U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

IPO Underpricing Prediction: Integrating Machine Learning Models With Financial And Non-Financial Metrics

João Valente Martins

Dissertation Report

Master in Finance

Supervised by

Professor Miguel Sousa

2025

Acknowledgements

This thesis would not have been possible without the support, guidance, and encouragement of many important people, to whom I am deeply grateful.

First and foremost, I am sincerely thankful to my supervisor, Professor Miguel Sousa, for his invaluable guidance and availability throughout the entire research process. His expertise and insightful feedback have been fundamental to the successful completion of this work.

I am profoundly grateful to my parents for their unconditional support, sacrifices, and unwavering belief in me, not only during this period but throughout my entire life. You truly are my role models. To my brother, thank you for always being there to motivate and cheer me up in every challenge I faced.

To my friends, your companionship, laughter, and encouragement have brought the much-needed balance and strength during this journey. And to my girlfriend, thank you for your understanding and patience during the long hours and difficult moments of this process, and for always reminding me to believe in myself.

Finally, I wish to extend my heartfelt thanks to everyone who contributed, directly or indirectly, to this thesis. Your kindness and support have truly made this journey so special.

Abstract

The **IPO underpricing** phenomenon consists of newly issued stock closing at a higher price on its first trading day than its issuing price (at IPO). It represents a missed opportunity for firms to raise additional capital while incurring substantial costs for pre-IPO shareholders. Traditional econometric methods have provided valuable insights into the matter. However, they fail to capture the complex nature of the data involved. This study takes on advanced **machine learning models**, such as Random Forest, Decision Trees, Support Vector Machines and Extreme Gradient Boosting, to **predict** IPO Underpricing in the **US market**, improving upon prior research by incorporating firm-specific and market sentiment variables.

By employing data from 304 U.S. IPOs between 2020 and 2024, the study aims to address two research questions: (1) Can some machine learning techniques predict whether an Initial Public Offering will be underpriced? and (2) Can some machine learning techniques help Initial Public Offering investors boost their results?

This research holds significant implications for all the involved stakeholders in the IPO process. Predicting whether an IPO will be underpriced can help issuers leave less money on the table and boost investors' returns.

The best-performing model would allow investors to achieve a return of 10.37% (7.71 percentage points higher than an uninformed investor, for the same allocation case) on choosing whether to invest in the last 30 IPOs of our sample.

Keywords: **IPO Underpricing; machine learning models; predict; US market**

Resumo

O fenómeno da **subvalorização da oferta pública inicial** (*IPO*) consiste no facto de as ações recém-emitidas fecharem a um preço mais alto no seu primeiro dia de negociação do que o seu preço de emissão (na *IPO*). Representa uma oportunidade perdida para as empresas angariarem capital adicional, enquanto acarreta custos substanciais para os acionistas pré-IPO. Os métodos econométricos tradicionais forneceram informações valiosas sobre o assunto, mas não conseguem captar a natureza complexa dos dados envolvidos. Este estudo utiliza **modelos de aprendizagem automática**, como *Random Forest*, *Decision Trees*, *Support Vector Machines* e *Extreme Gradient Boosting*, para **prever** a subvalorização das ações em oferta pública inicial no **mercado dos EUA**, melhorando as pesquisas anteriores ao incorporar variáveis específicas da empresa e do sentimento do mercado.

Ao utilizar dados de 304 *IPO*s nos EUA entre 2020 e 2024, o estudo visa abordar duas questões de investigação: (1) Podem algumas técnicas de aprendizagem automática prever se uma oferta pública inicial será subvalorizada? e (2) Podem algumas técnicas de aprendizagem automática ajudar os investidores de ofertas públicas iniciais a melhorar os seus resultados?

Esta investigação tem implicações significativas para todas as partes envolvidas no processo de *IPO*. Prever se uma *IPO* será subvalorizada pode ajudar os emissores a perder menos dinheiro e aumentar os retornos dos investidores.

O modelo com melhor desempenho nas melhores circunstâncias alcançou um retorno de 10,37% (7,71 p.p. superior ao caso base, de um investidor desinformado, para o mesmo caso de alocação), ao escolher se deve investir ou não nas últimas 30 *IPO*s extraídas para o estudo.

Palavras-Chave: **subvalorização da oferta pública inicial**; **modelos de aprendizagem automática**; **prever**; **mercado dos EUA**

Table of Contents

1. Introduction....1
2. Literature Review....3
   2.1 Theoretical Explanations of IPO Underpricing....3
   2.1.1 Information Asymmetry Theories....3
   2.1.2 Institutional Theories....4
   2.1.3 Behavioural Theories....5
   2.2 Key Determinants of Underpricing....5
   2.3 Underpricing Globally....6
   2.4 Machine learning techniques in IPO Underpricing prediction....8
3. Methodology....10
   3.1 Research Design....10
   3.2 Data Collection....11
   3.3 Variables Description....11
   3.3.1 Target Variable....11
   3.3.2 Explanatory Variables....12
   3.4 Data Cleaning....15
   3.5 Data Division....15
   3.6 Data Preparation....15
   3.7 Models....16
   3.8 Normalisation....17
   3.9 Handling Data Imbalance with SMOTE....17
   3.10 Model Development....18
   3.10.1 Train Test Split....18
   3.10.2 Performance Measures....19
   3.10.3 Hyperparameter Tuning....20
   3.10.4 Cross-Validation....21
   3.11 Practical Test to Verify the Impact of the Models....21
4. Results and Discussion....23
   4.1 Preliminary Statistical Results....23
   4.1.1 Descriptive Statistics....23
   4.1.2 Correlation Matrices....27
   4.1.3 Mutual information....29
   4.2 Regression Models....31
   4.3 Classification Models and Respective Accuracies....34
   4.4 Results for the Practical Test to Verify the Impact of the Models....39

4.4.1 First Allocation Case 39
4.4.2 Second Allocation Case 41
4.4.3 Third Allocation Scenario 43
4.4.4 Fourth Allocation Case 44
5. Conclusion 47
6. References 49

List of Figures

Figure 1 – Correlation matrix for Investors' predictive models' variables without outliers. 28

Figure 2 – Correlation matrix for Underwriters' predictive models' variables without outliers. 29

Figure 3 – Violin charts of the target variable of training and testing datasets for the Underwriters' predictive models alternative with outliers. 32

Figure 4 – Parity plots of the SVR models. Left: with outliers, right: without outliers.....34

Figure 5 – Confusion matrices, left: best-performing test set for Investors' predictive model without outliers (model RF SG); right: worst-performing test set for Investors' predictive model with outliers (model DT SG). 38

List of Tables

Table 1 – IPO Underpricing Globally: A sample from the Loughran et al. (1994) study. 7

Table 2 – IPO underpricing prediction studies with machine learning approaches. 8

Table 3 – Descriptive statistics of the Investors' predictive models with outliers. 24

Table 4 – Descriptive statistics of the Underwriters' predictive models with outliers. 25

Table 5 – Descriptive statistics of the Investors' predictive models without outliers. 26

Table 6 – Descriptive statistics of the Underwriters' predictive models without outliers. ...27

Table 7 – Relevance of explanatory variables of Investors' predictive models, with and without outliers. 30

Table 8 – Relevance of explanatory variables of Underwriters' predictive models, with and without outliers. 31

Table 9 – Values of R2 and RMSE for training and testing datasets for Underwriters' predictive models. 33

Table 10 – Optimised hyperparameters for each model. 35

Table 11 – Training and testing accuracies of the optimised models. 36

Table 12 – ROI of the different predictive models for different SMOTE settings with and without outliers in the first allocation case (with percentages of 5 %, 30 %, 50 %, 70 % and 100 %) 40

Table 13 – ROI of the different predictive models for different SMOTE settings with and without outliers in the second allocation case (with percentages of 5 %, 30 %, 50 %, 70 % and 100 %). 42

Table 14 – ROI of the different predictive models for different SMOTE settings with and without outliers in the third allocation case (with percentages of 15 %, 30 %, 50 %, 70 % and 100 %). 44

Table 15 – ROI of the different predictive models for different SMOTE settings with and without outliers in the fourth allocation case (with percentages of 15 %, 30 %, 50 %, 70 % and 100 %). 45

Table 16 – Summary table of the ROIs of Uninformed Investor, and average and best scenario of Investors' and Underwriters' predictive models for the four allocation cases. ..46

List of Acronyms and Abbreviations

AS	Acquired Stake
DT	Decision Trees
DV	Deal Value
EBIT	Earnings Before Interest and Taxes
FS	Firm Size
IPO	Initial Public Offering
MK RT	Market Return
NPROF	Net Profit
OP	Offering Price
OR	Over Range
PP	Predicted Price
R^2	Coefficient of Determination
RF	Random Forest
RFN	Random Forest Normalised
RMSE	Root Mean Square Error
ROA	Return On Assets
ROI	Return On Investment
RR	Relation to Range
SG	SMOTE Global
SM	Simple Models
SMOTE	Synthetic Minority Oversampling Technique
ST	SMOTE in Training
SVC	Support Vector machines in Classification
SVR	Support Vector machines in Regression
U	Underpricing
UR	Under Range
XGB	Extreme Gradient Boosting

1. Introduction

The phenomenon of Initial Public Offering (IPO) underpricing has been a subject of interest and debate in the corporate finance field since Ibbotson (1975) documented the short-run underperformance of IPOs and observed the frequent initial return premium (i.e., underpricing). This phenomenon consists of the tendency of newly issued stocks to close at a higher price on their first trading day than their original issuing price at their IPO. This underpricing, while beneficial to initial investors, represents a missed opportunity for the issuing companies to raise additional capital and imposes substantial costs on their pre-IPO shareholders. The magnitude of this phenomenon ranges widely across markets, as noted by Loughran et al. (1994). Their study shows that developed countries like Austria and Denmark show single-digit percentages on average initial returns (a measure of IPO underpricing that consists of the percentual difference between the price at which the first trading day closes and the IPO issuing price per share), and some emerging markets present values of over 100%, meaning that these companies could have duplicated the raised capital.

Historically, underpricing has been explained by many theories, such as Information Asymmetry and Signalling. Rock (1986) introduced the winner's Curse model to explain the Information Asymmetry theory, which suggests that uninformed investors face greater risks compared to their informed counterparts. He describes the need for the underpricing phenomenon to ensure that uninformed investors don't quit the market, exposing it to liquidity problems. Welch (1989) expanded on this theory with one of its own, the Signalling theory. It consists of the idea that firms deliberately underprice their IPO to signal their value to the market in the hope of recovering this cost on subsequent financing rounds.

The advancements in data science and machine learning techniques have offered a new panoply of possible approaches to analyse data. Despite being valuable, traditional econometric models often fail to capture nonlinear relationships and interactions between variables in financial data. On the other hand, machine learning techniques can process large amounts of data and identify hidden patterns in it. Quintana et al. (2017) exemplify this potential by using the Random Forest algorithm to predict IPO underpricing in the USA market. However, this study relied only on IPO characteristics as its variables, while other similar studies provided a more complete approach by also incorporating firm characteristics and market sentiment variables.

The study conducted in this dissertation consists of a more complete predictive approach to the IPO underpricing phenomenon in the USA compared to that of Quintana et al. (2017) by overcoming the main limitations stated by the authors, which consist of relying only on IPO characteristics. This work will include firm characteristics and market sentiment variables to provide a better-performing predictive analysis.

The relevance of this study lies in its practical implications for the various stakeholders involved in the IPO process. For issuers, better estimations of underpricing can lead to more accurate pricing strategies, leaving less money on the table, thereby reducing one of the IPO's most significant costs. For investors (whom the study will focus on), especially those less informed, these models can help them make safer investment decisions. Lastly, insights into underpricing dynamics can refine advisory services and pricing mechanisms for the underwriters, enhancing their market credibility.

The dissertation structure will be characterised by the introduction (current section), where the theme is introduced, the purpose and relevance of the work are presented, and the structure of the work is described. A literature review section follows, where the study is situated in the existing literature, and a gap is found, containing also a description of the main theoretical and empirical studies on the theme. Then, the methodology part describes all the data processing, from its collection to the model building and output estimation. In sequence, a section presenting the results obtained and the respective discussion. Lastly, the conclusion includes a final overview of the work as well as its limitations and improvement possibilities.

2. Literature Review

2.1 Theoretical Explanations of IPO Underpricing

The IPO underpricing phenomenon has been a central focus of academic research, with various theories proposed to explain the persistent gap between the IPO issuing price and the closing first-day trading price. These theories can be categorised into three categories: informational asymmetry theories, institutional theories, and behavioural perspectives.

2.1.1 Information Asymmetry Theories

The information asymmetry theory is one of the most influential and accepted explanations of the phenomenon. Originally proposed by Rock (1986), the Winner's Curse model argues that the disparity in knowledge between informed and uninformed investors creates adverse selection. Informed investors, typically institutional players, have, as the name suggests, better information or a better capacity to process the information than their counterparts, bidding only on undervalued IPOs, leaving overpriced issues disproportionately allocated to uninformed investors. To encourage this second group of investors not to leave this market, issuers underprice their offering to provide a compensatory return.

This model has been empirically supported by various studies, such as Carter and Manaster (1990), Keloharju (1993), and Chhabra and Kiran (2022). However, its applicability in explaining the magnitude of underpricing in developing markets has been questioned by Keasey and Short (1992) and Beatty and Welch (1996).

The Signalling Theory takes a different approach, suggesting that firms use underpricing as a deliberate strategy to signal their quality to the market. Allen and Faulhaber (1989) and Welch (1989) argue that high-quality firms incur the underpricing cost to "leave a good taste in investors' mouths" to enhance their reputation and increase the likelihood of favourable future equity offerings. The Francis et al. (2010) study empirically supports the theory, which showed that IPO firms with substantial underpricing at the initial stage are more inclined to engage in subsequent equity issues. However, Ritter (2011) criticised the inefficiency of this theory, labelling it as a "silly academic theory". He argues that underpricing is an expensive signalling strategy, especially compared to alternative methods such as advertising and marketing.

The certification hypothesis, proposed by Booth and Smith (1986), emphasises the role of third-party certifications, such as reputable underwriters, in reducing information asymmetries, as their involvement suggests a guarantee of the quality of the issuing firm.

Empirically, the theory is supported by Carter and Manaster (1990) and Habib and Ljungqvist (2001), who indicate that the involvement of reputable underwriters is associated with lower levels of underpricing, as they certify the firm's value to potential investors. However, some critics, such as Marisetty and Subrahmanyam (2010), argue that this hypothesis fails to explain the extreme levels of the phenomenon observed in many developing markets, where reputable underwriters are still associated with significant underpricing.

Ex-ante Uncertainty theory provides another perspective on the phenomenon, suggesting that firms characterised by higher uncertainty regarding their value present greater underpricing. Beatty and Ritter (1986) demonstrated that smaller firms and those that lacked detailed specifications of what they would do with the proceeds are more likely to face investor scepticism, leading to higher underpricing levels. Subsequent research by Leone et al. (2007) further supports this hypothesis, showing that transparency in the use of IPO proceeds can reduce uncertainty and, consequently, underpricing.

