Investigating Underpricing in Venture-Backed IPOs Using Statistical Techniques

Michael Tan

April 2020

Abstract

This paper concerns applying statistical methods to investigate underpricing in VC-backed technology Initial Public Offerings (IPOs) since the great recession. In particular, firm, market, and IPO-specific variables were explored to determine if there were any significant relationships to underpricing. The paper focused on the Bank Preference theory of underpricing, where underpricing is said to occur because investment banks running IPO processes are incentivized to underprice to decrease the risk that they will not be able to allocate all the issuance to price-sensitive public markets investors.

Acknowledgements

First and foremost, I would like to thank my advisors, Professor Shawn Santo from the Statistical Science department and Professor Daniel Xu from the Economics department. This paper was guided in large part through my weekly meetings with them, and they have exposed me to new ideas and techniques which have been applied in this paper. I'd like to thank them for their continuous and kind support.

1 Introduction

This study aimed to investigate underpricing in venture-backed tech Initial Public Offerings (IPOs) in the US after the 2008 financial crisis. The precise period investigated is 2009-2020. This paper sought to make the following unique contributions to this field of study: first, demonstrating the usefulness of using random forest techniques to analyze IPO data, in contrast to most of the historical literature which relied on classical techniques such as linear regression; second, while most of the historical literature has focused on IPOs of a certain time period or in a certain geographic region, this paper seeks to analyze a unique set of IPOs with differentiating characteristics that there is limited literature about: venture-backed tech IPOs; and third, to investigate if particular investment banks being associated with the IPO company had any effect on underpricing – much of the historical literature has focused on market or firm factors instead of placing more of a focus on the investment banks.

Companies tend to IPO for a variety of reasons. A startup may seek to tap the public markets for financing, especially if the private markets become a less accessible source of capital for any reason. IPOs also provide an exit opportunity for early-stage investors and employees, allowing them to cash out on any equity they may own in the company. For instance, in the failed attempt by WeWork to IPO in 2019, the company sought to raise more money through the public markets after it had already raised billions from the private markets and VC-investors such as Softbank were looking to exit their investment in WeWork for a profit.

The three key players in an IPO are (1) the company going public, (2) the investment bank(s) running the IPO process, and (3) the investors in the IPO. Each have their own roles and incentives. The company that is going public is seeking to raise as much money as possible through selling their stock in the public market. This means that they would theoretically prefer the IPO price to be as high as possible, but not so high that no investors would buy at that price. Meanwhile, the investment banks make fees off of the total size of the IPO and are in charge of finding institutional buyers for the IPO stock. The banks

seem to have incentives aligned with the issuing company, in that the investment banks are also incentivized to maximize price so that the fees will be larger), but also need to be cognizant of investor demand since it is inversely related to price. The investors are seeking to maximize the potential return they may get from investing in an IPO. A lower IPO price initially could offer more return upside in the future (in either the short-term or long-term), and so their demand is inversely related to price.

In these IPOs, although the theoretical goal of the issuance is to establish a fair price reflecting the true value of the company, there do exist incentives to underprice or overprice the issuance. Overpricing naturally means that more money will be raised upfront through the IPO (because it will be raised at a higher valuation). For example, a possible incentive for underpricing is that it allows for the possibility of a post-IPO "pop" in the share price, which attracts great PR and increased investor interest for the future.

The current venture-backed startup landscape is awash with activity. A venture-backed startup is simply a startup that has raised capital from venture capital firms. There has been much media coverage of large Initial Public Offerings, much talk of overvaluation, and the raising of many multibillion dollar venture funds dedicated to catalyzing and sustaining the growth of startups towards a profitable exit, such as an IPO. In 2018, there were 134 IPOs, 66% of which were VC-backed and 38 of which were tech IPOs. In 2018, tech IPOs generated almost $12 billion in proceeds (proceeds are the dollars received by the issuer from the issuance and sale of its common equity), but only 16% of those companies were profitable at time of IPO. The median market price to sales ratio for 2018 IPOs was at its highest mark since 2002 at 11.3 (Ritter 2018). These macro trends have been in play for around a decade, ever since the world recovered from the financial crisis of 2008.

Take, for example, the IPO of Zoom Video Communications (Ticker: $ZM), the popular video conferencing software company. On April 18, 2019, Zoom IPO'ed at a price of $36 per share, raising $356.8 million for the company by selling 9.91 million shares in the IPO. However, the shares closed up 72% at $62 by the end of the first trading day. This indicates a very large underpricing effect of 72%. It seems that Zoom could have IPO'ed at a much

higher price than just $36 a share, as by the end of the day, investors were willing to buy the stock at a price as high as $62 a share. If the IPO had been priced higher initially (i.e less underpriced), it would have raised more money off of selling those 9.91 million shares.

Much has been made of the phenomenon of underpricing, in which issuers can be seen to be "leaving money on the table" – with an underpriced IPO, a company is not raising as much money as it could be. From a theoretical finance perspective, an IPO is considered underpriced when there is an increase in stock value from the initial offering price to the first-day closing price. This one-day increase is often called an IPO "pop." In the example of Zoom, the pop was 72%. An IPO "pop" indicates that investors were actually willing to pay a higher price than the official IPO offer price. If the IPO had originally been offered at that higher price, then the startup would have raised more capital. Ritter (2018) notes that IPOs of US companies were underpriced by an average of 18 percent from 1980-2017, and at the height of the 2000 dot-com bubble, underpricing rose to extreme levels, and the average IPO was underpriced by over 70%.

Venture capitalists who back companies looking to IPO have expressed discontent over this phenomenon. Notably, Bill Gurley of Benchmark accuses the investment banks who run the IPO processes of consistently underpricing IPOs, leading to a cumulative $170bn left on the table for issuers. The VC's who back these companies would also make less money from selling the shares that they own in the issuing company if the issue is underpriced.

Investment banks play a key role in the IPO process – they have responsibilities in making sure that the issuer meets all the regulatory and logistical requirements for an IPO, and arguably more importantly, setting an IPO price and then allocating all the shares being offered at that price. They also collect fees (usually around 7% of the total IPO size, and this is known as the spread) for IPO's that they underwrite. The size of spread depends both on negotiations between the underwriters and the issuers as well as amount of risk the underwriters take on (i.e. number of shares they decide to allocate). Generally, this

spread is broken down 20%/20%/60% between management fee, underwriting fee, and selling concession (Torstila 2001). These fees usually come in the form of the bank being able to purchase the issuer's shares at a small discount to the price that will be offered to the public (Public Offering Price, or POP). This margin between the price that the bank can buy at and that the public can buy at is effectively the fee that the banks get paid. The size of the IPO is one of the primary determinants for the amount of fees that the banks. This paper sought to explore IPO underpricing with a focus on the role of these investment banks by investigating if any market or firm-specific factors were significant predictors of underpricing.

The remaining sections of this paper are organized as follows. Section 2 details the preceding the literature on the subject of underpricing in IPOs. Section 3 introduces a mathematical framework between the three players in the IPO: the issuing company, the investment bank, and the investors. Section 4 goes over the dataset, data cleaning, and exploratory data analysis. Section 5 describes the methods used in this analysis. Section 6 provides an exploration of the results of the analysis. Section 7 makes conclusions about the study.

2 Literature Review

There have been many theories put forward that seek to explain the phenomenon of IPO underpricing. Below I describe a few of these theories:

2.1 Explanations of Underpricing

2.1.1 Winner's Curse

Rock (1986) proposes that, in a world with two types of investors, informed and uninformed, IPO's are underpriced so that uninformed investors become willing to take the risk to buy the stock at IPO. Seeing the adverse selection problem, uninformed investors will rationally assume that if they are able to purchase shares at IPO, it is only because the

informed investors passed up on those shares. Thus, underpricing is used to compensate the uninformed investors for this.

2.1.2 Marketing

Issuers may see an IPO "pop" resulting from an underpriced IPO as a good way to generate buzz and publicity from investors following the IPO. This could lead to more investor interest and demand for the stock. Thus, the IPO is used primarily as a marketing event rather than a capital raising event.

2.1.3 Prospect Theory

Loughan and Ritter (2002) posit that issuers may prioritize the psychological benefit they gain from seeing a dramatic increase in their wealth post-IPO resulting from an IPO "pop." This draws from Kahneman and Tversky's (1979) idea of prospect theory that individuals care more about relative changes in their wealth rather than the absolute level of the wealth.

2.1.4 Market Cyclicality

Underpricing happens as a result of hot markets where investor demand is irrationally high, leading to an IPO "pop." Ritter (1984) shows evidence that underpricing during a hotter time in the markets in the 1980s was three times more server than during a colder time in the markets in the 1980s.

