Who issues stock? Insights from predicting SEOs using machine learning

A dissertation presented in part fulfilment of the requirements for the degree of *Master of Commerce* in the Department of Accounting and Finance at The University of Auckland.

Zachariah Ryan

2021

Table of Contents

List of Figures

List of Tables

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and guidance. I want to thank the following people who have been instrumental in helping me to complete my research.

I would first like to thank my supervisor, Dr Paul Geertsema, whose expertise in using machine learning in finance research was invaluable throughout the entirety of the research process. From the formulation of the research questions to the construction of the algorithms, this project's completion would not have been possible without Dr Geertsema's supervision. I also appreciate Dr Geertsema's willingness to extensively discuss any issues that I faced and offer insightful feedback. Additionally, completing my research would not have been nearly as enjoyable without Dr Geertsema's passion for the field, which was a source of inspiration.

I would also like to thank Professor Henk Berkman for his help at the beginning of the project. His research expertise and comprehensive instruction allowed me to explore a novel research area while ensuring I could contribute meaningfully to the field. I also value the countless discussions that I have had with Professor Berkman over the past year, which have given me a much better understanding of finance and academia in general.

Abstract

I used gradient boosted machine (GBM), a machine learning (ML) algorithm, to predict whether a firm will engage in an seasoned equity offering (SEO), using lagged data that would have been available at the time of prediction. The GBM model achieved an AUC score of 0.78, substantially outperforming a logistic regression model. The outperformance suggests that the GBM model can make sense of the non-linear and complex relationships that characterise a firm's decision to issue seasoned equity. The most important variables are determined using SHAP values, a feature importance method. I find a firm's recent stock return to be the most important driver of its decision to conduct an SEO by a considerable margin. Other important variables include whether the firm has issued seasoned equity in the past, its retained earnings, and the dividends it pays per share. I find that 10 variables are responsible for most of the model's predictive ability, and numerous theoretically-motivated variables are found to be unimportant. Overall, the ML approach taken identifies variables that should be part of theory, offering unique insights into the drivers of SEOs.

Keywords— *seasoned equity offering, capital structure, machine learning, gradient boosted trees*

1 | Introduction

1.1 Background and motivation

Seasoned equity offerings (SEOs) are a significant corporate activity which allow firms to raise capital to finance their investments. They are a clear signal of future corporate events and are thus of interest to investors and competitors. SEOs also provide substantial fees to the investment banks that act as advisors and underwriters to the issuing firm. Prior academic research has shown a stock price run-up before an SEO, a negative response from investors to the public announcement of the transaction, and the tendency for issuers to underperform their non-issuing peers for up to five years after the issue (Loughran & Ritter, 1997). Given that SEOs are of interest to a variety of parties, signal future corporate activity, and are related to stock returns, being able to predict whether a firm will issue seasoned equity is a worthwhile pursuit.

As firms choose whether to finance their assets with debt or equity, corporate finance research has attempted to explain how firms decide on the debt/equity composition of their capital structure. There is a striking lack of academic consensus on which capital structure theory is correct, which has motivated continued research into the topic. As the decision of whether to issue equity is closely related to capital structure, the study of SEOs yields insights into capital structure theory.

An SEO is an equity issuance by a company that is already publicly traded, that is, has already had an initial public offering (IPO). SEOs are typically facilitated and underwritten by investment banks and can be either dilutive or non-dilutive.

Dilutive SEOs occur when new shares are issued, increasing the total number of shares on the secondary market. Non-dilutive SEOs occur when existing shareholders, such as venture capital firms, sell all or some of their shares, resulting in the total number of shares remaining fixed and thus not having dilutive effects.

Machine learning (ML) has become the dominant computing method for predictive tasks across many disciplines. However, it is only in recent years with advancements in computing power, the development of better algorithms, and the explosion of big data that ML has shown promise to transform finance research (Lopez de Prado, 2019). As my research aims to predict whether a firm will conduct an SEO, it makes sense to go beyond traditional statistical methods and employ ML's predictive power for the task. The novelty of this study is, thus, derived from the ML method I employ. Moreover, I further the academic conversation regarding the role of ML in academic research.

1.2 Research questions

Advancements in ML over the past decade have led to the creation of immensely successful prediction algorithms (Schmidt, Marques, Botti, & Marques, 2019). My research aims to test whether ML algorithms can predict SEOs, using lagged data that would have been publicly available at the time of prediction. If SEOs are found to be predictable, these findings can be built upon in future research or used by practitioners who are interested in knowing what firms are most likely to issue seasoned equity. Additionally, by analysing feature importance, I hope to highlight the important drivers of SEO prediction, which may then yield a basis

for economic theory construction.

Question 1: *Can ML predict whether a firm will engage in an SEO?*

Question 2: *What are the most important determinants of a firm's decision to conduct an SEO?*

1.3 Contribution

My research contributes to two primary strands of literature. First is the body of research concerned with using ML to predict corporate events. Recent studies have employed ML in a corporate finance setting, for example, to predict bankruptcy (Barboza, Kimura, & Altman, 2017; Lahmiri & Bekiros, 2019; Wang et al., 2017; Yu, Miche, Séverin, & Lendasse, 2014) or to value firms (Geertsema & Lu, 2019). However, this literature is insubstantial and still maturing. As such, this paper contributes to the nascent discussion regarding the role of ML techniques in corporate finance research. The importance of investigating ML's use in finance should not be overlooked. In recent years, ML and artificial intelligence (AI) have transformed countless industries and research areas. Finance researchers must investigate the potential of ML if they are to be at the forefront of their field. Furthermore, practitioners such as investment managers and investment bankers may be interested in the prediction of SEOs for a variety of reasons, such as generating trading strategies or identifying potential origination targets.

In his paper 'Beyond Econometrics: A Roadmap Towards Financial Machine Learning', Lopez de Prado (2019) suggests that ML offers a set of tools well suited to overcome the complex relationships in financial markets. Lopez de Prado

(2019) states that recent scientific breakthroughs show that ML's uses extend beyond simple prediction to questions that also require theoretical understanding. The paper provides a list of tasks where ML can be used to aid scientific research, one of which is particularly relevant and worth quoting.

> *'ML algorithms can determine the relative informational content of variables (features, in ML parlance) for explanatory and/or predictive purposes... [A]lthough [feature importance] does not uncover the underlying mechanism, it discovers the variables that should be part of the theory'* (Lopez de Prado, 2019, p. 6).

Discovering the variables that should be part of theory motivates the second strand of literature to which my research will contribute, that is, the body of corporate finance research concerned with the determinants of firms' equity issuances, which is directly related to capital structure theory. My research aims to generate economically interpretable insights by analysing what 'features' (ML parlance for independent variables) are the most important determinants of SEOs. Due to their ability to cope with many features, ML models can address many of the limitations of empirical capital structure research. For example, Bradley, Jarrell, and Kim (1984) assert that a significant limitation in testing trade-off capital structure models is the problem of missing variables. This is a common research problem that arises when excluded variables are correlated with included variables, causing misleading interpretations of the regression results. ML models can partially address this omitted variable bias by including many features, motivating the use of ML in corporate finance research. ML models do lack some of the interpretability and rigour of traditional regression models. However, my goal is not to test the

underlying mechanisms of how certain features drive a firm's decision to conduct an SEO. Instead, I hope to discover all of the important features, providing a basis for future, more traditional, research.

1.4 Results

My results show that an ML-based GBM model can predict SEOs with an AUC score of 0.74, thus performing substantially better than chance alone. The GBM model compares favourably to an ML-based logistic regression model, which has an AUC score of 0.63. As features can only enter linearly into a logistic regression, this difference in performance suggests that there are complex and non-linear relationships that characterise a firm's decision to issue seasoned equity.

The model is rerun using the 10 most important features. Its performance does not change significantly. Thus, the 10 features are the key determinants of SEOs for the GBM model. The model's feature importance is calculated using SHAP (SHapley Additive exPlanations) values, a method with roots in game theory and is detailed in Section 6.1. The most important features include a firm's past annual stock return (*trt1y*), whether it has issued equity in the past (*past_seo*), its dividends per share (*dvpspq*), its retained earnings (*req*), its Altman Z-Score (*alt_z*), and the market-wide number of SEOs (*num_seo*). As capital structure theory and a firm's decision to conduct an SEO are directly related, the notable lack of importance of theoretically motivated features suggests that firms' financing decisions are not as simple as conventional capital structure theories suggest. While the findings cannot provide any evidence about a firm's long-term capital structure,

they indicate that a multitude of factors play into firms' equity financing decisions. Furthermore, as the model has noteworthy predictive ability, the feature importance findings provide a set of variables that should be part of the theory related to SEOs.

Overall, the results provide a novel contribution to the literature by using modern ML methods in finance research. The contribution's importance stems from the ML model's predictive ability, which cannot be found using traditional statistical techniques. As the predictive power corresponds to the model's ability to uncover and understand the relationships that drive SEOs, an implication is that the resulting feature importance is grounded in reality and paints the best available picture into what determines SEOs. Furthermore, the model's predictive ability suggests that ML-based research could be used to predict other corporate events such as mergers and acquisitions (M&A) and debt issuances. Additionally, a trading strategy could be tested to see whether the model's predictive ability could be used in conjunction with related stock return findings, for example, that issuers tend to underperform their peers (Loughran & Ritter, 1997).

2 | Literature Review

2.1 Capital structure theory

In this first section, I summarise the literature on capital structure theory. It would be an oversight not to provide a thorough review of this literature, as it gives context to my research and places it in the broader body of academic work. Crucially, capital structure theories suggest which variables may have relevance for predicting SEOs. While my research does not focus on studying the determinants of a firm's long-term capital structure, indirect evidence can be gleaned by analysing which factors are the most important in a firm's decision to conduct an SEO. However, it is worth noting that, as shown by Hovakimian (2006), decisions to issue debt or equity do not necessarily lead to persistent changes in a firm's capital structure.

Firms fund their assets by raising capital from various sources, including debt, equity, and hybrid securities. Capital structure research attempts to explain how firms choose their capital structure and what their optimal capital structure is. There is a notable lack of academic consensus around which capital structure theory is correct, with Bradley et al. (1984) describing capital structure theory as "one of the most contentious issues in the theory of finance" (p. 1). Many of the leading theories are contradictory, and a large body of empirical work exacerbates this uncertainty by presenting mixed support for the theories. However, it is slightly bewildering that capital structure theory is thought of as a 'puzzle'. The idea that

there is a single governing theory that determines a firm's optimal and realised capital structure assumes that all managers and shareholders have the same set of preferences and are not in any way deficient in relevant academic knowledge. In reality, managers are subject to bounded rationality (Simon, 1990) and thus make capital structure and financing decisions which are unlikely always to follow a particular set of theoretical principles, potentially even guided by investment bankers seeking to generate fees.

Arguably, some of the best evidence on how managers determine their capital structure is from survey findings such as those of Graham and Harvey (2001) and Pinegar and Wilbricht (1989). Surveys like these allow for insights into what managers do in practice, in contrast with theories that dictate how they ought to act. With my research, I hope to take an empirical approach akin to a survey, where I let the data 'speak for itself' without imposing a particular model. By determining which variables are most important for predicting SEOs, indirect evidence can be garnered on capital structure theory.

2.1.1 Capital structure irrelevance

The origin of capital structure theory can be traced back to Modigliani and Miller (1958) who asserted that capital structure is irrelevant and that the mixture of debt and equity chosen by a firm does not affect its value. A first version of the theory assumed perfectly efficient markets, free of taxes, bankruptcy costs, and asymmetric information. Critics were quick to point out that the theory's assumptions were too demanding and unrealistic. Arguments about the failure of the theory to correctly take into account the tax benefits of debt prompted Modigliani

and Miller (1963) to revise their paper, which conceded the point and made the theory's estimates more realistic. Modigliani and Miller's (1958) paper spurred serious academic research into capital structure, thus playing an essential role in developing the field and influencing subsequent theories.

2.1.2 Pecking order

In 1961, Donaldson (1961) described firms' preferences for using internal funds over external funds and a preference for issuing debt instead of equity. Myers and Majluf (1984) explained these preferences with a theoretical pecking order model of capital structure. Myers and Majluf's (1984) pecking order model suggested that firms prioritise their sources of financing, first using internal financing, then debt, and only then issuing equity as a last resort.

The pecking order theory is based on the idea of asymmetric information between managers and investors. As information asymmetry increases, investors demand a higher rate of return to mitigate the risk arising from the asymmetry. To reduce the costs of financing, a company should thus adhere to the pecking order. Firms should prioritise internal financing, which has the least amount of information asymmetry. They should then issue debt, as doing so signals that the board is confident that an investment is profitable and that the current stock price is undervalued, thus reducing information asymmetry. Only then should a firm resort to equity financing, which has the highest cost associated with it.

Fama and French (2002) critique the pecking order model. Their study found, contrary to what the model would suggest, that a large proportion of companies issued equity, and that over half of their sample violated the pecking order. Fur-

thermore, Frank and Goyal (2008) described that, in research following Myers and Majluf (1984) (Eckbo & Masulis, 1992; Eckbo & Norli, 2004; Halov & Heider, 2011), adverse selection models are found to be very 'delicate', not necessarily imply a pecking order of capital structure.

2.1.3 Trade-off models

Modigliani and Miller's (1963) correction to their capital structure theory took into account the tax benefits of debt. By itself, this tax shield of debt suggested that a firm should be 100% debt-financed to maximise its value. However, this correction to the theory failed to factor in the potential costs of debt. Kraus and Litzenberger (1973) introduced the first trade-off model in which the optimum leverage ratio was a trade-off between the tax shield benefits of debt and the 'deadweight costs of bankruptcy', that is, the cost of debt. Following Kraus and Litzenberger (1973), many trade-off models were developed, which can be broadly categorised as either static or dynamic.

Static trade-off models predict that firms maintain a target leverage ratio that optimally balances the costs and benefits of debt, thus maximising a firm's value (Myers, 1984). Several early studies provide indirect evidence in support of static trade-off theory, such as Schwartz and Aronson (1967), Long and Malitz (1985), Smith and Watts (1992), and MacKie-Mason (1990). More direct evidence that firms adjust toward a target leverage ratio is provided by Taggart (1977), Marsh (1982), Auerbach (1985), Jalilvand and Harris (1984), and Opler and Titman (1994). Trade-off theories have been developed over the years to consider additional costs and benefits of debt, and Bradley et al. (1984) synthesise this body

of research. The authors develop and test a trade-off model, finding support for a trade-off theory of capital structure. Nevertheless, there is also a substantial amount of evidence that is inconsistent with static trade-off models. For example, Myers's (1984) empirical tests find that static trade-off models do a poor job at explaining firms' financing behaviour. Evidence for a significant negative relationship between past profitability and leverage ratios, inconsistent with static trade-off models, is also found by Kester (1986), Titman and Wessels (1988), Rajan and Zingales (1995). Trade-off theories would have predicted this relationship to be positive instead, as past profitability lowers the likelihood of bankruptcy, thus reducing the cost of debt.

The shortcomings of static trade-off theory gave rise to dynamic trade-off models. These dynamic models factor in firms' costs of moving towards their target leverage ratios, which involves the costly issuance or repurchase of securities. These costly adjustments imply that firms' capital structures may not always be aligned to their target leverage ratios, thus offering a potential reconciliation between trade-off theory and the empirically observed fact that many firms do not maintain optimal leverage ratios. Specifically, firms will only adjust their capital structure when the benefits of moving towards their target leverage ratio outweigh the costs. Examples of dynamic trade-off models include those of Fischer, Heinkel, and Zechner (1989) and Lucas and McDonald (1990).

