PREDICTING STARTUP SUCCESS IN THE U.S

by

Felipe Veloso

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Economics

Charlotte

2020

Approved by:

_____

Dr. Craig A. Depken II

_____

Dr. Matthew Metzgar

_____

Dr. Sonia Jarvis

ABSTRACT

FELIPE VELOSO. Predicting startup success in the U.S. (Under the direction of DR. CRAIG A. DEPKEN II)

This thesis consists of the creation of a reliable model to predict the success of Startups located in U.S. Previous researches have been focused either on the accuracy of different algorithms, without investigating the real effect of the risk and success factors, or on explaining the effects of those factors for Startup success/failure, such as the education of the entrepreneurs, financing methods, and timing. Another difference is that this research is focused only on Startups located in the U.S. This research combines the findings of many studies of the determinants of startup success and models focused on comparing different predictive methods, like Logistic Regressions, Linear Discrimination Analysis and Machine Learning Algorithms. The final output is a reliable and predictive model that could help Angel Investors and Venture Capitalists to build more consistent and measurable startup portfolios.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Sonia Jarvis for guiding me in this research and encouraging me to strive to be my best, personally and professionally, by always telling what I needed to hear, no matter how bad it was. I would like to thank Dr. Craig Depken for the practical advice and for teaching me most of what I know about econometrics. Also, I would like to thank my bosses during my time as a Teaching Assistant, Dr. Connaughton, for teaching me how to think like an economist, and Dr. Metzgar, for all the advice and kindness during my very stressful first year in the United States. Most importantly, I am more than grateful to have the unconditional and unfailing support of my family, who always trusted in my potential and taught me to have faith in God and that hard work pays off.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Startups are companies with high growth potential and usually innovative Business Plans and Technologies. In the past fifty years, they have entirely changed the global business environment and the way people spend their time. There are many examples of companies that started as a startup and today figure among the Fortune 500 companies, such as Apple, Facebook, Alphabet (Google), Amazon, Microsoft, Stripe, and Uber, among many others. The technology enables companies to provide products/ services that are cheaper, faster, and more convenient for the customers. This advantage, generally caused by economies of scale, lower labor costs, and better accessibility, has the potential to generate fast traction and very high returns for investors. However, the possibility of high returns comes along with a much higher risk. According to Startup Genome, over ninety percent of startups fail, increasing the uncertainty regarding investment decisions to Angel Investors and Venture Capitalists. This pattern of a majority of failures and very few huge successes have many reasons. I refer to most of them in this thesis. One of the most compelling reasons for this phenomenon is called "The winner-takes-all economy" , which posits that:

> Economists have long described winner-takes-all markets in which small differences in performance lead to significant differences in rewards. Such markets include those for star entertainers and athletes. Julia Roberts and Tiger Woods, for example, make vastly more than average members of the Screen Actors Guild and the Professional Golfers' Association, respectively, make. By contrast, in business, small differences in performance have traditionally generated only small differences in rewards. But the situation is currently changing. Notable examples over the past

> decade include the outsized performance generated by single product lines such as Microsoft's Windows operating systems, Intel's microprocessors, and Nokia's mobile telephones. Now this dynamic is spreading to other sectors.
>
> In the future, this gap between winners and underperformers is likely to keep widening. In this new economic landscape, companies that are not the very best in their segment are likely to face competition trouble, even if they have the foresight or good fortune to pick attractive industry segments. Campbell and Hulme (2001)

There are several pieces of research about the factors that influence startup success. Wong (2002) studies the characteristics and differences in terms of support and influence between Angel investors and Venture Capitalists to startup management and network, Solomon et al. (2008) empirically analyzes the effect of education on entrepreneurial success, Kenney and Von Burg (1999) study the regional characteristics that affect the companies within it, using the Silicon Valley as an example. There are other models with different approaches to predict startup success,Krishna et al. (2016) utilizes financial aspects of startups and an artificially created "severity score" to proxy entrepreneur's characteristics as predictors and Logistic Regression and some machine learning algorithms as the methods. Shah (2019) utilizes diversified risk and success factors, such as the entrepreneur's education, amount and source of investments, number of milestones achieved, regional variables and the category of the industry, Lussier (1996) utilizes variables such as financial control, timing, staffing and others that are generally not used in other researches, it also utilized the Linear Discrimination analysis as the benchmark algorithm. This research differs from all of those in four aspects. First, it only considers startups from the U.S. Secondly, it considers a success not only good exits (IPO and Acquisition) but also startups that have milestones published in the news after five years operating (they are more likely to

have traction and return the early investments). Thirdly, it utilizes transformations and selection bias corrector to guarantee that the results are not only predictive and statistically significant but also are consistent with economic intuition and previous literature.

This thesis provides a detailed step-by-step Model Development approach to predict startup success utilizing different types of algorithms. Chapter 2 defines the desired outcome of the model, the mainstream statistical and machine learning techniques, and the evaluating methods to compare their accuracy.

Chapter 3 details the data and their sources, the modifications required and their reasons, the definition of success/ failure and the intuition behind that, the success and risk factors to be tested, the method to define the training and testing data and the limitations of the research caused by the data selection.

Chapter 4 compares the expected and estimated results, excludes and modifies variables using the results of the benchmark algorithm (Logistic Regression) to guarantee better accuracy and interpretation, implements the rejection inference, and compare the accuracy of all the algorithms.

Chapter 5 summarizes the results and suggests ideas for future research.

CHAPTER 2: MODEL FRAMEWORK AND THEORY

Startup success is a tricky definition, some researches, such as Bento (2018) and Huang (2016), consider a successful startup those with a successful exit. A successful exit, as explained by Guo et al. (2015), can take two routes, an IPO or a Merger/ Acquisition. He says:

> There are two main exit routes for a successful startup. The company can go through an Initial Public Offering (IPO), or it can be sold to an existing firm via an acquisition. First, under an IPO, the venture obtains a stock market listing, which enables the company to receive additional financing for its projects and enables insiders to sell their shares to the public eventually. If the startup is acquired, the insiders obtain immediate cash in return for their shares. Understanding the leading trade-offs faced by startups at the exit stage is crucial because this understanding not only allows one to determine how venture capitalists and entrepreneurs divest their companies but also how the decisions are taken at the onset of the venture. – *Guo, Bing and Lou, Yun and Perez Castillo, David. Journal of Economics & Management Strategy; v. 24, n. 2 1, (2005): 415-455.*

Other researches, such as Shah (2019) and Krishna et al. (2016), who both utilize data from Crunchbase, consider successful startups labeled as "operating" in the database (although it may generate misleading conceptual results in the model, explained in Chapter 3).

This thesis evaluates success in the investor's perspective, which is the likelihood to get a return from the investment made in a startup. A successful exit surely could

represent a very high likelihood of return from the investment. However, those are not the only ways that an investor could profit with an investment in a startup. A startup with significant traction that becomes a trustworthy company, consequently, tends to pay back its shareholders through dividends, and so also could be considered a success.

This thesis considers successful startups, those that 1- Had a successful IPO, 2- Merged or were Acquired, 3- Are operating and still have successful results after five years in operation (Chapter 3 explains this selection method.) I classify the successful startups as "1" and the failure startups as "0".

2.1 Model Framework

The data provides the status of startups (operating, closed, IPO, M&A). The idea is to distinguish between the successes and failures utilizing other variables from the data set (e.g., milestones, funding rounds, funding amounts, education of the entrepreneur) and external data (e.g., industry growth and GDP growth) to predict whether or not a startup is likely to succeed. The predicted values fall into two categories, 0 for failure and 1 for success.

In the estimation phase, I randomly divided the data into two data sets: training data (corresponding to 70% of the observations) and testing data (corresponding to 30% of the observations). The idea is to fit the model in the training data using the benchmark method (Logistic Regression) and test the performance of the model and the significance of the variables in the testing data.

After evaluating the suitability of alternative frameworks, this research compares the performance of methods such as Logistic Regression, Linear Discrimination Analysis, XGBoosting, and Support Vector Machine with Kernel.

2.2 Method - Logistic Regression

The Logistic Regression is a specific case of linear regression where the response, Y, is a binomial variable. Logistic regression models the probability that Y belongs to one of the two categories Berkson (1944):

$$log(\frac{p(x_i)}{1-p(x_i)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.1}$$

where

$$p(x_i) = P(y_i = 1|x_i) = \frac{e^{\sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}}} \tag{2.2}$$

By design, the logistic regression function gives output between 0 and 1, which is the probability of belonging to one of the binary classes, $P(x_i)$. The coefficients of Equation (7) are fit using the maximum likelihood estimation Fisher (1912).

2.3 Alternative Method 1 - Linear Discriminant Analysis (LDA)

This method generates a linear discriminant score based on the linear combination of the input variables that maximizes the ratio of the variance between the classes to variance within the classesFisher (1936).

Assume a linear model that looks like a regression $y = w_1 x_1 + ... + w_i x_i + ... + w_n x_n$, where x represents the variables so that $x = (x_1, ..., x_i, ..., x_n)$ and $w = (w_1, ..., w_i, ..., w_n)$ represents a weight vector. The results of the linear model is represented as $y = w^T x + w_0$. The classification is obtained by placing a threshold on y, i.e. w_0, which is the mid-point of distance between the means. The goal is to select the projected results that best separates the two groups.Kennedy (2013)

Assuming that the two groups have a common sample variance, Fisher (1936) suggests the use of a sensible measure of separation as:

$$M = \frac{distance\ between\ sample\ means\ of\ two\ groups}{sample\ variance\ of\ each\ group} \tag{2.3}$$

Where M is the separating distance, this guarantees a more significant separation while also enforces a small variance within each class, minimizing the class overlap.

