PREDICTING IPO UNDERPRICING

A STUDY ON THE PREDICTABILITY OF IPO UNDERPRICING THROUGH MACHINE LEARNING ALGORITHMS

AXEL TARDELL
VICTOR HOLGERSSON

Bachelor thesis
Stockholm School of Economics
2023

Abstract

This paper primarily serves to examine whether a specific subset of variables, derived from publicly available pre-IPO data, can be effectively modeled to predict and classify if an IPO will be underpriced using non-linear machine learning (ML) models. Secondly, we analyze whether the performance of ML-based models is greater compared to conventional linear models. Specifically, the focus has been on: linear, neural network (NN), random forest (RF), and gradient-boosting tree (GBT) approaches, including both regression and classification tasks. The findings indicate that given our input data, predictability is attainable solely through a classification approach, albeit with moderate support. Additionally, the evidence of this study favors machine learning models and their ability to capture complex patterns in financial data. This paper may be complemented with further analysis in order to reach a conclusion regarding the actual predictability of IPO underpricing.

Keywords: Machine Learning, Neural Network, Random Forest, Gradient-Boosting Trees, IPO Underpricing

Authors:

Axel Tardell (25414)

Victor Holgersson (25192)

Tutor:

Milda Tylaite, Assistant Professor, Department of Accounting

Date:

December 5th, 2023

Acknowledgements:

First and foremost, we would like to extend our appreciation to Milda Tylaite, our supervisor, for her invaluable insights during the research journey. Additionally, we express gratitude to our peers in the supervision group for their consistently thought-provoking feedback. Lastly, we extend special thanks to Vladimir Bril for his invaluable support and industry-specific insights.

Table of contents

1. Introduction

1.1 Purpose

Initial public offerings (IPOs) often exhibit underpricing, a phenomenon where newly issued shares generate positive initial returns during their first day of trading. This study examines whether it is possible to apply machine learning models to predict and classify IPO underpricing using a specific subset of publicly available information prior to the IPO. Furthermore, this study will compare the performance of machine learning models to linear models in this field of research. The study will be conducted on data from the US market during the period 2010-2020. In this paper, positive initial returns and underpricing are used interchangeably. Previous research on the topic has primarily relied on linear models as standard methodology, particularly multilinear linear regression analysis. However, the emergence of machine learning techniques has gained traction beyond the academic literature, with financial practitioners employing sophisticated machine learning models in order to develop quantitative investment strategies (Krauss, Do & Huck, 2016). This study aims to bridge the gap between established academic literature and financial practitioners by exploring the application of machine learning models to predict and categorize IPO underpricing.

1.2 Background

An IPO is a process in which a firm, for the first time, offers their shares on a stock exchange. There are several factors motivating an IPO, but the main purpose is that it enables improved access to capital and enhanced liquidity for the company. Likewise, the IPO can also serve as an exit strategy for existing shareholders that want to divest and cash out. The underpricing of IPOs is a phenomenon where the stock price of newly issued shares rises significantly on the first day of trading in a public market. In essence, IPO shares are priced lower than their market value, and are consequently corrected by the market, generating a “pop” in the stock price. In turn, this results in the issuing company “leaving money on the table”. That is, not receiving as much

equity from their issue as they could have raised. In order to go public through an IPO, companies generally contract an underwriter to manage the process, typically investment banks. Researchers like Reilly & Hatfield (1969), Stoll & Curley (1970), Logue (1973), and Ibbotson (1975) were among the first to record the systematic phenomenon of IPO stocks having positive initial returns, suggesting that IPOs are on average underpriced. For instance, research shows that IPO shares on the US market experienced average initial returns of +19.0% during the period 1980-2022 (Ritter, 2023). IPO underpricing has been studied by a plethora of researchers throughout the years, however most of the research has been devoted to explaining why it occurs. The aim of this study is not to explain why underpricing occurs, but rather to explore if it is feasible to predict and classify IPO underpricing by using a certain subset of publicly available parameters prior to the offering. Nonetheless, theories and research on why underpricing occurs will be presented in order to provide a theoretical framework that will aid in both contextual understanding and the interpretation of the results.

1.3 Contribution

A gap exists in the literature regarding whether machine learning can effectively predict and classify IPO underpricing using a specific subset of variables, and whether the performance of ML-based models is greater compared to conventional linear models in this field of research. Furthermore, there is a disparity between financial practitioners and financial research, with academia lagging behind in the application of machine learning models.

To address this gap and contribute to the existing literature, this study will apply a similar methodology of Krauss et al. (2016), who applied neural networks (NN), gradient-boosting trees (GBT), and random forests (RF) to develop a trading strategy for statistical arbitrage. To our knowledge, very few studies using NN, GBT, or RF methods to estimate IPO pricing have been conducted. Following a methodology closely related to Krauss et al. (2016), this research integrates similar ML models for regression and classification tasks related to IPO underpricing. Additionally, this study will include linear models to enable comparison with the selected ML models. Furthermore, underpricing will be measured both as a continuous variable and as a binary variable. That is, to predict both the extent and occurrence of IPO underpricing. Notably,

literature suggests that binary classification excels in financial data modeling (Enke & Thawornwong, 2005). This paper aims to examine and contribute with the following:

1. Can a specific subset of publicly available variables prior to an IPO be effectively modeled to predict and classify if an IPO will be underpriced using conventional linear and non-linear ML models?
2. Subsequently, are ML-based models performance greater compared to conventional linear models?
3. Bridging the gap between financial practitioners and academic research by applying advanced machine learning methodologies, thereby improving the integration of sophisticated ML models from industry practices into academic literature.

2. Literature Review

This section is dedicated to the emergence of literature and empirical studies concerning IPOs. Firstly, it will present how IPOs are priced and seminal studies from the early years of IPO related research. Secondly, it will outline studies that estimate IPO underpricing based on linear regression, as well as non-linear approaches used in other financial studies. Thirdly, it will introduce the theoretical framework, providing the contextual basis from which our discussion and conclusions will derive. The aim of this literature review is to establish a comprehensive understanding of the subject, to provide a foundation for conducting the research efficiently and to formulate hypotheses.

2.1 The Pricing of an IPO

The typical process of conducting an IPO in the United States involves a series of structured steps. After selecting an investment bank to oversee the IPO, the company initiates the process by filing an S-1 statement with the Securities and Exchange Commission (SEC). Upon receiving approval from the SEC, the chosen investment bank conducts a “road show” where the company’s executives present the investment opportunity to institutional investors. During these presentations, investors offer feedback in the form of indications of interest, which are non-binding expressions of potential investment intent. These indications are recorded in a “book”. Considering the feedback received and the prevailing market conditions, the investment bank suggests an offering price to the issuing company. Following the pricing, investors are asked to confirm their earlier indications of interest, shares are allocated accordingly, and trading commences within a few hours (Ljungqvist, 2007).

2.2 History of IPO Studies

In the early stages of literature and studies on IPO underpricing, it became evident that estimating initial returns for IPOs was a complex endeavor. Ritter (1984) established that there is a positive relation between the risk related to the IPO, and its underpricing. Beatty & Ritter (1986) argue that even if IPOs typically exhibit underpricing on average, an investor subscribing to an IPO cannot know the value of an offering before it starts trading publicly. This uncertainty

is referred to as ex ante uncertainty. Additionally, they theorize that when the ex ante uncertainty increases, the expected underpricing will follow and also increase. Lowry, Officer & Schwert (2010) observed patterns in the volatility of initial returns for IPOs. They suggest that the observed patterns support the idea that valuing private companies with ambiguous futures are challenging, and consequently underpricing serves a reaction to the intricacies of this valuation dilemma. The valuation problem and the volatility of returns were particularly evident in the context of younger firms and those operating in complex industries. In essence, a great deal of the early studies within the field was based on how firm specific characteristics influenced IPO pricing.

2.3 Linear Regression Studies

Studies based on linear models have contradicting evidence on whether IPOs can be priced using publicly available information prior to the offering. Lowry and Schwert (2002) find evidence supporting that IPOs cannot be priced using publicly available information. Contradicting this, Butler et al. (2014) compiled a list of 48 variables that had been utilized in prior research and applied them in a multilinear linear regression analysis, and discovered that 15 out of the 48 variables proved to be statistically significant. Much of the literature preceding their study often focused on only a handful of variables. Their approach helped mitigate the risk of omitting crucial control variables that could bias the impact of various independent variables. By controlling for all 48 variables in their regression analysis, they obtained results that partly contradicted previous findings regarding the robustness of certain variables. Subsequently, they isolated the set of the 15 robust variables and conducted a linear regression analysis with them, which yielded an adjusted R-squared of 45.5% Furthermore, their findings have contributed significant insights to various theories explaining the occurrence of this phenomenon.

2.4 Machine Learning Adaptation

While there's a scarcity of robust studies using ML to estimate IPO underpricing, we turn to a reputable study, namely Krauss et al. (2016) that employed NN, GBT and RF models in a closely

related financial context. It is believed that these ML methods, known for their ability to capture complex patterns in financial data, can be adapted to address the challenges of estimating IPO pricing effectively. By leveraging the models' predictive power and tailoring them to our research objectives, we aim to fill an evident research gap while building on established methodologies.

Huck (2009) developed a statistical arbitrage strategy based on various ML algorithms. His methodology consisted of performing forecasts through NNs to generate a predictive future return. The algorithm in Huck's study bases this forecast on previous returns in a time-series vector. He found that the statistical arbitrage strategy produced weekly excess returns of more than 0.80% indicating the accuracy and strength of NNs. Kraus et al. (2016) adopted this methodology and developed ML algorithms such as NN, RF and GBT to create a statistical arbitrage trading strategy. They used the different methods to evaluate what profits each model had produced after a given period of time. Their results show promising evidence of the capability of ML models to model financial data, for instance, their NN, GBT and RF models generated approximately 0.30%, 0.38% and 0.44%, respectively, in daily returns. Additionally, they find that RF proved to be the best performing model in their study. However, their work is built on a data set based on a time series while the majority of the variables in this project are based on a cross sectional set of predetermined parameters.

2.5 Theoretical Framework

Ljungqvist (2007) reviewed the principal theories on why the underpricing phenomenon occurs. Ljungvist advocates that the explanatory theories can be based on four principles, and categorized followingly: asymmetrical information, institutional, control, and behavioral.[1] Out of these four categories, it is the former that has obtained the most empirical research and the most recognized theories. One of these theories is the winner's curse by Rock (1986). The theory holds that well-informed investors bid selectively for attractively priced IPOs, while uninformed investors bid indiscriminately for all IPOs. The result is a winner's curse, where the uninformed investors will be crowded out by informed investors in the good IPOs, but "win" the full

[1] See Ljungqvist (2007) "Chapter 7 - Handbook of Empirical Corporate Finance"

allocation of bad IPOs. This can result in negative average returns for the uninformed investor, consequently causing them to quit participating in the market. Thus, to avoid a withdrawal of uninformed investors, and ensure a functioning market, IPOs must be underpriced on average.