2.1.5 Bank Preferences

Underpricing may be driven by incentives experienced by the investment banks running the IPO processes. Because these banks are tasked with allocating shares for the IPO and they make money based on what they allocate, banks may be incentivized to underprice the IPO so that it is easier for them to sell/allocate more shares to public markets investors. Moreover, this implies that the health of the company at the time of IPO could influence the

degree of underpricing – a healthier company would not require as much underpricing for the banks to be able to ensure they meet their allocation targets, as the shares of a healthy company would already be in high demand by public markets investors. Conversely, an unhealthy company would require more underpricing to be able to ensure that allocation targets are met – public markets investors would need to be better incentivized by lower prices to buy the shares of an unhealthy company. This effect would also be impacted by the overall macroeconomic and market conditions of the time of IPO – a hotter market may make it easier for the IPO price to be set higher, as demand from investors is typically higher in a hot market, as opposed to a cold market. Additionally, overpricing may lead to poor stock performance post-IPO, which in turn can tarnish the reputation of the bank.

2.2 VC-Backed Companies

Literature also suggests that it would be appropriate to treat VC-backed IPOs as a separate category with distinct characteristics. Megginson and Weiss (1991) explain that venture capital backing serves to certify the quality of an IPO, allowing for the acquisition of higher-quality underwriters and greater institutional backing. Jain and Kini (2000) build upon this, saying, "further, VC-backed IPO firms allocate significantly higher resources to RD expenditures, attract prestigious investment bankers, achieve greater success in the road shows, and attract stronger analyst following compared to their non-VC-backed counterparts."

3 Mathematical Model of Agent Incentives

Exploring further the Bank Preference theory of underpricing, it would be instructive to examine the incentives that banks, issuers, and public markets investors experience.

Beginning with the banks, it can be said that they are seeking to maximize their fee proceeds subject to the constraint of being able to sell all of the allocation. The underlying factor in both of these is price. A higher price means that the total IPO amount on which

the spread is applied is higher, but it also increases the risk of not being able to sell all the issuance. Thus, this is an optimization problem where banks seek to maximize the total value of the IPO issuance but are constrained by the demand for that issuance on the part of the public investors. Meanwhile, the issuing company has the same incentives as the bank. They are also seeking to raise the most money possible – maximizing the value of the IPO while being constrained by public investor demand. On the public investor side, investors in any given IPO are betting that the value of the stock will appreciate in the future, leading to a financial return. This can all be written mathematically as:
Bank's optimization function:

$$\text{Maximize: } P_{ipo} * S$$
$$\text{Subject to: } S = D_{inv}$$

Issuer's optimization function:

$$\text{Maximize: } P_{ipo} * S$$
$$\text{Subject to: } S = D_{inv}$$

Investor's optimization function:

$$\text{Maximize: } (P_{future} - P_{ipo}) * S$$

Where:

- P_{ipo} is the price of IPO
- S is the total shares sold by the issuer / bought by the public investors
- D_{inv} is the quantity of shares demanded by the investor
- P_{future} is the price of the issued stock some time in the future

4 Data

4.1 Description of Dataset

The data set used in this study consists of first-day trading returns of 113 US public offerings between March 2009 and February 2020. All of these offerings involved venture capital-backed companies and all were involved in a technology space. The time period was chosen to eliminate the massive distortionary effect that the Great Recession of 2008 would have had on the dataset. Initial data was filtered and taken through the Thomson Reuters SDC Platinum New Issues database, which also flagged entries for whether or not they were venture-backed. After some data cleaning and pulling more important data from outside sources, a full dataset was created, with each row representing one new issue and the columns showing information such as issuer name, date of issue, financial details about the issuer before the IPO, the investment bank that was the lead underwriter for the IPO, the one-day return of the IPO, and more. It can be said that the dataset has 3 different segments of data: 1) firm-specific financial data (such as revenues and net income), 2) market-specific data (such as market performance and VIX prices around the time of the IPO), and 3) IPO-specific data (such as which investment bank lead the IPO and how much in proceeds was raised in the IPO).

A full list and description of variables that were ultimately used in creating the regression and random forest models are listed in Figure 1 on the next page.

Name of Variable	Type	Data Segment	Description
LogOneDayChange	numeric	IPO-specific	Log of the quotient of the closing price and the offer price, used to measure underpricing
market_performance	numeric	Market-specific	30-day performance of S&P 500 prior to IPO date
VIX_before	numeric	Market-specific	VIX price on IPO date
Total.Revenues.Last.Reported	numeric	Firm financials	Latest revenues reported in S-1 filing
Total.Operating.Expenses	numeric	Firm financials	Latest operating expenses reported in S-1 filing
Selling.and.Marketing	numeric	Firm financials	Latest selling and marketing expenses reported in S-1 filing
Net.Income	numeric	Firm financials	Latest net income reported in S-1 filing
Net.Income.Margin	numeric	Firm financials	Net income divided by revenue
RevGrowth	numeric	Firm financials	Year-on-year revenue growth from last reported revenues in S-1 relative to the year right before
Long.Term.Debt	numeric	Firm financials	Latest long-term debt reported in S-1 relative to the year right before
Total.Assets	numeric	Firm financials	Latest total assets reported in S-1 filing
Long.Term.Debt.to. Total.Asset.Ratio	numeric	Firm financials	Long term debt divided by total assets
Type.of.Security	binary	IPO-specific	Common Stock or Class A shares
Investment.Bank..Lead.Left.	factor	IPO-specific	The lead underwriting bank for the IPO
Additional.Lead.Underwriters	numeric	IPO-specific	Number of additional major underwriters for the IPO
Proceeds	numeric	IPO-specific	Proceeds raised in the IPO
Total.VC.Funding.Raised	numeric	Firm financials	Total venture capital raised by issuing company prior to IPO

Figure 1: List of Variables

4.2 Data Cleaning

In order to create a full dataset to be used for analysis, it was necessary to find the financial statistics of each firm directly before their IPO and information about market factors around the time of the IPO so that that information can be used as predictors of underpricing in any model. Additionally, it was necessary to gather the first-day change between the closing price and offer price of the new issue, as this is the response variable (the underpricing effect). This first day change was then converted into LogOneDayChange, defined as the log of the quotient of the closing price and the offer price:

$$\log \frac{P_{closing}}{P_{offer}}$$

The larger this value is, the greater the underpricing effect. If the value is negative, then this indicates that the offer price was actually lower than the closing price, indicating overpricing.

A logarithm was applied to the first day change because this is an often-used metric in the financial world and because it makes the data fit better for linear regression purposes, which will be further discussed in the Exploratory Data Analysis Section.

While the Thomson Reuters SDC Platinum New Issues database provided information such as the Issue Date, Issuer, Business Description, Ticker Symbol, Proceeds, and the Offer Price of the IPO, finding firm-specific financials such as Last Reported Revenue Before the IPO, Total Assets Before the IPO, Net Income Before the IPO, and more required sifting through the S-1 SEC filing that any firm seeking to IPO in the United States must file.

Additionally, market and stock performance were attained by using the BatchGetSymbols library in R and using the issue date listed for each IPO (each row of data) as the parameter to retrieve the performance of the stock, the S&P 500, and the VIX (volatility index) around that issue date. More specifically, the one-day change of the stock (underpricing), the S&P 500 change over the past 30 days before the issue date, and the VIX price at the issue date were attained. The market performance and VIX were thought to have the potential to be important to the underpricing effect because investor demand is

dependent on how hot or cold or volatile the market is at any given time.

After these steps, some new variables were created by manipulating existing variables, such as dividing Net Income by Revenue to calculate Net Income Margin. Margins, growth, and other key financial ratios are important for ascertaining the financial health of any company, and so these new variables were made.

Below is a full summary of the data sources:

- **From Thomson Reuters**: Proceeds, Type of Security, Company Name, Ticker, Date of IPO, Filing Date, Business Description, Total VC Funding Raised, One Day Change in Stock Price Post-IPO
- **From S-1 filing**: Total Revenues, Selling and Marketing Costs, Total Operating Expenses, Net Income, Long Term Debt, Total Assets, Investment Bank that served as the Lead Left Underwriter, Additional Lead Underwriters
- **From BatchGetSymbols API**: Market Performance, VIX before IPO
- **Partially from News Reports**: One Day Change in Stock Price Post-IPO
- **Calculated from other variables**: Log One Day Change, Net Income Margin, Revenue Growth, Long Term Debt to Asset Ratio

There were 25 instances of missing data in the selling and marketing expenses column, and the mice package was used with the random forest method to impute values for these missing entries by using the existing data from the other, fully filled entries. This process involved using the non-missing selling and marketing data and regressing that on the other variables, and then using that regression model to predict values which are then imputed to the missing rows.

4.3 Exploratory Data Analysis

Beginning with the time period that the dataset covers (2009-February 2020), it is evident that this is a time period with no major shocks to the economy. There were no great

anomalies or events that would cause distortions to the time period or the dataset. All of the IPOs in the dataset have occurred in a relatively stable bull market, and there are no major macroeconomic distortions to control for. This can be seen in Figure 2 below, which shows the S&P 500 Performance from 2009-2020.

Figure 2: S&P 500 Performance 2009-2020

Next, by looking at the response variable, the one day change of the stock price immediately after IPO, we saw that the distribution of this variable was skewed to the right and was not ideal for linear regression. The histogram of the One Day Returns is shown below in Figure 3. There were many points far to the right of the histogram (many IPOs that experienced very large first-day IPO pops).

