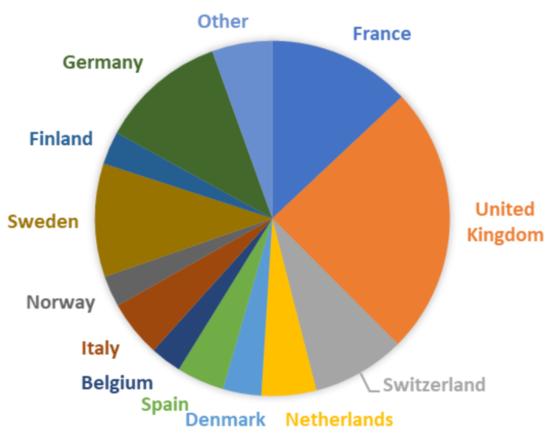


Figure 1. Number of constituents per country in STOXX600

As a supplement to the emission indicators, financial indicators, adjusted to logarithmic values, are gathered from datastream to broaden the scope of the model. These are gathered to represent a wide set of firms' financial characteristics and consists of: Book-To-Market, Price/Earnings and Market Value. Moreover, index constituents' daily price data are gathered stretching the whole period and used to calculate the quarterly returns. The cumulative quarterly return of the last date of each quarter is kept and constitutes the return variable in the paper. These returns are merged with the environmental indicators and lays the foundation of the following methodology and investment strategy.

3.2. Data Treatment

While many variables are already provided in a ratio form, some standardizations are still needed to provide comparable results. Some variables are gathered unadjusted in any sort of sense, as is the case for 'CO2 Equivalent Emission Total'. While the effect of the country and industry of the companies is considered, there is also the issue of the constituents' size. A larger firm will likely have a larger natural tendency to emit. Left unadjusted the model would therefore be largely biased by the size differences between firms.

There are multiple ways of treating size effects in the model to remove biases. As noted by De Franco et al. 2018, including a size factor into the model, regressing the companies separately based on the company size classification or adjusting the necessary variables would mitigate size bias. This paper follows the example set by Lanza, Bernardini, Faiella 2020 by adjusting the necessary variables by the firm's reported revenue of the same year. Adjusting for market cap could mean a bias in the form of differences across companies in how the company is valued. A higher market cap does not necessarily mean that the company is larger in the sense of their operations. By adjusting for revenue, only the variables which are necessary are adjusted and the size of the firm's operations is included in the following models.

As noted by G. Serafeim & S. Kotsantonis, ESG research are exposed to inconsistency in data reporting. This inconsistency notably takes form in missing values, which can arise because of multitude of reasons. Some companies might not report a specific variable due to lack of resources, that the firm prefers to report nothing as opposed to bad news, but also because of non-relevancy for the company's industry. As a result, when unfiltered, there are a large subset of missing values in the dataset. The treatment of missing values could potentially have a strong impact on the paper's results. Left are two realistic options: either drop all missing values, leaving a substantially smaller dataset with limited explanation power, or extend the previous known value over all missing values. A combination of the options is used, first extending the data forward wherever possible, dropping variables that have fewer than 50% reported values, and lastly dropping all missing values.

3.3. Feature Selection

A commonly used method in feature selection is a backward elimination wrapper method. The methodology consists of a loop through all variables, starting with a model including all relevant variables, then removing one at a time. As the loop proceeds, variables that fail to satisfy the set p-value constraint will be dropped. Applying a backward elimination wrapper method on a rolling window basis to estimate which features have the largest statistical explanatory value to explain returns over the whole time series show that a multitude of the emission indicator fail to provide sufficient explanatory value to estimate returns. Similarly, implementing a Recursive Feature Elimination method integrating a linear regression model of ranking the features according to their importance and finding the optimal number of features show that only a fraction of the variables together shows the highest accuracy score. Furthermore, when considering the Pearson Correlation heatmap shown in Figure 2, few, if any variables show a noteworthy correlation with return. The firms' Scope 3 emission weighted by revenue show the highest correlation among the emission variables and log(P/E) among all variables. However, with growing interest among investors to apply ESG integration in their investment decisions there is reason to believe that current market conditions are not effectively reflected in the visualization.

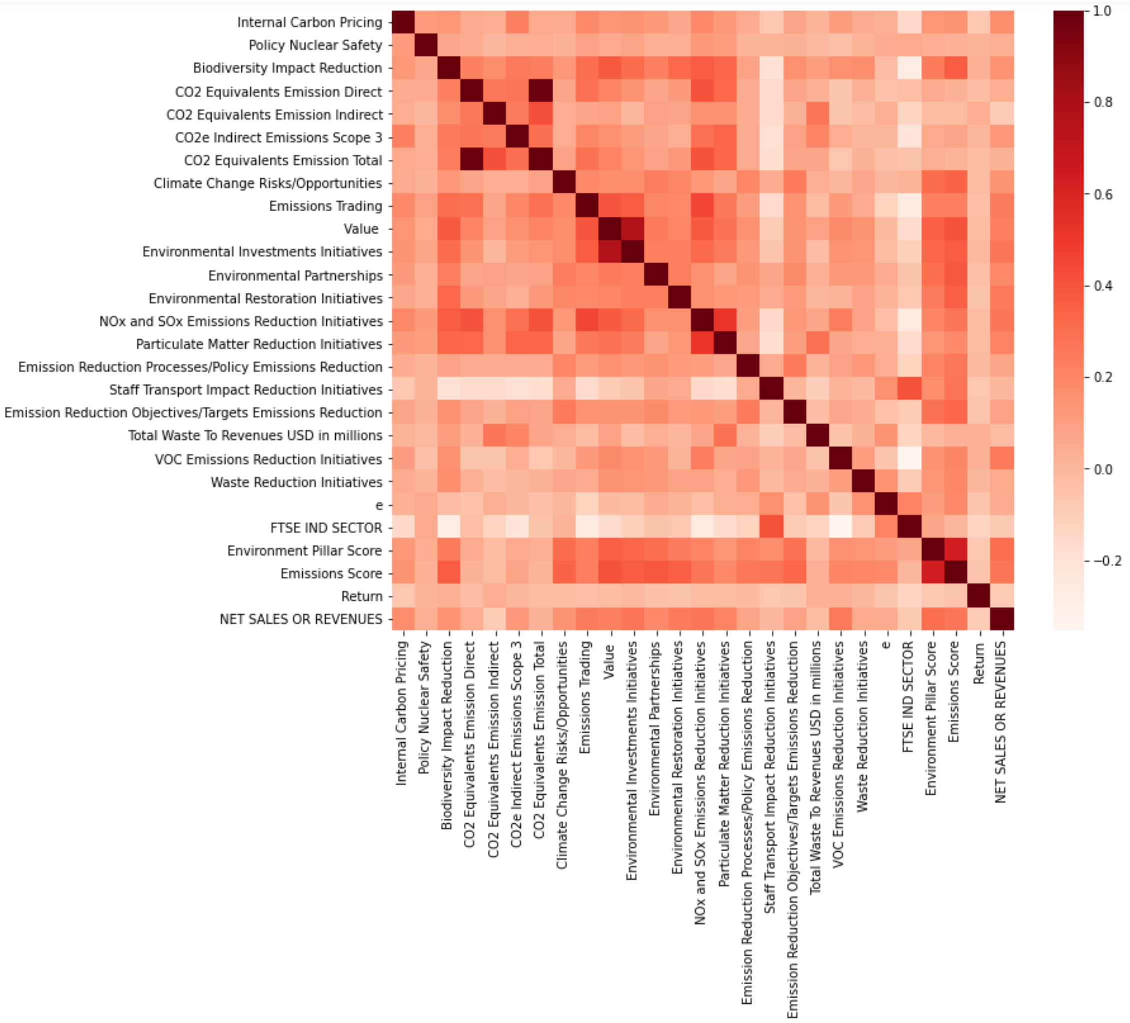


Figure 2. Pearson correlation heatmap matrix. Including data from the whole timeseries 2011-2020.

Variables that fail to prove to have sufficient explanatory value, defined as a p-value of 0.05 within the backward elimination, are dropped and not used in the subsequent models. The backward elimination method and the subsequent analysis results in two emission indicators showing results within the model's cut-off limits: Internal Carbon Pricing and NOx and SOx Emissions Reduction Initiatives. As a result, all approaches incorporate the following variables: Book-to-market, Price-earning, Market Value, Industry, Country, Internal Carbon Pricing, NOx and SOx Emissions Reduction. Furthermore, the Emissions Score and the C02 Equivalent Emission Total are included to better the representation of the companies' overall emission data.

3.4. Data Description

The STOXX 600 is constructed of the 600 largest companies written and active in Europe. The largest markets of constituents in the index consists of United Kingdom, France, Germany, and Sweden, together accounting for 59% of all constituents. Although no immediate and highly influential difference between the countries have been noted it is likely that factors such as the environmental sentiment, laws and financing options vary slightly between the countries. These differences are covered by country dummy variables. While accounting standards are to an extent similar across markets, ESG reporting has yet to see the same standardization despite aforementioned frameworks of reporting. However, there is little differences across the different industries and markets in tendencies to report on emission data, the number of companies within each industry largely correlates with the number of observations after cleaning the data.

