2022 2nd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 26-28, 2022

Predicting Success Rate of Startups using Machine Learning Algorithms

Malhar Bangdiwala
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
malhar.bangdiwala@spit.ac.in

Yashvi Mehta
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
yashvi.mehta@spit.ac.in

Smrithi Agrawal
Information Technology
Sardar Patel Institute of Technology
Mumbai, India
smrithi.agrawal@spit.ac.in

Sunil Ghane
Assisant Professor
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
sunil_ghane@spit.ac.in

2022 2nd Asian Conference on Innovation in Technology (ASIANCON) | 978-1-6654-6851-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASIANCON55314.2022.9908921

***Abstract*—It is essential for startups to assess if they are on the path to success. The failure rate of startups was around 90% in the year 2019 and hence it is necessary to know the success rate of a startup. Success for startups can be twofold: launching an IPO (Initial Public Offering) or getting merged/acquired by another company. This paper attempts to determine the success of a startup in terms of getting merged or acquired. With the help of historical data available on startups, five models have been built and compared to predict if a startup would get acquired or not. The models that have been used are Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural networks. The data used to train these models includes key features such as valuations, funding rounds, investments, etc. By using said models, one would be able to get an estimate of the trajectory of the company. After applying the models, we were able to get an accuracy of around 92% in all of them. This information would be vital to various stakeholders of the company as well as potential investors.**

***Index Terms*—Decision Trees, Random Forest, Gradient Boost, Logistic Regression, MLP Neural networks**

I. Introduction

Across the globe, startups have begun to garner a lot of attention in recent years. The number of new ventures are proliferating by the minute. Especially since the pandemic struck, startups and entrepreneurship have become the buzzwords of the time. The global epidemic has resulted in a true startup boom, with the number of new businesses throughout the world greatly exceeding last year's figures. Workers who were laid off and formed their own firms are credited for this boom in entrepreneurship. The number of applications filed for starting a business in July 2020 in the United States reached a record-high of 551,657. According to the Census Bureau[1], compared to the same period in 2019, this was an increase of 95%.

Startups are increasingly being recognised as major engines of economic growth and employment creation. Startups have become detrimental to the upliftment of society. They can generate meaningful solutions and thus, they operate as vehicles for socioeconomic development and transformation through innovation and scalable technology[2]. Startups are providing a new source of income for people. In the COVID timespan, one in five microbusinesses launches were by those who were categorized as non-employed. The pre-COVID figure was 12%[3]. This shows that microbusinesses and startups are providing a way out of the poverty cycle for a large number of people. Startup companies are the most dynamic economic organizations on the market since they provide additional dynamics and competitiveness to the economic system. This means that the economy stays healthy, vital, and diligent, while individual companies find it harder to fall asleep on their laurels[4].

However, not all the startups that have been founded in recent years succeeded. The failure rate of startups in 2019 was around 90%. Research concludes that 21.5% of startups fail in the first year, 30% in the second year, 50% in the fifth year, and 70% in their tenth year. This paints a rather grim picture for aspiring entrepreneurs. There are a plethora of reasons for such failures. One of the reasons is no market demand. Many startups revolve around a product for which the demand died down over the years or simply never existed. Another reason could be insufficient financial resources. Many startups simply run out of resources that are needed to sustain themselves in the future. Other reasons could also include wrong managerial decisions, acts of God such as Covid, or even strong competition.[5]

Thus, it is imperative for today's entrepreneurs to gauge if their company is on the right track. While there is no one specific definition or path to success, success for startups can be broadly classified into two categories. One option is for a firm to do an IPO (Initial Public Offering), in which the company is listed on a public stock exchange and shareholders have the option to sell their shares to the general public. Another is being Merged or Acquired (M&A) by another firm, in which the startup's owners or investors earn quick cash in exchange for their shares. This is also known as the ”exit strategy.”

In this paper, a startup is deemed to be successful if it undergoes an M&A. Traditional approaches based on vague statistics and estimates which take time cannot be used in this fast-paced world. Thus, an attempt at building five machine learning models for the same with high accuracy has been made. These machine learning models are Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural Networks. The data gathered for the same mainly revolves around financial numbers such as valuations and the amounts raised in each venture capital round. These models can be used not only by various stakeholders of the startup to assess their growth but also by potential venture capitalists who are looking to invest in the company.

The rest of the paper is organized as follows: Section II summarizes the recent research relevant to the problem, followed by a description of the methodology used in this study in Section III. Comparison of models and results are stated in Section IV. Section V discusses the conclusion and future work is presented in Section VI.

