Norwegian School of Economics

Bergen, Spring 2023

Are seasoned equity offerings predictable?

Predicting future SEOs with machine learning algorithms

Peder Hernholm and Andreas Ore Wormsen

Supervisor: Walt Pohl

Master thesis in Economic and Business Administratioin

Major in Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

This master thesis explores the predictability of seasoned equity issuance in the United States using the machine learning methods based on logistic regression, decision-trees, random forest and XGBoost. In addition, we investigate the practical value of predicting seasoned equity offerings.

Our results show a benefit from employing machine learning for this purpose, with the best performing model (XGBoost) achieving an AUC of 0.72. The random forest model demonstrated similar capabilities with an AUC of 0.71, indicating that sophisticated non-linear models are suited for this type of prediction problem. Further, the impact of seasoned equity offerings on stock returns is analyzed to identify the possible benefits our models provide. Our efforts included two linear regressions using separate data samples, and one difference-in-differences estimation. These tests failed to provide conclusive evidence; however existing literature implies a negative effect on stock returns from seasoned equity offerings.

This thesis contributes to the extensive research conducted on the topic of seasoned equity offerings. While there are no directly comparable publications, we utilize existing literature to improve our thesis and to reflect on our findings. With this thesis we facilitate and encourage further research on this relatively unexplored area of seasoned equity offerings.

Acknowledgement

This thesis is the final project of our Master of Science degree in financial economics at NHH. The research question was intentionally developed to combine finance and machine learning, as we both are fascinated by this subject area. Working on this thesis for the past semester has allowed us to improve on an important and quickly developing area of finance, for which we are grateful.

Finally, we would like to thank our supervisor, Walt Pohl, for his quality guidance and commitment to the project. In addition, we would like to thank family and friends for supporting us in our academical endeavors.

Contents

1. Introduction

Seasoned equity offerings (SEOs) are a common way for firms to raise capital in the financial markets. The choice of a firm to conduct an SEO can be indicative of their prospects and has been shown to significantly impact shareholders and the perceived value of the company. Therefore, it is desirable for investors to identify the firms that are likely to conduct an SEO in the future. In this thesis we use applied machine learning methods, to improve investor's ability to assess the risk of an SEO occurring. There is existing research on this topic, however these efforts have mainly been concerned with predicting the choice between issuance of equity and debt. Our intention is to only analyze share issuance, but we still draw inspiration and insight from previous studies. In addition to developing machine learning algorithms, we investigate the economic implications these models can contribute toward. This is interesting to identify ways in which the machine learning models can provide value in practice. In this section we rely on existing literature that studies the relationship between SEOs and stock returns to design and interpret our experiments.

Despite the economic incentive for financial actors, there are few published efforts aimed at predicting SEOs. Our intuition is that modern machine learning techniques could produce significant prediction results and complement the existing body of research on the field of SEOs. Thus we try to answer the following research question: can machine learning models predict firms that will conduct an SEO in the future?.

In answering this question, we used recognized and applied machine learning models. The models selected were logistic regression, decision trees, random forest, and XGBoost. These were designed to make predictions on the likelihood of a firm to conduct an SEO, using a panel data set with information on US listed companies. Then each model was compared based on their prediction power using area under the ROC-curve (AUC) as the evaluation metric. The best performing model was XGBoost, followed by random forest. These had a respective AUC of 0.72 and 0.71.

To investigate the economic implications from using the prediction models, we analyze the relationship between stock returns and SEOs. Determining this dynamic can increase the value of our models. Through multiple linear regressions and a difference-in-differences (DiD) estimation we do not find conclusive evidence on the impact of SEOs on stock

returns. However, a large body of research indicates a negative impact of additional share issuance on perceived company value.

In summary our results produce encouraging evidence for the use of machine learning algorithms for this task. The models we create are beneficial in predicting the firms that will conduct an SEO, as our models greatly outperform random guessing. However, we encourage further studies on this issue and view our thesis as a fundament for improvements to be made. Related to the economic implications of our findings we are unable to produce compelling evidence. There is however existing evidence indicating a negative impact of SEOs on stock returns, thus we are comfortable in stating that our models can supplement financial risk assessment and decision making.

1.1 Litterature review

Seasoned equity offerings are an area that's well explored in corporate finance. Introducing machine learning to predict its occurrence, however, is unchartered territory. The existing literature on this topic is concerned with how an SEO impacts a firm and its stock value. As well as the underlying performance of a company that's issuing seasoned equity. Aspects of this literature are of interest for this thesis as it highlights the motivation behind predicting companies that will conduct an SEO. In the following section we review some literature that's interesting and relevant for our thesis on the topic of SEOs.

Masulis and Korwar (1986) published a study where they investigated the price reactions of a company's common stock around an SEO announcement. They conducted an event study analyzing stocks in a period where they announced a seasoned equity offering. Then, compared their returns prior to and following the announcement date with those of the general market. The study found that common stock has a negative reaction to an announcement of an additional equity offering. They also found that the negative effect varied in magnitude for different sectors, with industrials firms seeing a greater negative response than public utilities (Masulis & Korwar, 1986). Their conclusion has been strengthened by several studies identifying the same relationship. A study within the same area was conducted by Brav, Geczy & Gompers (2000). They found that underperformance after a SEO or IPO was most profound in smaller firms with lower book-to-market ratios (Brav, Geczy, & Gompers, 2000).

In 1982, Paul Marsh published the article “The Choice Between Equity and Debt: An Empirical Study”. Marsh analyzed determining factors in the choice between debt or equity issuance for companies in the UK from 1959 to 1974. The study found several important factors explaining preferred financing sources. Specifically, the results pointed at company size, bankruptcy risk, target debt level, market conditions and historical security prices as the most important factors impacting the choice. The study predicted the choice between the two financing methods using logistic regression on a sample of firms in need of long-term capital. Historic security prices were identified as the most important explanatory variable. The author also points out that companies behave as if they have target debt levels, and that this is an important factor when choosing funding source (Marsh, 1982). Although Marsh does not aim to predict the timing of issues, the study still provided us with valuable insights.

In a scientific article published by Eckbo and Masulis (1995), the authors are concerned with understanding the causes and effects of seasoned equity offerings. The article is a general exploration of the SEO topic, as well as an evaluation of relevant studies. Eckbo and Masulis reach interesting conclusions that are relevant for our thesis. They find that firms issue more equity compared to debt in expansive business cycles, and the opposite in periods of contraction. Another interesting observation made in the article is that SEOs often lead to a decrease in management percentage ownership of common stock. A reduction is seen as a negative sign for investors. (Eckbo & Masulis, 1995)

2. Theory

2.1 Sources of funding

Company financing through equity issuance is a relatively expensive and permanent capital source, as funding is raised through the sale of ownership stakes in the company (CFA Institute, 2023). The process in which firms raise equity capital is either through retained earnings, initial public offerings or seasoned equity offerings. Debt, however, is a cheaper and finite source of capital where the company lend funds from a counterpart with an obligation to repay it at some point, usually with interest. An important difference in these two capital sources is related to their associated risk. As an investment, debt is viewed as less risky than equity. That's because a company must honor its obligations to creditors, before attributing profits to the shareholders. Investors demand compensation for this risk, leading the cost of equity capital to be greater than that of debt (Damodaran, 2012, pp. 182-183).

2.2 Seasoned equity offering

Seasoned equity offerings (SEOs) refer to companies raising equity capital, having already conducted an initial public offering. As with every source of funding it is associated with both costs and benefits. Therefore, it is viewed as a strategic business decision. We will highlight some of the main implications of SEOs for a company.

First off, SEOs have a direct impact on investors. The issuance of new shares leads to an increase in the number of outstanding shares in the company. Thereby diluting the existing shareholders assuming they don't acquire new shares in the offering. In addition, studies find that the stock of publicly traded firms generally underperform after an SEO (Masulis & Korwar, 1986).

SEOs have associated costs that are important to understand and assess for companies. The main direct cost is the underwriter fee, claiming about 5 % of the proceeds raised (Berk & Demarzo, 2017, p.889). In addition, there are legal and transaction costs that occur in the process.

The market value of the company greatly impacts the effectiveness of an SEO. Myers-Majluf (1984) argues that SEOs indicate that the company stock is overvalued, if the manager acts

in the best interest of existing shareholders (Myers & Majluf, 1984). The reason being that an SEO becomes more effective as the market value of the firm increases in value. Conversely, it is negative for existing shareholders if a company conducts an SEO when the company stock is undervalued (Berk & Demarzo, 2017, p. 888). Therefore, it is logical to assume that a firm that favors an SEO to other sources of funding is at a minimum priced at a fair valuation. The study from Eckbo & Masulis (1995) strengthen this idea as they find an increase in equity financing relative to debt in expansive business cycles (Eckbo & Masulis, 1995). These periods are normally associated with inflated stock valuations.

2.3 Machine learning models

This section contains a summary of the different machine learning models used to create our prediction models. In addition to the models explained below we also create prediction models using logistic regression and random assignment. Whether a firm conducts an SEO is quantified as a binary variable which means we have a classification problem. This impacts the choice of models, as some are more suited to handle such problems.

2.3.1 Decision trees

Decision trees are a well-established and widely used statistical tool for predictive modeling and decision making. This supervised learning algorithm can be used in both regression and classification problems. In this thesis we will be exploring a classification problem, thus our focus is on classification trees. The models have a hierarchal structure consisting of a root node, branches, internal nodes, and leaf nodes (IBM, Decision Trees, u.d.), hence the name decision trees.

In decision trees the predictor space is stratified by features in the data (James, Witten, Hastie, & Tibshirani, 2022, pp. 303-315). The predictions are split based on how the training set distributes relative to each variable. To build a decision tree, the algorithm starts with the entire dataset and selects the most informative feature to split the data. This initial split corresponds to the “root node” in figure 1. The initial split creates two nodes, each containing a subset of the data (Internal Node in figure 1). The process of splitting is repeated until a stopping criterion is met. As an example, such a criterion could be that each leaf node must contain at least 10 observations.

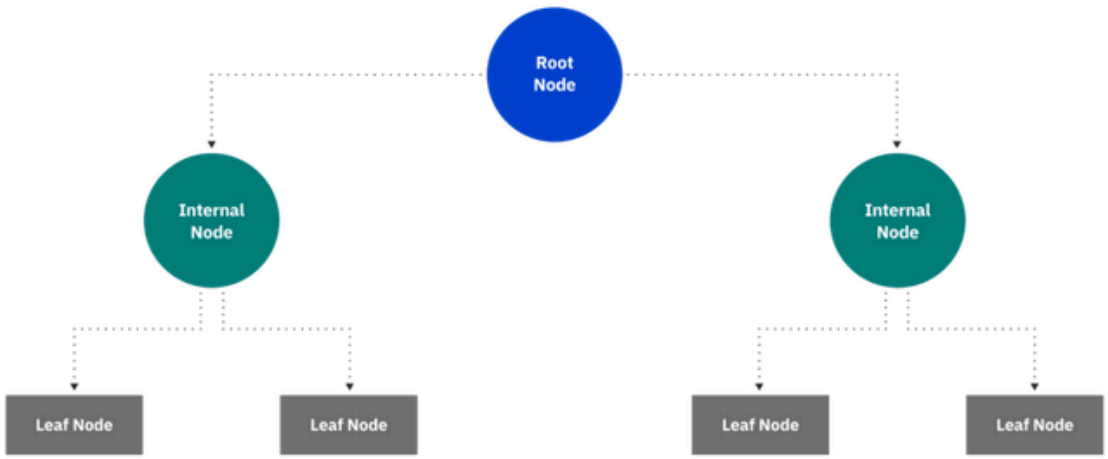

Figure 1: structure of a decision tree (IBM, u.d.)

Figure 2 is an illustration of a classification decision tree that tries to predict if a person has a heart decease. The initial split divides the data based on age. Then observations are split based on other features. In the leaf nodes the observations are classified as heart decease (0) and no heart decease (1).

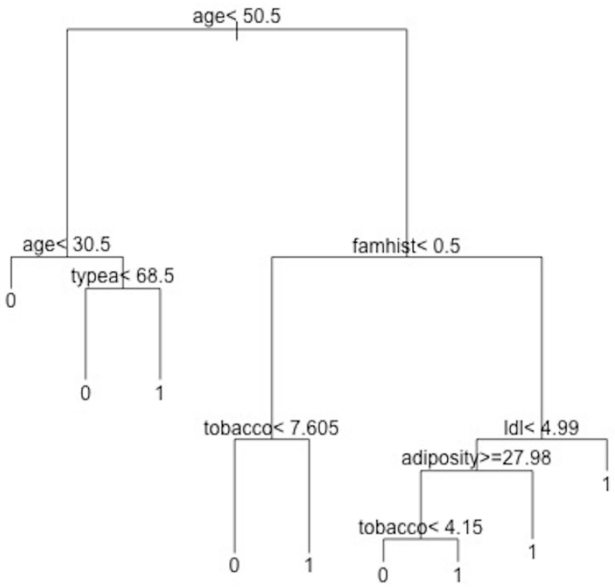

Figure 2: decision tree using data on heart decease.

2.3.2 Random forest

Random forest is a commonly used machine learning algorithm that combines multiple decision trees to reach a single prediction model, making it an ensemble learning algorithm. In contrast to regular decision trees this model can produce more general, accurate and robust results by decreasing variance. (IBM, u.d.)

Random forest improves on the concept of bagging, which is the process of fitting decision trees on a set number of generated samples based on the training data (James, Witten, Hastie, & Tibshirani, 2022, pp. 316-343). The method of creating multiple samples from one set of training data is called bootstrapping. Each generated tree is then merged into one averaged model, lowering the variance in the final model. Thus, each tree can be optimized for minimum bias letting variance increase, as the averaging of each tree reduces the variance. The improvements from such ensemble methods does come at a price, as interpretation of the model gets increasingly difficult (James, Witten, Hastie, & Tibshirani, 2022, pp. 316-343).