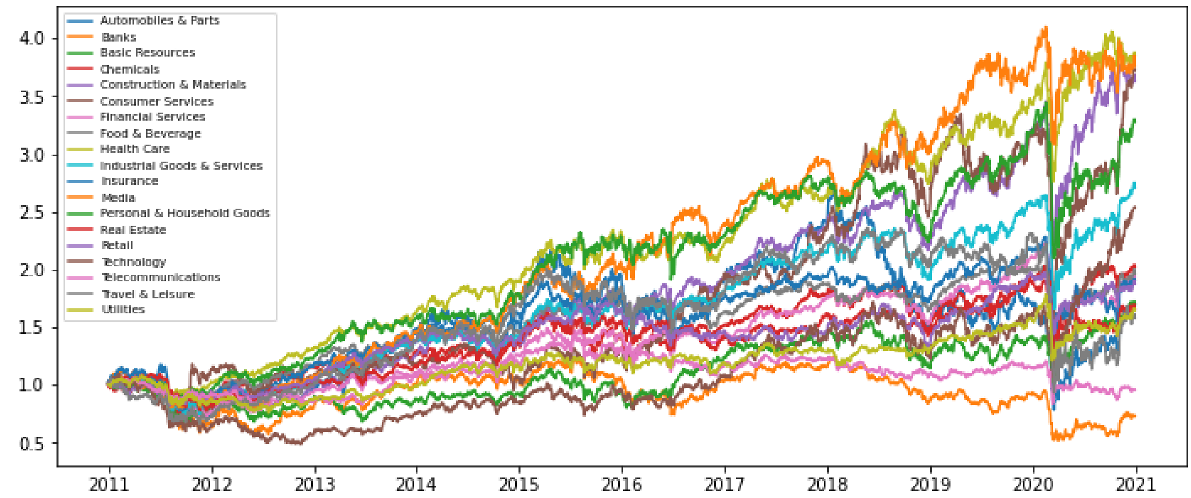


Figure 3. Cumulative return per industry 2011 - 2020.

Industrial Goods & Services account for the better part of the observations of the cleaned data, as well as the largest part of the constituents in STOXX 600, followed by Banks, Utilities, Health Care and Personal & Household Goods. While there is little difference across industries in tendencies to report on emission data, in order to understand the following results, it is important to understand the differences in industries regarding to what extent they emit and differences in how far they have come in emission reduction. Certain industries, although they do not represent the larger part of the constituents, show a larger percentage of companies that report on having implemented internal carbon pricing and NOx and SOx Emission Reduction Initiatives (NSERI). Specifically, Consumer Services, Basic Resources, Utilities, Travel & Leisure and Automobiles & Parts all show 50% or more of constituents reporting that they have implemented reduction initiatives. Internal Carbon Pricing, considered a larger commitment of companies, show a lesser percentage of constituents having implemented across all industries. Similar to the case of NSERI, consumer services show the comparably largest percentage of companies reporting on implementation of Internal Carbon Pricing, followed by Chemicals and Basic Resources.

Once cleaned there are 7365 observations for each variable. Due to the recent growth in companies that report on emission indicators, the dataset is negatively skewed towards more recent years and naturally to companies that report on emission indicators. Before 2017, no companies had reported that they had implemented internal carbon pricing. NOx and Sox Emission Reduction Initiatives does however not show a similar tendency. Instead, as many industries see decreased percentages of companies reporting on implemented NSERI as industries that see increased percentages of companies reporting on implemented NSERI when comparing the period 2011-2016 to 2017-2020.

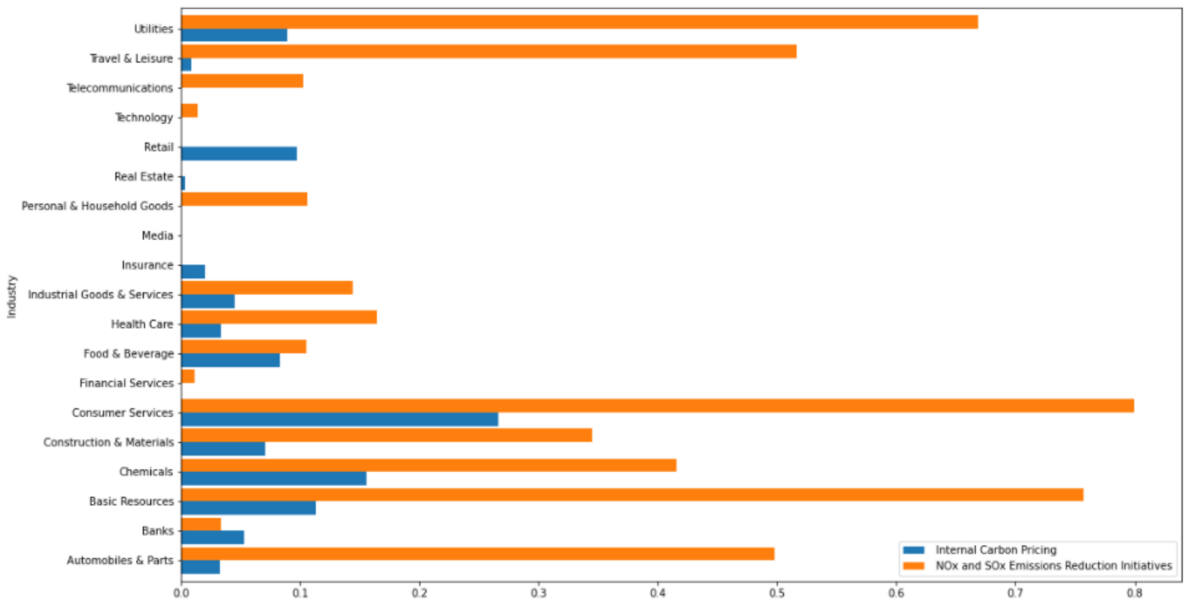


Figure 4. Percentage of companies reporting on Internal Carbon Pricing and NOx and SOx emissions reduction initiatives.

Variable	Description
Internal Carbon Pricing	Does the company have an internal price on carbon?
NOx and SOx Emissions Reduction	Does the company report on initiatives to reduce, reuse, recycle, substitute, or phase out SOx (sulfur oxides) or NOx (nitrogen oxides) emissions? - Any new project undertaken or initiated to reduce NOx (nitrogen oxide) & SOx (sulphur oxide) emissions - General legal compliance is not qualified data - Inline with the legal compliance or governement imosed processes to reduce SOx (sulfur oxides) or NOx (nitrogen oxides) which are well described are qualified - Follows green house gas (GHG) protocol for all our emission classifications by type"
Emissions Score	Emission category score measures a company's commitment and effectiveness towards reducing environmental emission in the production and operational processes.
Emissions Total	Total Carbon dioxide (CO2) and CO2 equivalents emission in tonnes. - Following gases are relevant : carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), hydrofluorocarbons (HFCS), perfluorinated compound (PFCS), sulfur hexafluoride (SF6), nitrogen trifluoride (NF3) - Total CO2 emission = direct (scope1) + indirect (scope 2) - Follows green house gas (GHG) protocol for all our emission classifications by type - - Weighted by revenue in USD
Market Value	Market value on Datastream is the share price multiplied by the number of ordinary shares in issue. The amount in issue is updated whenever new tranches of stock are issued or after a capital change. - For companies with more than one class of equity capital, the market value is expressed according to the individual issue. - Gathered in USD
Book to Market	$\frac{Book\ value\ per\ share \times Number\ of\ Ordinary\ Shares}{Market\ Value}$
Price Earnings	$\frac{Price\ per\ share}{Earnings\ per\ share}$

Table 1. Description of featured emission variables. Description gathered from the notes of Thomson Reuters Datastream. All variables displayed in currencies are gathered in USD. Financial variables are used in logarithmic form.

4. Methodology

This section outlines the methodology used in the research. The methodology consists of two parts. First the methodology regressing or classifying the indicators on returns using a regression, k-Nearest Neighbors and Gradient Boosting approach. Second, the methodology of creating and understanding the portfolios.

4.1. Regression Approach

The first portfolio formations follow a simple ordinary least squares regression approach to access the validity of an investment strategy based on the emission and financial indicators gathered from the feature selection. As seen in formula (1) the return is regressed on Book-To-Market, Market Value, Price-Earning, Internal Carbon Pricing, NOx and SOx Emission Reduction Initiatives, CO2 Equivalent Emissions Total, Emissions Score as well as industry and country dummy variables of the previous quarter. The return of the following quarter is then predicted based on the included indicators and used to form percentiles portfolios based on the predicted return ranking, where the 10th percentile portfolio is the portfolio consisting of the stocks with the highest predicted returns. To reduce time varying differences in how the independent variables are reported the method is applied on a rolling basis with four year rolling windows. The portfolios are then rebalanced on a quarterly basis. A long-short strategy would consist of buying the portfolio with the highest predicted returns (10th portfolio) and shorting the portfolio with the lowest predicted returns (1st portfolio).

$$\begin{aligned} R_{it} = \alpha_i &+ \beta_{1i} \log(BM)_{t-1} + \beta_{2i} \log(MV)_{t-1} + \beta_{3i} \log(PE)_{t-1} \\ &+ \beta_{4i} ICP_{t-1} + \beta_{5i} NSERI_{t-1} + \beta_{6i} CO_2\ Emission\ total_{t-1} \\ &+ \beta_{6i} Emissions\ Score_{t-1} \\ &+ \beta_{6i} Country\ Dummy\ Variables_{t-1} \\ &+ \beta_{6i} Industry\ Dummy\ Variables_{t-1} + \varepsilon_{it} \end{aligned} \tag{1}$$

4.2. Classification Strategies

Unlike the regression approach in which a quantitative return is predicted and used to form portfolios, this paper's machine learning approaches use a classification approach to forming the percentile portfolios. Two models are used in this sense: k-nearest neighbors and gradient boosting. Although both models could be implemented in regression problems, the infinitively large possible returns each iteration can take makes the models unsuitable in this case. Instead, a novel approach of predicting the companies' up and down signals on a four-year rolling window basis are used and the probability of an up signal in next period is used to form percentile portfolios. The portfolios are therefore not formed based on the predicted size of the coming return, but the likelihood of a positive return. The best predicted percentile portfolio is the portfolio consisting of stocks with the highest predicted probability of showing positive returns. This is reflected in the 10th percentile portfolio. The following two chapters explain the k-Nearest Neighbor and Gradient Boosting approaches in more detail.