II. Literature Survey

Startup failures have attracted a lot of interest, and most companies are focused on developing various prediction models to accurately foretell the fate of a new company. A few researchers have done some intriguing studies into startup success and failure trends. One of the articles [5] examines the success and risk elements in the pre-startup phase. The authors want to know how important different approaches and variables are in explaining pre-startup success. They developed a paradigm that argues that start-up efforts differ in terms of the individual(s) who start the endeavour, the organisation they form, the environment in which the new venture is launched, and the method by which the new venture is launched.

Various studies have been conducted in order to determine various characteristics of entrepreneurship and how some of them can lead to a successful business. Similar difficulties are addressed by Begley and Tan [6]. Another well-known book is by R. Dickinson, who examines crucial success elements for small enterprises in his piece [7]. He also teaches how to alter these characteristics to produce a profitable business. Gartner [8] highlights a variety of issues that innovators face. This article focuses on the challenges that innovators confront in terms of capital, management, and so on. Malhotra [9] addresses the market orientation for entrepreneurs. The factors that can lead to successful businesses are discussed by McClelland [10].

Felgueiras et al.[11] discuss multi-label text classification tests performed on a dataset collected from CrunchBase. Three classification algorithms were used, including Multinomial Naive Bayes, SVM, and Fuzzy fingerprints, as well as various combinations of text representation features. The trial results in a precision of 70% and a recall of 42%. The accuracy of the multi-class technique is greater than 65%.

The work by Pan et al. [12] is based entirely on machine learning, and the data set was obtained via CrunchBase. The main purpose is to forecast the state of start-ups, whether they have gone through M&A (mergers and acquisitions) or an initial public offering (IPO) (Initial public offering). Logistic Regression, Random Forests, and K Nearest Neighbors are the ML methods that have been used. The study examines the various ML algorithms listed above and determines which algorithm performs best with the dataset. F1 scores are utilised as the key criterion, and KNN was shown to have the highest F1 score.

The research by Arroyo et al. [13] focuses not just on the traditional two classification categories of M&A and IPO, but also on other conceivable outcomes such as a second investment round or the company's closure. ML algorithms that are used in this study are Support Vector Machines, Decision Trees, Random Forests, Extremely Randomized Trees, and Gradient Tree Boosting. The popular database CrunchBase was used in this investigation once again. The focus on early stage enterprises, time-aware analysis, and the multiclass prediction issue are the major elements of their approach. The study's findings suggest that the best algorithm, Gradient Tree Boosting[14], has a worldwide accuracy of roughly 82%.

III. Methodology

A. Dataset

In order to use our models effectively and get comprehensive results, we chose datasets having a variety of attributes present. The dataset has information and statistics not only about public companies but also about private companies around the world. The data was obtained from CrunchBase, a platform for finding business information about companies. This information includes funding and investment details, details of founding individuals, information about acquisitions, mergers, latest trends, and news. The dataset contains information on numerous start-up businesses from around the world. These details include start date, start location, seed funding, total rounds of funding, funding details, valuation, market value, investment and acquisition details, country, city, etc.

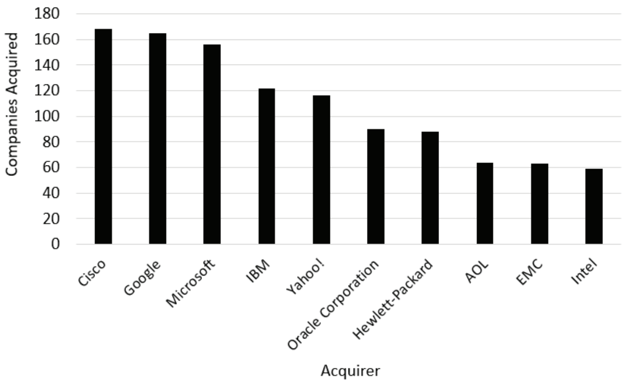

Fig. 1. Number of companies acquired vs acquirer name

Figure 1 depicts the number of firms that a single acquirer has bought. This visualization is essential to this research study since it is based on acquisition data and the main yardstick of the success of a startup in this paper is whether it was bought by a larger company or not. Cisco has acquired the most number of startups: 168, while Google is second with 165 companies acquired.