2.1.2 Institutional Theories

Institutional Theories provide an alternative perspective on the phenomenon, focusing on the influence of regulatory and market-specific factors on IPO underpricing.

The Lawsuit Avoidance Hypothesis was introduced by Ibbotson (1975), who proposed that US IPO issuers deliberately underprice issued securities to avoid potential litigation risks from disappointed investors due to poor post-IPO performance. A lawsuit from the IPO investors not only directly inflicts financial damages, resulting from incurred legal fees and waste of management time, but it also indirectly harms the company's reputation, capital and likelihood of incurring higher costs when raising capital in the future. This hypothesis is studied and modelled by Tinic (1988), Hughes and Thakor (1992) and Hensler (1995). It is empirically validated by Lowry and Shu (2002), who showed that 6% of IPO firms in the US were sued, causing damages averaging 13.3% of the proceeds of IPOs from 1988 to 1995.
The Price Stabilisation Hypothesis highlights the underwriter's role in managing aftermarket volatility. Ruud (1993) showed that underwriters engage in post-IPO price stabilisation to prevent dramatic price decreases, which could damage their reputation and future business prospects. This activity is particularly valuable in highly volatile markets with significant speculative trading.

2.1.3 Behavioural Theories

Behavioural theories provide a psychological explanation for the phenomenon, focusing on irrational investor behaviour. Welch (1992) introduced the Informational Cascades Model, where he argues that later investors of an IPO base their decisions on the actions of the previous participants, disregarding their own information and opinions. This sequential behaviour amplifies demand, prompting issuers to underprice as a strategy to generate early investors' interest, creating traction and positive momentum for the issuance. Amihud et al. (2003) and Pollock et al. (2008) empirically confirmed the presence of informational cascades in the IPO market, concluding that underpricing can serve as a mechanism to generate favourable market demand, contributing to a successful issuance.

2.2 Key Determinants of Underpricing

IPO underpricing is influenced by a combination of firm, issue, and country-specific factors and broader market conditions that possess significant explanatory significance in estimating the initial return magnitude (underpricing).

2.2.1 Company-Specific Factors

Company characteristics, such as size, age, and ownership structure, strongly influence the IPO underpricing magnitude. Smaller and younger companies present greater ex-ante uncertainty, increasing underpricing levels to compensate for the perceived risk (Beatty & Ritter, 1986). Empirically, Sahoo and Rajib (2010) showed that less established companies consistently experience higher underpricing levels, resulting from investors' scepticism about their long-term viability. Financial information prior to the IPO can also be valuable for this estimation. Basti et al. (2015) used net income, a profitability ratio (return on asset, ROA), and a debt-to-equity ratio (DER) to conduct their study on IPO underpricing prediction.

2.2.2 Issue-Specific Factors

The pricing mechanism is itself a significant issue-specific determinant. The book-building method is widely adopted due to its ability to incorporate investors' feedback during the price determination part of the process, leading to relatively lower underpricing on these issuances. However, in emerging markets like Indonesia, even book-building IPOs exhibit high underpricing levels due to information asymmetry and speculative behaviour (Utamaningsih, 2017). Offering size is also a significant determinant of the phenomenon, as larger offerings may require higher underpricing to attract sufficient demand.

2.2.3 Country-Specific Factors

Legal and institutional frameworks are important in shaping IPO underpricing. Countries with developed markets, like the US, exhibit lower underpricing levels due to strong regulatory oversight, investor protections, and transparent disclosure standards (Lowry and Shu, 2002). On the other hand, emerging markets such as Bangladesh and Saudi Arabia, due to weaker legal systems, limited market maturity, and significant information asymmetry, experience higher underpricing levels (Islam et al., 2010; Alanazi and Al-Zoubi, 2015).

2.2.4 Market Conditions

The market conditions under which the issuance is conducted significantly influence its underpricing, particularly during periods of great market volatility and characterised by speculative investor behaviour. During "hot" IPO markets (periods characterised by increased investor enthusiasm and oversubscription of IPOs), the underpricing levels tend to rise as issuers aim to capitalise on this period of favourable conditions to ensure successful offerings.

Empirical studies, such as Ritter and Welch (2002) and Coakley et al. (2009), showed that in the US and Europe, respectively, "hot" market periods, in fact, lead to higher initial returns due to aggressive bidding and elevated investor demand. The opposite happens during economic downturns ("cold" markets) when underpricing decreases as issuers and investors adopt more cautious approaches. Furthermore, speculative bubbles in emerging markets like China and the GCC (Gulf Cooperation Council) amplify the impact of market sentiment, contributing to extreme underpricing during bullish periods (Tian, 2011; Alanazi & Al-Zoubi, 2015).

2.3 Underpricing Globally

It has been proven empirically that IPO underpricing reveals significant variations across countries with developed and emerging markets due to differences in institutional frameworks, market conditions, and firm characteristics. However, there is a clear and identifiable pattern where the underpricing phenomenon is larger in countries characterised by emerging markets. Loughran et al. (1994) article provides the equally weighted average initial returns for 55 countries. A sample of the study conducted is presented in Table 1.

Table 1 – IPO Underpricing Globally: A sample from the Loughran et al. (1994) study.

This table presents the Average Initial Return of the IPOs, the sample size (number of IPOs) and the time period of the respective issuances for the countries: China, Qatar, United Arab Emirates, Austria, Denmark, France and the United States. The information was retrieved from the Loughran et al. (1994) study

County	Sample Size	Time Period	Avg. Initial Return
China	5,219	1990-2023	**155.7%**
Qatar	17	2003-2021	**257.2%**
United Arab Emirates	35	2003-2021	**186.4%**
Austria	106	19712018	**6.2%**
Denmark	190	1984-2021	**7.6%**
France	923	1983-2023	**9.3%**
United States	13,811	1960-2023	**17.5%**

From Table 1, the observable average initial return (a measure of IPO underpricing that consists of the percentual difference between the price at which the first trading day closes and the IPO issuing price per share) is way higher for countries with emerging markets such as China, Qatar, and United Arab Emirates (155.7%, 257.2%, and 186.4%, respectively) than the ones of more developed markets, such as Austria, Denmark, and France (6.2%, 7.6%, and 9.3%, respectively). These differences are due to a more regulated IPO process, strong investor protections, more frequent involvement of reputable underwriters, and efficient pricing mechanisms in developed markets, which do not happen in emerging markets or happen to a smaller extent.

Despite being a developed market with its own unique characteristics, the U.S. still exhibits a high average initial return (17.5%) and a substantial number of IPOs each year, making it a suitable and relevant setting for the development of this study.

2.4 Machine learning techniques in IPO Underpricing prediction

Previous IPO Underpricing studies have predominantly employed some variants of simple regression analysis and statistical t-test. As examples, Boonchuaymetta and Chuanrommanee (2013) used OLS and t-statistics, Chorruk and Worthington (2010) used multivariate regression, Vong and Trigueiros (2010), and Deb and Marisetty (2010) used cross-sectional regression models.

Machine learning techniques can capture and represent complex relationships between input and output variables hidden in the data, identifying patterns without being subject to constraining assumptions such as linearity, normality and multicollinearity (Basti et al., 2015).

The reduced number of studies with this innovative approach to the phenomenon can be considered a gap in the existing literature. The studies that began bridging that gap, presented in Table 2, applied distinct machine learning techniques that try to predict whether the IPO will be underpriced and to what extent.

Table 2 – IPO underpricing prediction studies with machine learning approaches.

This table summarises the studied country, the methodology applied and the considered variables of the relevant studies that used machine learning techniques to predict IPO underpricing. Those studies being: Basti et. al (2015), Baba and Sevil (2020), Dhini and Sondakh (2024), Esfahanipour et. al (2016) and Quintana et al. (2017)

Authors	Country	Methodology	Variables
Basti et. al (2015)	Turkey	• Decision Trees • Support Vector Machine	• Company Characteristics • IPO Characteristics • Market Sentiment
Baba and Sevil (2020)	Turkey	• Random Forrest • Robust regression methods	• Company Characteristics • IPO Characteristics • Market Sentiment
Dhini and Sondakh (2024)	Indonesia	• Random Forrest • Gradient Boost Tree	• Company Characteristics • IPO Characteristics • Market Sentiment
Esfahanipour et. al (2016)	Iran	• Neural Network • Fuzzy Regression	• Company Characteristics • IPO Characteristics • Market Sentiment
Quintana et al. (2017)	United States	• Random Forrest • Instance-based learning • Least median of squares	• IPO Characteristics

The studies presented in Table 2 explore IPO performance and underpricing prediction across different regions using advanced machine learning methods. They utilise variables related to company characteristics, IPO features, and market sentiment to enhance predictive accuracy. Across all studies, machine learning techniques such as Random Forest, Gradient Boosted Trees, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks consistently outperform traditional regression models, demonstrating their robustness in handling complex financial datasets with high variability and outliers.

The best performing methods were Random Forest and Gradient Boosted Trees for their ability to combine strong and weak predictors. Nonetheless, all the ensemble methods showed superior performance compared to the most traditional linear regression models due to their ability to handle large datasets, optimise hyperparameters, and capture non-linear relationships.

The variables used in the studies can be classified into three categories: company characteristics, IPO characteristics, and market sentiment. From the first group (company characteristics), the most common variables across the presented studies were size, as measured by total assets, profitability captured by ROA, solvency as the Debt to Asset Ratio, and industry type. From the second one (IPO characteristics), the more usual ones were offering price, IPO proceeds, retained shares, and underwriter reputation. Lastly, from the market sentiment group of variables, the most frequent were pre-IPO market Index performance and investor optimism captured in various ways.

The studies conducted using machine learning techniques have been concentrated in emerging markets (Turkey, Indonesia and Iran) as they show higher levels of underpricing. Quintana et al. (2017) explore the phenomenon in the US as it is a developed market but, unlike European countries, still presents significant IPO underpricing levels, with a 17.5% average initial return from 1960-2023, according to Loughran et al. (1994).

3. Methodology

3.1 Research Design

This work builds upon the existing literature that aims to predict the Underpricing phenomenon (Quintana et al., 2017) by incorporating firm-specific and market sentiment variables, such as Return on Assets (*ROA*) and investor optimism, respectively, with a machine learning predictive approach. The models used will be those that have shown the best performance in previous similar studies (Basti et al., 2015; Baba & Sevil, 2020; Dhini & Sondakh, 2024): Random Forest, Decision Trees, Support Vector Machines and Extreme Gradient Boosting.

The study will consider two different scenarios, one that includes the variables known at the time an investor needs to express its interest in the offering, designated by Investors' predictive models, and the other, which includes the variables known at the time of the IPO, such as the offering price, denominated Underwriters' predictive models. The first scenario can be valuable to an investor as it aims to predict whether an IPO will be underpriced or not at the time to show its intention to participate in the deal. The second one can be valuable to the underwriter of the IPO, as it aims to predict if the offering will be underpriced or not, according to the offering price they set.

The study intends to answer the following research questions:

Q1. Can some machine learning techniques predict whether an Initial Public Offering will be underpriced?

Q2. Can some machine learning techniques help Initial Public Offering investors boost their results?

To answer the above questions, several activities must be performed, which are presented below.

3.2 Data Collection

The sample was collected from various sources, as the lack of a well-constituted extractable database was one of the major obstacles this work faced. Most of the information was extracted from the Orbis M&A database (Moody's Corporation, 2025), more specifically, the pre-deal financials variables. However, the respective IPO prospectus (U.S. Securities and Exchange Commission, 2025) was also used to complete some information. Other variables, such as Range and First-day Closing Price, were collected from other sources, such as IPO Scoop (IPOScoop, 2025), Yahoo Finance (Yahoo Finance, 2025) and related news.

The retrieved sample consists of USA IPOs that happened between the 11th of June 2020 and the 3rd of December 2024, as, despite including the Covid-19 period, the 5-year window presents a good balance between the number of observations that are needed to construct these types of models while still reflecting the current regulatory environment, socioeconomic state and investor behaviour. The sample period ended in 2024 as the Orbis M&A database (Moody's Corporation, 2025) didn't have the 2024 reports of the companies, presenting the 2023 annual reports as the year prior pre-deal financials for the companies that did their IPOs in 2025.

The USA IPOs were chosen because, despite having a developed market, it still possesses a significant average initial return and number of IPOs (17.5% and 13,811, respectively), according to Loughran et al. (1994), maximising the potential return that investors could get, especially compared to other developed countries.

3.3 Variables Description

3.3.1 Target Variable

According to the finance literature (Beatty and Ritter, 1986; Daily et al., 2003; Loughran and Ritter, 2004), IPO underpricing can be defined by the relation of the difference between the stock's price at the end of its first trading day and the initial offering price defined by the underwriter on its IPO, and this initial offering price. If the computed value is positive, it indicates that the underwriter undervalued the issuing company compared to the market, leaving money on the table (the issuing company could have earned more money) but providing returns to the IPO investors. This relation is given by Equation 3.1.

$$Underpricing_i = \frac{(1st\ Day\ Closing\ Price_i - \ Offer\ Price_i)}{Offer\ Price_i} \times 100 \tag{3.1}$$

The machine learning models will be applied in two ways: in regression and in classification modes. In regression, the models will try to predict the Underpricing variable (*U*; continuous variable) as presented above, and in classification, the models will try to predict a binary variable that takes the value 1 if the *U* is positive and 0 otherwise. The study will focus on the classification problem, as its objective is to predict if the IPO will be underpriced or not (answering Q1) and to compute what returns investors would get investing in them (answering Q2). Nonetheless, the regression was applied to verify the quality of these types of models in this subject.

3.3.2 Explanatory Variables

The explanatory variables were chosen in accordance with the existing predictive literature (see Section 2.4) and can be categorised into three groups: company characteristics, IPO characteristics, and market sentiment.

Company characteristics

The company characteristics group includes the firms' pre-IPO financials, based on the full-year report from the year prior to the IPO. This metric is used because it is the most consistent and comparable across all firms. The IPO prospectus typically includes the full-year report for the previous year, along with financials from the most recent 3, 6, or 9 months, depending on the IPO date. However, only the full-year report is available for all firms, making it the most reliable basis for comparison.