The findings of Butler et al. (2014) suggest that historical data on previous IPOs and publicly available information prior to the IPO significantly impact their initial returns, thus weakening the support for explanations relying on asymmetric information. Even if asymmetrical information theories are the most established, it becomes evident that other explanations and forces are at play. Consequently, other explanations should be included in the discourse regarding why the phenomenon occurs.

Additional theories that are of relevance are share allocation theories, which explore how the allocation among investors is executed. Share allocation theories fall under Ljungqvist's (2007) "control" category of explanation theories. Ritter & Welch (2002) suggest that one of the reasons this has become an interesting theory is due to the perceived unfairness in how shares are allocated. They argue that studies have shown evidence that underwriters in certain cases intentionally have left more money on the table than necessary, in order to favor buy side clients. The evidence of Booth & Chua (1996), Brennan & Frank (1997), Mello & Parsons (1998), and Stoughton & Zechner (1998) all suggest that intentionally underpricing an IPO creates an excess demand and allows the issuer, and the underwriter to control the allocation of shares. In line with this, studies on the US market by Hanley & Wilhelm (1995) and Aggarwal, Prabhala, & Puri (2002) showed that institutional clients were given preference in the allocation of IPOs. Concluding, there are many theories explaining the phenomenon. Asymmetrical information explanations were previously used as the basis to understand IPO underpricing, however, given the results of Butler et al. (2014) this study will also take into account other explanations such as share allocation theories, as the literature support their influence and role in understanding IPO underpricing.

2.6 Hypothesis

The objective of Butler et al.'s (2014) study was to establish a benchmark, identifying the key variables that exert a significant impact on IPO underpricing. Due to limited accessibility and time constraints, this study is not able to replicate the same set of variables as them. Furthermore, the variables proved significant in their study were examined during an earlier time period, potentially worsening significance in the time period examined in this paper. Yet, similar variables are retrieved as the literature supports these variables' significance. Additionally, a method similar to Krauss et al. (2016) is adopted due to their successful application of ML models in a similar financial context, as seen by their promising results. Furthermore, underpricing is measured both as a continuous and as a binary variable. The former is due to the fact that a majority of IPO underpricing studies are based on regression analysis. The latter is because research has shown that binary classification tasks excel in financial data modeling (Enke & Thawornwong, 2005). In light of the above, we formulate the following research question: *"Can a subset of publicly available variables prior to an IPO be effectively modeled to predict and classify if an IPO will be underpriced using non-linear ML models? Subsequently, is the performance of ML-based models greater compared to conventional linear models?*

Given the literature, we hypothesize that IPOs, to some degree, can be predicted and classified using a subset of public pre-IPO data. Based on the empirical findings of Enke & Thawornwong (2005), we believe that the classification will provide better results than the regressions. Furthermore, given the promising results of Krauss et al. (2016), and the complex nature of IPO data, we believe that the machine learning models will outperform the linear models.

3. Data and Software

3.1 Data and Variables

3.1.1 Data Collection Process

The primary source for the data collection is Refinitive Eikon. The variables *total assets, net income, age, industry dummy, proceeds amount and offer price* were all gathered from Eikon's database. The variable *prior 30 day returns* was retrieved from Capital IQ. In an ideal scenario, not limited by any resources or time constraints, the data should be collected directly from each firms' IPO prospectus to ensure the highest quality of data.

In Eikon's database, we gather data on IPO issues spanning from January 1, 2010, to December 31, 2020. The focus was on listings exclusively from US stock exchanges. Following previous studies, all IPOs with an offer price below five dollars were excluded.[2] This process generated a data set consisting of 1148 observations, which was subsequently imported for analysis.

To improve the efficiency of our analysis and improve the accuracy of our regression and classification models, the data was processed and cleaned from transactions including no values or extreme values. The data was processed by removing two security types: units and American depositary shares, leaving us with issues of common and ordinary stock. This resulted in a reduction of 329 IPOs. Next, we removed IPOs on other listings than Nasdaq or NYSE, which amounted to a removal of 15 data points. Further, 6 and 62 points were removed as they contained NaN inputs for total assets and net income, respectively. The variable age had missing values on 154 data points, which were removed. At this point the data set contained 582 data points. Next, we looked at the distribution of the dependent variable and found that the data set predominantly exhibits characteristics of a normal distribution curve, although there is a slight

[2] Shares trading below five dollars on the US market are recognized as penny stocks. There are several reasons to exclude penny stocks, for example, penny stocks may not align with the typical characteristics or behaviors of IPOs with higher offer prices. Excluding them helps maintain data consistency and reliability for the analysis.

rightward skew observed. To remove any outliers we calculated the mean and standard deviation, and filtered out outliers in the dependent variable by using the upper bound $\mu + 3\sigma$ and the lower bound $\mu - 3\sigma$.[3] Accordingly, we deleted 5 data points above the upper bound (215%) and 0 data points under the lower bound (-101%), since initial returns cannot take on a value below -100%. After having processed the data, we obtained a set containing 577 data points.

3.1.2 Variable Selection

TABLE 1

VARIABLES

The variables listed below have demonstrated significance in prior literature and are utilized in this study as predictors for IPO underpricing. Further detailed definitions and explanations of these independent variables are provided below.

Variable Characteristics	**Variable Name**	**Name (in Python)**	**Database**
1) Firm characteristics			
a) Size	Total assets	total_assets	Eikon
b) Profitability	ROA	roa	Eikon
c) Age	Age	age	Eikon
2) Industry characteristics	Industry dummy	dummy_sum	Eikon
3) Offering characteristics			
a) Size	Proceeds amount	proceeds_amount	Eikon
b) Price	Offer price	off_price	Eikon
4) Market characteristics	Prior 30 day return	market_return	Capital IQ
5) IPO Underpricing (DV)	IPO underpricing	first_day_return	Eikon

[3] The positive skewness of the dependent variable indicates that we have a presence of a few unusually large outliers, which will cause the mean of our dependent variable to become greater. That is, underpricing will on average become greater. Statistical theory suggests that any observation outside three standard deviations from the mean is an outlier.

Age

Firm age is calculated as the difference between the year of the IPO and the founding year of the firm. Beatty & Ritter (1986) theorized that IPO underpricing increases in the ex ante uncertainty of the firm. Loughran & Ritter (2004) exhibit a pattern between initial returns and age, and suggest that younger firms are riskier, and thereby, investors subscribing to IPOs of younger firms must be compensated for that risk in the form of underpricing. In line with these studies, Chambers & Dimson (2009) use age, size and valuation as proxies for firm risk in their study. Similar to Chambers & Dimson we use age as one of our proxies for firm characteristics and uncertainty.

Total assets

Similar to the age variable, size is commonly used as a proxy for firm risk. There are several approaches to measure size. Loughran & Ritter (2004) use sales and assets as measures for size, in their proxy for firm risk. However, their findings show that sales lack significance as a predictor variable, while assets show significance. Given the evidence of Loughran & Ritter (2004) we use total assets before offering as a measure of size. Total assets is defined as total assets on the balance sheets, including current assets, long term investments and funds, net fixed assets, intangible assets and deferred charges, before offering.

ROA

When analyzing a company's profitability, various measurements come into play. In this study, we focus on one such metric, namely return on assets (ROA). ROA is calculated by dividing a company's net income by its total assets, both of which are imported from Eikon. This variable aims to gauge how effectively a firm utilizes its assets to generate profits. Beatty & Ritter (1986) proposed a theory suggesting that underpricing tends to rise when there is greater uncertainty before an IPO. Factors such as the firm's performance and quality contribute to this uncertainty in valuing a company. Moreover, findings from a study conducted by Purnadanandam & Swaminathan (2004) indicate that IPOs experiencing significant initial returns often possess

lower profitability compared to those without such high initial returns. Therefore, by incorporating this variable, the aim is to capture the firm's profitability and risk before its IPO.

Prior 30 day returns

Butler et al. (2014) show that the prior 30 day return of the stock market has shown to be a robust predictor of IPO pricing. Noticing the significance of how previous market conditions influence the choice of issuing equity and its effect on initial returns we include it in our analysis. To control for this, and to measure the market sentiment before the offering, the prior 30 day return of the S&P 500 index is used.

Industry dummy

Loughran & Ritter (2004) suggest that technology stocks see higher uncertainty. High uncertainty, or riskier IPOs, have according to Ritter (1987) indicated higher levels of underpricing, than firms with less risk. In addition, Hunt-McCool, Koh & Francis (1996) further evaluate six industries in which underpricing is more prominent, one of them being the healthcare industry. Therefore, we create a dummy variable that will take the value '1' if the issuing firm is categorized to belong to the technology or healthcare industry, and '0' otherwise.

Proceeds amount

Carter, Dark & Singh (1998) hypothesize that larger IPOs will in general be issued by more established firms, subsequently risk should be lower as well as the level of underpricing. The study by Pirayesh Neghab, Bradrania & Elliott (2023) suggests that proceeds amount was the most effective variable in their estimation of deliberate premarket underpricing. Proceeds amount is defined as expected proceeds amount excluding overallotment, measured in millions USD.

Offer price

Firms often manipulate their stock price by setting specific prices in IPOs, executing share repurchases and stock splits. Previous research such as Roger, Bousselmi, Roger, & Willinger (2018) find that stock price levels exhibit a significant impact on investor behavior. Furthermore, Baker & Powell's (1993) paper on managerial motives for stock splits suggest that the primary motive for stock splits is to position the price into a better trading range in order to increase

trading liquidity. Weld, Michaely, Thaler, & Benartzi (2009) found that the average nominal price of a share of stock on the New York Stock Exchange has stayed relatively stable, hovering around $35, since the Great Depression. In essence, the literature suggests that stock price levels influence the trading of a stock. In the IPO process, the offer price will equal the opening price in which the newly issued shares trade at. The inclusion of the offer price variable takes into consideration how behavioral factors might influence IPO underpricing.

IPO Underpricing (dependent variable)

The initial first day return is what IPO underpricing is defined as. It is defined as the percentage change between the offering price and the closing price of the first day of trading given by the following equation:

$$Underpricing = \left(\frac{Closing\ Price_{T=1} - Offer\ Price_{T=1}}{Offer\ Price_{T=1}}\right) \cdot 100.$$

3.2 Software

All data handling is performed in Python in the source-code editor Visual Studio Code. For the modeling, we apply the SkLearn library within the broader Sci-Kit library which is a simple and efficient tool for predictive data analysis. It includes multilinear regression and classifier, neural network regression and classifier, random forest regression and classifier and gradient-boosting tree regression and classifier. For each of the regression models, there are a set of incorporated hyperparameters that can be changed to optimize the performance of the models. For performance evaluation, we employ several standard routines in the package SkLearn to calculate statistical metrics such as adjusted R-squared, Mean-Squared-Error as well as precision, recall, F1-score and prediction accuracy.