Academic testing of trade-off models has been difficult as these models make similar predictions to other capital structure theories, such as pecking order (Fama & French, 2002). For example, Hovakimian, Opler, and Titman (2001) suggest that, consistent with firms trading off the risks of bankruptcy with the benefits of debt,

more profitable firms are more likely to issue debt instead of equity. However, this finding is also consistent with pecking order models. Additionally, several studies, such as those of Leary and Roberts (2005) and Strebulaev (2007), show that the market timing findings in Baker and Wurgler (2002) and Welch (2004), which will be discussed in the next section, are also consistent with dynamic trade-off theory. Overall, these mixed conclusions motivate additional research and may suggest that the various capital structure theories may not be mutually exclusive.

2.1.4 Market timing

A third major theory of capital structure emerged when Baker and Wurgler (2002) presented their market timing theory. This theory asserts that a firm's capital structure is the cumulative result of past attempts to time the equity market, with managers issuing equity when they perceive their firm's shares to be overvalued, and not subsequently adjusting the firm's capital structure back towards a target. Specifically, Baker and Wurgler (2002) find a negative relation between historical average market-to-book ratios and leverage, leading them to their market timing theory. The idea of timed equity issuances is consistent with Graham and Harvey (2001) survey findings, in which managers time the market by avoiding issuing equity if they perceive it to be undervalued. Further evidence for market timing is provided by Jenter (2005), "ELLIOTT2008175" (n.d.), and Huang and Ritter (2009). Additionally, Bancel and Mittoo (2004) and Henderson, Jegadeesh, and Weisbach (2006) suggest that debt issues are also timed to market conditions.

Baker and Wurgler's (2002) United States-based findings have been replicated in other markets, such as G-7 countries (Mahajan & Tartaroglu, 2008), The Nether-

lands (De Bie & De Haan, 2007), and China (Zhao, Lee, & Yu, 2020). Overall, the international studies agree with there being short-term market timing effects, but evidence for persistent impacts on long-term capital structure is divided. The most significant blow to the market timing theory comes from Hovakimian (2006), who does not question Baker and Wurgler (2002) empirical findings but disagrees with their interpretation. Specifically, Hovakimian (2006) finds that, although equity issuances may be timed, they do not have persistent effects on capital structure, and debt transactions have timing patterns that are unlikely to cause the negative relationship between market-to-book and leverage found by Baker and Wurgler (2002). Additionally, historical average market-to-book ratios are found to have an impact on a firm's current financing and investment decisions, inconsistent with market timing and suggesting that these ratios contain information about growth opportunities not captured by current market-to-book ratios, possibly due to being less noisy.

2.2 Determinants of SEOs

Capital structure theories give rise to a set of variables that may drive a firm's decision to conduct an SEO. Capital structure research analyses variables that affect a firm's leverage decisions, meaning that the variables are thus directly related to a firm's decision to issue equity. These theoretically motivated predictor variables are discussed in this section and summarised in Table 2.1.

Given the progressive nature of academic research, a theory or related variables may be shown not to be relevant or be disproved to some degree. When compiling

the set of features, I did not restrict myself to features that align with the current academic consensus on SEO determinants or capital structure theory. Instead, I included as many theoretically motivated features as possible and then let the ML model determine which features are the most useful predictors of SEOs.

2.2.1 Pecking order determinants

Based on the pecking order, features related to a firm's retained earnings and debt capacity are likely to be relevant predictors, as firms issue equity once they have exhausted retained earnings and debt financing. Thus, I include retained earnings and debt to equity ratio as features in my dataset.

Additionally, Frank and Goyal (2003) test pecking order theory, using a firm's financing deficit as the variable of interest. While the authors' regressions do not try to predict SEOs, I include a financing deficit variable, as it may prove interesting. Additionally, Frank and Goyal (2003) assert that, according to Harris and Raviv (1991), under pecking order theory, firms with fewer tangible assets may have greater information asymmetry problems, and thus tend to accumulate more debt over time. In other words, firms with high asset tangibility may have lower information asymmetry problems, face a lower cost of equity, and be more inclined to conduct an SEO. Alternatively, having more collateral supports debt, and thus the relationship may go the other way. Regardless, asset tangibility may have predictive power.

In their research on equity issuances and pecking order theory, Fama and French (2002) provide key variables that are likely to be related to firms' equity offering decisions. These variables include book leverage, market-to-book ratio, profitabil-

ity (which can be measured using ROE and ROA), and volatility (proxied for by firm size using a log assets variable).

2.2.2 Trade-off determinants

Trade-off theory suggests that a firm's leverage will be a function of the costs and benefits of debt. Therefore, under trade-off models, one expects firms with higher costs of debt, or lower benefits of debt, to be more likely to finance their investments with equity.

In their study on capital structure, Frank and Goyal (2009) provide a comprehensive set of variables that proxy for the costs and benefits of debt. Profitable firms face lower costs of financial distress and are benefited more by interest tax shields. Thus, trade-off models would predict that more profitable firms should be more inclined to use debt financing. Larger firms, as measured by the log of total assets, face a lower risk of default and are more likely to use debt. Firms with greater growth face higher financial distress costs and increased debt-related agency problems; thus, they are predicted to use more equity financing. Growth can be measured by market-to-book ratio, change in log assets, and CAPEX/assets ratio. Firms with greater tangible assets are easier to value by investors, which reduces distress costs, so trade-off theory would predict that, as asset tangibility increases, leverage does too. Similarly, firms with greater research and development (R&D) and selling, general administrative (SG&A) expenses are likely to have more intangible assets and may be more prone to financial distress, so are predicted to have lower leverage. High tax rates increase the tax shield of debt; thus, trade-off theory predicts that firms would issue more debt when tax rates

are higher. DeAngelo and Masulis (1980) show that depreciation tax shields are a substitute for the tax shield benefits of debt, so firms with high depreciation may be less inclined to use debt financing.

2.2.3 Market timing determinants

Whether or not market timing has persistent effects on capital structure, as suggested by Baker and Wurgler (2002), research has shown that managers time their equity issuances to market conditions. Graham and Harvey (2001) find that equity undervaluation/overvaluation is the second most important consideration for managers when choosing whether to issue stock. This finding implies that a higher market-to-book ratio, which is a proxy for overvaluation, will increase the probability of an SEO. Similarly, Graham and Harvey (2001) find that managers issue stock during a 'window of opportunity' occurring after a recent share price increase. This survey evidence is in line with Loughran and Ritter (1995) finding that most firms conducting SEOs had significant share price increases in the prior year. That finding also supports Lucas and McDonald's (1990) model, which predicts that equity issues are preceded by an abnormal positive return on the stock. These studies suggest that a firm's stock return may be an important predictor of SEOs.

2.2.4 SEO and IPO cycles

IPO volume has been found to fluctuate over time, exhibiting a cyclical pattern (Lowry & Schwert, 2002). Lowry and Schwert (2002) find that IPO cycles are driven by companies learning positive information from other firm's IPOs, that is,

other IPOs suggesting that they can raise more money in an IPO than they had initially thought.

Howe and Zhang (2010) find that SEOs exhibit cycles similar to those of IPOs; however, the cycles are found to be less volatile. Since equity issuances appear to occur in waves, the numbers of IPOs or SEOs that have occurred in prior months could be useful conditioning variables and are thus included in my dataset. Additionally, firms that have conducted an SEO in the past may be more likely to conduct another one. Thus, a dummy variable capturing whether a firm has previously issued seasoned equity is included in my dataset.

2.2.5 Investor sentiment

Baker and Wurgler (2007) state that investor sentiment can affect the cost of capital, which may potentially cause equity issuances to fluctuate with investor sentiment. Baker and Wurgler (2007) construct a sentiment index based on the proxies for investor sentiment used by Baker and Wurgler (2006). Baker and Wurgler's (2007) proxies comprise equity share in new issues, number and first day-returns of IPOs, closed-end fund discount, and NYSE turnover. Howe and Zhang (2010) investigate whether investor sentiment impacts SEO volume but do not find statistically significant results. Investor sentiment will not vary in the cross-section by stock, but it may still be a useful conditioning variable; thus I included it in my dataset.

2.2.6 Demand for capital

Intuitively, when firms have better investment opportunities, they are more likely to issue equity to finance these opportunities. Thus, when overall economic conditions are good, and firms as a whole have good growth opportunities, firms are more likely to undertake SEOs. While firms could use debt financing, there is support for firms preferring equity in periods of good market conditions (Baker & Wurgler, 2000). There is also substantial research that links IPOs to variables that proxy for investor demand for capital. For example, Lowry (2003) shows that IPO volume is positively related to demand-for-capital proxies. She cites Mikkelson, Partch, Shah, et al. (1997), whose intuitive evidence shows that the majority of firms undertake an IPO to raise working capital and money for new investments. Furthermore, Harjoto and Garen (2003) show that firms with higher unanticipated growth are more likely to conduct an SEO following their IPOs. Therefore, I expect SEO probability to have a positive relation with demand-for-capital proxies such as a firm's market-to-book ratio.

2.2.7 Information asymmetry

Related to pecking order theory, but not inextricably linked, is the idea that information asymmetry will affect a firm's decision to conduct an SEO. When information asymmetry increases, so do the adverse selection costs of issuing equity. Myers and Majluf (1984) suggest that firms can time their equity issuances to periods when information asymmetry is low, thus reducing the cost. Bayless and Chaplinsky's (1996) results provide empirical support to this idea, finding that firms

issue equity at least partially due to reduced levels of asymmetric information. By definition, information asymmetry is impossible to measure. Still, research has suggested that market microstructure variables such as bid-ask spread (Copeland & Galai, 1983) and trading volume (Chae, 2005) may be useful proxies.

Table 2.1: Theoretically-motivated predictor variables

This table provides a summary list of SEO predictor variables suggested by the literature. Many of the variables are discussed in various sources and based on different theoretical motivations. Given that there is no extensive literature on predicting SEOs, many of the variables are only indirectly suggested to be relevant predictors. Furthermore, it is worth noting that no particular care was taken to select only literature (and thus predictors) that are consistent with the prevailing academic consensus. The purpose was to develop as comprehensive a list as possible by including anything that may have predictive power.

Variable	Pecking order	Trade-off	Market timing	Other
Retained earnings	✓			
Debt-to-equity ratio	✓	✓		
Financing deficit	✓			
Asset tangibility	✓	✓		
Debt-to-asset ratio	✓	✓		
ROE	✓	✓		
ROA	✓	✓		
Log Assets	✓	✓		
Market-to-book		✓	✓	✓
CAPEX / assets		✓		
R&D expense		✓		
SG&A expense		✓		
Tax rate		✓		
Depreciation tax shield		✓		
Past year's stock return			✓	
Past month's stock return			✓	
Market-wide number of SEOs				✓
Market-wide number of IPOs				✓
Past issuer				✓
Investor sentiment				✓
Bid-ask spread	✓			✓
Trading volume	✓			✓

2.3 Machine learning in finance

ML-based finance research is still in its infancy, but the field has been growing steadily in recent years. ML-based research is starting to make it into the top finance journals, such as the studies of Gu, Kelly, and Xiu (2020a) and Bianchi, Büchner, and Tamoni (2020), which were published in the Review of Financial Studies.

Two facts emerge consistently across the literature. First, neural networks and tree-based methods tend to be the best performing ML algorithms with financial data. Second, ML models outperform traditional econometric models consistently across prediction tasks. These facts provide a basis for my research and are unsurprising as they hold across other scientific disciplines.

Given my research's focus on firm-level financial data, studies such as Geertsema and Lu (2019) and Amel-Zadeh, Calliess, Kaiser, and Roberts (2020) are of particular interest. Geertsema and Lu (2019) use gradient boosted trees and financial statement data to produce firm enterprise valuations. The authors' model compares favourably to sell-side analysts' forecasts and final-year finance students' valuations. Amel-Zadeh et al. (2020) use financial statement data to investigate whether a range of ML models can be used to forecast the sign and magnitude of abnormal stock returns around earnings announcements. The authors note that

> *"machine-learned relationships between the set of financial statement variables and stock returns follow fundamental valuation intuition for predicting free cash flows and are consistent with established return re-*

> *lationships of accounting-based regularities found in prior accounting research"* (p. 34).

This example is noteworthy as it demonstrates how theoretical insights can be derived from ML models.

Additionally, there is a substantial and well-developed literature on ML in equity pricing. To give some examples, Messmer (2017), Gu et al. (2020a), and Gu, Kelly, and Xiu (2020b) use ML models for cross-sectional return prediction. Chen, Pelger, and Zhu (2019) use deep neural networks to estimate the stochastic discount factor, and Feng, Polson, and Xu (2020) use deep neural networks to generate risk factors. Rossi (2018) and Gu et al. (2020a) use boosted regression trees in their equity pricing research. Moritz and Zimmermann (2016) use tree-based conditional portfolio sorts to study the cross-section of returns, focusing on showing that ML is not necessarily a 'black box' and that interpretable information can be extracted from tree-based conditional sorts. Bryzgalova, Pelger, and Zhu (2019) use tree-based models to build interpretable basis assets that capture the information contained in stock characteristics.

2.4 Predicting SEOs

There is no substantial body of research that focuses on predicting SEOs, and few studies have had similar goals to the present study. DeAngelo, DeAngelo, and Stulz's (2010) goal is somewhat similar to mine, and the authors use a logit regression to assess the determinants of a firm's decision to conduct an SEO. In their study, a firm's decision to conduct an SEO in a given year is regressed against

market-to-book ratio, prior stock return, and future stock return. Similarly, Virolainen et al.'s (2009) thesis investigates the micro and macro determinants of SEOs using a wide range of firm-level and macroeconomic variables. Aside from using ML techniques, my research is distinct from these studies by the fact that it focuses primarily on SEO prediction and uses historical (lagged) data that would be publicly available in the prediction period.

3 | Data

3.1 Features

To create the features that were used to predict SEOs, variables from several databases were combined.

1. The Compustat Fundamentals Quarterly database was used to gather firm-level accounting data (see Table A.1).
2. The CRSP monthly file was used for stock return data (see Table A.3).
3. Firm-level financial ratios were obtained from the WRDS ratio suite (see Table A.1).
4. Investor sentiment data were sourced from Jeffery Wurgler's database (see Table A.3).
5. Data on the number of IPOs and SEOs each that occur each month were taken from Jay Ritter's database (see Table A.3).

As in Geertsema and Lu (2019), all accounting and ratio data were assumed to be available for use four months after the quarterly balance sheet date. Practically, this involved lagging the accounting data by four months, and lagging the ratio data by two months, as upon collection the ratio data is already lagged by two months. The lagging of the accounting and ratio data prevented the use of data before it would have been available to the public, as accounting data is usually released around two months after the balance sheet date. Likewise, features rep-

resenting investor sentiment and the number of IPOs in a given month were lagged by one month. I chose a one month lag was because the monthly value for both of these features can only be calculated once a given month has ended. Furthermore, I scale many of the accounting variables by total assets, *atq*, as shown in Table A.1 in the appendices.

The Compustat accounting data ranges from January 1980 to November 2020, as this was the period for which SEO data were available on SDC Platinum. Since the accounting data were quarterly, they were converted into a monthly frequency before being merged with the other datasets. The conversion process used filled the quarterly data forward across the subsequent months, before the next quarterly data were released. Additionally, several requirements were imposed on the dataset. Each observation is required to have a valid 6-, 7-, or 8-digit CUSIP and be listed on the Amex, Nasdaq, or NYSE. Additionally, to restrict the dataset to firms of meaningful size, a requirement for a cross-sectional percentile rank of total assets above the 20th percentile of firms for a given month in the dataset is imposed. This approach is based on Geertsema and Lu (2019), albeit using a 20^{th} percentile cut-off instead of a 10^{th} percentile cut-off. This filtering process ended up yielding a dataset of firm-level accounting data comprising 1,904,549 monthly observations across 14,298 firms. I then merged the Compustat accounting data with the CRSP return data, retaining only observations with non-missing monthly returns. Finally, the datasets containing the Compustat ratios, investor sentiment and the number of IPOs and SEOs were merged with the accounting and return data. When merging the ratio data with the other data, no requirement for valid ratio data was imposed, so as not to overly restrict the dataset.