However, random forest deviates from regular bagging in that each decision tree being fitted on randomly sampled data is only allowed to consider a certain number of explanatory variables. The benefit is a decrease in correlation of the trees, leading to a greater variance reduction in the final model. Thus, avoiding that a strong predictor will end up in the top split (root node) every time, leading all trees to look similar. Averaging across many similar trees offers less benefit in terms of reducing variance. (James, Witten, Hastie, & Tibshirani, 2022, p. 320)

2.3.3 XGBoost

The final model we will discuss is Extreme Gradient Boosting (XGBoost), which is a popular library within supervised machine learning. XGBoost was released in 2014 by Tianqi Chen and has since become widely applied due to its powerful results (Chen & Guestrin, 2016).

The model builds on the idea of gradient boosting, which constructs additive regression models by sequentially fitting a simple parameterized function to current “pseudo”-residuals by least squares at each iteration (Friedman, 2002). The principal idea is fitting multiple trees on the residuals of an initial model. For each iteration the results are used to incrementally

decrease the residual of the entire model (Natelkin & Knoll, 2013). The improved prediction of each tree is multiplied by a learning rate to slow down the learning process and avoid overfitting. The iterations are concluded when the marginal residual reduction from additional iterating is negligible.

The XGBoost library is an optimized version of a gradient boosting model, providing improvements in some areas. It is more efficient than regular gradient boosting by allowing for parallel processing (Chen & He, 2017). The model is widely applicable on different data structures, and customizable. Most importantly it has proved highly effective and is currently considered one of the most powerful machine learning frameworks. Similar to random forest, XGBoost suffer drawbacks related to the interpretation of the model (James, Witten, Hastie, & Tibshirani, 2022, p. 343). However, they both facilitates the analysis of the most important variables in the prediction model, called variable importance. The variable importance is derived from the Gini index, illustrating the impact of each variable on the classifications (James, Witten, Hastie, & Tibshirani, 2022, p. 343).

2.3.4 Threshold value

Prediction models regarding a classification problem produce an output between 0 and 1. To classify these values a threshold is required. The default threshold value is 0.5, resulting in predictions below 0.5 being assigned to the class 0 (false), and those above being assigned to 1 (true). However, the threshold can be manually set to any value between 0 and 1. Modifying the threshold may be useful in increasing the accuracy of the model. It may also be beneficial if the nature of the prediction problem favor accuracy in one of the classification groups over the other. E.g., by setting a very large threshold, only instances where the model has high conviction are classified as true. This will however decrease the accuracy amongst the false predictions.

3. Data

The models in this thesis are created using a dataset consisting of 630 firms from united states in the period from 2010 to 2020 with total of 6930 observations. The firms included are firms that have either conducted an SEO or debt issuance in the observation period. The data set also consists of accounting-, stock- and macro data that we deem relevant in predicting SEOs.

3.1 Data description

The data on SEO's were collected from the SDC Platinum database. This source provides data on financial transactions such as equity offerings, debt offerings and M&As globally. The database contains observations from 1985 and is recognized as the industry standard for investment banking and deals data (libguides, u.d.). SDC Platinum also consist of identification variables as ticker symbol allowing us to merge with data from other sources.

The accounting data was extracted through the Compustat database, which consists of several financial variables for companies globally. Through this resource we can incorporate data as accounting measures (from income statement and balance sheet) and valuation metrics (financial ratios). (Wharton Research Data Services)

Using the CRSP database, we collected return and volume data for individual companies' stock and a respective industry classifier. Both Compustat and CRSP data were gathered from the Wharton research data services website. (Wharton Research Data services)

The macroeconomic data from FRED stems from various data sources. They were collected on a yearly basis. The variables we extracted from the FRED database were the effective federal funds rate, US gross domestic product (GDP) and a measure of both consumer and investor sentiment. (FRED)

3.2 Feature selection

Feature selection is an important part of creating a prediction model, and one must carefully consider the included variables to mitigate bias. Our method for feature selection started

with collecting many variables based on relationships reported in existing literature and our intuition.

Several of our variables were based on stock price changes and financial performance. This idea came from the study by Marsh which stated, “companies are heavily influenced by market conditions and the past history of security prices in choosing debt and equity.” (Marsh, 1982, p.23). To incorporate this in our models, we chose a variety of changes in stock prices, pe-ratios, and other financial performance metrics. The same study also mention that overall market performance did play a role in deciding whether to issue equity or debt, thus we included variables such as market index performance, fed-rate, GDP, VIX and consumer sentiment. Marsh also found that companies below long term or above short-term debt targets are more likely to issue debt (Marsh, 1982). Therefore, we included variables representing deviances in debt levels. The studies also mention size and asset composition as important predictors, which we incorporated in our model. The survey by Eckbo and Masulis previously mentioned in literature review also found similar findings as Marsh which supports the variable choices. (Eckbo & Masulis, 1995)

When selecting variables, we reduced the number of predictors based on two criteria. One being considerations to collinearity, leading us to remove some predictors that correlated too heavily with another predictor. However, we let some variables have high correlations if they served different purposes. One example is the pe-ratio and the change in pe-ratio. Where both provide respective insight, but still have high correlation. Variables with high correlation can lead to bias and complicate interpretation of the variables.

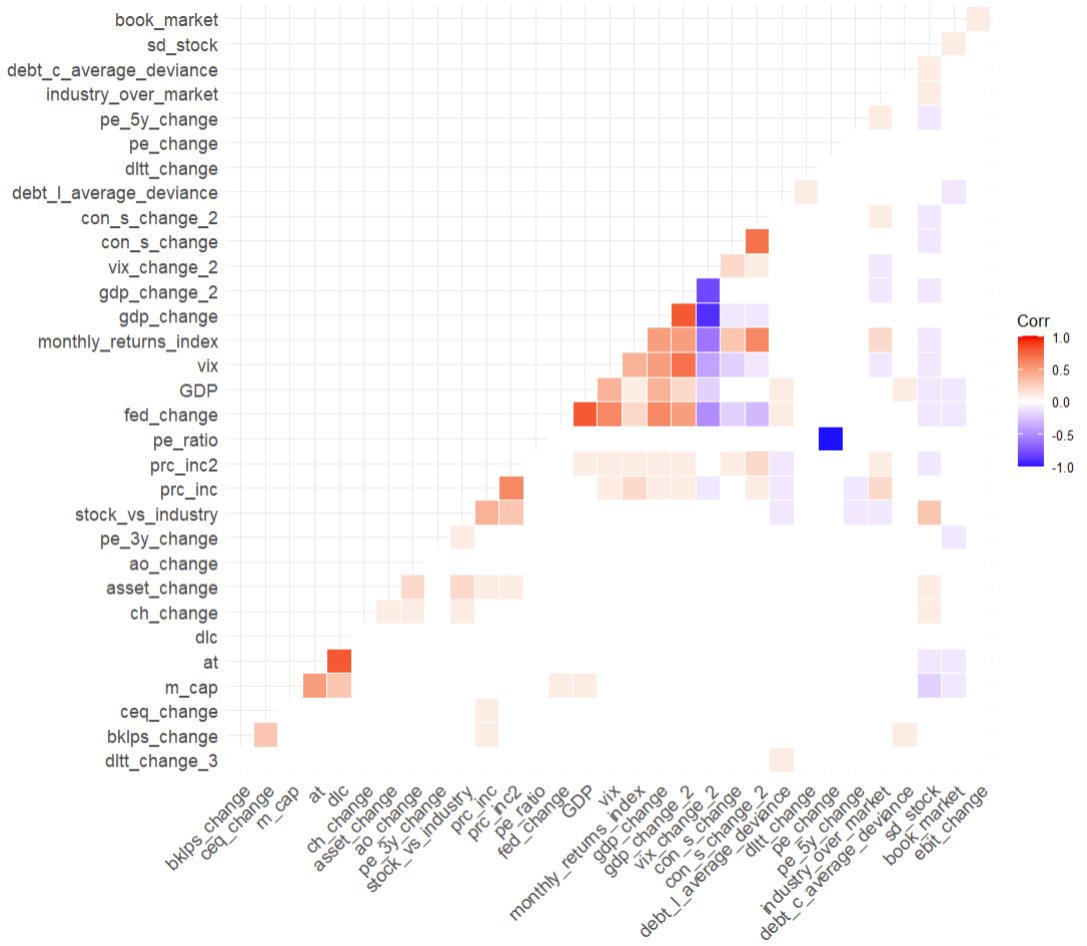

Figure 3: correlation matrix

The second criteria were related to improvement in model performance. This included grouping variables with similar properties together, to conduct a grouped backward stepwise selection. By training and testing our models on a validation set, holding out an entire group of predictors, we could analyze the benefit these types of variables added to our models. In addition, we investigated the number of predictors to include. This was done by comparing the performance of the models we used on three variable samples with varying sizes. One was trained using 10 variables, another on 22 and a third on 33 (table A.2.2), and tested on the validation set. The models using the highest number of variables outperformed the others. Therefore, we continued with 33 predictors for all our models except the logistic model which had a better performance with 22 variables. Below we illustrate the ROC-curve for the three different XGBoost models tested on the validation set.

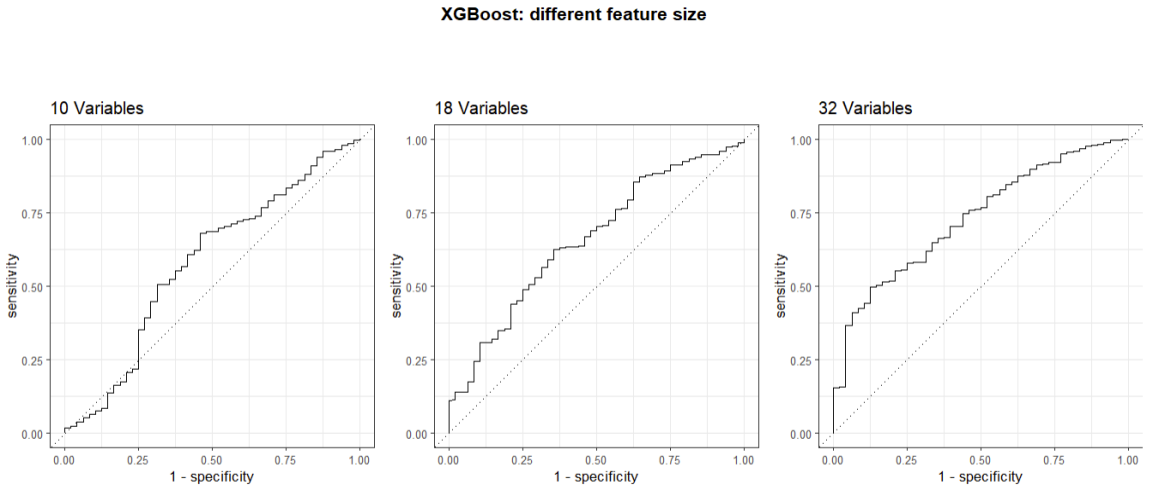

Figure 4: ROC-curves for XGBoost model on different subsamples of predictors

3.3 Variable explenation

In the table below (table 1) we explain all variables present in this analysis. The dependent variable is called “shares” and is a dummy indicating whether a firm issues equity in the upcoming year. There was a total of 777 share issues in our data. With 11.65% of observations as issues in training data and 9.12% in the test data.

VARIABLE	DESCRIPTION
pe_ratio	Common shares outstanding * price closed year / Net income. PE represents price over earnings for the companies.
pe_change	Changes in P/E ratio from last year
pe_5y_change	Changes in P/E ratio from 5 years ago to now.
pe_3y_change	Changes in P/E ratio from 3 years ago to now
prc_inc	Percentage increasing in stock price
prc_inc2	Percentage increase in stock price from 2 years ago until now
debt_l_average_deviance	Long term debt / total assets: Difference between today's ratio and the average ratio of past 5 years
debt_c_average_deviance	Same as above, only for current debt.
GDP	Gross Domestic Product
GDP_change	Percent change in GDP past year.
GDP_change_2	Percent change in GDP past 2 years.
con_s_change	Change in consumer sentiments index past year. (data from FRED)
con_s_change_2	Same as above, but from past 2 years.
sd_stock	Standard deviation of company stock returns
monthly_return_index	A value weighted monthly return index of all the stocks in the CRSP data set
industry_over_market	Used CRSP to obtain industry returns. Shows excess returns in industry.
fed_change	Change in fed funds rate in current year
asset_change	Yearly change in company asset value
at	Total assets
BKVLPS_change	Book Value per share yearly change
ch_change	Change in cash balance
ceq_change	Change in common equity
EBIT_change	Change in EBIT
vix	The VIX index
vix_change_2	Change in VIX index last 2 years
dltt_change	Change in total long-term debt
dltt_change_3	Change in long term debt last 3 years
stock_vs_industry	Excess return from a single stock based on the industry average return
m_cap	Market capitalization
book_to_market	Book value divided by market capitalization

Table 1: Independent variable names and description

4. Methodology

All the analyses in this thesis were conducted using the statistical programming language R (version 4.2.2), in the developing environment: RStudio. Several libraries were utilized throughout the project, with heavy reliance on the “tidymodels” framework.

4.1 Data collection, cleaning and processing

The database: SDC platinum, was used to retrieve the companies that issued additional equity (SEO), as well as the period in which the event occurred. We chose to only include companies whose stocks are noted on US exchanges and collected observations from 2011 until 2021. These decisions are based on a concern with the availability of data, as our models rely on relevant, detailed, and abundant data.

The identification variable: ticker symbol, was used to match data from SDC Platinum with data from Compustat, FRED and CRSP. Then we merged the chosen variables into one panel data set of yearly observations, requiring some variables to be annualized. The monthly return data from CRSP, was averaged within each year. To quantify the occurrence of an SEO, we created a dummy variable taking the value 1 in the year of a company’s SEO, and 0 otherwise. This variable was then modified to match with the prior year of the predictor variables. The reason being that we predict whether a company will issue additional equity next year. Thus, we do not possess data from the year of the SEO at the time of the prediction.