4. Methodology

Our methodology consists of four steps. First, we create a second set of data, in which the variables ‘offer price’, ‘ROA’ and ‘proceeds amount’ are log transformed, using the natural logarithm (see Appendix I). Second, we split the two data sets, both containing 577 observations, into non-overlapping training and testing sets for the regression and classification models. Third, we run the multilinear regression and classifier (MLR and MLC respectively) on the training set containing the log transformed variables. Followingly, we train a neural network regression and classifier (NNR & NNC), random forest regression and classifier (RFR & RFC) and a gradient-boosting tree regression and classifier (GBR & GBC) on the original training set containing no log transformed variables. Fourth, we use the trained models to make out-of-sample predictions on the remaining observations in the test sets to measure the performance of each regression and classification model.

4.1 Regression & Classification

The regression models measure IPO underpricing as a continuous variable, while the classification models aim to measure IPO underpricing as a binary variable. That is, the classifier is given a threshold for the dependent variable initial returns, and all IPOs that are predicted to be greater than the threshold are classified as ‘good IPOs’ while values predicted below the threshold are classified as ‘bad IPOs’. Suppose the regression model predicts an underpricing of 1% whilst the actual underpricing was at 10%, then the statistical accuracy would be low and the regression model would be evaluated as inadequate. However, seeing as though it is indicating a good IPO, an investor might want to invest in such a deal. This is another reason for implementing a classification task in addition to the regression task.

4.2 Generation of Data sets

4.2.1 Log Transformations

One of the objectives of this study is to examine whether ML models are better than linear models at predicting and classifying IPO underpricing. Linear and non-linear regressions are based on different sets of assumptions, therefore, to avoid restricting any model's performance, the data sets have been altered slightly to fit the respective model's assumptions. Among other assumptions, the linear model assumes that the residuals are normally distributed. A common practice by previous scholars is to transform certain variables that do not fulfill this assumption by using a logarithmic scale.[4]

Following tests of the variables (see appendix I), results show that ROA, proceeds amount and offer price should be log transformed. Though the results of these tests are weak, there is enough underlying significance to justify the performed transformations. To level the playing field for each of the models, an additional data set solely for the linear model was created, in which ROA, proceeds amount and offer price have been log transformed. Seeing as though the variable age contains zero-values, we measure firm age in *years begun* by adding +1 to all observations of the age variable. This is done to remedy against the mathematical issue of taking the log transformation of zero.

4.2.2 Test & Training Data

Data is divided into training and testing data: 80% is used for training whilst 20% is used for testing the models. This amounts to 461 training data points and 116 testing data points. Therefore, the training sets consist of a 461 x 7 matrix and a 461 x 1 vector while our testing sets consist of a 116 x 7 matrix and a 116 x 1 vector, for the input and output variables respectively, which is the same for each of the non-linear regression and classification algorithms. Similar

[4] West (2022) suggests that the purpose of transforming data is to conform the data towards following a relatively symmetric distribution. Furthermore, the altered distribution does not have to be entirely normal, but if it happens to be, it could boost confidence in tests with smaller sample sizes and potentially streamline the process of statistical modeling.

matrices and vectors are constructed for the linear models, with the difference that the variables ROA, proceeds amount and offer price are log-transformed.

4.3 Model Training

From the SkLearn library there are predetermined functions and corresponding hyperparameters that can be altered for each model. The hyperparameters and their values have been tested through trial and error to minimize the MSE and maximize the accuracy and R-squared score. All hyperparameters started from default values which are regular starting points when adjusting such models, and were therefore changed to find the desired outcome. When optimizing hyperparameters, a parameter grid is created in which GridSearchCV is used to iterate over a multitude of different hyperparameter values to find those that maximize R-squared and minimize MSE.[5] Within each regression model, there are several alternatives for each hyperparameter that are structured in various ways, meaning that they are not arbitrarily chosen, and are instead somewhat default. The exact values used for each hyperparameter are organized in Appendix II.

4.3.1 Multilinear Regression (MLR)

MLR is a simple statistical method that models the relationship between independent and dependent variables. Typically, this linear model takes the form:

$$Y_i = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ \ldots \ + w_n \cdot x_n = w_0 + \sum_{i=1}^{n} w_i \cdot x_i .$$

This multilinear regression model aims to find the coefficients $w = (w_1, w_2, \ldots, w_n)$ that minimizes the sum of squared differences between the observed values in the data set and the predicted values. The objective function to be minimized is the mean-squared-error (MSE), which provides the optimal values for the coefficients that define the linear relationship between the input and output variables.

[5] See documentation from SKLearn
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

4.3.1 a) Linear Regression Assumptions

MLR and MLC rely on several assumptions to be valid models, namely the Gauss - Markov assumptions for linear regression models. Firstly, they assume a linear relationship between the dependent variable and each independent variable, Secondly, independence between independent variables is needed to avoid multicollinearity issues, assessed through methods like the Variance Inflation Factor. Thirdly, a constant variance of residuals (homoscedasticity) is examined by statistical tests such as White's test as well as the residuals being normally distributed. Fourthly, the independence of observations, where the Durbin Watson statistic helps identify autocorrelation in residuals. These assumptions form the foundation for reliable interpretations in linear regression analysis. The results (see Appendix I) do not entirely support the Gauss - Markov assumptions for linear regression. However, the data set and models were altered to optimize conditions, specifically for linear models, to provide a fair evaluation between all models.

4.3.2 Neural Network (NN)

A NN is a ML algorithm inspired by the human brain structure. It is made up of different layers including the input, hidden and output layer depending on the parameter's weight and bias. The connections between the nodes are associated with biases and weights such that each node applies an activation function (non-linear) to its inputs. This method is divided into forward and backward propagation in which data is fed through the network producing predictions. Thereafter, the model is trained where the weights and biases are adjusted, to minimize the loss. The loss function varies depending on the data set (and its distribution) and is minimized by setting its partial derivatives, with respect to the different weights and biases, to zero and solving the system of equations. A typical NN takes the following form:

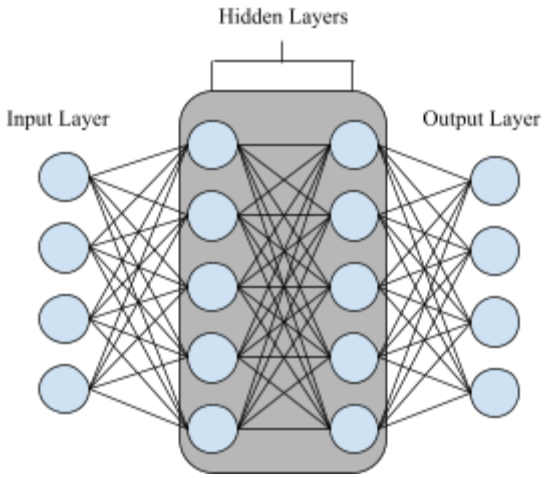


Figure 1. Neural network structure

4.3.3 Random Forest (RF)

A RF is a powerful ML algorithm used for regression and classification tasks, meaning that the output is either a continuous or binary variable. It operates by creating multiple decision trees, each trained on different subsets of data and features. When making predictions, each tree "votes", and the final prediction is determined by the majority. The two primary parameters that are predetermined when performing a RF are the number of trees (x) and the maximum depth (y) of each decision tree. This means that the ML algorithm will create x independent trees, all of depth y, and find the average prediction of each tree to predict the dependent output variable. Such a tree with x = 2 and y = 3 may look like:

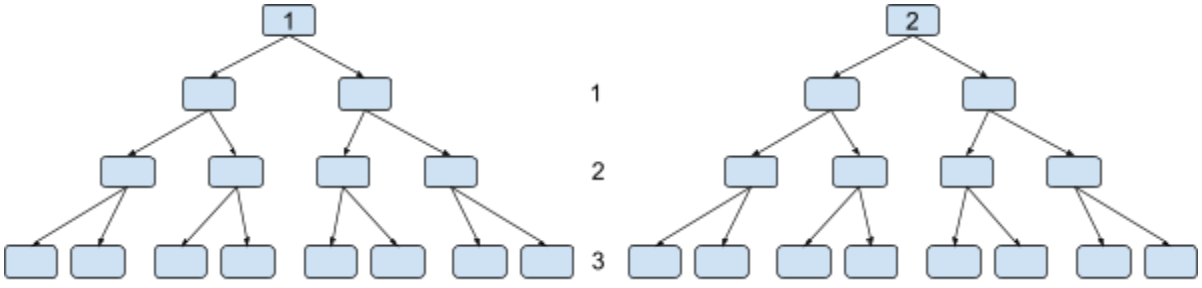


Figure 2. Random forest structure

4.3.4 Gradient-Boosting Trees (GBT)

GBT is similar to the RF in terms of it creating ‘n’ trees to create a prediction. However, in the RF model ‘n’ *independent* trees are created to then calculate the mean of each prediction whereas the GBT builds upon the results of the previous tree to create the next tree. Each tree is used to create a prediction whether an underpricing occurs by building a model that minimizes the loss function. The loss function, oftentimes MSE, is once again varied depending on the data set and what type of distribution it generally follows.

4.4 The Classification Task

This study includes a classification task based on prior positive findings favoring binary classification in financial data modeling (Enke & Thawornwong, 2005). The data set’s skewed distribution (see Appendix IV) of initial returns, with a mean of 21.27% and a median of 10.76%, suggests that the majority of IPOs are underpriced, which will impact the results and prediction accuracies of this study. Arriving at this, two thresholds were set, 0% and 10%, to classify IPOs as either good or bad. The weak threshold will primarily measure whether an IPO is deemed good (underpriced) or bad (not underpriced). However, under the weak threshold (0%), a majority of observations will be above 0%, causing a class imbalance that potentially interferes with the correct classifications of good and bad IPOs. The strong threshold will be approximately set towards the median, thereby mitigating any issues that the class imbalance might cause. Adjusting the thresholds offers valuable insights into different prediction scenarios, considering the data’s skewness. Furthermore, the two hypotheses will provide some economic intuition to the discussion on how these types of classifications of IPOs could be adopted by the potential investor.

4.4.1 The Weak Threshold

For the weak threshold, the threshold value for the classifier is set to 0% initial returns. This means that the output variable is classified as ‘1’ (good IPO) when initial returns are greater than 0%, and classified as ‘0’ (bad IPO) when the value is equal to or less than 0%. This specific

threshold value is established to test the theoretical efficiency of the model in categorizing IPOs as either underpriced or not underpriced. This is considered a weak threshold since it classifies all IPOs generating a gross return greater than 0% as good. However, in practice, financial frictions such as courtage costs and tax payments may reduce net earnings. Thus, the net returns from the IPO can be significantly reduced if the gross returns are considerably low from the beginning. To remedy against this real-world risk, the threshold must be increased, which is partly why the strong threshold is introduced.

4.4.2 The Strong Threshold

For the strong threshold, the threshold value for the classifier is set to 10% initial returns. This means that the output variable is classified as 1 (good IPO) when initial returns are greater than 10%, and classified as 0 (bad IPO) when the value is equal to or less than 10%. This threshold value is close to the median of the data set (10.7%) meaning that the data is now more centered with almost as many values below and above the threshold, thus mitigating any potential issues that a class imbalance might cause. Furthermore, this is interesting as it gives an indication of how the models perform when data is centered, making the classification more difficult. Here, all costs associated with the IPO transaction can be mitigated as the first day return will, with almost all certainty, exceed the costs for each allocation.