When selecting what accounting data and ratios to include, shown in Tables A.1 and A.2 respectively (see Appendix A), a data-driven approach was taken, that is, including as many potentially useful features as possible. Many of the features included may not offer a great deal of predictive power; however, I choose to let the data 'speak for itself', as the ML algorithm used can handle a large number of features. This method avoids the possibility of missing important features. Furthermore, I ensured that all of the theoretically-motivated predictors, as discussed in the literature review, were included, as these variables are likely to offer the greatest amount of predictive power. Additionally, using the accounting variables, I constructed additional features which were motivated by the literature review. These additional features are shown in Table A.3. Furthermore, several of the accounting variables are year-to-date variables. Year-to-date variables can be misleading and are dependent on the fiscal quarter of the variable. A better approach for getting a yearly value using the quarterly data is to take a four-quarter rolling sum of the quarterly values. I have used this rolling sum approach with the *dpq*, *niq*, *oibdpq*, *revtq*, *xrdq*, and *xsgaq* variables, and list the rolling sum variables in Table A.3 in Appendix A. Unfortunately, some accounting variables are only available in the year-to-date form. While this form is suboptimal, I have included the variables regardless, as they may offer predictive power.

Overall, 120 features are included in the final dataset. Tables A.1 and A.2 contain the accounting and ratio variables from WRDS and Table A.3 comprises the features that I have created, along with the corresponding calculations. Additionally, I describe the 10 most important features [1] in Table 3.1 below.

[1] Feature importance is based on SHAP values. The importance methodology and outputs are found in Section 6

Table 3.1: 10 most important features

This table lists the 10 most important features, as determined by SHAP values. In other words, these 10 features contribute most to the model's ability to predict SEOs. I further discuss these features and the importance ranking methodology in Chapter 6.

Variable	Description
trt1y	Yearly total return
past_seo	Whether the firm has previously issued seasoned equity (dummy)
req	Retained Earnings
dvpspq	Dividends per Share - Pay Date - Quarter
alt_z	Altman Z-Score (Altman, 1968)
num_seo	Market-wide number of SEOs
tstkq	Treasury Stock - Total (All Capital)
trt1m	Monthly total return
sstky	Sale of Common and Preferred Stock
apq	Account Payable/Creditors - Trade

3.2 Target variable

The target variable, *seo*, is a dummy variable representing whether or not a firm has carried out an SEO in a given month. SDC Platinum was used to gather the required SEO data. I searched for SEOs by United States (US) firms between January 1980 and November 2020, which spans the entire available sample period. I required that the SEO value was greater or equal to US$5m to retain only economically significant transactions. Furthermore, I required that the transactions were specifically SEOs, excluding all IPOs and convertible issues. This process yielded a dataset comprising 24,503 SEOs. After merging the SEO dataset with the features and restricting the sample to common domestic US equities (share codes 10 and 11), the number of SEOs was reduced to 8,505. SEO months comprise 0.46%

of the total dataset, with 99.54% of months being non-SEO months.

4 | Methodology

4.1 Machine learning

In general, ML is a method for learning patterns from data without the researcher imposing a structure. Arguably, the best known and insightful article on the applications of ML to finance is Gu et al. (2020a). The authors provide a comparative analysis of several ML techniques for the task of asset pricing. Notably, the authors give a definition of ML tailored to finance, as follows:

1. A diverse collection of high-dimensional models for statistical prediction, combined with
2. So-called 'regularisation' methods for model selection and mitigation of overfitting, and
3. Efficient algorithms for searching among a vast number of potential model specifications. (Gu et al., 2020a, p. 3).

Two main qualities make ML well suited to predicting SEOs. First, ML algorithms can detect patterns in high-dimensional space, potentially even using more explanatory variables than observations (Lopez de Prado, 2019). Traditional econometric models fail when the predictor count approaches the observation count, or when predictors are highly correlated. Since many highly correlated firm characteristics may affect the probability that a firm engages in an SEO, the ability for ML algorithms to handle a large number of explanatory variables makes it extremely useful for SEO prediction.

Second, ML allows for complex patterns from data to be understood with little human guidance or model specification (Lopez de Prado, 2019). Therefore, ML analysis can reveal relationships in multidimensional data that were not known a priori (Dhar, 2013), which can enhance model performance and be used to provide insights into the relationships that are most important for predicting a phenomenon. As discussed in the literature review, there is debate surrounding the reasons that firms engage in SEOs.

By allowing the data to speak for itself, I can model the relationships that predict SEOs. This model may subsequently be used in conjunction with traditional statistics and the prevailing corporate finance theories to develop theoretical insights.

Given that ML models do not need to be specified by the researcher, many conclude that ML must, therefore, be a 'black box', from which no understanding can be extracted. Lopez de Prado addresses this concern in his paper 'Beyond Econometrics: A Roadmap Towards Financial Machine Learning' (Lopez de Prado, 2019). This paper points out that ML models can be interpreted through a number of procedures.

4.2 Philosophical approach

The process of fitting an ML model follows inductive reasoning, where general principles are derived from specific examples. An ML model generalises the training data (specific examples), that is, the data that is used to train the ML model, and the model is assumed to hold over new unseen data when it is tested or used

(ML models are tested using a separate set of data from that on which they are trained). Galindo and Tamayo (2000) cite Asmis (1984), stating that, in line with an ML approach, one of the oldest principles of induction advocated by Epicurus is that all models or theories should be consistent with data.

Using an inductive approach has its merits in finance research as financial data are full of complex and non-linear patterns (Lopez de Prado, 2019) that traditional econometrics can struggle to make sense of. However, philosophers such as Hume (1739) and Popper (2005) have questioned the validity of inductive reasoning, with Hume's 'problem of induction' being one of the most famous arguments in philosophy. In very general terms, the problem of induction is based on the idea that induction assumes the future will behave like the past, but we may have no good reason for thinking that this is true. Arguing that one believes the future will behave like the past, because it has done so in the past, uses induction to justify induction, resulting in circular reasoning.

While these concerns with following an inductive approach and using ML in scientific research are valid, I intend to use my predictive ML model to answer a practical research question, and, in conjunction with established theory, to lay a foundation for more traditional research.

4.3 Sample splitting

Related to the philosophical approach detailed above, Geertsema and Lu (2019) also describe how a dataset can be used for either hypothesis search or testing. Traditional empirical research tests hypotheses based on established theory, thus

utilising the entire dataset for testing. Geertsema and Lu (2019) describe how an ML approach of splitting a sample into training and testing sets allows the researcher to carry out a hypothesis search on the training set and test the hypothesis on the testing set.

In my research, I divide the sample into training, validation and testing subsets. The training set is the dataset used to train the ML models; these are the data that the model learns from. The validation set is then used to tune the model's hyperparameters, so the model will 'see' these data but not learn from them. Therefore, the validation data can indirectly affect the model, hence being kept separate from the testing set. Finally, the test set is used to evaluate the model's predictive performance, after the model is fully trained and tweaked using the training and validation sets. The performance and insights from the test set are what I will use for drawing conclusions.

Specifically, I divide the dataset into 60% for training, 20% for validation, and 20% for testing. I split the dataset on a chronological basis. The rationale behind this approach is that a practitioner does not have access to data points in the future when they develop models. Thus, the model's predictive ability reflects what would have been available to a practitioner who cannot see into the future.

4.4 Gradient Boosting Machine

4.4.1 Decision trees

To predict SEOs, I use an ML-based algorithm called Gradient Boosting Machine (GBM). I describe the algorithm in detail in Section 4.4.3, but some background is necessary.

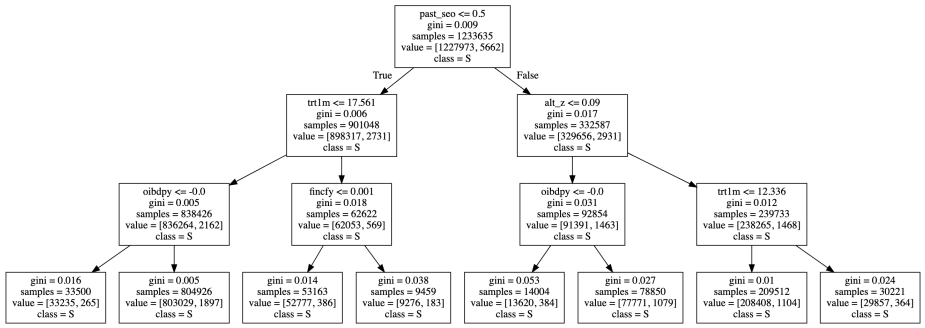
GBM is an ML ensemble method (see Section 4.4.2) based on decision trees. Decision trees are an ML algorithm which can solve both classification and regression problems by recursively splitting a dataset based on various conditions, reaching an output for a target variable. This splitting process is layered, and at each layer the sample is split into two or more groups comprising observations that are most similar to each other (homogeneity) and most different from the other groups (heterogeneity). Decision trees achieve this separation by using metrics such as the Gini Index, Chi-square, and Information Gain. The process is defined as recursive because each sub-population may be split a number of times until a particular stopping criterion is reached, such as a maximum number of splits. The groups that are created from the splitting process are referred to as the 'nodes' or 'leaves' of the tree. The splits occur based on a particular feature.

Figure 4.1 provides an example of a simple decision tree model trained on my dataset. The figure shows how decision trees work and also highlights the algorithm's interpretability. The first split occurs on the *past_seo* variable, i.e. whether a company has issued equity in the past. Depending on the result of the *past_seo* variable, the next splits occur on the *trt1m* (past month's stock return) and *alt_z*

(Altman Z-Score) variables. A possible economic interpretation of the model's splitting decisions may be that, if a company has a history of raising equity and experiences a large stock return in the month prior, managers may take advantage of this window of opportunity and issue additional equity (Graham & Harvey, 2001). Conversely, suppose a company does not have a history of issuing equity. In that case, its Altman Z-Score, that is, its likelihood of bankruptcy, may be the most significant driver of SEOs so that the firm can meet its financing needs and avoid bankruptcy, in line with the trade-off and pecking order models.

Figure 4.1: Decision tree example

This figure plots the splits of a simple decision tree trained on the training dataset. The key bits of information in each node (the boxes) are the name of the feature and the gini value used for the splitting decision.



Decision trees are very popular as they can be trained quickly and are easy to interpret. One of the primary disadvantages of decision trees is that they are prone to overfitting, that is, fitting the training data too well and generalising poorly to new data. Overfitting can be resolved through the use of ensemble learning, which will be detailed next.

4.4.2 Ensemble methods

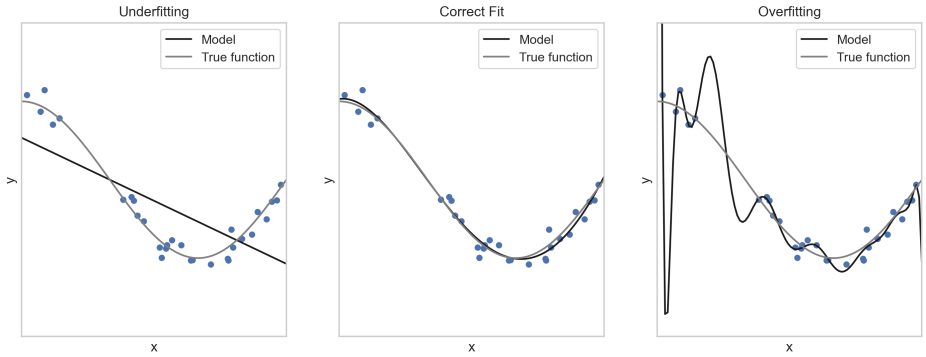
In ML, ensemble methods refer to the combination of multiple base models (often referred to as 'weak learners') into a single model, to obtain better performance than could be achieved from any constituent model. For example, in a decision tree ensemble, rather than just relying on a single tree model and hoping it makes the correct splitting decisions, multiple models' results are aggregated into a final prediction.

To best understand ensemble learning, it is important to explain the types of ML errors and the concepts of overfitting and underfitting. Brief definitions of these concepts are given below.

1. **Overfitting:** Occurs when a model is almost perfectly accurate when handling training data but generalises poorly to unseen data
2. **Underfitting:** Occurs when a model is too simple and does not fit the training data adequately, thus cannot make accurate predictions.

Figure 4.2: Under and overfitting

This figure provides a simple graphical illustration of what underfitting and overfitting look like in 2D space. The plot on the left shows underfitting, the plot in the middle shows a good fit, and the plot on the right shows overfitting.



The total prediction error of an ML model is a combination of three main types of error - bias error, variance error, and noise.

Mathematically, the relationship between the predicted values and the inputs values can be written as $y = f(x) + e$, where e is total error term. Let $\hat{y} = \hat{f}(x)$ be the ML model prediction for some input data x. Then the expected squared error can be written as:

$$E\left[e^2\right] = E\left[(\hat{y} - y)^2\right] \tag{4.1}$$

The expected squared error can be decomposed into bias, variance, and noise (irreducible) error components, as shown below.

$$\begin{aligned} E[e] &= E\left[(y - \hat{y})^2\right] \\ &= (E[\hat{y}] - y)^2 + VAR[\hat{y}] + VAR[\varepsilon] \\ &= Bias + Variance + Noise \end{aligned} \tag{4.2}$$

These errors are described in general terms below.

1. **Bias error:** The difference between the model's prediction and the actual result. A model with high bias is underfitting, and increasing model complexity would reduce bias.

2. **Variance error:** The amount the model's estimate would change if different training data were used, that is, how sensitive the model is to small fluctuations in the training data. High variance means the model generalises poorly to unseen data and is overfitting; thus, using a less complex model

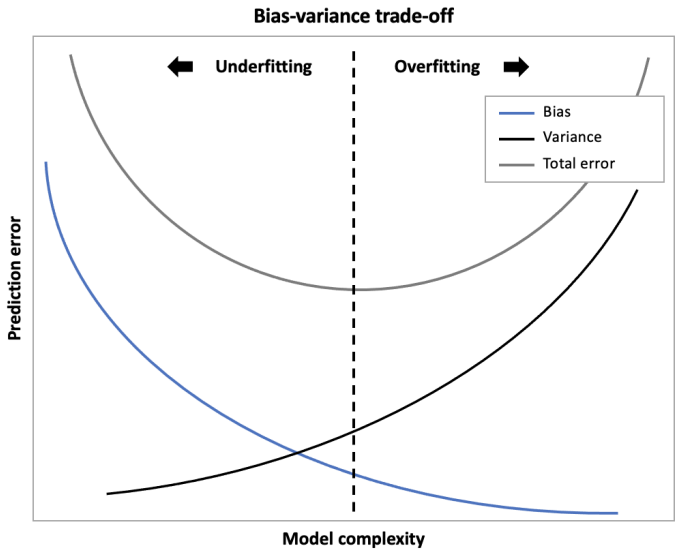
would reduce variance.

3. **Noise error:** The irreducible error in a prediction.

There is a trade-off between bias and variance in ML models, because reducing model complexity reduces variance but increases bias, and vice versa. Ensemble methods allow for bias and variance to be managed in a balanced way by using multiple models and combining them without underfitting or overfitting the data.

Figure 4.3: Bias-variance trade-off

This figure plots the bias-variance trade-off, which is an essential consideration in ML. On the x-axis is model complexity, and on the y-axis is prediction error. The legend in the top right shows which type of error each line corresponds to.



