Paper 3878-2019

Predicting the success of a startup company

Vrushank Shah, Oklahoma State University; Dr Miriam Mcgaugh,Oklahoma State University

ABSTRACT

More than 50% of startup companies fail in the initial four years. Further, three out of every four venture-backed firms fail. The algorithm proposed in this paper will help to predict the success of a startup company based on financial and managerial variables. This prediction will help investors to get an idea whether the investing in a startup will be successful or not? Apart from implementing a model consisting of all the factors mentioned below and predicting the success of a startup company, various other models will be created representing various milestones achieved by the company. This paper will help startup companies to know which factors are essential for getting an investment. The algorithm will be based on more than 15,000 companies' data collected from crunchbase.com. The financial variables include: investments in each funding rounds, valuation after each round of funding, current market value, total funds, investments and acquisitions by the company, financial background of key people and the managerial variables includes: Number of employees, competitors, location, age of the company, founders background, burn rate and various news articles on the company scrapped from internet. A variety of methods will be used to determine the best model such as random forest, text parsing, logistic regression, decision tree and survival analysis.

INTRODUCTION

According to Adora Cheung, co-founder and CEO of Homejoy," Startup is a state of mind, it's when people join your company and are still making the explicit decision to forgo stability in exchange for the promise of tremendous growth and the excitement of making an immediate impact." More than 100 million startups are launched per year, which is about 3 startups per second. But more than 50% of startups fail in the initial four years. There are various reasons for a startup to fail for example lack of focus, raising too much money too soon, lack of general and domain-specific business knowledge, etc. There are very few studies which are performed to understand the reasons for the success of a startup company. There are various articles which describe the reason for failure for the startup companies but without the backing of data. This paper tries to create an accurate predictive model to predict whether a startup firm will succeed or fail.

The data for this paper was taken from crunchbase.com. More than 15,000 companies' data was analyzed and used to build a model. All the companies that started between years 2000-2014 were used in this paper. The key factors included the amount of seed funding, the time taken for seed funding, the company's valuation and other managerial variables. The results were explained using the survival model and logistic regression.

KEY FACTORS AND DATA EXPLORATION

1. **Seed funding:** It is a form of initial investment by an investor in a company in exchange for an equity stake in the company. This is the most important stage for a startup firm.

2. **Series funding:** After seed funding is achieved the startup companies start getting series funding. There could be series A,B,C,D,E,F and G fundings. Here the

alphabets correspond with the development stage of the companies that are raising the capital.

3. **Rounds of funding:** Total rounds of funding received by a company.

4. **Time to get seed funding:** Getting seed funding is very essential for a startup firm but no company will like to wait for months to get its initial funding. The more time it takes for a company to get the initial funding the value of its product decreases. This parameter measures the number of months required for a startup company to get seed funding.

5. **Valuation after each round of funding:** Valuation after seed funding is calculated using the formula: (100*(Seed amount)/15). Valuation after series A funding is calculated using the (100*(Series A amount)/8). Valuation after series B,C,D,E,F,G is calculated using the formula: (100*(Amount)/5).

6. **Number of Milestones:** Number of achievements by the company as per the company.

7. **Average time taken to achieve each milestone:** This parameter gives the average months taken to achieve each milestone. If the number is large then it is considered bad as it indicates that a company is taking a longer time to achieve milestones.

8. **Average time taken to achieve funding:** Number of months taken to receive each round of funding.

9. **Region:** The city where the company is located.

10. **Degree:** The highest education completed by the core-committee of the company.

11. **University:** The University from which the highest education was completed by the core-committee members.

12. **BurnRate:** Amount of time taken by the company to burn all its funds. It is calculated using the formula: Total fund/Number of Months Company was active.

13. **Total funding:** The total amount of money received by the company.

14. **Category_code:** The domain of the startup company.