4.2.1. k-Nearest Neighbors Approach

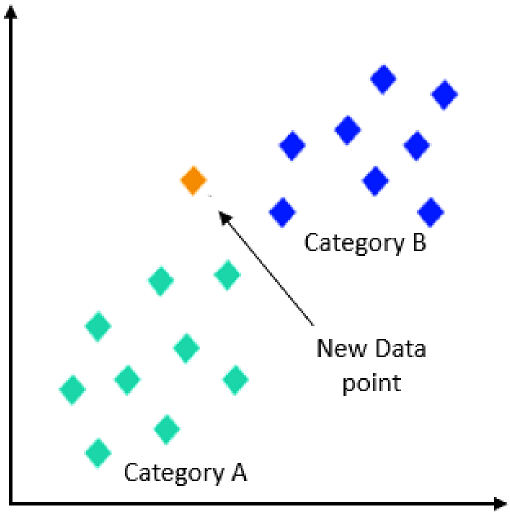


Figure 5. Visualization of a k-Nearest Neighbors classification problem with two clusters.

k-Nearest Neighbors (kNN) is one of the oldest machine learning methodologies implementable in both regression and classification problems. The basic concept relies on the notion that similar things are close to each other in characteristics. As visualized in Figure 5, in a dataset with two categories visualized through a scatterplot, a new uncategorized data point is put in a category through the k-nearest neighbors approach based on its x and y value. As such, the kNN method classifies a new data point into category A or category B based on the characteristics of category A and category B datapoints in the training dataset. As a result, the Euclidean distance, as seen in equation (2), between observations is of essence in a kNN model. kNN is not a probabilistic model, although the probability of the new data point being in category A can be induced through an underlying probability formula as seen in equation (3). In the practical kNN implementation done in this research the probability is gathered using the sklearn library in python. The kNN algorithm moreover differs from some other machine learning approaches in the sense that the model does not weigh the importance of the input variables, instead treating all variables as equally important in determining the output.

$$d(x_i, x_k) \stackrel{\text{def}}{=} \sqrt{(x_i^{(1)} - u_k^{(1)})^2 + (x_i^{(2)} - u_k^{(2)})^2 + \ldots + (x_i^N - u_k^N)^2} \tag{2}$$

The kNN approach implemented in this paper consists of a classification of the quarterly return signals and is compared to the regression and Gradient Boosting approaches. The quarterly returns are formed into a dummy variable dependent on if the observation shows a positive or negative return. The methodology follows a similar approach to that of Martinez, Frias, Pérez & Rivera, using a rolling window of four years to train the model and the subsequent years to test the model's ability to predict the return up/down signal. (Martinez, Pérez & Rivera 2019) The portfolios are formed using the model's classification probability of showing a positive return for the specified observation. Similar to the regression method, the probabilities are ranked according to their size and placed into decile portfolios dependent on their rank.

$$\mathbf{Pr}(\mathbf{Y} = \mathbf{j} \mid \mathbf{X} = X_0) = \frac{1}{k} \Sigma \sum_{i \in N_0} I(y_i = j) \tag{3}$$

Implementing a k-NN algorithm on a time series dataset presupposes that thorough data treatment and selection is done. Beyond the treatment explained in detail in chapter 3.2, special attention needs to be paid to the dataset's tendencies of global trends, seasonality and selecting k. (Martínez, F., Frías, M.P., Pérez, M.D. et al, 2019) Global trends and seasonality is in large considered by the rolling window. Selecting the number of neighbors (clusters (k)) are usually done by one of two ways. One commonly used approach consists of setting the number of neighbors to the square root of the number of training observations. The elbow method, another commonly used method, is however a more time-consuming approach that involves optimizing the forecast accuracy through a minimization method to determine the optimal number of clusters. To optimize the model and provide the best possible results, this paper relies on the elbow method in choosing the number of neighbors.

4.2.2. Gradient Boosting Classifier Approach

Gradient boosting classifier is a supervised machine learning method used to classification problems. Its objective consists of minimizing the difference between the predicted value and the real value, I.e minimization of a logarithmic loss function. The algorithm's concept relies on a decision tree like approach in which several iterations of the classification prediction is combined to optimize the model. The target variable is predicted using the training dataset, then checked against the real value and trained again until the model correctly predicts all classification, as visualized in Figure 6. The sum of the prediction thereby becomes increasingly accurate the more iterations it has done. As such the model comes with two threats. As the sum of the prediction becomes increasingly accruate, it also becomes increasingly complex, leaving the model exposed to problems of overfitting the data. This leaves the model exposed and reliant on what inputs are used.

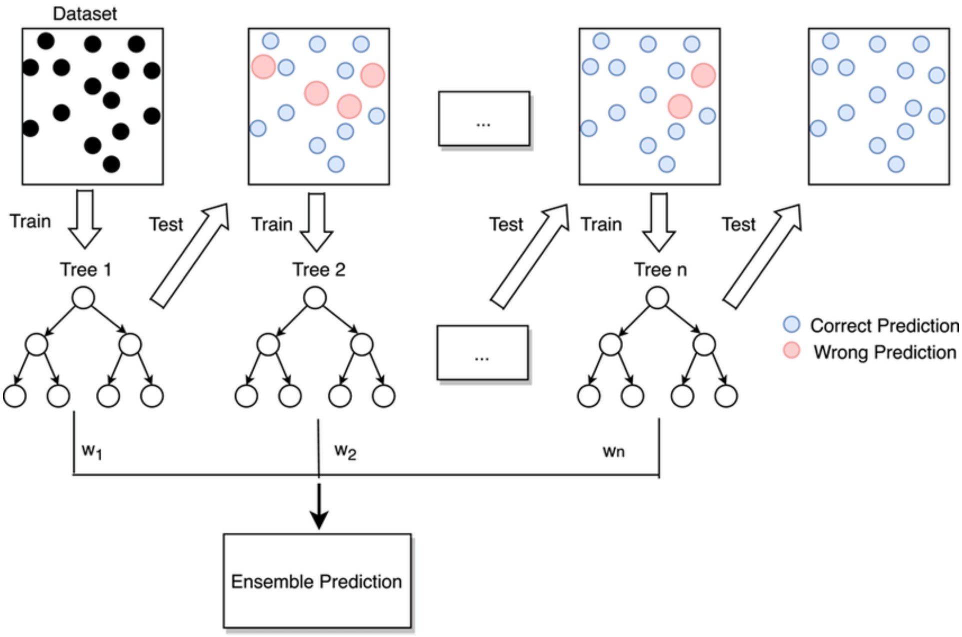


Figure 6. Visualization of a gradient boosting classifier. (Zhang, T., Lin, W., Vogelmann, A. M., Zhang, M., Xie, S., Qin, Y., & Golaz, J.-C. 2021)

The model is heavily dependent on the constrains put on the decision trees, the learning rate, and the sampling methodology. The decision trees can be constrained by several inputs: the number of trees, tree depth, nodes, observations, and minimum improvement to loss by one more tree. More tree constraints generally limit overfitting, but also limits the scope of model. As such, only the depth constraint, which limits the complexity of the trees as visualized in Figure 6, will be constrained regarding the decision trees. The learning rate however determines the size of each step taken to minimize the loss function. A too large learning rate risks the learning overshooting and missing the minimization of the loss function while a too small learning rate requires more computer power and risks finding a local minimum as opposed to the global minimum. The sampling methodology characterizes how the sample data is selected. To reduce correlation between trees a random sampling methodology picks a subsample at random to train the data.

A variety of methods are used to determine the optimal values of the algorithms to optimize the results, while still avoiding overfitting. To effectively limit overfitting, the decision trees are limited to a depth of 4. This allows the model to include the basic components of the dataset, while limiting it from overestimating the testing data's true explanatory power on the return signal. The models time series characteristics makes random sampling difficult to implement without simultaneously creating new biases and limiting the model's accuracy. The learning rate is adjusted and set to 0.1 based on manually testing which learning rate provides the better accuracy.

Similar to the k-Nearest Neighbor approach, the model's probability metric of a positive return in the coming quarter lies as the foundation in the percentile portfolio creation. The probability is ranked and equally sorted into 10 portfolios based on the observations ranking. The stocks which correspond to each portfolios constituents are then invested in and rebalanced on a quarterly basis.

4.3. Evaluation Metrics

		Predicted Values	
		Down Signal	Up Signal
Real Value	Down Signal	True Negative (TN)	False Positive (FP)
	Up Signal	False Negative (FN)	True Positive (TP)

Figure 7. Visualization of confusion matrix k-NN and gradient boosting approach.