There are two main approaches to ensemble learning - bagging and boosting. Bagging, which stands for bootstrap aggregating, gives each model in the ensemble an equal weight when the predictions are aggregated. Models are trained using a random subset of the training set to promote model variance. This process involves

creating many 'copies' of the training data and then applying a single weak learner to each copy to obtain multiple predictions. These predictions are then aggregated into a single final prediction in order to obtain a model with a lower variance.

Boosting is similar to bagging, but the weak learners are trained iteratively, with each new learner helping to correct the errors made by the previously trained learner. The predictions of these iteratively trained models are then combined into a single prediction. This process allows high bias weak learners to be combined in an ensemble that has lower bias than the individual models.

4.4.3 GBM

GBM was developed initially by Friedman (2002), based on observations from Breiman (1997). GBM applies a decision tree-based ensemble learning algorithm that uses the boosting method described previously. GBM is one of the most popular ML algorithms and has been used with success for a great variety of tasks Ke et al. (2017). Like other boosting methods, GBM sequentially trains each decision tree, with each subsequent model taking into account the shortcomings of the previously trained models. This consecutive fitting of models provides a more accurate estimate of the target variable. GBM constructs new base-models that minimise the overall prediction error when combined with the previously constructed models. Each subsequently constructed model's target outcome is to minimise prediction error by optimising the current model's loss function. In generalised terms, this loss function evaluates how well an algorithm models the data. If predictions deviate significantly from actual results, the loss function would have a high value, and vice versa. GBM uses a different loss function depending on the

task; for example, a regression problem may use the model's residuals, and a classification problem may use the model's multinomial deviance (Hastie, Tibshirani, & Friedman, 2009). Since predicting SEOs is a binary classification task, my GBM model uses a binary classification loss function.

A mathematical description of a general binary classification GBM model is explained below, with notation taken from Burkov (2019).

Assume there are M regression decision trees, i.e. the constituent trees that make up the GBM ensemble. The ensemble (combination) of decision trees uses a sigmoid function, as shown in Equation (4.4), similar to a logistic regression model. This output is the probability that an observation x belongs to class 1, which in the present case is an SEO (remember, class 0 = non-SEO and class 1 = SEO).

$$\Pr(y = 1 \mid \mathbf{x}, f) \stackrel{\text{def}}{=} \frac{1}{1 + e^{-f(\mathbf{x})}} \tag{4.3}$$

In Equation (4.4), $f(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x})$, and f_m is a regression tree. Simply, $f(\mathbf{x})$ is the sum of the constituent decision trees, f_m.

Similar to a logistic regression, the goal is to find a model, f, that maximises $L_f = \sum_{i=1}^{N} \ln\left(\Pr\left(y_i = 1 \mid \mathbf{x}_i, f\right)\right)$.

The GBM algorithm starts by assuming an initial constant model $f = f_0 = \frac{p}{1-p}$, where $p = \frac{1}{N}\sum_{i=1}^{N} y_i$. As mentioned, this model is a constituent decision tree model, which is referred to as a weak leaner. At each boosting iteration m, a new tree f_m, is added. Remember, the goal of these new trees is to explain the shortcomings of the past trees. To do so, the first partial derivative g_i, of the

current model is calculated for each $i = 1, \ldots, N$:

$$g_i = \frac{dL_f}{df} \tag{4.4}$$

When calculating the partial derivatives of the current model, f is the decision tree model built at the past iteration m - 1. To calculate g_i one needs to find the derivatives of $\ln(\Pr(y_i = 1 \mid \mathbf{x}_i, f))$ with respect to f for all i. Note, i corresponds to each firm-month observation in the training set, with y_i being the class label (SEO vs non-SEO) for each of these observations.

The training set is then modified by replacing the original class label y_i with the corresponding partial derivative g_i. A new tree f_m is then built using the modified training set. Again, the point of this process is for the sequentially trained trees to correct the shortcomings, measured by the gradient, of the previous tree. Once the new tree f_m is built, an optimal update step ρ_m is determined as:

$$\rho_m = \arg\max_{\rho} L_{f+\rho f_m} \tag{4.5}$$

At the end of iteration m, the ensemble (combined) model f is updated by adding the new tree f_m:

$$f \leftarrow f + \alpha \rho_m f_m \tag{4.6}$$

We iterate until $m = M$, resulting in the ensemble model f, which can then be used to make predictions.

I have chosen to use GBM due to its strong performance and flexibility found in similar research that uses financial accounting data (Amel-Zadeh et al., 2020; Geertsema & Lu, 2019). Geertsema and Lu (2019) detail several features of GBM that make it well-suited to financial research. I will briefly discuss the most important of these features and how they complement my research.

First, GBM is robust to missing data. This attribute is important, as my dataset contains a large amount of missing data. Second, GBM is robust to outliers, which is crucial given that outliers can be extremely common in financial accounting data. For example, a firm may have a negative net income one year due to large and unexpected write-downs. Third, GBM is robust to irrelevant variables. Since I have included as many features as possible to ensure that I do not miss any important predictors, it is essential that I use an algorithm that can handle a large number of features. Fourth, GBM is robust to multi-collinearity, which is necessary as many accounting variables may be highly correlated. Fifth, GBM is able to accommodate interaction effects, which is vital given the complex non-linear relationships that characterise corporate events. Finally, GBM is fast, with Geertsema and Lu (2021) stating that implementations such as LightGBM can train models in seconds, around 100 times faster than neural networks.

Neural networks are another type of popular ML algorithm that have shown strong predictive performance in past research and real-world applications, but fall short of GBM in a few of the features outlined above. For example, neural networks cannot handle missing data, and as described by Geertsema and Lu (2021) are much slower. Given that these shortcomings are significant with my dataset, GBM is the clear choice between the two algorithms for my research.

4.5 Performance evaluation

In my dataset, there is a significant class imbalance problem. Specifically, there are many more monthly observations where a firm does not conduct an SEO than months where a firm does. This imbalance can result in misleading performance if unsuitable measures are used. For example, one could have an ML algorithm that classifies every observation as not an SEO, and doing so would result in having a high accuracy score, as there are so few months where a firm issues seasoned equity. However, as this research aims to predict the months when a firm does carry out an SEO, the class imbalance needs to be addressed.

There are three primary means of dealing with imbalanced data: rebalancing the classes by removing observations or creating new synthetic observations; using performance measures that are robust to imbalances; or using weight scaling hyperparameters. For my research, I have chosen to use weight scaling in conjunction with carefully selecting and analysing performance measures.

Weight scaling works by weighting rare labels using a weight multiplication. In the GBM model employed in this study, a parameter called 'is_unbalance' is used. Section 4.6 details how this hyperparameter works. Additionally, in the Results chapter (p. 62), I analyse the model's performance using random undersampling, finding that this method of dealing with class imbalances is inferior to the hyperparameter adjustments used in the main model.

Brownlee (2020) describes several performance measures that can tell a more truthful story than traditional accuracy when dealing with imbalanced data. One

of the more robust performance measures is AUC-ROC, which stands for 'Area Under The Curve - Receiver Operating Characteristics curve'. Using AUC-ROC is well suited for my research, given that I am studying a binary classification problem.

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier. The curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR).

$$TPR = \frac{TP}{TP + FN} \tag{4.7}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.8}$$

The AUC score indicates how good the model is at distinguishing between classes. In my research, the higher the AUC, the better a model is at distinguishing between firms that will engage in SEOs and firms that will not. To interpret the AUC value, one notes that chance alone would give a score of 0.5, and a perfect classifier would score 1. A good AUC score is context-dependent. For example, Hosmer, Lemeshow, and Sturdivant (2013) suggest that in the health sciences an AUC of greater than 0.9 is considered outstanding. However in a field such as finance lower AUC scores are acceptable. Since a company's decision to issue equity is likely a response to unpredictable outside events, such as crises or an excellent investment opportunity arising, I anticipate that my model's AUC scores will be comparatively low.

However, a low AUC score is not as much of an issue in my research as in other studies or fields. To illustrate this claim, imagine a firm wants to use ML to predict whether a credit card transaction is fraudulent, or a medical professional wants to use ML to determine whether a tumour is benign or cancerous. If the algorithm predicts incorrectly, the adverse effects can be significant. Conversely, in my research, nothing is lost or harmed by the algorithm predicting incorrectly, and any predictability above chance may be useful. For example, if a hedge fund could profitably trade on predicted SEOs, having a small edge using an ML algorithm could be beneficial. Additionally, one can still derive economic insights from feature importance, even with non-perfect predictably.

4.6 Hyperparameter tuning

An essential part of creating an ML model is tuning its hyperparameters. Hyperparameters are parameters whose values are used to control the learning process. They differ from model parameters in that they cannot be directly trained from the data. Conversely, model parameters are learned by the algorithm when training, for example, by optimising a loss function using gradient descent. While model parameters specify how the input data is transformed into the desired output, hyperparameters define how the model is structured.

The best way to select the optimal hyperparameters is through experimentation, for example, by analysing which hyperparameter settings result in the highest performance. This experimentation can either be manual or automated through an algorithm such as GridSearchCV. The hyperparameters must be tuned using a

validation set, in which the model's performance is assessed. Since this validation set is part of the model fitting process, it cannot be used to evaluate how well the model generalises to unseen data and a separate test set must be used instead.

I use the GridSearchCV algorithm to determine the optimal values for several hyperparameters. The algorithm loops through predefined hyperparameters and fits the model on the training set. GridSearchCV tries all combinations of the specified values and evaluates the model for each combination, using cross-validation. The best performing combination of hyperparameters can then be selected when creating the model. The cross-validation process splits the training data into several sections and uses one section as a testing set and the others as training sets, over multiple iterations. The error estimation is averaged over the iterations to get the total effectiveness of the model. Cross-validation ensures that the model has correctly learned most of the data patterns and does not just pick up on noise.

Related to hyperparameter tuning is training the model with early stopping. Early stopping involves training the model until the validation score (the performance score on the validation set) stops improving. Early stopping ensures that the model does not overfit the training data and generalises well to unseen data. As described in an earlier section, I have created a separate validation set to use for this process.

Table 4.1 displays the selected values for each of the hyperparameters, along with whether the respective hyperparameter was tuned using GridSearchCV. In Appendix B, I describe each of the hyperparameters. These hyperparameters can be broadly categorised into gradient boosting methods, regularisation, and training parameters.

Table 4.1: Hyperparameter tuning

This table presents the final value to which each hyperparameter was set, and whether the hyperparameter was tuned using GridSearchCV. Not every hyperparameter was altered from its default value. Only those which were adjusted or considered important are listed in this table. Descriptions of what each parameter does are listed in Section 9.1 in Appendix B.

Parameter	Value	Tuned with GridSearchCV
seed	52	No
boosting_type	gbdt	No
max_depth	2	No
objective	binary	No
num_leaves	6	Yes
learning_rate	0.005	No
max_bin	512	No
subsample	0.7	Yes
subsample_freq	1	No
colsample_bytree	0.65	Yes
lambda_1	1	Yes
lambda_2	1	Yes
is_unbalance	True	No
metric	auc	No
num_iterations	1000	No
early_stopping_rounds	100	No

5 | Results

5.1 Main results

After training and tuning the model, it is tested on the testing data to determine to what degree it can predict SEOs. As discussed earlier, corporate events such as SEOs are tough to predict; they are driven by numerous factors that may not be picked up by accounting- or economic-based predictor variables. For example, a manager may, by chance, come across a great investment opportunity that they need to raise capital for, or an unexpected crisis may arise, requiring capital to sustain the business.

Given the difficulty in predicting corporate events, the performance results are very positive. As shown in Figure 5.1, the model has an AUC score of 0.78, which is significantly better than chance alone, which would give a score of 0.50. Figure 5.2 plots a confusion matrix that provides a summary of the predictive results. Correct and incorrect predictions are summarised in the table with their values and broken down by each class - non-SEO (Class 0), and SEO (Class 1). The bottom right quadrant shows that 62% of SEO predictions were correct. This result is impressive given the highly significant class imbalance. It shows that the model can sort through the massive number of non-SEOs to predict SEOs.

Additionally, the confusion matrix shows that the model has predicted 77% of non-SEOs correctly. Because of the class imbalance, this relatively high degree of non-SEO predictive ability is essential, as most months are likely to be non-SEO

months.

Figure 5.1: GBM model AUC-ROC curve

This figure presents the AUC-ROC curve for the GBM model. The graph is constructed by plotting the TPR against the FPR. The model's AUC score, which is the primary performance metric in this research, is 0.78.

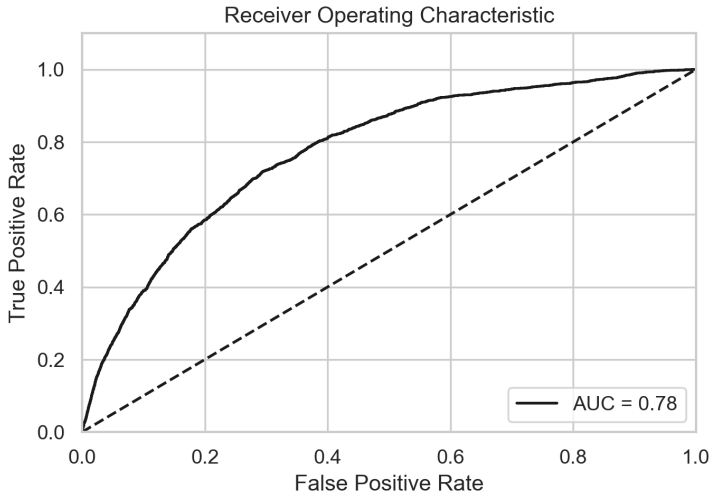
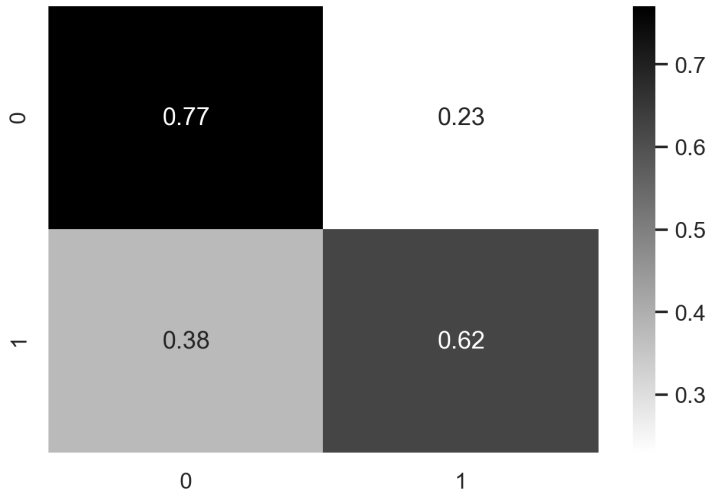


Figure 5.2: GBM model confusion matrix

The figure below presents the confusion matrix for the GBM model. The top two quadrants show the correctly and incorrectly classified non-SEOs respectively. The bottom two quadrants show the incorrectly and correctly classified SEOs respectively.



Overall, SEOs are predictable to some degree using my ML algorithm. As with other corporate events, it is unlikely that SEOs could be entirely predictable, and considering the large class imbalance, the model's results are noteworthy. Thus, extracting feature importance from the model will provide valuable insights into the determinants of a firm's decision to issue seasoned equity. Furthermore, practitioners interested in predicting SEOs would have an edge using this ML model that could not be found elsewhere, highlighting the model's utility. As described earlier, what can be considered acceptable model performance is very context-dependent. Section 5.3 builds a traditional logistic regression model that provides a benchmark for my GBM model. As will be discussed, the findings support the notion that my ML model can uncover essential relationships that drive SEOs which linear models fail to identify.