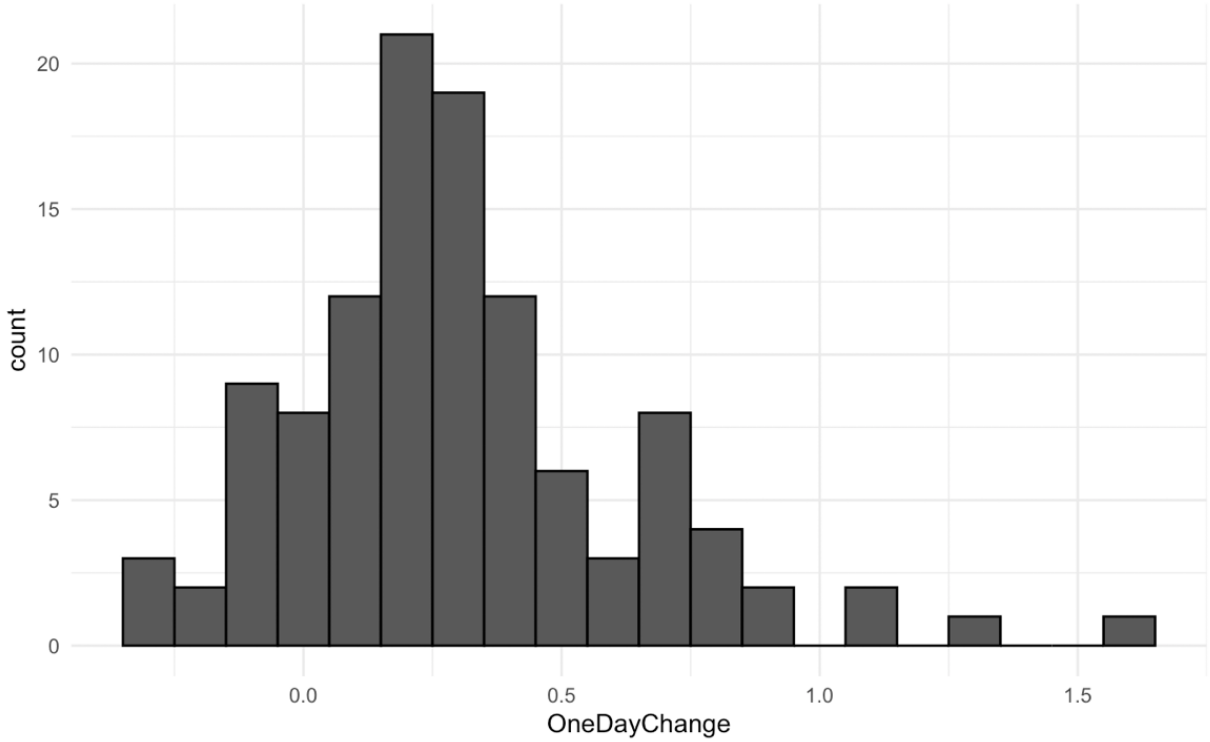

Figure 3: Histogram of One Day Returns

To fix this, we can instead use the log return, defined mathematically as:

$$\log \frac{P_{closing}}{P_{offer}}$$

This log return metric is used often in financial for many reasons – amongst them is the reason of having better normality (other reasons include mathematical ease and greater utility when working with time series, which almost all stock price charts use).

Indeed, as can be seen in the Figure 4 Below, the log of the one-day change of the stock price on the IPO date looked to be distributed approximately normally, and so was an ideal response variable to use for linear regression analysis. The log of the one-day change was centered around a mean of 0.2371, which translates to a one-day percentage change of +26.76% (the median was +26.00%). The average IPO in this dataset of VC-backed tech IPOS since 2009 was overpriced by 26.76%. A Jarque-Bera test, wish tests for skewness and kurtosis relative to a normal distribution, resulted in a p-value of 0.5557, thus giving

more evidence of normality (the p-value was much to high to be able to reject the null hypothesis that the distribution was normal).

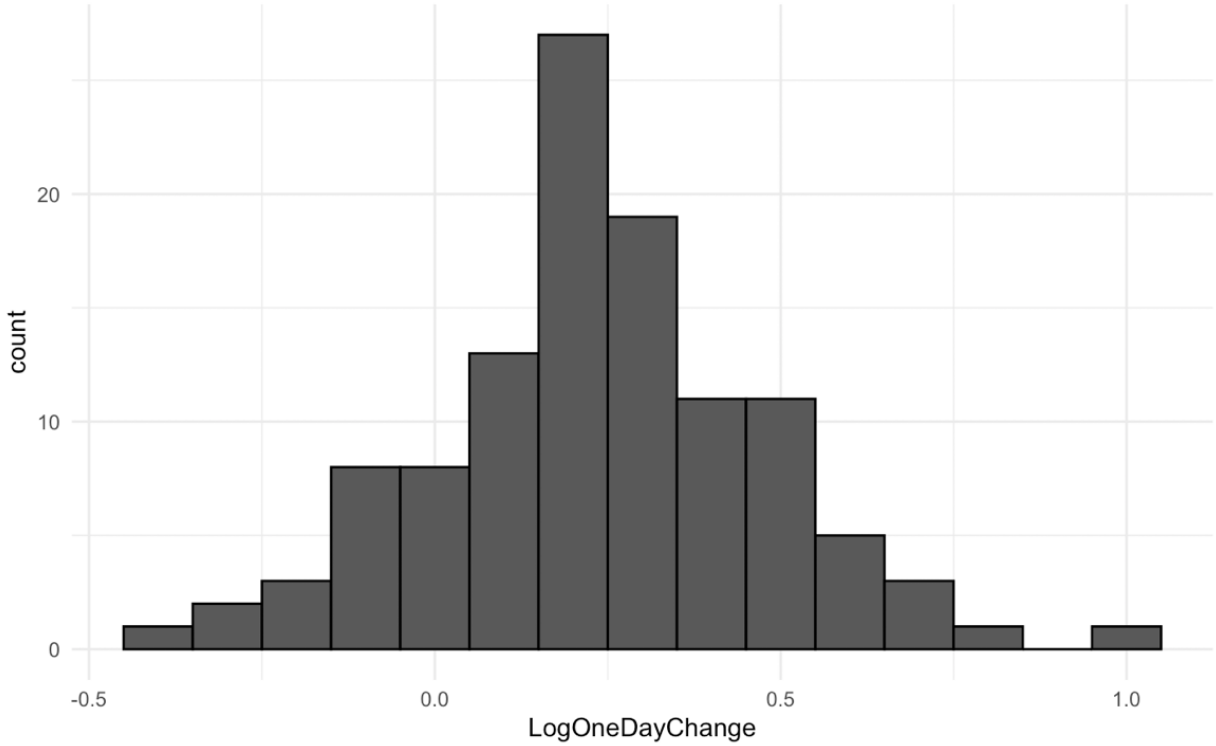

Figure 4: Histogram of Log One Day Returns

Now with the response variable set, the next step was to preliminarily examine some of the potential predictor variables.

To further explore some of the relationships between the response variable of log return and the other predictor variables in the hopes that this would guide the regression analysis, we examined scatterplots like the following, which plotted the response variable on the y-axis and predictor variables on the x-axis (Figure 5 below):