4.5 Statistical Tests

For each algorithm a model has been trained on the training data, consisting of 80% of the entire data set. These models are then used to test the remaining 20% of the data set and compare the modeled/predicted results with the observed/actual values in addition to a handful of statistical tests.

4.5.1 Adjusted R-squared

The adjusted R^2 (R-squared) test is an adjustment of the statistical R-squared test. The R-squared test analyzes how well the independent (input) variables in a regression model predict the variability in the dependent (output) variable. In other words, it gives the proportion of variance in the dependent variable that is predictable from the independent variables. This value

typically ranges from 0 to 1 where 0 indicates that the model does not predict the dependent variable at all and a value of 1 indicates that the model perfectly predicts the dependent variable. In practice, the R-squared gives an indication of how well the model fits the data. Mathematically, it can be calculated as:

$$R^2 = 1 - SSR/SST,$$

where $SSR = \sum(y_i - \hat{y}_i)^2$ represents the residual sum of squares and $SST = \sum(y_i - \bar{y})^2$ represents the total sum of squares. In this case, y_i represents the observed values, $\hat{y}_i$ represents the fitted/modeled values and $\bar{y}$ is the mean of the observed values. To calculate the adjusted R-squared, which takes into account the number of variables in the model, the following equation is used:

$$Adjusted\ R^2 = 1 - (1 - R^2) \cdot \frac{(N-1)}{(N-p-1)},$$

where N is the sample size (of the test data) and is the number of parameters in the model.

4.5.2 Mean-Squared-Error (MSE)

MSE provides a measure of how well the model predicts the output variable. It gives an indication of how far off the predicted values are from the actual values whilst penalizing larger errors more heavily. In this case it is important to understand that dealing with investments is of risky nature and therefore it is reasonable that greater deviations would have a greater impact on the overall score. The MSE is calculated through the following equation:

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 = \frac{1}{n} \cdot SSR$$

4.5.3 Precision, Recall & F1-score

The classification is structured so that binary variable 1 defines a good IPO, and binary variable 0 defines a bad IPO. Precision is calculated as the number of true positive predictions divided by

the total number of positive predictions made by the model, which is the sum of true positives and false positives. A high precision score, for binary variable 1, indicates that the model is good at avoiding false positives; that is, not mistakenly classifying IPOs as underpriced when they in fact are not underpriced. Seeing as we are looking at possible investment decisions, it is imperative that a false positive is not particularly prevalent in the models, indicating the importance of aforementioned tests in addition to the ordinary prediction accuracy measurement. Precision is calculated by:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall measures the ratio of correctly predicted positive instances to the total actual positive instances in the data set. A high recall score indicates that the model is good at identifying most of the positive instances in the data set without missing many positive instances. Recall is calculated by:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The F1-score is a comparative metric used in classification tasks to provide a balance between precision and recall, calculated from the harmonic mean of the two. The F1-score ranges from 0 to 1, with higher values indicating better performance. The formula for F1-score is:

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4.5.4 Prediction Accuracy

Prediction accuracy is used to evaluate the performance of the different classification models on the test data set. It represents the proportion of correctly classified instances out of the total number of instances in the test set, for both of the binary variables. The formula for prediction accuracy is:

$$Prediction\ Accuracy\ =\ \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

5. Results

This section consists of three parts: regression results, regression statistical results and classification results. The models are trained using the data from the training set, and the results are produced when the models perform predictions on the test set. The regression results are presented in a regression plot and a histogram of residuals. The statistical results from the regressions are compiled in a table to present the statistical test values from the regression models. Additionally, the classification results are presented in tables, featuring metrics such as precision, recall, F1-score, and prediction accuracy.

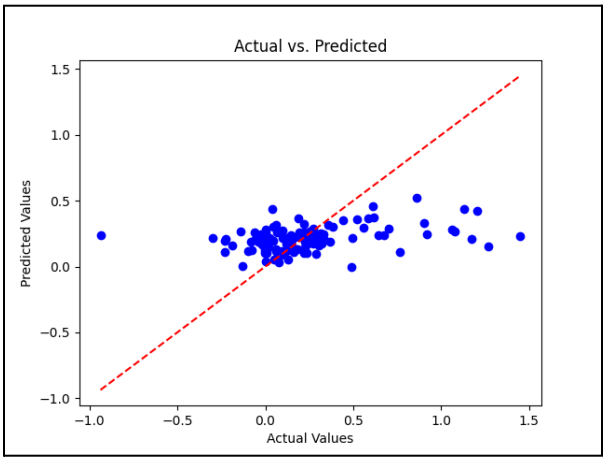
5.1 Regression Model Plots

5.1.1 Multilinear Regression

FIGURE 3 & 4

REGRESSION MODEL PLOTS

Figure 3 portrays the MLR's predicted values against the actual observed values. The red line exhibits where predicted values are the same as the actual observed values, indicating perfect predictions. Figure 4 shows a histogram of the residuals which portrays the distribution of the residuals and potential outliers.



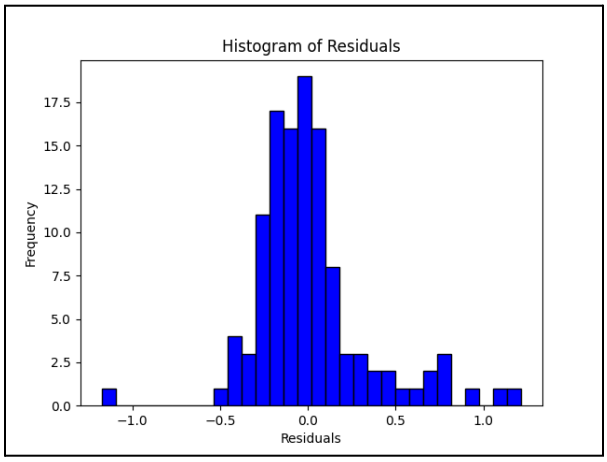


Figure 3: Actual vs Predicted MLR

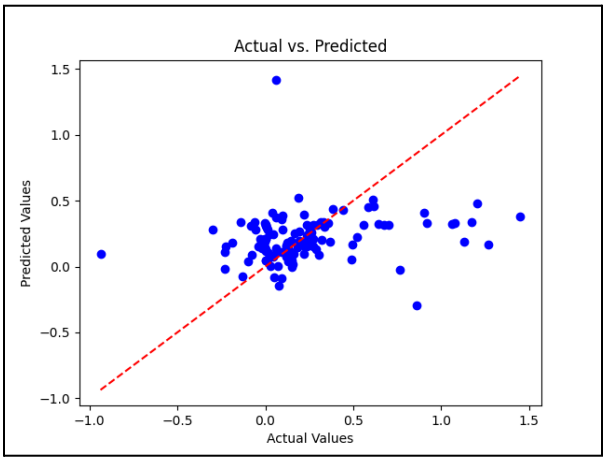
Figure 4: Histogram of Residuals MLR

5.1.2 Neural Network Regression

FIGURE 5 & 6

REGRESSION MODEL PLOTS

Figure 5 portrays the NNR's predicted values against the actual observed values. The red line exhibits where predicted values are the same as the actual observed values, indicating perfect predictions. Figure 6 shows a histogram of the residuals which portrays the distribution of the residuals and potential outliers.



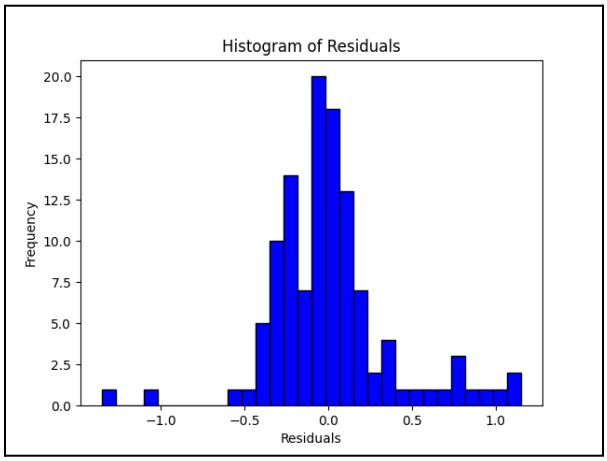


Figure 5: Actual vs Predicted NNR

Figure 6: Histogram of Residuals NNR

5.1.3 Random Forest Regression

FIGURE 7 & 8

REGRESSION MODEL PLOTS

Figure 7 portrays the RFR's predicted values against the actual observed values. The red line exhibits where predicted values are the same as the actual observed values, indicating perfect predictions. Figure 8 shows a histogram of the residuals which portrays the distribution of the residuals and potential outliers.

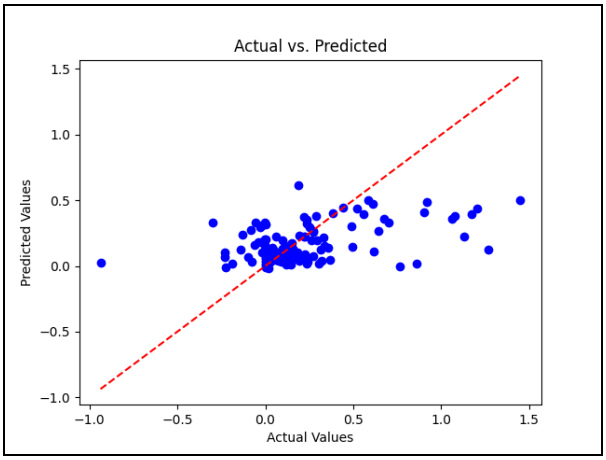


Figure 7: Actual vs Predicted RFR

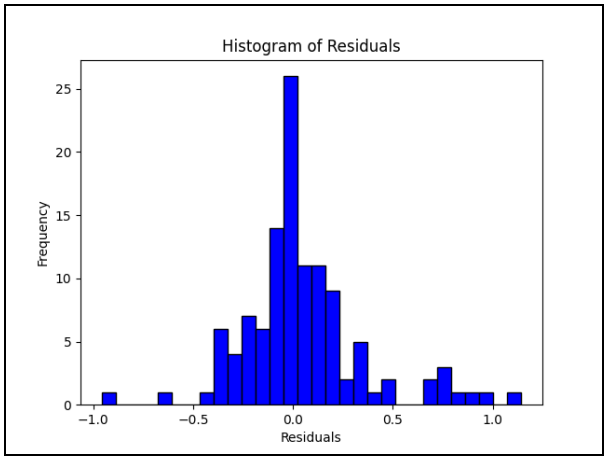


Figure 8: Histogram of Residuals RFR

5.1.4 Gradient-Boosting Tree Regression

FIGURE 9 & 10

REGRESSION MODEL PLOTS

Figure 9 portrays the GBR's predicted values against the actual observed values. The red line exhibits where predicted values are the same as the actual observed values, indicating perfect predictions. Figure 10 shows a histogram of the residuals which portrays the distribution of the residuals and potential outliers.