Machine learning classification problems are typically evaluated according to their performance metrics as indicated by recall, precision, F-Score and accuracy. These measures can be calculated through the models' confusion matrix, as shown in Figure 7. A perfect confusion matrix would be reflected by 100% of predicted up signals also being real value up signals, and 100% of predicted down signals also being real value down signals. The main metric of concern in the analysis of the results will however be the models' accuracy, as shown in formula (4). Furthermore, the k-NN algorithm and gradient boosting approach will also be reviewed according to its weighted average precision and recall to deeper understand the underlying tendencies of the models' predictions.

$$\boldsymbol{Accuracy = \frac{Correct\ predictions}{Total\ number\ of\ predictions}} \quad (4)$$

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \quad (5)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{6}$$

Precision is defined as the ratio of correct positive predictions to the overall number of positive predictions. and recall, defined as the ratio of correct positive predictions to the number of positive examples.

$$\boldsymbol{F_1} = \boldsymbol{2} * \frac{\boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \tag{7}$$

The F-Score, showing the harmonic mean between the recall and precision, is calculated. (Burkov. A., 2019) As with recall and precision, the F-Score is stronger the closer to 1 it is.

$$R_{\mathrm{it}} - RF_{\mathrm{t}} = \alpha_i + \beta_{1i}\,(RM_t - RF_t) + \beta_{2\mathrm{i}}\mathrm{SMB}_t + \beta_{3\mathrm{i}}HML_t + \beta_{4i}RMW_t + \beta_{5i}\mathrm{CMA}_t + \varepsilon_{it} \tag{8}$$

$$R_{\mathrm{it}} - RF_{\mathrm{t}} = \alpha_i + \beta_{1i}\,(RM_t - RF_t) + \beta_{2\mathrm{i}}\mathrm{SMB}_t + \beta_{3\mathrm{i}}HML_t + \varepsilon_{it} \tag{9}$$

To review the models' value as investment strategies, the resulting percentile portfolios' alpha from the Fama French 3 factor model (9) (Fama and French 1993) and 5 factor model (8) is calculated using data from Kenneth French's Data Library. As a complement, the Sharpe Ratio is used to compare and understand the differences between the regression, k-nearest neighbors and Gradient Boosting strategies.

5. Results and Discussion

This section describes the main results following the investment strategy. First, there is a brief discussion regarding the baseline regressions and the emission indicators explanatory value and robustness. Next, the results of the specified strategies are explored.

5.1. Baseline Regressions and Feature Selection

An initial exploration of the dataset is made to determine what emission indicators show the best explanatory value. Although this paper gathers and starts with a large amount of emission indicators, the feature selection quickly eliminates most of the features from the model. Left are only Internal Carbon Pricing and NOx and Sox Emissions Reduction Initiatives. These are complemented with Emissions Score and CO2 Equivalents Emissions Total to better reflect the full spectrum of relevant emission information.

	Regression						
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-0,03 (-1,24)	-0,03 (-1,46)	-0,02 (-0,78)	-0,02 (-1,02)	-0,03 (-1,28)	-0,03 (-1,6)	-0,03 (-1,65)
Emissions Score	0,00 (-1,84)		0,00 (-2,18)			0,00 (-1,51)	0,00 (-1,59)
Internal Carbon Pricing	-0,04 (-5,56)	-0,04 (-5,68)				-0,04 (-5,46)	-0,04 (-5,46)
NOx and SOx Emissions Reduction Initiatives	0,01 (2,34)	0,01 (2,09)	0,01 (2,09)	0,01 (1,78)			
CO2 Equivalents Emissions Total	0,00 (0,12)	0,00 (0,37)	0,00 (0,17)	0,00 (0,47)	0,00 (0,83)	0,00 (0,62)	
logMV	0,01 (5,07)	0,01 (4,72)	0,01 (4,5)	0,01 (3,98)	0,01 (4,38)	0,01 (5,35)	0,01 (5,44)
logBM	-0,02 (-9,92)	-0,02 (-10,24)	-0,02 (-9,93)	-0,02 (-10,29)	-0,02 (-10,14)	-0,02 (-9,68)	-0,02 (-9,73)
logPE	0,04 (15,49)	0,04 (15,56)	0,04 (15,53)	0,04 (15,61)	0,04 (15,51)	0,04 (15,36)	0,04 (15,34)

Table 2. Baseline regression results of next quarter return on emission and financial indicators. Emission indicators are dropped across regression. Betas are shown with the respective t-value in parenthesis. All regressions include industry and country dummy variables. All regression includes all gathered data 2011 – 2020.

An initial baseline regression seen in Table 2 of the feature variables shows a statistical significance at the 1% level across all financial indicators but varying economic significance. logPE shows the strongest explanatory power in the model with a beta of 0.04 across all regressions, logBM the second strongest with -0.02 and logMV 0.01. The table reinforces the results from the feature selection showing that only two emission indicators having explanatory value in the models. Internal Carbon Pricing shows results of statistical significance at 1% level while maintaining an effect on par with logPE. NSERI is significant at 1% level in combination with Internal Carbon Pricing but 5% if ICP is omitted. Furthermore, the differences of the emission score between the regressions is worth noting. The specific emission indicators show a stronger result than the emission score itself and Internal Carbon Pricing carries some of the explanatory power of the emission score, as seen when considering regression 3, in which the emissions score is statistically significant at a 5% level when Internal Carbon Pricing is omitted.

5.2. Regression Strategy

	3 Factor Model		5 Factor Model		
Portfolio	Alpha	T-Value	Alpha	T-Value	Sharpe Ratio
1 (low)	0,05%	1,91	0,05%	1,90	0,38
2	0,04%	1,48	0,04%	1,47	0,29
3	0,02%	0,82	0,02%	0,80	0,17
4	0,02%	0,68	0,02%	0,67	0,16
5	0,01%	0,50	0,01%	0,49	0,12
6	0,01%	0,45	0,01%	0,44	0,09
7	0,00%	-0,09	0,00%	-0,10	-0,01
8	-0,01%	-0,23	-0,01%	-0,24	-0,04
9	0,00%	0,16	0,00%	0,15	0,04
10 (high)	0,00%	-0,11	0,00%	-0,12	-0,03

Table 3. Results per percentile portfolio implementing a regression strategy to estimating percentile portfolios. The table reports alphas (in percentage) and t-values of the three-factor model and five-factor model. The annul Sharpe ratio of each portfolio is reported.

A regression approach to forming the portfolios show results in line some of the previous research although not statistically significant in this research. As seen in Table 3 the portfolios alphas are formed in reverse. The predicted worst performing portfolios as formed by the model show the highest real alpha and Sharpe ratio, while the best predicted performing portfolios show the lowest real alpha and Sharpe ratio. More specifically, the 1st portfolio shows an alpha of 0.05% from both factor models and is statistically significant at 10% level. The 10th portfolio shows an alpha of 0% without any statistical significance. These results are consistent with those of Trinks & Scholtens (2017). However, the low signifcance and limited Sharpe ratio by the regression approach to estimating returns and forming portfolios could suggest a bias in the form of underfitting the data. Such a problem would not be farfetched considering the simplicity of an OLS regression and could mean that even a simple machine learning model would better understand the data at hand and form percentiles portfolios with better performance.

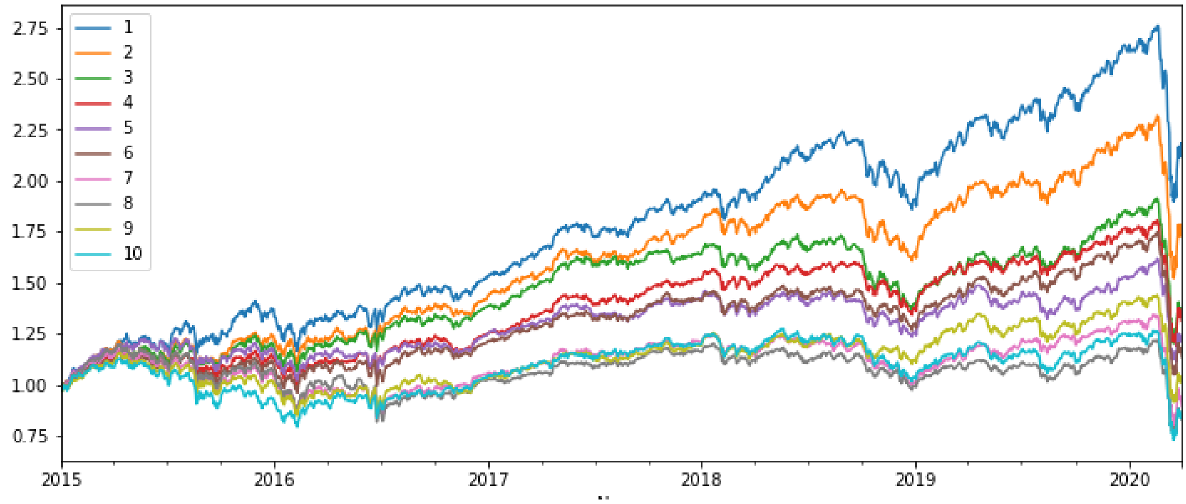


Figure 8. Cumulative returns of percentile portfolios (Regression) **as** formed by the return prediction using a regression model as seen in (1) on a four-year rolling window basis. March 2015 – July 2020.

The portfolios' returns as seen in Figure 8 do otherwise not differ considerably from the returns from the index in general as seen in Figure 3, although showing more limited returns for the worst performing portfolios. Furthermore, portfolios 3-10 all follow similar trends while the 1st and 2nd percentile portfolios show considerably higher return by the end of the timeseries. All portfolios follow similar overall market trends and are subject to similar systematic risks as seen in Figure 8 by decreases in late 2015, late 2018 and in 2020.