DATA EXPLORATION

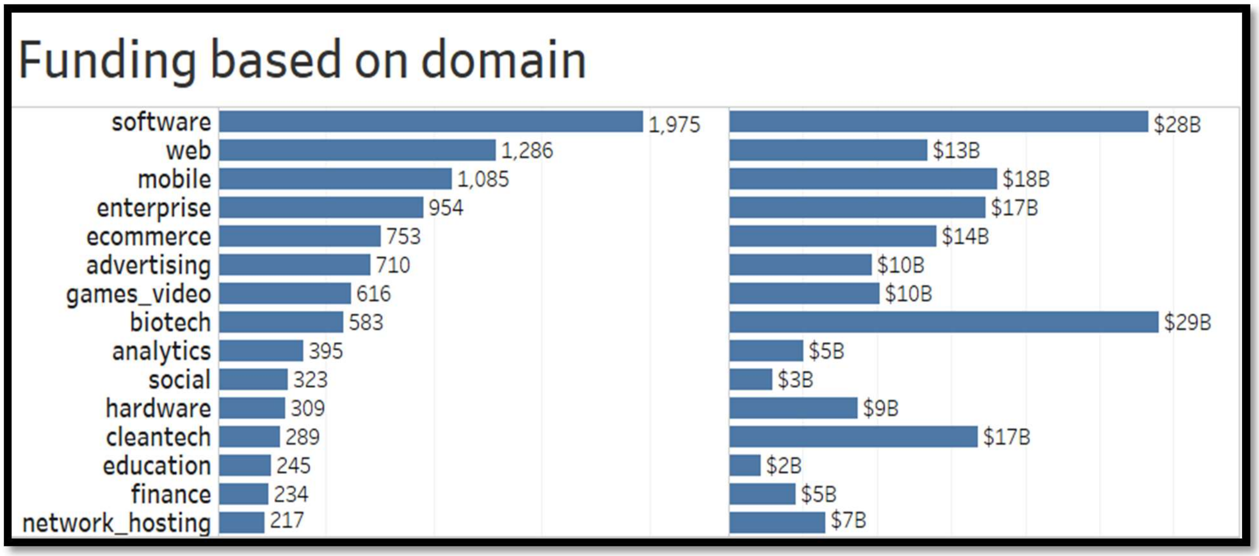
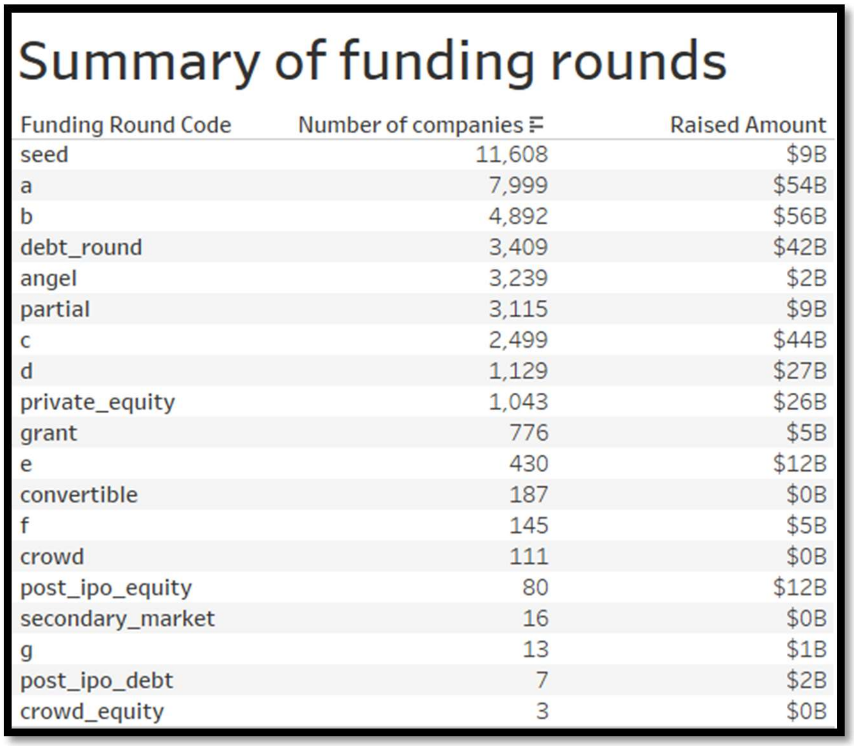