B. Data Cleaning

Data preprocessing is one of the key steps in the whole data mining process. In this paper, there are several datasets: acquisitions, companies, funding, etc. that have been cleaned and merged into the final dataset. Usually, data cleaning consists of processes like determining outliers and removing or imputing outliers, removing or replacing missing values, removing duplicate values, and removing values with less or no importance. The number of null values in the companies dataset is represented in Table I.

TABLE I
NUMBER OF NULL VALUES IN DATA

Sr. No.	Column name	Number of null values
1	Permalink	4856
2	Name	4856
3	Homepage_url	8305
4	Category_list	8817
5	Market	8824
6	Funding_total_usd	4856
7	Status	6170
8	Country_code	10129
9	State_code	24133
10	Region	10129
11	City	10972
12	Funding_rounds	4856
13	Founded_at	15740
14	Founded_month	15812
15	Founded_quarter	15812
16	Founder_year	15812
17	First_founding_at	4856
18	Last_founding_at	4856

Data cleaning plays an immense role in our models. We have used SparkSession so as to programmatically create PySpark RDD and DataFrame. The reason to use this is that queries can be executed to retrieve the data and get the result back as a DataFrame.After removing rows with null data or missing data, we have used spark to create dataframes with defined labels.

Some visualisations can be drawn after the final dataset has been curated. These can be seen in Figure 2. The target variable is depicted by the 'status' column of the dataset. The value of 1 is assigned to those startups that have been acquired. All other companies have been assigned the value of 0. As can be seen, only 14% of the companies have been acquired.

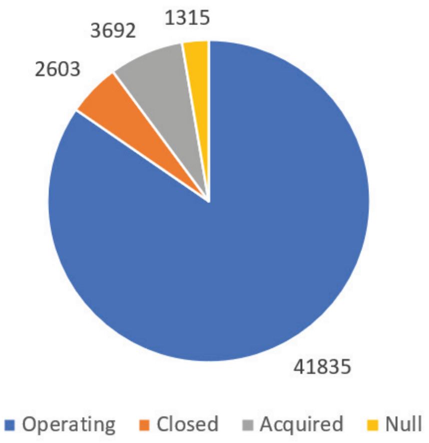

Fig. 2. Distribution of Target Variables

Post splitting the dataset into training and testing, we defined a vector assembler with the features for modeling.

C. Models Used

1) Decision Tree: Decision tree displays predictions from a succession of feature-based splits using a flowchart that looks like a tree structure. To make a decision it starts with the root node and concludes with the leaf nodes. The algorithm is divided into two classes: the main class, which includes the algorithm, and a helper class, which defines a node.

2) Random Forest: Random Forests train several trees by utilizing a random sample of the statistics. This makes the model more robust and less overfitting. The final class of an instance is assigned by outputting the class that is the mode of individual tree outputs, which can create robust and accurate classification and manage a huge number of input variables. This helped in handling our dataset that had extremely uneven class distributions. We have made use of 25 trees in our model. We tried the same with 15 and 20 models as well. Although the accuracy remained the same, the AUC obtained was marginally different. The AUC values are tabulated in Table II.

TABLE II
NUMBER OF TREES & AUC OF MODELS USED

Sr. No.	Number of trees	AUC
1	15	0.6720
2	20	0.6833
3	25	0.6839

3) Gradient Boost: Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. Here the algorithm started by building a decision stump and then assigning equal weights to all the data points. Then it increased the weights for all the points which are misclassified and lowered the weight for those that are easy to classify. A new decision stump is made for these weighted data points. The idea behind this is to improve the predictions made by the first stump. By using this algorithm, errors made by previously trained trees were corrected since the trees were build one at a time.

4) Logistic Regression: Logistic regression is a type of linear regression in which the response variable, Y, is a binary variable. The logistic regression function is designed to produce a value between 0 and 1, which represents the chance of belonging to one of two dichotomous classes, p, (xi).

After removing the variables with approximately zero variance, full simple logistic regression (M0) evaluates the remaining variables. M0 exposes the inconsequential variables and proves the existence of the dummy trap. In the simplified logistic regression model, one level of dummy variables and statistically unimportant variables are omitted in the second phase (M1). The near-zero estimate of overall funding is the most surprising conclusion. This is the result of several factors, some of which are positively connected with success rate and others which are adversely correlated with success rate. As a result, the larger the amount of financing, the higher their anticipation that the firm has future potential. The cash-burning of a company explains the insignificant effect of total funding (USD) on success. According to Ooghe and De Prijcker [15], startups that got large investments during their rapid-growth era frequently fail due to bad management decisions, including misallocation of cash received. This data reveals that unsuccessful businesses have significant burn rates.