From this group, the explanatory variables are:

1. *ASSETS*, which consist of the total of the accounted assets. They were used as a measure of the size of the firm. IPOs of smaller firms are generally considered more speculative than those of larger firms and are, therefore, expected to exhibit higher levels of Underpricing (Basti et al., 2015);
2. Firm Size (*FS*), computed as the logarithm of total assets, serves as the size factor for the issuing company. Loughran and Ritter (2004) and Vong and Trigueiros (2010) have found a negative and significant relationship between the size of the companies and the Underpricing levels;
3. Net profit (*NETPROF*) and *EBIT*, which are two accounting standards that can reflect the profitability of a company, exhibiting its quality. These types of firm-specific fundamentals can explain variations in the Underpricing levels. Loughran and Ritter (2004) suggest that firms with stronger financial positions are perceived as

less risky by investors due to their signals of quality, which reduce information asymmetry and attract investors' interest without the need for large Underpricing;

4. Return on Asset 1(*ROA1*), defined by the relationship between *EBIT* and the total value of Assets, and Return on Asset 2 (*ROA2*), represented by Net Profit and the total value of Assets. The *ROA* is also used as an explanatory variable in the IPO Underpricing literature, having Grammenos and Papapostolou (2012) discovered a significant and negative relationship between *ROA2* and IPO Underpricing. The authors concluded that underwriters set the offer price of high *ROA2* firms closer to their intrinsic value, and therefore experience lower Underpricing. Studies like Basti et al. (2015) use both computations of the *ROA* as explanatory variables. The variables *ROA1* and *ROA2* are given by Equations 3.2 and 3.3, respectively[1].

$$ROA1 = \frac{EBIT}{ASSETS} \tag{3.2}$$

$$ROA2 = \frac{NET\ PROFIT}{ASSETS} \tag{3.3}$$

IPO characteristics

This group of variables present the ones that describe the structure, context, and financial dimensions of the IPO. Being them the following:

1. Deal Value (*DV*), which refers to the total monetary value of the IPO, is calculated by multiplying the offering price by the number of shares offered. According to Kennedy et al. (2006) and Grammenos and Papapostolou (2012), IPO with higher deal values generally exhibit lower levels of Underpricing. This is because larger IPO are typically associated with more established and financially stable firms, which investors perceive as less risky. As a result, there is less needed to underprice the offering to attract investors' interest;
2. Acquired Stake (*AS*), consisting of the percentage of the company undergoing IPO that will be sold in the deal. A larger acquired stake may signal stronger investor access, reduced insider control, and greater market transparency. As noted in prior literature (Zheng et al., 2005), higher ownership dispersion post-IPO is often

[1] The use of these variables (ASSETS, FS, NETPROF, EBIT, ROA1 and ROA2), which are highly correlated, would be a problem if the aim of the study was to interpret their regression coefficients, as multicollinearity would occur. However, as the aim is to predict the outcome variable, this is not a problem, as machine learning techniques handle correlated variables effectively.

associated with lower levels of underpricing due to reduced information asymmetry and perceived risk;

3. Offering Price (*OP*), which is the price set for the IPO shares. Studies like Booth and Chua (1996) and Ibbotson et al. (1994) suggest that the price is a proxy of uncertainty, and as it increases, the Underpricing should decrease. Quintana et al. (2017) state that the offering price is considered a key value, not only as a component of previous indicators but in its own right, and suggest that the price itself can be viewed as a quality indicator by investors. In the investors' predictive models (scenario with known information at the time the investor needs to make its decision to invest in the IPO), this variable is replaced by the predicted offering price as the middle point of the IPO range (*PP*- Predicted Price);
4. Relative Range (*RR*) is defined as the difference between the Offering Price and the Predicted Offering Price in relation to the latter (Predicted Offering Price). Studies such as Benveniste and Spindt (1989) and Hanley (1993) argue that this relationship can be interpreted as a signal by investors. The greater the value, the larger the expected Underpricing of the IPO. In addition to this variable, two binary ones were created: Over Range (*OR*), which takes the value 1 if the offering price is greater than the maximum value of the range and 0 otherwise, and Under Range (*UR*), which takes the value 1 if the offering price is lower than the minimum end of the range and 0 otherwise. These variables are dependent on the offering price, therefore being excluded from the first created scenario (Investor's Predictive Models). The *RR* variable is calculated by Equation 3.4.

$$RR = \frac{(Offering\ Price - \ Middle\ Range\ Point)}{Middle\ Range\ Point} \tag{3.4}$$

Market Sentiment

For the last group, a single variable was captured, the Market Return (*MKRT*), which consists of the return of the USA stock market from the 30-day period before the IPO. To represent the USA market, the returns computed were those of the S&P 500 (^GSPC) (Yahoo Finance, 2025). This approach has already been implemented by other studies in the past, such as Basti et al. (2015), Baba and Sevil (2020), Dhini and Sondakh (2024), Esfahanipour et al. (2016), and others, who computed that variable for the respective markets of their studies.

3.4 Data Cleaning

For the period between the 11th of June 2020 and the 3rd of December 2024, 636 Initial Public Offerings that occurred in the USA were extracted from the Orbis M&A database (Moody's Corporation, 2025). Then, an elimination procedure was implemented, where observations that lacked needed variables and that were not possible to complete with other sources, such as the respective IPO prospectus (U.S. Securities and Exchange Commission, 2025), Yahoo Finance (Yahoo Finance, 2025) and related news, were removed, as well as the Unit IPOs, IPOs where a warrant, or a fraction of a warrant, is offered with one share. This type of IPO is different from the other in that only a share is offered, as they are usually smaller in offer size, riskier, marketed by less prestigious underwriters, and present lower initial returns when compared to the other type (Jain, 1994). After this removal, the data sample was composed of 304 valid observations.

3.5 Data Division

The dataset was first divided into two datasets. A dataset to develop the models, corresponding to the first 274 observations and a dataset with the last 30 observations to simulate a real-world investment situation, allowing the evaluation of the model's predictive performance. In this situation, the investor is hypothetically positioned on June 29, 2023 (the date of the first of the 30 IPOs) and invests in those predicted by the models to be underpriced, continuing through to the final IPO on December 3, 2024.

3.6 Data Preparation

IPOs can be very unpredictable and present extreme levels of underpricing, even when it is not foreseeable. So, whether to exclude possible outliers from the study was a tough choice to make. To overcome this dilemma, two alternatives were made. One where the outliers were excluded following the Mahalanobis distance method at a 5% significance level, which identifies multivariate outliers based on their distance from the mean while accounting for the covariance structure of the data (De Maesschalck et al., 2000). Observations with a squared Mahalanobis distance exceeding the chi-squared critical value were considered outliers and removed. This procedure was performed with the data of explanatory variables, with 21 and 25 observations removed in the first and second scenarios, respectively. The Mahalanobis distance squared is given by Equation 3.5.

$$D^2(x) = (x - \overline{\boldsymbol{x}})\Sigma^{-1}(x - \overline{\boldsymbol{x}})^T \tag{3.5}$$

Where $\boldsymbol{x}$ is a row vector representing an observation, $\overline{\boldsymbol{x}}$ is the mean row vector (sample mean) computed from a set of observations, and the $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix, used to normalise the distances by accounting for correlations and scaling among the variables.

3.7 Models

In this study, the predictive models will be explored by implementing supervised machine learning methods, which consist of algorithms that learn from input features (explanatory variables) to predict an output variable (target variable). The fundamental principle behind these models is that they are trained on a dataset where the outcomes are known (in this case, they learn from IPOs where the Underpricing is known), allowing them to identify patterns and then make predictions on unseen data. Supervised learning encompasses two types of problems: regression, where the target variable is continuous (the specific percentage of Underpricing), and classification, where the target variable is categorical (in the case of this study, it is binary, 1 if underpriced and 0 otherwise). The study will allocate both approaches but will focus on the classification one, as for an investor who aims to maximise their return, the most important question is whether the IPO will be underpriced or not (to invest in the ones that he predicts will be).

In this work, four of the most popular and efficient supervised machine learning models were applied, namely Random Forest (RF), Decision Trees (DT), Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) in both regression and classification modes (Subasi, 2020; Zollanvari, 2023).

Random Forest (RF)

Random Forest is an ensemble learning method constructed on several decision trees during the training phase and outputs the mean prediction on the regression approach or the most voted class on the classification approach. This method is effective in high-dimensional datasets and non-linear relationships.

Decision Trees (DT)

Decision Trees are intuitive, tree-structured models that split the data into subsets based on the value of input features. In each node, the algorithm selects the feature and threshold that best separate the data according to a certain criterion. In the regression approach, the tree predicts a continuous outcome by averaging the target values obtained in the leaf nodes. In classification, the tree assigns the class based on the most chosen class within each leaf.

Support Vector Machine (SVM)

Support Vector Machines are powerful algorithms that aim to find the optimal boundary (hyperplane) that fits the data (in regression) or divides the data (in classification). For regression (also known as Support Vector Regression, SVR), the models seek a function that deviates from the actual targets by a value inferior to a specific margin (ε), while also being as flat as possible. For classification, SVM (also called Support Vector Classification, SVC) finds the hyperplane that maximises the margin between two classes. These models are particularly effective in high-dimensional spaces and in cases where the relationship between features and output is complex and nonlinear.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is an advanced method of implementing gradient boosting, which consists of an ensemble of weak learners (Decision Trees, normally) in a sequential manner. Each new tree is trained to correct the residual errors of the previous one, optimising a loss function through gradient descent. In a regression approach, the model minimises a continuous loss function such as mean squared error. In classification, it optimises a classification loss, such as logistic loss, for example. XGBoost is known for its scalability, regularisation mechanisms, and high predictive performance.

3.8 Normalisation

Normalisation was applied to the dataset to ensure consistency in preprocessing across all models and to support more reliable performance evaluation. However, RF was applied to both datasets (original dataset and normalised one) to compare whether a marginal increase in performance exists. Nonetheless, this process (normalisation) is required for SVM, as this method calculates the distance between data points, to construct the hyperplane, and features with larger scales (Assets, for example) can dominate the calculation, leading to biased results. Although normalisation is not mandatory for DT and XGBoost models, the models can obtain marginally better results.

3.9 Handling Data Imbalance with SMOTE

The existence of data imbalance is a common problem that appears when applying supervised machine learning techniques. This is more critical for classification than for regression.

The dataset of this study is characterised by a significant class imbalance, with a notably higher number of underpriced IPOs compared to overpriced ones. This imbalance has a

negative impact on the performance of classification models by biasing predictions towards the majority class. To address the issue, the Synthetic Minority Oversampling Technique (SMOTE) was employed (Chawla et al., 2002). This generates synthetic observations of the minority class to create a balanced distribution through the implementation of the k-neighbours procedure (k=5, in this study).

Thus, in this work, three settings for developing models were used. The first one, without SMOTE (SM), establishes a baseline using the original imbalanced dataset. The second one applies SMOTE to the training set (ST), preserving the original distribution of the test set. The third one corresponds to the application of the SMOTE procedure to the entire dataset (train and test sets) (SG), to understand how the models would behave in a fully balanced environment.

3.10 Model Development

The model development process involved several key steps to ensure the reliability and robustness of the results. These included splitting the dataset into training and testing sets, evaluating model performance using appropriate metrics and tuning hyperparameters by applying cross-validation. These steps are applied to this work as follows.

3.10.1 Train Test Split

A common approach in regression and classification is to randomly split the dataset into training (70–80% of observations) and test sets. While simple, this method can produce uneven distributions, leading to biased models that perform poorly on new data.

In regression, the Kennard-Stone (KS) method improves on this by selecting samples based on pairwise distances, ensuring uniform coverage of the feature space (Kennard & Stone, 1969). This reduces overfitting and enhances generalisation, especially in small or imbalanced datasets.

In classification, a strategy to overcome uneven distributions of the classes in train and test sets is to apply a stratified strategy, which consists of applying the random split to each class (Zollanvari, 2023).

In this study, both train-test splitting methods are employed in regression and classification tasks.

3.10.2 Performance Measures

This study implements different performance measures for evaluating the quality of the models, which depend on the type of supervised learning technique: regression or classification.

In the regression, the performance was evaluated using two measures: *RMSE* (Root Mean Square Error) and the R^2 (Coefficient of Determination). *RMSE* quantifies the performance of the predictions by measuring the deviation between predicted and reference values, with *RMSE* being more sensitive to larger errors due to its squared nature. R^2 indicates the proportion of variance in the reference data explained by the models as an indicator of overall predictive capability.

The *RMSE* is given by Equation 3.6:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{3.6}$$

where $\hat{y}_i$ is the predicted value by the model for sample i, y_i is the value obtained by the reference method for sample i, and *N* is the number of samples. The aim is to have RMSE values that are as low as possible.

The *RMSE* values are calculated for different data subsets, leading to variations such as *RMSECV* and *RMSEP*, depending on the dataset used in the calculation:

RMSECV (Root Mean Square Error of Cross-Validation): This is the *RMSE* obtained during cross-validation, where subsets of the training data are systematically left out and predicted by the model.

RMSEP (Root Mean Square Error of Prediction): This is the *RMSE* calculated using an independent test set, separate from the training dataset. *RMSEP* reflects the model's actual predictive performance on new samples and is typically higher than *RMSECV*.

The coefficient of determination R^2 is given by Equation 3.7:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.7}$$

The aim is to achieve values of R^2 near 1. The values of R^2 are also calculated for the train and the test data sets.

For classification models, the evaluation measures differ from those used for regression models. The most common measures are the accuracy and the confusion matrix, which is also a visualisation tool.

The accuracy quantifies the proportion of correctly classified observations out of the total number of observations. It is calculated as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ observations} \tag{3.8}$$

The confusion matrix evaluates the performance of a classification model by displaying the number of correct and incorrect predictions for each class. Important metrics can be derived from this matrix, including:

Correct Predictions per Class: Represented by the diagonal elements of the matrix, these values indicate the number of accurately classified samples for each class.

Incorrect Predictions per Class: The off-diagonal elements reveal misclassifications, showing samples where the model confused one class for another.

The confusion matrix offers valuable information about which classes are frequently misclassified, helping to identify patterns and areas for improvement in the model's performance.

3.10.3 Hyperparameter Tuning

Machine learning models can be represented by mathematical models with several parameters that must be learned from data. During training, these parameters are fine-tuned to fit the data correctly. However, there are additional parameters known as hyperparameters, which are not learned through the standard training process and influence the model's performance and complexity. These are configuration settings that control the learning process of the model (Zollanvari, 2023).

The process of determining the best values for the hyperparameters is called hyperparameter tuning, and the most used strategy is called Grid Search (Subasi, 2020).

In the strategy of Grid Search, the technique consists of creating a grid of potential values for each hyperparameter and training the model for each combination. It records the performance measures for each setup as well as the combination for best results. However, this technique can be slow and computationally intensive, since it requires training the model for every possible hyperparameter combination (Zollanvari, 2023).

3.10.4 Cross-Validation

Hyperparameter tuning requires defining a scoring measure to identify the best combination of hyperparameter values using the previously mentioned Grid Search procedure. The most common scoring measure for regression tasks is RMSE, while for classification, the standard scoring measures are the Accuracy or the Balanced Accuracy. These performance measures are typically calculated using k-fold cross-validation, which is especially recommended when working with small to medium datasets, as in this study (Bruce et al., 2020).

The training dataset is partitioned into k equally sized folds. In each of the k iterations, one fold is held out as the validation set while the remaining k–1 folds are used for training. This process is repeated k times so that every observation is used once for validation.