Figure 1. Funding based on domain

Figure 1 shows the relationship between number of companies under each domain and total amount of funding received by that particular domain. The highest number of startup firms were under software domain but the highest funding was achieved by companies under biotech domain.



Summary of funding rounds

Funding Round Code	Number of companies	Raised Amount
seed	11,608	$9B
a	7,999	$54B
b	4,892	$56B
debt_round	3,409	$42B
angel	3,239	$2B
partial	3,115	$9B
c	2,499	$44B
d	1,129	$27B
private_equity	1,043	$26B
grant	776	$5B
e	430	$12B
convertible	187	$0B
f	145	$5B
crowd	111	$0B
post_ipo_equity	80	$12B
secondary_market	16	$0B
g	13	$1B
post_ipo_debt	7	$2B
crowd_equity	3	$0B

Figure 2. Total funding received under each round of funding

The above figure illustrates the trend of funding raised for each rounds. According to the table, 11,608 companies were able to achieve seed funding and 7,999 companies were able to make it to series A funding. The highest amount was raised for series B funding. As the company develops, the amount of funding starts increasing hence the amount for seed funding is less and from series A onwards the total raised amount is more even though the number of companies decreases.

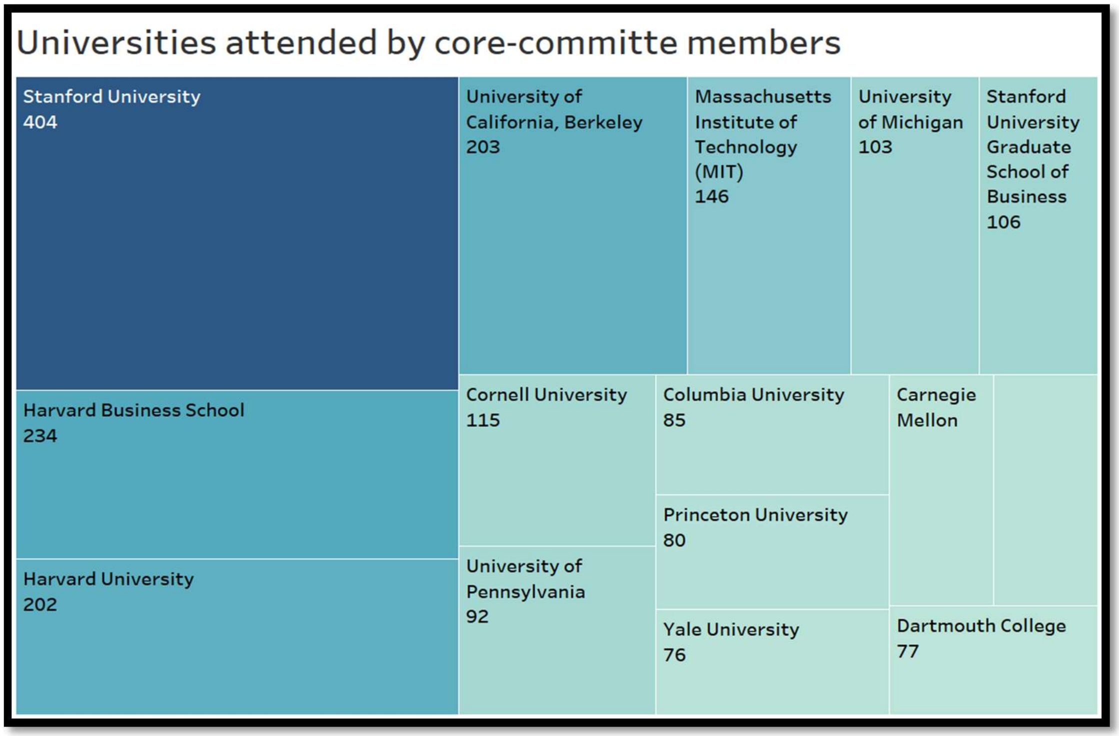


Figure 2. Universities attended by core-committee members

The core-committee members of a company includes CEO, CFO, CTO, Founders and Co-Founder. The chart gives the universities where these people had received their education. Most of the members of the core-committee were from IVY league colleges.

EFFECT OF THE DEGREE ON THE SUCCESS OF THE COMPANY

Survival Analysis: Survival analysis is a statistical method used to analyze data where the outcome variable is the time until the occurrence of an event. The event means either the death, occurrence of disease, churn,etc. The time to the event can be measured in days, weeks, months and years. The test is conducted for a specific period and subjects are kept under focus till the event occurs or the end of the period is reached. The observations are called censored when the survival time of these observations is not known during the course of the test.

In this section, analysis was done to find whether there is an association between the educational degree of the core-committee people and the survival of the company using survival analysis. The time period of the test was for 60 months. And the event occurred if a company had completely shut down. The survival analysis was done for seven major domains which had large numbers of startup companies. The seven domains are: Advertising, E-commerce, Enterprise, Analytics, Software, Mobile, Video and web-domains. Some of the examples of the resulting survival analysis probabilities have been shown below:

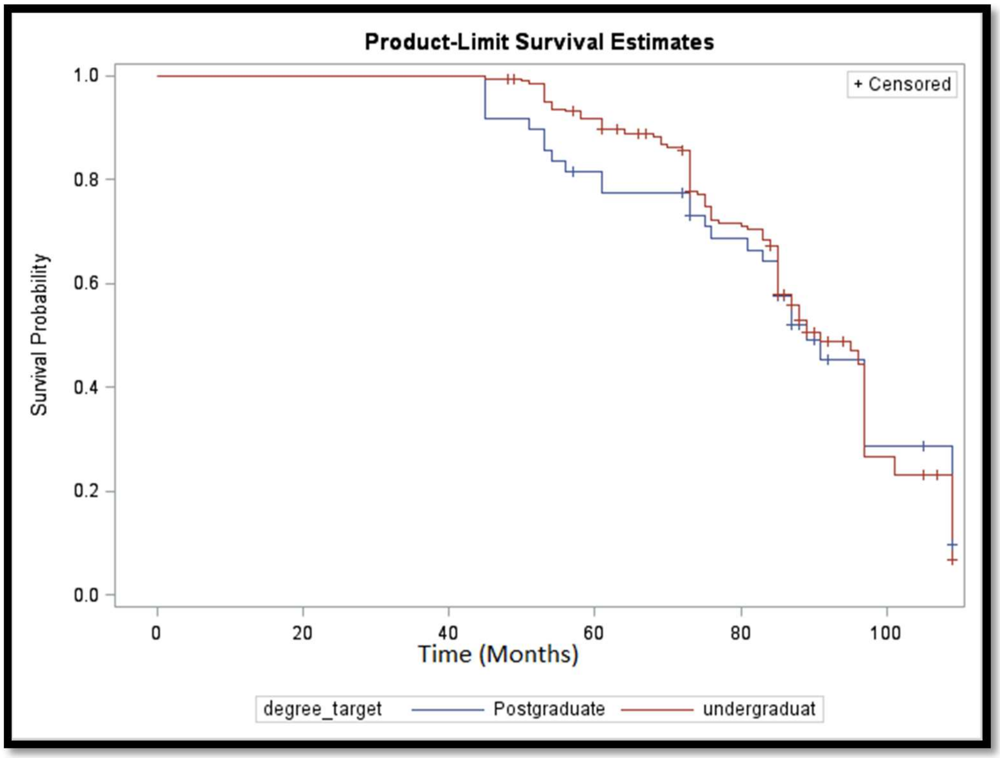


Figure 3. Survival Analysis in Advertising domain

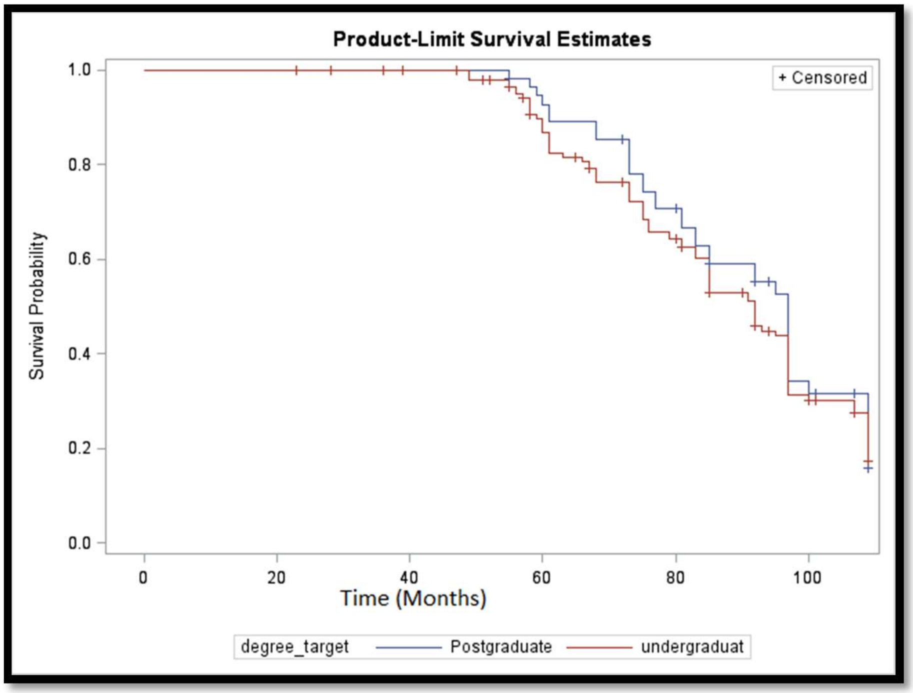


Figure 4. Survival Analysis in Mobile domain

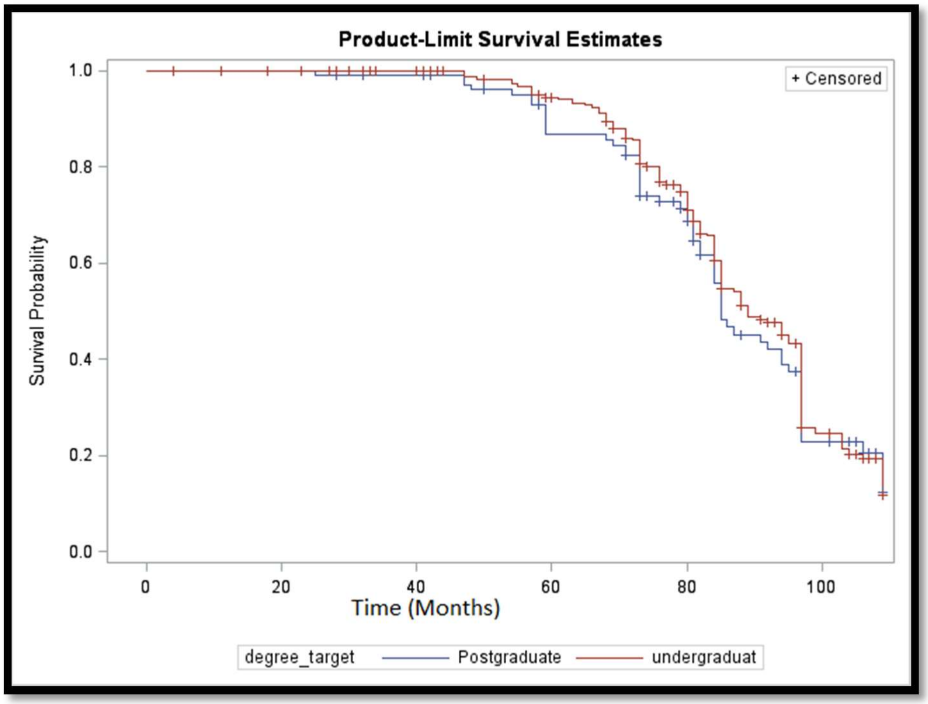


Figure 5. Survival Analysis in Web domain

Figures 3, 4 and 5 help to show how the degrees of core-committee people affect the success of the company. In the above graphs a red line means core-committee people have a postgraduate degree which includes Ph.D., MS, MBA, MD, etc and the blue line indicates core-committee people have an undergraduate degree which includes BS, BE, BA, etc. As the time period increases the survival probability for the company decreases. There is a clear difference between red line and blue line on the graph which means the success of a company does depend on the degree of the core-committee people.

Summary of Survival Analysis for all the domains:

Category	Survival chances for Post-graduate	Survival chances for Under-graduate
Advertising	70% ↓	90% ↑
Analytics	80% ↑	60% ↓
E-commerce	95% ↑	80% ↓
Enterprise	95% ↑	85% ↓
Video	95% ↑	85% ↓
Mobile	90% ↑	80% ↓
Software	80%	80%
Web	85% ↓	95% ↑

Table 1. Summary of survival Analysis for each domain

According to the above table, most of the successful domains have post-graduate people in the core-committee. Thus, having post-graduate core-committee people helps to get more funding and also leads to high survival chances of the company.

