ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*
*Journal DOI: 10.21474/IJAR01*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# STATISTICAL AND MACHINE TECHNIQUES FOR ASSESSING THE STATUS OF STARTUPS

**T. Leo Alexander[1] and Marion Nikita Joseph[2].**

1. Associate Professor, Department of Statistics, Loyola College, Chennai – 600 034.
2. Research Scholar, Department of Statistics, Loyola College, Chennai – 600 034.

***Manuscript Info***

***Manuscript History:***

Received: 18 March 2016
Final Accepted: 26 April 2016
Published Online: May 2016

***Key words:***
Big Data, Business Organization, Data Mining, Startup Company, Statistical techniques.

***Corresponding Author**

**T. Leo Alexander.**

***Abstract***

We are in the midst of an entrepreneurial revolution that is spreading to nearly every nook and corner on the planet. Even countries plagued with political strife or in the midst of a deep recession are seeing a surge in start-up activity. In this paper, we have used some of the statistical and data mining techniques for predicting the status of startups. Sections 3, 4, 5 and 6deal with Logistic Regression, Decision Trees, Ensemble Learning and Boosting methods respectively for finding some of the factors that determining the successes or failures of the Startups and comparison of the methods.

## Introduction

A startup is a young company that is beginning to develop and grow[(8)], is in the first stages of operation, and is usually financed by an individual or small group of individuals and which could be an entrepreneurial venture or a new business, a partnership or temporary business organization designed to search for a repeatable and scalable business model. We are in the midst of an entrepreneurial revolution that is spreading to nearly every nook and corner on the planet. Even countries plagued with political strife or in the midst of a deep recession are seeing a surge in start-up activity[(9)].

There are a few Challenges for Startups, which are Culture, Awareness, Social Issues, Technology, Financial issues, Sustainability Issues and Regulatory Issues.

In the following Sections we will be discussing a quick look into the main objectives of the study and the data sources, methods adopted and Statistical Analysis.

## Objectives of the study

1. To assess the factors that is influencing the success or failure of Startups globally using advanced data mining techniques.
2. To identify the industries that has potential to be successful worldwide.
3. To assess the status of startups in India and their growth in the recent years.

In other words this study aims to analyze the trends in the startups worldwide and identify influential factors and potential industries that have been promising in the past.

The data was collected from one of the leading data science competitions website***(***https://www.kaggle.com/***)***. It consists of 472 rows containing the information of startups along with their current status: “Success” or “Failed” (response variable). The data consists of 116 variables among which 76 are categorical and 40 are numerical. 22 of the variables contain information pertaining to the founders and co-founders of these startups since these might affect the status of the company.

Data cleaning was done by accounting for missing values and treating for outliers. Further feature engineering was done to create some new variables by combining some of the existing variables and dummy variables were also created for the variables with many levels like industry of the company etc. and about 70 variables have been used for the modeling.

In the following Sections, we discussed Statistical Techniques and Data Mining methods which have been applied to study the factors influencing the success or failure of the Startups. The applications and findings of the modeling and data mining techniques pertaining to the 472 Startups included in the study are discussed in detail in the following Sections.

## Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).In logistic regression, the dependent variable is binary or dichotomous(6),(7).Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest

$$\log it(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \,, \qquad (3.1)$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} \qquad (3.2)$$

and

$$\log it(p) = \ln\left(\frac{p}{1-p}\right). \qquad (3.3)$$

### Classification table

The classification table is another method to evaluate the predictive accuracy of the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off value) are cross-classified.

### ROC curve analysis

Another method to evaluate the logistic regression model makes use of ROC curve analysis. In this analysis, the power of the model's predicted values to discriminate between positive and negative cases is quantified by the Area under the ROC curve (AUC). The AUC, sometimes referred to as the c-statistic (or concordance index), is a value that varies from 0.5 (discriminating power not better than chance) to 1.0 (perfect discriminating power).