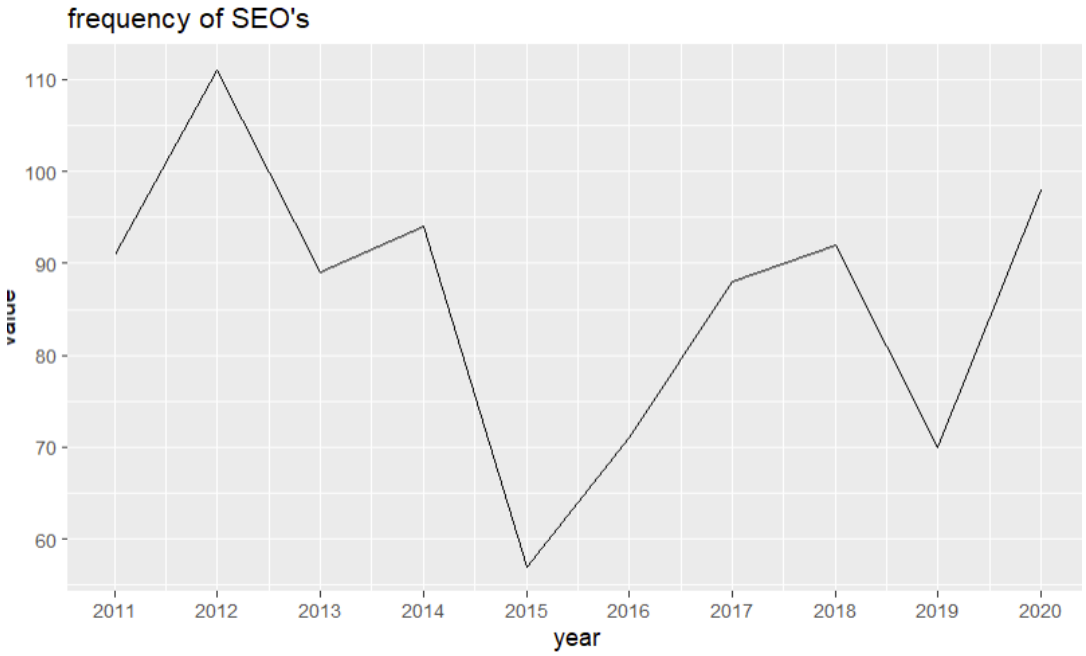

Figure 5: frequency of SEOs in our data across the sample period

It was necessary to conduct some cleaning of the data to continue with the analysis. There were several companies without matching tickers across the data sources, which were removed. In addition, firms with less than 15 years of data were excluded. This period is greater than the sample period, to ensure the availability of lagged variables. The final data set contained a total of 7,560 observations, consisting of 630 unique listed companies and 777 cases of SEOs.

In creating a robust model that mitigates the risk of overfitting it is necessary to operate with training and test data sets. This important idea allows the model to be fitted on one data set and tested on a separate. Testing the model on an unseen set of data improves the generalizability of the results (James, Witten, Hastie, & Tibshirani, 2022, p. 30).

For this task we split the data into three sub samples. This is due to aspects related to our approach and the problem making it desirable. An additional split allows us to modify and specialize the model parameters and features, through performance evaluation on one test set. The final test set has been held completely out of the training process and can be used to evaluate the optimized models. This method was chosen to increase prediction power of the models while mitigating risks related to overfitting.

The data was split into the following periods: (training set) 2010-2017, (validation set) 2018 and (test set) 2019-2020. This is a somewhat rough way to split the data, however necessary

due to the annual frequency of observations. The distribution of observations across the samples is approximately: (training set) 70 %, (validation set) 10 % and (test set) 20 %. The training set is large to ensure the model is fed enough observations. The validation set is smaller than the test set because we prioritized a large test sample.

4.2 Benchmark model

The performance of a prediction model is relative and depends on the phenomenon one is trying to capture and how the results are evaluated. There are problems that are easily predictable, and a strong model would then require a high performance. In other cases, it could be considered a success to get a few correct predictions. In our case there is a lack of comparable studies making it difficult to objectively determine the performance of a prediction model. To handle this, we designed a simplistic model to establish a performance benchmark. This was done by calculating the probability that an observation in the training set contained an issue of shares which was equal to 11.65 %. The model randomly predicts SEOs with the probability found in the training data, yielding results that could be expected from random guessing.

4.3 Tuning of models

The validation set explained in section 4.1. was used to draw preliminary insight from the models and identify an optimal feature sample. That is a type of model tuning but will not be discussed further in this section. There is another powerful method to improve the predictions in random forest and XGBoost models, which is to tune the hyperparameters. The two models include several hyperparameters, whose value influences the prediction results. In this report we use k-fold cross validation to optimize these hyperparameters based on the area under the ROC-curve metric. The cross-validation is conducted using 6 folds on the training data set. The grids we used for the tuning contained 450 unique combinations of hyperparameter and are presented in table A.1.1. and A.1.4.

In the random forest model, the hyperparameters we tuned were the size of feature sample included in each tree, number of trees in the forest and minimum size of nodes. The size of the feature sample decides the number of variables that are randomly selected from the entire variable sample at each iteration. The optimal size of the predictor sample was 5. The

number of trees is the number of decision trees that are fitted to make up the model. Increasing the number of trees makes the model more computationally demanding, and the marginal effects of increasing the number of trees generally taper off at high values. For this reason, we ran a cross validation to explore the range of trees we would use in our grid (figure A.1.3). From this exploratory analysis we chose the amount of trees to range from 1 to 500. The best performing model used 230 trees. The hyperparameter that controls the minimum number of observations in a leaf node regulates the stopping point for each generated tree. In our model a node must contain a minimum of 23 observations. This means that nodes approaching this limit of observations will not develop further branches.

In the XGBoost model the hyperparameters that were tuned were the number of trees generated, the depth of each tree, the minimum size of nodes, the learning rate, and the loss reduction. Tree depth relates to the complexity at each iteration by determining a maximum of node levels in each decision tree. This parameter was set to 10, as it was the optimal value. The learning rate was discussed in section 2.7.4. and decides how much the final model learns from each iterated tree. The optimal rate in our case was 0.014. Loss reduction is a parameter that tries to minimize the difference between the predictions and the actual values. The optimal value we found was 0.0534. The remaining hyperparameters have the same purpose as in the random forest model discussed above. The number of trees was tuned as for random forest with a range from 1:500 (figure A.1.4) and the results was 478 trees. The minimum size of nodes was 35.

4.4 Optimal threshold

In the selection of an optimal threshold, we chose to maximize the evaluation metric precision. The precision of our model is a measure illustrating the percentage of positive SEO predictions that are correct.

$$Precision = \frac{TP}{TP + FP}$$

Equation 1: Calculating model precision (TP = True Positive, FP = False Positive)

The rationale being that our main interest is to identify firms that perform an SEO, as this can benefit the economic and practical usability of our findings. One of the targets of this thesis to offer insights for investors on the risks of a company performing an SEO. Therefore, it is of importance that predictions stating that an SEO will occur are precise. Besides, most of our observations do not represent an SEO, thus never predicting an SEO (threshold = 1) would yield an "accurate" model due to the unbalanced data. It would however not provide any practical or theoretical benefit. For these reasons we chose to prioritize high conviction in the positive predictions.

To further guide our choice of threshold due to several thresholds which gave the same precision, we incorporated recall as a secondary target to maximize. Recall score is a measure of how many of the actual positive instances that the model captures. Our reason for using recall was to favor a threshold that facilitate a larger amount of correct positive SEO predictions.

$$Recall = \frac{TP}{TP + FN}$$

Equation 2: Calculating model recall (TP = True Positive, FN = False Negative)

4.5 Evaluation

The evaluation of the prediction models is an important process where we review their performance and reliability. It is conducted by extracting an evaluation metric from the test results of the models. There are several metrics that can be used, depending on the nature of the problem. The fact that we have a classification problem excludes some. Our models will be evaluated using the area under the ROC-curve (AUC) and precision. We also use confusion matrices as an illustrative tool of performance.

The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various thresholds (Hoo, Candlish, & Teare, 2017). It illustrates the tradeoff between correctly predicted positive instances and falsely predicted positive instances, for changing threshold values. The ROC curve is thus concerned with positive predictions, which in our

case is the prediction of an SEO occurring. In general, a high performing model has a high TPR and a low FPR.

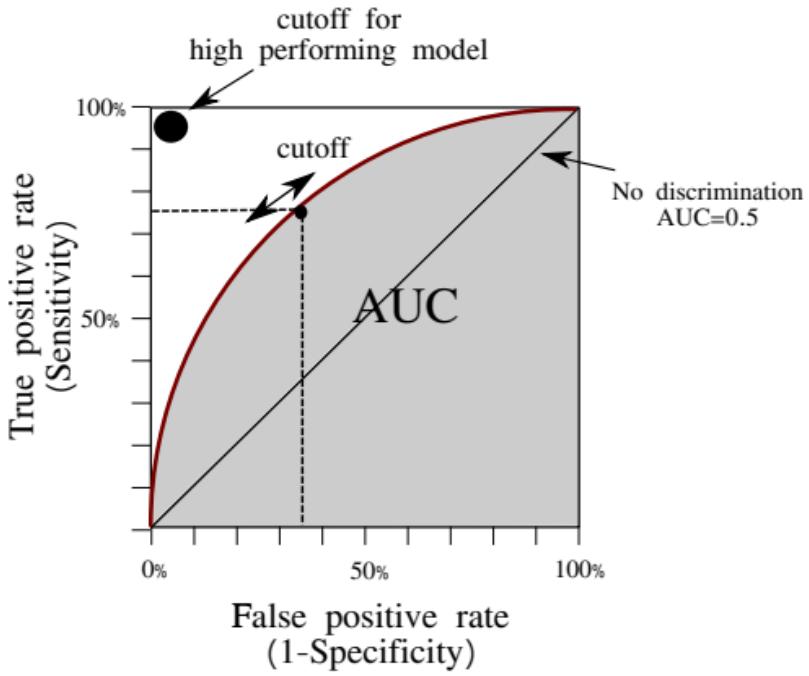

Figure 6: ROC curve (Rodriguez-Hernandez, Pruneda, & Rodriguez-Diaz, 2021)

The AUC is a measure of the general performance of the model across all possible thresholds. It illustrates the ability of a model to identify whether a specific condition is present or not (Hui Hoo, Candlish, & Teare, 2017). A model with an AUC value of 0.5 indicates that the model is no better at correctly predicting outcomes than random guessing. A value of 1 indicates a perfect model with 100% correct predictions. The AUC measure was used to compare model performance. In the continuing analysis we only continued with the best model.

After using the method explained in section 4.4 to select a threshold, we constructed a confusion matrix to further evaluate the best performing model. A confusion matrix compares the prediction results with the actual outcomes, illustrated in figure 7. From the confusion matrix the model accuracy could be calculated. This measure states the percentage of classifications that was predicted correctly.

	Predicted Negative	Predicted Positive
Actually Negative	True Negative (TN)	False Positive (FP)
Actually Positive	False Negative (FN)	True Positive (TP)

Figure 7: confusion matrix

$$Accuracy = \frac{TP + TN}{Total\ amount\ of\ predicitons}$$

Equation 3: Calculating model accuracy (TP = True Positive, TN = True Negative)

Then we proceed to test the robustness of our results, with the intention of strengthening the external validity. The intention of the tests is to control whether the model produces similar results when it is trained and tested on different sample periods. Therefore, we trained the best performing model on two new periods. One was trained on data from 2010 to 2014 and tested on data from 2015. The other was trained on data from 2016 to 2019 and tested on data from 2020. The AUC values obtained from the test results could be compared to those of the original model.

In addition, we conducted a separate robustness test using a k-fold cross validation with 8 folds. This method leaves one-fold out and trains a model on the remaining data. The fold that's held out is used as a test set. This method is repeated 8 times holding out a unique fold at each iteration. The output yields eight sets of test results, with unique AUC values that can be compared to our original model.

The test models discussed above should produce similar AUC values as seen in our main model, to strengthen the robustness of the results and validate our results. These tests evaluate the reliability and reproducibility of our best performing prediction model.

4.6 Implications of results

After creating the prediction models and reviewing the results we were still interested in the practical implications of the findings in our thesis. Investigating this further is important to determine the outcome from practical application of the model and to identify whether it provides any value. The thesis was designed with the assumption that an SEO has a negative impact on the value of a company. A notion found in existing literature. The applicability of our results rests on this relationship holding true, thus we are interested in investigating this further. Using the same data set as for the prediction models we try to identify the impact of SEOs on company value.

To conduct this analysis, we use linear OLS regression. The dependent variable is returns in the year of the SEO, and the independent variable is whether a firm conducted an SEO in that respective year or not. The coefficients related to the SEO variable express the impact of an SEO on firm value. Additionally, we created models using returns in the following year as the dependent variable. The reason being that stock returns in the year of the SEO are impacted by returns prior to the event, while the stock returns in the following year must be after the SEO. It is also interesting to investigate the development in returns over time. For both models we conducted two separate regressions adding controls for industry specific returns, book value and size of firm, as these variables are highlighted in the literature and discussed in section 1.1. (Masulis & Korwar, 1986) (Brav, Geczy, & Gompers, 2000). In addition, we included the standard deviation of the stock returns, as we thought general variation in returns might be an important factor affecting the response from an SEO.

There is a possibility that there are underlying characteristics besides share issuance amongst the companies conducting SEOs that impact firm value. This notion is discussed in existing publications e.g. by Carlson, Fisher and Giammarino (2010), and would create a biased model (Carlson, Fisher, & Giammarino, 2010). To account for this, we analyze a subsample of similar observations based on the feature sample included in the prediction models. The sub sample only consists of the observations predicted to conduct an SEO, due to the assumption that these firms possess similar characteristics. Again, we use linear OLS regression with returns of the company's stock in the year of the prediction as the dependent variable. The independent variable is whether a firm has conducted an SEO or not, given that the model predicted that they would. This method is an attempt to better isolate the effect of the SEO on stock returns. We expect to see a decrease in market value of the firms that are

predicted to conduct an SEO and do it, compared to the firms predicted to do an SEO that don't do it. Similarly, to the previous regression we created four models, only changing the data sample.