Applying the idea of this research and assuming m as the sample mean and $m_{success}$ and $m_{failure}$ the sample means of the observations of the startups considered successful and failures, also assuming V as the common sample variance, the corresponding separating distances would be:

$$M = w^T . \frac{m_{success} - m_{failure}}{(w^T . V . w)^{\frac{1}{2}}} \tag{2.4}$$

where w^T is the transpose of w. After differentiating the equation above the result is:

$$w^T \propto (V^-1(m_{success-m_{failure})^T}) \tag{2.5}$$

LDA assumes that the inputs are measured on an interval scale or ratio scale, requiring in many situations data manipulation. Bishop (2006)

2.4 Alternative Method 2 - Support Vector Machine (SVM)

Support Vector Machines (SVM) Vapnik (1998) is a statistical pattern classifier that utilizes a kernel technique Burges et al. (1999). It uses training data to learn a classifier (i.e., similar to what the coefficients mean for logistic regression), then, using those parameters, it classifies the test data Cortes and Vapnik (1995). SVM constructs a high-dimensional plane to separate groups in a n-dimensional feature space, where n is the number of features Sacchet et al. (2015). The goal is to find the hyperplane that maximizes the margins between the two support vectors; Figure 2.1 illustrates the method. To be specific, to classify successful startups vs. failed startups, using the risk and success factors as the features. The hyperplane is defined by the following equation:

$$< w, x > + b = 0 \tag{2.6}$$

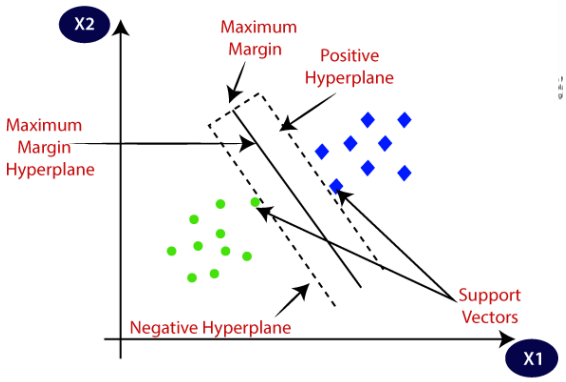

Figure 2.1: SVM illustration – *Source: Javapoint*

where $x_i \in R_d$ represents the vector of the risk/ success factors with length d, and $w \in R_d$ separates the groups (success and failure) by maximizing the margin between the hyperplane and each group. The optimal hyperplane is identified through the L2-norm problem:

$$\min \left(\frac{1}{2} \langle w, w \rangle + D \sum_i v_i^2 \right) \tag{2.7}$$

with the following constrains:

$$\begin{aligned} y_i \big(\langle w, x_i \rangle + b \big) &\geq 1 - v_i \\ v_i &\geq 0 \end{aligned} \tag{2.8}$$

where D is a penalty parameter, v_i represents slack variables, and $y = \pm 1$ represents the group label, -1 for startups that failed and 1 for startups that succeeded. The value of D is scaled for each data point, depending on the group size:

$$D = \frac{N}{N_G} \tag{2.9}$$

where N_G is the number of observations in each group. The weights applied to the risk/ success factors were computed based on their relation to the hyperplane.

The SVM is usually a very accurate method that also applies to graphical and image related models; there are plenty of examples of this method used in medical

analysis, Sacchet et al. (2015) provides an excellent example of this application.

2.5 Alternative Method 3 - Extreme Gradient Boosting (XGBoost)

XGBoost Chen and Guestrin (2016) is a modern tree boosting system that won numerous Kaggle machine learning tournament in recent years, because of its high predictive power Huang et al. (2019). XGBoost (Extreme Gradient Boosting) implements gradient boosting decision trees with highly efficient system optimization to provide predictions that are scalable, portable, and accurate. In addition to traditional gradient boosting methods to decision trees Hauskrecht (2019), XGBoost includes stochastic gradient boosting with sub-sampling at the row, column, and column per split levels. The objective during training is to minimize the following function:

$$Obj = L + \Omega \tag{2.10}$$

L is the loss function, which in this case (binary model) it is a binary classification log loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i log(p_i) + (1-y_i) log(1-p_i)\Big] \tag{2.11}$$

Ω is the regularization term, controlling the complexity of the model.

$$\Omega = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{2.12}$$

where T is the number of leaves and w_j^2 is the score on the j^{th} leaf Huang et al. (2019).

Besides the excellent accuracy and attention in the competitions, XGBoosting is still considered a black-box model, which means that the coefficients and their significance are not well detailed, as it is in the logistic regression and other methods.

CHAPTER 3: MODEL DEVELOPMENT DATA

This research utilizes mostly data from Crunchbase, an online network created to connect startups and investors; Crunchbase is a complete database about startups in the market. For this reason, most of the researches about the subject utilize it. To complement, I add macroeconomic variables like GDP and Industry situation from the databases FRED and CBP (Counting Business Patterns).

The Crunchbase database contains data from over 100k startups of diverse industries and locations across the globe. As I use the academic license, the database restricts to startups founded before 2012, as the time range includes the financial crisis of 2008, it may generate conservative results, which is a plus to a good model.

There are eleven tables ("degrees", "IPO","funds", "objects","people","acquisitions", "funding rounds","investments","milestones", "offices", and "relationships") in the database, connected by the ID of each company. This raw format of the data enables the testing of many hypothesis and different combinations of variables to create a predictive model. However, as the database is "user-created," I had to make some modifications to assure that the data utilized would correctly illustrate the real startup environment, correctly capturing the risk and success factors to startup outcome. The data selection followed the next steps:

3.1 Data preparation

The data preparation consisted of the following steps:

1. Keep only startups founded from 2000 to 2012;

2. Keep only startups located in the U.S.;

3. Keep only startups with Angel or Venture capital investments higher than zero

(It is nearly impossible to operate a business without at least $1 of initial investment, so any startup under the operating, acquired, or IPO status and no initial investment may be omitting information);

4. Drop startups under the status "operating" without any milestone after 4 years of operation (Uncertainty regarding the real current status and likelihood to pay back the investments, as this research classify success in the investor perspective, it would be difficult to classify as either a success or a failure).

5. Integrate the necessary information contained in various tables and data sources into a unique DataFrame using the software Python.

3.2 Target Variable

As explained in the Model Framework, the target or dependent variable is "Successful startup - likely to generate a return to the investors." The steps to create the binary (0 and 1) target variable are:

- Startups under status "Closed" are classified as a failure "0".
- Startups with a successful exit (IPO or Acquired) are classified as successful "1".
- Startups under the status "Operating" and a new milestone (advancement with media coverage) after 4 years of operation are classified as successful "1".

Classification of the 927 final observations:

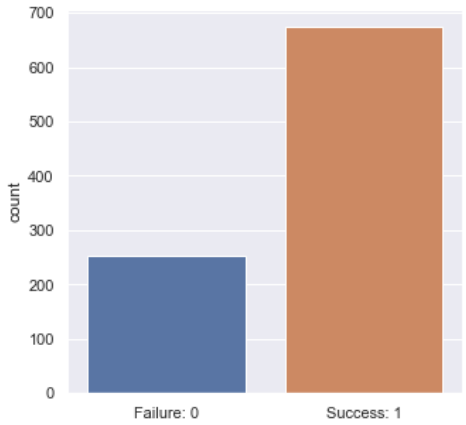

Figure 3.1: *Counts for 'Successes' and 'Failures' of the target variable*

3.3 Independent Variables

3.3.1 Candidate Pool

Following all the variables considered to build this model:

- Number of Angel Funding Rounds: This variable could explain the evolution of startups, the intuition is that the more funding rounds a startup have, the more likely it is to have a great product or market traction.

- Number of Milestones: This variable consists of the number of times the media released something good about the company. Intuitively one of the best predictors of startup success.

- Months from the foundation to the first milestone: This variable is a way to understand the psychological characteristics of the entrepreneur and the development process of the company.

- Percent of firms in the industry before the year of foundation: Out of all the operating companies in a specific industry, the percent founded before the year of each company's foundation, it shows the stage of the competition. A higher percent means a more established competition.

- Type of education and years from graduation: Those variables represent the type of degrees of the entrepreneurs (Undergrad, Masters, MBA, Ph.D.) and how long ago their graduated. These variables capture the effects of education and experience in entrepreneur success.

- The size of the team: The number of people on the team is an excellent way to capture the size of the business, which can add predictability to the model, taking into consideration the other variables, like industry and investment.

- Participants in the Angel investment rounds: The total number of angel investors shows how fragmented the equity is, which influences the management and, consequently, the outcome of the company.

- Amount of Venture capital investment: Just the fact that a startup reached venture capital rounds demonstrates that the company has potential, the amount of investment shows how significant this potential is.

- Amount of Angel investment: The amount of angel investment could demonstrate the potential of the idea and the energy and drive of the entrepreneur.

- Months to raise angel investment: The time to raise angel investment captures how fast the companies break-even or upgrade to a venture capital investment, it is essential to see whether the company is evolving or wasting capital.

- GDP growth: The GDP growth affects the investment and consumption in an economy; it could be a risk/ success factor for a startup to obtain early-stage traction.

- Type of industry: Some specific industries present peculiarities hard to measure by the financial aspects of a company or economic situation of a country/ industry, those peculiarities are measured here.

- The situation of the industry: The growth in the number of companies and average salary in the industry could explain trends that affect the future ability to find skilled human capital and the demand in each industry.

- Location: Some regions differ from others in terms of skilled human capital, the concentration of knowledge, and market acceptance to innovation, it is essential to capture them.

3.3.2 Transformations

Some transformations are necessary to address collinearity among variables and correct any possible non-linear behavior of any variable alone or group of variables. Following the transformations that I tried and the intuition behind them:

- **Number of Angel Funding Rounds and Time (Years) to raise angel investment:** There are two possible issues to address about these variables. One is the relationship between them, intuitively, the more rounds of investments a startup raises, the longer it takes to raise it. To fix this possible issue and address any possible collinearity or coefficient bias, I tried an interaction variable *Time (Years) to raise the angel investment multiplied by the Number of Angel rounds*. The second issue is that those variables may not have a linear effect on startup success, so I tried to add a second squared term of each, *Number of Angel Funding Rounds squared* and *Time (Years)to raise angel investment squared*, because intuitively they may present a decreasing return per scale.