5.3. Classification Strategies

	k-NN				Gradient Boosting			
	F1-Score	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy
2016-04-01	0,54	0,57	0,58	0,59	0,55	0,57	0,58	0,58
2016-07-01	0,54	0,56	0,58	0,58	0,56	0,57	0,58	0,58
2016-10-01	0,50	0,59	0,57	0,58	0,53	0,58	0,58	0,58
2017-01-01	0,46	0,61	0,56	0,57	0,51	0,59	0,57	0,57
2017-04-01	0,50	0,59	0,56	0,58	0,50	0,59	0,56	0,56
2017-07-01	0,48	0,58	0,55	0,54	0,49	0,59	0,55	0,55
2017-10-01	0,48	0,55	0,53	0,55	0,45	0,59	0,53	0,53
2018-01-01	0,48	0,54	0,52	0,54	0,48	0,60	0,54	0,54
2018-04-01	0,51	0,60	0,56	0,55	0,52	0,59	0,56	0,56
2018-07-01	0,53	0,57	0,55	0,53	0,53	0,58	0,56	0,56
2018-10-01	0,54	0,55	0,54	0,53	0,52	0,59	0,55	0,55
2019-01-01	0,57	0,59	0,59	0,59	0,57	0,57	0,58	0,58
2019-04-01	0,53	0,56	0,54	0,58	0,54	0,59	0,55	0,55
2019-07-01	0,54	0,57	0,55	0,56	0,54	0,58	0,55	0,55
2019-10-01	0,53	0,55	0,53	0,54	0,51	0,54	0,51	0,51
2020-01-01	0,67	0,92	0,57	0,56	0,57	0,94	0,47	0,47

Table 4. kNN & Gradient Boosting model evaluation data for start of testing data between 2016 – 2020. The testing data includes all succeeding observations following the testing data start as indicated by the row.

Although the models differ in methodology the k-Nearest Neighbor and gradient boosting approaches show similar results across all evaluation metrics. The kNN classification approach sees marginally higher accuracy and f-score compared to the gradient boosting approach. The mean accuracy of kNN is 0.56 and the mean f-score is 0.53 which can be compared to the mean accuracy of 0.55 and mean f-score of 0.52 of the gradient boosting. Both models are as such slightly better than randomly allocating the return signal. The gradient boosting approach see somewhat higher precision for the rolling windows between 2017 and 2018 but is also characterized by lower f-scores during the period as compared to the kNN approach.

With ESG factors gaining a prominent voice in finance in recent years there could be a case for that the models would show a better precision the latter in the dataset the model was applied. Despite the U-shaped growth of the F1-Score, this evolution is not reflected in any of the evaluation metric for the kNN and Gradient Boosting model. Instead, the Gradient Boosting and k-Nearest Neighbor approach alike varies around a f-score of 0.6 for the whole timeseries with exception of the last rolling window, in which the precision increased considerably. However, this increase is not noticeably in the model’s accuracy.

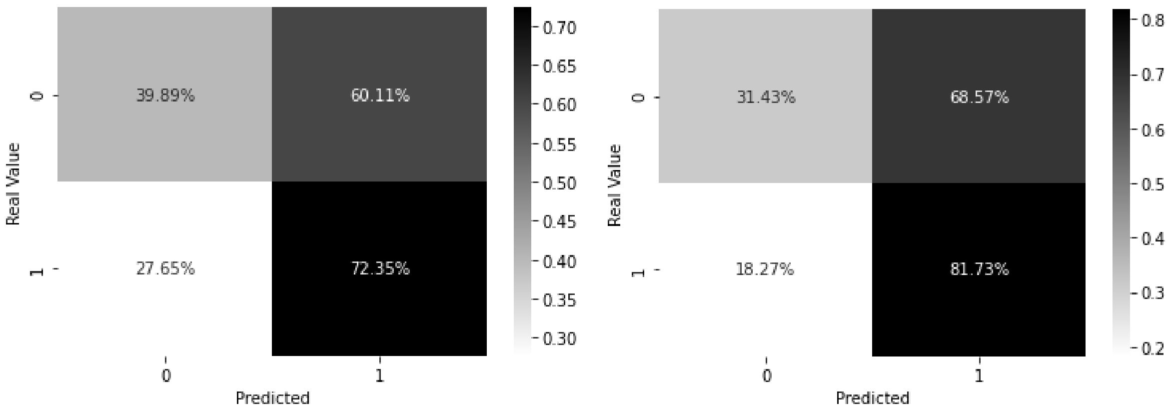


Figure 9. Mean confusion matrix of the k-Nearest Neighbors and Gradient Boosting. The kNN confusion matrix is shown on the left side and the Gradient Boosting on the right side. The values show the percentage of predictions in relative to the Real value.

From studying the mean confusion matrix of the k-nearest neighbors' approach one can gain a deeper insight into the models' tendencies in its prediction. In Figure 9 the kNN and the gradient boosting model to a large extent predict up signals as compared to down signal. There are relatively few instances in which the models predict a down signal when the real value is an up signal. Although the pattern is clear for both models, the Gradient Boosting classifier differs slightly from the kNN model in that it predicts up signals to an even larger extent. It managed to correctly predict 81.73% of up signals as compared to the k-Nearest Neighbors correctly predicting 72.35% of all up signals. Although the Gradient Boosting on average has a smaller accuracy rate, when considering Figure 9 it is possible to conclude that the Gradient Boosting model better understood the classification problem as opposed to the kNN model. This discrepancy could be caused by the kNN model's relative simplicity while still being optimized through the selection of the number of neighbors each iteration, while the Gradient Boosting model was tuned to standard values.

5.3.1. k-Nearest Neighbors Strategy

Portfolio	3 Factor Model Alpha	3 Factor Model T-Value	5 Factor Model Alpha	5 Factor Model T-Value	Sharpe Ratio
1 (low)	-0,04%	-0,87	-0,04%	-0,86	-0,20
2	-0,01%	-0,24	-0,01%	-0,23	-0,09
3	0,00%	0,11	0,00%	0,12	0,00
4	0,00%	0,01	0,00%	0,02	-0,01
5	-0,02%	-0,49	-0,02%	-0,48	-0,09
6	0,01%	0,28	0,01%	0,29	0,07
7	0,01%	0,49	0,01%	0,50	0,13
8	0,03%	0,87	0,03%	0,88	0,20
9	0,04%	1,39	0,04%	1,40	0,32
10 (high)	0,06%	2,06	0,06%	2,07	0,42

Table 5. Results per percentile portfolio implementing a kNN probability strategy to estimating percentile portfolios. The table reports alphas (in percentage) and t-values of the three-factor model and five-factor model. The annul Sharpe ratio of each portfolio is reported.

Using a k-Nearest Neighbor approach to forming percentile portfolios based on the probability of up signals show slightly more promising results as compared to the regression strategy. The approach successfully allocates the stocks to portfolios such that the portfolios are non-inversely ranked according to their returns. However, only the best ranked portfolio shows statistically significant alphas at the 5% level. An alpha of 0.06% of both the 3-factor model and the 5-factor model could furthermore be argued to be material and non-existent when applied in practice including transaction costs.

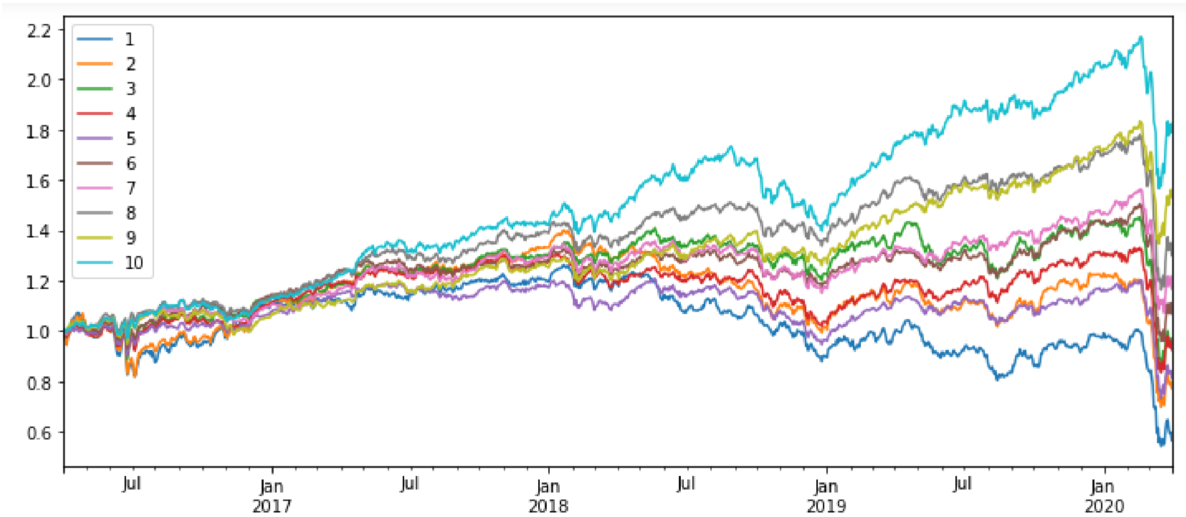


Figure 10. Cumulative returns of percentile portfolios (kNN) formed using the probability of a up signal as predicted by a k-Nearest Neighbors approach. March 2016 - March 2020

The 10th portfolio, similar to the regression approach, seemingly outperforms the other portfolios by a margin. However, the kNN approach differs from the regression in the sense that the worst performing portfolio also deviates from the other portfolios by performing visibly worse as seen in Figure 10. The 10th portfolio sees a Sharpe ratio of 0.42 and a 3 and 5 factor model alpha of 0.06% while the 1st portfolio sees a Sharpe ratio of -0.20 and a 3 and 5 factor alpha of -0.04%. The kNN does otherwise not differ significantly from the overall market trends noticed during the timeseries. All portfolios see a similar bearish trend at the end of 2018 and beginning of 2020.