In a separate analysis to establish the relationship between SEOs and market value we conducted a difference-in-differences (DiD) regression. The test creates a quasi-experiment structure allowing us to compare the stock price response in two groups of firms (The World Bank, u.d.). In our case we compare firms that do an SEO with those who don't. The validity of a DiD estimation is threatened by the possibility that underlying factors in the selected groups, other than the variables studied, explain the differences observed. We control that the two groups possess similar traits and are expected to perform similar if neither were to issue shares (parallel trend assumption) (Abadie, 2005). The estimation is conducted on the subsample of firms that were predicted to issue shares, because our best performing prediction model has deemed these similar based on 33 variables. With respect to the parallel trend assumption, the two groups are analyzed further to uncover potential differences. Then we conduct the DiD estimation and reflect on the results relative to the two previous regressions.

5. Results

In this section we present the results from the prediction models, using the methods discussed in section 4. The first part is a performance evaluation of each prediction model where the XGBoost had the best performance. Subsequently we present the results of our regression analyses which build on the prediction model in an effort to gain economic insights.

5.1 Benchmark model

The purpose of this model is to provide a benchmark for the other models. The result was as expected, achieving an AUC of 0.5. Out of 115 cases of SEOs, the model correctly predicted 9 and produced 131 false positives, yielding a precision of 6.4%.

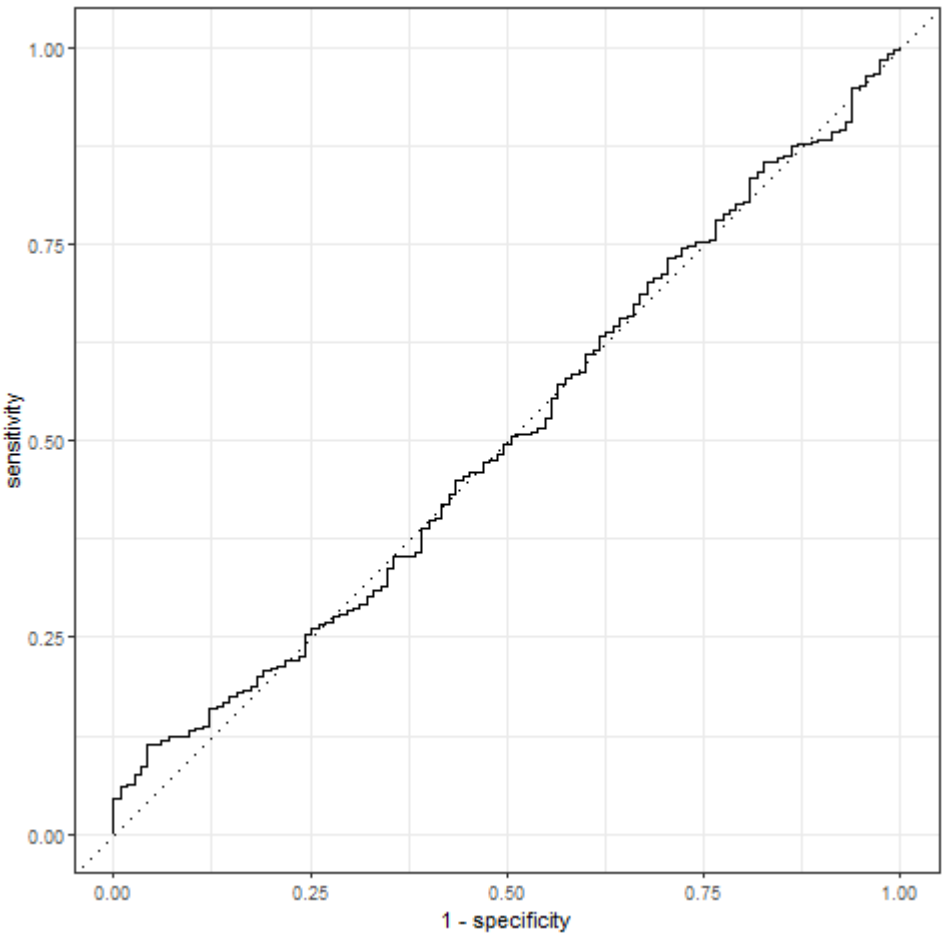

Figure 8: ROC curve for decision tree

5.2 Decision tree

The decision tree model showcased a poor ability to predict SEOs. With an AUC value of 0.56, it performed slightly better than the benchmark model. This model was expected to be one of the weaker ones, due to its simplicity. However, more sophisticated decision tree models are explored next.

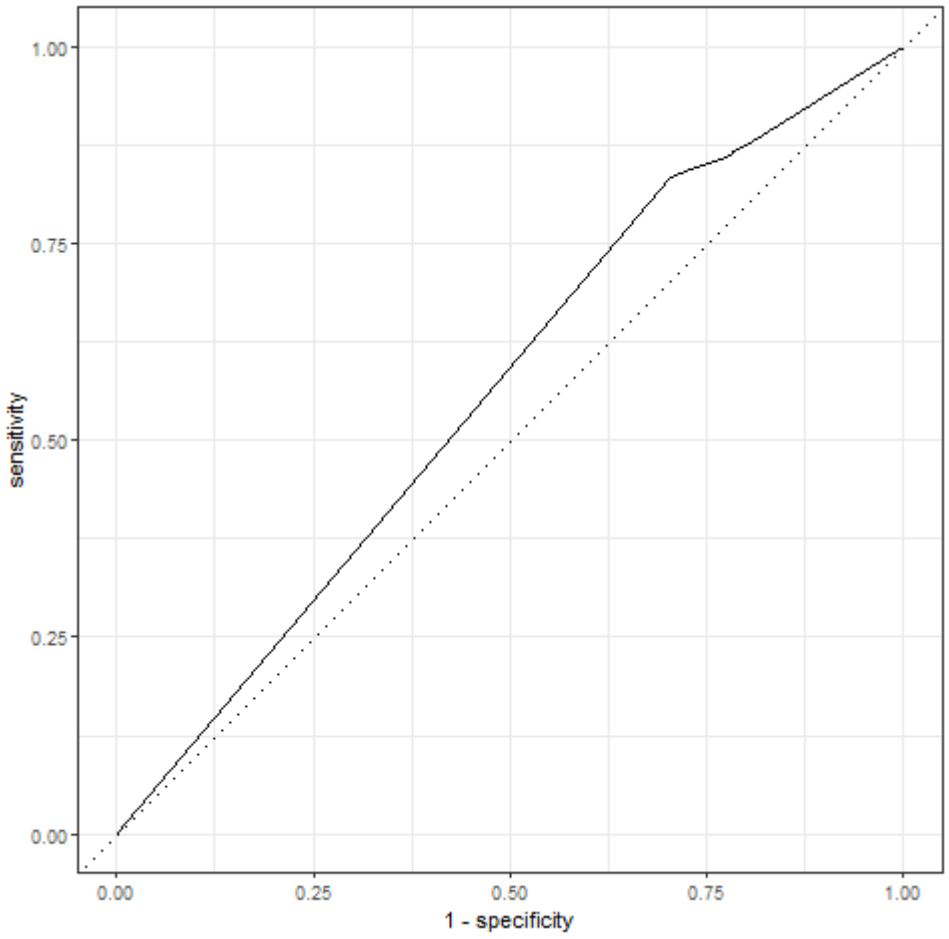

Figure 9: ROC curve for the decision tree

5.3 Logistic regression

The prediction model using logistic regression performs better than the decision tree with an AUC of 0.63. This is a significant increase in performance compared to both previous models. The logistic model is clearly better than the benchmark model, indicating some prediction ability. However, like the decision tree this is a relatively simple model using only one iteration in the learning process. In addition, logistic regression has a linear relationship.

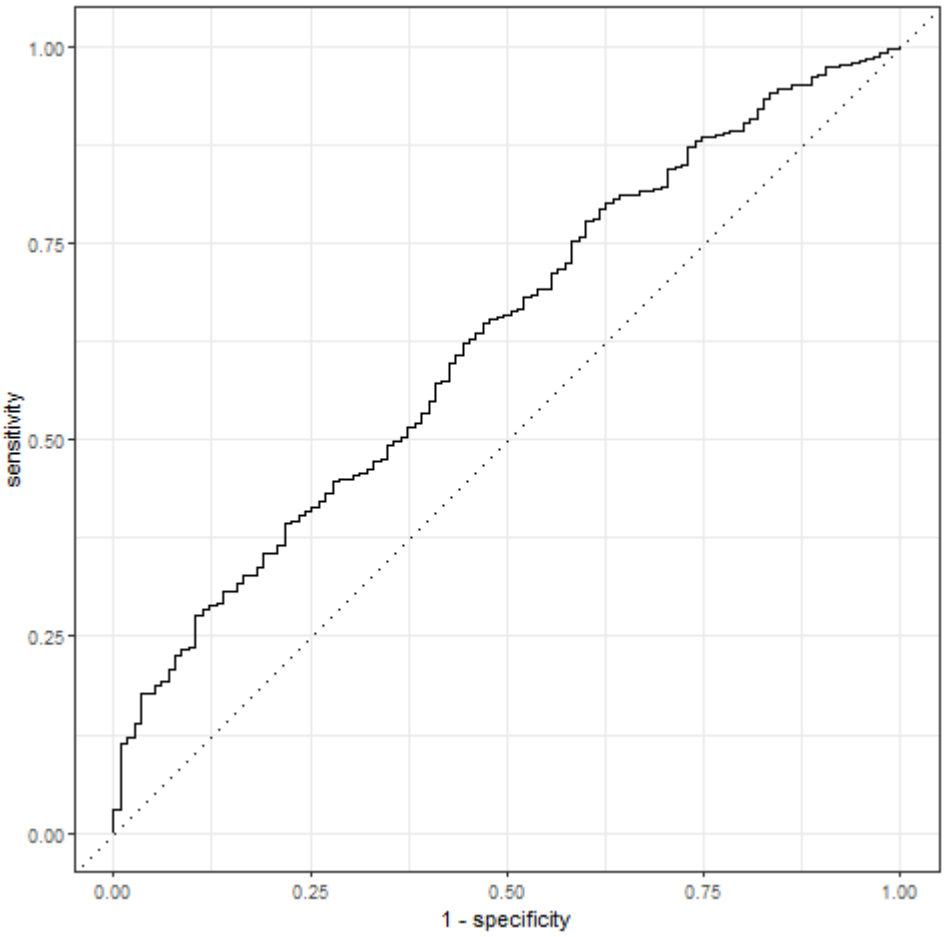

Figure 10: ROC curve for the logistic model

5.4 Random forest

The Random Forest model is one of the more sophisticated models we used, and it performed second best on our test data. The AUC of the model equaled 0.71, which is a significant improvement compared to the logistic regression. The performance of this model

depends on the choice of hyperparameters, which we explain in section 4.3 (figure A.1.2). Improvements compared to the previously presented model results were expected.

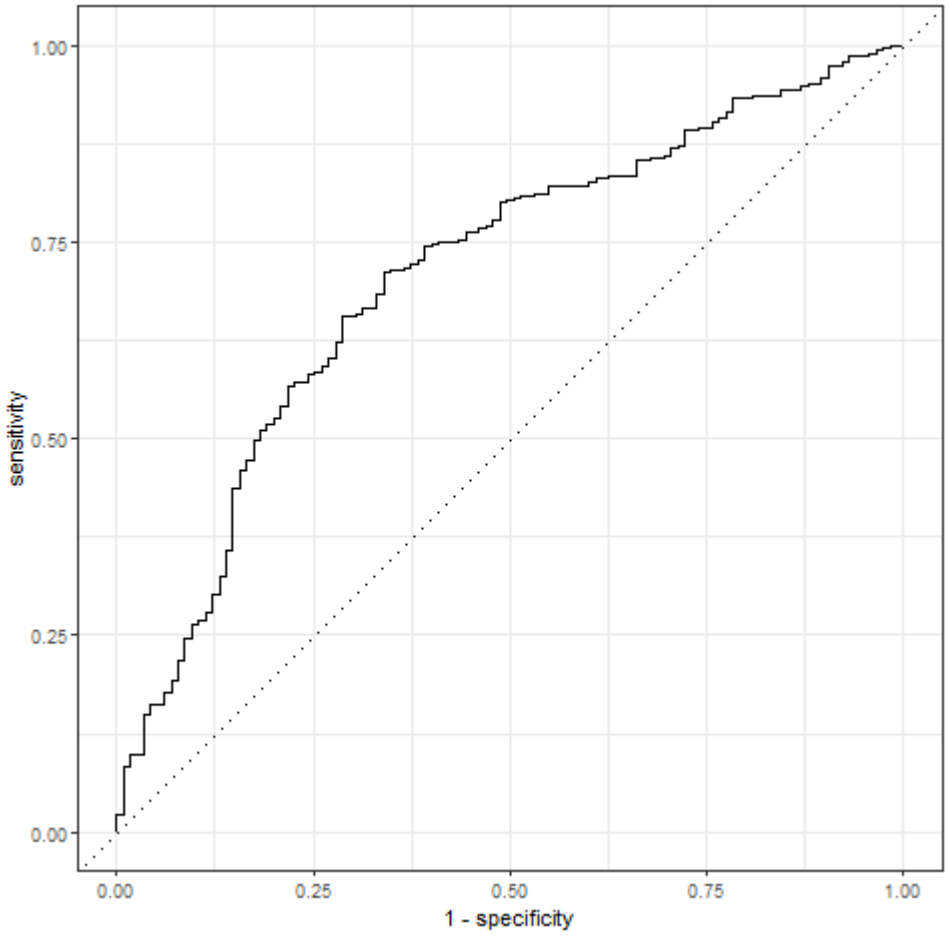

Figure 11: ROC curve of the Random Forest model

5.5 XGBoost

The final model we created used the XGBoost library. This is another sophisticated model known for strong performance as stated in the theory section. This model had the best performance and resulted in an AUC of 0.72. A slight improvement compared to the Random Forest model, and significantly better than the benchmark. This model also depends on the chosen hyperparameters explained in section 4.3 (figure A.1.5). The improvement from random forest may come from the ability of XGBoost models to learn from its mistakes to reduce residuals.