5.2 Predicting with top ten most important features

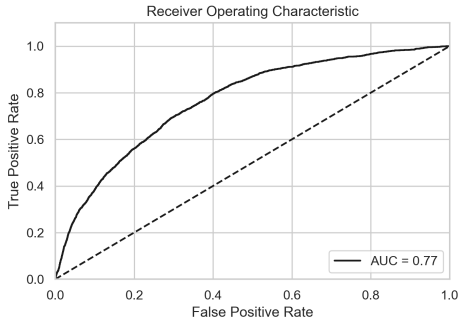
I construct a GBM model using a set of the 10 most important features, selected from the full dataset. The feature importance ranking uses the SHAP importance measure, which is described in Chapter 6. This analysis allows me to assess how important the omitted features are compared to the top 10 features. The results thus aid the generation of economic insights; for example, if the top 10 most important features drive the predictive performance, these features are the key determinants of whether a firm chooses to issue seasoned equity.

I rerun the GBM model with these 10 variables. The AUC score of 0.77 remains roughly the same as the full model. The confusion matrix shows that the model is still reasonably good at distinguishing between SEOs and non-SEOs, with the number of correctly predicted SEOs falling only slightly from 62% to 59%.

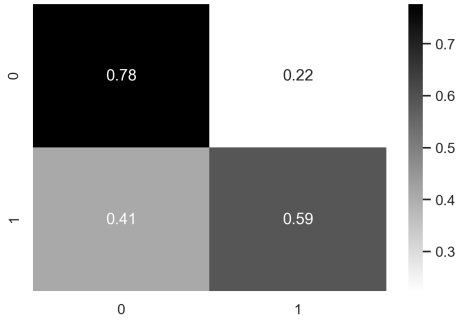
Overall, these results show that the 10 most important features drive most of the model's predictive performance. This result is interesting as it suggests that these 10 features are the primary drivers of a firm's decision to issue seasoned equity. Thus, the focus will be on these 10 features when analysing feature importance.

Figure 5.3: GBM model performance using top ten most important features

The figures below show the GBM model's performance using the 10 most important features, as measured by SHAP values. The model's construction is unaltered from the full model, and the only difference is using a reduced set of features. Panel A presents the AUC-ROC curve, showing that the model has an AUC score of 0.77. Panel B presents the confusion matrix, showing the correctly and incorrectly classified non-SEOs (0), and SEOs (1).



Panel A: GBM model's AUC-ROC (using top ten features)



Panel B: GBM model's confusion matrix (using top ten features)

5.3 Predicting with logistic regression

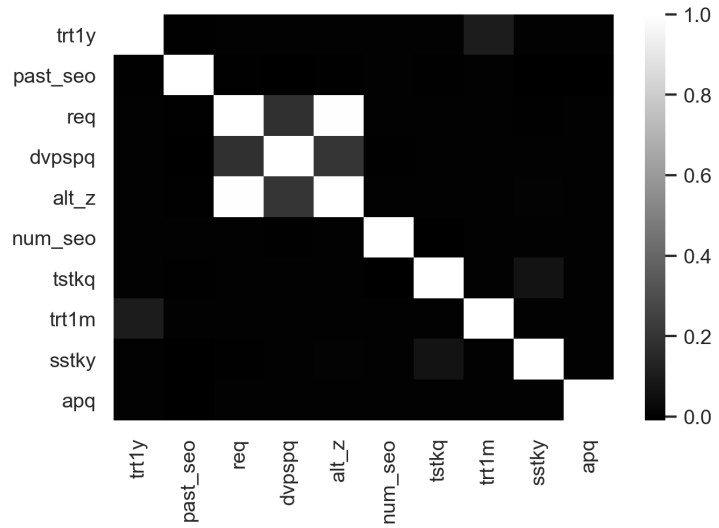
I test whether a logistic regression model can be used to predict whether a firm will engage in an SEO using the set of the 10 most important features described above. Specifically, I am using an ML-based logistic regression model. As variables can only enter linearly into a logistic regression, this analysis allows me to investigate how much of the predictability is due to GBM's ability to uncover the complex and non-linear relationships between the predictor variables and a firm's decision to conduct an SEO. Moreover, these findings can provide a benchmark with which to compare my GBM model's performance.

I start by ensuring that none of the features are highly correlated with one another,

to avoid multicollinearity issues, which are a problem in linear models. Figure 5.4 plots a correlation heatmap which shows that the *req* and *alt_z* features are highly correlated. To avoid multicollinearity issues, I drop *alt_z* from the analysis using linear models. I choose to drop *alt_z* instead of *req* as it is the least important of the two features.

Figure 5.4: Correlation heatmap between important features

The figure plots a correlation heatmap between the 10 most important features, as measured by SHAP values. The white squares represent highly correlated features, and the black squares represent highly uncorrelated features, as indicated by the legend. The only variables that are highly correlated are *req* and *alt_z*.



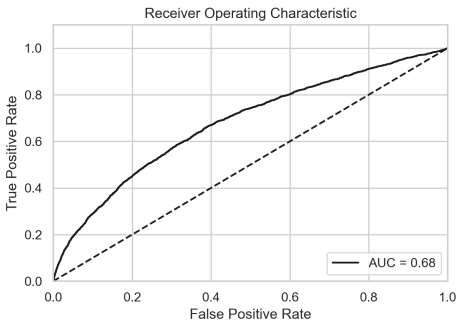
The logistic regression's AUC score of 0.68, shown in Figure 5.5 Panel A, is substantially lower than the full GBM model's score of 0.78. The confusion matrix shows that the reduced performance stems from the model's inferior ability to predict both non-SEO and SEO months, relative to the main GBM model. Correctly classified SEOs fall from 62% to 57% and correctly classified non-SEOs fall from

77% to 70%.

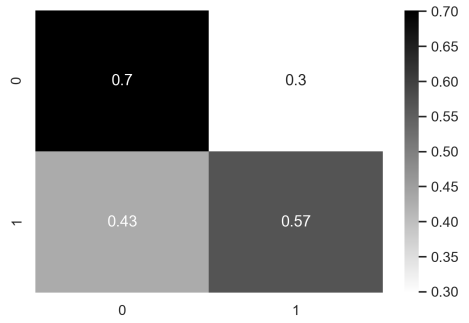
Overall, these findings indicate that the GBM model's ability to include non-linearities and complex interactions among variables is essential for SEO prediction. Furthermore, the results support Lopez de Prado's (2019) assertion that ML offers a set of tools useful for understanding the complex relationships in finance.

Figure 5.5: Logistic regression model performance

The figures below show the performance of the logistic regression model trained using the nine most important features. Only nine features were used, as the *alt_z* variable was very highly correlated with *req*, which may have caused multicollinearity issues if both variables were included in the regression. Panel A presents the AUC-ROC curve, showing that the model has an AUC score of 0.68. Panel B presents the confusion matrix, showing the correctly and incorrectly classified non-SEOs (0), and SEOs (1).



Panel A: Logistic regression model's AUC-ROC (using top ten features)



Panel B: Logistic regression model's confusion matrix (using top ten features)

I also run a traditional non-ML logistic regression in a Python statistics package to get a regression output with interpretable coefficients and t-statistics. The summary table below shows that *trt1y* and *trt1m* are insignificant. However, as shown previously, all of these 10 variables are important predictors of SEOs. Thus, their insignificance in the logistic regression model exposes weaknesses in limiting fi-

nance research to only using linear models. For example, *trt1y* is the most important feature in the main GBM model, and its relevance is backed by economic theory, yet it is insignificant in the traditional logistic regression model. This insignificance is likely due to the fact that its relationship with a firm's decision to conduct an SEO is non-linear, as will be discussed in Section 6.3.

Table 5.1: Traditional logistic regression output

This regression output is from a traditional (non-ML) logistic regression run using a Python-based statsmodels package. It uses the top nine most important features, as measured by SHAP values, and omits *alt_z* due to how highly correlated it is with *req*, leading to possible multicollinearity problems.

	Dependent variable:
	(1)
trt1y	0.000
	(0.000)
past_seo	1.183***
	(0.022)
req	-0.010***
	(0.001)
dvpspq	-29.321***
	(9.782)
num_seo	0.004***
	(0.000)
tstkq	-0.300***
	(0.078)
trt1m	-0.000
	(0.000)
sstky	0.008***
	(0.002)
apq	-0.684***
	(0.057)
const	-5.873***
	(0.024)
Observations	1,904,549
R^2	
Pseudo R^2	0.03038
Residual Std. Error	1.000(df = 1904538)
F Statistic	(df = 10.0; 1904538.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

5.4 Predicting with random undersampling

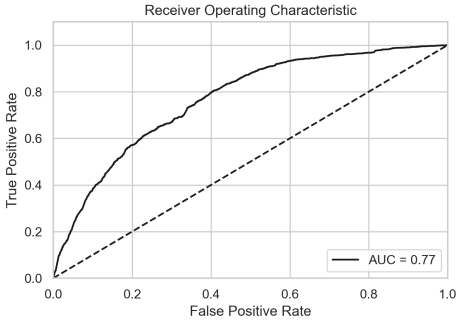
Another way to address the class imbalance in the dataset is to use random undersampling. As mentioned, a bias towards one class in the training dataset may negatively affect ML algorithms, sometimes even resulting in the model completely ignoring the minority class. In the main GBM model, this imbalance was dealt with by setting the 'is_unbalance' parameter to 'True'. Since random undersampling is a commonly used alternative for addressing class imbalances, it is worth investigating how it would affect the model's results, and whether it is more or less preferable to simply changing the model's 'is_unbalance' parameter setting.

In random undersampling, examples from the majority class, non-SEO in this case, are randomly selected and deleted from the training dataset. Using this method, the result is a training set comprising an equal number of SEO and non-SEO transactions. As there are 4,512 SEOs in the original training set, the total undersampled training set comprises 4,512 SEOs and 4,512 non-SEOs. The validation and testing dataset remain the same, as undersampling these sets would not accurately represent the real SEO/non-SEO ratio.

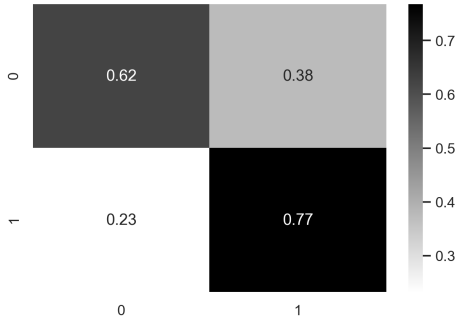
The resulting AUC score and confusion matrix are shown in Figure 5.6. The findings show that the model's performance scores are only slightly lower than those of the main model. This result indicates that the ability to predict SEOs using a GBM model is robust to alternative methods of addressing class imbalances; however, I conclude that the use of the 'is_unbalance' parameter is the optimal choice.

Figure 5.6: Model performance using random undersampling

The figures below show the GBM model's performance using random undersampling to address the class imbalances, rather than using the 'is_unbalance' hyperparameter. Panel A presents the AUC-ROC curve of the model, which has an AUC score of 0.77. Panel B presents the confusion matrix, which shows the number of correctly and incorrectly predicted non-SEOs and SEOs, using the testing data set.



Panel A: GBM model's AUC-ROC curve using random undersampling



Panel B: GBM model's confusion matrix using random undersampling

6 | Feature importance

6.1 Methodology

By analysing the drivers of SEO predictability, I can develop economic insights and contribute to corporate financing theory. This section will discuss the methodology of determining feature importance and then relate the findings to the broader finance literature.

Lundberg, Erion, and Lee (2018), when outlining the challenges of determining feature importance in complex ML models, propose the use of SHAP values which were developed by Lundberg and Lee (2017). The GBM model has its own default feature importance measures; however, the drawbacks to using these measures are summarised in Appendix C. The SHAP model is an ML interpretability method based on the Shapely value, a solution concept in co-operative game theory (Shapley, 1953).

In general terms, in co-operative game theory, a group of players can form a joint strategy to create value. The Shapely value provides a solution to how the value/reward should be fairly divided amongst a group of diversely skilled actors working in a coalition. The value applies primarily in situations where the contributions of the participants working in cooperation are not equal. Importantly, this concept can be applied to ML. By using a predictive model such as the one developed in my research, the 'game' is reproducing the model's outcome, and the 'players' are the features. The Shapely value calculation in an ML context

essentially gets the marginal contributions of a feature across all permutations and then takes the average.

Lundberg and Lee's (2017) SHAP value is inspired by several methods and uses Shapely values, the main concept of which remains the same when calculating SHAP values. This method is built on by Lundberg et al. (2020), who further develop the SHAP value by creating tree-based SHAP values. These tree-based SHAP values have the advantage of being able to be calculated precisely, unlike traditional SHAP values which can only be approximated, since computing them is NP-hard.

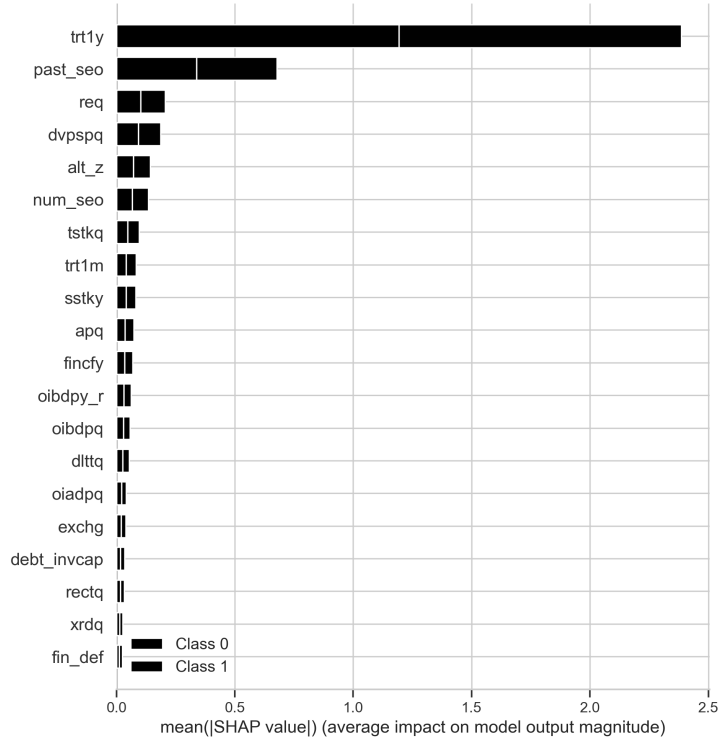
6.2 Main model feature importance

As measured by the SHAP values, feature importance results are shown in Figure 6.1. The 10 most important features are *trt1y*, *past_seo*, *req*, *dvpspq*, *alt_z*, *num_seo*, *tstkq*, *trt1m*, *sstky*, and *apq*. The definitions of these variables can be found in Table 3.1. These findings can be used in conjunction with traditional finance theory to develop insights into how firms make capital structure and financing decisions. As discussed, although ML-based research does not provide robust evidence about the underlying mechanisms that drive SEOs, it does identify the variables that should be part of the theory (Lopez de Prado, 2019). However, I will provide some basic economic intuition behind how the important features may drive SEOs. The discussion aims to provide broad ideas and link the results back to past literature, but I do not claim to know precisely how the model uses the features for its predictions. Future research could build on the results by using

them to form testable hypotheses further investigate. A notable caveat to feature importance interpretation is that, just because a particular feature is a good predictor, it does not mean it is necessarily responsible for the firm's decision to issue seasoned equity. The feature may instead be a proxy for other characteristics. In saying this, the ability to include many features in a ML model is a step in the right direction for uncovering the true determinants of SEOs. When analysing these results, it is important to remember that the data were lagged, as detailed in Chapter 3.