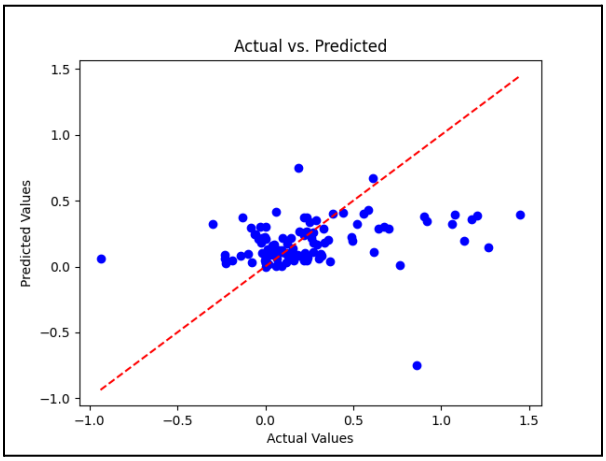


Figure 9: Actual vs Predicted GBR

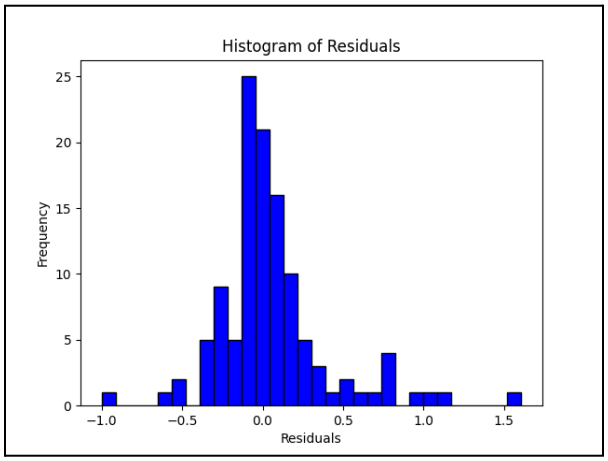


Figure 10: Histogram of Residuals GBR

5.1.5 Regression Plots Patterns

Figure 3 demonstrates the unsatisfactory performance of the MLR, predominantly predicting values between 0% and 50%, likely influenced by the data set's average underpricing. However, its prediction accuracy diminishes for values beyond this range. Figure 4 partly supports the linearity assumption, with residuals being approximately normally distributed. Yet, some predictions exhibit significant deviations of over 100 percentage points. In conjunction with this, the correlation matrix for the MLR (see Appendix III) indicates that most variables have relatively low correlations with the dependent variable. Low correlations indicate that the magnitude of the values for the inputs will have a low effect on the modeled dependent variable. This explains why there is a low spread in the predictions. Retrieving the equivalent to coefficients from ML models is a complex endeavor since they operate in a non-linear sense, but given the closely similar results (Figures 5, 7 & 9) it is reasonable to believe that the independent variables have low explanatory significance towards the dependent variable in the ML models as well. Similar to the MLR, the NNR (Figure 5) predicts most values within the same range, with a slightly broader spread. The residuals for the NNR (Figure 6) also approximate a normal distribution, while displaying outliers of over 100 percentage points. The RFR (Figure 7) aligns with prior models, largely predicting values within the range of 0% to 50%, presenting a poor fit visually. However, Figure 8 indicates better-centered residuals around 0, displaying a distribution approximating normality. The GBR (Figure 9) visually resembles other models in addition to the residuals (Figure 10) once again being approximately normally distributed.

In essence, it is evident that all models primarily forecast IPO underpricing within the span of 0% to 50%. All models demonstrate similar visual performances with some minor variations in residual distributions, for instance, the GBR had an extreme outlier of over 150 percentage points. All models struggle to accurately forecast values outside the aforementioned range, with multiple residuals reaching over 100 percentage points. Overall, there is a general pattern of normally distributed residuals which is sought after when conducting regression analysis, particularly for the linear model.

5.2 Regression Results

TABLE 2

STATISTICAL RESULTS FOR THE REGRESSION MODELS

The following table depicts the results for the statistical tests conducted for the regression models. It represents the results from applying the trained model on the remaining test data and analyzing the differences between the predicted and actual values. Adjusted R-squared measures the proportion of variation in the dependent variable explained by the model while adjusting for the number of parameters, whereas Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values, reflecting the model's predictive accuracy.

Method	**MLR**	**NNR**	**RFR**	**GBR**
MSE	0.1101	0.1327	0.1003	0.1235
Adjusted R-squared	0.0735	-0.1160	0.1562	-0.0392

From Table 2, in respect to MSE, the RFR is the most accurate model. The second best model is the MLR, third best is the GBR and the least accurate model is the NNR. However, the adjusted R-squared results exhibit greater variability among the models, which demonstrates that the RFR is the best performing regression model. To put the results into context, Butler et al. (2014) conducted regression analysis on a sample of U.S. IPOs spanning from 1981 through 2007, achieving an adjusted R-squared of 45.5%. In comparison to their study, all the models in this study demonstrate unsatisfactory performance. The regression plots, histograms and Table 2 suggest that the chosen independent variables in this study have a limited explanatory power in describing the variation of the dependent variable. The low adjusted R-squared values coupled with relatively high MSE values across all regression models imply an insufficiency in predicting IPO underpricing as a continuous variable. Thus, the results fail to support the notion that IPO underpricing is predictable through a regression task, using the subset of publicly available variables.

5.3 Classification Results

5.3.1 Classification Results for the Weak Threshold

TABLE 3

CLASSIFICATION RESULTS FOR THE WEAK THRESHOLD

The following table depicts the results from the classification tasks. The classifier has been given a threshold based on the dependent variable initial returns. The classifier is set to take on the variable '1' (good IPO) if the IPO's estimated initial return is greater than the threshold value. Additionally, the classifier is set to take on the variable '0' (bad IPO) if the IPO's estimated initial return is equal to, or less than the threshold value. The weak threshold is set to 0%. The performance of the classification models is measured through precision, recall, F1-score and prediction accuracy. Precision is calculated as the number of true positive predictions divided by the total number of positive predictions made by the model, which is the sum of true positives and false positives. Recall measures the ratio of true positive instances to the total actual positive instances in the data set, which is the sum of true positives and false negatives. The F1-score is calculated from the harmonic mean of precision and recall. Prediction accuracy is the proportion of correctly classified instances out of the total number of instances in the test set. A true positive/negative depends on which binary variable is analyzed. True positive for '1' indicates the correct prediction of a good IPO. True positive for '0' indicates the correct prediction of a bad IPO.

	Precision		**Recall**		**F1-score**		**Prediction Accuracy**
	0	**1**	**0**	**1**	**0**	**1**	
MLC	0.00	0.78	0.00	1.00	0.00	0.87	0.78
NNC	0.00	0.77	0.00	0.96	0.00	0.85	0.74
RFC	0.30	0.78	0.12	0.92	0.17	0.85	0.74
GBC	0.19	0.77	0.15	0.81	0.17	0.79	0.66

5.3.2 Classification Results for the Strong Threshold

TABLE 4

CLASSIFICATION RESULTS FOR THE STRONG THRESHOLD

The following table depicts the results from the classification tasks. The classifier has been given a threshold based on the dependent variable initial returns. The classifier is set to take on the variable '1' (good IPO) if the IPO's estimated initial return is greater than the threshold value. Additionally, the classifier is set to take on the variable '0'

(bad IPO) if the IPO's estimated initial return is equal to, or less than the threshold value. The strong threshold is set to 10%. The performance of the classification models is measured through precision, recall, F1-score and prediction accuracy. Precision is calculated as the number of true positive predictions divided by the total number of positive predictions made by the model, which is the sum of true positives and false positives. Recall measures the ratio of true positive instances to the total actual positive instances in the data set, which is the sum of true positives and false negatives. The F1-score is calculated from the harmonic mean of precision and recall. Prediction accuracy is the proportion of correctly classified instances out of the total number of instances in the test set. A true positive/negative depends on which binary variable is analyzed. True positive for '1' indicates the correct prediction of a good IPO. True positive for '0' indicates the correct prediction of a bad IPO.

	Precision		**Recall**		**F1-score**		**Prediction Accuracy**
	0	**1**	**0**	**1**	**0**	**1**	
MLC	0.54	0.67	0.63	0.58	0.58	0.62	0.60
NNC	0.57	0.70	0.67	0.60	0.61	0.64	0.63
RFC	0.55	0.70	0.71	0.54	0.62	0.61	0.61
GBC	0.57	0.68	0.61	0.65	0.59	0.66	0.63

5.3.3 Prediction Accuracy

Tables 3 and 4 show the results of the classification models and their corresponding thresholds. The weak threshold and strong threshold each have different values for what constitutes a 'good IPO' or a 'bad IPO' based on its predicted level of initial returns. Given the positive skewness of the dependent variable, and the class imbalance, it is particularly interesting to see how the performance of the models changes as the threshold is altered since it affects the distribution of the class balance.

When initially reviewing the models' prediction accuracy under the two thresholds, it becomes evident that each model performs worse under the strong threshold, in which the threshold approximately equals the median. This is in line with what was theorized in section 4.4.2, that is, the skewness of the distribution in the dependent variable will increase the complexity of the classification task. This influenced the prediction accuracy and overall performance of the

models. The variability in prediction accuracy across different models and thresholds also suggests that no single model is best for all scenarios, and the choice of model may depend on the specific context and requirements of the IPO prediction task.

From Table 3 it is evident that MLC, NNC, and RFC exhibit the highest prediction accuracies. Specifically, the MLC achieves the highest prediction accuracy of 0.78. Nevertheless, prediction accuracy can display an incomplete picture, especially in scenarios with imbalanced classes, where one class outnumbers the other. In these instances, a model might achieve a seemingly high accuracy by consistently predicting the dominant class, neglecting to recognize the minority class (bad IPOs) altogether. This seems to be the case for the MLC, as it shows 100% accuracy in predicting good IPOs but misclassifies all bad IPOs as good, indicating a tendency to label everything as a good IPO. The NNC follows a similar trend to the MLC but demonstrates more flexibility in its classifications, identifying data as both good and bad IPOs. This is supported by the NNC's recall of 0.96 for the binary variable '1', indicating a few false negative predictions.

Considering the MLC's 0.78 prediction accuracy, and its stationary classification of good IPOs (the weak threshold) one can assume that the class imbalance has strongly affected the models. Furthermore, the results of the MLC suggest that roughly 78% of the test data set may have initial returns greater than 0%. This is in line with the observed distribution of the dependent variable (see Appendix IV). One could argue that all models under the weak hypothesis with a prediction accuracy below 0.78 do not show any predictive capability, as it does not beat random chance, where random chance is to classify everything as a good IPO (underpriced).