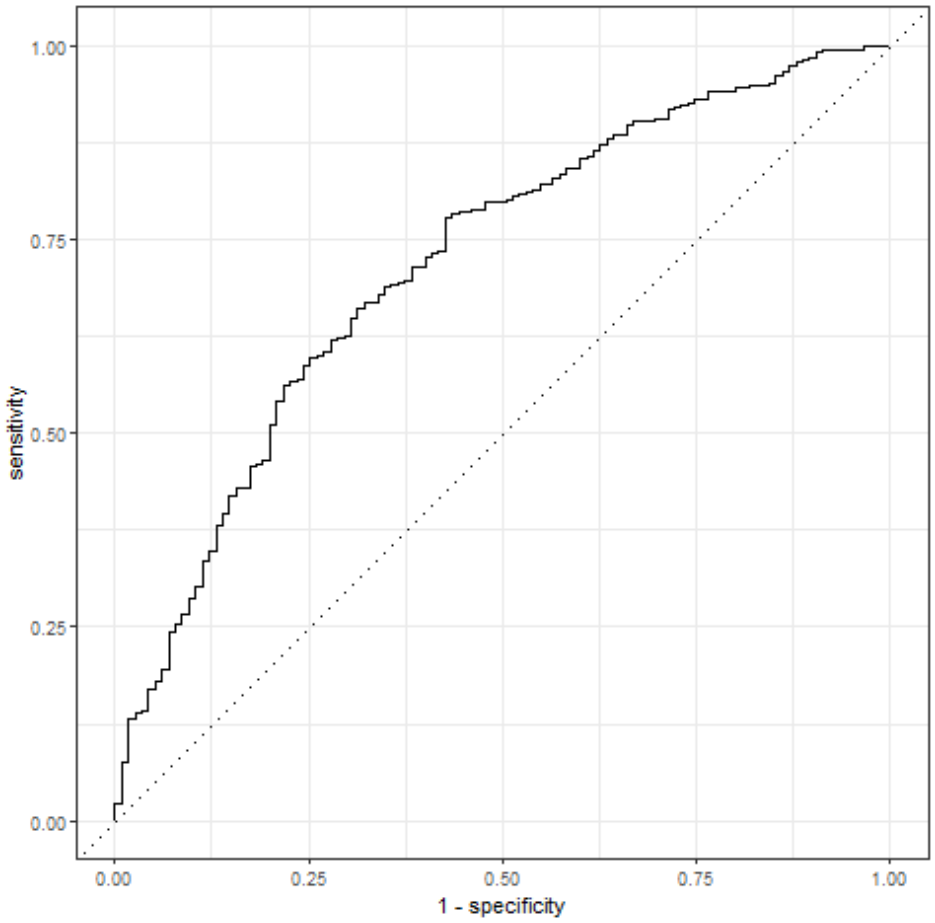

Figure 12: ROC curve of the XGBoost model

For further use of the XGBoost model we had to decide a threshold value, and as discussed in section 4.4. We chose to optimize the model for maximum precision. This resulted in a threshold value of 0.7. Lower thresholds resulted in the identical precision but chose 0.7 as it gave the highest recall score. We used recall score to favor the thresholds that resulted in the largest amount of actual SEO occurrences correctly predicted.

Threshold and performance		
threshold	precision	recall
0.65	0.40	0.04
0.66	0.40	0.04
0.67	0.40	0.04
0.68	0.33	0.04
0.69	0.37	0.06
0.70	0.40	0.08
0.71	0.33	0.08
0.72	0.30	0.08
0.73	0.27	0.08
0.74	0.30	0.13
0.75	0.29	0.15

Table 2: precision and recall for varying threshold levels, tested on the validation set

The confusion matrix below (figure 13) illustrates the model performance on the test data at the chosen threshold. Our model attained a precision of 34% on the test data, which is a significant increase compared to the benchmark model with 6.4% (from section 5.1). The accuracy of the model at the given threshold is 90%, but as stated in methodology, this is a bad measure when dealing with unbalanced outcomes.

	Predicted Negative	**Predicted Positive**
Actual Negative	1120 (TN)	25 (FP)
Actual Positive	102 (FN)	13 (TP)

Figure 13: confusion matrix of XGBoost model

To investigate the robustness of our models we used two new periods as training data (2010-2014 & 2016-2019) and tested the model on the year after each period (2015 & 2020). Below we present the AUC values of these models (Table 3). Both perform slightly worse than the original model with AUC's of 0.71 and 0.67 (compared to 0.72).

Robust test	
Test Period	AUC
2015	0.71
2020	0.67

Table 3: AUC from robust tests based on test periods

To further investigate the robustness, we conducted k-fold cross validation on the model. The AUC value produced from each iteration is illustrated in table 4. The results from this test were slightly below our original model with the average AUC equal to 0.71 (compared to 0.72).

AUC for different data folds	
Fold	AUC
1	0.73
2	0.71
3	0.70
4	0.71
5	0.69
6	0.69
7	0.71
8	0.71

Table 4: AUC from k-fold cross validation robust test

5.6 Impact of SEOs

To infer economic insight from our predictions we conducted a series of investigative analyses on the impact of SEOs on stock value. The investigations comprise of two linear regressions on different data samples, and a DiD regression. We tried to investigate the relationship between issues of equity and financial performance in the form of average monthly returns in both the year of issue and the subsequent year.

5.6.1 Linear regression using full sample

The initial regression analysis is conducted on all our data and investigates the relationship between SEOs and stock returns. This data sample consists of the full test data sample (1260 observations). The purpose is to identify whether there are statistical differences in financial performance between stocks that issue new shares and those who do not. Table 5 presents a regression analysis including four models. In model (1) and (2) the dependent variable is average monthly returns in the present year. The difference between the models is the introduction of control variables in model (2). In model (3) and (4) the dependent variable is the average monthly return in the following year. Similarly, we created models with and without controls. The variable *share issued* is a dummy indicating the occurrence of an SEO.

	Dependent variable:			
	monthly_returns_stock		monthly_ret_stock_lead	
	(1)	(2)	(3)	(4)
share issued	0.022***	0.009***	-0.003	-0.008*
	p = 0.00000	p = 0.005	p = 0.456	p = 0.054
Industry over market		0.526***		-0.158***
		p = 0.000		p = 0.003
log(m_cap)		0.003***		-0.001***
		p = 0.000		p = 0.008
bkvlps		-0.00001		-0.00001
		p = 0.596		p = 0.562
sd_stock		0.189***		0.116***
		p = 0.000		p = 0.000
Constant	0.020***	-0.051***	0.023***	0.030***
	p = 0.000	p = 0.000	p = 0.000	p = 0.001
Observations	1,260	1,260	1,260	1,260
R^2	0.024	0.437	0.0004	0.112
Adjusted R^2	0.023	0.435	-0.0004	0.108
Residual Std. Error	0.040 (df = 1258)	0.030 (df = 1254)	0.041 (df = 1258)	0.039 (df = 1254)
F Statistic	30.493*** (df = 1; 1258)	194.759*** (df = 5; 1254)	0.557 (df = 1; 1258)	31.632*** (df = 5; 1254)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5: Linear regression results on full sample

Significant results were found in three out of four models related to the variable *share issuance*. The data showed a positive correlation with returns in the year of the SEO, in both model (1) and (2). These findings were statistically significant at a 1% level. However, a negative coefficient is observed for the subsequent year, in both model (3) and (4). In this case only model (4), which included control variables, had statistically significant coefficient at a 10% level. This indicates a negative price reaction for stocks in the subsequent year of an SEO.

In model (2) and (4) both the standard deviation (*sd_stock*) and industry returns over market returns (*Industry over market*) gave statistically significant coefficients. In addition, model (2) found a statistically significant coefficient related to market cap (*mcap*).

The adjusted R-squared was highest for model (2) with a value of 0.435 and lowest for model (3) with a negative value of -0.0004. Increasing the number of control variables increased R-squared in both cases. This suggests that the inclusion of additional variables improves the regression model in its ability to explain return.

5.6.2 Linear regression using sub-sample

The next regression is performed identical to the previous one, except for a change in the data sample in which the analysis is conducted on. The sub-sample only consists of firms where we predicted an SEO to occur (true positives and false positives) inside the test period. Due to this, the regression only contains 68 observations.

	Dependent variable:			
	monthly_returns_stock		monthly_ret_stock_lead	
	(1)	(2)	(3)	(4)
share issued	0.010	-0.013	-0.004	-0.007
	p = 0.704	p = 0.316	p = 0.876	p = 0.781
Industry over market		0.613***		-0.597
		p = 0.003		p = 0.137
log(m_cap)		0.007		-0.007
		p = 0.111		p = 0.432
bkvlps		0.0001		0.0005
		p = 0.828		p = 0.676
sd_stock		0.232***		0.159**
		p = 0.000		p = 0.023
Constant	0.055***	-0.096*	0.043***	0.094
	p = 0.00001	p = 0.096	p = 0.0005	p = 0.417
Observations	68	68	68	68
R^2	0.002	0.779	0.0004	0.140
Adjusted R^2	-0.013	0.761	-0.015	0.071
Residual Std. Error	0.084 (df = 66)	0.041 (df = 62)	0.086 (df = 66)	0.082 (df = 62)
F Statistic	0.146 (df = 1; 66)	43.778*** (df = 5; 62)	0.025 (df = 1; 66)	2.020* (df = 5; 62)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Linear regression results on sub-sample

We do not observe any significant results in the coefficients related to the *share issued* variable. However, a negative coefficient appears in model (2) and (3) for the variable *share issued*. Due to the lack of statistical significance, the results do not allow us to draw conclusions about the true relationship. In model (2) and (4) the *industry over market* and *sd_stock* variables have statistical significance. The adjusted R -squared are improved for both (2) and (4) when we increase numbers of control variables to 0.761 and 0.071. This

suggests that adding control variables improves the regression model similar to the previous model.

5.6.3 Difference-in-differences

5.6.3.1. Discriptive analysis

To further investigate the relationship between share issuance and returns we conduct a DiD regression. In this experiment we also used the sub-sample of firms that we predicted to conduct an SEO, for reasons discussed in section 4.6. The sample is split into two groups based on whether they actually conducted an SEO. In table 7 we present a summary of the differences between the two groups, as the validity of a DiD regression depends on similarity between the control and treatment group. To do this we conducted a t-test.

Variable differences t-test for treatment and control group	
Variables	p-values
at	0.33
sd_stock_lag	0.32
monthly_returns_stock	0.15
m_cap_lag	0.94
industry_over_market_lag	0.39
Book value per share	0.68

Table 7: t-test investigating differences in control and treatment groups (using p-values)

From the t-test there is not a statistically significant difference in the variable means between the two groups. In addition, we supplement the descriptive analysis by illustrating the

difference in variable means for the control and treatment group. As presented in table 8 there are some differences in the means, even though they are not statistically significant.

Summary statistic for control and treatment group						
treatment	mean_assets	mean_SD_stock	mean_monthly_returns	mean_marketcap	mean_industry_ret	mean_book_value
0	1770.95	0.17	0.04	1609707	-0.01	6.71
1	2592.11	0.23	0.08	1549329	0.00	5.44

Table 8: Means for variables in control and treatment group

Lastly, we plot the historic returns in the control and treatment group from 2011 to 2019. Overlaying patterns with respect to historic returns strengthen the validity of the DiD model, as it strengthens the parallel trends assumption. From figure 14, historic returns in the groups seem to be correlating, with some variance.

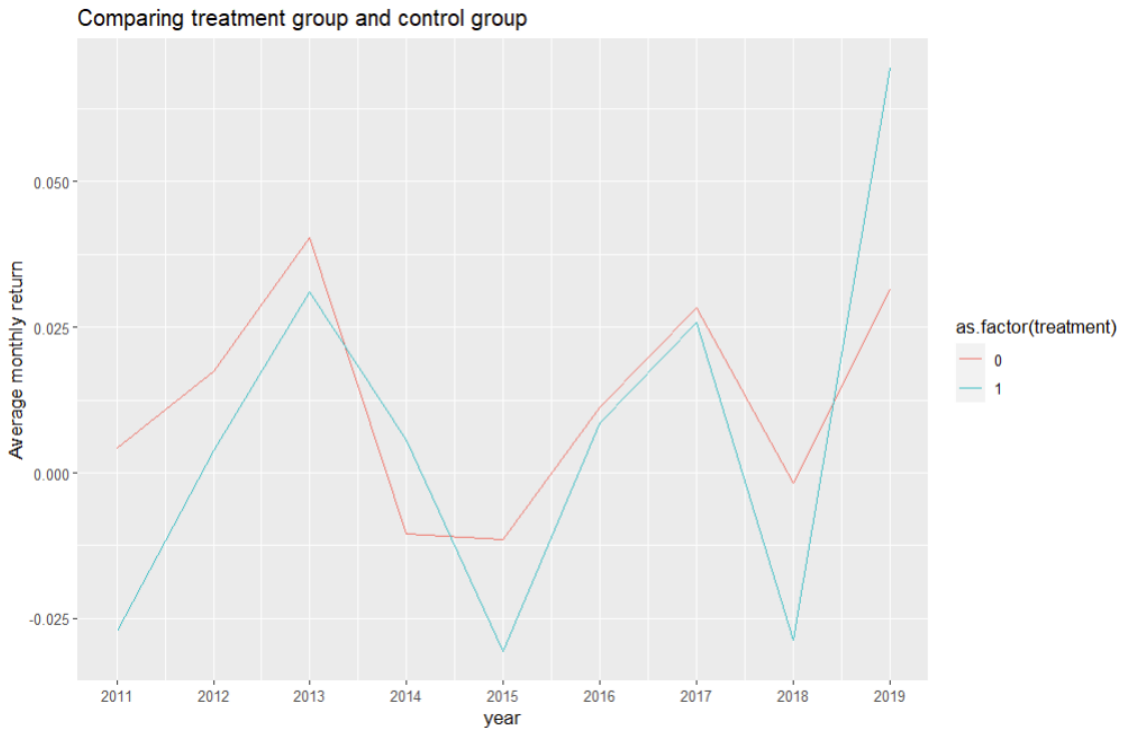

Figure 14: Historical returns for control and treatment group

5.6.3.2 DiD regression

In our DiD model the treatment group consists of firms the firms that conducted SEOs, while the control group consists of those that did not. As in the previous regressions we looked at

two different dependent variables and created models with and without control variables. The results of the DiD regression is presented in table 9.