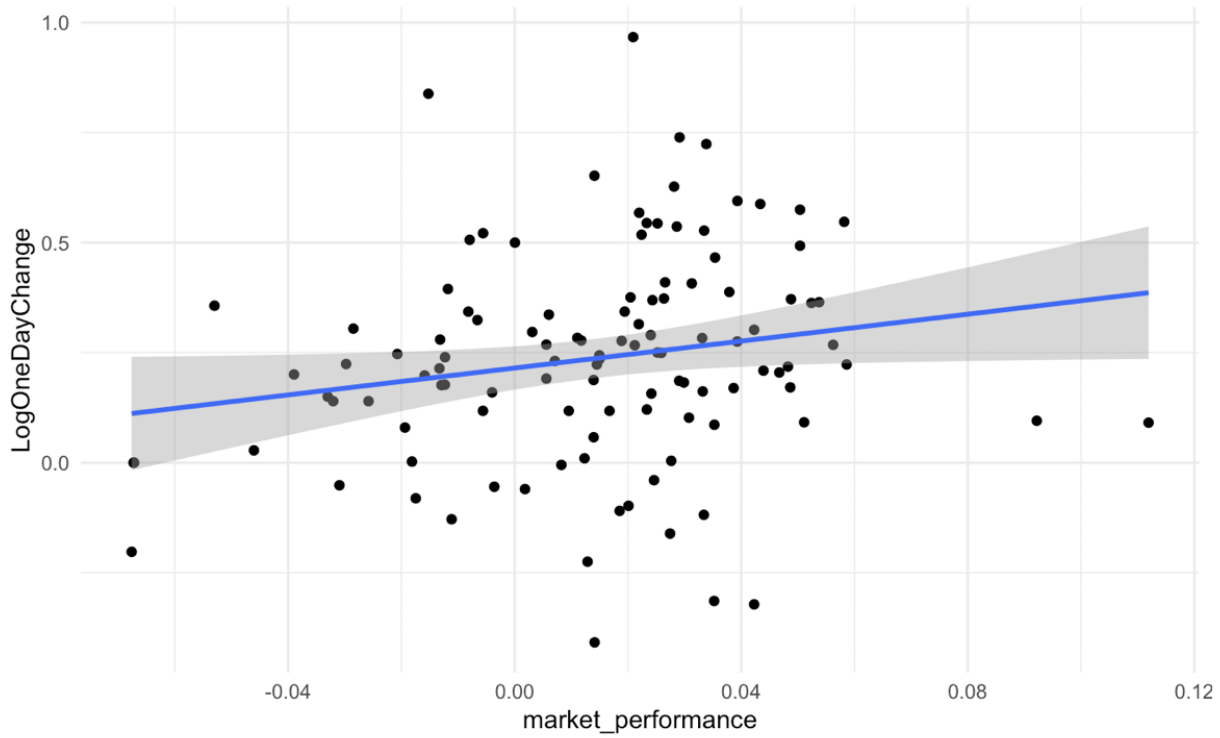

Figure 5: Log Returns vs Market Performance

The scatterplots showed that there were indeed trends to be analyzed between the log return and other variables such as the market performance before IPO. This can be seen in the pronounced trendlines of nonzero slope across the scratterplots (the figure shown is just one example). Given these preliminary findings, the analysis was continued.

5 Methods

We used the following methods to explore underpricing, and this exploration was split into two analyses: regression and classification. Using multiple regression methods such as linear regression and random forest was important so that we could triangulate between results to better pinpoint what factors are indeed important to underpricing. For classification, logistic regression with three different cutoffs on log returns to define whether something is underpriced was done so that we could see if there exists some kind of dis-

cernable demarcation that separates IPO pops that are clearly too high from the rest.

5.1 Linear Regression

Linear regression was used to investigate the relationship between certain firm-specific and market-specific variables and the underpricing effect, as measured by the log return. Linear regression is a relatively simple method that allows for high interpretability.

5.1.1 Stepwise Regression

Stepwise regression uses either forward, backward, or bi-directional selection to determine what variables to use in a linear regression model.

- Forward selection starts with 0 variables in the model and keeps adding new variables that are statistically significant according to some criterion such as AIC, BIC, or R^2. It does this until there are no further statistically significant variables to be added. In this case, the metric used was AIC.
- Backward selection starts with a full model and proceeds by removing the least statistically significant variable. It does this until no more removals can occur without losing goodness of fit.
- Bidirectional selection combines forward and backwards selection and tests at every step whether variables should be added or removed.

5.2 Random Forest

Random forest regression was used to predict the degree of underpricing (Log One Day Returns) given the different predictor variables such as the firm-specific and market information detailed in the data section. The random forest algorithm constructs and merges multiple decision trees to create more accurate predictions. As Baba and Sevil (2019) state, “empirical analyses of IPO initial returns are heavily dependent on linear regression

models. However, these models can be inefficient due to its sensitivity to outliers which are common in IPO data... the machine learning method random forest is introduced to deal with the issues the linear regression cannot solve." The random forest regression method does lack the interpretability that a linear regression would have, but it nevertheless does show what the most influential predictor variables are. Random forest classification was also used to perform the task of predicting the binary result of whether an IPO would be underpriced or not.

5.3 Logistic Regression

Logistic regression is another method that was used for binary classification, in addition to the random forest classification. It was used to investigate the relationship between certain firm-specific and market-specific variables and the underpricing effect, this time having underpricing be binary. The logistic regression method allows for more interpretability than the random forest method does.

6 Results and Analysis

6.1 Regression

6.1.1 Stepwise Model

Beginning with the regression analysis (both linear and random forest), this paper investigated a few models that provided informative results.

To begin, stepwise regression was used as a starting point from which conclusions could be drawn.

The step-wise regression results show the following coefficient effects (Figure 6):

Coefficient	Estimate	Std. Error	t-value	P-value
Intercept	3.078e-01	4.802e-02	6.410	4.02e-09
market_performance	1.890	7.232e-01	2.614	0.0103
Total.Assets	-1.281e-04	6.543e-05	-1.958	0.0529
Long.Term.Debt.to.Total.Asset.Ratio	-2.351e-01	1.296e-01	-1.815	0.0723
Net.Income Margin	-1.038e-01	4.814e-02	-2.157	0.0333
Type.of.SecurityCommon Shares	-8.792e-02	4.783e-02	-1.838	0.0688

Figure 6: Stepwise Regression Output

These preliminary findings offer important insights: All of the variables listed in the table above are significant at a 0.1 significant level or lower. At a 0.05 significance level, market_performance and Net.Income.Margin are significant, and Total.Assets is very close to being significant with a p-value of 0.0529. Conveniently, the signs (positive/negative) of the coefficients make sense within the context of the data. A negative sign means smaller One Day Log Return, while a positive sign means greater One Day Log Return. Since the One Day Log Return is the measure of underpricing, a negative sign means less underpricing and a positive sign on the coefficient means more underpricing. There are a few interpretations of these coefficients that are significant at a 0.05 significance level:

- **Higher returns in the S&P 500** before the IPO (market_performance is higher) lead to a significant increase in the underpricing effect. More precisely, a one percent increase in the S&P 500 30 days before the IPO leads to a 1.019x multiplicative effect on the one-day percentage (not log) return of the issued stock. A possible explanation of this effect is that a hotter market leads to unexpectedly high levels of investor demand. Investors are encouraged in hot markets, possibly because of behavioral factors, to demand more stock. IPOs may experience greater demand as a result, and so the closing price will be higher than the offer price, leading to an underpricing effect.

- A company with a **greater net income margin** will experience a larger underpricing effect. Net income margin is a measure of profitability, defined as the net income divided by the topline revenue. More precisely, a ten percent increase in the net income margin before IPO leads to a 0.897x multiplicative effect on the one-day percentage return of the issued stock. A possible explanation of this effect is that a greater net income margin is indicative of a healthier company. With a healthier company, the investment bank in charge of the IPO process has less risk of public markets investors shying away from the company's IPO, and as a result they would not have to underprice as heavily to ensure that all issuance is allocated. Additionally, a higher net income margin means that the company is more profitable and can produce greater cash flows, which are ultimately what investors are entitled to as equity holders. Thus, a higher net income margin would attract greater investor demand before IPO and consequently would require less underpricing on the part of the investment bank. This is in line with the bank preference theoretical framework detailed in Section 3.

- A company with **more total assets** before the IPO leads to a slightly smaller underpricing effect. More precisely, a one million dollar increase in the total assets of the company before the IPO leads to a 0.99987x multiplicative effect on the one-day percentage return of the issued stock. A possible explanation of this effect is that more assets are indicative of a healthier company. With a healthier company, the investment bank in charge of the IPO process has less risk of public markets investors shying away from the company's IPO, and as a result they would not have to underprice as heavily to ensure that all issuance is allocated. This is in line with the bank preference theoretical framework detailed in Section 3.

The stepwise model displayed good homoscedasticity, linearity, and normality. There were no extremely high leverage points; while many other studies of IPO underpricing suffer from outliers and high leverage points, it seems that restricting our attention to just VC-backed IPOs could have made it so the dataset did not have very high leverage points,

as the dataset is composed of companies with very similar characteristics, all operating in the same VC-backed tech startup mold. The residuals vs fitted and scale-location plot did not show many abnormalities, except that data was more clustered and residuals were higher in the middle of the fitted values. This is shown shown by the following diagnostic plots (Figure 7)[1]:

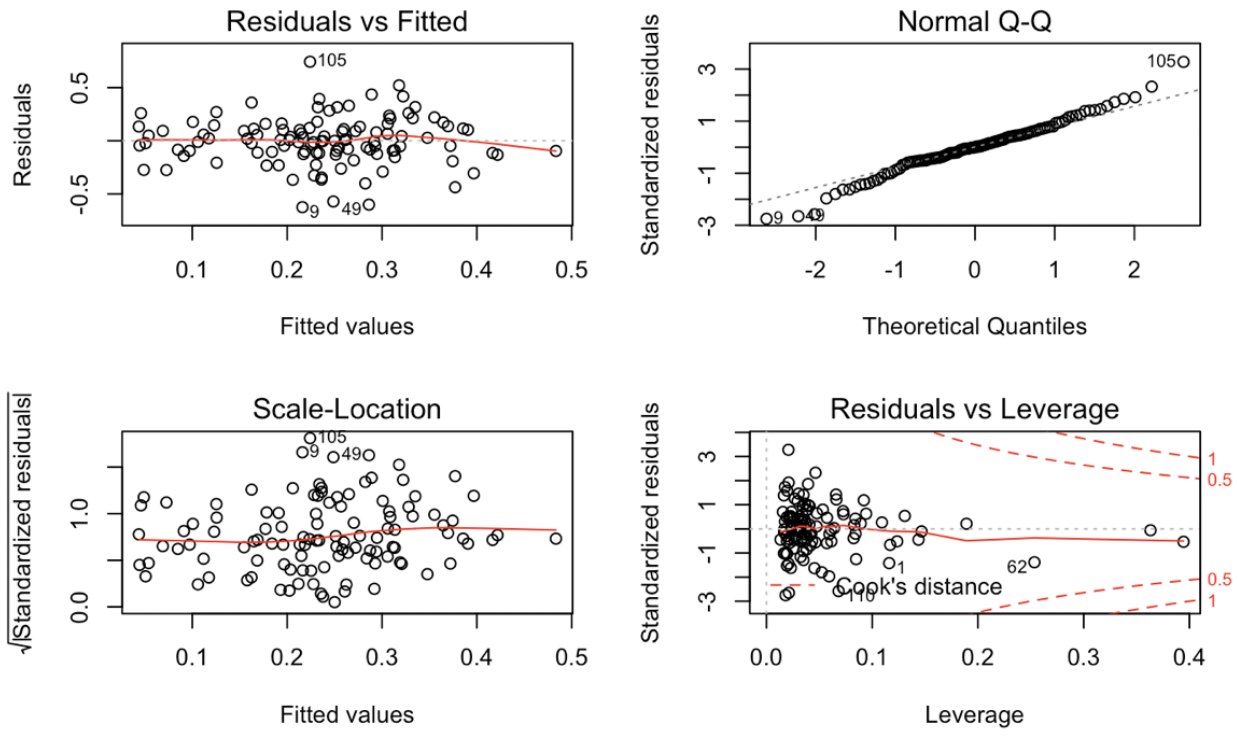

Figure 7: Stepwise Model Diagnostic Plots

Additionally, the stepwise model had an adjusted R^2 of 10.45%.

6.1.2 Generalized Additive Model

The next step in the analysis was fitting a generalized additive model (GAM). We chose to fit this with GAM because we wanted to use it as a possible exploratory tool for our

[1]However, there do exist mild concerns about independence, as IPOs can indeed influence each other: for example, the IPO and performance of a company like Dropbox might influence the IPO of another very similar company like Box.

parametric regression model. We hoped to identify a possible transformation on one of the predictors. GAM applies smoothing functions to predictor variables in order to achieve a better fit for the regression. The GAM achieved an adjusted R^2 of 42.2%. The results of the GAM corroborated the result of the stepwise model that market performance before IPO was a significant variable, and also provided strong evidence that the Proceeds raised in the IPO and Total VC Funding Raised before IPO were also very significant predictors for underpricing. The Proceeds term had a p-value of 0.00158 in the GAM, and the Total VC Funding Raised term had a p-value of 0.0290).

6.1.3 Final Linear Regression Model

Using the guidance from the stepwise model and the GAM (mainly by implementing relatively significant variables from those models), the linear regression model that was ultimately used was the following:

$$\begin{aligned}\text{LogOneDayChange} = \beta_0 + \beta_1\text{*VIX\_before} + \beta_2\text{*market\_performance} + \\ \beta_3\text{*Selling.and.Marketing} + \beta_4\text{*Net.Income.Margin} + \beta_5\text{*Rev.Growth} + \\ \beta_6\text{*Total.Assets} + \beta_7\text{*log(Proceeds)} + \beta_8\text{*Type.of.Security} + \\ \beta_9\text{*Selling.and.Marketing:RevGrowth} + \\ \beta_{10}\text{*market\_performance:Selling.and.Marketing}\end{aligned}$$

This final linear regression model built on the variables suggested by the stepwise model and made logarithmic transformations on certain variables following guidance from exploratory data analysis and from smoothing functions in the GAM. Additionally, interactions were explored and relatively significant ones were added into the regression model.

The model had an adjusted R^2 of 24.1 %, and the following variables were significant at at least a 0.05 significance level (See Figure 8 below).

In particular, these results confirm and build on the conclusions from the stepwise model. Namely, the following can be seen:

	LogOneDayChange		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.40	-0.82 – 0.02	0.061
VIX_before	-0.01	-0.01 – 0.00	0.122
market_performance	2.89	1.01 – 4.78	**0.003**
Selling.and.Marketing	0.00	-0.00 – 0.00	0.518
Net.Income.Margin	-0.10	-0.20 – -0.01	**0.037**
RevGrowth	0.08	0.00 – 0.16	**0.041**
Total.Assets	-0.00	-0.00 – -0.00	**0.010**
log(Proceeds)	0.16	0.08 – 0.24	**<0.001**
Type.of.Security: Common Shares	-0.07	-0.18 – 0.04	0.186
Selling.and.Marketing:RevGrowth	-0.00	-0.00 – -0.00	**0.002**
market_performance:Selling.and.Marketing	-0.03	-0.05 – -0.00	**0.045**
Observations	113		
R^2 / R^2 adjusted	0.309 / 0.241		

Figure 8: Final Linear Regression Model Output

- **Market Performance**: This model corroborates the conclusions from the stepwise model about the effect of S&P performance before the IPO on the underpricing effect. Market performance was one of the most significant variables in this regression, with a p-value of 0.003. Going public in hotter markets seems to lead to a greater first-day trading pop, and thus a larger underpricing effect. The size of the effect in this

regression was about two times larger than the size of the effect in the previous stepwise regression.

- **Total Assets**: This model corroborates the conclusions from the stepwise model about the effect of the total assets of a company before the IPO on the underpricing effect. Total assets was also a very significant variable in this regression, with a p-value of 0.01.

- **Net Income Margin**: This model corroborates the conclusions from the stepwise model about the effect of the net income margin of a company before the IPO on the underpricing effect. Net income margin was also a significant variable in this regression at a 0.05 significance level, with a p-value of 0.037.

- **Proceeds**: In additional to the three effects above which were already identified in the stepwise model, the log of the total proceeds raised during the IPO was the most significant predictor, with a p-value of 0.00019. An IPO that raised more proceeds will be underpriced more. A possible explanation of this effect is that larger IPOs lead to a greater amount of proceeds raised, and these larger IPOs may be more well-known to retail investors as they would be IPOs of more well-known companies. These retail investors may then drive up the price on the first-trading day, even if a fair price was decided by the investment based on the demand of institutional (not retail) investors, and this would lead to greater underpricing.

- **Revenue Growth**: Another additional insight identified by this linear regression model was the effect of revenue growth experienced by the company in the year before IPO on the underpricing effect. The p-value for this variable was 0.041. Surprisingly, stronger revenue growth was correlated with more underpricing. This may be explained by the trend of venture-backed companies prioritizing revenue growth above all else, causing them to spend inordinate amounts on selling and marketing to boost growth. This means that perhaps some of the revenue growth experienced is actually unhealthy and financially unsustainable (driven only by increased selling

costs), and so could be a signal of unhealthiness, which would require more underpricing on the part of the investment banks. More of this is explored and evidenced by the Selling and Marketing Costs and Revenue Growth interaction term, which is discussed next.

- **Selling and Marketing Costs and Revenue Growth Interaction**: For higher levels or revenue growth, increased selling and marketing costs will actually cause less underpricing. A possible explanation of this is that a company's selling and marketing costs are spent for the primary purpose of increasing revenue. If indeed there is more revenue growth, then it would appear that the investment in selling and marketing has paid off in an efficient manner, which would signal to investors that the company is functioning well and allocating resources effectively. This would provide more confidence to these investors that the company is a healthy one that would indeed continue to grow. Consequently, an investment bank would have to underprice less to be able to sell all the issued shares to these investors. This was a very significant term, with a p-value of 0.00181.

  Moreover, this interaction effect can be used to help explain the positive effect that increased revenue growth has on underpricing. The interaction effectively caveats this effect by emphasizing that if the revenue growth was obtained efficiently by the company (appropriate levels of selling and marketing costs were used to achieve that revenue growth), then the revenue growth is seen in a much healthier light and so there will be less underpricing on the part of the investment banks as they try to sell the issuance of a company with healthier growth to investors. This interaction effect can be seen in more detail in the graph in the Appendix captioned: "Selling and Marketing Costs and Revenue Growth Interaction Plot."

- **Selling and Marketing Costs and Market Performance**: For greater levels or selling and marketing costs, increased market performance cause less underpricing. Meanwhile, for lower levels of selling and marketing costs, increased market performance

leads to more underpricing. A possible explanation of this is that in a hotter market, companies are expected to try to grow faster and investors are also less wary of higher costs. Low spending while in good market conditions can indicate underlying issues with the company, such as an inability to raise funds to finance the spending. An inability to raise funds in such good market conditions would reflect negatively on the company's health. This interaction effect can be seen in more detail in the graph in the Appendix captioned: "Selling and Marketing Costs and Market Performance Interaction Plot."