For regression problems, the RMSE is calculated for the k validation sets, and their average is computed, which is the final value of the model's performance.

Accuracy is the default scoring measure for classification tasks, which reflects the proportion of correct predictions. However, when dealing with imbalanced datasets, Balanced Accuracy is preferred, as it accounts for class imbalance by averaging across all classes. As for regression, these measures are calculated for k validation sets, and their average value is also the final value of the model's performance.

These approaches help minimise overfitting and ensure the models generalise better to unseen data.

3.11 Practical Test to Verify the Impact of the Models

A practical test was conducted to understand the impact that the classification models could have in a real situation. In this situation, the Return on Investment (*ROI*) is computed by trying to invest the same amount in the IPO that the models will predict to be underpriced. In this study, as referred to above, this practical test is performed with the last 30 observations that were previously selected from the original dataset. This is applied to the two scenarios (investors' and underwriters' predictive models, considering the alternatives of presence or absence of outliers) with the three different SMOTE settings.

The investment amount was divided by the offering price per share to compute the number of shares that the investor would want in that offering. Then comes one of the major factors in play while investing in an IPO: the allocation percentage. If the demand for a deal is high, the investor can only get a percentage of the shares he wants, which usually happens in IPOs that go on to be underpriced (informed investors better predict that and decide to invest).

On the other hand, if a deal is bad (overpriced), the informed investors wouldn't want to participate in it, and the allocation percentage would be higher. In this practical test, the allocation percentages were conducted in four cases, two that relate the offering price to the IPO range and the others that relate the first-day closing price to the same range. In all cases, percentages are allocated considering those relationships and are then multiplied by the number of shares that the investor would want to buy and rounded down to the nearest whole share to compute how many shares he would get.

With the number of shares computed and the underpricing known, it is possible to calculate the return that the investor would get from the IPO that he wishes to participate in (according to the predictive models), as well as the amount invested in them, allowing the computation of the *ROI*, given by Equation 3.9. The models and different cases are benchmarked against an investment in an all-offerings strategy that an uninformed investor could put into practice. As far as it is known, this part represents the innovation this work brings to the field, as it shows the benefit that these models can have on a real-life scenario, something that hasn't been done in the literature.

$$ROI = \left(\frac{\sum_{i=1}^{N} \text{Underpricing}_i \times \text{Offer Price}_i \times Shares_i}{\sum_{i=1}^{N} \text{Offer Price}_i \times Shares_i} \right) \times 100 \tag{3.9}$$

4. Results and Discussion

This chapter is divided into four main sections. Section 4.1 presents the main results of the preliminary statistical analysis of the involved variables, making a division between the original dataset (without the last 30 observations for the practical test), with and without outliers using Mahalanobis distance with a 5% significance level. The results are presented for the two scenarios: Investors' and Underwriters' Predictive Models. Section 4.2 summarises the performance measures of using regression to predict underpricing, considering different regression models. Section 4.3 shows the performance measures of classification models applied to only predict whether underpricing exists. Lastly, Section 4.4 presents the impact of using the generated classification models on the practical test to verify the impact of the models on the returns of an IPO investment strategy.

4.1 Preliminary Statistical Results

4.1.1 Descriptive Statistics

Firstly, the descriptive statistics were performed for Investors' and Underwriters' predictive models with outliers to comprehend the original dataset (excluding the 30 observations used for the practical test, as described in Section 3.5) and what adjustments were needed.

According to Table 3, the variables *DV*, *EBIT*, and *ASSETS* have very high ranges, meaning that their distribution is very dispersed. The target variable, Underpricing (*U*), presents a mean value of 0.289 and a minimum and maximum value of -0.459 and 2.011, respectively.

Table 3 – Descriptive statistics of the Investors' predictive models with outliers.

This table presents the descriptive statistics of the Investors' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Predicted Price (PP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 274 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was the original one, including outliers, but without the 30 observations used for the practical test. The information regarding the IPO was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	mean	std	min	25%	50%	75%	max
DV	4.716E+05	9.331E+05	3.163E+03	1.179E+05	2.224E+05	4.739E+05	1.193E+07
AS	0.215	0.138	0.022	0.120	0.192	0.275	1.000
EBIT	1.914E+07	3.747E+08	-1.021E+09	-4.020E+07	-1.350E+07	7.407E+06	4.450E+09
NPROF	-3.114E+07	2.487E+08	-1.618E+09	-4.783E+07	-1.747E+07	-1.493E+06	2.181E+09
ASSETS	1.061E+09	2.969E+09	1.500E+04	4.306E+07	1.453E+08	7.293E+08	2.732E+10
PP	20.413	18.401	3.500	15.000	17.000	20.000	250.000
ROA1	-1.375	9.305	-128.433	-0.505	-0.165	0.018	0.834
ROA2	-1.722	11.488	-128.433	-0.583	-0.204	-0.011	14.218
FS	18.735	2.427	9.616	17.578	18.795	20.408	24.031
MK RT	0.021	0.035	-0.108	0.002	0.025	0.043	0.121
U	0.289	0.412	-0.459	0.010	0.190	0.428	2.011

Table 4 presents the descriptive statistics of the Underwriters' predictive models for the same dataset. It includes all the variables that are present in Table 3 (except PP) and adds the variables that are known at the time of the IPO, such as the Offering Price (OP), Relative Range (RR), Over Range (OR), and Under Range (UR).

Table 4 – Descriptive statistics of the Underwriters' predictive models with outliers.

This table presents the descriptive statistics of the Underwriters' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Offering Price (OP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), RR (Relation to Range, as the relative difference between the OP and the middle point of the IPO range and the middle point of the same range), UR (Under Range, takes value 1 if OP is lower than the minimum end of the IPO range and 0 otherwise), OR (Over Range, takes value 1 if OP is higher than the maximum end of the IPO range and 0 otherwise), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 274 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was original, including outliers, but without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	mean	std	min	25%	50%	75%	max
DV	4.716E+05	9.331E+05	3.163E+03	1.179E+05	2.224E+05	4.739E+05	1.193E+07
AS	0.215	0.138	0.022	0.120	0.192	0.275	1.000
EBIT	1.914E+07	3.747E+08	-1.021E+09	-4.020E+07	-1.350E+07	7.407E+06	4.450E+09
NPROF	-3.114E+07	2.487E+08	-1.618E+09	-4.783E+07	-1.747E+07	-1.493E+06	2.181E+09
ASSETS	1.061E+09	2.969E+09	1.500E+04	4.306E+07	1.453E+08	7.293E+08	2.732E+10
OP	21.9	20.3	4.0	15.0	18.0	23.0	250.0
ROA1	-1.375	9.305	-128.433	-0.505	-0.165	0.018	0.834
ROA2	-1.722	11.488	-128.433	-0.583	-0.204	-0.011	14.218
FS	18.735	2.427	9.616	17.578	18.795	20.408	24.031
MK RT	0.021	0.035	-0.108	0.002	0.025	0.043	0.121
RR	0.049	0.136	-0.429	0.000	0.053	0.118	0.500
UR	0.084	0.278	0.000	0.000	0.000	0.000	1.000
OR	0.336	0.473	0.000	0.000	0.000	1.000	1.000
U	0.289	0.412	-0.459	0.010	0.190	0.428	2.011

The variables of the Underwriters' Predictive Models (Table 4) that are also included in the Investors' Models are present in the alternative with outliers, with the same distribution as in Table 3. The Predicted Price (*PP*) variable, which, in the first scenario, replaces the Offering Price (*OP*) (variable of the second scenario), shows a lower mean value, meaning that, on average, the offering prices of the IPOs are higher than the expected price of the IPOs given the ranges that are presented for them.

The application of Mahalanobis distance with a 5% significance level, as presented in Section 3.6, was implemented to identify outliers in both scenarios. This procedure removed 21 and 25 observations in the first and second scenario, respectively.

In the distribution of the Investors' Predictive Models variables (Table 5), the exclusion of the outliers did not produce significant changes, with the target variable (*U*) having a slightly higher mean value (0.296).

Table 5 – Descriptive statistics of the Investors' predictive models without outliers.

This table presents the descriptive statistics of the Investors' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Predicted Price (PP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 253 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	mean	std	min	25%	50%	75%	max
DV	3.558E+05	4.318E+05	3.163E+03	1.150E+05	2.131E+05	4.162E+05	3.360E+06
AS	0.208	0.109	0.022	0.126	0.192	0.275	0.658
EBIT	-9.700E+06	1.120E+08	-5.580E+08	-3.703E+07	-1.378E+07	6.074E+06	1.074E+09
NPROF	-3.020E+07	8.488E+07	-3.602E+08	-4.735E+07	-1.768E+07	-3.260E+06	6.070E+08
ASSETS	6.068E+08	1.143E+09	1.320E+05	4.405E+07	1.411E+08	6.482E+08	7.379E+09
PP	19.022	11.071	3.500	15.000	17.000	20.000	90.000
ROA1	-0.547	2.182	-32.750	-0.505	-0.178	0.018	0.834
ROA2	-0.641	2.243	-32.697	-0.589	-0.213	-0.012	0.345
FS	18.697	2.077	11.791	17.601	18.765	20.290	22.722
MK RT	0.021	0.035	-0.108	0.002	0.024	0.043	0.121
U	0.296	0.419	-0.459	0.020	0.213	0.438	2.011

From Table 6, it is observable that the elimination of outliers also did not significantly change the distribution of the model's variables, especially the target variable (*U*), which only saw small changes in the reference values and a slight increase in its mean value to 0.291. So, it can be concluded that the outliers may have eliminated some noise without significantly impacting the variables' distribution.

Table 6 – Descriptive statistics of the Underwriters' predictive models without outliers.

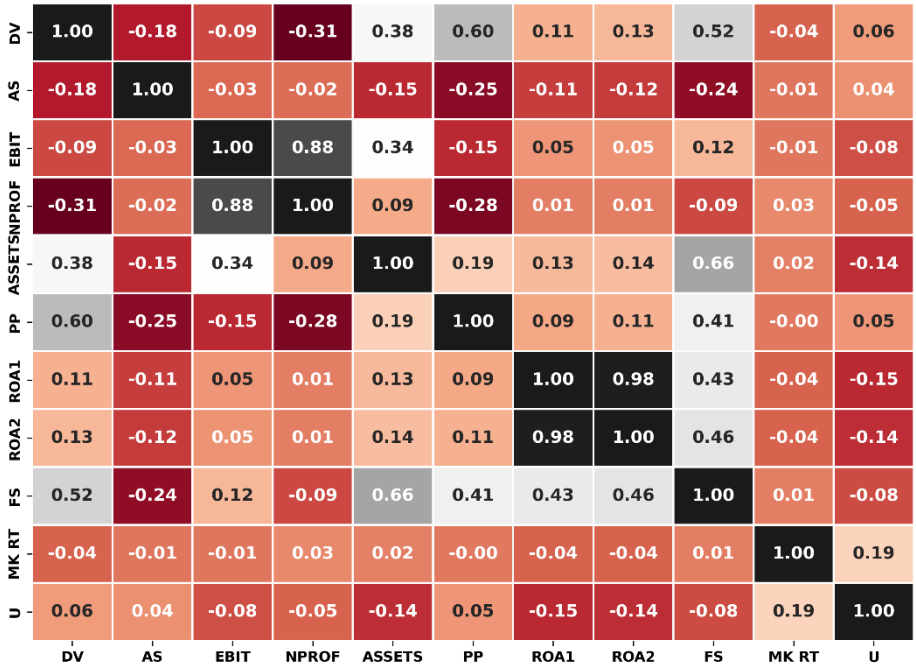
This table presents the descriptive statistics of the Underwriters' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Offering Price (OP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), RR (Relation to Range, as the relative difference between the OP and the middle point of the IPO range and the middle point of the same range), UR (Under Range, takes value 1 if OP is lower than the minimum end of the IPO range and 0 otherwise), OR (Over Range, takes value 1 if OP is higher than the maximum end of the IPO range and 0 otherwise), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 249 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	mean	std	min	25%	50%	75%	max
DV	3.612E+05	4.604E+05	3.163E+03	1.176E+05	2.131E+05	4.162E+05	4.200E+06
AS	0.208	0.112	0.022	0.126	0.192	0.274	0.801
EBIT	-1.423E+07	9.210E+07	-5.580E+08	-3.703E+07	-1.378E+07	6.074E+06	3.872E+08
NPROF	-3.253E+07	7.631E+07	-4.632E+08	-4.735E+07	-1.985E+07	-3.260E+06	2.186E+08
ASSETS	5.965E+08	1.103E+09	3.012E+05	4.405E+07	1.335E+08	6.467E+08	6.155E+09
OP	19.8	10.9	4.0	15.0	18.0	23.0	80.0
ROA1	-0.424	0.799	-5.636	-0.505	-0.178	0.018	0.834
ROA2	-0.519	0.974	-8.628	-0.589	-0.216	-0.012	0.345
FS	18.703	2.034	12.616	17.601	18.710	20.287	22.541
MK RT	0.022	0.034	-0.108	0.002	0.025	0.043	0.121
RR	0.050	0.130	-0.429	0.000	0.059	0.118	0.471
UR	0.072	0.259	0.000	0.000	0.000	0.000	1.000
OR	0.337	0.474	0.000	0.000	0.000	1.000	1.000
U	0.291	0.410	-0.459	0.021	0.213	0.434	2.011

4.1.2 Correlation Matrices

The correlation matrices were computed for Investors' and Underwriters' Predictive Models variables without outliers to evaluate the existence of linear relationships between explanatory variables and the target variable.

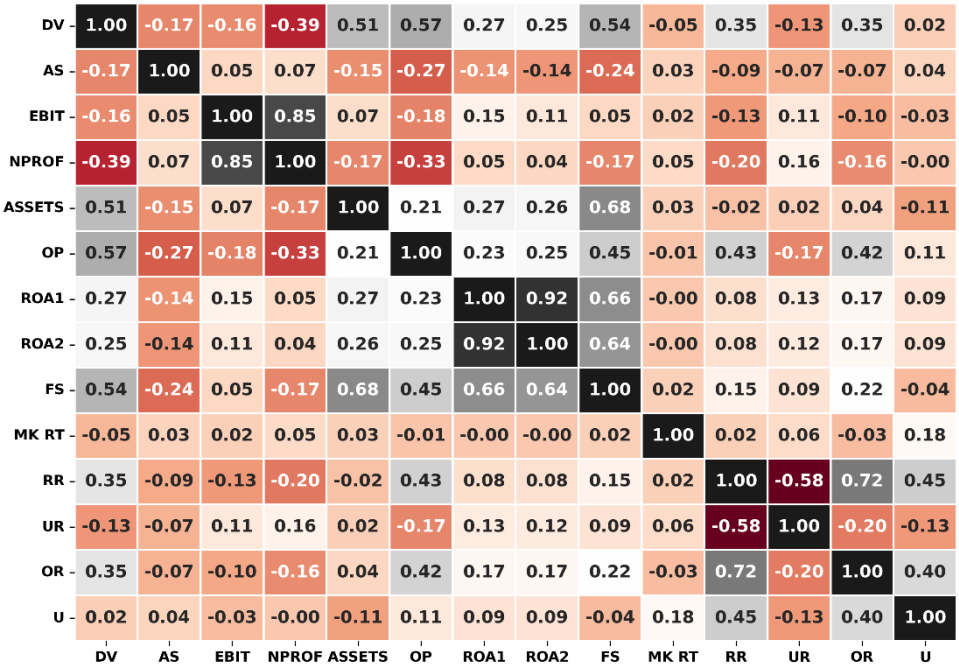
Figure 1 shows the correlation matrix for the Investors' Predictive Models variables without outliers. Considering correlation values greater than or equal to 0.5 and lower than or equal to -0.5, there are correlations between *DV* and *PP*, *DV* and *FS*, *EBIT* and *NPROF*, *ASSETS* and *FS*, *ROA1* and *ROA2*. There are no correlations between the target variable (using a continuous variable) and the explanatory variables with values greater than or equal to 0.5 and lower than or equal to -0.5.