Differnce in Differnce model

	Dependent variable:			
	monthly_returns_stock		monthly_ret_stock_lead	
	(1)	(2)	(3)	(4)
share issued	-0.024	-0.033***	-0.015	0.0001
	p = 0.455	p = 0.006	p = 0.693	p = 0.995
industry_over_market		-0.169		-0.602*
		p = 0.383		p = 0.069
log(m_cap)		0.064***		-0.091***
		p = 0.00000		p = 0.00001
sd_stock		0.273***		0.015
		p = 0.000		p = 0.771
bkvlps		-0.006**		0.006
		p = 0.025		p = 0.166
Observations	68	68	68	68
R^2	0.017	0.899	0.005	0.796
Adjusted R^2	-0.996	0.768	-1.021	0.529
F Statistic	0.572 (df = 1; 33)	51.893*** (df = 5; 29)	0.159 (df = 1; 33)	22.623*** (df = 5; 29)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: Results from difference-in-differences regression

In model (2) we found that the treatment group that conducted an SEO, underperformed in the year of the SEO compared to the firms that did not. The coefficient, with a value of -0.033, was statistically significant at a 1% level, with a p-value of 0.006. The remaining models failed to identify a statistically significant difference in returns related to the variable *share issued*.

6. Discussion

6.1 Machine learning model performance

Through the evaluation of our results XGBoost was identified as the best performing model in our study. In this section we compare the models and discuss the differences in performance. In addition, we reflect on our findings in relation to existing literature.

The purpose of the benchmark model was to provide an unsophisticated baseline, that makes predictions based on a single static metric: probability. This is not a machine learning model and is included as a comparison tool. The test results from the model indicated that the model had no predictive power, with an AUC of 0.5.

Amongst the machine learning models, the decision tree was the worst performing with an AUC of 0.56. The model is simple, and the poor relative performance was expected. It is a slight improvement compared to the benchmark. The logistic model, which is a linear model, showed a significant increase in performance with an AUC of 0.63. This is another relatively simple model, however the improvements in performance are significant compared to the benchmark and decision tree.

The more sophisticated machine learning models show an improved ability to predict SEOs. The random forest model achieved an AUC of 0.71, which is a substantial leap in performance from the previously discussed models. The XGBoost model produces the strongest results with an AUC of 0.72. These findings indicate that sophisticated non-linear models are useful in predicting SEOs. The complexity of these models makes them hard to interpret, therefore we cannot explain the exact reason for the outperformance observed from the XGBoost model. An explanation could be that it stems from its ability to learn and adjust to previous mistakes, an ability that suits complex data structures.

As previously explained, we used a validation set to modify and optimize the models. The validation results indicate the same relative differences in model performance as the test results, presented in table A.2.2. The XGBoost model provided superior performance followed by random forest.

From the robustness tests conducted on the XGBoost model we found similar performance using the cross-validation method (table 4). The other robustness test (table 3) saw a greater

variance (decrease) in performance relative to the main results. This may be explained by the smaller training sets. However, they are still performing on a similar level, and some variation in results should be expected. We believe these tests indicate sufficient robustness strengthening the generalizability of our model.

The AUC was helpful as a relative performance measure identifying the strongest model. Further on, we aimed to maximize the precision of the best performing model, as explained in section 4.4. The XGBoost model achieved a precision of 34 % on the test data. That means that 34 % of the positive SEO predictions produced by our machine learning model were correct. This seems low; however, one must assess the relative performance. In comparison the benchmark model has a precision of 6.4 %. Lack of research fundament on the predictability of SEOs make an objective performance conclusion difficult to reach. What these results do indicate is that SEOs are difficult to predict with high conviction.

6.1.1 Interpretation of XGBoost model

As stated, the XGBoost model has a shortcoming related to the interpretation of results. We employ variable importance analysis to obtain insights into the variables that have the greatest impact on the predictions. Conducting such an analysis is advantageous for comprehending the primary factors that drive the prediction model and identifying the most influential determinants for SEOs.

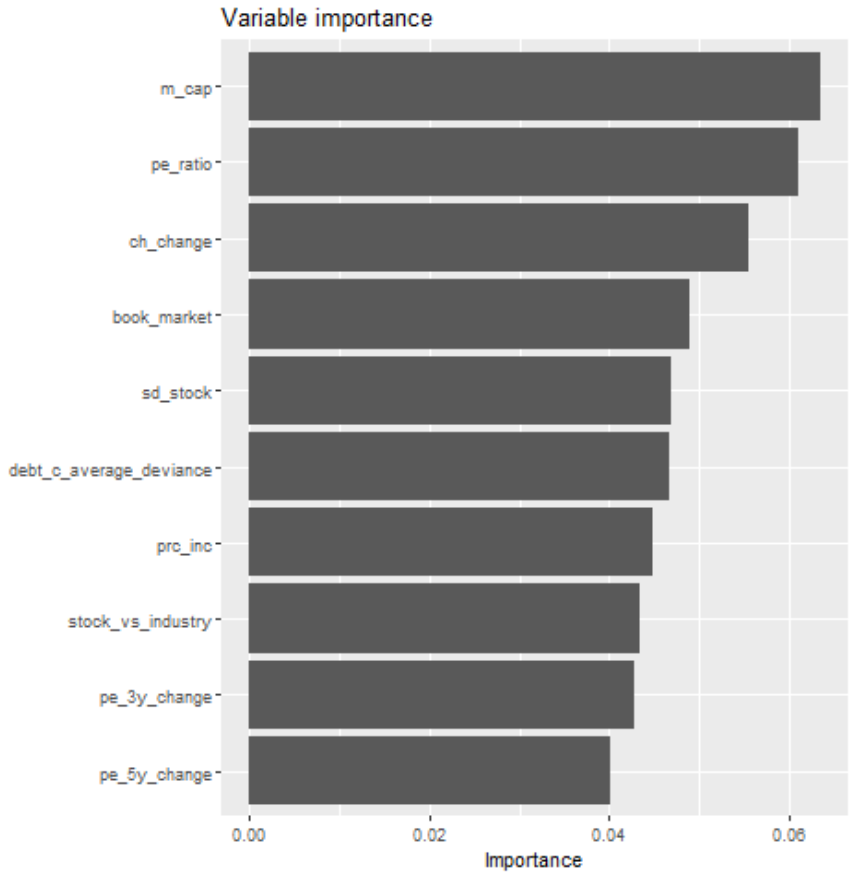

Figure 15: Variable importance plot

The most impactful features on the classifications made by the XGBoost model is market capitalization, PE-ratios, changes in cash balance, book-to-market value, standard deviation of the stock returns, deviances in current debt and stock performance. An interesting observation is the importance of several PE-ratio metrics. These findings are in line with previous studies investigating determinants of the choice to conduct an SEO. Paul Marsh (1982) and Eckbo and Masulis (1995) highlight similar impacting factors, strengthening both our results and the conclusion they reach in their studies (Marsh, 1982) (Masulis & Eckbo, 1995).

6.2 Regression models

In the second part of our research, we investigated whether we could find a relationship in our data indicating that SEOs have a negative impact on company value. As mentioned in the literature review, Masulis and Korwar (1986) found that SEOs have a negative impact on stock returns (Masulis & Korwar, 1986). The thesis is founded on this being true, and we

imply that an SEO propose a risk for investors. To increase the value of our research it is interesting to investigate whether this notion holds true, using our data.

The two linear regressions we presented in table 5 and 6, gave differing results both in relation to the impact of SEOs on returns, and the significance of the findings. The first regression, where we used the entire data sample, indicated a significant impact of SEOs on returns in 3 out of 4 models. The models using the subset data failed to find a significant relationship. This is likely due to the decreased number of observations in the second regression. These results support the idea presented by Carlson, Fisher and Giammarino (2010) that the negative stock performance around an SEO could be explained by other underlying factors (Carlson, Fisher, & Giammarino, 2010).

In the regression using the full data sample, firms that conducted an SEO had an increase in that years returns, which is contrary to our hypothesis. These findings were significant at the 1% level. It is important to note that since we use yearly data, we do not capture the reaction at the exact point of the SEO. The effect we observe is for that reason affected by movements in the stock price before the SEO was conducted. It is reasonable to believe that some of the firms that issue additional shares have seen a positive stock price development prior to the event, as SEOs are most beneficial when the company stock is highly valued (Berk & Demarzo, 2017, p.888). However, using the same data set we found that firms conducting an SEO experienced a decrease in stock returns in the following year. These findings were only significant in the model including controls at the 10 % level. In contrast to the model analyzing stock performance in the year of the SEO, these findings only illustrate the impact on stock returns prior to the SEO. It is however a weakness that this model use returns of a period not directly after the share issuance. The contradicting findings in the models analyzing different return periods each have its strengths and weaknesses. Therefore, the analysis does not imply a clear relationship between SEOs and returns.

To find more concise and strong evidence on the impact of SEOs on stock returns we conducted a separate linear regression illustrated in table 6. This time we used a sub-sample of stocks, including only the observations where we predicted the firm would issue additional shares within the test period. As explained in section 4.6., we chose this method to obtain a more homogenous group of firms whose variation could be more definitely attributed to the issuance of shares. Using this method, we did not find any statistically significant impact of SEOs on returns. This is likely due to the low number of observations

included in the regression (68). These results do however support the idea presented by Carlson, Fisher and Giammarino (2010) that the negative stock performance around an SEO could be explained by other underlying factors (Carlson, Fisher, & Giammarino, 2010). From table 6 we see that the model including control variables, looking at returns in the year of the SEO indicate a negative impact on returns. Both the models using returns in the year prior to the SEO also indicate a negative impact on returns from share issuance. As the lack of statistical significance makes these results inconclusive, we argue that they at least do not weaken the hypothesis that SEOs are negative for shareholders. If possible, it would have been desirable to replicate this regression analysis with more observations.

In a final attempt to determine whether firms that issue additional shares experience a different response in returns, we conducted a DiD regression. As explained in section 5.6.3., we conducted the analysis on the same sub-sample as in the previously discussed regression. The DiD regression resulted in a highly significant (1% level) difference in returns for the two groups in the year of the SEO, when including control variables (table 9). The difference was negative indicating a decrease in returns in the group that conducted an SEO, compared to those who only were predicted to do so. These results strengthen the initial assumption that SEOs negatively impacts stock performance, and contradict the ideas argued by Carlson, Fisher and Giammarino (2010).

6.3 Resuslts in relation to litterature

In this section we will discuss our findings in relation to previous studies. It's not straightforward to accurately interpret the XGBoost model's decision-making process. However, as discussed in section 6.1.1., we analyzed the 10 most important variables for the predictions. The variable importance does not explain which direction the relationship between a variable and the classification goes. However, we can compare the ranked variables to aspects of the studies discussed in the literature. The regressions investigating the relationship between SEOs and stock returns is another part of our thesis that we will compare with established ideas and evidence.

From the variable importance presented in figure 15 we observe the 10 most influential variables in the prediction model. As briefly mentioned, these findings are consistent with studies explained in the literature review. Historical stock prices, bankruptcy risk, company size and debt ratios were all highlighted in the study by Marsh (1982), as factors impacting

the choice of conducting an SEO (Marsh, 1982). The 10 most important variables in the XGBoost model represent aspects of all these factors, supporting the conclusions drawn by Marsh (1982). The article by Eckbo and Masulis (1995) argues that firms issue more equity compared to debt in expansive business cycles. To capture this effect, we implemented variables as effective federal funds rate, market returns and changes in GDP. These were not present amongst the top 10 important variables. The fact that stock performance versus industry performance is amongst the most important variables could be argued to somewhat reflect business cycles. However, we conclude by stating that our findings do not support or resemble those of Eckbo and Masulis (1995).

When making comparisons to existing literature it is important to note that previous studies were investigating the choice between equity or debt, whereas we are investigating SEOs only. Thus, we cannot expect their findings to perfectly manifest in our results, weakening our ability to draw a conclusion between our model and the previous studies. Either way aspects of the existing literature proved useful in building prediction models.

The results from the two linear regressions and the DiD estimation is comparable to similar existing research. As previously stated, the consensus belief on the relationship between firm value and additional equity issuance is that SEOs have a negative impact (e.g. (Masulis & Korwar, 1986)). Unfortunately, our investigative efforts failed to produce compelling evidence on the impact of SEOs on returns. The analysis we conducted had some weaknesses, resulting in the conclusion reached by Masulis and Korwar being more compelling. The study by Masulis and Korwar analyzed the variations in returns on a daily frequency around the SEO announcement. We analyzed yearly returns in response to the SEO announcement, leading to a more rough and less detailed investigation around the event.

6.4 Limitations of study

This thesis has some limitations which we would like to address in this section. The critique we will highlight is related to the intervals of our data, the machine learning models used and the applicability of our results.

The fact that we used yearly data resulted in limited possibilities to analyze effects near the SEO in time. We do not view this as a problem for the prediction models, but when

estimating the effect of an SEO on returns it would have been interesting to view data in the days and weeks surrounding the event. The relationship might then be more precisely attributable to the SEO. Conducting such an analysis was not feasible due to the nature of our data. It would require collecting new variables that are updated on a weekly or daily basis. This problem would be an interesting subject for further study. However, the failure to prove a relationship between SEOs and returns, makes it uncertain whether predicting firms that issue new equity is of benefit for investors.