### Analysis based on Logistics Regression

First, all the variables were used in the LR model. A rather anomalous result was obtained where all the variables were significant. Now we see the model accuracy measures with classification matrix means that the predicted values of the dependent variable is status of startups. Also we see the AUC value and ROC curve for the model.

**Accuracy measures:** The Classification matrix of final logistic model on test data is as follows:

| | Failure | Success |
|---|---|---|
| Failure | 43 | 7 |
| Success | 16 | 75 |

The accuracy was 0.836 in this case which was higher than the previous models. This was obtained over a threshold range of 0.37.

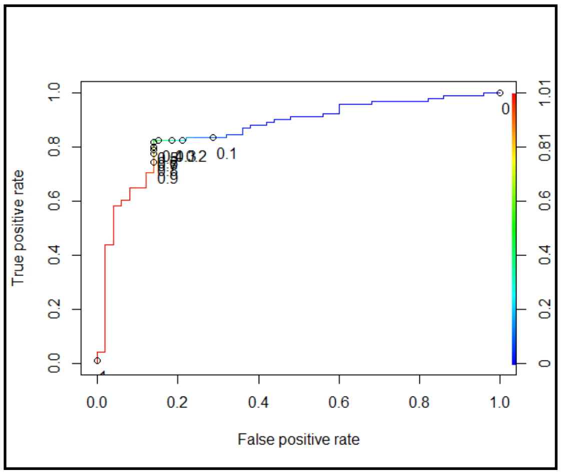

**Fig 3.1 ROC curve LR final model**

**Area Under the Curve (AUC):**The AUC obtained for train data for final Logistic regression model was0.83.The AUC obtained for test data was 0.84 which signifies the stability of the model.

## Decision Trees

Decision trees are a class of predictive data mining tools which predict either a categorical or continuous response variable. They get their name from the structure of the models built. A series of decisions are made to segment the data into homogeneous subgroups. This is also called recursive partitioning. When drawn out graphically, the model can resemble a tree with branches.

### Classification and Regression Trees (CART)

CART, a recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition(1).

### Impurity Measures

Used by the cart (classification and regression tree) algorithm, gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

Estimation of Node Impurity: Gini Measure

The Gini measure is the measure of impurity of a node and is commonly used when the dependent variable is a categorical variable, defined as:

If costs of misclassification are not specified,

$$g(t) = \sum_{j \neq i} p(j\,|\,t)\,p(i\,|\,t) \qquad (4.1)$$

If costs of misclassification are specified,

$$g(t) = \sum_{j \neq i} C(i\,|\,j)\,p(j\,|\,t)\,p(i\,|\,t)\,, \qquad (4.2)$$

where the sum extends over all $k$ categories. $p(j\,/\,t)$ is the probability of category $j$ at node $t$ and $C(i\,/\,j\,)$ is the probability of misclassifying a category $j$ case as category $i$..

### Analysis based on Decision Trees

CART(Classification and Regression Tree) model in R is a decision tree model which takes train data as input with 71 variables. It performs a univariate split with respect to independent which gives maximum information gain from root node to child node. “class” method deals with the case when response variable is a categorical variable. We set the complexity parameter as 0.03 which specifies that when the split does not improve by 0.03 amounts, the model will not perform the split.