This figure presents the correlation matrix of the Investors' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Predicted Price (PP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 253 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Figure 1 – Correlation matrix for Investors' predictive models' variables without outliers.

For the scenario of Underwriters' Predictive Models variables without outliers (Figure 2), and using the same threshold values, it is observed that correlations exist between *DV* and *ASSETS*, *DV* and *OP*, *DV* and *FS*, *EBIT* and *NPROF*, *ASSETS* and *FS*, *ROA1* and *ROA2*, *ROA1* and *FS*, *ROA2* and *FS*, *RR* and *UR*, and *RR* and *OR*. As in the former scenario, there are no correlations between the target variable (using a continuous variable) and the explanatory variables with values greater than or equal to 0.5 and lower than or equal to -0.5.



This figure presents the correlation matrix of the Underwriters' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Offering Price (OP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), RR (Relation to Range, as the relative difference between the OP and the middle point of the IPO range and the middle point of the same range), UR (Under Range, takes value 1 if OP is lower than the minimum end of the IPO range and 0 otherwise), OR (Over Range, takes value 1 if OP is higher than the maximum end of the IPO range and 0 otherwise), and U (Underpricing, relative difference between the OP and the first day closing price). The sample consists of 249 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Figure 2 – Correlation matrix for Underwriters' predictive models' variables without outliers.

Both Models show similar correlation patterns between their respective variables. From the variables present in both scenarios, the only correlation relationships that gain significance are the *DV* and *ASSETS*, *ROA1* and *FS* and *ROA2* and *FS*. Furthermore, the target variable does not present any significant (greater than the thresholds) relationships with any of its explanatory variables in any of the two scenarios.

4.1.3 Mutual information

The relevance of explanatory variables for a binary classification was estimated with the module of mutual information classification of Python's scikit-learn package, which was developed based on the work of Kraskov et al. (2004). This method estimates the mutual information between each explanatory variable and the target variable, capturing non-linear

dependencies without assuming any specific distribution. It allows us to identify which explanatory variables contribute most to distinguishing between the two classes, guiding feature selection and improving model interpretability. Unlike p-values, mutual information has no built-in significance level, so interpretation is relative, higher mutual information indicates stronger predictive power, and zero values mean the explanatory variables provide no information about the target.

Tables 7 and 8 present the relevance of the explanatory variables of Investors' and Underwriters' predictive models, respectively, with and without outliers.

Table 7 – Relevance of explanatory variables of Investors' predictive models, with and without outliers.

This table presents the relevance of the Investors' predictive models variables (estimated with the mutual information module of Python): Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Predicted Price (PP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), and U (Underpricing, relative difference between the OP and the first day closing price). The samples consist of the 274 US IPOs of the original dataset (WithO), and the 253 US IPOs obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (WithoutO), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	DV	**AS**	**EBIT**	**NPROF**	**ASSETS**	**PP**	**ROA1**	**ROA2**	**FS**	**MK RT**
WithO	0.021	0.000	0.020	0.020	0.000	0.000	0.000	0.000	0.000	0.000
WithoutO	0.023	0.000	0.035	0.019	0.005	0.020	0.000	0.001	0.000	0.000

According to Table 7, the most relevant explanatory variables are EBIT, DV and NPROF in both outlier alternatives, and PP and ASSETS in the alternative without outliers.

Table 8 – Relevance of explanatory variables of Underwriters' predictive models, with and without outliers.

This table presents the descriptive statistics of the Underwriters' predictive models variables: Deal Value (DV in thousand $), AS (Acquired Stake), EBIT (in $), Net Profit (NPROF in $), ASSETS (in $), IPO Offering Price (OP in $), ROA1(Return On Asset 1, as the relation between EBIT and ASSETS), ROA2 (Return On Asset 2, as the relation between NPROF and ASSETS), FS (Firm Size, as the logarithm of ASSETS), MK RT (Market Return, as the return of the S&P 500 in the prior 30 day period), RR (Relation to Range, as the relative difference between the OP and the middle point of the IPO range and the middle point of the same range), UR (Under Range, takes value 1 if OP is lower than the minimum end of the IPO range and 0 otherwise), OR (Over Range, takes value 1 if OP is higher than the maximum end of the IPO range and 0 otherwise), and U (Underpricing, relative difference between the OP and the first day closing price). The samples consist of the 274 US IPOs of the original dataset (WithO), and the 249 US IPOs obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (WithoutO), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

	DV	AS	EBIT	NPROF	ASSETS	OP	ROA1	ROA2	FS	MK RT	RR	UR	OR
WithO	0.021	0.000	0.020	0.020	0.000	0.023	0.000	0.000	0.000	0.000	0.041	0.000	0.032
WithoutO	0.029	0.000	0.038	0.022	0.008	0.027	0.007	0.013	0.000	0.000	0.049	0.000	0.031

According to Table 8, the most relevant explanatory variables are *RR*, *OR*, *EBIT*, *OP*, *DV*, *NPROF*, *ROA1*, *ASSETS* and *ROA2*, corroborating the findings of Basti et al. (2015), Dhini and Sondakh (2024) and Quintana et al. (2017).

Comparing the two scenarios, the Investors' predictive models do not include the two most relevant variables from the Underwriters' models (*RR* and *OR*), and from the variables included in both scenarios, the *ROA* gains importance in the second one (Underwriters') in the alternative without outliers. The *PP* (Predicted IPO Price) variable, which is a predicted proxy of the *OP* (Offering Price) variable included in the Underwriters' predictive models, only appears to be important in the alternative without outliers. From the analysis of the two scenarios, the conclusion that the Underwriters' predictive models are more complete than their counterparts (Investors' predictive models) can be reached due to the inclusion of more variables of significant importance. This was expected, as some of the Underwriters' model variables (OP, RR, OR and UR) consist of information that is only available later than the one used to compute the variables of the Investors' predictive models. So, it can be expected that the Underwriters' predictive models outperform the Investors' ones, something that will be discussed next.

4.2 Regression Models

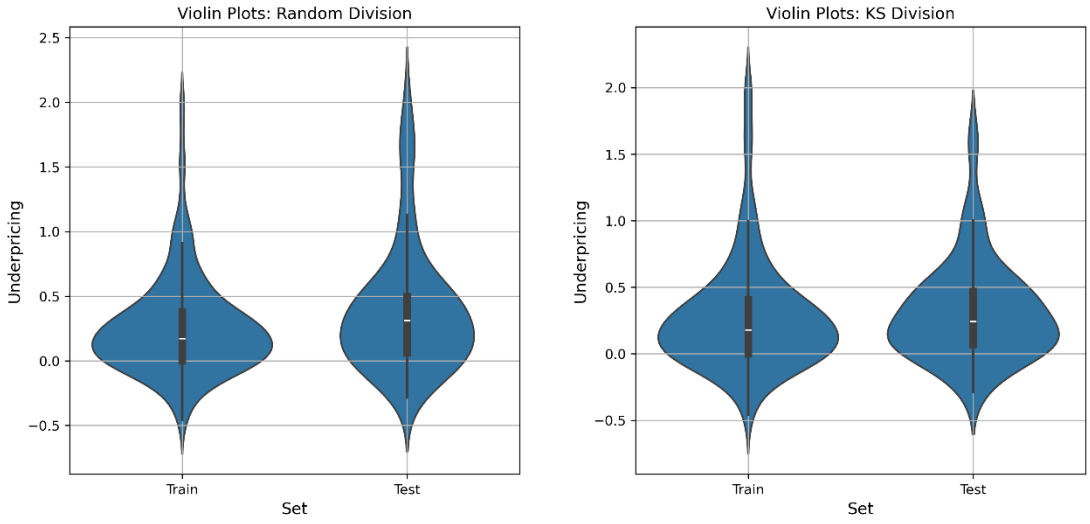
This section presents the main results achieved for predicting the Underpricing by applying the following regression models: Random Forest with original explanatory variables,

Random Forest, Support Vector Machines and Extreme Gradient Boosting with normalised explanatory variables.

The main goal here is to perform a base analysis of the potential of regression models to predict Underpricing. As referred to in Section 3.7, this study focuses on applying classification models due to their practical relevance for creating an investment strategy.

The regression models were developed for the scenario of Underwriters' predictive models, as it is the most complete scenario (as described in Section 4.1.3). The first step in generating these models is to perform the data division in the training and testing datasets, which in this study was 75% for training and 25% for testing.

According to Section 3.10.1, the Kennard-Stone method was implemented instead of random split division to ensure uniform coverage of the feature space, creating the two datasets with more similar distributions, as can be seen in Figure 3, which presents the violin charts of the target variable on training and testing datasets for the alternative with outliers.



This figure represents the violin charts of the training and test datasets for the two different test/train split methods. On the left, the distribution of the target variable (Underpricing) is presented when the random division method is applied to split the train and test datasets. On the right is presented the distribution of the target variable (Underpricing) when the Kennard-Stones (KS) division method is applied to split train and test datasets. The sample consists of 249 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Figure 3 – Violin charts of the target variable of training and testing datasets for the Underwriters' predictive models alternative with outliers.

Table 9 summarises the regression performance measures (R^2 and RMSE) for training and testing datasets for Underwriters' predictive models. It is worth noting that the models were developed using k-fold cross-validation (with k=5), as described in Section 3.10.4.

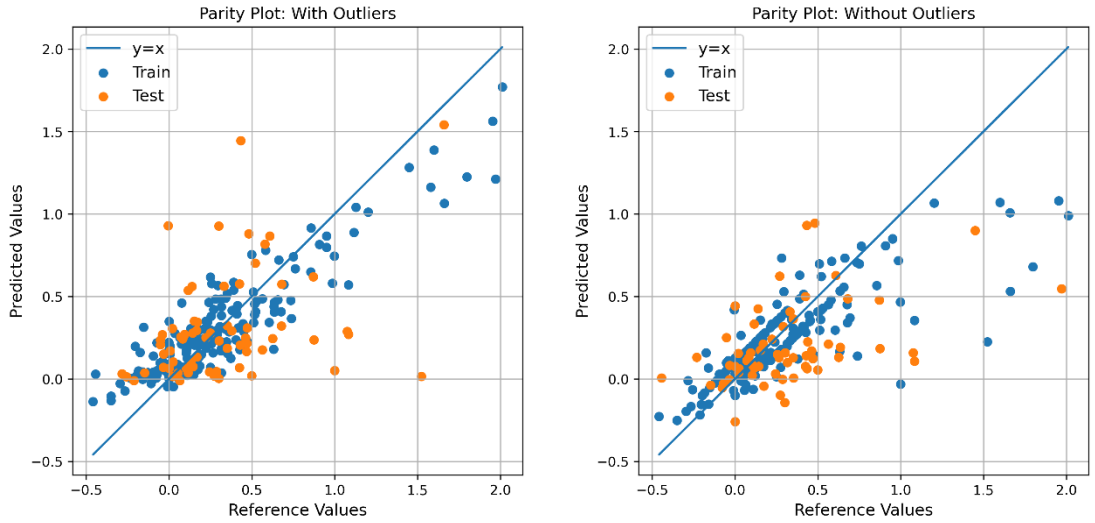
Table 9 – Values of R^2 and RMSE for training and testing datasets for Underwriters' predictive models.

This table presents the values of the R2 and the RMSE of the Train and Test (split on a 75/25 ratio using Kennard-Stone's) subsets of the Underwriters' predictive models: Random Forest (RF), RFN (Random Forest Normalised), Support Vector Machines in Regression (SVR) and Extreme Gradient Boosting. The samples consist of the 274 US IPOs of the original dataset (WithO), and the 249 US IPOs obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (WithoutO), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Accuracy		With Outliers				Without Outliers			
		R^2		RMSE		R^2		RMSE	
		Train	Test	Train	Test	Train	Test	Train	Test
Underwritters' Predictive Models	**RF**	0.80	-0.08	0.19	0.39	0.90	0.01	0.13	0.38
	RFN	0.80	-0.08	0.19	0.38	0.78	0.03	0.20	0.38
	SVR	0.59	0.10	0.27	0.35	0.62	0.09	0.26	0.37
	XGB	0.72	-0.09	0.22	0.39	0.80	0.03	0.19	0.38

In general, the values for R^2 and *RMSE* can be considered reasonable for all models in the training dataset. However, despite the use of the Kennard-Stone division technique, the models don't generalise well for the test dataset. The R^2 values in the testing dataset are close to 0, showing that there is no enhancement in considering these models instead of using the average value to estimate the Underpricing. In addition, the values of *RMSE* for the testing dataset are also significantly different and higher than the values obtained for the training dataset, which demonstrates the weaker predictive ability of the models. Nevertheless, these values are similar to those presented in the literature related to the subject of this study Quintana et al., 2017).

Furthermore, Figure 4 shows the parity plots to visualise the performance for the SVR model (model with the best *RMSE* values for the testing dataset) for the alternatives with and without outliers (left and right windows, respectively). Evidently, the models' bad performance is more specific to the testing dataset.



This figure presents the parity plots of the Underwriters' SVR (Support Vector Machines in Regression) predictive model. The Y vector is the values predicted by the model (Predicted Values), and the X vector is the actual values of the observations. The samples consist of the 274 US IPOs of the original dataset (With Outliers, on the left), and the 249 US IPOs obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (Without Outliers, on the right), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Figure 4 – Parity plots of the SVR models. Left: with outliers, right: without outliers.

4.3 Classification Models and Respective Accuracies

In the classification models, the dataset was divided into 75% for training and 25% for testing. As referred to in Chapter 3 (Section 3.10.3), the best hyperparameter values for each model are presented in Table 10 and were obtained using a 5-fold cross-validation, supported in the training dataset, with balanced accuracy as scoring. The hyperparameters presented in this table are the most relevant for each model. These values were found by implementing the grid search technique as described in Section 3.10.3.