Through tinkering with the regression and analyzing the outputs, a few more effects were revealed to not be statistically significant. These effects, which pertain mainly to the effect of particular investment banks conducting an IPO, are explored briefly below:

- **Investment Bank**: The particular investment bank who served as the lead left bookrunner, whether it was Goldman Sachs or Morgan Stanley or JP Morgan or others, did not have a statistically significant impact on the underpricing effect.
- **Additional Lead Underwriters**: The number of lead underwriters in addition to the lead left bookrunner did not have a statistically significant impact on the underpricing effect.
- These two results seem to indicate that there is no significant signalling advantage to be gained from using a particular investment bank or a particular number of investment banks to conduct the IPO, at least as it pertains to underpricing. Even though Morgan Stanley and Goldman Sachs were the lead left bank on a majority of the IPOs in the dataset, it did not seem that either they or any other bank affected the degree of underpricing for IPOs. (Oftentimes it is remarked that having a prestigious bank like Goldman Sachs or Morgan Stanley leading an IPO gives that IPO a mark of quality (signals to the market that this is a good IPO to invest in), and this is what I refer to as a signalling advantage).

The linear regression model displayed good homoscedasticity, linearity, and normality. There was were no high leverage points that affected the regression significantly. The residuals vs fitted and scale-location plot did not show many abnormalities, and was less clustered around the middle than the stepwise regression. This is shown shown in the following diagnostic plots (Figure 9):

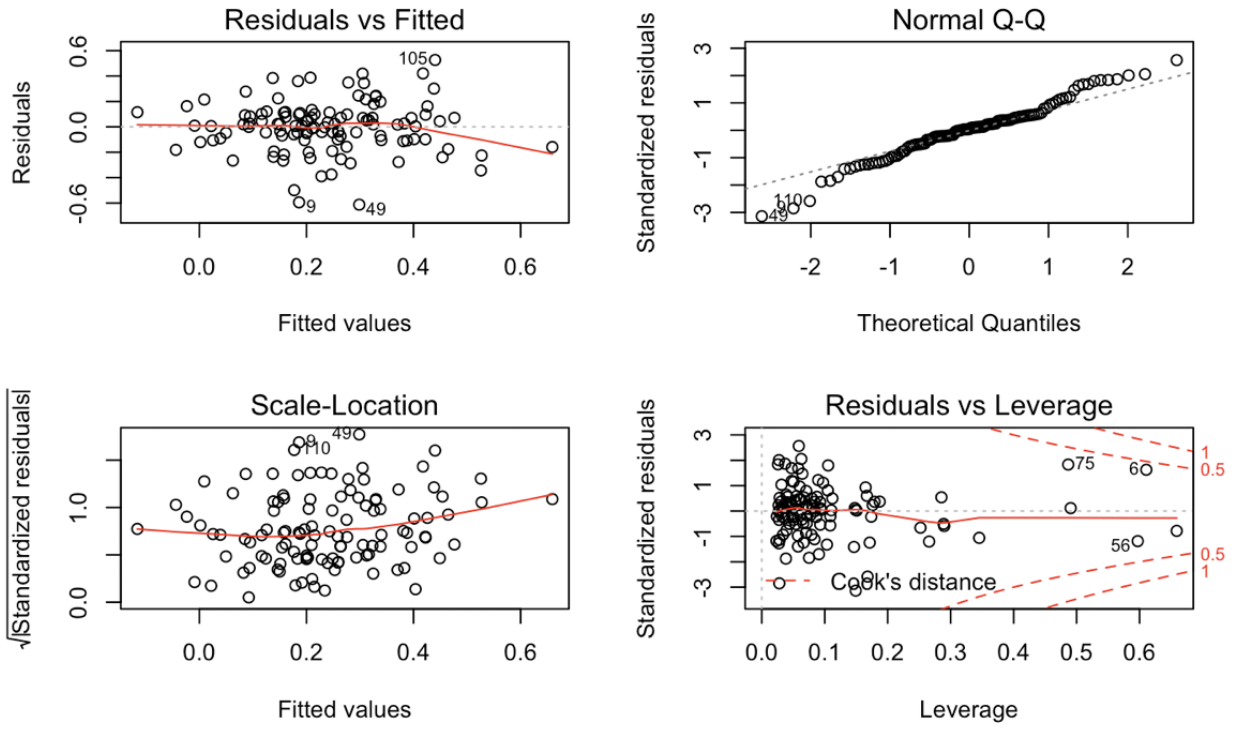

Figure 9: Linear Regression Model Diagnostic Plots

6.1.4 Random Forest Regression

To further inform the linear regression results, a random forest regression was fitted using the total dataset (no variable selection occurred). In theory, the random forest regression is more resilient to outliers. The random forest regression performed well, with RMSE being limited to a few percentage points of One Day Returns (See Figure 10 below):

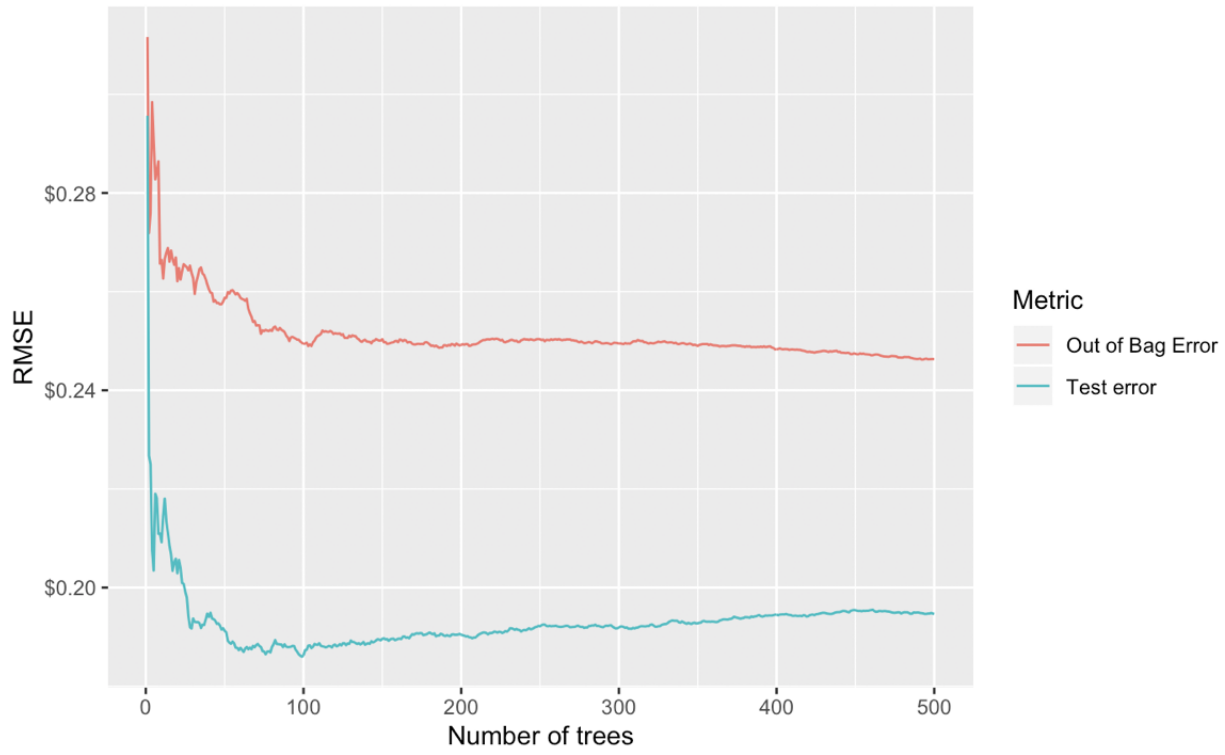

Figure 10: RMSE

The random forest regression output below shows a list of variables ordered by importance (how much they impact the response variable of log one day returns). However, the random forest regression output cannot be interpreted directionally.

In line with the linear regression results is the conclusion that proceeds raised, net income margin, revenue growth, total assets, and market performance are very important to underpricing. Proceeds raised was by far the most important predictor. Additionally, the total VC funding raised before IPO and selling and marketing costs were of relatively high importance. Though the random forest regression does not (and neither does any regression) provide an explanation for the mechanism through which the financials affect the underpricing effect, the results do provide evidence to support the theoretical framework that these firm-specific financials are important to the degree of underpricing.