- **Number of Milestones and Time (Months) to reach those Milestones:** Those two variables present the same behavior of the ones in the last item; the more milestones raised, the longer it is expected to take to raise those milestones. As in the last item, I tried an interaction term between those two variables *Number of Milestones multiplied by the Time (Months) to reach those Milestones*. Another aspect that needs attention is the role that those variables

play on one another. As an example, early media attention represented by a negative value of the time to reach the first milestone (milestone before that foundation of the company) only could represent something solid if followed by other milestones, meaning that a possible disruptive technology/ idea proved to be valid through next milestones. A late first milestone, followed by many other milestones, could mean that the company took its time to understand the market and adapt before expanding. A practical way to capture this behavior is through the substitution of the variables by a function; here I tried squared and cubic functions. $(a+b)^2 = a^2+2ab+b^2$ and $(a+b)^3 = a^3+2a^2b+2ab^2+b^3$.

- **Education of the Entrepreneur and Time in years from Graduation:** The type of education and years from graduation to the foundation of the startup (a proxy for experience) intuitively have a non-linear relation to successful entrepreneurship. Different degrees of education requires a different amount of experience to reach the optimum combination of knowledge and experience. The effect of the amount of experience on startup outcome is itself is not linear but rather squared, in which it has an upward slope for startup success in the first years, and then the slope decreases. To capture this relationship I converted those variables into a squared function, such as I did in the last item, $(a+b)^2 = a^2+2ab+b^2$.

Literature review of variables		
Variables	**Exp. effect**	**Benchmark research**
Number of funding rounds	Positive	Krishna et al. (2016)
Number of milestones	Positive	Shah (2019)
Time in months from foundation to the first milestone	Positive	Graham (2012)
Percent of firms in the industry before the year of foundation	Negative	Finkelstein (2002)
Education times months from graduation	Positive	Van Gelderen et al. (2005) Gimeno et al. (1997)
Small team -Team composed by three or fewer people (Dummy)	Negative	Roach and Sauermann (2015)
Participants in the angel investment rounds	Positive	Van Gelderen et al. (2005)
Startup reached Venture capital rounds of investment (Dummy)	Positive	Hellmann and Thiele (2015)
Time in months to raise Angel investment	Not clear	Feinleib (2011)
GDP	Positive	Yrle et al. (2000)
Dummy if the industry is expanding	Positive	Porter and Strategy (1980)
Type of industry (Dummy)	Both	-
State - Region (Dummy)	Both	-
Startup is location in the Silicon Valley and other micro regions (Dummy)	Positive	Krajcik and Formanek (2015)

3.4 Rejection Inference

As explained in the Model Selection section and Chapter 2 (Data Framework), I excluded the startups with neither Angel nor Venture capital investments from the sample. The reason for this exclusion is the lack of trust in those observations; a startup without some initial investment, seed, grant or angel, is unlikely to operate and obtain success, so the probability of missing information is high. Because of this exclusion, I did not have information regarding the startups with no investment and their likelihood to succeed based on the other factors. It generated a censoring problem. According to the literature (explained in Chapter 4), the amount of angel investment has a convex effect on the likelihood of success; it means that angel investment has a linear positive effect until it reaches some point, after it decreases. The intuition is that after some investment, the company needs to get enough traction to migrate to more advanced types of investment (e.g., Venture Capital), in which they receive a higher level of support. Too much angel investment could mean that either the company did not allocate well the resources utilizing too much capital without results or sold out too much equity too early Feinleib (2011). Because of the censoring, in the first times that I fitted a Logit model, the amount of angel investment presented a strictly negative linear effect, so it was necessary to include in the data startups that did not receive any investment. The method utilized is the Rejection Inference, as explained by Mok (2009)

> Reject inference is the process of estimating the risk of default for loan applicants rejected under the acceptance policy. The problem solved by the Rejection Inference is a data bias problem. When the data are missing completely at random, then there is no reject inference problem at all. If the data are missing at random, then the selection mechanism is ignorable. But when the data are missing not at random, then the selection mechanism is nonignorable. – *Jie-Men Mok (2009)*

The exact Reject Inference method applied is called "Re-Classification" ,Sergiu Luca (2018) explains it more details:

> Re-classification makes use of the accept/reject model as well, but under a different form and for a different purpose. The goal pursued under Re-classification consists of adding to the observed bads a sub-set of rejects most likely to be bads (note, no goods are added). The accept/reject model serves the purpose of proxy identification of rejects most likely to be bads. Once identified, these rejects-turnedbads ('RTB') are assigned a weight so that their contribution to resulting total bads (observed + inferred bads) is controlled. An important rule of thumb is that the inferred bads should not account for more than the observed bads. A popular value for the proportion of inferred bads to total bads (p) is in fact 30%. In addition to 30%, I have tested a proportion of 50%. Given that the difference between the two is negligible, I have only kept the 30%. – *Sergiu Luca & Desjardins - SAS Global Forum Proceedings (2018)*

The following figure demonstrates the balance between Success and Failure startups if I apply the threshold and classify the rejects (startups without any investment), adding it to the original database. Figure 3.2 shows the total sample in regards to success and failures after combining the data estimated in the rejection inference and the original sample.

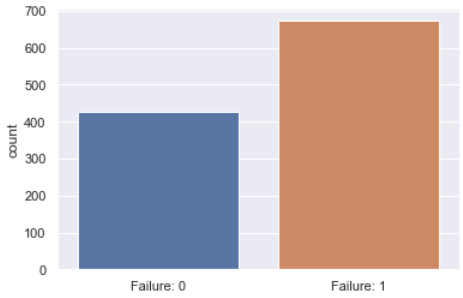

Figure 3.2: *Counts for 'Successes' and 'Failures' of the target variable after the rejection inference method*

3.5 Training and Testing Data

To fit and test the accuracy of models, I randomly separate the sample in two parts, training (in sample), corresponding to 70% of the observations, and testing (out of sample), corresponding to 30% of the observations. To ensure that the results are reliable, I drop the target variable from the testing sample.

3.6 Limitations

The following limitations of the data and its modifications have the potential to influence the results:

1. The research utilizes data about companies with at least some investment input in the database; I assume that they represent the entire population. However, it may be the case that just the fact that the companies provide complete information differentiates them from the others.

2. This model considers the "operating" companies with milestones after 4 years as successful and drops the others under the status of "operating," it may affect the coefficients, specifically overestimating the effect of variable "milestones."

3. This model utilizes rejection inference to adjust the coefficient of the binning variable "amount of angel investment," however, as I do not know for sure the outcome of those startups after a possible investment, some coefficients could

be over or underestimated.

4. Crunchbase as a Crowdfunding database, relying on correct inputs from Entrepreneurs, Investors, and Accelerators. This research assumes that those actors input the right information into the database and so the results represent the reality.

5. As this research bases its conclusions on the information from the companies inserted in the database, if the act of creating an account to reach investors represents a decisive factor about the company, the model is likely to overestimate the final probability of success, comparing to the entire population (counting the startups that have never inserted data into Crunchbase).

6. The number of observations for each type of degree of education (Undergraduate, Msc/Ma, MBA, and Ph.D.) and the number of milestones is not substantially large, so the results are valid to understand the relationship of related variables. However, the levels may not be accurate.

CHAPTER 4: MODEL ESTIMATION AND RESULTS

As explained in the previous chapters, this model utilizes the Logistic Regression as the benchmark technique for variable selection, due to its straightforward interpretation and statistical approach, it is possible to compare the estimated effect and the statistical significance of each risk factor. To accomplish the objective of this research, I need to be sure that the final model is predictive and that the estimated coefficients reflect the real impact on startup success. For this reason, it is fundamental to check whether all the variables are well specified, such that the results match the economic intuition, and there are previous researches that back the conclusions. This thesis is the first to study deeply the conceptual soundness of the coefficients and not focus only on predictability. After validating the variables and the reliability of the model, this research tests the predictability, comparing its accuracy with it of other well-known methods, in this case, XGBoosting, Supporting Vector Machines(SVM), and Linear Discrimination Analysis (LDA). After obtaining the final results, I compare them with those of other researches.

Following the detailed steps of this process:

1. Apply the Logistic regression using all hyper-parameters and their transformations that conceptually affect startup outcomes.

2. Understand the random effect of each variable, make the required modifications, and remove the statistically insignificant risk factors based on their p-value.

3. Apply all the classification methods on the final model using the training data.

4. Score the test data using the output of each classification method.

5. Calculate the performance of each classification method using Somer's D, Accuracy, Confusion Matrix, and AUC (Area under the curve.)

This thesis utilizes the Python programming language and the following libraries to run the analysis:

- Pandas, DateTime, NumPy, and Scipy for data manipulation.
- Sklearn, Statsmodels, and Xgboost for running the statistics and machine learning prediction methods.
- Matplotlib, Seaborn, and Graphviz for plotting the graphics.

4.1 Model Performance Metrics

The performance metrics are the following: Somers' D, Accuracy, Confusion Matrix, ROC curve, and AUC score.

4.1.0.1 Somers' D

The value of Somers' D Somers (1962) is defined as

$$D_{XY} = \frac{\tau_{XY}}{\tau_{XX}} \tag{4.1}$$

where τ_{XY} is the difference between the number of concordant (the predicted is equal to the observed) and discordant (the predicted is different than the observed) pairs and τ_{XX} is the total number of pairs. The higher is the value of Somers' D, the better is the performance of the model.

4.1.0.2 Accuracy

The accuracy represents the percent of the total estimations that were correct (the predicted are equal to the observed). Following the accuracy equation:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \tag{4.2}$$

Such as the Somers' D, the higher is the accuracy, the more predictive is the model.