Table 10 – Optimised hyperparameters for each model.

This table presents the best hyperparameters for the Investors' and Underwriters' predictive models for every machine learning method (RF - Random Forest; RFN- Random Forest Normalised; DT - Decision Trees; SVC - Support Vector Machines in Classification; XGB - Extreme Gradient Boosting), for every SMOTE setting (SM - Simple Model; ST - SMOTE in Training; SG - SMOTE Globally) and different outliers alternatives (With Outliers and Without Outliers). The samples consist of the 274 US IPOs of the original dataset (With Outliers), and the 253 and 249 US IPOs, for the Investors' and Underwriters' predictive models respectively, obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (Without Outliers), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Model	Hyperparameters	With Outliers						Without Outliers					
		Investors			Underwriters			Investors			Underwriters		
		SM	ST	SG	SM	ST	SG	SM	ST	SG	SM	ST	SG
RF	n_estimators	7	8	10	7	10	10	8	10	10	6	10	10
	max_depth	500	750	250	500	500	500	150	250	500	100	750	150
RFN	n_estimators	7	8	7	7	10	8	7	6	8	6	8	6
	max_depth	200	750	200	500	500	250	750	200	500	100	500	500
DT	ccp_alph	1.00E-10	1.00E-10	0.01	0.001	1.00E-14	1.00E-05	0.01	0.001	0.001	0.01	1.00E-10	1.00E-10
	max_depth	8	10	7	7	11	8	5	9	7	6	8	8
SVC	C	30	5	15	45	3	10	15	10	10	5	5	15
	gamma	0.1	1	1	0.1	2	0.85	0.5	0.5	0.9	1	1	0.9
XGB	learning_rate	0.5	0.05	0.05	0.5	0.05	0.5	0.05	0.05	0.05	0.1	0.05	0.05
	max_depth	2	11	10	3	6	10	5	10	10	10	3	5
	n_estimators	125	125	75	100	100	125	100	100	100	100	75	75
	subsample	0.5	0.7	1	0.5	0.9	0.5	1	1	1	0.5	1	0.5

After, the predictive abilities of the optimised models were measured by their accuracies (as given by Equation 3.8) for training and testing datasets. Table 11 shows the accuracy values of the different models in both scenarios, with and without outliers and considering the three different SMOTE settings.

Table 11 – Training and testing accuracies of the optimised models.

This table presents the accuracy (computed by the relation between the well predicted observations and the total number of observations) in both train and test (with a 75/25 split) for the Investors' and Underwriters' predictive models for every machine learning method (RF - Random Forest; RFN- Random Forest Normalised; DT - Decision Trees; SVC - Support Vector Machines in Classification; XGB - Extreme Gradient Boosting), for every SMOTE setting (SM - Simple Model; ST - SMOTE in Training; SG - SMOTE Globally) and different outliers alternatives (With Outliers and Without Outliers). The samples consist of the 274 US IPOs of the original dataset (With Outliers), and the 253 and 249 US IPOs, for the Investors' and Underwriters' predictive models respectively, obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one (Without Outliers), in the period between 11th of June 2020, and the 20th of June 2023, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Accuracy		With Outliers						Without Outliers					
		SM		ST		SG		SM		ST		SG	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Investors' Predictive Models	**RF**	0.97	0.83	1.00	0.64	1.00	0.72	1.00	0.84	1.00	0.75	1.00	0.92
	RFN	0.98	0.83	1.00	0.64	0.99	0.71	0.99	0.84	0.95	0.73	1.00	0.89
	DT	0.98	0.65	1.00	0.68	0.94	0.60	0.85	0.78	0.93	0.62	0.89	0.74
	SVC	0.88	0.86	0.97	0.75	0.98	0.77	0.98	0.70	0.96	0.64	0.98	0.88
	XGB	1.00	0.68	1.00	0.62	1.00	0.67	0.99	0.81	1.00	0.78	1.00	0.86
Underwritters' Predictive Models	**RF**	0.99	0.84	1.00	0.75	1.00	0.77	0.94	0.84	1.00	0.76	1.00	0.83
	RFN	0.99	0.84	1.00	0.75	1.00	0.76	0.94	0.84	0.98	0.79	0.96	0.82
	DT	0.95	0.80	0.99	0.65	0.99	0.72	0.85	0.81	0.94	0.71	0.94	0.74
	SVC	0.93	0.83	1.00	0.83	0.99	0.81	0.87	0.81	1.00	0.75	1.00	0.85
	XGB	1.00	0.71	1.00	0.72	1.00	0.74	0.98	0.79	0.94	0.73	0.97	0.85

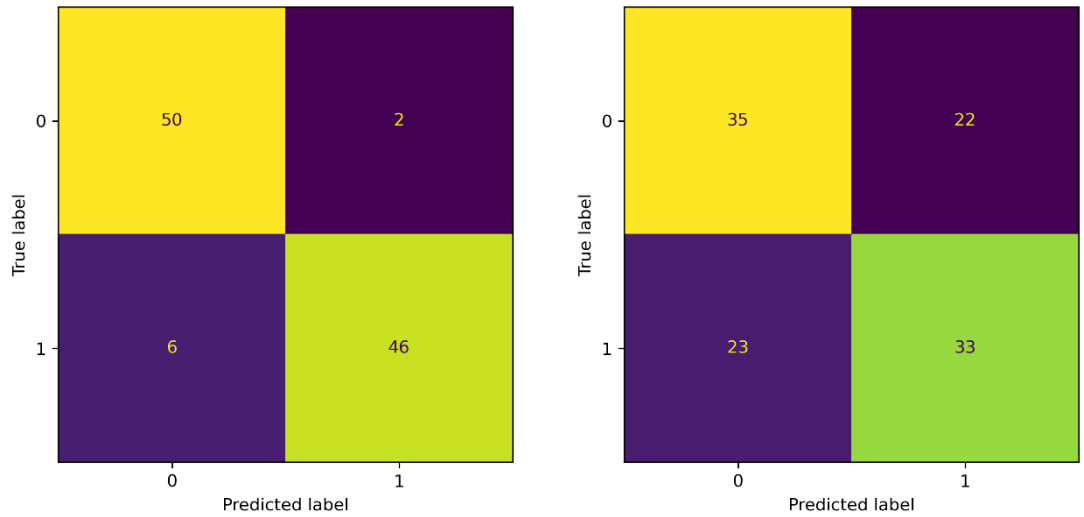
In training, all models for different settings demonstrate high predictive abilities, suggesting that they are learning the training data extremely well. However, this cannot be interpreted by itself to analyse the quality of the models. Table 11 shows the existence of high differences between training and testing accuracies, which may indicate overfitting, despite the procedure used to generate the models (cross-validation). This phenomenon occurs when a model is too tailored to the training set and fails to generalise to unseen data (testing dataset). It can create a misleading impression of model quality, looking like it is performing well, but, in fact, is memorising rather than learning, compromising its goal of performing well for different observations, possibly leading to poor decisions. This problem can be seen particularly in the ST SMOTE setting, as the training accuracy is 1.00 or almost 1.00 for all scenarios, but it generates way worse results in the testing accuracy.

The overall testing accuracy of the models improves when outliers are excluded. In the Investors' Models and Underwriters' Models, in only three and four instances, respectively, the accuracy does not improve compared to the respective instance with outliers. In the first scenario (Investors' Predictive Models), there is one instance where the accuracy increases

from 0.72 to 0.92 with the exclusion of outliers (Investors' Predictive Models SG RF: With Outliers-0.72; Without Outliers-0.92).

Under the best-performing settings (SG and without outliers), an unexpected outcome emerges: the Investors' Predictive Models consistently outperform the Underwriters' Models in terms of test accuracy, except for the Decision Tree model, where both achieve an accuracy of 0.74. This result was not expected, given that the Investors' Models include nearly all the variables used in the Underwriters' Models, excluding key features such as *RR* (Relative Range), *OR* (Over Range), and *UR* (Under Range), which were previously identified as important through mutual information analysis (Section 4.1.3). Additionally, the Investors' Models replace the *OP* variable (also deemed important) with *PP*, which only shows relevance when outliers are excluded. This relevance likely contributes to the notable performance variation observed between the Investors' Models with and without outliers.

An additional way to inspect the quality of these models is through the analysis of confusion matrices, which, for a binary classification, indicate 4 values corresponding to true negatives and false positives (1st row) and false negatives and true positives (2nd row). As examples, Figure 5 represents the confusion matrices of the best-performing test set (Investors' Predictive Model without Outliers RF SG, with testing accuracy of 0.92) (left window), and the worst-performing test set (Investors' Predictive Model with Outliers DT SG, with testing accuracy of 0.6) (right window), respectively.



This figure presents the best-performing test set scenario on the left, being the Investors' Random Forest (RF) predictive model with SMOTE applied to both train and test sets (SG), and without outliers. The sample consists of 253 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was obtained by excluding the outliers (using the Mahalanobis distance with a 5% significance) from the original one, without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

And the worst-performing test set scenario on the right, being the Investors' Decision Trees (DT) predictive model with SMOTE applied to both train and test sets (SG), and with outliers. The sample consists of 274 US IPOs in the period between the 11th of June 2020 and the 20th of June 2023. This sample was the original one, including outliers, but without the 30 observations used for the practical test. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus.

Figure 5 – Confusion matrices, left: best-performing test set for Investors' predictive model without outliers (model RF SG); right: worst-performing test set for Investors' predictive model with outliers (model DT SG).

For the first example (Figure 5, left), the model correctly predicts 96 observations, taking into consideration that due to the SMOTE, the number of observations in the training dataset was synthetically increased to produce a balanced set, with the same number of 0 and 1 observations. Within the incorrectly predicted observations (8), 6 are false negatives, which isn't as bad for the investor. Here, the model predicted that the observations were going to be overpriced (predicted label equal to 0) despite them being underpriced (true label equal to 1). In this case, the investor would lose an investment opportunity with a positive return rate. Considering the false positives, the model mispredicts 2 observations, where it predicts that they were going to be underpriced (predicted label equal to 1) despite them being overpriced (true label equal to 0). This would make the investors participate in those deals and provide negative return rates. In theory, in a bad IPO (an IPO that provides negative first-day return rates), investors get higher allocation percentages compared to the good ones (an IPO that

provides positive first-day return rates), making investments in bad IPO worse than not participating in good IPOs.

For the second example (Figure 5, right), the model also has SMOTE applied to both training and test (SG), and only predicts correctly 68 observations, mispredicting 45 observations, with 22 of them corresponding to false positives, which are the worst for investors.

4.4 Results for the Practical Test to Verify the Impact of the Models

The predictive models were applied to the last 30 IPO retrieved (as detailed in Section 3.5), evaluating the hypothetical returns that an investor would earn by using these models to select underpriced IPOs (as explained in Section 3.11). In this scenario, the investor is placed on June 29, 2023 (the date of the first IPO in the sample) and invests in the IPOs through to December 3, 2024 (the date of the last IPO in the sample of the 30 observations retrieved from the original dataset), trying to invest $1,000 only in those IPOs predicted to be underpriced.

In practice, one of the main challenges in IPO investing is the allocation percentages (the proportion of the desired investment amount an investor is able to acquire). If investors were able to obtain full allocations in each IPO and invested equal amounts across all of them, their returns would correspond to the average of the initial returns (average Underpricing), which can be quite high under certain circumstances.

However, as previously discussed in Section 3.11, allocation is typically inversely related to IPO quality. In high-demand IPOs (those with positive initial returns), investors generally receive smaller allocations due to oversubscription.

Thus, to compute the returns that the models could provide, four allocation cases, which take different allocation percentages regarding the relation between the offering price or the first day closing price and the range of the IPO (as mentioned in Section 3.11), were studied (Sections 4.4.1, 4.4.2, 4.4.3, and 4.4.4), some less optimistic and some more optimistic. In all cases, the returns provided by the models were benchmarked against an investment strategy of trying to invest $1,000 in every IPO (following the respective allocation metric), which mimics the strategy of an uninformed IPO investor.

4.4.1 First Allocation Case

In the first case, the allocation percentages are 100% if the offering price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point

of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 5% if the offering price is over the IPO range. Using these allocation percentages, an uninformed IPO investor would get an ROI of 2.33% by trying to invest $1,000 in each of the 30 IPOs between June 29, 2023 and December 3, 2024. This allows the understanding of how the results would react in a less optimistic scenario, with reduced exposure to IPOs with high demand.

The results of using models, taking into consideration the two different scenarios (Investors' Predictive Models and Underwriters' Predictive Models), the outlier alternatives (with and without outliers) and the three SMOTE settings (SM, ST and SG) in the third allocation case are presented in Table 12.

Table 12 – ROI of the different predictive models for different SMOTE settings with and without outliers in the first allocation case (with percentages of 5 %, 30 %, 50 %, 70 % and 100 %)

This table presents the returns of the Investors' and Underwriters' predictive models for the different machine learning methods (RF – Random Forest; RF – Random Forest Normalised; DT – Decision Trees; SVC – Support Vector Machine in Classification and XGB – Extreme Gradient Boosting), with and without outliers, and for the three different SMOTE settings (SM – Simple Model; ST – SMOTE in Training and SG – SMOTE Global) for the first allocation case. In this case the allocation percentages are: 100% if the offering price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 5% if the offering price is over the IPO range. The sample consists of the 30 observations retrieved from the original dataset, with US IPOs between 29th June 2023 and 3rd December 2024. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus. The values represent the ROIs and respective averages of the Investors' and Underwriters' predictive models, where the investor tries to invest $1,000 in each IPO he predicts to be underpriced, following the allocation percentages.

ROI (%)		With OP (First Allocation Case)						Average
		With Outliers			Without Outliers			
		SM	ST	SG	SM	ST	SG	
Investors' Predictive Models	RF	4.36	4.29	3.87	4.36	4.64	4.85	
	RFN	6.00	2.16	2.33	2.66	5.23	3.00	
	DT	5.95	0.89	-2.47	4.69	2.53	2.64	**4.34**
	SVC	4.45	5.12	8.03	**8.63**	7.39	7.16	
	XGB	2.97	2.95	6.61	4.64	4.94	5.21	
Underwritters' Predictive Models	RF	4.42	5.00	7.40	2.08	1.03	4.28	
	RFN	4.05	2.08	2.08	4.34	2.65	4.59	
	DT	2.74	3.79	4.97	4.64	6.03	8.59	**4.77**
	SVC	2.98	2.33	3.32	5.57	1.74	3.71	
	XGB	5.74	5.98	9.57	8.76	8.80	**9.97**	
Average			**4.13**			**4.98**		**4.56**

If the models are applied to the same 30 IPO and the investors try to invest $1,000 in those predicted to be underpriced, the average of the returns achieved is 4.56% (average of all values of Table 12), 95.71% higher than an uninformed investor (2.33%). In the best situation, the application of the model XGB provided a return of 9.97% (Underwriters' Predictive Models XGB Without Outliers SG), 327.90% higher than 2.33% (invest in all IPO strategy).