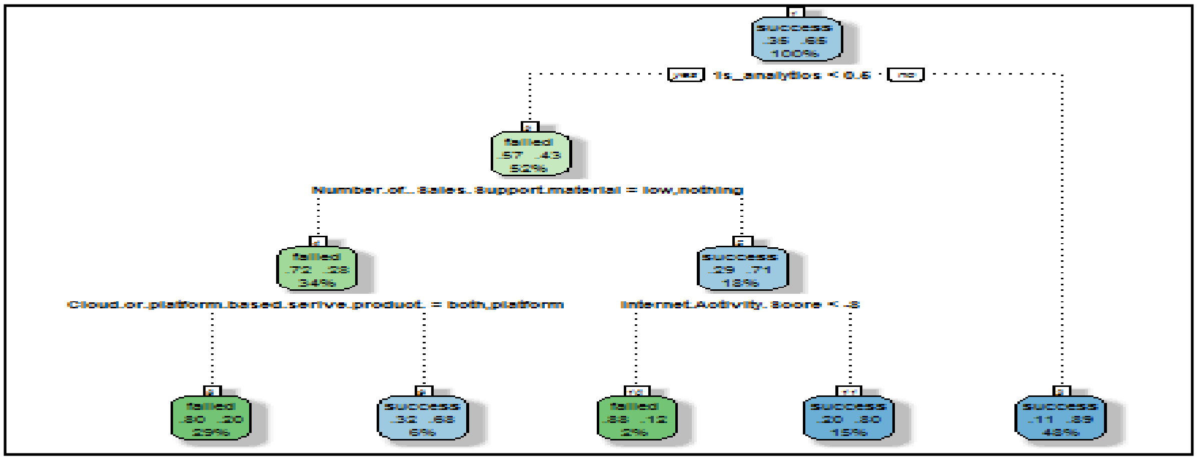

**Fig 4.1: Decision tree from CART model**

The above picture shows the decision tree. The relative importance of the independent variable can be visually seen from the model itself. The first split happened on the basis of “Is_analytics”. Further split happened on the basis of “Number of Sales Support Material”, “Cloud or platform based service product” and “Internet Activity Score”.

It can be seen that when then value of “Is_analytics” is 1 (i.e. > 0.6), the model indicates success. Similarly, when “Number of Sales Support material” is not low and “Internet Activity Score” is greater than -8, model predicts success. The accuracy vs threshold curve has been generated for train data to get an estimate of threshold for which the accuracy is maximized. For this case, it is taken as 0.5 (almost same for range 0.2 to 0.6) as evident from the graph. While predicting the test data, the confusion matrix has been created taking threshold as 0.5, which is as follows:

**Accuracy Measures:** The Classification matrix for decision tree (CART) on test data is as follows:

| | Failure | Success |
|---|---|---|
| Failure | 40 | 10 |
| Success | 13 | 78 |

The model attains accuracy of 0.8368794 which is almost similar as Logistic regression (model 3). The ROC has been generated by changing threshold from 0 to 1.

**Area Under the Curve:** The AUC obtained for this model while modeling on train data is 0.85.The AUC obtained for test data was 0.86 which signifies the stability of the model.

## Ensemble Learning

Recently there has been a lot of interest in “ensemble learning” — methods that generate many classifiers and aggregate their results. Two well-known methods are boosting of classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees — each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction.

### Random Forests

The random forests are an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node(5).

### Variable importance

This is a difficult concept to define in general, because the importance of a variable may be due to its (possibly complex) interaction with other variables. The random forest algorithm estimates the importance of a variable by looking at how much prediction error increases when (OOB) data for that variable is permuted while all others are left unchanged.

### Analysis based on Random Forest

Further we have used Random Forests to improve the prediction accuracies and understand the relative importance of the predictor variables influencing the status of startups.