4.1.0.3 Confusion Matrix

The Confusion matrix details number of true positives, true negatives, false positives, and false negatives. The following image perfectly illustrates it:

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 4.1: Confusion Matrix – *Source: Glass box artificial intelligence*

4.1.0.4 ROC curve - AUC score

The ROC curve (Receiver Operating Characteristic Curve) is a graph with the True positive rate in the Y-axis, and the False positive rate in the X-axis, the Area under this curve (Blue line in figure 2.3) is the AUC, the Redline in figure 2.3 represents the 0.5, every model with AUC above it shows that the model could be more predictive than randomly tossing a coin. As Kraus (2014) explains, Assuming two samples, of

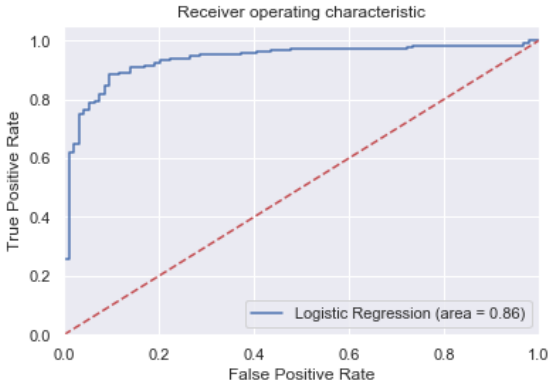

Figure 4.2: ROC – *Source: Sklearn - Python*

failures n_f and successes n_s, the possible scores from each sample correspond to:

$$S(x_f, x_s) = \begin{cases} 1 & \text{if} \quad x_f > x_s \\ 0.5 & \text{if} \quad x_f = x_s \\ 0 & \text{if} \quad x_f < x_s \end{cases} \tag{4.3}$$

where x_f and x_s are the scores from the failures and successes, respectively. By taking the average over the comparisons, the AUC can be written as following:

$$AUC = \frac{1}{n_f.n_s} \sum_{1}^{n_f} \sum_{1}^{n_s} S(x_f, x_s) \tag{4.4}$$

Such as all the metrics above, the result of the AUC score is expected to be higher for more predictive models.

4.1.0.5 Cross Validation

Cross-Validation is a method to test the sensibility of the model. In this research, I utilize the k-fold cross-validation, which randomly divides a data set into k disjoint folds with approximately equal size, and each fold is in turn used to test the model induced from the other $k-1$ folds by a classification algorithm. The performance of the classification algorithm is evaluated by the average of the k accuracies resulting from $k-fold$ cross-validation, and hence the level of averaging is assumed to be at fold Rodriguez et al. (2009). The performance metric used as the outcome of each fold is the ROC - AUC (Area Under the Curve) in this research. The range between the fold with lower and higher performance is an indication of how sensible the model is, a big range means that the model is not constant and could be poorly specified.

4.2 Model Selection

The model selection consists of the process of determining the combination of variables from the candidate pool that generates an output consistent with the economic

intuition of the variables and presents a good out-of-sample performance.

4.2.1 Removing Non-predictive Variables

After running the entire first proposed model using Logistic Regression, some variables demonstrated to be weak predictors of startup outcome based on their p-values (not significant in a 5% significance level). The method calculated heteroscedasticity-consistent (HC) standard errors to avoid including non-significant variables in the final model Cramer (2007). Those variables are GDP growth at the year of foundation, the situation of the industry (expansion, stagnation, and recession) in the years before the foundation, The state in which the company locates, the micro-region in the U.S where the company locates(the Silicon Valley area as the unique exemption), and most of the specific industries as individual factors. After removing those variables, the model demonstrated much better predictability in the testing data. It does not mean that those variables do not affect startup success in any way; it just means that in this model, they are not very good predictors. For future researches, it could be valuable to test those factors in a different approach and alone to understand their effect on startup outcomes.

4.2.2 Training and Testing data

The data is divided into two categories, training, and testing, accounting for 70%, and 30% of the observations, respectively. I used the training set to fit the model, while the test, also known as the holdout sample, is used to calculate the performances. In this research I also use an approach called k-fold cross-validation, that consists in fitting the model with n observations and testing with the remaining, then fitting again with $n + 1$ and testing with the remaining, I repeat until I have enough performances to evaluate the consistency of the model.

4.3 Applying Rejection Inference

After running the model with the optimum variables, most of the random effects were as expected. However, the variable "Amount in $ of Angel Investment" demonstrated a negative linear effect, which is in disagreement with the economic intuition and literature. According to the literature (details in the next section), the amount of angel investment should represent a convex effect on startup success. It means that some investment is better than none, but as I keep increasing the value, at a certain point (depending on the industry, target market, revenue stream), the company should stop getting more angel investment and go for Venture Capital rounds, which provide better support to bring the company to the next level. If the company keeps raising more angel investment, the return per scale starts decreasing.

The reason for a linear negative effect in this variable relies on the exclusion of observations with zero total investment (explained in Chapter 3). As I only have startups with some amount of investment, the startups that did not receive any angel capital only received venture capital rounds of investments, and, as explained in the next chapter, have a higher likelihood to succeed, concluding, the absence of observations with zero total investments biased the effect of angel investment.

The solution for this problem is to apply a correction for sample selection bias called Rejection Inference. As explained in Chapter 3, it adds observations with zero investment using a different threshold.

4.4 Check Logit Assumptions

4.4.1 Multicollinearity

Before analyzing the results of a Logistic Regression, it is important to check whether the model is free of perfect multicollinearity, which means that two or more variables are not perfectly correlated. Otherwise, all the results are invalid. I used the Variance Inflation Factor(VIF) for this step, which quantifies how much the vari-

ance inflates with each variable. When there is multicollinearity in the model, the standard errors, and, consequently, the variance, inflates. The "rule of thumb" for understanding whether there is a model presents multicollinearity differs from authors. According to Hair et al. (2006), the tolerance level is 10, after that collinearity is a problem, meanwhile, Rogerson (2001) considers the tolerance level 5 and Pan and Jackson (2008) even consider it 4.

In our model, before the addition of functions $F1 = (Education\ +\ Experience;)^2$ and $F2 = (Time\ to\ reach\ the\ first\ Milestone\ +\ Number\ of\ Milestones\ )^3$ all the VIF values were under 4, which are under the tolerance level. After the addition of the variables that compose the functions (F1 has 3 variables and F2 4 variables), the VIF factors of some variables composing the functions surpass 5, which the highest is of "2 x $Education$ x $Time$ $(Years)$ $from\ Graduation\ -2ab-$", with a value of 6.77. The cause of those VIF values higher than 5 is the format utilized to capture the effect of the variables; it does not imply multicollinearity, besides they are still lower than 10.

	VIF Factor	Variables
0	4.31355	Percent of companies before foundation(ind)
1	1.5535	Located at the Silicon Valley (Dummy)
2	1.48166	Industry: Business Services (Dummy)
3	1.08032	Industry: Social and Nonprofit (Dummy)
4	1.42706	Reached Venture Capital rounds (Dummy)
5	1.64307	Industry: Business use Technology (Dummy)
6	1.16296	Amount of Angel investment (in Millions) x N Participants
7	4.83431	F1: Time (Years) from Graduation #squared#
8	2.02884	F1: Education (0-4) #squared#
9	6.77073	F1: 2 x Education x Time (Years) from Graduation
10	2.97629	F2: Milestones cubic
11	1.84714	F2: Time (Months) to first milestone cubic
12	5.49895	F2: 3x Milestones squared x Time (Months) to first milestone
13	4.19376	F2: 3x Milestones x Time (Months) to first milestone squared
14	2.66551	Industry: Popular use Technology (Dummy)
15	1.45915	Total Funding rounds x Time (Years) to raise Investment

Figure 4.3: *Variance Inflation Factor - all variables from the final model*

4.4.2 Heteroskedasticity

The presence of Heteroskedasticity in the disturbances of an otherwise properly specified linear model leads to consistent but inefficient parameter estimates and inconsistent covariance matrix estimates, it could compromise the model selection process as we utilize the t-values to determine the variables to remove. To avoid this problem, I calculated the standard errors using an HC1 - Heteroskedasticity-consistent covariance matrix estimator. Cribari-Neto and Galvão (2003)

4.4.3 Independent observations

Each observation in the database has a unique identification number; each observation refers to a different startup, which is unique and not related to others in the dataset. In the data cleaning process, all the observations were verified and checked for possible repeated cases. The observations used in the final model are independent and represent the development of a unique company.

4.4.4 Linearity

Most of the variables are linearly related to the Log Odds of success of a startup. The variables that present a non-linear effect were modified to adapt to a Logistic Regression approach; it worked as the individual variables and functions utilized are either individually statistically significant or jointly significant, and their coefficients conceptually sound correct both theoretically and intuitively.

4.4.5 Sample Size

The final dataset has 1250 observations, in which 850 are for training, and 365 are for testing. As Bujang et al. (2018) suggests, I calculate the minimum number of observations in a training dataset to deliver the statistics that represent the parameters using the following formula $n = 100 + 50i$, where i represents the number of variables. The final model has 15 variables, so a reasonable number of observations would be 850, exactly the number of observations in our training dataset.