5.3.2. Gradient Boosting Classifier Strategy

	3 Factor Model		**5 Factor Model**		
Portfolio	Alpha	T-Value	Alpha	T-Value	Sharpe Ratio
1 (low)	-0,02%	-0,43	-0,02%	-0,43	-0,12
2	-0,01%	-0,39	-0,01%	-0,38	-0,11
3	0,00%	0,03	0,00%	0,04	0,01
4	0,01%	0,28	0,01%	0,29	0,07
5	0,00%	0,07	0,00%	0,08	0,02
6	0,01%	0,47	0,01%	0,47	0,11
7	0,01%	0,32	0,01%	0,33	0,09
8	0,03%	1,05	0,03%	1,07	0,24
9	0,02%	0,86	0,03%	0,87	0,21
10 (high)	0,03%	0,91	0,03%	0,93	0,16

Table 6. Results per percentile portfolio implementing a Gradient Boosting Classification probability strategy to estimating percentile portfolios. The table reports alphas (in percentage) and t-values of the three-factor model and five-factor model. The annul Sharpe ratio of each portfolio is reported.

When implementing the Gradient Boosting classification approach to ranking and forming percentile portfolios based on their probability of up signals, the results are neither economically significant nor statistically significant. The highest ranked portfolio shows an insignificant alpha of 0.03% in both factor models and the lowest ranked portfolio show an insignificant alpha of -0.02% in both factor models. These results compare unfavorably when compared to the regression and k-nearest neighbors approaches alike. The poor performance are furthermore reflected in the Sharpe ratios, with the highest ranked portfolio showing a Sharpe ratio of 0.16 and the lowest ranked portfolio showing a Sharpe ratio of -0.12.

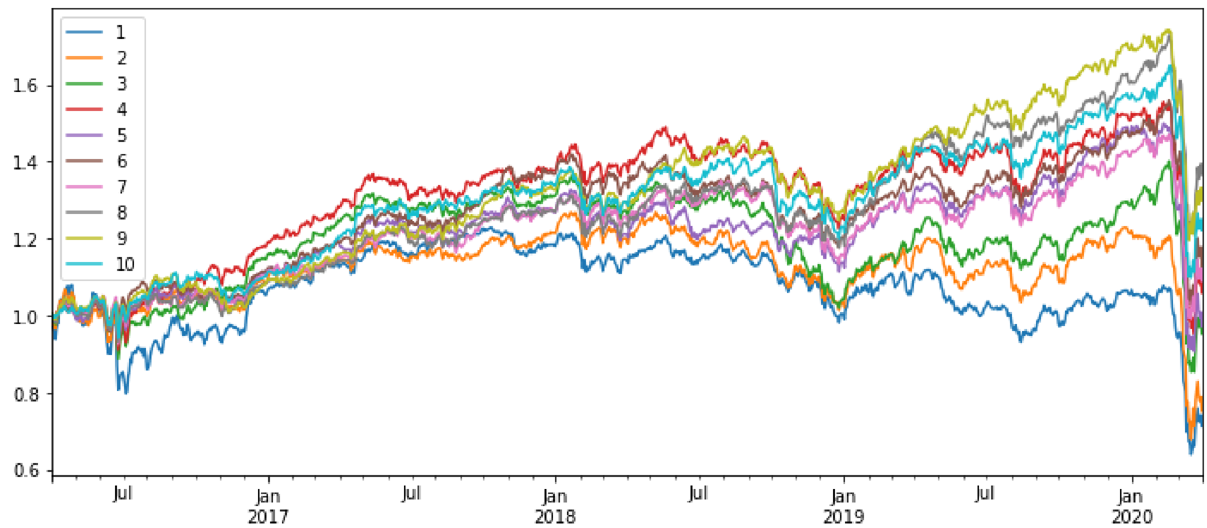


Figure 11. Cumulative returns of percentile portfolios (Gradient Boosting) formed using the probability of a up signal as predicted by the Gradient Boosting classifier approach. March 2016 – March 2020.

The Gradient Boosting approach does not only differ from the Regression and k-Nearest Neighbors approaches through the Sharpe ratio. The overall results of all percentile portfolios show relatively poor performance. The lowest ranked portfolios see similar results in terms of performance magnitude and volatility as the lowest ranked portfolios of the kNN approach. In fact, when considering Figure 11 all portfolios seemingly follow a similar trend as that of the kNN approach in Figure 10. The main difference lies in the highest ranked portfolios, in which the kNN clearly outperform the Gradient Boosting approach both in terms of Sharpe Ratio and cumulative returns. These comparably indecisive results could be derived from multiple factors. The portfolios are formed not by if the signal prediction is correct, but by the probability of the signal being positive in the coming quarter. As such, the similarity between the kNN and Gradient Boosting models in accuracy means that the difference in portfolio performance derives from the probability metric. The Gradient Boosting model is furthermore formed in a standardizes manner, with no dynamic adjustments of decision tree depth or learning rate in the timeseries, thereby being less agile as compared to the kNN approach where the elbow method is applied adjusting the number of neighbors dynamically in each iteration.

6. Conclusion

This paper aimed to examine the answers of 3 main research questions. Which emission indicators are especially relevant in ESG-integration? Are emission indicators viable to use in investment decisions from a risk-return perspective? Can a simple machine learning approach better the results of the investment strategy as compared to a traditional regression approach? – Through a feature selection and three different approaches to forming emission-integrated portfolios, this dissertation has provided insight into all three research questions, although to varying degrees.

The first research questions are explored and answered as part of the feature selection analysis. Although the dissertation starts with a large set of emission indicators, the feature selection shows a very limited set of indicators having a statistically significant effect on next quarter returns. Consistent with previous research finding negative effects of environmental scores on returns, the feature selection shows a negative effect of Internal Carbon Pricing, and a positive effect of NOx and SOx Emission Reduction Initiatives on next quarter returns. Internal Carbon Pricing is one of the more ambitious initiatives that companies can undertake to reduce emissions. It has a material, direct and measurable impact on the companies' operating costs, thereby being more likely to negatively affect returns. These results are in line with previous research that show the existence a carbon premium from the European Trading Scheme (ETS). (A.M. Oestreich & I. Tsiakas., 2015) Comparably, NOx and Sox Emission Reduction Initiatives is not necessarily an ambitious undertaking. It does not necessarily have a material impact on operating costs and a "Yes" does not necessarily mean that the costs of the initiative exceed the positive goodwill effects.

The regression approach provides similar results to previous research although they are not statistically significant in this paper. The portfolios are formed reversely, with the worst ranked portfolio (1^{st} portfolio) showing the best results and the best ranked portfolio (10^{th} portfolio) showing the worst results. This effect could be explained by the relatively large explanatory power of internal carbon pricing and that firms see increased investments and operating costs but little increased firm value from implementing emission initiatives.

The paper furthermore explores machine learning classification approaches to determining percentile portfolios. Through two different classification approaches in the form of a kNN algorithm and a Gradient Boosting algorithm, differences in how the ESG-integration strategy is implemented is discussed and analyzed.

The k-nearest neighbor classification approach to forming percentile portfolios show promising results when compared to a regression approach. The relative simplicity and equally weighted variable nature of the machine learning model could benefit the results. The accuracy consistently shows slightly better results than a random assignment and the percentile portfolios outperform the regression approach. However, when considering the classification metrics, the model shows a strong tendency to predict up signals across the whole time series. This could suggest that the approach fails to understand the subtle differences between the causes of up and down signals.

The Gradient Boosting approach does not outperform an OLS regression approach and shows slightly lower model accuracy as compared to the k-nearest neighbor approach. The comparably lacking returns when implemented in an investment strategy are likely influenced by several factors. The relatively small differences between the accuracy of the k-nearest neighbors and gradient boosting model when compared to the differences in significance and Sharpe ratio shows that the poor performance in terms of returns of the gradient boosting model lies in the probability metric. Although the gradient boosting model does not differ considerably from the kNN approach in terms of accuracy, in fact showing tendencies of understanding the machine learning problem better, the poor probability results does, in this research, not make it a more suitable investment strategy than a regression or kNN approach.

Altogether this paper has found that, from a large selection of emission indicators, Internal Carbon Pricing and NOx and SOx Emissions Reduction Initiatives are the most material indicators from a return perspective. Through three different approaches this paper concludes that a k-Nearest Neighbors classification approach can outperform the paper's regression approach to forming percentile portfolios.

This paper could be expanded upon in several ways. First, this research examines the viability of 2 supervised models on emission integration. This could be expanded to include a larger scope of both supervised and unsupervised models to further the insight into what models are viable when implemented into ESG integration. Second, non-emission features in the regression models could be expanded to reduce bias from omitted variables, and further exploration of how the feature variables affect the results of the machine learning models would deepen the insight into the models and features. Lastly, to gain a better insight into emission integration in an ESG-integration context, portfolios with emission-integration could be compared to similarly formed portfolio without emission-integration.