Focusing on the Investors' Predictive Models (scenario with the available information at the time for the investor to show its interest), as it is the most realistic scenario to implement an investment strategy, the models average a return of 4.34% (lower than the Underwriters' Predictive Models average of 4.77%), but 86.27% higher than the uninformed investor, and reaching under certain circumstances a return of 8.63% (Investors' Predictive Models SVC Without Outliers With OP SM), 270.39% higher than an uninformed investor (2.33%).

4.4.2 Second Allocation Case

In the second case, the same allocations are considered (as in Section 4.4.3), however, the market approach to the demand is used, replacing the offering price with the first-day closing price. Being the allocation percentages 100% if the first-day closing price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 5% if the first-day closing price is over the IPO range. Using these allocation percentages, an uninformed IPO investor would get an ROI of -7.27% by trying to invest $1,000 in each of the 30 IPOs between June 29, 2023 and December 3, 2024.

The results of using models, taking into consideration the two different scenarios (Investors' Predictive Models and Underwriters' Predictive Models), the outlier alternatives (with and without outliers) and the three SMOTE settings (SM, ST and SG) in the fourth allocation case are presented in Table 13.

Table 13 – ROI of the different predictive models for different SMOTE settings with and without outliers in the second allocation case (with percentages of 5 %, 30 %, 50 %, 70 % and 100 %).

This table presents the returns of the Investors' and Underwriters' predictive models for the different machine learning methods (RF – Random Forest; RF – Random Forest Normalised; DT – Decision Trees; SVC – Support Vector Machine in Classification and XGB – Extreme Gradient Boosting), with and without outliers, and for the three different SMOTE settings (SM – Simple Model; ST – SMOTE in Training and SG – SMOTE Global) for the second allocation case. In this case the allocation percentages are: 100% if the first-day closing price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 5% if the first-day closing price is over the IPO range. The sample consists of the 30 observations retrieved from the original dataset, with US IPOs between 29th June 2023 and 3rd December 2024. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus. The values represent the ROIs and respective averages of the Investors' and Underwriters' predictive models, where the investor tries to invest $1,000 in each IPO he predicts to be underpriced, following the allocation percentages.

ROI (%)		With FDCP (Second Allocation Case)						Average
		With Outliers			Without Outliers			
		SM	ST	SG	SM	ST	SG	
Investors' Predictive Models	RF	-6.48	-7.00	-6.35	-6.48	-5.81	-5.77	
	RFN	-6.68	-7.32	-7.27	-6.34	-5.26	-7.07	
	DT	-5.57	-6.41	-6.66	-4.41	-6.45	-6.80	**-5.61**
	SVC	-6.52	-5.53	-1.46	**-0.26**	-0.66	-2.42	
	XGB	-6.58	-8.03	-5.25	-6.44	-6.21	-4.79	
Underwritters' Predictive Models	RF	-6.15	-6.71	-1.97	-8.14	-8.35	-5.83	
	RFN	-5.79	-8.14	-8.14	-4.42	-7.58	-6.20	
	DT	-8.48	-6.15	-4.37	-6.44	-5.48	-1.98	**-5.60**
	SVC	-8.50	-7.27	-5.07	-4.42	-7.76	-5.75	
	XGB	-6.21	-6.23	-2.83	-1.95	**-0.51**	-1.31	
Average			**-6.17**			**-5.04**		**-5.61**

The models, despite producing better results (-5.61% on average) than what an uninformed investor would get (-7.27%), are still negative. Nonetheless, the application of the models enables investors to achieve better results than uninformed ones. In the best Underwriter's scenario, the model XGB can provide a return of -0.51% (Underwriters' Predictive Models XGB Without Outliers ST). The Investors' Predictive Models provided an average of returns of –5.61% (lower than the Underwriters' Predictive Models average of -5.60%), and reached, under the best scenario, a return of -0.26% (Investors' Predictive Models SVC Without Outliers SM).

The market approach to demand (using the first-day closing price), presented in Table 13, enables worse results than the approach using the offering price, presented in Table 12, as it diminishes the access to IPOs with high Underpricing levels, as they translate to a higher first-day closing price. An example of this is the Apogee Therapeutics Inc. IPO (one of the 30 observations of the practical test, which happened on July 14, 2023), where the range was

between $16 and $18, and the offering price was $17, meaning that in the first case, this IPO would allow an allocation percentage of 50%. So, if the models predict it to be underpriced, from the $1,000 that the investor aims to invest, divided by the price per share of $17, the investor would want 58 shares (58.8 rounded down), from which, following the allocation rule, he would only be able to buy 29 shares. In the second allocation case, the IPO experienced a significant level of Underpricing (24.9%), bringing the first day closing price to $21.23, over the range of the IPO, meaning that following the market approach of demand used in this case, the investor would only get an allocation percentage of 5%, being only able to buy 2 shares (2.9 rounded down). This example demonstrates that the market approach to demand used in the second allocation case negatively affects the returns, as the first day closing price is directly related to the Underpricing of the IPO, making the ones with greater Underpricing less available to investors due to lower allocation percentages.

4.4.3 Third Allocation Scenario

In the third case, the allocation percentages are 100% if the offering price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 15% if the offering price is over the IPO range. Using these allocation percentages, an uninformed IPO investor would get an ROI of 2.66% by trying to invest $1,000 in each of the 30 IPOs between June 29, 2023 and December 3, 2024.

The returns of the models, taking into consideration the two different scenarios (Investors' Predictive Models and Underwriters' Predictive Models), the outlier alternatives (with and without outliers) and the three SMOTE settings (SM, ST and SG) in the first allocation case are presented in Table 14.

Table 14 – ROI of the different predictive models for different SMOTE settings with and without outliers in the third allocation case (with percentages of 15 %, 30 %, 50 %, 70 % and 100 %).

This table presents the returns of the Investors' and Underwriters' predictive models for the different machine learning methods (RF – Random Forest; RF – Random Forest Normalised; DT – Decision Trees; SVC – Support Vector Machine in Classification and XGB – Extreme Gradient Boosting), with and without outliers, and for the three different SMOTE settings (SM – Simple Model; ST – SMOTE in Training and SG – SMOTE Global) for the third allocation case. In this case the allocation percentages are: 100% if the offering price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 15% if the offering price is over the IPO range. The sample consists of the 30 observations retrieved from the original dataset, with US IPOs between 29th June 2023 and 3rd December 2024. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus. The values represent the ROIs and respective averages of the Investors' and Underwriters' predictive models, where the investor tries to invest $1,000 in each IPO he predicts to be underpriced, following the allocation percentages.

ROI (%)		With OP (Third Allocation Case)						Average
		With Outliers			Without Outliers			
		SM	ST	SG	SM	ST	SG	
Investors' Predictive Models	RF	4.70	4.66	4.27	4.70	5.01	5.21	**4.70**
	RFN	6.37	2.49	2.66	3.03	5.59	3.33	
	DT	6.30	1.29	-2.47	5.17	3.12	3.01	
	SVC	4.80	5.48	8.37	**8.94**	7.76	7.51	
	XGB	3.40	3.35	6.97	4.99	5.35	5.54	
Underwritters' Predictive Models	RF	4.75	5.40	7.85	2.43	1.43	4.66	**5.16**
	RFN	4.45	2.43	2.43	4.73	3.00	4.97	
	DT	3.11	4.20	5.51	4.99	6.41	9.04	
	SVC	3.36	2.66	3.69	5.88	2.10	4.04	
	XGB	6.10	6.41	9.96	9.09	9.21	10.37	
Average			**4.50**			**5.35**		**4.93**

The Investors' Predictive Models achieved an average return of 4.70% (lower than the Underwriters' Predictive Models average of 5.16%), 76.69% higher than an investment in all IPO strategy, which in this case would yield a return of 2.66%. The optimal circumstances in this scenario would allow the investor to achieve a return of 8.94% (Investors' Predictive Models SVC Without Outliers SM), 236.09% higher than the uninformed investor.

From Table 14, it is observable that the returns are, as expected, higher than in Table 12, due to more exposure to high-demand IPOs.

4.4.4 Fourth Allocation Case

In the fourth case, the same allocations are considered, however, there is more of a market approach to the demand, replacing the offering price with the first-day closing price. Being the allocation percentages 100% if the first-day closing price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO

range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 15% if the first-day closing price is over the IPO range. Using these allocation percentages, an uninformed IPO investor would get an ROI of -4.11% by trying to invest $1,000 in each of the 30 IPOs between June 29, 2023 and December 3, 2024.

The results of using models, taking into consideration the two different scenarios (Investors' Predictive Models and Underwriters' Predictive Models), the outlier alternatives (with and without outliers) and the three SMOTE settings (SM, ST and SG) in the second allocation case are presented in Table 15.

Table 15 – ROI of the different predictive models for different SMOTE settings with and without outliers in the fourth allocation case (with percentages of 15 %, 30 %, 50 %, 70 % and 100 %).

This table presents the returns of the Investors' and Underwriters' predictive models for the different machine learning methods (RF – Random Forest; RF – Random Forest Normalised; DT – Decision Trees; SVC – Support Vector Machine in Classification and XGB – Extreme Gradient Boosting), with and without outliers, and for the three different SMOTE settings (SM – Simple Model; ST – SMOTE in Training and SG – SMOTE Global) for the fourth allocation case. In this case the allocation percentages are: 100% if the first-day closing price is lower than the minimum end of the IPO range, 70% if it is between the minimum end and the middle point of the IPO range, 50% if it is in the middle point of the IPO range, 30% if it is between the middle point and the maximum end of the IPO range, and 15% if the first-day closing price is over the IPO range. The sample consists of the 30 observations retrieved from the original dataset, with US IPOs between 29th June 2023 and 3rd December 2024. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus. The values represent the ROIs and respective averages of the Investors' and Underwriters' predictive models, where the investor tries to invest $1,000 in each IPO he predicts to be underpriced, following the allocation percentages.

ROI (%)		With FDCP (Fourth Allocation Case)						Average
		With Outliers			Without Outliers			
		SM	ST	SG	SM	ST	SG	
Investors' Predictive Models	RF	-2.89	-3.28	-2.80	-2.89	-2.06	-2.09	
	RFN	-2.35	-4.26	-4.11	-3.16	-1.68	-3.51	
	DT	-1.77	-5.09	-5.94	-0.68	-2.58	-3.68	**-2.19**
	SVC	-2.84	-2.15	2.24	**3.32**	2.58	1.01	
	XGB	-3.40	-4.17	-0.99	-2.76	-2.44	-1.30	
Underwritters' Predictive Models	RF	-2.68	-2.31	2.36	-4.73	-5.10	-2.32	
	RFN	-2.30	-4.73	-4.73	-1.09	-3.96	-2.29	
	DT	-4.67	-2.52	-1.46	-2.76	-0.47	2.32	**-1.81**
	SVC	-4.59	-4.11	-1.97	-1.05	-4.59	-2.44	
	XGB	-2.22	-1.72	2.08	2.03	4.04	3.80	
Average		**-2.65**			**-1.35**			**-2.00**

From Table 15, it is observable that the returns are, as expected, higher than in Table 13, due to more exposure to highly underpriced IPOs.

The Investors' Predictive Models achieve a return average of -2.19% (lower than the Underwriters' Predictive Models average of -1.81%), and reach, under the best scenario, a return of 3.32% (Investors' Predictive Models SVC Without Outliers SM).

The Underwriters' Predictive Models, contrary to what is described in Section 4.3, outperform the Investors' ones, behaving as expected, due to the inclusion of more important variables (described in Section 4.1.3) in the former. The Underwriter's Models achieve, on average, better ROI results than the Investors' ones in all four allocation cases, as presented in Table 16. The exclusion of outliers is also helpful to the model's performance, as the averages of ROI without Outliers are superior to their respective counterparts (including Outliers) in all four cases, as well.

Table 16 – Summary table of the ROIs of Uninformed Investor, and average and best scenario of Investors' and Underwriters' predictive models for the four allocation cases.

This table summarises Section 4.4, exhibiting the ROIs obtained by an Uninformed Investor, and the average and best scenario of the Investors' and Underwriters' predictive models for the four distinct allocation cases. The sample consists of the 30 observations retrieved from the original dataset, with US IPOs between 29th June 2023 and 3rd December 2024. The information regarding the IPOs was retrieved from the Orbis M&A website, and complemented by Yahoo Finance, IPO Scoop, and eventual news and IPO prospectus. The uninformed Investor tries to invest $1,000 in each of the 30 IPOs following the allocation rule of the respective allocation case, and the average and best scenario consists of the average and highest value of the ROIs of the Investors' and Underwriters' predictive models for each case, where the investor tries to invest $1,000 in each IPO he predicts to be underpriced.

ROI (%)	Uninformed Investor	Average		Best Scenario	
		Investors' Predictive Models	Underwritters' Predictive Models	Investors' Predictive Models	Underwritters' Predictive Models
First Case	2.33	4.34	4.77	8.63	9.97
Second Case	-7.27	-5.61	-5.60	-0.26	-0.51
Third Case	2.66	4.70	5.16	8.94	10.37
Fourth Case	-4.11	-2.19	-1.81	3.32	4.04

From Table 16, it can be concluded that the predictive models enable better results than what an uninformed investor could achieve in all four allocation cases, reaching in the best scenario an improvement of 7.71 p.p. (on the third case using the Underwriters' predictive models; 10.37-2.66=7.71 p.p.). Focusing on the more realistic models for the investors, investors' predictive models, it is observable that they enable better results than an uninformed investor, answering positively to the second research question of this study.

5. Conclusion

This study is situated in the predictive analysis niche of the extensively studied area of IPO Underpricing and builds upon the Quintana et al. (2017) predictive study of US IPO Underpricing. This work overcomes the main limitations identified by the authors and builds a practical test to compute what returns an uninformed retail IPO investor could achieve by implementing a machine learning approach. The USA were chosen because, despite being a developed country, the Underpricing levels are still quite high (enabling good returns) and the number of deals is significant.

In this study, some machine learning techniques (Random Forrest, Random Forrest Normalised, Decision Trees, Support Vector Machines and Extreme Gradient Boosting) are applied to the 304 US IPO observations retrieved. The models were applied in two approaches: regression, where the predicted variable is continuous, and classification, where the predicted variable is binary, taking the value 1 if underpriced and 0 otherwise.

In the regression approach to predict the exact extent of the IPO Underpricing (predicted continuous variable), the models were weak. The best-performing model, in testing, only got an R^2 of 0.10 and an RMSE of 0.35. Despite being similar to the values achieved in other studies (Quintana et al., 2017), the models show weak predictive ability. Nonetheless, the study gave more emphasis to the classification models, as for an investor, the most important question is whether an IPO will be underpriced (answering question 1).