4.5 Interpreting the Coefficients

The final Logistic Regression presented the following output:

	Coef.	Std.Err.	z	P>|z|
Percent of companies before foundation(ind)	-0.0265	0.0026	-10.3552	0.0000
Located at the Silicon Valley (Dummy)	0.6909	0.1980	3.4895	0.0005
Industry: Business Services (Dummy)	0.5692	0.2764	2.0591	0.0395
Industry: Social and Nonprofit (Dummy)	-1.7373	0.8512	-2.0410	0.0412
Reached Venture Capital rounds (Dummy)	3.4313	2.4964	1.3745	0.1693
Industry: Business use Technology (Dummy)	0.4448	0.2927	1.5196	0.1286
Amount of Angel investment (in Millions) x N Participants	0.1452	0.0564	2.5751	0.0100
F1: Time (Years) from Graduation #squared#	-0.5967	0.2899	-2.0582	0.0396
F1: Education (0-4) #squared#	-0.0499	0.0587	-0.8499	0.3954
F1: 2 x Education x Time (Years) from Graduation	0.2702	0.1473	1.8347	0.0666
F2: Milestones cubic	0.0580	0.0190	3.0575	0.0022
F2: Time (Months) to first milestone cubic	-0.2008	0.0679	-2.9570	0.0031
F2: 3x Milestones squared x Time (Months) to first milestone	0.0990	0.0423	2.3423	0.0192
F2: 3x Milestones x Time (Months) to first milestone squared	0.2930	0.1043	2.8099	0.0050
Industry: Popular use Technology (Dummy)	-0.3702	0.2137	-1.7326	0.0832
Total Funding rounds x Time (Years) to raise Investment	8.9604	2.6065	3.4378	0.0006

Figure 4.4: *Output of Logistic Regression*

The signs of the regression coefficients are consistent with our expectations and previous empirical researches on the effect of each variable on start-up success/failure.

4.5.1 Function 1: (Education+ Time from graduation)^2

The expected effect of education on startup success is positive; high education could represent technical skills, general knowledge, good network, and high I.Q.; all those factors could influence the entrepreneurship process positively. In the model developed by Van Gelderen et al. (2005), the variable education, classified as either low or high, has a positive impact on startup success. In the study of Gimeno et al. (1997), the years of education the entrepreneur has gone through effects negatively the likelihood of the company being discontinued within the first five years after foundation.

Such as education, the expected effect of experience on the outcome of entrepreneurship is positive. It could indicate managerial expertise, industry know-how, network, personal savings to back the beginning of the company without giving up too much

equity to early investors, credibility, among other aspects. According to Van Gelderen et al. (2005), the overall experience (work, management, and expertise in starting a business) has a positive effect on the success of a startup.

As explained in Section 3.3.2- Transformations, the weight of experience on startup success depends on the type of education. "x" years of experience after finishing an undergraduate degree is not the same as "x" years of experience after a master, MBA, or Ph.D. To capture this behavior, I transformed those two variables in a quadratic function. The following graph represents the result:

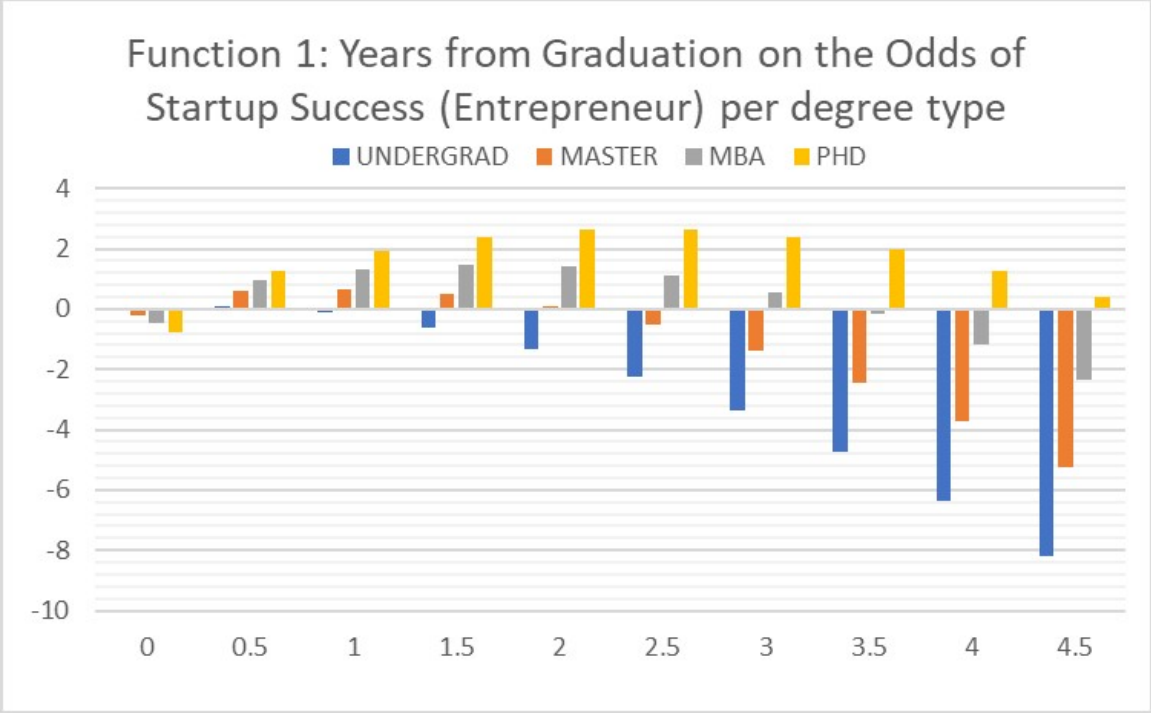

Figure 4.5: *Y = Effect of Function 1 on Odds of success, X = Years after graduation*

In the graph above, I can see that some experience after graduation is better than none for all the types of degrees, and each type has a different "optimum" amount of experience before entrepreneurship. Higher degrees require more experience to Increase the Odds of entrepreneurial success. As stated in section 3.6 -Limitations, I don't have enough observations of each degree type, and time from graduation to generate a very accurate assertion about the level of effects, however, the format of

them seems assertive.

4.5.2 Function 2: (Milestones + Time to first milestone)^3

According to the Cambridge dictionary, a milestone is an important event in the development or history of something or someone's life. In our database, milestones reflect positive news about the company with media coverage. A milestone is related to an actual event demonstrating the progress of the company. Most of the time, it happens when the company launches a new technology, opens new offices, reaches outstanding sales, or creates new partnerships. The effect of a new milestone is positive to the probability of success of a startup. In the model of Predictingthesucess, the number of milestones is one of the significant variables to predict startup success. However, the author does not state the signal of the coefficient on the paper.

The time (in months) between the foundation and the first milestone demonstrates how fast the startup reached the first accomplishment. My first intuition was that the effect of this variable was negative so that the faster a startup gets media coverage, the more likely the company would be to succeed. However, many researches say that the longer it takes to reach the first milestone, the more likely the company is to succeed.

Graham (2012) explains this effect:

> The growth of a successful startup usually has three phases: 1. there's an initial period of slow or no growth while the startup tries to figure out what it's doing. 2. As the startup figures out how to make something lots of people want and how to reach those people, there's a period of rapid growth. 3. Eventually, a successful startup is likely to grow into a big company; then this growth is likely to slow, partly due to internal limits and partly because the company is starting to bump up against the limits of the markets it serves.

it means that successful startups may take a longer time to mature and understand their businesses before starting to grow. This fact could also be related to the psychological aspects of the entrepreneur, more specifically, resilience. According to Masten (2001):"resilience is the aspect of psychological capital that involves coping and adapting to risks or adversity".

Intuitively, a resilient entrepreneur would not give up after a challenging beginning and instead, adapt, understanding the market and going successfully through the first phase of a successful startup mentioned above. The research of Baluku et al. (2016) confirms the assumption that resiliency is positively related to startup success.

As explained in Section 3.3.2- Transformations, our understanding is that the above researches are well structured so that combining those variables is likely to generate reliable results. The following graph represents it:

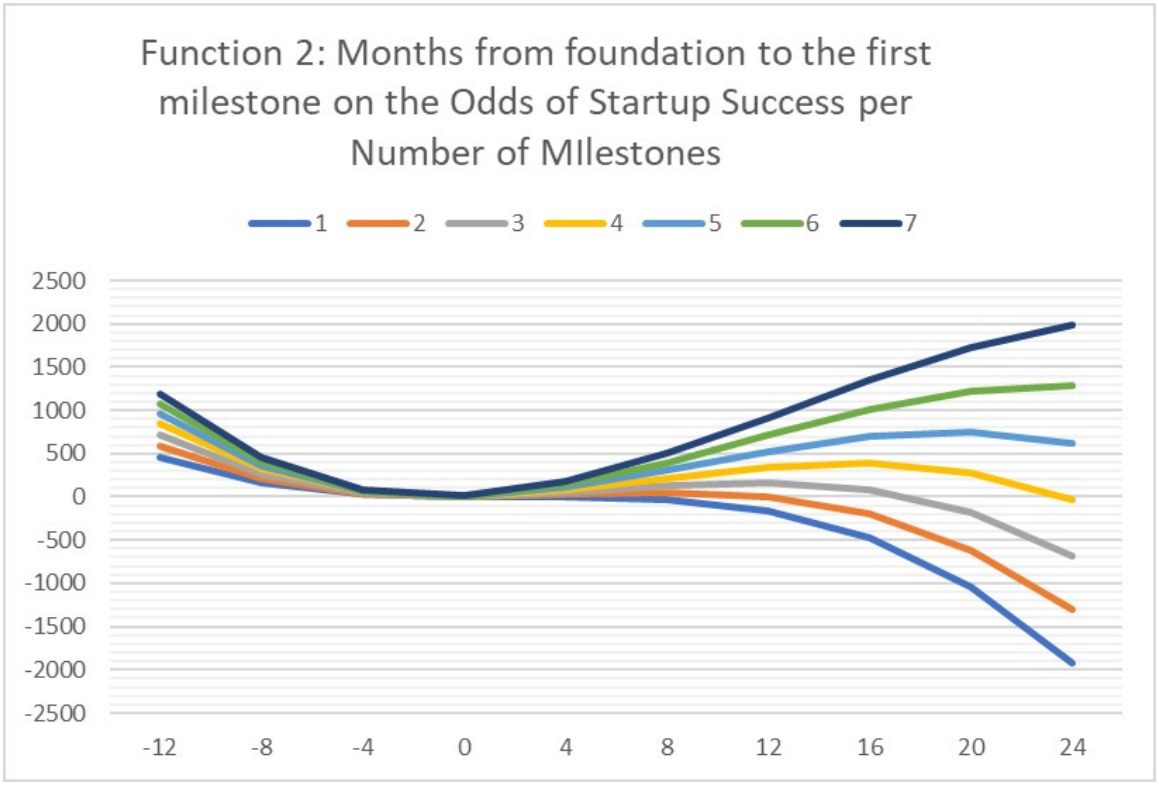

Figure 4.6: *Y = Effect of Function 2 on Odds of success, X = Months to reach the first milestone*

In the graph above, I can see that startups that get media coverage long before the foundation are more likely to succeed than ones that get in the short time after the foundation, no matter how many milestones those startups ended up getting. For those startups that get the first milestone after the foundation, the number of milestones obtained later determines whether the early obtained media coverage reflected pure speculation / paid advertising or real disruptive prospective. Startups that took longer to get the first milestones and ended up getting less than four is likely to be negatively impacted. The opposite happens with the companies that take longer to get the first and obtain more than 4 milestones, each additional milestone after the fourth also increases the upward slope of the function.