Figure 6.1: Feature importance using SHAP values

The figure plots the importance of each feature using SHAP values. The 20 most important features are included in the graph. We can clearly see that the most important feature is *trt1y*, a firm's stock return over the past year. Whether a firm has issued equity in the past, represented by *past_seo*, is also highly important, but the importance of other variables quickly begins to diminish. Each importance bar is split into the importance of the feature for predicting non-SEOs and SEOs respectively. Each bar is split approximately equally, suggesting that each feature is roughly as important for predicting non-SEOs as it is for predicting SEOs.



As shown in Figure 6.1, the *trt1y* variable, which is the cumulative return over the previous year, is the most significant feature by a substantial margin. This finding is in line with Graham and Harvey's (2001) finding that managers issue stock during a 'window of opportunity', occurring after a large share price increase, and Lucas and McDonald's (1990) model, where abnormal positive returns are predicted to preceded SEOs. While no conclusions regarding long-run capital

structure can be made from these findings, the importance of *trt1m* is also consistent with Baker and Wurgler's (2002) market timing theory, where managers time equity issuances to favourable market conditions. It is interesting how much more significant this variable is than the rest; its value implies that managers primarily focus on their stock price when deciding to issue additional equity. However, one must remember that, given the model's ability to pick up on complex interactions between predictors and how the SHAP values are calculated, it may be the interaction of past stock return with other factors, for example, financing deficit or bankruptcy probability, that gives the variable its importance.

The next two most important variables from my GBM model are *past_seo* and *req*. The importance of the *past_seo* variable - whether the firm has previously conducted an SEO in the sample period - is interesting, as it is not often discussed as being a key determinant of equity issuances. However, it is makes sense that a firm that has issued seasoned equity in the past is more likely to do so again. This could be due to its size, structure, managerial tendencies, and growth opportunities. A firm's retained earnings, *req*, is related to pecking order theory, which suggests that a firm will only issue equity once it has exhausted its retained earnings and debt financing capacities, respectively. Thus, the pecking order would imply that a deficient retained earnings value would raise the chances that a firm issues equity, and vice versa.

Next in terms of importance, are *dvpspq*, *alt_z*, and *num_seo*. Dividends per share, *dvpspq*, could help predict SEOs for a variety of reasons. For example, firms which pay dividends are likely to be more mature with stable cash flows (Damodaran, 1996). These firms may have more debt capacity, and thus be less

likely to issue equity under both pecking order and trade-off models. The *alt_z* variable is the firm's Altman Z-Score, which gauges its likelihood of bankruptcy (Altman, 1968). A firm with a higher probability of bankruptcy can strengthen its financial position or raise capital for investments by issuing equity. Both the trade-off and pecking order theories indirectly suggest that *alt_z* is an important driver of a firm's decision to issue equity. Under trade-off theory, a firm will choose to use equity rather than debt if the costs of debt are higher than the benefits, and these costs increase with higher *alt_z* scores. Likewise, under pecking order theory, a firm will finance its investments with equity only once it can no longer use retained earnings or debt. Thus, a higher *alt_z* score is related to the depletion of retained earnings and debt capacity, providing theoretical backing for the relation between *alt_z* and a firm's decision to issue seasoned equity. The *num_seo* variable is the number of SEOs that were conducted market-wide in the month prior. This result supports Howe and Zhang's (2010) finding that there are SEO cycles, similar to the IPO cycles documented by Ibbotson and Jaffe (1975) and Ibbotson, Sindelar, and Ritter (1994).

Finally, in terms of importance, there are *tstkq*, *trt1m*, *sstky*, and *apq*. Total Treasury Stock, *tstkq*, is the total amount of stock bought back by a firm. The process of buying back stock reduces the number of shares outstanding on the open market. Managers may buy back stock if they believe it to be overvalued or sometimes to protect against a takeover. Since firms are likely to issue equity when they believe their shares are overvalued and repurchase shares when they believe their stock to be undervalued, a large amount of treasury stock may indicate that managers are unlikely to conduct an SEO. A firm's past month's stock return,

trt1m, tells a similar story to *trt1y*. Given its importance, it implies that it has the unique predictive ability not captured fully by *trt1y*. Sale of Common and Preferred Stock, *sstky*, is the amount of both common and preferred stock a company has sold in the past. Its importance is interesting, as it suggests that it has predictive power beyond that of *past_seo*, that is, whether a company has issued equity in the past. The predictive power may arise from the magnitude of past equity issuances, with firms that have raised more in the past being likely to conduct an SEO to fund growth opportunities or financing deficits. A high value of Trade accounts payable, *apq*, may indicate that a firm is in a stable financial position as suppliers are willing to extend the firm credit in the form of accounts payable. This may be linked to a firm's debt or internal financing capacity, meaning that they are less likely to need to issue equity.

Overall, the majority of the important features for predicting SEOs have intuitive explanations, and many are in line with traditional financing theory. One of the key contributions of extracting feature importance from my ML model is that it clarifies what they key variables are for predicting whether a firm will issue seasoned equity. Furthermore, while the intuition behind the predictors is up for discussion, their importance is not. As mentioned, one must be careful when interpreting the findings, as the model can capture multi-way interactions among the variables, and these interactions may not always be apparent. In Chapter 7, I further discuss the implications of the feature importance findings and how they support various capital structure theories.

6.3 Feature comparison between classes

This section compares SEO and non-SEO firms' characteristics, using the 10 most important features identified by the SHAP values in the previous section. The aim is to derive additional economic intuition around how the important features relate to a firm's decision to conduct an SEO. Table 6.1 presents summary statistics, with the means for each feature grouped by whether the observation corresponds to an SEO or not. Figure 6.2 then plots the distributions of the important features. When there is separability between feature distributions of classes, it is an indication that an ML model can distinguish between them and use the respective features to make correct predictions. Given the algorithm's ability to model complex relationships, there is likely more to the important features' usefulness than merely class separability, but it nevertheless provides some understanding. Additionally, no rigorous statistical testing is carried out; instead, these findings aim to present general ideas to aid the discussion.

Table 6.1: Summary statistics

This table provides summary statistics for each of the 10 most important features, grouped by non-SEO (0) and SEO (1). Additionally, t-statistics for differences in means are reported in the fourth column.

seo	0	1	t-stat
trt1y	21.58	41.91	11.77
past_seo	0.24	0.52	51.08
req	0.81	-0.46	-12.69
dvpspq	0.0004	0.0002	-6.54
alt_z	1.91	-0.23	-6.83
num_seo	29.58	31.74	8.65
tstkq	0.09	0.04	-2.18
trt1m	2.46	2.52	0.09
sstky	0.05	0.12	12.46
apq	0.45	0.12	-15.01

Figure 6.2: Histogram comparisons between non-SEOs (0) and SEOs (1)

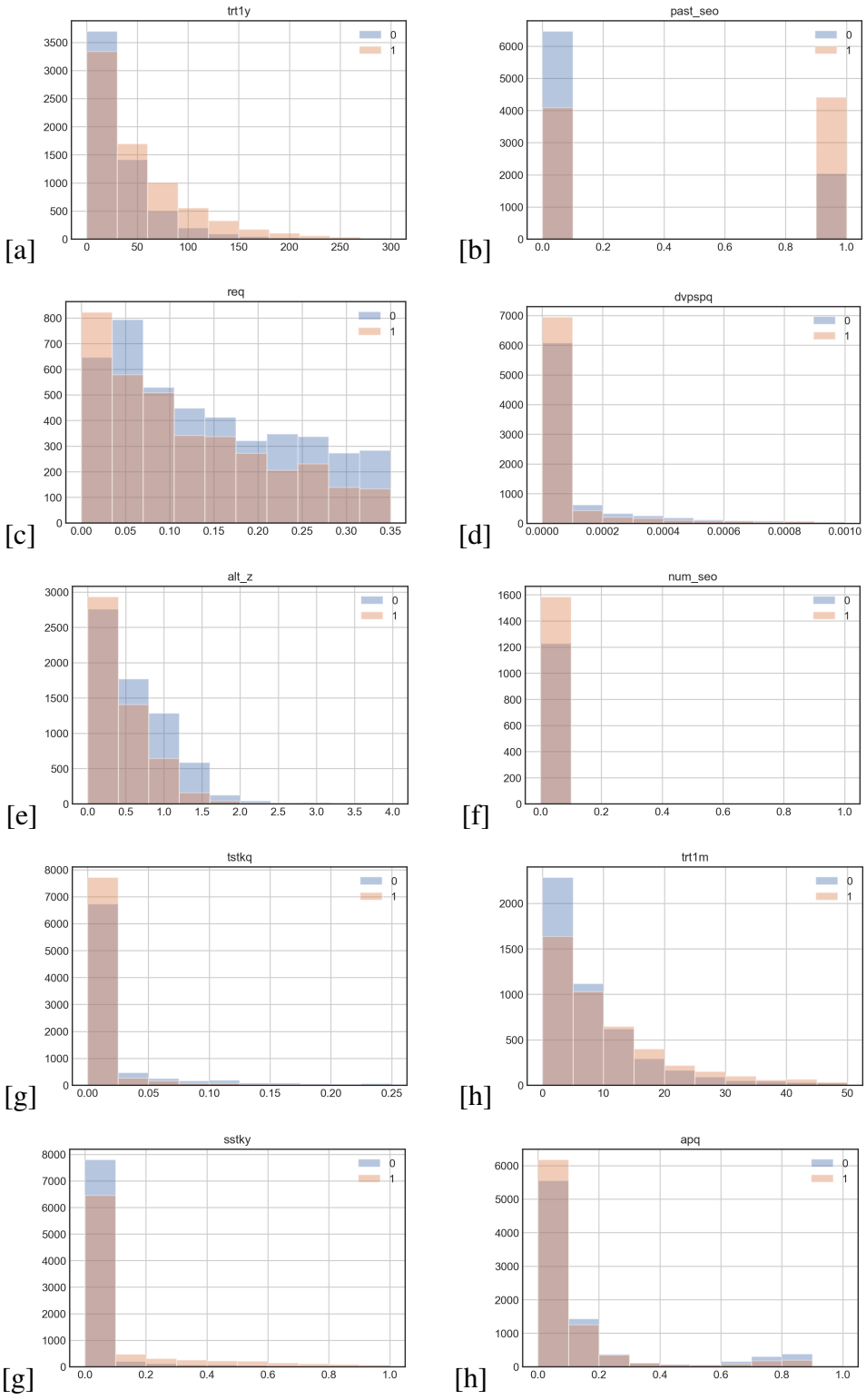


Table 6.1 and Figure 6.2 show that SEO and non-SEO firms have meaningful differences in their mean values for all important features. T-statistics for differences in means also shown in Table 6.1[1]. All differences are significant, *trt1m*. This lack of significance could be indicative of the feature's importance arising from a non-linear relationship with a firm's decision to conduct an SEO. Suppose the feature has a non-linear relationship with a firm's decision to conduct an SEO and complex interactions with other features. In that case, it is not surprising that the GBM model can use the feature to discriminate between classes, even if the differences in means are not significant.

With the most important feature, *trt1y*, the mean value for SEO firms is 94% higher than that of non-SEO firms, consistent with prior research findings that firms issue equity after run-ups in share price. The differences in the monthly return, *trt1m*, tell the same story; however, the difference is insignificant.

SEO firms have a higher mean for *past_seo*, suggesting that SEO firms are more likely to have been past issuers. The values for retained earnings, *req*, suggest that issuers tend to have negative retained earnings, in line with theories such as pecking order where firms turn to equity as a last resort once they have exhausted retained earnings and debt capacity. Similarly, dividends per share, *dvpspq*, is significantly higher for non-issuers, implying they may have more stable cash flows that can support debt or internal financing. *alt_z* is interesting, as it appears that issuers have lower scores, that is, a lower probability of bankruptcy. The number of market-wide SEOs is significantly higher for SEOs than non-SEOs, implying that seasoned equity offerings occur in cycles. Total Treasury Stock,

[1]Given unequal variances, a Welch's t-test was used.

tstkq, is significantly lower for SEOs, in line with the fact that high treasury stock may correspond to undervalued equity. SEO firms have a significantly higher sale of common and preferred stock, *sstky*, indicating that the amount of equity sold in the past is informative about a firm's future equity financing decisions. Accounts payable, *apq*, is significantly lower for SEO firms, supporting the assertion that firms with high accounts payable may be more stable and able to use debt or internal financing instead of resorting to costly equity.

Overall, these results help to give some idea of the relationships that drive SEOs. The results suggest that, at a firm level, there are differences between the firms who engage in SEOs and those who do not. While they do not provide any robust evidence, they do motivate further discussion. I also run traditional univariate logistic regressions, with the results being displayed in Table 6.2. These findings, along with the multivariate regression results in Table 5.1, provide further insights into how these top 10 features impact a firm's decision to issue equity, and whether the respective relationships are positive or negative.

Given the *trt1y* features' lack of significance in the univariate regression (Table 6.2), I analyse the relationship between a firm's likelihood of conducting an SEO by *trt1y* deciles. The *trt1y* feature's lack of significance could be due to non-linearities, as mentioned, and analysis by decile provides a rough idea of whether this is the case. Table 6.3 confirms the presence of non-linearities. As expected and consistent with market-timing, the bottom-most decile is significantly less likely to raise equity, and the top-most decile is significantly more likely to issue equity. In the middle deciles, the relationship directions are mostly as expected, but the relationship is not quite linear. Given that this is just a simple univari-

ate regression, further analysis would be needed to investigate this relationship's dynamics properly. However, there are a few potential explanations. To give an example, we see that decile 0 firms are actually more likely to conduct an SEO than decile 1 firms, inconsistent with market timing. One possible explanation could be that firms that have expected very negative stock returns have exhausted equity, which needs to be replenished via SEO to avoid the threat of bankruptcy or takeover. Overall, one can conclude that the relationship is non-linear, and the ability of the ML algorithm to pick up on this is what drives the feature's usefulness.

Table 6.2: Univariate regression output

The output below provides univariate regression results for each of the 10 most important features. The regressions are logistic regressions that use the same SEO dependant variable as in the main model.

	Dependent variable: SEO									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
trt1y	0.000									
	(0.000)									
past_seo		1.218***								
		(0.022)								
req			-0.005***							
			(0.001)							
dvpspq				-79.229***						
				(13.250)						
alt_z					-0.002***					
					(0.000)					
num_seo						0.005***				
						(0.000)				
tstkq							-0.240***			
							(0.068)			
trt1m								0.000		
								(0.000)		
sstky									0.007***	
									(0.002)	
apq										-0.828***
										(0.058)
const	-5.407***	-5.862***	-5.407***	-5.386***	-5.406***	-5.545***	-5.396***	-5.407***	-5.407***	-5.292***
	(0.011)	(0.016)	(0.011)	(0.011)	(0.011)	(0.019)	(0.011)	(0.011)	(0.011)	(0.013)
Note:										*p<0.1; **p<0.05; ***p<0.01

Table 6.3: *trt1y* univariate regressions by decile

The output below shows several univariate logistic regressions. The independent variable in each regression is a particular decile of the *trt1y* variable. This analysis allows one to see whether the direction of the relationship between *trt1y* and a firm's likelihood of conducting an SEO changes based on decile, highlighting any non-linearities in the relationship. Decile 0 is the decile of the lowest annual stock returns, and decile 1 is the decile of the highest annual stock returns

	Dependent variable: SEO									
Decile	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	-0.755***									
	(0.050)									
1		-1.028***								
		(0.057)								
2			0.193***							
			(0.027)							
3				-0.635***						
				(0.091)						
4					-0.783***					
					(0.051)					
5						-0.417***				
						(0.043)				
6							-0.244***			
							(0.040)			
7								-0.011		
								(0.036)		
8									0.336***	
									(0.032)	
9										1.184***
										(0.025)
const	-5.352***	-5.340***	-5.443***	-5.394***	-5.351***	-5.372***	-5.385***	-5.406***	-5.446***	-5.610***
	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.013)

Note: *p<0.1; **p<0.05; ***p<0.01

7 | Discussion

7.1 Overview

This study asked two questions: can ML successfully predict whether a firm will engage in an SEO and what are the most important determinants are for this decision? For the model to be a good predictor, it would have to correctly model the relationships that drive SEOs. Since my GBM model has predictive power, it implies that it has correctly 'learned' most of the determinants of SEOs, at least to the degree to which they are predictable using historical data. Thus, by analysing the model's feature importance, one gains real insight into what makes a firm choose to issue seasoned equity. As numerous features are included in the dataset, and no model is specified, these relationships and important features are solely based on reality. One can be confident that the important features are true drivers of equity issuances or at least good proxies.