PREDICTIVE MODELING

In this section, we will use all the variables mentioned in the key factors section and build a supervised classification model. There are various techniques available to build a binary classifier model such as decision trees, logistic model, neural network, etc. The predictive model will help us to understand which key factors are essential for a company to be successful. In this paper we have made used of neural network and logistic regression model.

Neural network models are the part of the machine learning literature. They work like biological neural networks. Neural networks are approximation functions hence they can map any complex input to output space. It does not compute sequentially like other machine learning algorithm but it has a style of parallel computation.

Logistic regression is a machine learning model that deals with binary target variables. It helps to explain the relationship between the binary target variable and the nominal, interval ,ordinal or ratio level independent variables. Its quite easy to explain the results of logistic model to non-technical people.

DATA PREPARATION

After performing initial data exploration variables, total_funds and total_valuation had high skewness and kurtosis. A log transformation was performed on total_funds and total_valuation to decrease the skewness and kurtosis. The companies that were started between year 2000-2014 were used for analysis. Also, only those companies who received first round funding were examined. A target variable was created and 1 was assigned to the companies which are closed or acquired and 0 was assigned to the companies which were still operating.

PROCESS FLOW

SAS® Enterprise Miner was used to build the predictive model. A stratified sampling was performed on the target variable. Data was divided into 60% training, 20% validation and 20% testing dataset. A transform node was used to do log transformation on total_funds and total_valuation variables. A stepwise regression model was run on the key factors variables and neural network was also run on the same variables. SAS Enterprise Miner has a model selection node that selects the best model based on the value of a single statistic. For this analysis, we will use ROC index as our selection criteria. ROC curve is one of the common method used to measure the performance of a logistic model.ROC plots true positive rate against false positive rate. And then the area under the curve is measured to check the binary classification of our logistic model. A perfect ROC index is 1.