4.5.3 Percent of older companies in industry

This variable is a way to understand the competition outlook for each company. It represents the percentage of companies in the startup's industry founded in years before the year of the foundation of the startup. The idea is to proxy the size, market share, and know-how of the competitors. According to Finkelstein (2002), the first companies to enter a market have an advantage compared to the later entrants, as he explains that

> First mover advantage is nothing more than a competitive advantage that accrues to companies by virtue of their being the first entrant to a market.

Consequently, the sign of this variable should be negative, which means that the higher the percent of companies that existed before the foundation of a startup, the more rash environment it is likely to face, and consequently the less likely it is to succeed. After running the model, this variable presented the expected negative coefficient and was also statistically significant at one percent significance level.

4.5.4 Located at the Silicon Valley

After testing the effect of U.S. states, macro-regions (e.g. Mountains, Mid-Atlantic, New England, and the others), and micro-regions (N.Y., Texas, Silicon Valley, Raleigh, and other "startup hubs."), the only one statistically significant was Silicon Valley. I suppose that with more observations, New York City and Dallas would also be significant, as they are the other two most import startup regions in the U.S. As expected, the "Silicon Valley effect" is positive, and companies located there are 99% more likely to succeed. As Armour et al. (2004) explains, the environment created by a disruptively creative culture and the concentration of highly skilled and eager employees, investors, and creative entrepreneurs from all around the world attracts the most promising early-stage companies and boosts their odds of success.

4.5.5 Reached Venture Capital Rounds

This Dummy variable indicates that the company went through a successful round of venture capital investment. In my model, a startup that successfully reached venture capital rounds of investments were 30 times more likely to be successful. Even if this variable is not statistically significant, it adds predictability, and its coefficient matches the economic expectation. I also tested the variables Amount and Log of the amount of Venture capital investment. However, they were not as statistically significant and predictive as the dummy. As the amount of money invested varies according to the type of market and business structure, the amount of investment represents mostly the required capital to run the strategy . For this reason, those the amount of investment alone does not indicate any higher probability of success for a startup. The reason for considering venture capital investments as a success factor relies on the higher maturity of the companies usually approved in those rounds of investment.

4.5.6 Angel Investment (in millions) x Number of Investors

The product of Angel investment amount (in millions) and the number of angel investors has a positive effect on startup success. This variable is also statistically significant at 1% significance level. The reason for arranging those variables as a product and not separated bases is the relation between them; more investors tend to result in a greater amount of angel investment. As Wong (2002) explains:

> Since many angels are former entrepreneurs or industry executives, they may derive some private benefits from assisting in the development of a new firm. Two of the more identifiable ways of providing help is assistance in procuring a management team and procuring additional funds.

It means that the more angel investors, the more angel capital is raised. Another factor to consider is that as angel investors take a higher risk, they would only increase the amount of money invested or actively use their network to capture funds if the companies and their entrepreneurs present good perspectives. A startup with poor management and a low perspective would not keep attracting new investors or more investments from the previous investors.

4.5.7 Total Funding Rounds x Time to Raise Investment

The product of the total number of funding rounds and time in years to raise the investment is positive and statistically significant. The reason for arranging those variables as a product relies on their relationship. More funding rounds tend to increase the total time to raise investment. This variable captures two essential aspects of startup success. First is the idea that raising too much capital too fast is a risk factor because the company needs some time to adjust its products/services to the market and prepare a more efficient strategy. Too much initial capital would induce the entrepreneurs to burn too much money at a fast rate and lose equity, limiting future rounds, as Feinleib (2011) explains. Second is that new funding rounds

only happen if the investors still believe in the future of the company. When the circumstances indicates that a startup is going to fail, the investors are likely to stop investing in there.

4.5.8 Industries (Dummies)

The only four industries that are statistically significant and jointly significant in the sample are Business use technology (positive effect), Personal use technology (negative effect), Social/ Nonprofit (negative effect), and Business Services (positive effect). Those effects reflect the market behavior in the time interval used in this research (2000 - 2012).

Intuitively, the reason behind the difference in sign between Business use and Personal use technologies seems to be a term called "winner-takes-all." As Marmer et al. (2011) explains, it means that in some markets, especially with large scale standard products/services, which generally describes the personal use technology, the company that obtains more traction becomes a monopoly and the others leave the market (e.g., Google, Youtube, Facebook, Uber/Lyft). Business use technology tends to be more focused on niches and particular business problems, as there are many gaps in this market; there are spaces for many companies to succeed.

The Social/Nonprofit industries are more vulnerable to political and economic changes as, in most cases, they rely on government support and donations, generally easily affected by changes of government and financial crisis.

The Business services industry is heterogeneous and difficult to monopolize, composed of many different niches, products, and services. It can focus on either local, regional, national, or global businesses. The rising demand for IT and online services by the companies boosted opportunities for many startups. Another peculiar aspect of this industry is the likelihood of acquisitions, startups that offer business solutions to big companies are likely to be acquired by them.

4.6 Comparing Alternative Algorithms

4.6.1 Algorithms Performance

Performance			
Metrics	**Accuracy**	**AUC**	**Somers' D**
Logistic Regression	0.9068	0.9098	0.8137
Linear Discriminant Analysis	0.8274	0.8387	0.6548
Support Vector Machine	0.9068	0.9080	0.8137
Extreme Gradient Boosting	0.9014	0.8987	0.8027

After testing the fitted models in the out-of-sample data (30% of the total observations), the Logistic Regression, Support Vector Machine (using linear kernel), and Extreme Gradient Boosting presented a very similar performance, much better than that of Linear Discriminant Analysis. The results indicate that those three models have around 90% of accuracy, which is kept almost constant in the Area Under the Curve, this means that the percent of true positives and true negatives are constant with the accuracy. The Somers' D is over 0.8, which, considering that the Somers' D discounts the false positives and true negatives, demonstrates that the models significantly differentiates between successful and unsuccessful startups.

Analyzing the confusion matrix, the Logistic Regression predicts better than the others true negatives (Predicted label = 0 and Actual label =0) but under-performs the Support Vector Machine and the Extreme Gradient Boosting in predicting the true positives (Predicted label =1, Actual label = 1). Extreme Gradient Boosting classifies fewer startups as unsuccessful, and, consequently, predicts fewer unsuccessful observations, it means that this algorithm is less conservative for this model then Logistic Regression and Support Vector Machine is.