Upon analyzing Table 4, a significant improvement is observed in the models' capacity to classify bad IPOs, since the measures regarding the binary variable '0' have increased significantly. This enhancement is likely due to adjusting the threshold closer to the median, resulting in an almost equal distribution of classes above and below the threshold. The results of the strong threshold underscore that the ML-based models gain traction and perform marginally better than the MLC. This seems reasonable, as when moving the threshold closer to the median, the complexity of the classification task should increase. In Table 4, the NNC and the GBC have the highest prediction accuracy values, but their F1-scores differ. NNC is modestly better at

predicting bad IPOs, while the GBC is modestly better at predicting good IPOs, thus, neither model can be considered superior. Given the balanced distribution, one could argue that a prediction accuracy score above 0.50 supports predictability, as random chance becomes approximately 0.50. The results under the strong hypothesis thereby support that it is possible to predict and classify IPO underpricing using a subset of publicly available variables prior to the offering, albeit the level of predictability is moderate.

5.3.4 Precision, Recall and F1-Score

When the data is imbalanced, or when the impact of false negatives or false positives are not equal, other metrics such as precision, recall and F1-score provide notable insights. There is no universal rule of thumb for what is considered a good F1-score, as it depends on the context and what topic is being studied. However, F1-scores above 0.50 are considered acceptable, and above 0.70 are generally considered good.[6]

From the weak threshold it is difficult to provide a definitive answer regarding the models are able to predict and classify IPO underpricing with the given variables. The reason is because the prediction accuracy is flawed by the imbalanced class distribution. Additionally, measuring predictiveness via precision, recall and F1-scores is an ambiguous task, as they are dependent on subjective interpretations. While the MLC produces the highest prediction accuracy, its zero-sum values of precision and recall when labeling bad IPOs imply that this model is not reliable. Under the weak threshold, the RFC exhibits the same prediction accuracy as the NNC, yet its F1-score for predicting bad IPOs is higher. Deciding which model has the most accurate performance under the weak hypothesis will thus depend on whether an investor values predicting a good IPO higher than predicting a bad IPO, which fundamentally has to do with the level of risk aversion. Under the weak threshold, many of the models have superior results for classifying good IPOs, but fail in classifying bad IPOs.

[6] More information about class imbalance and the F1-score: https://spotintelligence.com/2023/05/08/f1-score/?fbclid=IwAR1i1Z3xrMOTGSjwMNzinQom_5_DZpqKz-RypNccrp10S5rQJVsDvZcAdqs#:~:text=The%20F1%20score%20ranges%20from,for%20the%20binary%20classification%20task

Under the strong threshold, all models perform similarly well. The NNC and RFC exhibit the highest accuracy measures for both good and bad IPOs. They are consistently over 0.6 for both binary variables, but do not meet the threshold of 0.70 for the F1-score. Once again, it is difficult to give a precise answer as to which model is preferred as it once again depends on the type of investor at hand.

5.4 Summary of Results

5.4.1 Regression Results

Section 5.1 exhibits the results of each model's regression. Figures 3, 5, 7 & 9 indicate that, on average, all models predict initial returns to be within the span of 0-50%, which likely is related to low correlations between the independent variables and the dependent variable. Visually, it seems that the residuals of all models approximately follow a normal distribution. From Table 2, the adjusted R-squared and MSE values indicate that RFR is the best performing model, followed by the MLR. The results of the regression analysis fail to support that IPO underpricing can be predicted using the given subset of pre-IPO variables. However, the results indicate that one of the ML models (RFR) performs better than the linear model (MLR) in predicting IPO underpricing when it is treated as a continuous variable. On the other hand, the MLR performs better than two out of three machine learning models, the NNR & GBR, which contradicts the hypothesis that ML-based models would perform better than linear models when predicting IPO underpricing.

5.4.2 Classification Results

The weak threshold will at first glance indicate that all models succeed in predicting IPO underpricing as they all achieve prediction accuracies well above 0.50. Moreover, they show that the MLC is better at predicting IPO underpricing than ML-based methods. Yet, given the fact of the imbalanced class distribution, and that the MLC is static in its classification, the initial observation is rendered inaccurate. Therefore, providing a determinate answer under the weak threshold is ambiguous, as the interplay of several statistical measures should be accounted for.

Furthermore, the results of the strong threshold support that IPO underpricing is, to some degree, predictable through classification with the given variables. Thus, the research question is only supported when IPO underpricing is measured as a binary variable in classification tasks, given a threshold of 10%. Summarized, IPO underpricing is only predictable using a subset of pre-IPO data under the strong hypothesis, and ML-based models exhibit marginally better performance in predictive classification tasks.

6. Discussion

This section aims to explore the practical implications of these models for potential investors, alongside a comparison between the performance of ML-based models and linear models. The objective is to thoroughly analyze and contextualize the results, offering a deeper understanding and interpretation of the findings.

6.1 Practical Implications for an Investor

To put the results into context, an investment perspective is applied to the discussion, focusing on how a potential investor might interpret the results and subsequently value the models. Consequently, the relatively poor performance of the regression models reduce their relevance for this discussion. Followingly, this part will focus on how the potential investor would evaluate the results of the classification results.

It is well established that the average IPO is underpriced. For instance, if the IPO market functioned similarly to the stock market, then an investor who ”bought the market” could generate the same returns as the average underpricing of all IPOs. The fundamental difference between a regular stock market and the IPO market is that in a stock market, security prices are set where supply meets demand. This is not the case for the IPO market. IPOs are generally priced by underwriters, who in turn attempt to estimate the demand for the offered shares. Investment banks generally set this price below the equilibrium price, causing the market to become underpriced on average. This is adjusted for during the first day of trading, subsequently,

IPO shares will on average generate positive initial returns on average. Assuming this theory holds, why is researching the topic of predicting IPO underpricing relevant, seeing as buying the market could guarantee a profit?

The reason is spelled share allocations. Theories like the winner's curse by Rock (1986) argue that good IPOs will experience a higher demand, as both informed and uninformed investors will subscribe to them. On a regular stock market, prices would adjust until settling in equilibrium. However, IPOs experience a fixed offer price, possibly resulting in demand exceeding supply. Rock argues that this will be the case for good IPOs, which in turn results in the crowding out of both informed and uninformed investors. On the contrary, Rock suggests that a winner's curse emerges when an uninformed investor subscribes to a bad IPO, as the low demand will result in the investor receiving a full allocation. This gives rise to a situation where a potential investor buying the market would not be able to capture the average underpricing of the market, as the potential investor will become crowded out in the good IPOs and win the full allocation of bad IPOs.

Seeing as an investor in theory only receives parts of the good IPOs, but a full allocation of the bad IPO shares it becomes imperative to avoid bad IPOs. From the results of the weak threshold, it seems that the MLC performs the best but delving deeper into the results shows that it classifies everything as a good IPO (see 5.3.3). Having a model classify everything as a good IPO would be similar to buying the market. However, as stated above, great value lies in the potential investors' ability to avoid subscribing to the bad IPOs, given the theory of the winner's curse. In light of this theory, buying the market is not a sound approach to investing in IPOs. Essentially, being able to differentiate between good and bad IPOs and only subscribing to the former would largely facilitate the navigation of the treacherous IPO landscape.

If having a prediction accuracy below 0.78 under the weak threshold is not better than buying the market, and seeing that the ML models have a prediction accuracy below this, is the performance of the ML models worthless? The answer is twofold. Yes, the prediction accuracy of the ML models under the weak threshold indicate that they will be wrong in more instances, than if one would buy the market. However, this isn't necessarily a fair measure for the models, as their

power is derived from their flexibility to predict and classify both good and bad IPOs. Subscribing to bad IPOs could be detrimental for the potential investor, therefore, the ability of the models to classify bad IPOs must be given fair consideration. The results from both thresholds underscore the abilities of the ML-based models to classify both good and bad IPOs to a greater extent than the linear model. Consequently, a potential investor seeking accuracy in classifying both '0' and '1' instances would lean towards ML models as the preferred choice.

Under the weak threshold, an IPO classified as good may still capture minimal underpricing, potentially resulting in negative *net* profits. The tradeoff between high prediction accuracy and the ability to detect bad IPOs will influence the potential investors' choice of the optimal model. Given that each investor holds different assumptions, there is no single model in this study that is optimal for each individual. Therefore, given the assumptions of each investor an individual investigation is needed to pinpoint the "optimal strategy" —a balance where crowding out, costs, and classification accuracy converge to maximize returns for a potential investor.

Other aspects to consider are that even if it were possible to create a model that is 100% accurate in its predictions, share allocation theories would limit the value of such a model. Share allocation theory posits that IPOs are intentionally underpriced because when an excess demand occurs, investment banks manage the allocation of the subscription rights. The investment banks tend to allocate shares to large institutions and significant clients with whom they wish to maintain a favorable relationship. Consequently, while predicting underpriced IPOs may be feasible, volume and allocation constraints significantly curtail the practical application of these models in real-world contexts, such as investment strategies.

6.2 Linear vs Non-Linear Models

The subquestion in this study regarded whether machine learning models are more efficient in predicting and classifying IPO underpricing compared to traditional linear models. In this paper's regression analysis, the adjusted R-squared of RFR (15.62%) is substantially higher than that of MLR (7.35%). Even if neither is considered well performing, the results still support that when

conducting regression analysis within the field of IPO underpricing, ML-based models, particularly random forest models, are the preferable ones. The results from the classification tasks, as seen in section 5.3, support that on average, ML-based methods are preferred. At its core, depending on the nature of the task, regression or classification, and depending on the threshold for the classification task, 0% or 10%, the answer will differ on which model is preferable. However, considering both the predictive regression and classification tasks, the RFC seems to perform most satisfactorily.

Another intriguing aspect involves the marginal benefit from employing non-linear over linear models. Existing literature indicates that more intricate models, like ML techniques, can capture complex and non-linear data patterns. The findings in this study reinforce this idea, and this becomes particularly evident in the examination of the results for the two thresholds in section 5.3. When the threshold is set in a fashion that makes the classification task more complex (10%), the ML methods show greater performance, affirming the theory that ML-based models excel in complex tasks. Nevertheless, when classifying IPO underpricing, the complexities associated with ML, and the marginal differences in our results, imply that linear approaches might be preferable for research objectives. Therefore, considering the complexity involved in operating and evaluating machine learning models, their benefits and potential advantages might not necessarily outweigh the disadvantages.

7. Conclusion

In this study, our research question centered on the prediction and classification of IPOs using a subset of publicly available variables prior to the offering. We anticipated the feasibility of this prediction and hypothesized that classification models would yield more precise predictions compared to regression models. Additionally, given the complex nature of IPO data and promising findings in financial studies, we expected ML-based models to outperform linear models.

The study analyzed a data set of 577 observations from the US market spanning 2010 to 2020. Following the methodology of Krauss et al. (2016), three machine learning models were applied, namely neural networks, random forest and gradient-boosting trees, additionally, a multilinear model was applied. The examination of the regression results highlighted the favorability of non-linear regression models, particularly favoring RFR as the prime choice for modeling IPO underpricing. However, its efficacy in modeling IPO underpricing lacked significance for both research and investment purposes evident in the notably low adjusted R-squared scores. RFR exhibited the strongest performance, yet its adjusted R-squared of approximately 0.16 signifies a limited ability to accurately explain the variability of underpricing. These findings underscore RFR as the most fitting model among the four examined models when predicting IPO underpricing through a regression task.