TABLE III
LOGISTIC REGRESSION PARAMETERS & ACCURACY

Sr. No.	**α value**	**Accuracy**
1	0.01	0.9162
2	0.05	0.9205
3	0.1	0.9209

For this research, we trained three Logistics Regression models with different α values. α is the penalty that is used to keep the overfitting in check. The different results obtained by using various α values have been tabulated in Table III. We have used that α value which has given the highest accuracy.

5) MLP Network: A fully connected multi-layer is known as a multi-layer perceptron (MLP). There are three components: an input layer, hidden layers, and an output layer. The computation of MLP is the arbitrary number of hidden layers inserted between the input and output layers. Data flow from the input to the output layer in the forward direction, similar to a feed-forward network. The MLP neurons are trained using the backpropagation algorithm. In our research, we looked at a variety of scenarios with varying numbers of hidden layers and 30 neurons. We discovered that when the number of concealed layers is the highest, the accuracy is the best (in this case, the number of hidden layers was 3). Accuracy and AUC have been tabulated in Table IV.

TABLE IV
NEURAL NETWORK PARAMETERS & PERFORMANCE

Sr. No.	**No. of hidden layers**	**Accuracy**	**AUC**
1	1	0.9121	0.8950
2	2	0.9175	0.8950
3	3	0.9217	0.8972

IV. RESULTS

When it comes to evaluating model performance and deciding on a metric to compare the five models that were implemented, there are numerous alternatives. By comparing True Positive Rate (TPR) vs. False Positive Rate (FPR) at all classification thresholds, a Receiver Operating Curve (ROC) depicts the performance of a classification model. The integral of the ROC curve between 0 and 1 is used to calculate the AUC, which is an aggregate assessment of performance at various threshold levels. Because each metric has advantages and disadvantages, a combined evaluation technique is used. The above-mentioned equations have been written below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$True\,Positive\,Rate = \frac{TP}{TP + FN} \tag{2}$$

$$False\,Positive\,Rate = \frac{FP}{FP + FN} \tag{3}$$

After appropriate data preparation for each model, and running the respective models, the accuracy that has been achieved for each is depicted in Figure 3. All models had an accuracy of over 91.75%. The logistic regression model gave the highest accuracy of around 92.5%. The lowest is the Neural network model with the accuracy being a little over 91.75%. Decision tree and random forest classifiers performed similarly, with accuracies over 92.25%.

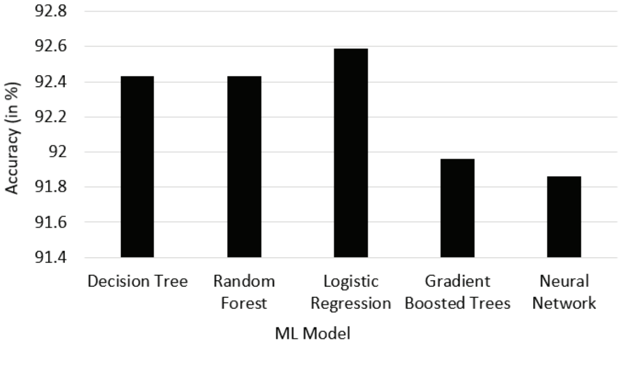

Fig. 3. Accuracy of Models

When compared to other models, the AUC value of the Multilayer Perceptron Neural Network is greater. Although Random Forest and Logistic Regression with PCA methods have greater accuracy, their AUC values are much lower. This can be seen in Figure 4.

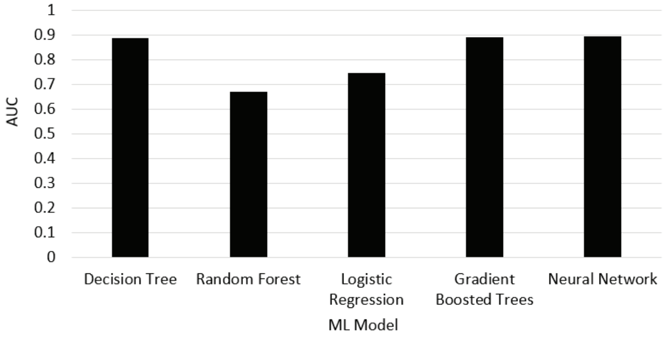

Fig. 4. AUC of Models

Table V shows the exact values of accuracy and AUC of the models used. In conclusion, all the five models perform satisfactorily in determining whether the company is successful or not.