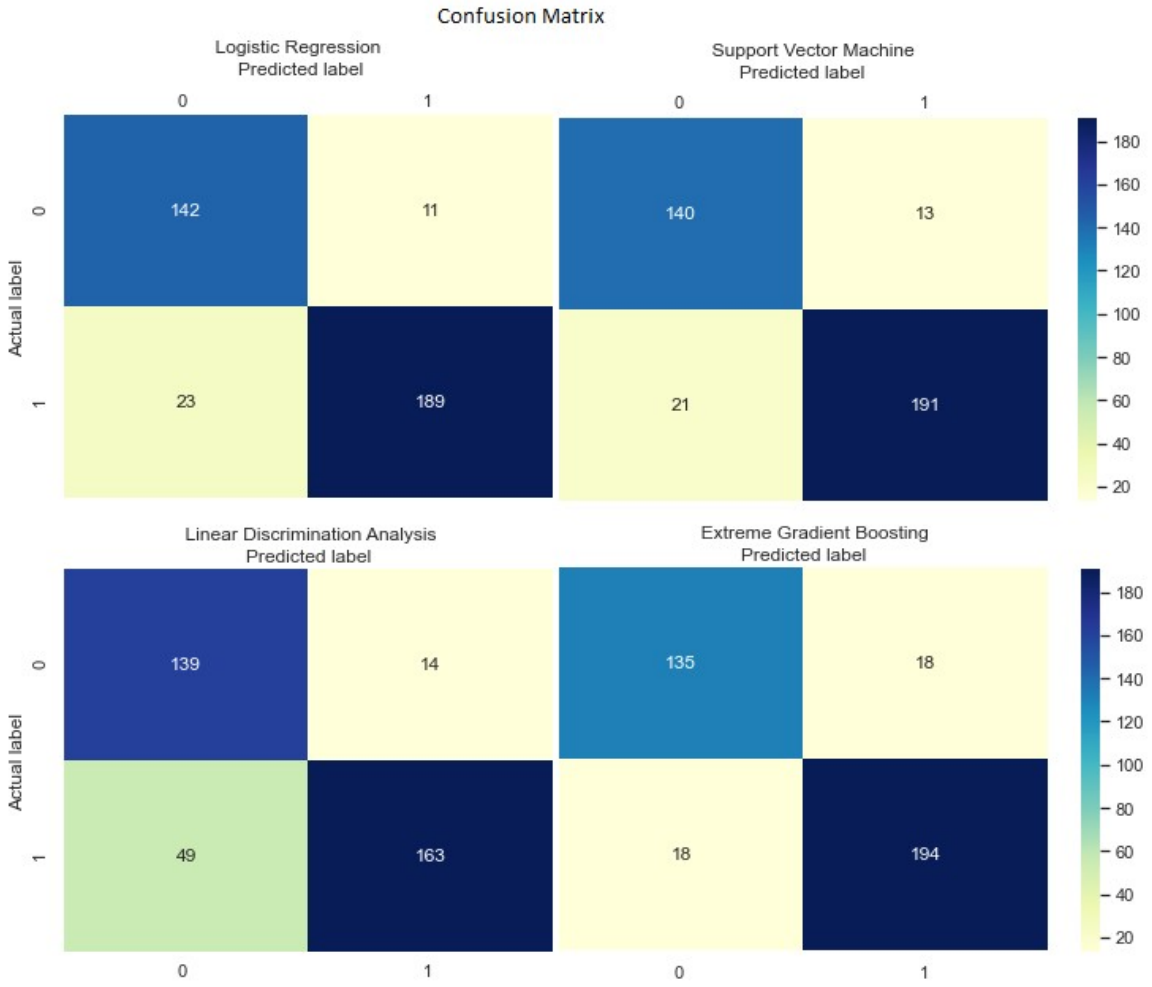

Figure 4.7: *Confusion Matrix*

4.6.2 Sensitivity Analysis

In the figure below, the boxes represent the values between 25% and 75% of the observations, the lines inside the boxes represent the average AUC of each fold of the Cross-validation (explained in section 4.1) and the lines above and under those boxes represent the highest and lowest performances of the folds. The objective of this analysis is to check whether the models generate a stable performance across the data set; the lower the range of those boxes, the more sensible and consistent the models are.

Even though the performance of the Logistic Regression, Support Vector Machine,

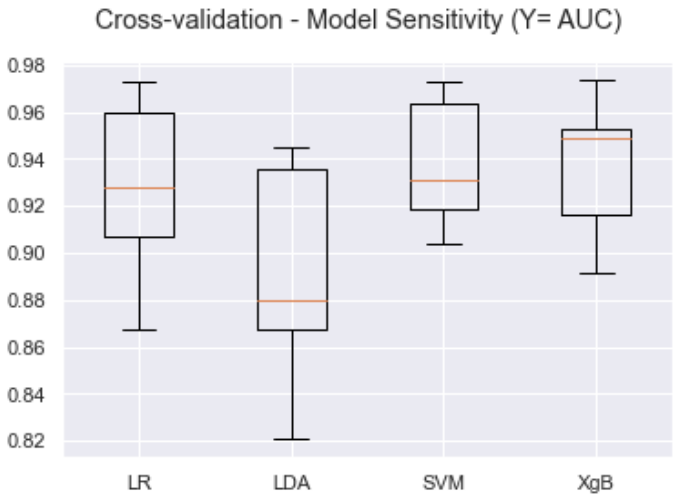

Figure 4.8: *Sensitivity analysis; LR: Logistic Regression, LDA: Linear Discriminant Analysis, SVM: Support Vector Machine, XgB: Extreme Gradient Boosting*

and Extreme Gradient Boosting, are very similar, the performances of Support Vector Machine deviate less from the mean, while the Extreme Gradient Boosting presents an average performance slightly superior. From a general perspective, these three models have acceptable sensibility, the lowest AUC value of the three is over 0.86.

4.6.3 Feature Importance

Linear Discrimination Analysis does not perform well and Extreme Gradient Boosting, as a decision tree, does not provide well defined coefficients. For those reasons, I only analyze the feature importance of the Logistic Regression and Support Vector Machine to the Odds of startup success. As the Logistic Regression reports Log Odds, it is necessary to take the exponent of the coefficients to see the effects on the Odds.

The method utilized to develop the bar chart above is basically multiplying the coefficients by the average non-zero value of each variable (in the Logistic Regression I use the exponent of the coefficient to convert Log Odds into Odds.)

In both algorithms, by magnitude, the *Function* 2 is the most critical variable and *Reached Venture Capital Rounds* (*Dummy*) is the second most important. The Support Vector Machine relies much more on *Function* 2 as all the other positive

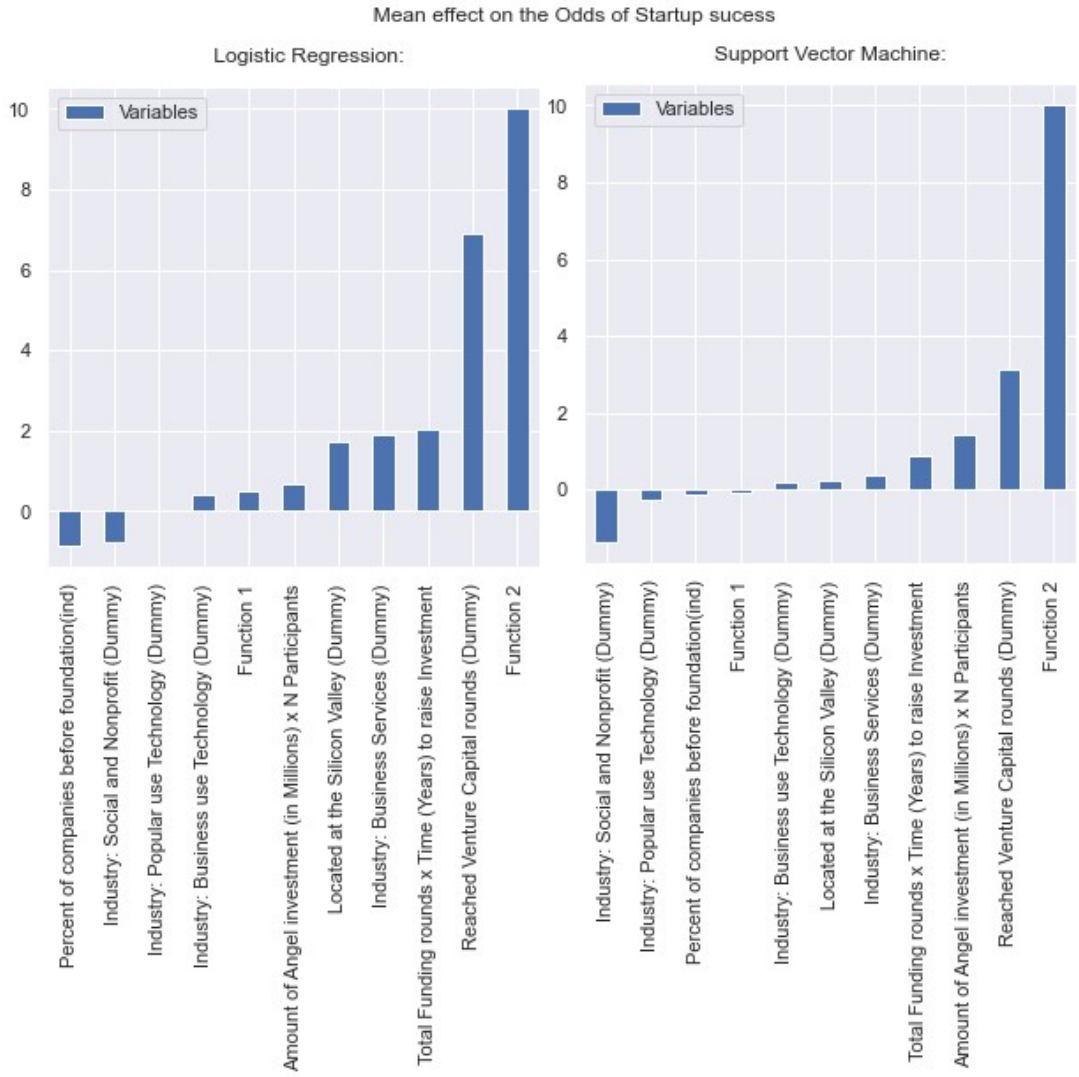

Figure 4.9: *Feature importance to the odds of startup success; Logistic Regression and Support Vector Machine*

variables that have the same sign effect in the Logistic Regression have a lower value compared to *Function* 2. Another interesting difference is that in Support Vector Machine, the effect of $Function\ 1 : (Education + Years\ from\ graduation)^2$ is negative, which matches neither our references nor the economic intuition explained in Section 4.5.1.

4.7 Advantages of Logistic Regression

A good scorecard should be predictive, interpretative, conservative, and trustworthy. Linear Discriminant Analysis is not very predictive and under-performs the other algorithms by a significant amount; for this reason, it is not worth to use it in the final model. Extreme Gradient Boosting is very predictive, slightly out-performing the others in the Cross-Validation analysis. However, it is not conservative as it classifies considerably more startups as successful than the other models, holding similar results in the performance metrics. Because it is a type of Decision tree, it does not provide well-defined coefficients, making it difficult to study the effect of each variable. As one of the methods considered a "black box" because of the lack of transparency, only an outstanding performance would make it worth to use this algorithm in the final model, which is not necessarily true in our research as the Logistic Regression and the Support Vector Machines have similar performance and more conservative outcomes. Support Vector Machine (with the linear kernel) is predictive and more transparent than the Extreme Gradient Boosting. Because it uses a linear approach, it is possible to obtain a coefficient for each variable. However, the Support Vector Machine does not provide indicators of statistically significance, depends heavily on the variable "Function 2", and calculates a misleading effect of education on startup success. For these reasons, it is not the winner algorithm on this thesis.

The Logistic Regression, as explained in Chapter 2, provides a robust analysis of the coefficients and their statistical significance, and generates predictive results. For those reasons, it is the most utilized method to create scorecards to predict credit defaults, using almost the same methods applied in this thesis. In this research, the Logistic Regression presented an excellent performance, predicting well the successes and out-performing the other algorithms in predicting the failures, it means that it generates more conservative results. All the coefficients of the Logistic Regression matched with economic intuition and literature. I was also able to check whether all

the variables were individually or jointly statistically significant. Besides a slightly weaker stability (bigger range in the Cross-Validation analysis), Logistic Regression is still the safest and most reliable choice to use in the final model.