In the classification results it became evident that the models achieved high prediction accuracies, well over 50%. However, given the class imbalance under the weak threshold,

prediction accuracy is deemed to not be an accurate measure in determining whether the models show predictable capability or not. Under the strong threshold, which is not subject to the class imbalance, the results were promising, indicating that all models have a prediction accuracy well over 50%. Thus, the results under the strong hypothesis support that our classification task is to some degree successful in predicting IPO underpricing using a subset of pre-IPO data. Nonetheless, the different classification results, discussed in the Results and Discussion sections of this paper, can be evaluated differently, depending on what the desirable outcome and focus points are.

Furthermore, the subquestion within this study—whether ML-based models outperform conventional linear models—poses a challenge for definitive determination. This study partly supports the efficiency and accuracy of non-linear ML algorithms in predicting IPO underpricing using a subset of publicly available variables in comparison to linear models. It is clear that ML algorithms are an important feature in econometrics and finance and the results support the use of models like this for future research on the topic of predicting and classifying IPO underpricing. The financial markets are complex and ever changing, and variables that have proven to have a significant influence in the past, might not hold any robustness in future predictions. Nonetheless, our findings provide a solid foundation for further research in this area and offer valuable insights for investors and other market participants seeking to better anticipate IPO underpricing predictability.

7.1 Limitations and Extensions

One issue that surfaced is how to treat the data sets for the linear vs the non-linear models. In order to maintain a systematic approach, this study aimed to use the same inputs for all four models. However, as the assumptions for linear and non-linear models differ, it would be reasonable to transform the data so that the model can make the most of the inputs. An example of this is that linear regression assumes normality, thus, variables that are skewed could for example be logarithmic transformed to better approximate a normal distribution. This process was performed, and a second data set was produced with a few log-transformed variables solely for the multilinear regression and classification tasks. However, in this case, we reasoned that

transforming the data to optimize the efficacy of the multilinear models was inherently important, as otherwise the comparison between our models would become inadequate.

The study by Butler et al. (2014) conducted regression analysis with a data set of 5382 observations ranging from 1981 to 2007 on the US market. They started with 48 variables and determined that 15 of them were the most significant in explaining the variability of underpricing. Consequently, they applied them in a regression analysis and achieved an adjusted R-squared of 45.5%. Their study not only contained a more extensive data set, but also examined a different time period, including several years of distress and volatility in the overall economic climate, compared to ours. Given these differences, further research that conducts a comparative analysis using the machine learning algorithms employed in this study on their data set could offer valuable insights and enhance the overall understanding of IPO underpricing within the literature. Furthermore, seeing that their linear regression achieved a relatively high adjusted R-squared of 45.5%, it would be interesting to see if machine learning models, using the same inputs, could generate a result greater than 45.5% for the adjusted R-square value.

Further extensions could also involve examining how to optimize the models to benefit IPO investing. With respect to financial frictions, prediction accuracy, allocation theories and the winner's curse phenomenon, it would be of interest to extend this research to a more practical standpoint. For this study to become useful in investment opportunities, it would be interesting to quantify the theories presented in this study and align them with the predictive models to achieve a comprehensive understanding of portfolio construction and risk management. Ultimately, the objective, to render this research practically useful in investment contexts, lies in translating the findings of this study into insights that can guide future research regarding the predictability of IPO underpricing.

References

Aggarwal, R., Bhagat, S. & Rangan, S. 2009, "The Impact of Fundamentals on IPO Valuation", *Financial Management,* Vol. 38(2), pp. 253-284.

Aggarwal, R., Prabhala, R.N. & Puri, M. 2002, "Institutional Allocation in Initial Public Offerings: Empirical Evidence", *The Journal of Finance,* vol. 57, no. 3, pp.1421-1442.

Baker, H.K. & Powell, G.E. 1993, "Further Evidence on Managerial Motives for Stock Splits", *Quarterly journal of business and economics*, vol. 32, no. 3, pp. 20-31.

Beatty, R.P. & Ritter, J.R. 1986, "Investment banking, reputation, and the underpricing of initial public offerings", *Journal of Financial Economics,* vol. 15, no. 1, pp. 213-232.

Booth, J.R. & Chua, L. 1996, "Ownership dispersion, costly information, and IPO underpricing", *Journal of Financial Economics,* vol. 41, no. 2, pp. 291-310.

Brennan, M.J. & Franks, J. 1997, "Underpricing, ownership and control in initial public offerings of equity securities in the UK", *Journal of financial economics,* vol. 45, no. 3, pp. 391-413.

Butler, A.W., Keefe, O.C. & Kieschnick, R. 2014, "Robust determinants of IPO underpricing and their implications for IPO research", *Journal of Corporate Finance.* vol. 27, pp. 367-383.

Carter, R.B., Dark, F.H. & Singh, A.K. 1998, "Underwriter Reputation, Initial Returns, and the Long-Run Performance of IPO Stocks", *The Journal of Finance,* vol. 53, no. 1, pp. 285-311.

Chambers, D. & Dimson, E. 2009, “IPO Underpricing over the Very Long Run”, *The Journal of Finance,* vol. 64, no. 3, pp. 1407-1443.

Enke, D. & Thawornwong, S. 2005, “The use of data mining and neural networks for forecasting stock market returns”, *Expert Systems with Applications,* vol. 29, no. 4, pp. 927-940.

Hanley, K.W. & Wilhelm, W.J. 1995, “Evidence on the strategic allocation of initial public offerings”, *Journal of Financial Economics,* vol. 37, no. 2, pp. 239-257.

Hubbard, R. G. (1990) Asymmetric Information, Corporate Finance, and Investment. University of Chicago Press.

Huck, N. 2009, “Pairs selection and outranking: An application to the S&P 100 index”, *European Journal of Operational Research,* vol. 196, no. 2, pp. 819-825.

Hunt-McCool, J., Koh, S.C. & Francis, B.B. 1996, “Testing for Deliberate Underpricing in the IPO Premarket: A Stochastic Frontier Approach”, *The Review of Financial Studies,* vol. 9, no. 4, pp. 1251-1269.

Ibbotson, R.G. 1975, “Price performance of common stock new issues”, *Journal of Financial Economics,* vol. 2, no. 3, pp. 235-272.

Krauss, C., Do, X.A. & Huck, N. 2016, “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500”, *European journal of operational research*, vol. 259, no. 2, pp. 689-702.

Ljungqvist, A. 2007, "Chapter 7 - IPO Underpricing" in Handbook of Empirical Corporate Finance, ed. B.E. Eckbo, Elsevier, San Diego, pp. 375-422.

Logue, D.E. 1973, “On the pricing of Unseasoned Equity Issues: 1965-1969”, *The Journal of Financial and Quantitative Analysis*, vol. 8, no. 1, pp. 91-103.

Loughran, T. & Ritter, J. 2004, “Why has IPO underpricing changed over time?”, *Financial Management,* Vol. 33, no. 3, pp. 5-37.

Lowry, M., Officer, M.S. & Schwert, G.W. 2010, “The Variability of IPO Initial Returns”, *The Journal of Finance,* vol. 65, no. 2, pp. 425-465.

Lowry, M. & Schwert, G.W. 2002, “IPO Market Cycles: Bubbles or Sequential Learning?”, *The Journal of Finance*, vol. 57, no. 3, pp. 1171-1200.

MacKie-Mason, J. K. (1990) ‘Do Firms Care Who Provides Their Financing?’, in Asymmetric Information, Corporate Finance, and Investment. University of Chicago Press, pp. 63–104.

Marsh, P. 1982, “The choice between debt and equity: An empirical study”, *Journal of Finance,* vol. 37, no. 1, pp. 121-144.

Mello, A.S. & Parsons, J.E. 1998, “Going public and the ownership structure of the firm”, *Journal of financial economics,* vol. 49 no. 1, pp. 79-109.

Pirayesh Neghab, D., Bradrania, R. & Elliot, R. 2023, “Deliberate premarket underpricing: New evidence on IPO pricing using machine learning”, *International Review of Economics & Finance,* vol. 88, pp. 902-927.

Purnanandam, A.K. & Swaminathan, B. 2004, “Are IPOs really underpriced?”, *The Review of Financial Studies*, vol. 17, no. 3, pp. 811-848.

Reilly, F.K. & Hatfield, K. 1969, “Investor experience with new stock issues”, *Financial analysts journal,* vol. 25, no. 5, pp. 73-80.

Ritter, J.R. & Welch, I. 2002, “A review of IPO activity, pricing, and allocations.”, *The Journal of Finance,* vol. 57, no 4, pp. 1795-1828.

Ritter, J.R. 1984, “The "Hot Issue" Market of 1980”, *The Journal of Business,* vol. 57, no. 2, pp. 215-240.

Ritter, J.R. 1987, “The costs of going public”, *Journal of financial economics,* vol. 19, no. 2, pp. 269-281.

Ritter, J.R. 2023, “Initial Public Offerings: Underpricing”. Available through: Warrington College of Business, University of Florida.

Rock, K. 1986, "Why new issues are underpriced", *Journal of Financial Economics,* vol. 15, no. 1, pp. 187-212.

Roger, T., Bousselmi, W., Roger, P. & Willinger. M. 2018, “Another Law Of Small Numbers: patterns of trading prices in experimental markets”, *CEE-M Working Papers,* hal-01954921.

Stoll, H.R. & Curley, A.J. 1970, “Small business and the new issues market for equities”, *The Journal of Financial and Quantitative Analysis*, vol. 5, no. 3, pp. 309-322.

Stoughton, N. & Zechner, J. 1998, “IPO-mechanisms, monitoring and ownership structure”, *Journal of financial economics*, vol. 49, no. 1, pp.45-77.

Weld, W.C., Michaely, R., Thaler, R.H. & Benartzi, S. 2009, “The Nominal Share Price Puzzle”, *The Journal of Economic Perspectives,* vol. 23, no. 2, pp. 121-142.

West, R.M. 2022, “Best practice in statistics: The use of log transformation”, *Annals of clinical biochemistry,* vol 59, no. 3, pp. 162-165.

Appendix:

I. Linear Assumptions

MLR and MLC build on various assumptions that have been tested in a plethora of different statistical tests such as Variance Inflation Factors, White’ test as well as the Durbin-Watson test. The conclusions from these results are that the following variables have been log-transformed to primarily normalize skewed data: roa, proceeds_amount and off_price.

Variance Inflation Factors (VIF) measure multicollinearity among independent variables in a regression model. High VIF values, typically above 10, indicate strong correlation between predictors, which essentially mitigates the reliability of regression coefficients. In the analysis conducted on the data set, VIF values were calculated for each predictor. The variables such as off_price, roa, and proceeds_amount exhibited the highest VIF values, indicating relatively high multicollinearity among these factors. On the other hand, market_return showed a lower VIF, suggesting lower correlation with other predictors. High VIF values might affect the precision of coefficients, influencing the accuracy of individual predictor effects on the dependent variable, resulting in the mentioned variables being log-transformed to reduce multicollinearity.