7. Appendix

Appendix 1. List of initially included emission indicator features

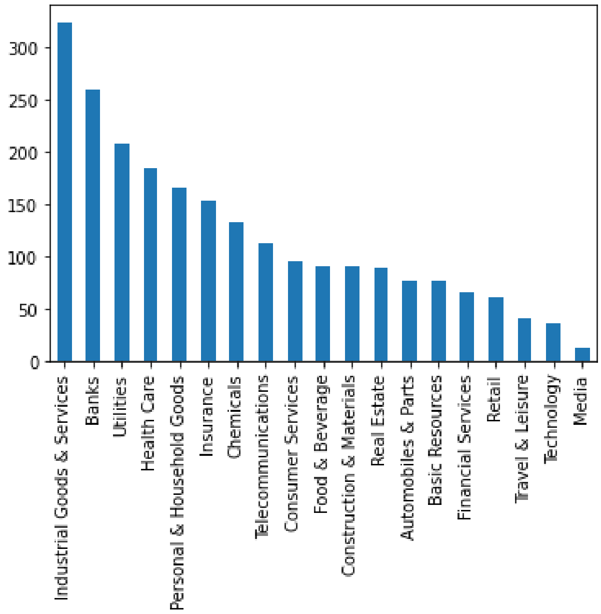
- Accidental Spills To Revenues USD in millions
- CO2 Equivalent Emissions Indirect, Scope 3 To Revenues USD in millions
- Emission Reduction Target Percentage
- Emission Reduction Target Year
- Internal Carbon Price per Tonne
- Internal Carbon Pricing
- NOx Emissions To Revenues USD in millions
- Ozone-Depleting Substances To Revenues USD in millions
- Policy Nuclear Safety
- SOx Emissions To Revenues USD in millions
- Self-Reported Environmental Fines To Revenues Local in millions
- VOC Emissions To Revenues USD in millions
- Biodiversity Impact Reduction
- CO2 Equivalent Emissions Direct, Scope 1
- CO2 Equivalent Emissions Indirect, Scope 2
- CO2 Equivalent Emissions Indirect, Scope 3
- Carbon Offsets/Credits
- Climate Change Commercial Risks Opportunities
- EMS Certified Percent
- Emissions Trading
- Environmental Expenditures Investments
- Environmental Investments Initiatives
- Environmental Partnerships
- Environmental Restoration Initiatives
- ISO 14000 or EMS
- NOx and SOx Emissions Reduction
- Particulate Matter Emissions Reduction
- Policy Emissions
- Staff Transportation Impact Reduction
- Targets Emissions
- Total Hazardous Waste To Revenues USD in millions
- Total Waste To Revenues USD in millions
- VOC Emissions Reduction
- VOC or Particulate Matter Emissions Reduction
- Waste Recycling Ratio
- Waste Reduction Initiatives
- Water Pollutant Emissions To Revenues USD in millions
- e-Waste Reductions

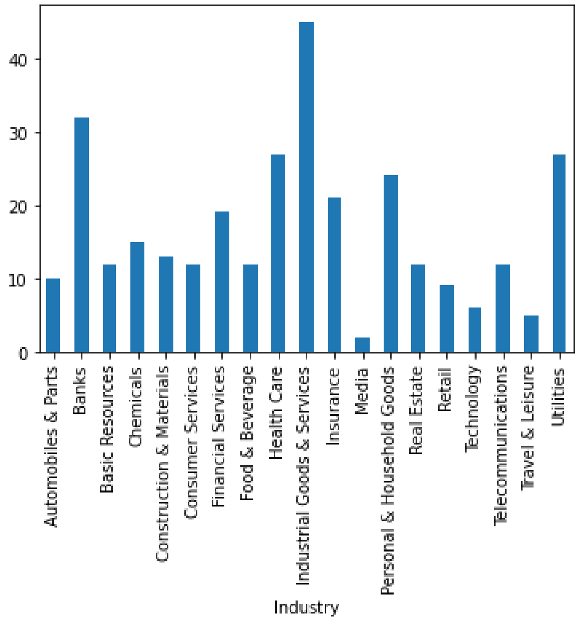
Industry	N	Emission Score							CO2 Equivalents Emission Total							Revenue						
		x̄	σ	Min	25%	50%	75%	Max	x̄	σ	Min	25%	50%	75%	Max	x̄	σ	Min	25%	50%	75%	Max
Automobiles & Parts	273	83,48	1,50	80,69	83,11	84,22	84,38	85,31	0,04	0,01	0,03	0,03	0,04	0,04	0,05	94521408	14341576	61425294	89155887	95972869	104338168	117564550
Banks	781	82,54	1,50	79,44	81,89	82,58	83,75	84,82	0,00	0,00	0,00	0,00	0,00	0,00	0,01	37782662	6023689	26823484	33044323	35826053	44187987	48257691
Basic Resources	300	89,73	2,28	86,30	87,50	90,24	91,77	93,25	0,43	0,05	0,35	0,38	0,44	0,47	0,60	25572804	7165948	14705205	18528053	27111181	33116723	36217936
Chemicals	438	75,25	3,42	70,06	72,89	74,43	77,00	82,39	0,29	0,08	0,17	0,20	0,32	0,36	0,40	19727396	3936419	13897736	16450252	18027789	23121582	27544413
Construction & Materials	296	82,13	3,13	76,81	81,28	81,95	83,37	88,47	0,54	0,28	0,17	0,25	0,60	0,80	0,94	21335230	3481066	16049538	18727371	20574468	22649586	29637370
Consumer Services	259	90,74	3,14	83,90	89,19	91,02	93,57	94,38	0,25	0,04	0,19	0,21	0,24	0,27	0,35	95932981	33876352	39838580	77354804	86107170	129874014	155448896
Financial Services	176	69,60	10,97	55,11	61,04	66,94	78,43	85,10	0,02	0,02	0,00	0,00	0,01	0,02	0,07	2048858	694801	974151	1460194	2079353	2397654	3383527
Food & Beverage	314	79,86	3,73	72,40	77,94	80,33	82,18	85,07	0,07	0,01	0,05	0,06	0,07	0,07	0,08	29423948	8457571	20100845	22579420	25749815	38441557	47455336
Health Care	650	73,93	3,55	66,52	71,41	75,46	76,60	78,34	0,03	0,01	0,02	0,03	0,03	0,04	0,05	15507913	3658371	10769979	12535595	14179527	18416248	23830530
Industrial Goods & Services	1075	78,57	2,40	74,78	75,93	79,35	79,98	83,72	0,05	0,01	0,03	0,04	0,04	0,06	0,07	17220934	2507662	11786912	15463778	15985308	19579768	21700735
Insurance	544	82,15	3,37	75,40	80,06	81,32	85,32	86,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	56881360	13078361	32671931	48106279	55518820	70130602	77082937
Personal & Household Goods	464	85,33	3,08	80,53	83,24	84,26	87,43	90,50	0,02	0,00	0,01	0,01	0,02	0,02	0,02	16792237	2163780	13266482	14995008	16905577	18555093	20429560
Retail	185	78,86	3,90	71,36	76,40	79,13	81,03	86,96	0,03	0,01	0,02	0,03	0,03	0,04	0,05	32327782	7980470	20881062	24770316	29840947	40493284	46708029
Technology	71	73,06	9,32	49,32	68,54	71,88	80,61	84,20	0,06	0,04	0,01	0,04	0,05	0,06	0,15	17218339	10625996	4995176	6321590	16632857	20558628	35971228
Telecommunications	409	78,71	4,07	72,83	75,62	76,84	81,32	87,58	0,03	0,00	0,02	0,03	0,03	0,03	0,03	26244783	2959118	20335573	24600138	25847792	28039254	33105213
Travel & Leisure	120	88,85	6,52	79,08	87,24	89,60	91,47	98,51	0,61	0,13	0,36	0,52	0,66	0,70	0,86	12677602	2463483	4130787	10928961	13041054	14284873	16061597
Utilities	694	86,46	1,56	84,11	85,39	86,58	86,97	89,59	0,79	0,15	0,48	0,71	0,79	0,87	1,09	49628480	19572989	18233522	35581677	39164507	69155616	85429429
Media	42	90,24	9,10	68,93	81,61	94,74	98,99	99,12	0,01	0,00	0,01	0,01	0,01	0,01	0,01	8462587	1007107	6933302	7547746	8800656	9217233	9995469
Real Estate	274	89,70	0,93	87,98	89,15	89,35	90,35	91,94	0,06	0,02	0,03	0,04	0,05	0,07	0,10	670432	108732	470284	603747	659611	755888	878203