In the classification approach, the techniques are applied in two scenarios (Investors' and Underwriters' Predictive Models), with two outlier alternatives (with and without outliers) and three SMOTE settings (SM, ST and SG). The alternative without outliers and the SG SMOTE setting performed better than their counterparts, as expected, reaching a maximum testing accuracy of 0.92. However, within this combination, something unexpected happened, the Investors' Predictive Models outperformed the Underwriters' ones, despite not having the two most significant variables (RR and OR) following the mutual information analysis.

The classification models were used to conduct the practical test, where they were applied to the last 30 retrieved observations to simulate an IPO investment strategy (invest $1000 in the IPOs predicted to be underpriced). Due to the ambiguous nature of IPO allocation percentages (inversely related to the quality of the IPO), four allocation cases were computed. Two relate the Offering Price and the IPO range, and the others relate the first day closing

price with the same IPO range, with two different allocation percentages (the second with higher exposure to the IPOs with greater demand). Under the best circumstances, the best-performing model achieved a return of 10.37% (7.71 p.p. higher than the base case, of an uninformed investor, for the same allocation case). In this test, the results are as expected, better in the alternative without outliers, and the Underwriters' Predictive Models outperform the Investors' ones. This practical test not only demonstrates the real-life impact that these models can have, but it also answers question 2.

The main limitations of this work are the relatively small dataset on which it is applied (304 observations) and the lack of some important financial variables. The Orbis M&A database from which the observations were extracted presented some problems, such as missing values from needed variables in certain observations, and the presence of a majority of Unit-IPO (which were excluded from the study), which led to the elimination of more than half of the retrieved dataset (636 observations were originally retrieved). The database also did not include any financial variables directly related to the debt of the companies that underwent the IPOs. The debt-related variables are a crucial aspect of the financials of a company, and are present in similar literature (Basti et al., 2015; Esfahanipour et al., 2016; Dhini & Sondakh, 2024), their absence being the major limitation of this work, as the model's performance would certainly benefit from their inclusion.

This work presents an innovation to the IPO prediction literature, specifically the one focused on the US market, building on the study of Quintana et al. (2017). This study implements the innovative approach of a practical test in which allocation percentages (one of the most relevant and ambiguous aspects of IPO investing) are considered. As a pioneer in this approach, the work presents many future research opportunities, starting with overcoming its above-mentioned limitations (a more extensive dataset and inclusion of debt-related explanatory variables) and a deeper study of the ambiguous nature of IPO allocation, especially to retail investors, as this study implemented a very simplistic one, enabling more realistic results' simulations.

6. References

Alanazi, A.S., & Al-Zoubi, H.A. (2015). Extreme IPO underpricing and the legal environment in wealthy emerging economies. *Journal of Multinational Financial Management,* 31, 83-103. https://doi.org/10.1016/j.mulfin.2015.05.004

Allen, F., & Faulhaber, G. R. (1989). Signalling by underpricing in the IPO market. *Journal of Financial Economics*, 23(2), 303-323. https://doi.org/10.1016/0304-405X(89)90060-3

Amihud, Y., Hauser, S., & Kirsh, A. (2003). Allocations, adverse selection, and cascades in IPOs: Evidence from the Tel Aviv stock exchange. *Journal of Financial Economics*, 68(1), 137-158. https://doi.org/10.1016/S0304-405X(02)00251-9

Baba, B., & Sevil, G. (2020). Predicting IPO initial returns using random forest. *Borsa Istanbul Review*, *20*(1), 13-23. https://doi.org/10.1016/j.bir.2019.08.001

Basti, E., Kuzey, C., & Delen, D. (2015). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems*, 73, 15-27. https://doi.org/10.1016/j.dss.2015.02.011

Beatty, R. P., & Ritter, J. R. (1986). Investment banking, reputation, and the underpricing of initial public offerings. *Journal of Financial Economics*, *15*(1-2), 213-232. https://doi.org/10.1016/0304-405X(86)90055-3

Beatty, R. P., & Welch, I. (1996). Issuer expenses and legal liability in initial public offerings. The Journal of Law and Economics, 39(2), 545-602. https://doi.org/10.1086/467359

Benveniste, L. M., & Spindt, P. A. (1989). How investment bankers determine the offer price and allocation of new issues. *Journal of Financial Economics*, *24*(2), 343-361.

Boonchuaymetta, E., & Chuanrommanee, W. (2013). Management of the IPO performance in Thailand. *Journal of Multinational Financial Management*, 23(4), 272-284. http://dx.doi.org/10.1016/j.mulfin.2013.05.004

Booth, J. R., & Chua, L. (1996). Ownership dispersion, costly information, and IPO underpricing. *Journal of Financial Economics*, *41*(2), 291-310.

Booth, J. R., & Smith II, R. L. (1986). Capital raising, underwriting and the certification hypothesis. *Journal of Financial Economics*, 15(1-2), 261-281. https://doi.org/10.1016/0304-405X(86)90057-7

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media. https://doi.org/10.1080/00401706.2021.1904738

Carter, R., & Manaster, S. (1990). Initial public offerings and underwriter reputation. *The Journal of Finance*, 45(4), 1045-1067. https://doi.org/10.1111/j.1540-6261.1990.tb02426.x

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. https://doi.org/10.1613/jair.953

Chhabra, S., & Kiran, R. (2022). Impact of information on winners' curse and long run performance of initial public offerings. *International Journal of Finance & Economics*, 27(1), 975-992. https://doi.org/10.1002/ijfe.2186

Chorruk, J., & Worthington, A.C. (2010). New evidence on the pricing and performance of initial public offerings in Thailand, 1997–2008. *Emerging Markets Review*, 11(3), 285-299. http://dx.doi.org/10.2139/ssrn.1421598

Coakley, J., Hadass, L., & Wood, A. (2009). UK IPO underpricing and venture capitalists. *The European Journal of Finance*, 15(4), 421-435. http://dx.doi.org/10.1080/13518470802560915

Daily, C. M., Certo, S. T., Dalton, D. R., & Roengpitya, R. (2003). IPO underpricing: A meta–analysis and research synthesis. *Entrepreneurship Theory and Practice*, *27*(3), 271-295. https://doi.org/10.1111/1540-8520.t01-1-00004

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems,* 50(1), 1-18. https://doi.org/10.1016/S0169-7439(99)00047-7

Deb, V.B. Marisetty, Information content of IPO grading, *Journal of Banking & Finance*, 34 (9) (2010) 2294–2305. https://doi.org/10.1016/j.jbankfin.2010.02.018

Dhini, A., & Sondakh, L. (2024, February). Predicting stock return of initial public offering in Indonesia stock exchange. In AIP Conference Proceedings (Vol. 2710, No. 1). AIP Publishing. https://doi.org/10.1063/5.0145600

Esfahanipour, A., Goodarzi, M., & Jahanbin, R. (2016). Analysis and forecasting of IPO underpricing. *Neural Computing and Applications*, 27, 651-658. https://doi.org/10.1007/s00521-015-1884-1

Francis, B. B., Hasan, I., Lothian, J. R., & Sun, X. (2010). The signaling hypothesis revisited: Evidence from foreign IPOs. *Journal of Financial and Quantitative Analysis*, 45(1), 81-106. https://doi.org/10.1017/S0022109010000037

Grammenos, C. T., & Papapostolou, N. C. (2012). US shipping initial public offerings: do prospectus and market information matter?. *Transportation Research Part E: Logistics and Transportation Review*, *48*(1), 276-295. https://doi.org/10.1016/j.tre.2011.07.009

Habib, M. A., & Ljungqvist, A. P. (2001). Underpricing and entrepreneurial wealth losses in IPOs: Theory and evidence. The Review of Financial Studies, 14(2), 433-458. https://doi.org/10.1093/rfs/14.2.433

Hanley, K. W. (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. *Journal of Financial Economics*, *34*(2), 231-250. https://doi.org/10.1016/0304-405X(93)90019-8

Hensler, D. A. (1995). Litigation costs and the underpricing of initial public offerings. *Managerial and Decision Economics*, 16(2), 111-128. https://doi.org/10.1002/mde.4090160203

Hughes, P. J., & Thakor, A. V. (1992). Litigation risk, intermediation, and the underpricing of initial public offerings. *The Review of Financial Studies*, 5(4), 709-742. https://doi.org/10.1093/rfs/5.4.709

Ibbotson, R. G., Sindelar, J. L., & Ritter, J. R. (1994). The market's problems with the pricing of initial public offerings. *Journal of Applied Corporate Finance*, *7*(1), 66-74. https://doi.org/10.1111/j.1745-6622.1994.tb00395.x

Ibbotson, R.G. (1975). Price performance of common stock new issues. *Journal of Financial Economics*, 2(3), 235-272. https://doi.org/10.1016/0304-405X(75)90015-X

IPOScoop.com LLC. (2025). *IPOScoop*. Retrieved June 24, 2025, from https://www.iposcoop.com/

Islam, M.A., Ali, R., & Ahmad, Z. (2010). An empirical investigation into the underpricing of initial public offerings in the Chittagong Stock Exchange. *International Journal of Economics and Finance*, 2(4), 36. https://doi.org/10.5539/ijef.v2n4p36

Jain, B. A. (1994). The underpricing of 'unit' initial public offerings. *The Quarterly Review of Economics and Finance*, *34*(3), 309-325. https://doi.org/10.1016/1062-9769(94)90030-2

Keasey, K., & Short, H. (1992). The Winner's curse model of underpricing: A critical assessment. *Accounting and Business Research*, 23(89), 74-78. https://doi.org/10.1007/BF01074315

Keloharju, M. (1993). The Winner's curse, legal liability, and the long-run price performance of initial public offerings in Finland. *Journal of Financial Economics*, 34(2), 251-277. https://doi.org/10.1016/0304-405X(93)90020-C

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*(1), 137-148. https://doi.org/10.2307/1266770

Kennedy, D. B., Sivakumar, R., & Vetzal, K. R. (2006). The implications of IPO underpricing for the firm and insiders: Tests of asymmetric information theories. *Journal of Empirical Finance*, *13*(1), 49-78. https://doi.org/10.1016/j.jempfin.2004.10.003

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, *69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

Leone, A. J., Rock, S., & Willenborg, M. (2007). Disclosure of intended use of proceeds and underpricing in initial public offerings. *Journal of Accounting Research*, 45(1), 111-153. https://doi.org/10.1111/j.1475-679X.2006.00229.x

Loughran, T., & Ritter, J. (2004). Why has IPO underpricing changed over time?. *Financial Management*, 5-37. http://dx.doi.org/10.2139/ssrn.331780

Loughran, T., Ritter, J. R., & Rydqvist, K. (1994). Initial public offerings: International insights. (updated version of 2024)

Lowry, M., & Shu, S. (2002). Litigation risk and IPO underpricing. *Journal of Financial Economics*, 65(3), 309-335. https://doi.org/10.1016/S0304-405X(02)00144-7

Marisetty, V. B., & Subrahmanyam, M. G. (2010). Group affiliation and the performance of IPOs in the Indian stock market. *Journal of Financial Markets*, 13(1), 196-223. https://doi.org/10.1016/j.finmar.2009.09.001

Moody's Corporation. (2025). *Orbis M&A: Deals list* [Database page]. Retrieved February 10, 2025, from https://orbismanda-r1.bvdinfo.com/version-20250320-1-0/OrbisMA/1/Deals/List

Pollock, T. G., Rindova, V. P., & Maggitti, P. G. (2008). Market watch: Information and availability cascades among the media and investors in the US IPO market. *Academy of Management Journal*, 51(2), 335-358. http://dx.doi.org/10.5465/AMJ.2008.31767275

Quintana, D., Sáez, Y., & Isasi, P. (2017). Random forest prediction of IPO underpricing. *Applied Sciences*, 7(6), 636. http://dx.doi.org/10.3390/app7060636

Ritter, J. (2011). Equilibrium in the initial public offerings market. *Annual Review of Financial Economics*, 3(1), 347-374. http://dx.doi.org/10.1146/annurev-financial-102710-144845

Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. *The Journal of Finance*, 57(4), 1795-1828. https://doi.org/10.1111/1540-6261.00478

Rock, K. (1986). Why new issues are underpriced. *Journal of Financial Economics*, 15(1-2), 187-212. https://doi.org/10.1016/0304-405X(86)90054-1

Ruud, J. S. (1993). Underwriter price support and the IPO underpricing puzzle. *Journal of Financial Economics*, 34(2), 135-151. https://doi.org/10.1016/0304-405X(93)90015-4

Sahoo, S., & Rajib, P. (2010). After market pricing performance of initial public offerings (IPOs): Indian IPO market 2002–2006. *Vikalpa*, 35(4), 27-44. http://dx.doi.org/10.1177/0256090920100403

Subasi, A. (2020). Practical machine learning for data analysis using python. *Academic Press*. https://doi.org/10.1016/C2019-0-03019-1

Tian, L. (2011). Regulatory underpricing: Determinants of Chinese extreme IPO returns. *Journal of Empirical Finance*, 18(1), 78-90. https://doi.org/10.1016/j.jempfin.2010.10.004

Tinic, S. M. (1988). Anatomy of initial public offerings of common stock. *The Journal of Finance*, 43(4), 789-822. https://doi.org/10.1111/j.1540-6261.1988.tb02606.x

U.S. Securities and Exchange Commission. (n.d.). *EDGAR: Electronic Data Gathering, Analysis, and Retrieval System* [Database]. Retrieved February 10, 2025, from https://www.sec.gov/edgar.shtml

Utamaningsih, A. (2017). The Influence of Underpricing to IPO Aftermarket Performance: Comparison between Fixed Price and Book Building System on the Indonesia Stock Exchange. *International Journal of Economics and Financial Issues*, 7(4), 157-161. https://www.econjournals.com/index.php/ijefi/article/view/4564

Vong, A. P., & Trigueiros, D. (2010). The short-run price performance of initial public offerings in Hong Kong: New evidence. *Global Finance Journal*, *21*(3), 253-261. http://dx.doi.org/10.1016/j.gfj.2010.09.003

Welch, I. (1989). Seasoned offerings, imitation costs, and the underpricing of initial public offerings. *The Journal of Finance*, 44(2), 421-449. https://doi.org/10.1111/j.1540-6261.1989.tb05064.x

Welch, I. (1992). Sequential sales, learning, and cascades. *The Journal of Finance*, 47(2), 695-732. https://doi.org/10.1111/j.1540-6261.1992.tb04406.x

Yahoo Finance. (n.d.). *S&P 500 Index (^GSPC) historical data*. Yahoo. Retrieved February 10, 2025, from https://finance.yahoo.com/quote/^GSPC/history

Zheng, S. X., Ogden, J. P., & Jen, F. C. (2005). Pursuing value through liquidity in IPOs: underpricing, share retention, lockup, and trading volume relationships. *Review of Quantitative Finance and Accounting,* 25, 293-312. DOI:10.1007/s11156-005-4769-z

Zollanvari, A. (2023). Machine Learning with Python: Theory and Implementation. Springer Nature https://doi.org/10.1007/978-3-031-33342-2