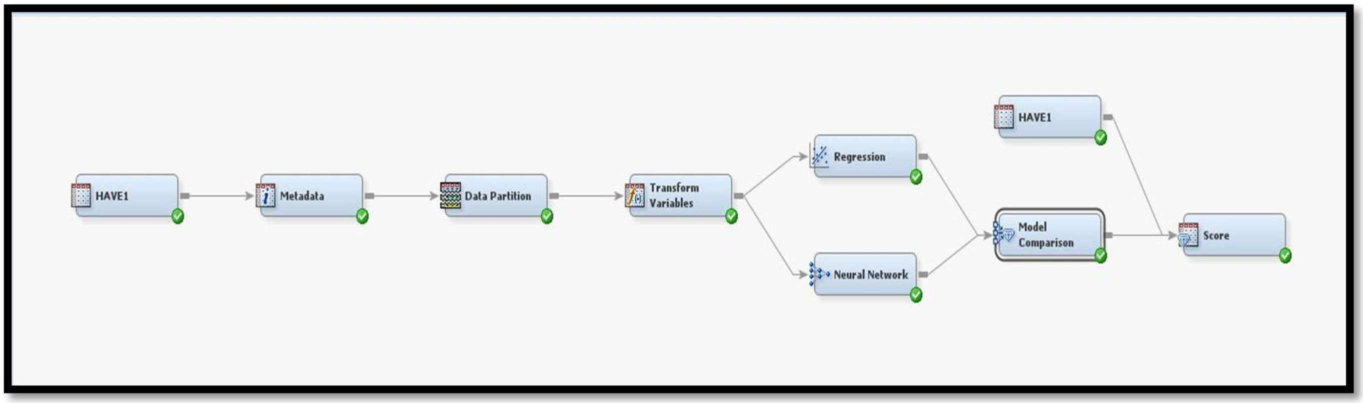


Figure 6. Process Flow

RESULTS

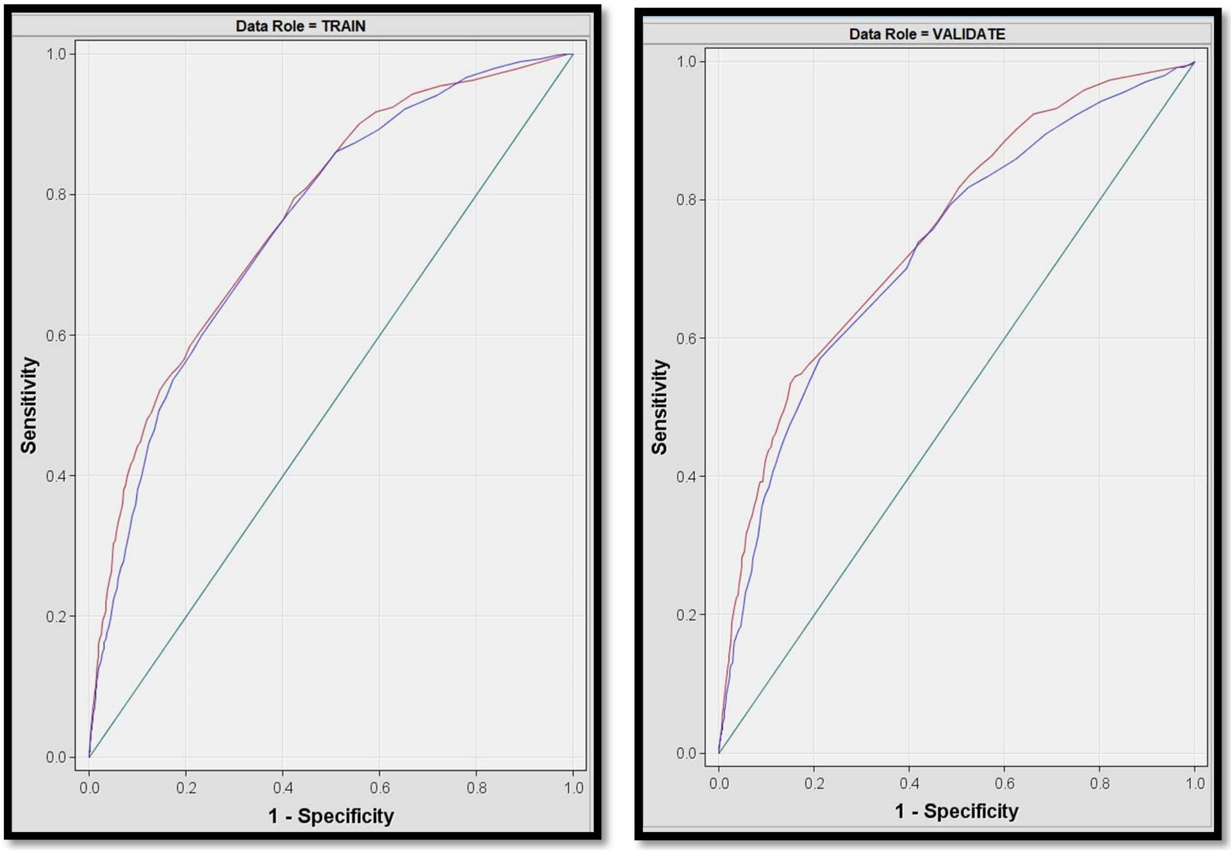


Figure 7. ROC Curve

Based on the ROC index Logistic Model was selected.It had 0.81 ROC index. A perfect ROC index is 1. Hence our ROC index of 0.81 means our model is performing well.

SIGNIFICANT VARIABLES:

According to logistic model all the variables except total funding were significant.

The significant variables are : Burnrate, Total Valuation, Total number of Milestones, Average days between each Milestone, Total Funding rounds, Average Days between each Funding Rounds, Time to Get Seed Funding, Domain and Location.

CONFUSION MATRIX FOR LOGISTIC MODEL (VALIDATION DATA):

		Actual	
		Survived	Failed
Predicted	Survived	1125 (True Positive)	92 (False Positive)
	Failed	348 (False Negative)	1553 (True Negative)

Table 2. Confusion Matrix

Confusion matrix helps to describe the performance of our logistic model. Based on the above table we can calculate three important metrics:

Accuracy: The accuracy of the model is 85.9%.

Sensitivity: Sensitivity indicates how often our model designated a company survived when it actually did survive. Our model's sensitivity is 76%.

Specificity: Specificity indicates how often our model predicted a company to fail when it did actually fail. Our model's specificity is 94%.

SCORING

Using the scoring node in Enterprise Miner the logistic model was tested on a new dataset. The following were the two examples showing how the model performed:

1) A closed company named Minekey, satisfied all the parameters of the model, such as it raised total of $36M ,2 milestones achieved, 2 funding rounds but had high burnrate and took around 23 months to achieve first funding and The model showed a the probability of surviving was 0.35 indicating it would not be successful.

2) An operating company named Rubicon project, was founded in 2007 and has raised $261M, 7 funding rounds, achieved 5 milestones and took around 4 months on an average to achieve each milestone. The model predicted its probability of surviving would be 0.87 hence indicating a successful company.

CONCLUSION

Based on the survival analysis, we can conclude that there is a strong relationship between degree and being a successful startup company. Because getting funds based on the idea does not lead to a successful company there should be people in the core-committee that have general and business-specific knowledge. The predictive model has got an accuracy of 86%, hence using the significant variables one can predict whether a company in an initial stage will be successful in future or not?

REFERENCES

Amar Krishna, Ankit Agarwal, Alok Choaoudhry, 02/02/2017,”Predicting the outcome of startups:Less failure ,More Success”,IEEE Xplore

SAS notes : http://support.sas.com/kb/22/601.html

ACKNOWLEDGMENTS

The data for this paper was obtained from crunchbase.com. We are thankful to crunchbase for providing the API to access the data.

CONTACT INFORMATION:

Vrushank Shah

Master of Science in Business Analytics

+1 405-334-3622

Vrushank.shah@okstate.edu