7.2 Performance

Overall, the model's predictive ability is significant. However, it is unclear whether its performance is good enough to use in a practical context. Since SEOs are related to future stock performance, with issuers underperforming their peers, it is feasible that a trading strategy could be created based on the results of this paper. This strategy would entail going short stocks most likely to conduct an SEO, and long stocks least likely to conduct an SEO. However, the model makes enough

incorrect predictions that a trading strategy may not be profitable. Additionally, another practical use of the ML model's predictive power is that investment banks could target their equity capital markets (ECM) origination towards firms that are most likely to need to issue equity. While being appointed to advise or underwrite an equity issue is mostly a result of relationships with firm management, having some idea of the firms most likely to need an SEO advisor could help direct or aid origination.

7.3 Insights

As described previously, the decision to issue seasoned equity is related to a firm's capital structure choice. Specifically, if a firm needs capital, it can turn to internal funds, debt, or equity. Thus, if a firm conducts an SEO, it implies that it has chosen to use equity instead of alternative financing options. Therefore, examining the drivers of SEOs provides insight into the factors that may be important to capital structure decisions.

When examining the GBM model's feature importance, it is evident that the most important feature is a firm's recent stock return. This finding suggests that managers take advantage of their stock being highly valued, issuing equity in a window of opportunity, as suggested by Graham and Harvey (2001). An obvious question is whether the extreme importance of prior stock return supports Baker and Wurgler's (2002) market timing hypothesis. While this would be a tempting conclusion, Hovakimian (2006) finds that equity issuances do not have persistent capital structure impacts even if managers time their equity issuances.

However, the variables related to other capital structure theories, such as effective tax rates, non-debt tax shields, and debt capacity, are notably unimportant compared to factors such as recent stock return and whether the firm has issued equity in the past. If pecking order or trade-off theories of capital structure hold, one would expect that features related to these should be more important predictors of SEOs than what the GBM model suggests. For example, suppose trade-off theory was the most accurate capital structure theory. In that case, variables such as effective tax rates (*efftax*), depreciation tax shield (*nd_ts*) and debt to equity ratio (*de_ratio*) should be important predictors. Yet that is not the case. While there are a few important features related to pecking order and trade-off theories, most notably retained earnings (*req*) and the firm's Altman Z-Score (*alt_z*), these features are substantially less important than recent stock return and equity issuance history.

One possibility is that interactions between the traditional capital structure theory variables and stock return or market condition variables give the model predictive power. Since the traditional linear model was shown to be an inferior predictor to the GBM model, it indicates that complex interactions and non-linear relations drive predictions. To picture this using a very general example, imagine that pecking order theory held, and retained earnings and debt capacity were important variables. Perhaps managers still time the market when issuing equity, and there is a non-linear relationship across all three variables that the model uncovers to form its predictions.

When trialling the model using only the 10 most important features, nearly all of the performance remained, indicating that these 10 features were responsible for

most of the model's predictive ability. This result suggests that all of the other features offer little predictive power, hence are negligible in a firm's decision to issue equity. Given this, if interactions between features give the model predictive power, it would only be the interactions between these 10 features that are important, implying that many of the theoretically motivated features are unimportant in a firm's SEO decisions.

The lack of support for any particular capital structure theory is consistent with the lack of consensus in traditional corporate financing and capital structure research. Importantly, it suggests that firms' financing decisions are complex and cannot be described by a single theory. If this is the case, it is very difficult for researchers to model the relationships that drive financing or leverage decisions using econometric methods, as it is unclear what features or interactions need to be included in the model. My ML-based approach's most significant contribution is that it allows the true relationships to be discovered and then analysed. These relationships that drive SEOs must, to some extent, be linked to a firm's capital structure.

By analysing the feature importance, one can form an idea of what a firm's decision-making process may look like when issuing seasoned equity. Based on the initial feature importance discussion and feature comparison between SEOs and non-SEOs, the following story emerges.

It is very apparent managers heavily focus on recent stock return when choosing to conduct an SEO, that is, they time the market. Whether the firm has issued equity in the past is also important, likely related to specific firm characteristics or growth opportunities. It may also imply that some managers may have a prefer-

ence for issuing stock. A firm's retained earnings is also an important determinant of SEOs. Intuitively, firms with high retained earnings are less likely to need to raise external capital. Dividends per share are also important. Dividends are likely an indicator that a firm generates sufficient internal funds not to need external financing. Alternatively, dividends may proxy for lack of growth opportunities or debt capacity. The idea of SEO cycles is also important, perhaps due to their relation with market-wide stock valuations, behavioural tendencies to copy other managers in their decisions, or information learned from other firms' SEOs.

7.4 Future research

As mentioned previously, future research could investigate whether a trading strategy could successfully use the GBM model's predictions. Additionally, using the feature importance findings, hypotheses could be developed and tested, possibly generating new ideas of capital structure theory.

Given the ML model's predictive ability, novelty, and insights into SEO drivers, it would be interesting to investigate which other corporate events ML models can predict. Possibilities include mergers and acquisitions, and debt issues. Both of these would offer similar types of insights and be interesting to practitioners. The GBM model would likely not have to be altered a great deal, and the dataset could remain mostly the same, albeit with a different target variable. As with SEOs, a trading strategy could be a possible extension for research that uses ML to predict other corporate events. For example, acquirers tend to experience negative stock returns upon M&A transaction announcement, and target firms tend to experience

positive stock returns. A trading strategy could short stocks that are most likely to be acquirers and go long firms that are likely to be targets. Additionally, M&A bankers may also be able to use these predictions for origination.

8 | Conclusion

ML-based predictive models can predict SEOs with a noteworthy level of performance. The GBM model considerably outperforms a logistic regression model, indicating that there are complex relationships between predictor variables and a firm's decision to issue seasoned equity. The GBM model's performance suggests that an ML-based approach has a place in finance research, as it can aid the discovery of complex and non-linear relationships that characterise many economic phenomena.

Based on SHAP values, a concept from game theory that allows feature importance to be better understood, 10 features contribute to most SEO predictability. The most important feature is a firm's past year's stock return, suggesting that managers focus heavily on timing their equity issuances. Other important features include whether a firm has issued seasoned equity in the past and how many SEOs occurred market-wide in the month prior.

Many variables suggested by capital structure theories to be important and found to offer no predictive power. Additionally, the important features do not tell a story consistent with any particular capital structure theory, supporting the idea that the lack of consensus in capital structure literature is due to no one theory explaining firms' financing decisions. Instead, SEOs appear to be driven by an eclectic set of variables, and heavily dominated by market timing considerations.

Overall, my results show that ML-based approaches can be useful for their superior predictive ability and the economic insights generated by analysing their

feature importance. Furthermore, the research shows how ML could also be used as the first step of research to understand better the key variables that should be part of a theory.

References

(n.d.).

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Amel-Zadeh, A., Calliess, J.-P., Kaiser, D., & Roberts, S. (2020). Machine learning-based financial statement analysis. *Available at SSRN 3520684*.

Asmis, E. (1984). *Epicurus' scientific method* (Vol. 42). Cornell University Press Ithaca, NY.

Auerbach, A. J. (1985). Real determinants of corporate leverage. In *Corporate capital structures in the united states* (pp. 301–324). University of Chicago Press.

Baker, M., & Wurgler, J. (2000). The equity share in new issues and aggregate stock returns. *the Journal of Finance*, *55*(5), 2219–2257.

Baker, M., & Wurgler, J. (2002). Market timing and capital structure. *The journal of finance*, *57*(1), 1–32.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, *61*(4), 1645–1680.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives*, *21*(2), 129–152.

Bancel, F., & Mittoo, U. R. (2004). Why do european firms issue convertible debt? *European Financial Management*, *10*(2), 339–373.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Bayless, M., & Chaplinsky, S. (1996). Is there a window of opportunity for seasoned equity issuance? *The Journal of Finance*, *51*(1), 253–278.

Bianchi, D., Büchner, M., & Tamoni, A. (2020). Bond risk premiums with machine learning. *The Review of Financial Studies*.

Bradley, M., Jarrell, G. A., & Kim, E. H. (1984). On the existence of an optimal capital structure: Theory and evidence. *The journal of Finance*, *39*(3), 857–878.

Breiman, L. (1997). Pasting bites together for prediction in large data sets and on-line. *preprint*.

Brownlee, J. (2020, Aug). *8 tactics to combat imbalanced classes in your machine learning dataset.* Retrieved from `https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/`

Bryzgalova, S., Pelger, M., & Zhu, J. (2019). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1). Andriy Burkov Canada.

Chae, J. (2005). Trading volume, information asymmetry, and timing information. *The journal of finance*, *60*(1), 413–442.

Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. *Available at SSRN 3350138*.

Copeland, T. E., & Galai, D. (1983). Information effects on the bid-ask spread. *the Journal of Finance*, *38*(5), 1457–1469.

Damodaran, A. (1996). *Corporate finance*. Wiley.

DeAngelo, H., DeAngelo, L., & Stulz, R. M. (2010). Seasoned equity offerings, market timing, and the corporate lifecycle. *Journal of financial economics*, *95*(3), 275–295.

DeAngelo, H., & Masulis, R. W. (1980). Optimal capital structure under corporate and personal taxation. *Journal of financial economics*, *8*(1), 3–29.

De Bie, T., & De Haan, L. (2007). Market timing and capital structure: Evidence for dutch firms. *De Economist*, *155*(2), 183–206.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

Donaldson, G. (1961). Corporate debt capacity, harvard. *Harvard University Press. Downs, TW,(1991).“An alternate approach to fundamental analysis: The asset side of the equation." Journal of Portfolio Management*, *17*(2),

6–17.

Eckbo, B. E., & Masulis, R. W. (1992). Adverse selection and the rights offer paradox. *Journal of financial economics*, *32*(3), 293–332.

Eckbo, B. E., & Norli, O. (2004). The choice of seasoned-equity selling mechanism: theory and evidence. *Tuck School of Business Working Paper*(2004-15).

Fama, E. F., & French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. *The review of financial studies*, *15*(1), 1–33.

Feng, G., Polson, N., & Xu, J. (2020). Deep learning in characteristics-sorted factor models. *Available at SSRN 3243683*.

Fischer, E. O., Heinkel, R., & Zechner, J. (1989). Dynamic capital structure choice: Theory and tests. *The Journal of Finance*, *44*(1), 19–40.

Frank, M. Z., & Goyal, V. K. (2003). Testing the pecking order theory of capital structure. *Journal of financial economics*, *67*(2), 217–248.

Frank, M. Z., & Goyal, V. K. (2008). Trade-off and pecking order theories of debt. In *Handbook of empirical corporate finance* (pp. 135–202). Elsevier.

Frank, M. Z., & Goyal, V. K. (2009). Capital structure decisions: which factors are reliably important? *Financial management*, *38*(1), 1–37.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, *15*(1-2), 107–143.

Geertsema, P., & Lu, H. (2019). Machine valuation. *Available at SSRN 3447683*.

Geertsema, P., & Lu, H. (2021). The cross-section of long-run expected stock returns.

Graham, J. R., & Harvey, C. R. (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of financial economics*, *60*(2-3), 187–243.

Gu, S., Kelly, B., & Xiu, D. (2020a). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gu, S., Kelly, B., & Xiu, D. (2020b). Autoencoder asset pricing models. *Journal of Econometrics*.

Halov, N., & Heider, F. (2011). Capital structure, risk and asymmetric information. *The Quarterly Journal of Finance*, *1*(04), 767–809.

Harjoto, M., & Garen, J. (2003). Why do ipo firms conduct primary seasoned equity offerings? *Financial Review*, *38*(1), 103–125.

Harris, M., & Raviv, A. (1991). The theory of capital structure. *the Journal of Finance*, *46*(1), 297–355.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business

Media.

Henderson, B. J., Jegadeesh, N., & Weisbach, M. S. (2006). World markets for raising new capital. *Journal of Financial Economics*, *82*(1), 63 - 101. doi: https://doi.org/10.1016/j.jfineco.2005.08.004

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hovakimian, A. (2006). Are observed capital structures determined by equity market timing? *Journal of Financial and Quantitative analysis*, 221–243.

Hovakimian, A., Opler, T., & Titman, S. (2001). The debt-equity choice. *Journal of Financial and Quantitative analysis*, 1–24.

Howe, J. S., & Zhang, S. (2010). Seo cycles. *Financial Review*, *45*(3), 729–741.

Huang, R., & Ritter, J. R. (2009). Testing theories of capital structure and estimating the speed of adjustment. *Journal of Financial and Quantitative analysis*, 237–271.

Hume, D. (1739). *A tretise on human nature*. Clarendon Press.

Ibbotson, R. G., & Jaffe, J. F. (1975). 'hot issue' markets. *The journal of finance*, *30*(4), 1027–1042.

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.

Jalilvand, A., & Harris, R. S. (1984). Corporate behavior in adjusting to capital structure and dividend targets: An econometric study. *The journal of Finance*, *39*(1), 127–145.

Jenter, D. (2005). Market timing and managerial portfolio decisions. *The Journal of Finance*, *60*(4), 1903–1949.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146–3154).

Kester, W. C. (1986). Capital and ownership structure: A comparison of united states and japanese manufacturing corporations. *Financial management*, 5–16.

Kraus, A., & Litzenberger, R. H. (1973). A state-preference model of optimal financial leverage. *The journal of finance*, *28*(4), 911–922.

Lahmiri, S., & Bekiros, S. (2019). Can machine learning approaches predict corporate bankruptcy? evidence from a qualitative experimental design. *Quantitative Finance*, *19*(9), 1569–1577.

Leary, M. T., & Roberts, M. R. (2005). Do firms rebalance their capital structures? *The journal of finance*, *60*(6), 2575–2619.

Long, M. S., & Malitz, I. B. (1985). Investment patterns and financial leverage. In *Corporate capital structures in the united states* (pp. 325–352). University of Chicago Press.

Lopez de Prado, M. (2019). Beyond econometrics: A roadmap towards financial machine learning. *Available at SSRN 3365282*.

Loughran, T., & Ritter, J. R. (1995). The new issues puzzle. *The Journal of finance*, *50*(1), 23–51.

Loughran, T., & Ritter, J. R. (1997). The operating performance of firms conducting seasoned equity offerings. *The journal of finance*, *52*(5), 1823–1850.

Lowry, M. (2003). Why does ipo volume fluctuate so much? *Journal of Financial economics*, *67*(1), 3–40.

Lowry, M., & Schwert, G. W. (2002). Ipo market cycles: Bubbles or sequential learning? *The Journal of Finance*, *57*(3), 1171–1200.

Lucas, D. J., & McDonald, R. L. (1990). Equity issues and stock price dynamics. *The journal of finance*, *45*(4), 1019–1043.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 56–67.

Lundberg, S., Erion, G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

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

MacKie-Mason, J. K. (1990). Do taxes affect corporate financing decisions? *The journal of finance*, *45*(5), 1471–1493.