Another aspect of the thesis that can be argued as a weakness is related to the interpretability and comparability of the prediction models. As mentioned in section 6.1.1., the XGBoost model is difficult to interpret and the same is true for the random forest model. This is however not crucial for the purpose of this thesis. The weakness related to comparability stems from the lack of existing literature predicting SEOs. The results from our prediction models are thus difficult to objectively deem either satisfactory or inadequate. To mitigate this issue, we included the benchmark model. However, it would have been desirable to view our results in relation to other studies regarding the predictability of SEOs. This thesis explores unchartered territory and can act as a basis and provide insight for further exploration.

7. Conclusion

Seasoned equity offerings have a direct impact on investors, who generally view it as a negative event. Therefore, we try to answer the research question: can machine learning models predict firms that will conduct an SEO in the future?. Then we investigate the economic implications of the prediction models, to identify whether they have beneficial practical use cases.

In conclusion our results indicate that machine learning models are beneficial in predicting the issuance of additional shares. The best performing model achieved an AUC of 0.72 and a precision of 34 %. This was a significant improvement compared to the benchmark model, which represents random guessing. The lack of comparable existing material on the predictability of SEOs makes our results difficult to evaluate objectively. Nonetheless we were able to present models in this thesis that possessed significant prediction power. In addition, we established that the sophisticated and non-linear models were superior in predicting SEOs. We also identified determinants that were proposed in the literature to impact a company's choice of conducting an SEO. Lastly, the relatively low precision from the best performing model (34%), leads us to believe that SEOs are difficult to predict with a high conviction. As this thesis delves into a relatively underexplored research question, we hope to provide a fundament for comparison and further improvements.

Through multiple regression analyses investigating the impact of SEOs on returns, we were not able to provide conclusive evidence on this matter. The reason being contradicting findings and lack of statistical significance. These tests were not the main objective of this thesis, resulting in some decisive flaws that have been discussed previously. Hence, we uphold the notion drawn in previous studies that SEOs have a negative impact on the market value of companies.

The models in this thesis can assist investors and stakeholders in identifying firms that will conduct an SEO in the future. Although we cannot provide proof of an economic benefit, we believe our findings to offer valuable insight for investors. The models may not be suited as a sole basis in decision making. However, they can indicate risk and supplement investment decisions.

References

Abadie, A. (2005, Januar 01). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*. doi:https://doi.org/10.1111/0034-6527.00321

(wrds) Wharton Research Data services . (u.d.). CRSP Monthly Stock. Pennsylvania, United States of America. Hentet March 30, 2023 fra https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/annual-update/stock-security-files/monthly-stock-file/

Berk, J., & Demarzo, P. (2017). *Corporate Finance Fourth Edition.* Pearson.

Brav, A., Geczy, C., & Gompers, P. (2000). Is the abnormal return following equity issuances anomalous? *Journal of Financial Economics*, 41.

Carlson, M., Fisher, A., & Giammarino, R. (2010, November). SEO Risk Dynamics. *Oxford University Press*. Hentet fra https://www.jstor.org/stable/40961307

CFA Institute. (2023). *Capital Structure*. Hentet fra CFA Institute: https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/capital-structure

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.*

Chen, T., & He, T. (2017). *xgboost: eXtreme Gradient Boosting.*

Damodaran, A. (2012). Investment Valuation. I A. Damodaran, *Investment Valuation* (ss. pp. 182-183). John Wiley & Sons, Inc.

Eckbo, B., & Masulis, R. W. (1995). Chapter 31 Seasoned equity offerings: A survey. *Elsevier B.V.*, Pages 1017-1072. doi:https://doi.org/10.1016/S0927-0507(05)80075-1

FRED. (u.d.). CBOE Volatility Index: VIX (VIXCLS). Saint Louis, Missouri, United States of America. Hentet May 1, 2023 fra https://fred.stlouisfed.org/series/VIXCLS?fbclid=IwAR2P9WiNNvSc0U7MT4NsHlzVr8zcd8sBbRQxMM59TgMNAEilHAh8mccSviQ

FRED. (u.d.). Equity Market Volatility Tracker: Macroeconomic News and Outlook: Business Investment And Sentiment (EMVMACROBUS). Saint Louis, Missouri, United States of America. Hentet May 1, 2023 fra https://fred.stlouisfed.org/series/EMVMACROBUS?fbclid=IwAR0_epOcAXjz08Myzl4-_OLzHcfocNWdFF_6I5y1957W2MsjW8qUBOYTf1c

FRED. (u.d.). Federal Funds Effective Rate (FEDFUNDS). Saint Louis, Missouri, United States of America. Hentet March 14, 2023 fra https://fred.stlouisfed.org/series/FEDFUNDS

FRED. (u.d.). Gross Domestic Product (GDP). Saint-Louis, Missouri, United States of America. Hentet March 25, 2023 fra https://fred.stlouisfed.org/series/GDP

FRED. (u.d.). University of Michigan: Consumer Sentiment (UMCSENT). Saint-Louis, Missouri, United States of America. Hentet March 27, 2023 fra https://fred.stlouisfed.org/series/UMCSENT

Friedman, J. (2002). *Stochastic gradient boosting.* CA 94305, USA: Stanford University.

Hoo, Z., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*.

Hui Hoo, Z., Candlish, J., & Teare, D. (2017, March 2017). What is an ROC curve? *Emergency Medicine Journal*.

IBM. (u.d.). *Decision Trees*. Hentet fra IBM: https://www.ibm.com/topics/decision-trees

IBM. (u.d.). *What is random forest?* Hentet fra IBM: https://www.ibm.com/topics/random-forest

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *Introduction to statistical learning.* Springer.

libguides. (u.d.). *A-Z Databases*. Hentet fra libguides: https://nhh.libguides.com/az.php?q=sdc&p=1

Marsh, P. (1982). The Choice Between Equity and Debt: An Empirical Study. *The Journal of Finance*, 121-144 (24 pages). doi:https://doi.org/10.2307/2327121

Masulis, R. W., & Korwar, A. (1986). Seasoned equity offerings: An empirical investigation. *Elsevier B.V.*, 91-118. doi:https://doi.org/10.1016/0304-405X(86)90051-6

Masulis, R., & Eckbo, E. (1995). Seasoned equity offerings: A survey. *Elsevier*.

Natelkin, A., & Knoll, A. (2013, December 04). Gradient boosting machines, a tutorial. *frontiers*.

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017, September 8). *Common pitfalls in statistical analysis: Logistic regression*. Hentet fra NAtional Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543767/

Rodriguez-Hernandez, Pruneda, & Rodriguez-Diaz. (2021). Statistical Analysis of the Evolutive Effects of Language Development in the Resolution of Mathematical Problems in Primary School Education. *Mathematics*, s. 8.

The R Foundation. (u.d.). *What is R?* Hentet fra r-project: https://www.r-project.org/about.html

The World Bank. (u.d.). *Difference-in-Differences*. Hentet fra Webområde for World Bank: https://dimewiki.worldbank.org/Difference-in-Differences

tidymodels. (u.d.). *Tidymodels.* Hentet fra tiidymodels: https://www.tidymodels.org/

Wharton Research Data Services. (u.d.). Compustat Daily Updates - Fundamentals Annual. Pennsylvania, United States of America. Hentet March 22, 2023 fra https://wrds-www.wharton.upenn.edu/pages/get-data/compustat-capital-iq-standard-poors/compustat/north-america-daily/fundamentals-annual/

What is random forest? (u.d.). Hentet fra IBM: https://www.ibm.com/topics/random-forest

Appendix

Appendix 1 - Tuning and hyperparameters

This section of our appendix shows figures and tables about the tuning of our models.

Below is snapshot of the random forest grid we used to tune using cross-validation. The Grid had a total of 500 different combinations of the variables, made by “tidy models” package.

grid for random forest tuning		
min_n	mtry	trees
11	21	251
16	22	151
37	8	9
28	25	39
26	11	402
36	24	350
14	23	469
4	13	466
28	2	27
8	2	138

Table A.1.1 - Random Forest grid (10 first)

Random Forest: Hyperparameters after tuning		
mtry	trees	min_n
5	230	23

Table A.1.2 - Random Forest hyperparameters

We used cross validation to figure out how many trees were we where going to tune with in XGBoost and Random Forest. Increasing the number of trees increases computer demand and time. Therefore we decided for both models to create a grid where we tune trees from 1-500. It seems from both plots that this seems reasonable. This cut off was made due to poor performance in our computers, making tuning with to many trees to be too time consuming. From the plots it seems like it does not help much to add more trees.

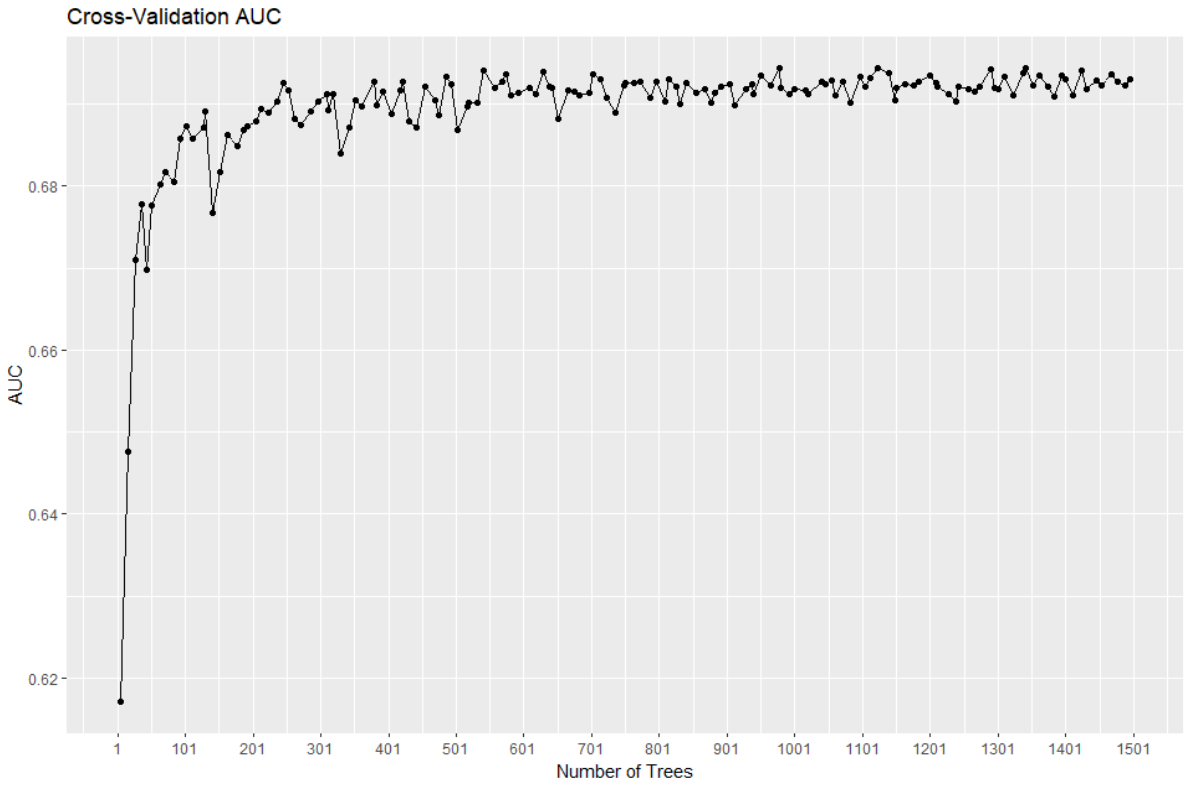

Figure A.1.3 – Random Forest trees: for range to tune

Below is two snapshots of the XGBoost grid we used to tune and our final hyperparameters after tuning.

XGBoost grid				
trees	min_n	tree_depth	learn_rate	loss_reduction
255	19	4	2.032760e-06	2.518268e-10
248	35	12	3.554604e-03	1.344985e+00
229	29	2	7.218426e-04	1.193542e-10
275	18	5	1.834449e-06	1.191397e-06
445	32	13	8.657415e-06	8.182065e-10
210	8	4	1.905286e-07	3.013254e+00
306	32	10	5.500218e-08	2.304834e-03
442	26	8	9.129951e-07	6.525553e-08
281	20	6	1.510237e-06	2.526071e+01
225	19	12	3.337065e-08	1.831075e-03

Table A.1.4 - table: XGBoost grid (10 first)

XGBoost: Hyperparameters after tuning				
trees	min_n	tree_depth	learn_rate	loss_reduction
478	35	10	0.014	0.0534