4.8 Predictability Benchmark With Other Researches

Performance of different researches		
Research	**Best algorithm**	**Best AUC**
Felipe Veloso, 2020	Logistic Regression	0.933 (Average CV)
Krishna et al. (2016)	AD Trees	0.972 (Highest CV)
Bento (2018)	Random Forest	0.932
Ünal (2019)	Extreme Gradient Boosting	0.929
Shah (2019)	Logistic Regression	0.810

The models above utilize the same Database (Crunchbase), but different time ranges, classifications of success, data selection, algorithms, and objective of analysis. The research of Krishna et al. (2016) considers successful the startups operating, and failures startups closed and acquired. This research has a data mining purpose, testing six different algorithms, adding variables, and reporting the performance (AUC), without evaluating the effects of each variable. Krishna et al. (2016) utilizes the highest AUC generated among all the Cross-Validation iterations as the final model predictability. In this research, the highest AUC generated in the Cross-Validation is also around 0.97, however, I used the average AUC to compare with other models because it reflects with more precision the expected predictability.

The research of Bento (2018) considers a startup successful as those with an IPO and M&A (Merger and Acquisition), the focus is data mining, testing many algorithms, mainly focused on performance, not interpretation of variables. Bento (2018)

analyzes the variables by their increment in the performance of the model, without explaining the sign and size of their coefficients.

The research of Ünal (2019) considers successful startups those either operating, acquired, or with an IPO. It only considers closed startups as failures. It tests more than six different algorithms and variables from the financial perspective, marketing, and industry, without any significant data transformation. Ünal (2019) utilizes Logistic Regression coefficients to make the data selection, after selecting the final model, compares the predictability of different algorithms.

The research of **?** consists of a short model published in the SAS Forum. It considers startups successful as those operating, while failures are those closed or acquired. This research focuses on the steps to create the model, applying many different statistical and machine learning methods. It offers some insights about the effects of each industry, without analyzing the coefficients of the other variables.

CHAPTER 5: SUMMARY AND FUTURE RESEARCH

This research develops a systematic model using Logistic Regression to predict startups that will potentially succeed, consequently applying other statistical and machine learning algorithms to test the adequacy and applicability of Logistic Regression.

This thesis analyzes the effects of the risk and success factors on startup outcome, making variable transformations when necessary. I discuss the theory and economic intuition of each variable, transformation, and iteration.

The data selection is based on statistical and logical approaches to ensure that the model correctly specifies the dependent and independent variables, such that the results can be interpreted, satisfying the economic intuition regarding the effects of the independent variables on startup outcome. All variables in the final model are both economic and statistically significant. I also determined training and testing data sets, including observations without successful funding rounds to correct selection bias, in a process called rejection inference.

I estimated four models, Logistic Regression, Linear Discriminant Analysis, Extreme Gradient Boosting (improvement of the decision tree), and Support Vector Machine with linear Kernel. To compare their performance, I utilized the following metrics: AUC, Accuracy, Sommers'D, and confusion matrix. To check the sensitivity of the models, I utilized the Cross-Validation, dividing the data in folders and computing the AUC generated by fitting each iteration, the idea is to check if the model is constant across the data set. I found that the Logistic Regression, Support Vector Machine, and Extreme Gradient Boosting have very similar performance in the 30% testing data set, generating AUCs around 0.93. However, Logistic Regression presented more reliable coefficients and more conservative results, it is also less complex

and more utilized to develop scorecards.

I developed a consistent model, statistically robust and conceptually sounding; the model is also easy to understand and replicate. However, certain aspects can be enhanced to improve the model's prediction and sensibility, capturing other risk factors that could interfere in startup outcomes. I want to propose the following enhancements:

1. Adding macroeconomic variables that capture the effects of the national and international economy on startup success, I tried GDP, but it was not statistically significant.

2. Including startup data from other databases, such as Pitchbook, to generate enough observations to analyze the effects of other micro-regions, such as New York and Dallas, on startup success.

3. Creating a variable that captures the effect of subjective aspects of the entrepreneur on startup success, such as motivation, leadership, personality, and resilience.

4. Creating a variable that captures the financial aspects of the entrepreneur on startup success, such as credit score, past years' income, and other aspects that could reflect how the entrepreneur deal with her/his private finances.

5. Utilizing the number of patents and universities per state/ city to check whether research/ education hubs are more likely to generate more successful startups.

REFERENCES

Armour, J., Cumming, D., et al. (2004). *The legal road to replicating Silicon Valley*. Citeseer.

Baluku, M. M., Kikooma, J. F., and Kibanja, G. M. (2016). Psychological capital and the startup capital–entrepreneurial success relationship. *Journal of Small Business & Entrepreneurship*, 28(1):27–54.

Bento, F. R. d. S. R. (2018). *Predicting start-up success with machine learning*. PhD thesis.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bujang, M. A., Saat, N., Bakar, T. A., et al. (2018). Sample size guidelines for logistic regression from observational studies with large population. *The Malaysian Journal of Medical Sciences*, 25(4):122.

Burges, C. J., Smola, A. J., and Scholkopf, B. (1999). Advances in kernel methods. *Support Vector Learning*.

Campbell, D. and Hulme, R. (2001). The winner-takes-all economy. *The McKinsey Quarterly*, (1):82–93.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA. ACM.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cramer, J. S. (2007). Robustness of logit analysis: Unobserved heterogeneity and misspecified disturbances. *Oxford Bulletin of Economics and Statistics*, 69(4):545–555.

Cribari-Neto, F. and Galvão, N. M. (2003). A class of improved heteroskedasticity-consistent covariance matrix estimators. *Communications in Statistics-Theory and Methods*, 32(10):1951–1980.

Feinleib, D. (2011). *Why Startups Fail*. Apress. p.175.

Finkelstein, S. (2002). First-mover advantage for internet startups: Myth or reality. *Handbook of business strategy*, 2002:39–46.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–156.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Gimeno, J., Folta, T. B., Cooper, A. C., and Woo, C. Y. (1997). Survival of the fittest? entrepreneurial human capital and the persistence of underperforming firms. *Administrative science quarterly*, pages 750–783.

Graham, P. (2012). Startup= growth. *Article source: http://www. paulgraham. com/growth. html, accessed October*.

Guo, B., Lou, Y., and Pérez-Castrillo, D. (2015). Investment, duration, and exit strategies for corporate and independent venture capital-backed start-ups. *Journal of Economics & Management Strategy*, 24(2):415–455.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006). Multivariate data analysis (vol. 6) pearson prentice hall upper saddle river.

Hauskrecht, M. (2019). Decision trees. *https://people. cs. pitt. edu/milos/courses/cs2750-Spring03/lectures/class19. pdf*.

Hellmann, T. and Thiele, V. (2015). Friends or foes? the interrelationship between angel and venture capital markets. *Journal of Financial Economics*, 115(3):639–653.

Huang, B. G. (2016). Predict startup success using network analysis and machine learning techniques. CS224 Stanford University.

Huang, J., Li, K., Yang, J., and Yu, X. (2019). Developing a credit spread trading strategy using tree-based machine learning methods.

Kennedy, K. (2013). Credit scoring using machine learning.

Kenney, M. and Von Burg, U. (1999). Technology, entrepreneurship and path dependence: industrial clustering in silicon valley and route 128. *Industrial and Corporate Change*, 8(1):67–103.

Krajcik, V. and Formanek, I. (2015). Regional startup ecosystem. *European Business & Management*, 1(2):14–18.

Kraus, A. (2014). *Recent methods from statistics and machine learning for credit scoring*. PhD thesis, University LMU Munich.

Krishna, A., Agrawal, A., and Choudhary, A. (2016). Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 798–805. IEEE.

Lussier, R. N. (1996). A business success versus failure prediction model for service industries. (number:2).

Marmer, M., Herrmann, B. L., Dogrultan, E., Berman, R., Eesley, C., and Blank, S. (2011). Startup genome report extra: Premature scaling. *Startup Genome*, 10:1–56.

Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American psychologist*, 56(3):227.

Mok, J.-M. (2009). Reject inference in credit scoring. *Amsterdam: BMI paper*.

Pan, Y. and Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in u.s adult males. *Epidemiology & Infection*, 136(3):421–431.

Porter, M. E. and Strategy, C. (1980). Techniques for analyzing industries and competitors. *Competitive Strategy. New York: Free*.

Roach, M. and Sauermann, H. (2015). Founder or joiner? the role of preferences and context in shaping different entrepreneurial interests. *Management Science*, 61(9):2160–2184.

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575.

Rogerson, P. A. (2001). Data reduction: factor analysis and cluster analysis. *Statistical methods for geography*, pages 192–197.

Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., and Gotlib, I. H. (2015). Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in psychiatry*, 6:21.

Sergiu Luca, D. (2018). Evaluation of different approaches to reject inference: a case study in credit risk. sas - paper 2731-2018.

Shah, V. M. (2019). Predicting the sucess of a startup company, 3878-2019. *Oklahoma State University*.

Solomon, G., Dickson, P. H., Solomon, G. T., and Weaver, K. M. (2008). Entrepreneurial selection and success: does education matter? *Journal of small business and enterprise development*.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American sociological review*, pages 799–811.

Ünal, C. (2019). Searching for a unicorn: A machine learning approach towards startup success prediction. Master's thesis, Humboldt-Universität zu Berlin.

Van Gelderen, M., Thurik, R., and Bosma, N. (2005). Success and risk factors in the pre-startup phase. *Small Business Economics*, 24(4):365–380.

Vapnik, V. (1998). Statistical learning theory. hoboken. *Wiley. Wang, K., Tsung, F.(2007). Run-to-run Process Adjust. using Categ. Obs. J. Qual. Technol.*, 39(4):312.

Wong, A. Y. (2002). Angel finance: the other venture capital. *Available at SSRN 941228*.

Yrle, A. C., Hartman, S. J., and Yrle-Fryou, A. R. (2000). Economic factors: Examining why small businesses fail. *Journal of Business and Entrepreneurship*, 12(3):67.