However, contrary to the linear regression results, the random forest regression shows that choosing a particular investment bank to serve in the lead left role is indeed important

to the underpricing effect. See Figure 11 highlighting variable importance:

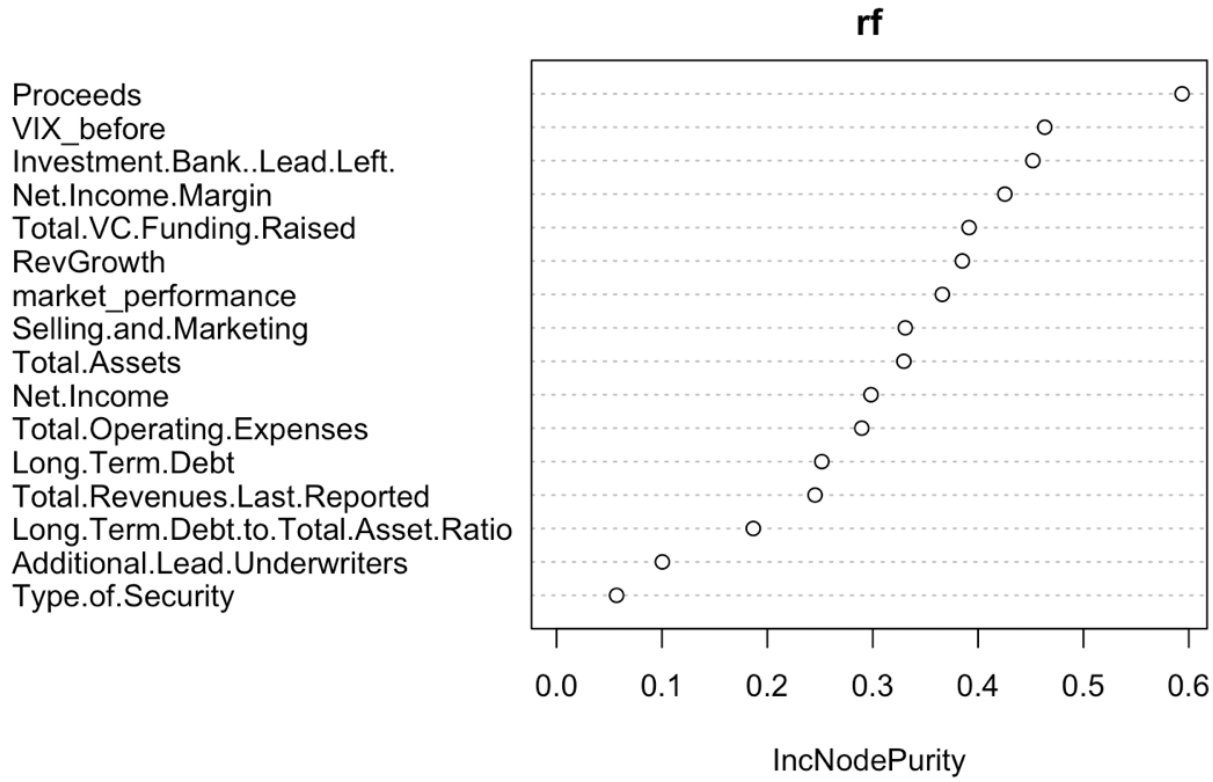

Figure 11: Importance of Variables in Random Forest

6.2 Classification

6.3 Logistic Regression Model

A logistic regression model was also used as a classification tool to further explore the underpricing effect. Given that the median one-day change in the dataset was 26% (as a log return this is 0.2311), this was used as the delineation point to determine what was underpriced and what was not[2].

The predictors used in the logistic regression model were the same as in the final linear

[2]In the dataset of 113 points, only 16 had a non-positive one-day change, which would be the technical dictionary delineation point for underpricing

regression model, and this was determined by using similar techniques as in the linear regression. The model is written below in mathematical form:

$$logit(P(y = 1)) = \beta_0 + \beta_1\text{*VIX\_before} + \beta_2\text{*market\_performance} +$$
$$\beta_3\text{*Selling.and.Marketing} + \beta_4\text{*Net.Income.Margin} + \beta_5\text{*Rev.Growth} +$$
$$\beta_6\text{*Total.Assets} + \beta_7\text{*log(Proceeds)} + \beta_8\text{*Type.of.Security} +$$
$$\beta_9\text{*Selling.and.Marketing:RevGrowth} +$$
$$\beta_{10}\text{*market\_performance:Selling.and.Marketing}$$

This model had an AIC of 100.55 and a McFadden's R^2 of 0.33. Significant variables at the 0.05 significance level were the following: market performance, selling and marketing costs, log(Proceeds), and the interaction term between selling and marketing costs and revenue growth. Market performance had a huge impact on increasing the odds that an IPO would be underpriced. For a for a five percent increase in market performance prior to IPO, we expect a 94% increase in the log-odds that the IPO is underpriced. See Figure 13 for an output of the significant variables:

Coefficient	**Estimate**	**Std. Error**	**t-value**	**P-value**
market_performance	38.810821	17.790617	2.182	0.02914
Selling and Marketing	-0.023204	0.010890	2.131	0.03311
log(Proceeds)	1.808543	0.687515	2.631	0.00852
Selling.and.Marketing:RevGrowth	-0.026101	0.009860	-2.647	0.00812

Figure 12: Logistic Regression Model Output

75% of the dataset was used as training data and 25% as test data. The `createDataPartition` function in the `caret` package ensured that the training and test sets would have close to equal proportions of underpriced vs not underpriced entries. Under 10-fold cross-validation, the AUC was 0.757. On the test dataset, the model performed admirably, with

an accuracy score of 0.8571. Sensitivity and specificity were both high, at 0.9412 and 0.7273 respectively. The resulting confusion matrix is below (Figure 13):

		Actual	
		Not Underpriced	Underpriced
Prediction	Not Underpriced	16	3
	Underpriced	1	8

Figure 13: Confusion Matrix

It seems that the model is very good at identifying companies that are not overpriced.

This classification task was also tried with other delineation points, such as classifying something as underpriced if the one day return was greater than 0%, or if it was greater than 30%. These models performed worse. A random forest classification with the total dataset using the 26% threshold also did not perform as well, having an accuracy score of just 0.6818 and specificity of 0.5.

7 Conclusion

In this study investigating IPO underpricing through the lens of the bank preference theory, linear regression, logistic regression, and random forest methods were used to explore predictors of underpricing, specifically in the context of VC-backed technology IPO's since the Great Recession. In this specific segment of companies, there exists a high degree of underpricing by the textbook definition – the average one-day stock price gain after IPO was 26.76%, and only 16 of the 113 companies in the dataset did not experience a positive gain in stock price on the first trading day. Though there is money left on the table, it seems that underpricing of 20-30% could be customary or even desired by issuers, perhaps for the marketing purpose of attracting more investor demand in the short or long term.

Almost all of the regression results provided at least some evidence to support the theory that underpricing occurs at least in part because the investment banks are setting

prices lower to decrease the risk that they will be unable to allocate all the issued stock. The health of the company at the time of IPO could influence the degree of underpricing – a healthier company would not require as much underpricing for the banks to be able to ensure they meet their allocation targets, as the shares of a healthy company would already be in high demand by public markets investors.

Indeed, the data shows evidence to support this – companies with better net income margins and more total assets experience smaller degrees of underpricing. Net income margin is particularly important, as it is closely tied to the cash flow of the company, which is what is theoretically the main driver of fundamental investment decisions.

Market performance was also a very powerful predictor of underpricing. It had an pronounced affect on underpricing in both the logistic and linear models. This could be because issuing companies tend to time their IPOs for hot markets where investor demand is high and where they can thus raise the most money in an IPO. This also provides evidence for the market cyclicality theory of underpricing.

Another interesting finding was that the linear regression found no strong evidence (though the random forest did) to support the idea that there is a significant signalling advantage to be gained from using a particular investment bank or a particular number of investment banks to conduct the IPO, at least as it pertains to underpricing. This is contrary to the popular belief that having a prestigious bank like Goldman Sachs or Morgan Stanley leading an IPO gives that IPO a mark of quality.

The use of random forest regression allowed for the triangulation of insights between it and the linear regression, and also for the discovery of other variables that may be relevant to underpricing, such as the total VC funding raised prior to IPO.

8 Appendix

8.0.1 Multicollinearity in Linear Regression

VIX_before	market_performance
1.206393	2.092815
Selling.and.Marketing	Net.Income.Margin
2.507734	1.319189
RevGrowth	Total.Assets
3.276779	1.926652
log(Proceeds)	Type.of.Security
3.128174	1.622168
Selling.and.Marketing:RevGrowth	market_performance:Selling.and.Marketing
5.127180	2.726646