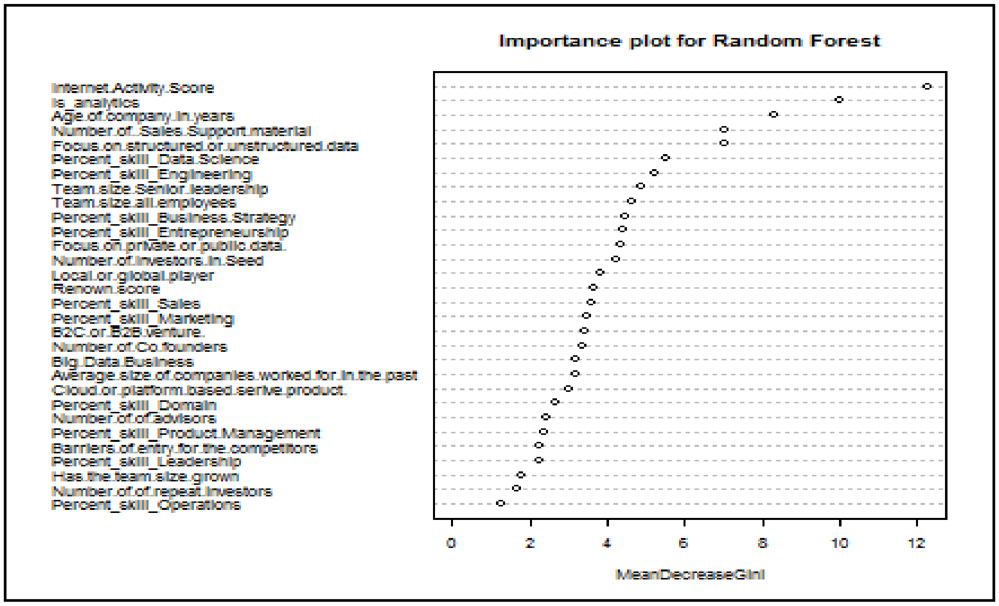

**Fig 5.1: Variable importance plot of variables in the model**

Now we see the model accuracy measures with classification matrix means that the predicted values of the dependent variable is status of Startups. Also we see the AUC value and ROC curve for the model.

**Accuracy Measures:**classification matrix for Random Forest model on test data is as follows:

| | Failure | Success |
|---|---|---|
| Failure | 44 | 6 |
| Success | 3 | 88 |

The model attains accuracy of **0.9361702** which is much higher than CART model and Logistic regression model. The Random Forest model has been trained with different number of trees ranging from 100 to 1200 and the number of randomly sampled variables ($m_{try}$) has been varied from 1 to 10. (Default value of $m_{try}$ is $\sqrt{No.of\ variable}$ which is$\sqrt{71} = 8.42\ \approx 8$).The following Table 5.1 enumerates accuracies for varying parametric values:

**Table 5.1**

| No of trees ↓ | m=1 | m=2 | m=3 | m=4 | m=5 | m=6 | m=7 | m=8 | m=9 | m=10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **100** | 0.837 | 0.908 | 0.901 | 0.922 | 0.908 | 0.908 | 0.929 | 0.922 | 0.929 | 0.915 |
| **200** | 0.844 | 0.901 | 0.908 | 0.894 | 0.894 | 0.915 | 0.929 | 0.922 | 0.929 | 0.908 |
| **300** | 0.837 | 0.901 | 0.901 | 0.908 | 0.901 | 0.908 | 0.929 | 0.922 | 0.922 | 0.915 |
| **400** | 0.865 | 0.894 | 0.901 | 0.908 | 0.908 | 0.922 | 0.915 | 0.922 | 0.922 | 0.908 |
| **500** | 0.851 | 0.901 | 0.887 | 0.894 | 0.901 | 0.915 | 0.908 | 0.922 | 0.929 | 0.929 |
| **600** | 0.879 | 0.894 | 0.887 | 0.901 | 0.915 | 0.922 | 0.915 | 0.936 | 0.936 | 0.929 |
| **700** | 0.872 | 0.894 | 0.894 | 0.908 | 0.915 | 0.908 | 0.929 | 0.922 | 0.922 | 0.922 |
| **800** | 0.865 | 0.901 | 0.901 | 0.887 | 0.915 | 0.915 | 0.922 | 0.915 | 0.922 | 0.929 |
| **900** | 0.865 | 0.887 | 0.894 | 0.887 | 0.908 | 0.901 | 0.929 | 0.922 | 0.929 | 0.929 |
| **1000** | 0.858 | 0.901 | 0.901 | 0.894 | 0.908 | 0.901 | 0.915 | 0.929 | 0.929 | 0.929 |
| **1100** | 0.865 | 0.894 | 0.894 | 0.887 | 0.901 | 0.908 | 0.915 | 0.922 | 0.929 | 0.922 |
| **1200** | 0.865 | 0.894 | 0.887 | 0.901 | 0.908 | 0.922 | 0.922 | 0.915 | 0.929 | 0.929 |

**Area Under the Curve:** The AUC obtained for this model on train data set is 0.89 and0.91 for the test data set which is much higher than the logistic regression model or CART model.