TABLE 1

VIF TABLE

Results of analyzing multicollinearity through VIF values between all the variables used in the MLR and MLC.

Variable	**VIF**
off_price	87.74

roa	56.03
market_return	1.17
age	10.21
proceeds_amount	62.01
total_assets	15.14

The p-value from White's test represents the probability of observing the test statistic (under the null hypothesis that there is homoscedasticity) as extreme as the one obtained if the null hypothesis were true. Both the LM-Test p-value (8.31e-05) and the F-Statistic p-value (2.53) are very low, providing strong evidence against the assumption of homoscedasticity. In simpler terms, the variance of the residuals might not be constant across different levels of the independent variables, potentially violating one of the assumptions of linear regression.

The Durbin-Watson statistic assesses the presence of autocorrelation in the residuals of a regression model. This statistic ranges between 0 and 4, where values closer to 0 indicate positive autocorrelation, a value around 2 suggests no autocorrelation, and values closer to 4 indicate negative autocorrelation.

In this case, a Durbin-Watson statistic of approximately 1.80 was received which suggests a tendency toward positive autocorrelation but still relatively close to the value of 2, which would indicate no autocorrelation. This suggests that there might be a mild tendency for the residuals to be positively correlated, but it is not strong.

II. SkLearn Optimization

The following is an in depth guide for how the regression and classification algorithms have been used in this project. The SKLearn library is used throughout as it is a standard built-in tool for Python code. The hyperparameters have been optimized through the function GridSearchCV to maximize R-squared whilst minimizing MSE and prediction accuracy for the regressions and classifications respectively. This found the best alternatives for each value defined below and

allows for reproducibility. In all regressions and classifications, the random state was set to ‘42’ which is useful if the desired results are to be reproducible. Not setting this to an exact integer value will result in varying results for each time the code is run.

II.a Multilinear Model

In this library, the default hyperparameters are used, i.e no hyperparameters are defined explicitly, as they result in the best desired outcome of the statistical variables. To implement MLC in python, the support vector classifier (SVC) library is used within the SKLearn library. It takes ‘kernel’ and regularization parameter ‘C’. The kernel defines what type of regression is to be fitted, which in this case is set to ‘linear’. The parameter ‘C’ is set to the default value of 1 since this is a linear classification. The value 1 means that the model assigns equal importance to maximizing the margin as minimizing the classification error, which is default for linear tasks.

TABLE 2

MLR & MLC HYPERPARAMETERS

Definitions and parameter values for MLR and MLC in the built-in SKLearn library.

Multilinear Regressor		
Standard/default parameters are used.		
Multilinear Classifier (SVC)		
kernel	linear	Kernel type to be used in the algorithm.
C	1	The ratio between importance of maximizing margins to minimizing classification error, default for linear classification.

II.b Neural Network Algorithms

The SkLearn Library includes both a neural network regression (NNR) and a neural network classifier (NNC) extension. NNR is created with inputs such as ‘hidden_layer_size’, ‘max_iter’, ‘random_state’, ‘solver’ and ‘activation’. NNC, however, is based on a multi-layer-perceptron (MLP) classifier. The MLP classifier is a so-called feed-forward neural network architecture that works similar to a NNR but with a binary output variable instead. The same parameters are used in NNC as for NNR and have been validated through optimization functions.

TABLE 3

NNR AND NNC HYPERPARAMETERS

Definitions and parameter values for NNR and NNC in the built-in SKLearn library.

Neural Network Regressor		
Variable	**Value**	**Explanation**
hidden_layer_sizes	128 x 64	This depicts the matrix form of the hidden layer.
solver	sgd	Specifies how the outputs are calculated from the hidden layer. The default solver ‘adam’ (stochastic gradient descent based) works well on relatively large data sets in terms of both training time and validation score (accuracy).
activation	tanh	Specifies how the nodes connect with one another. ‘Tanh’ is used because the derivatives of the tanh are larger than the derivatives of the (default) sigmoid function which helps us minimize the loss function faster.
learning_rate	0.01	The amount by which the contribution of each node is shrinked in response to the estimated error each time the weights are updated.
Neural Network Classifier		
hidden_layer_sizes	64 x 32	See explanations above.
max_iter	1000	
solver	sgd	
activation	tanh	

II.c Random Forest Algorithms

To implement a RFR and RFC on the data set, the SkLearn library is used once again. For RFR, four hyperparameters are used to fit the data in the best possible way: ‘n_estimators’, ‘max_depth’, ‘criterion’ as well as ‘max_features’. The RBC has been optimized by iterating over the same four variables. This means that ‘n_estimators’ trees are made with a maximum

depth of ‘max_depth’, each with a prediction value (continuous or binary) which is then averaged to get the most accurate prediction.

TABLE 4

RFR AND RFC HYPERPARAMETERS

Definitions and parameter values for RFR and RFC in the built-in SKLearn library.

Random Forest Regressor		
Variable	**Value**	**Explanation**
n_estimators	500	The number of trees created.
max_depth	6	The depth of each tree.
max_features	sqrt	The loss function to be optimized during the regression.
criterion	absolute_error	The number of features that each tree focuses on during regression.
Random Forest Classifier		
n_estimators	500	See explanations above.
max_depth	5	
max_features	log2	
criterion	gini	For more information about gini see Decision Trees document.[7]

II.d Gradient-Boosting Tree algorithms

To implement a GBR on the data set, the SkLearn library is used once again. Here, four hyperparameters are used to fit the data in the best possible way: ‘n_estimators’, ‘max_depth’, ‘loss’, ‘max_features’ as well as ‘learning_rate’. GBC has also been optimized by iterating over the same variables apart from ‘learning_rate’.

TABLE 5

GBR AND GBC HYPERPARAMETERS

[7] https://scikit-learn.org/stable/modules/tree.html#tree-mathematical-formulation

Definitions and parameter values for GBR and GBC in the built-in SKLearn library.

Gradient-Boosting Tree Regressor		
Variable	**Value**	**Explanation**
n_estimators	500	The number of trees created.
max_depth	2	The depth of each tree.
loss	huber	The loss function to be optimized during the regression.
max_features	sqrt	The number of features that each tree focuses on during regression.
learning_rate	0.01	The amount by which the contribution of each tree is shrinked in response to the estimated error each time each tree is updated.
Gradient-Boosting Tree Classifier		
n_estimators	1000	See explanations above.
max_depth	5	
loss	exponential	
max_features	sqrt	

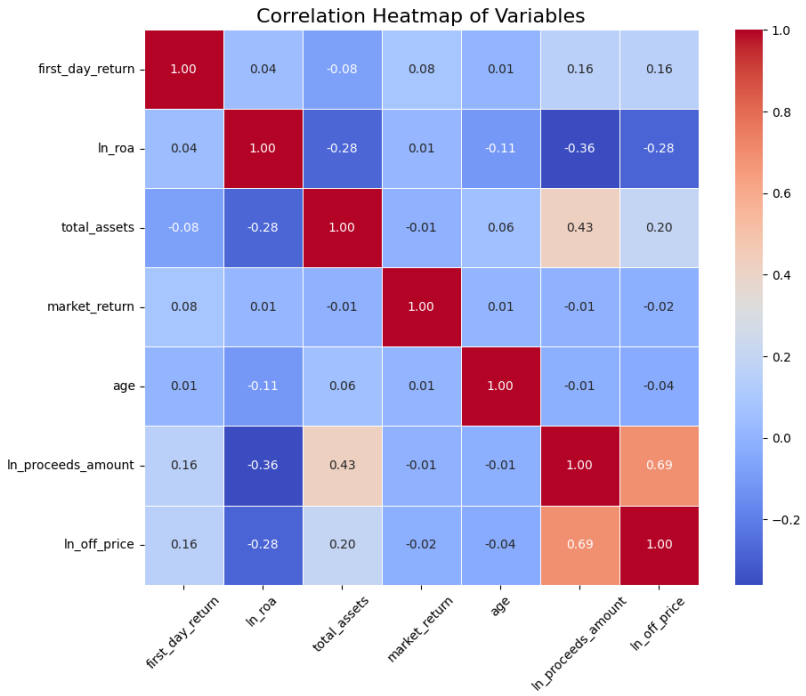
III. Correlation Matrix (linear model)

As one can see in the heatmap below, the explanatory variables have very little correlation to the dependent variable initial returns. Therefore, the linear regression will be heavily influenced by the intercept value in the linear model.

FIGURE 1

HEATMAP OF INDEPENDENT AND DEPENDENT VARIABLES

Figure 1 portrays a correlation heatmap of the variables for the linear model, where ‘ln’ before the variable means that it has been (naturally) log transformed.



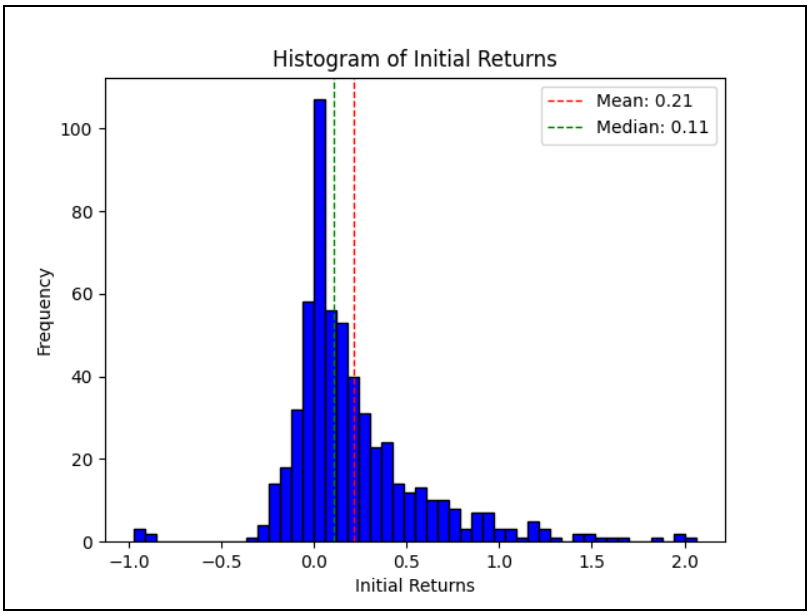
IV. Histogram of Initial Returns

The histogram of the initial returns illustrates the distribution of our dependent variable. The median (10,76%) takes on a lower value than the mean (21,27%), which indicates that the data is positively skewed.

FIGURE 2

INITIAL RETURN HISTOGRAM

Figure 2 portrays the distribution of the dependent variable. The green dotted line exhibits the median and the red dotted line exhibits the mean of the dependent variable.


Histogram of Initial Returns
Mean: 0.21
Median: 0.11
Frequency
Initial Returns
0
20
40
60
80
100
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0