Figure 14: Multicollinearity in Final Linear Regression

8.0.2 Interaction Effects in Linear Regression

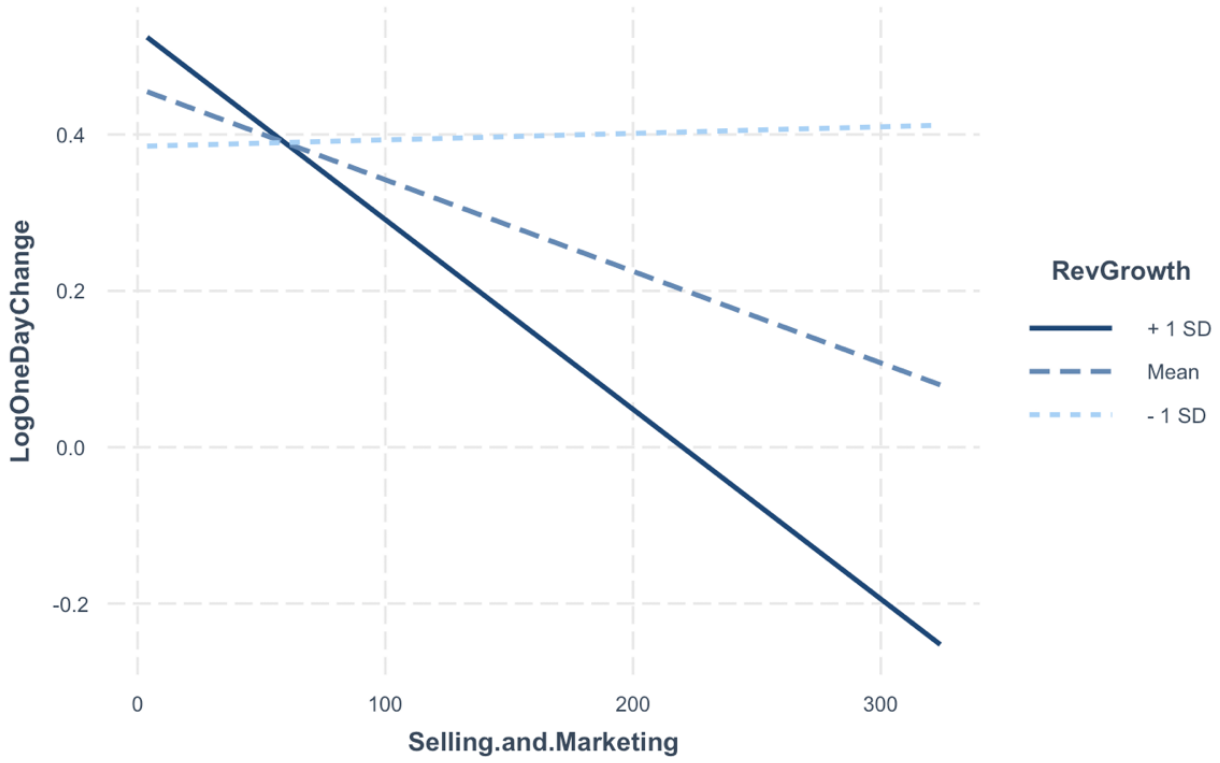

Figure 15: Selling and Marketing Costs and Revenue Growth Interaction Plot

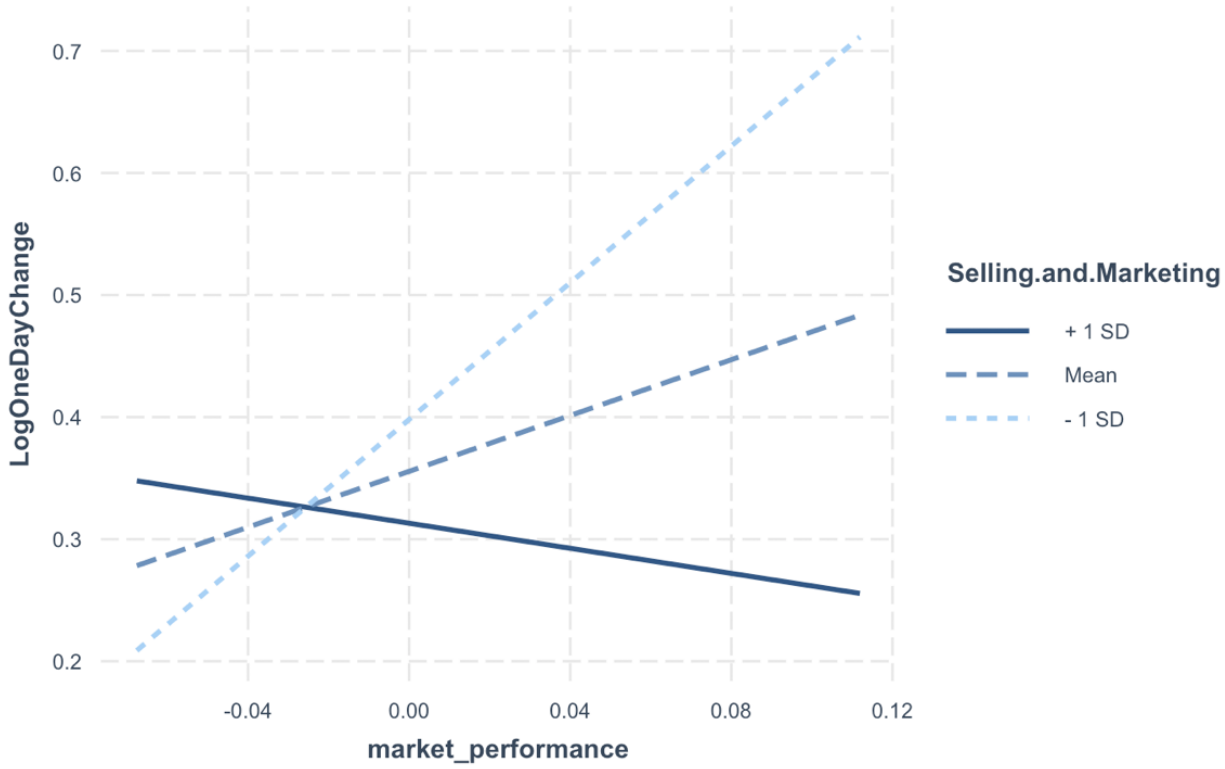

Figure 16: Selling and Marketing Costs and Market Performance Interaction Plot

8.0.3 Checking Independence in Linear Regression

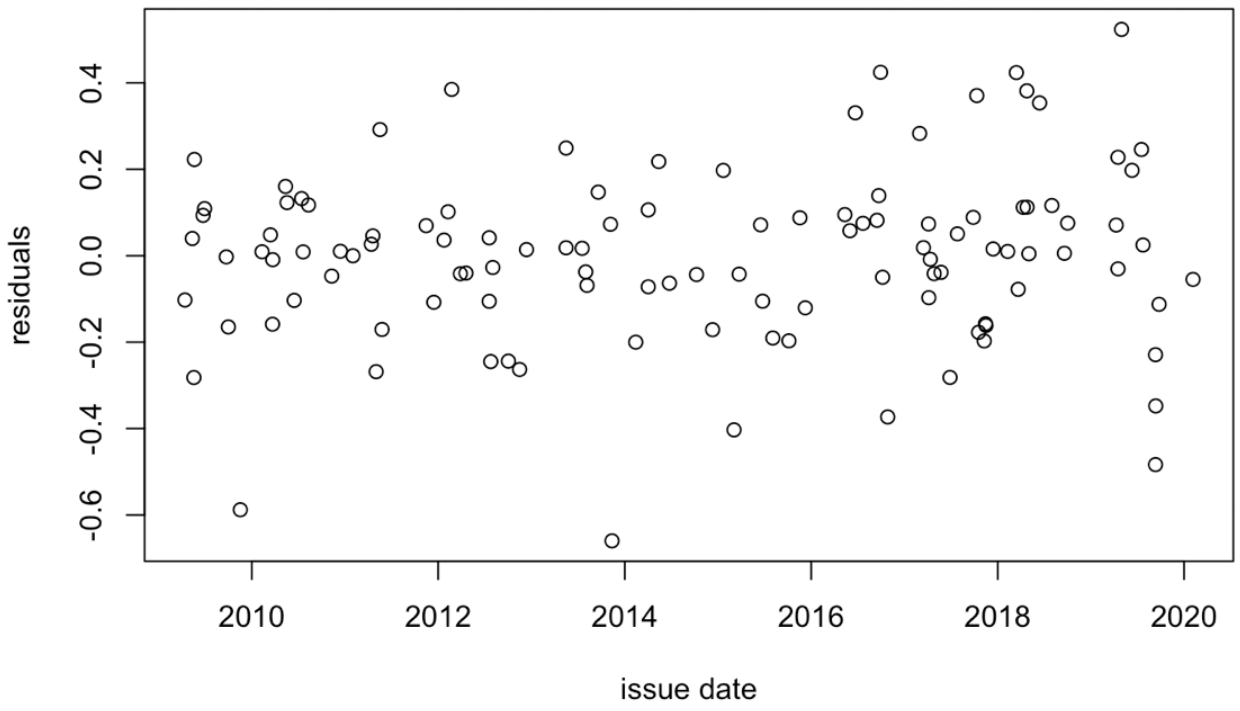

Figure 17: Residuals vs Issue Date

9 References

[1] Ritter, J (2018). Initial Public Offerings: Underpricing.

[2] Torstila, S (2001). What Determines IPO Gross Spreads in Europe? Blackwell Publishers.

[3] Ritter, J (1984). The ”Hot Issue” Market of 1980. The Journal of Business, 57(2).

[4] Megginson, W. L., Weiss, K. A. (1991). Venture Capitalist Certification in Initial Public Offerings. The Journal of Finance, 46(3), 879–903.

[5] Jain, B. A., Kini, O. (2000). Does the Presence of Venture Capitalists Improve the Survival Profile of IPO Firms? Journal of Business Finance Accounting, 27(9–10), 1139–1183.

[6] Baba, B., Sevil, G. (2019). Predicting IPO Initial Returns Using Random Forest.