Table A.1.5 - XGBoost hyperparameters

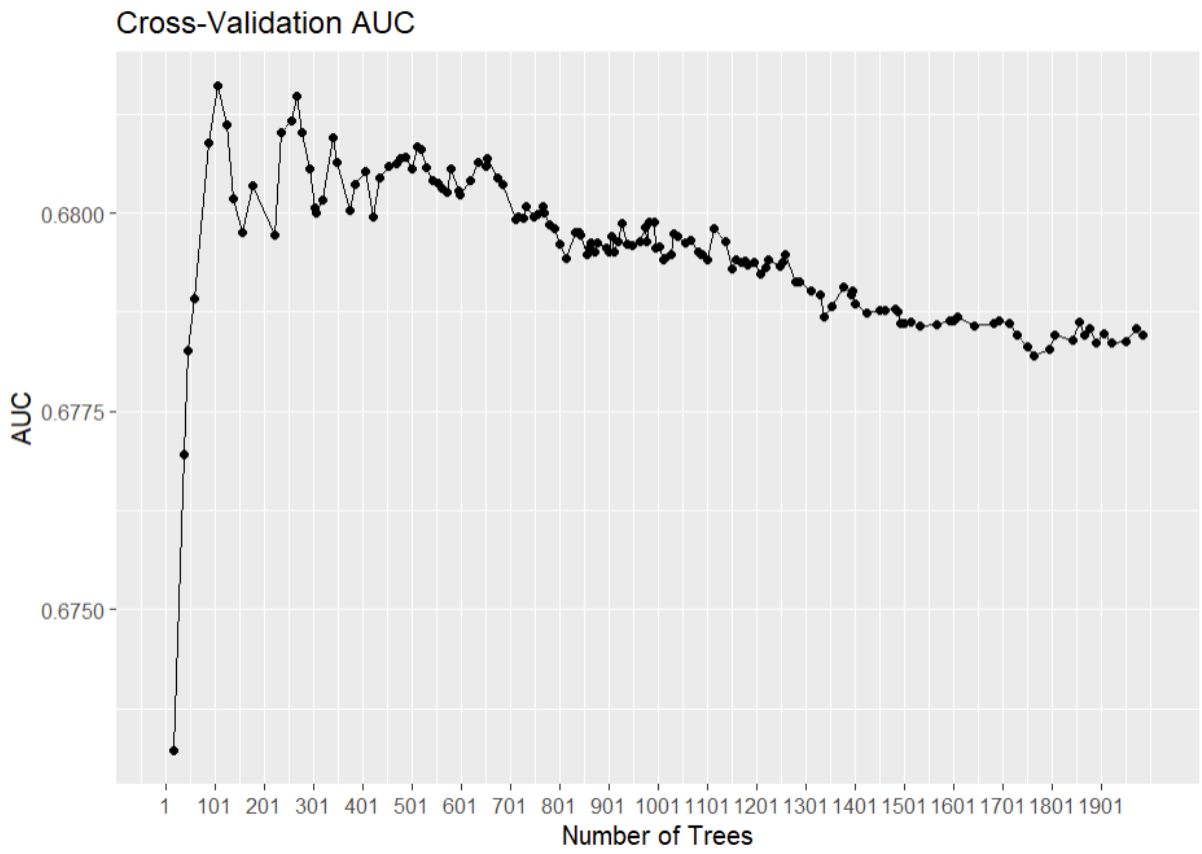

Figure A.1.6 plot - XGBoost trees to find range to tune

Appendix 2 - Validation and Robustness

This section of our appendix focuses on validation and robustness test of our models. Which are important for our thesis.

Below is the test AUC for each model after tuning and choosing the subset of variables.

Comparing model's AUC: test data	
model	AUC
XGBoost tuned	0.72
Random Forest tuned	0.71
Logistic model	0.60
Decision tree	0.56
Benchmark model	0.50

Table A.2.1 - Test data AUC

Below is the AUC on the validation set, from these results we made the decisions of variables featuring for the models

Comparing model's AUC: validation data	
model	AUC
XGBoost: 33 Variables	0.73
Random forest: 33 Variables	0.70
XGBoost: 22 Variables	0.66
Random forest: 22 Variables	0.67
XGBoost: 10 Variables	0.58
Logistic: 33 Variables	0.54
Logistic: 10 Variables	0.54
Random forest: 10 Variables	0.53
Decision tree model_1 validation	0.52
Decision tree model_1 validation	0.52
Logistic: 22 Variables	0.51
Decision tree model_1 validation	0.50

Table A.2.2 - Validation AUC

Our model worked with an AUC of 0.72 on the test data. To further verify results, we made robust test on different time periods to validate our results. The performance is plotted below.

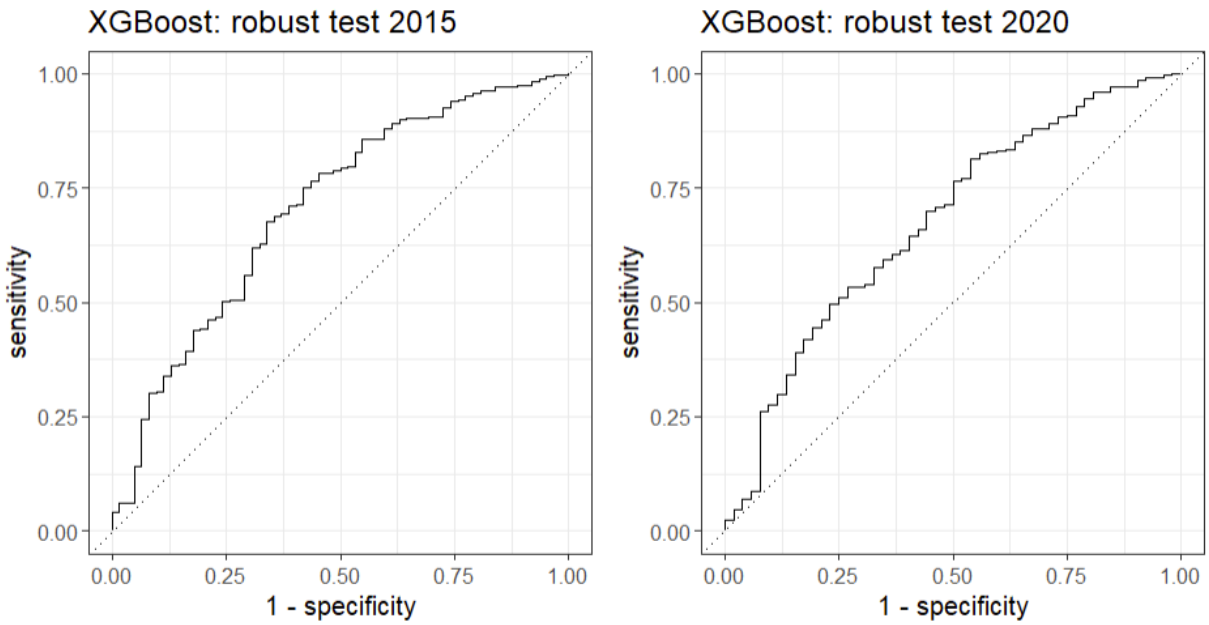

Figure A.2.2 - Robust test

Appendix 3 - Visualizations

This section we use to visualize different data which could be useful for our readers to further understand our methods and data.

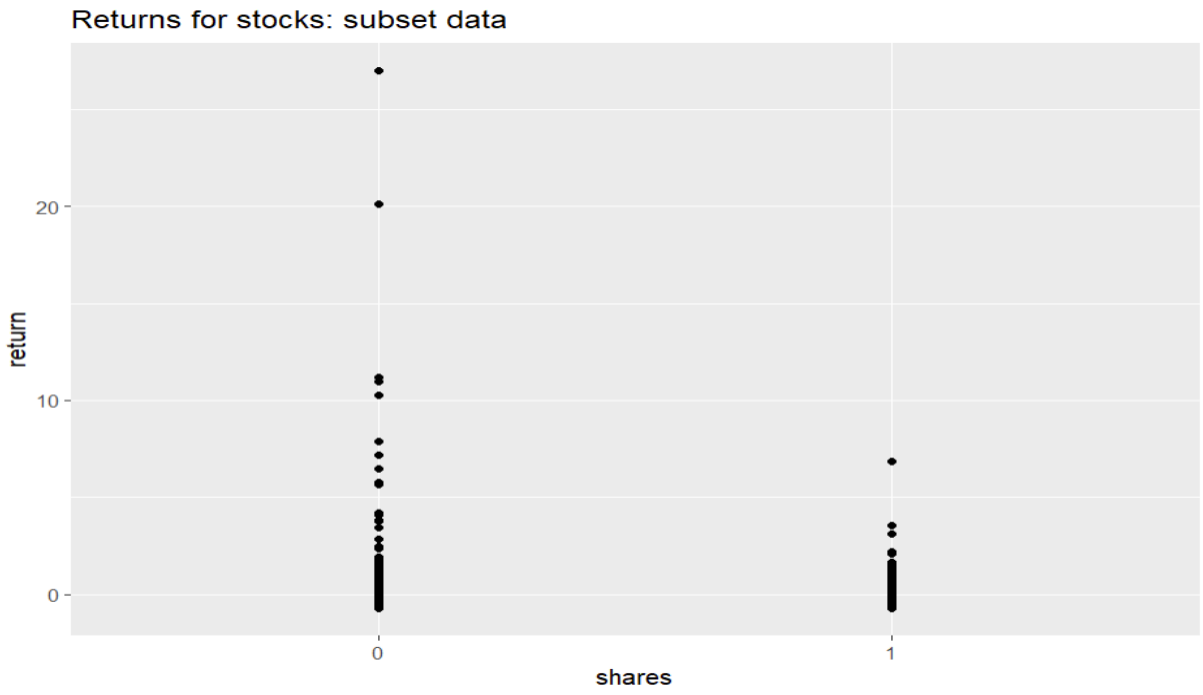

Figure A.3.1 - Returns for subset in Difference-in-Difference model

Example of exploratory analysis we did when creating model to investigate if there were any promising variables. It was hard to find any variables that seemed promising, we landed almost exclusively on variables that were mentioned in previous literature.

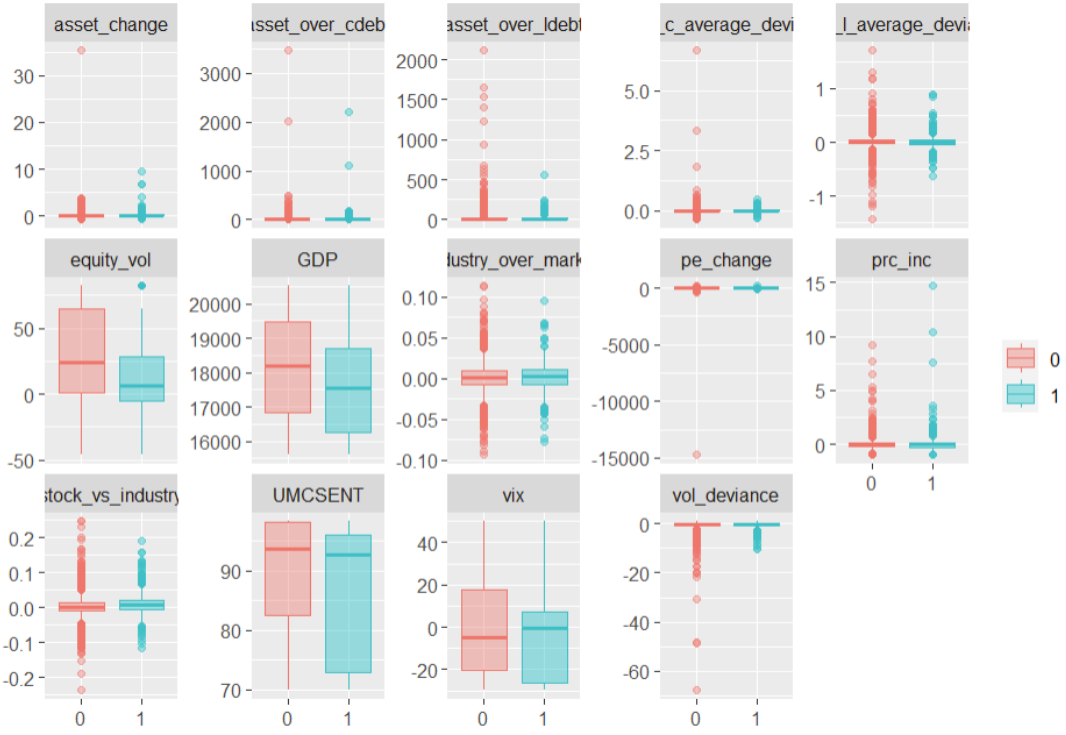

Figure A.3.2 – Investigation of subset of variables

From the plot below we see that most firms did not issue or issued one time in the whole period from 2010 – 2021. However, some companies issue a lot more, up to 11 times. If a company issued several times a year, we counted that as one issue.

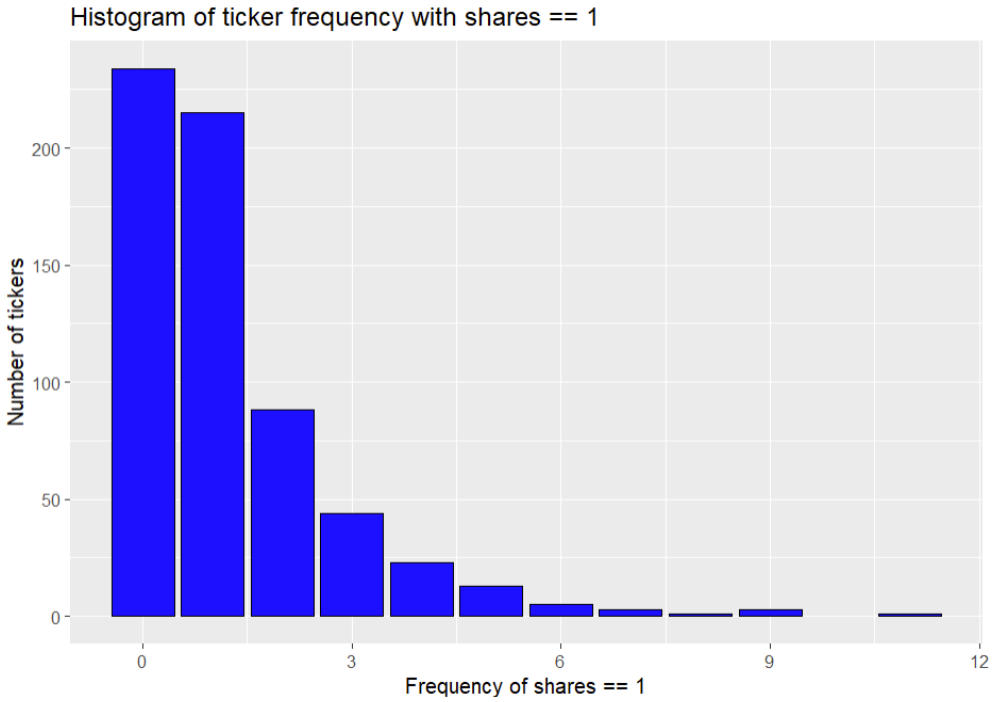

Figure A.3.3 - frequencies of issuance per firm