Mahajan, A., & Tartaroglu, S. (2008). Equity market timing and capital structure: International evidence. *Journal of Banking & Finance*, *32*(5), 754–766.

Marsh, P. (1982). The choice between equity and debt: An empirical study. *The Journal of finance*, *37*(1), 121–144.

Messmer, M. (2017). Deep learning and the cross-section of expected returns. *Available at SSRN 3081555*.

Mikkelson, W. H., Partch, M. M., Shah, K., et al. (1997). Ownership and operating performance of companies that go public. *Journal of financial economics*, *44*(3), 281–308.

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

Modigliani, F., & Miller, M. H. (1963). Corporate income taxes and the cost of capital: a correction. *The American economic review*, *53*(3), 433–443.

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

Myers, S. C. (1984). *Capital structure puzzle* (Tech. Rep.). National Bureau of Economic Research.

Myers, S. C., & Majluf, N. S. (1984). *Corporate financing and investment decisions when firms have information that investors do not have* (Tech. Rep.).

National Bureau of Economic Research.

Opler, T. C., & Titman, S. (1994). The debt-equity choice: An analysis of issuing firms. *Available at SSRN 5909*.

Pinegar, J. M., & Wilbricht, L. (1989). What managers think of capital structure theory: a survey. *Financial Management*, 82–91.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Rajan, R. G., & Zingales, L. (1995). What do we know about capital structure? some evidence from international data. *The journal of Finance*, *50*(5), 1421–1460.

Rossi, A. G. (2018). *Predicting stock market returns with machine learning* (Tech. Rep.). Working paper.

Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, *5*(1), 1–36.

Schwartz, E., & Aronson, J. R. (1967). Some surrogate evidence in support of the concept of optimal financial structure. *The Journal of Finance*, *22*(1), 10–18.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Shyam-Sunder, L., & Myers, S. C. (1999). Testing static tradeoff against pecking order models of capital structure. *Journal of financial economics*, *51*(2),

219–244.

Simon, H. A. (1990). Bounded rationality. In *Utility and probability* (pp. 15–18). Springer.

Smith, C. W., & Watts, R. L. (1992). The investment opportunity set and corporate financing, dividend, and compensation policies. *Journal of financial Economics*, *32*(3), 263–292.

Strebulaev, I. A. (2007). Do tests of capital structure theory mean what they say? *The journal of finance*, *62*(4), 1747–1787.

Taggart, R. A. (1977). A model of corporate financing decisions. *The Journal of Finance*, *32*(5), 1467–1484.

Titman, S., & Wessels, R. (1988). The determinants of capital structure choice. *The Journal of finance*, *43*(1), 1–19.

Virolainen, M., et al. (2009). Macro and micro determinants of seasoned equity offerings and issuer stock market performance.

Wang, N., et al. (2017). Bankruptcy prediction using machine learning. *Journal of Mathematical Finance*, *7*(04), 908.

Welch, I. (2004). Capital structure and stock returns. *Journal of political economy*, *112*(1), 106–131.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using extreme learning machine and financial expertise. *Neurocomputing*, *128*, 296–302.

Zhao, Y., Lee, C.-F., & Yu, M.-T. (2020). Does equity market timing have a persistent impact on capital structure? evidence from china. *The British Accounting Review*, *52*(1), 100838.

Appendices

A | Features

Table A.1: Compustat quarterly - accounting variables (72)

Variable	Description
actq	Current Assets - Total
ancq	Non-Current Assets - Total
apq*	Account Payable/Creditors - Trade
aqpq*	Acquisition/Merger Pretax
atq	Assets - Total.
capxy*	Capital Expenditures
ceqq*	Common/Ordinary Equity - Total
cheq *	Cash and Short-Term Investments
chq*	Cash
cogsq*	Cost of Goods Sold
cshiq*	Common Shares Issued
dd1q*	Long-Term Debt Due in One Year
dlcchy*	Changes in Current Debt
dlcq*	Debt in Current Liabilities
dltisy*	Long-Term Debt - Issuance
dltry*	Long-Term Debt - Reduction
dlttq*	Long-Term Debt - Total
dpq*	Depreciation and Amortization - Total

Table A.1: Compustat quarterly - accounting variables (72)

Variable	Description
dpy*	Depreciation and Amortization - Total
drltq*	Deferred Revenue - Long-term
dvpq*	Dividends - Preferred/Preference
dvpspq*	Dividends per Share - Pay Date - Quarter
dvy*	Cash Dividends
epspiq*	Earnings Per Share (Basic) - Including Extraordinary Items
exchg	Stock Exchange Code
fincfy*	Financing Activities - Net Cash Flow
gdwlamq*	Amortization of Goodwill
gdwlq*	Goodwill (net)
ibcomq*	Income Before Extraordinary Items - Available for Common
ibcy*	Income Before Extraordinary Items - Statement of Cash Flows
icaptq*	Invested Capital - Total - Quarterly
intanq*	Intangible Assets - Total
invtq*	Inventories - Total
ivltq*	Total Long-term Investments
ivstq*	Short-Term Investments- Total
lctq*	Current Liabilities - Total
lseq*	Liabilities and Stockholders Equity - Total
ltq*	Liabilities - Total
mkvaltq	Market Value - Total

Table A.1: Compustat quarterly - accounting variables (72)

Variable	Description
nimq	Net Interest Margin
niq*	Net Income (Loss)
niy*	Net Income (Loss)
oancfy*	Operating Activities - Net Cash Flow
oiadpq*	Operating Income After Depreciation - Quarterly
oibdpy*	Operating Income Before Depreciation
optvolq*	Volatility - Assumption (%)
piq*	Pretax Income
ppegtq*	Property, Plant and Equipment - Total (Gross) - Quarterly
ppentq*	Property Plant and Equipment - Total (Net)
pstkq*	Preferred/Preference Stock (Capital) - Total
rdipq	In Process R&D
rectq*	Receivables - Total
req*	Retained Earnings
revtq*	Revenue - Total
revty*	Revenue - Total
scstkcy*	Sale of Common Stock (Cash Flow)
sstky*	Sale of Common and Preferred Stock
teqq*	Stockholders Equity - Total
tstkq*	Treasury Stock - Total (All Capital)
txdbcaq*	Current Deferred Tax Asset
txdbclq*	Current Deferred Tax Liability

Table A.1: Compustat quarterly - accounting variables (72)

Variable	Description
txpdy*	Income Taxes Paid
txpq*	Income Taxes Payable
txtq*	Income Taxes - Total
wcapchy*	Working Capital Changes - Total
wcapq*	Working Capital (Balance Sheet)
xaccq*	Accrued Expenses
xidoy*	Extraordinary Items and Discontinued Operations
xoprq*	Operating Expense- Total
xrdy*	Research and Development Expense
xrdq*	Research and Development Expense - Quarterly
xsgay*	Selling, General and Administrative Expenses
xsgaq*	Selling, General and Administrative Expenses - Quarterly

* Feature has been scaled by total assets, *atq*.

Table A.2: WRDS ratio suite variables (30)

Variable	Description
capital_ratio	Capitalization Ratio
cash_debt	Cash Flow/Total Debt
cash_lt	Cash Balance/Total Liabilities
cash_ratio	Cash Ratio
cfm	Cash Flow Margin

curr_debt	Current Liabilities/Total Liabilities
curr_ratio	Current Ratio
de_ratio	Total Debt/Equity
debt_assets	Total Debt/Total Assets (Total debt as a fraction of total assets)
debt_ebitda	Total Debt/EBITDA
debt_invcap	Long-term Debt/Invested Capital
efftax	Effective Tax Rate
equity_invcap	Common Equity/Invested Capital
fcf_ocf	Free Cash Flow/Operating Cash Flow
gpm	Gross Profit Margin
int_debt	Interest/Average Long-term Debt
int_totdebt	Interest/Average Total Debt
intcov_ratio	After-tax Interest Coverage
invt_act	Inventory/Current Assets
lt_debt	Long-term Debt/Total Liabilities
lt_ppent	Total Liabilities/Total Tangible Assets
npm	Net Profit Margin
opmbd	Operating Profit Margin Before Depreciation
ptpm	Pre-tax Profit Margin
rect_act	Receivables/Current Assets
roa	Return on Assets
roce	Return on Capital Employed
roe	Return on Equity
short_debt	Short-Term Debt/Total Debt

totdebt_invcap	Total Debt/Invested Capital

Table A.3: Additional variables (12)

Variable	Description
past_seo	Whether the firm has previously issued seasoned equity (dummy)
num_seo	Market-wide number of SEOs
num_ipo	Market-wide number of IPOs
fin_def	Financing deficit (Shyam-Sunder & Myers, 1999)
alt_z	Altman's Z-Score (Altman, 1968)
nd_ts	Non-debt tax shied = depreciation / total assets
tobinq	Tobin's Q = (MV + liabilities)/(BV + liabilities)
asset_tang	Asset tangibility = PP&E / total assets (Rajan & Zingales, 1995)
trt1m	Monthly total return
trt1y	Yearly total return
sent	Investor sentiment (Baker & Wurgler, 2006)
SIC	SIC Industry Classification Code
dpy_r	Rolling four quarters Depreciation and Amortization - Total
niy_r	Rolling four quarters Net Income (Loss)
oibdpy_r	Rolling four quarters Operating Income After Depreciation
revty_r	Rolling four quarters Revenue - Total
xrdy_r	Rolling four quarters Research and Development Expense
xsgay_r	Rolling four quarters Selling, General and Administrative Expenses

B | Hyperparameter descriptions

B.1 Gradient Boosting methods

LightGBM can run different gradient boosting methods, and the parameter choice sets the method that the algorithm will use.

'boosting_type': The method chosen for this hyperparameter defines the model's algorithm for boosting/training the model. There are three different options, and the optimal choice depends on the problem one is trying to solve. I have selected the 'gbdt' (Gradient Boosted Decision Trees) method, which uses decision trees and stochastic gradient descent to train the model. This method is the most stable and reliable of the options but is time- and memory-consuming (Kaczmarek, 2020). However, the method's advantages outweigh the disadvantages for my research. Gradient boosted trees were described in Section 4.4.3.

B.2 Regularisation

In general, regularisation in ML is the process by which the coefficients of the model are regularised or shrunk towards zero. It prevents overfitting by discouraging model complexity, allowing the model to generalise better to unseen data. LightGBM has a set of hyperparameters related to regularisation, discussed below.

'lambda_l1 and lambda_l2': These parameters are used to prevent overfitting by adjusting the L1 (Lasso Regression) and L2 (Ridge Regression) regularisation

settings. Kaczmarek (2020) suggests using tuning methods to set this value; thus, I have used GridSearchCV to find the optimal value. The precise details of the methods are omitted but can be found in the LightGBM documentation.

'max_depth': Controls the maximum depth of each trained tree. A large value will likely result in overfitting the training data, resulting in the model generalising poorly to new data; however, increasing the value may improve performance. I have set max_depth to '2' after experimenting on the validation set.

'num_leaves': The maximum number of leaves for each trained tree. Large num_leaves increase the model's performance on the training set but also increases the likelihood of overfitting. According to the LightGBM documentation, a simple way of selecting the number of leaves is num_leaves = $2\hat{}$(max_depth). Regardless, it is necessary to tune num_leaves and max_depth together. I have optimised this parameter using GridSearchCV.

'max_bin': Specifies the maximum number of values for each feature - think of a traditional histogram and the bins for each bar. LightGBM uses a histogram-based algorithm to identify optimal split points when creating the weak learners. Thus, continuous features are split into discrete bins, on which this parameter imposes a maximum. Small values for max_bin increase the training speed of the algorithm, but may decrease performance. I have opted for a value of 512 after experimenting with the validation set.

'subsample': Specifies the percentage of rows used per iteration to build each tree. Then, a number of rows in accordance with the specified percentage are randomly chosen to fit each tree. Using a larger value may improve the model's

generalisation ability but may slow the speed of training. This parameter was optimised using GridSearchCV.

'colsample_bytree': Sets the percentage of the subset of features used to train each tree in each iteration. For example, if the hyperparameter is set to 0.50, LightGBM will use 50% of features for training each tree. There are similar trade-offs to the subsample parameter, and I use GridSearchCV to find the optimal value for this parameter.

B.3 Training parameters

'objective': Is the objective of the ML task, for example, regression, binary classification, or multiclass classification? Since I aim to predict whether a firm will engage in an SEO or not, the objective is binary classification.

'learning_rate': Is the multiplication performed on each boosting iteration. Specifically, each iteration in the GBM algorithm aims to improve the training loss. This improvement is multiplied by the learning rate in order to perform smaller updates. A lower learning rate will result in more iterations being required for training. The documentation advises that this parameter should not be tuned using an optimiser like GridSearchCV. I have followed this guidance. I chose to set the learning_rate to 0.005.

'is_unbalance': This is a crucial hyperparameter for my research and addresses the class imbalance in the dataset. It works by setting the weights of the dominant label to 1, and then using the ratio of the dominant/non-dominated observations to set the weight of the non-dominant label. In simple terms, it weights the non-

dominant class higher when training the model. Thus, the model is encouraged during training to over-correct the errors made on the positive class. This parameter only affects the training of the model and does not impact the calculation of performance evaluation metrics. This parameter is set to 'True' as my dataset is heavily imbalanced.

'metric': Specifies the metric used during the training of the model, which the model seeks to optimise. For my model, 'auc' is used, as the area under the curve is the performance measure I use to assess my model and aim to maximise.

'num_iterations': Sets the number of boosting iterations, i.e. the number of trees to build (see Methodology for details). The more trees the algorithm builds, the more accurate the model is likely to be, but this may be at the expense of increased training time and overfitting. I set num iterations to 1,000 in my model.

'early_stopping_rounds': Sets the number of rounds after which training will stop if the validation metric is not improving. Kaczmarek (2020) advise setting this parameter in conjunction with num_iterations, with a rule of thumb being to set it was 10% of the num_iterations value. If the value is set too high, it increases the chance of overfitting the training data. Given that I had set num_iterations to 1000, I set early_stopping_rounds to 100.

C | Feature importance

By default, LightGBM uses 'split' to determine feature importance. Split produces a result which contains the number of times a feature is used in the model, where the more times a feature is used, the more important it is. Using this default setting, Figure C.1 shows the most important features for SEO prediction. *trt1m*, *pas_seo*, *sstky*, *req*, and *alt_z* stand out as the five most important predictors of whether a firm will conduct an SEO.

There are other parameter settings than 'split', but these can provide conflicting results. Figures C.1 and C.2 compare the feature importance results using the 'split' and 'gain' measures, highlighting their differences. When using different importance measures, the inconsistency makes it unclear which features are the most important for SEO prediction. This lack of clarity arises due to the complexity of ML models such as LightGBM, which allow for multi-way interactions among features and non-linear relationships. In traditional finance research, this is not a problem as OLS models make it easy to determine which variables are the most important by analysing the model coefficients.

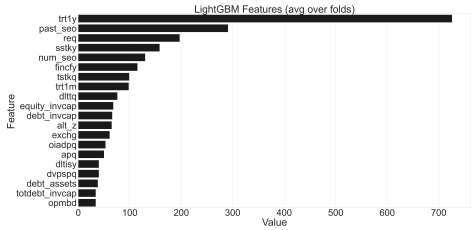


Figure C.1: Feature importance using the 'split' measure

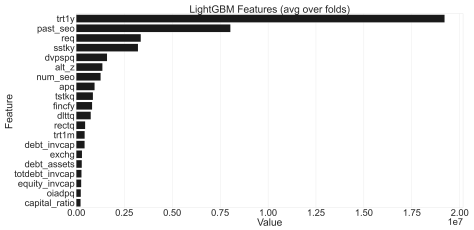


Figure C.2: Feature importance using the 'gain' measure