TABLE V
ACCURACY & AUC OF MODELS USED

Sr. No.	Model name	Accuracy (in %)	AUC
1	Decision Tree	92.43	0.88
2	Random Forest	92.43	0.67
3	Logistic Regression	92.59	0.74
4	Gradient Boosted Trees	91.96	0.89
5	Neural Network	91.86	0.89

V. CONCLUSION

This paper delves at how to forecast startup success in depth. The amount of literature on startup success highlighted the need for more research. The existing literature focuses on predicting established firm success rates. However, there are significant discrepancies between corporate and startup success prediction, rendering existing models useless for predicting startup success. The startup ecosystem's actors can greatly benefit from a quantitative strategy when it comes to making judgments in such a high-risk setting, due to the energy and time consuming nature of processing massive amounts of data.

We have used several machine learning algorithms to construct models for success/failure prediction of early stage startups. Precision accuracies of 92.4%, 92.3%, 92.6%, 91.9%, 91.8% for the respective models has been achieved. Further we have also studied the values for the ROC area. Given the prediction quality we can certainly say that any early stage startup can use our prediction models (at every milestone) to predict their outcome. Based on our analysis, we can also conclude that there is a strong relationship between the above mentioned features (TABLE I) 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.

VI. FUTURE SCOPE

Although successful models have been built for predicting whether a startup will get acquired or not, this can be extended to the other route of success for startups that have been discussed earlier, i.e. IPO. In the current models, the state of global economics and crises have not been considered. Moderate funding during a recession or an extraordinary global event such as the pandemic could point to a more successful startup than higher funding in normal times. Similarly, geographic location and the implication of a nation's economic scenario have not been considered. Thus, in the next iteration, the above-mentioned factors must be incorporated to build a more accurate predictor.

ACKNOWLEDGEMENT

This work was fully supported by Sardar Patel Institute of Technology. We would like to thank Prof. Sunil Ghane for his constant support and guidance and for pushing us to write this research paper.

REFERENCES

[1] Y.B. Altun "Pandemic Fuels Global Growth Of Entrepreneurship And Startup Frenzy" forbes.com. https://www.forbes.com/sites/forbestechcouncil/2021/04/09/pandemic-fuels-global-growth-of-entrepreneurship-and-startup-frenzy/?sh=292d4e3f7308 (accessed: May 7,2022)

[2] S. Korreck. The Indian Startup Ecosystem: Drivers, Challenges and Pillars of Support. ORF Occasional Paper #210

[3] S. Torkington "How the Great Resignation is driving a boom in startups from more diverse founders" weforum.org. https://www.weforum.org/agenda/2022/02/the-great-resignation-boom-in-startups-from-more-diverse-founders/ (accessed: May 7,2022)

[4] T. Eschberger "5 TOP reasons why startups fail" lead-innovation.com. https://www.lead-innovation.com/english-blog/reasons-startups-fail (accessed: May 7,2022)

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

[6] T. M. Begley and W.-L. Tan. The socio-cultural environment for entrepreneurship: A comparison between east asian and anglo-saxon countries. Journal of international business studies, pages 537–553, 2001.

[7] R. Dickinson. Business failure rate. American Journal of Small Business, 6(2):17–25, 1981.

[8] W. B. Gartner. Who is an entrepreneur? is the wrong question. American journal of small business, 12(4):11–32, 1988

[9] N. K. Malhotra. Marketing Research: An Applied Orientation, 5/E. Pearson Education India, 2008.

[10] D. C. McClelland. Characteristics of successful entrepreneurs*. The journal of creative behavior, 21(3):219–233, 1987.

[11] Marco Felgueiras, Fernando Batista, Joao Paulo Carvalho, Creating Classification Models from Textual Descriptions of Companies Using Crunchbase", IPMU 2020, CCIS 1237, pp. 695–707.

[12] Chenchen Pan, Yuan Gao, Yuzi Luo," Machine Learning Prediction of Companies' Business Success", CS229: Machine Learning, Fall 2018, Stanford University, CA,2018.

[13] Javier Arroyo , Francesco Corea, Guillermo Jiménez-Díaz, Juan A. Recio-García," Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments", IEEE Access 7,2019, pp. 124233-124243

[14] Amar Krishna, Ankit Agrawal, Alok Choudhary, “Predicting the Outcome of Startups: Less Failure, More Success”, IEEE, 2016.

[15] Ooghe, Hubert and De Prijcker, Sofie. Failure processes and causes of company bankruptcy: A typology. Management Decision. 46. 10.1108/00251740810854131. 2006.