Appendix 2. Descriptive Statistics

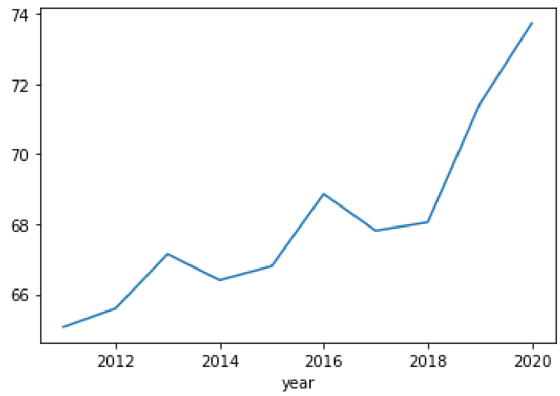
Industry	N	log(BM) x̄	σ	Min	25%	50%	75%	Max	log(MV) x̄	σ	Min	25%	50%	75%	Max	log(PE) x̄	σ	Min	25%	50%	75%	Max
Automobiles & Parts	273	6,64	0,25	6,20	6,46	6,65	6,81	7,17	9,99	0,29	9,11	9,76	10,02	10,22	10,49	2,27	0,34	1,55	2,07	2,27	2,43	2,99
Banks	781	7,16	0,21	6,88	6,98	7,11	7,29	7,78	10,21	0,23	9,52	10,07	10,21	10,40	10,53	2,69	0,37	1,78	2,40	2,63	3,05	3,38
Basic Resources	300	6,49	0,17	6,21	6,36	6,45	6,66	6,81	9,20	0,25	8,64	9,03	9,17	9,44	9,67	2,60	0,30	2,03	2,42	2,61	2,83	3,07
Chemicals	438	6,02	0,15	5,77	5,92	6,00	6,10	6,49	9,50	0,18	9,05	9,37	9,53	9,61	9,84	2,99	0,21	2,41	2,94	3,01	3,08	3,54
Construction & Materials	296	5,81	0,47	5,33	5,54	5,69	5,83	7,21	9,43	0,24	8,94	9,27	9,46	9,59	9,89	3,02	0,20	2,57	2,95	3,04	3,10	3,77
Consumer Services	259	6,20	0,36	5,37	5,99	6,17	6,50	6,78	10,33	0,23	9,66	10,25	10,32	10,50	10,66	2,87	0,48	1,81	2,47	2,89	3,19	3,78
Financial Services	176	6,31	0,30	5,80	6,18	6,27	6,44	6,90	8,79	0,27	8,14	8,61	8,83	9,00	9,14	2,57	0,25	2,29	2,39	2,52	2,71	3,14
Food & Beverage	314	5,72	0,26	5,38	5,54	5,59	5,96	6,34	9,97	0,43	9,46	9,62	9,83	10,45	10,79	3,06	0,20	2,66	2,93	3,07	3,15	3,60
Health Care	650	5,16	0,11	4,90	5,09	5,16	5,23	5,39	9,86	0,16	9,59	9,72	9,87	9,96	10,21	3,26	0,15	2,96	3,17	3,29	3,36	3,57
Industrial Goods & Services	1075	5,75	0,10	5,53	5,70	5,73	5,78	5,97	9,37	0,16	9,10	9,24	9,41	9,50	9,73	3,00	0,19	2,46	2,97	3,06	3,13	3,21
Insurance	544	6,54	0,19	6,23	6,40	6,46	6,68	6,96	9,86	0,15	9,48	9,78	9,87	9,98	10,20	2,55	0,16	2,19	2,42	2,59	2,67	2,88
Personal & Household Goods	464	5,45	0,27	4,89	5,48	5,53	5,60	6,00	9,87	0,14	9,64	9,74	9,88	9,99	10,11	3,02	0,15	2,61	2,91	3,04	3,12	3,31
Retail	185	5,85	0,19	5,48	5,73	5,85	5,96	6,24	8,99	0,23	8,32	8,92	9,04	9,15	9,35	2,89	0,22	2,35	2,73	2,87	3,06	3,24
Technology	71	6,02	0,37	5,32	5,72	6,15	6,30	6,57	9,92	0,40	9,11	9,61	10,01	10,23	10,68	3,15	0,38	2,48	2,98	3,15	3,30	4,49
Telecommunications	409	5,88	0,14	5,56	5,81	5,88	5,96	6,20	9,93	0,15	9,52	9,87	9,96	10,04	10,16	2,85	0,24	2,36	2,76	2,84	3,02	3,38
Travel & Leisure	120	6,15	0,43	5,29	5,94	6,14	6,33	6,88	9,15	0,30	8,39	8,88	9,24	9,38	9,51	2,77	0,37	1,96	2,60	2,76	3,02	3,66
Utilities	694	6,42	0,13	6,22	6,34	6,39	6,50	6,73	9,76	0,19	9,46	9,66	9,71	9,82	10,30	2,72	0,21	2,31	2,59	2,76	2,88	3,09
Media	42	4,71	0,29	4,35	4,48	4,67	4,83	5,38	9,92	0,23	9,50	9,78	9,87	9,94	10,35	3,01	0,15	2,70	2,90	3,02	3,11	3,27
Real Estate	274	6,91	0,10	6,74	6,82	6,91	6,99	7,17	8,55	0,15	8,31	8,44	8,58	8,63	8,86	2,68	0,23	2,26	2,57	2,65	2,91	3,05

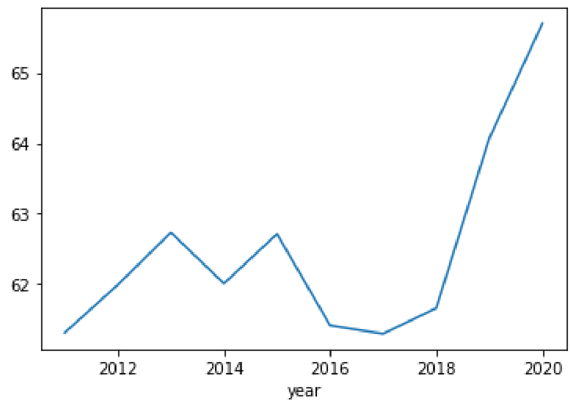
Appendix 2. Descriptive Statistics (continuation)





Appendix 3. Number of observations and companies per industry. Number of observations are viewed on the left and number of companies on the right.





Appendix 4. Average emission score and environmental score. Emission score is shown on the left and environmental score on the right.

		KNN Strategy							Gradient Boosting Classifier							Regression						
Industry	N	x̄	σ	Min	25%	50%	75%	Max	x̄	σ	Min	25%	50%	75%	Max	x̄	σ	Min	25%	50%	75%	Max
Automobiles & Parts	9	2	1	1	2	2	2	3	4	1	2	3	4	4	5	7	0	7	7	7	8	8
Banks	26	1	1	0	1	1	2	3	3	1	0	3	3	4	4	6	0	6	6	6	6	6
Basic Resources	10	5	1	3	4	4	6	6	4	1	3	4	4	5	6	4	0	4	4	4	5	5
Chemicals	14	5	0	4	5	5	6	6	5	1	4	5	5	5	7	4	0	4	4	4	4	5
Construction & Materials	11	6	0	6	6	6	6	7	5	1	4	5	5	6	7	5	0	5	5	5	5	5
Consumer Services	8	5	1	3	4	5	5	6	4	1	2	4	5	5	6	7	1	6	6	7	7	7
Financial Services	5	5	1	3	4	5	5	6	5	1	4	5	5	6	7	0	0	0	0	1	1	1
Food & Beverage	11	6	1	5	5	6	6	7	6	0	5	5	6	6	7	4	0	4	4	4	4	5
Health Care	23	7	0	6	6	7	7	7	5	1	4	5	5	5	6	3	0	3	3	3	3	3
Industrial Goods & Services	37	6	0	5	6	6	6	6	5	1	4	5	5	5	6	4	0	4	4	4	4	4
Insurance	19	3	0	2	3	3	3	3	5	1	4	4	5	5	6	6	1	5	6	6	6	7
Media	2	7	1	5	6	7	8	9	7	1	5	6	8	9	9	4	0	4	4	4	4	4
Personal & Household Goods	17	5	0	5	5	5	5	6	5	1	3	4	5	5	6	5	0	4	5	5	5	5
Real Estate	11	3	1	2	2	3	3	4	4	1	3	4	4	5	7	0	0	0	0	0	0	0
Retail	7	4	1	3	4	4	5	5	4	1	2	3	4	4	6	6	1	5	5	5	6	7
Technology	2	6	1	3	5	6	7	8	6	2	1	5	6	8	9	4	1	3	3	4	5	6
Telecommunications	12	4	1	3	4	4	5	6	4	1	2	3	4	4	6	5	0	5	5	5	5	5
Travel & Leisure	4	4	1	2	4	4	5	5	3	1	2	3	4	4	5	4	0	2	4	4	4	4
Utilities	22	5	0	4	4	5	5	6	5	0	4	4	5	5	5	5	0	4	5	5	5	5

Appendix 5. Descriptive statistics of appointed rankings per industry. The average number of firms, mean ranking, standard deviation, minimum value, maximum value and percentile values of each industry is presented for every implemented strategy.

8. References

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. International Journal of Business, Humanities and Technology, 3(3), 32-44.

Burkov, A. (2019). The hundred-page machine learning book (Vol. 1, pp. 3-5). Canada: Andriy Burkov.

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance. Strategic management journal, 35(1), 1-23.

Christensen, D. M., Serafeim, G., & Sikochi, S. (2021). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. The Accounting Review, https://doi.org/10.2308/TAR-2019-0506.

Coleman, J. S. (1988). Social capital in the creation of human capital. American journal of sociology, 94, S95-S120.

Dutta, A., Bouri, E., & Noor, M. H. (2018). Return and volatility linkages between CO2 emission and clean energy stock prices. Energy, 164, 803-810.

Dyck, A., Lins, K. V., Roth, L., & 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., & Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3), 1184-1216.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds (pp. 3-56). Journal of Financial Economics.

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

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

Giese, G., Lee, L. E., Melas, D., Nagy, Z., & Nishikawa, L. (2019). Foundations of ESG investing: How ESG affects equity valuation, risk, and performance. The Journal of Portfolio Management, 45(5), 69-83.

Gompers, P., Ishii, J., & Metrick, A. (2003). Corporate governance and equity prices. The quarterly journal of economics, 118(1), 107-156.

Gu, S., Kelly, B., & Xiu, D. (2018). Empirical asset pricing via machine learning (No. w25398). National bureau of economic research.

Halbritter, G., & Dorfleitner, G. (2015). The wages of social responsibility—where are they? A critical review of ESG investing. Review of Financial Economics, 26, 25-35.

Hart, O., & Zingales, L. (2017). Companies should maximize shareholder welfare not market value. ECGI-Finance Working Paper, (521).

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

Kotsantonis, S., & Serafeim, G. (2019). Four things no one will tell you about ESG data. Journal of Applied Corporate Finance, 31(2), 50-58.

Krokida, S. I., Lambertides, N., Savva, C. S., & Tsouknidis, D. A. (2020). The effects of oil price shocks on the prices of EU emission trading system and European stock returns. The European Journal of Finance, 26(1), 1-13.

Lanza, A., Bernardini, E., & Faiella, I. (2020). Mind the gap! Machine learning, ESG metrics and sustainable investment. Machine Learning, ESG Metrics and Sustainable Investment (June 26, 2020). Bank of Italy Occasional Paper, (561).

Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. Artificial Intelligence Review, 52(3).

Oestreich, A. M., & Tsiakas, I. (2015). Carbon emissions and stock returns: Evidence from the EU Emissions Trading Scheme. Journal of Banking & Finance, 58, 294-308.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442.

Stubbs, W., & Rogers, P. (2013). Lifting the veil on environment-social-governance rating methods. Social Responsibility Journal.

Trinks, P. J., & Scholtens, B. (2017). The opportunity cost of negative screening in socially responsible investing. Journal of Business Ethics, 140(2), 193-208.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(1), 1-20.