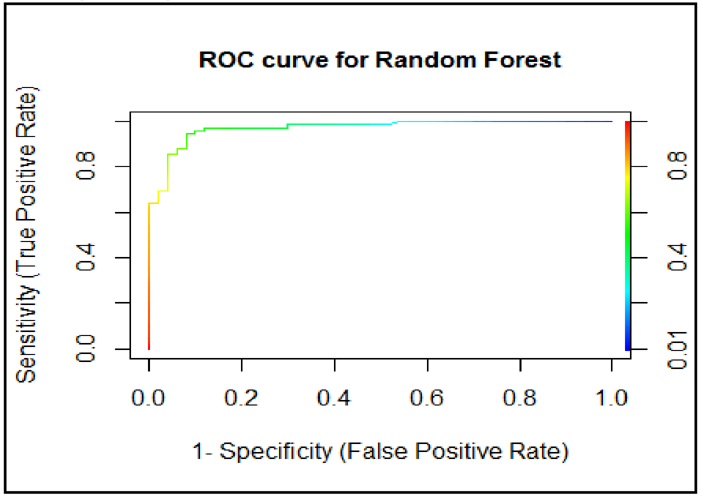

**Fig 5.2: ROC curve for Random Forest model**

## Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification (see also Bagging).

### Gradient Boosting

Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees[(2)]. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

### Analysis based on XGBoost- Xtreme Gradient Boosting/ Boosted Trees

A similar ensemble model, Gradient Boosted trees also have been applied on train data. We have applied Extreme gradient Boosting algorithms by the XGBoost package in R[(3),(4)].While modeling, the gain for each of the variables has been calculated for every tree in the forest and the mean of those gains is reported. Based on this metric, most important independent variables are obtained and top ten variables are shown below for XGBoost model.

**Table 6.1: Depicting Importance measures from XGBoost model**.

Now we see the model accuracy measures with classification matrix means that the predicted values of the dependent varial

| Feature | Gain | Cover | Frequency |
|---|---|---|---|
| **Is_analytics** | 0.1494962 | 0.0429849 | 0.0179211 |
| **Internet.Activity.Score** | 0.1482719 | 0.1740725 | 0.1684588 |
| **Percent_skill_Data.Science** | 0.1197457 | 0.0500478 | 0.0430108 |
| **Number.of..Sales.Support.material** | 0.0777968 | 0.0541701 | 0.0537634 |
| **Age.of.company.in.years** | 0.0726718 | 0.0955614 | 0.0609319 |
| **Number.of.Investors.in.Seed** | 0.0529553 | 0.0415104 | 0.0430108 |
| **Percent_skill_Engineering** | 0.0360754 | 0.0524166 | 0.046595 |
| **Team.size.all.employees** | 0.0342999 | 0.0468359 | 0.0824373 |
| **Percent_skill_Sales** | 0.030408 | 0.0597855 | 0.0394265 |
| **Local.or.global.player** | 0.0296384 | 0.0248622 | 0.0250896 |

**Accuracy Measures:** The Classification matrix for XGBoost model on the test data is as follows:

| | Failure | Success |
|---|---|---|
| Failure | 43 | 7 |
| Success | 5 | 86 |

The XGBoost model attains accuracy of 0.914 which is much higher than CART model or Logistic regression model but slightly lesser than Random Forest model.
The learning rate (eta) for the model has been kept default (0.3) and number of trees is set to be 100. The maximum depth of one tree of the ensemble was kept at 2.

**Area Under the Curve:-**
The AUC obtained for this model is about 0.885 for the train data and 0.901 for test data which is much higher than the Logistic regression model and CART model but is slightly lesser than Random Forest.

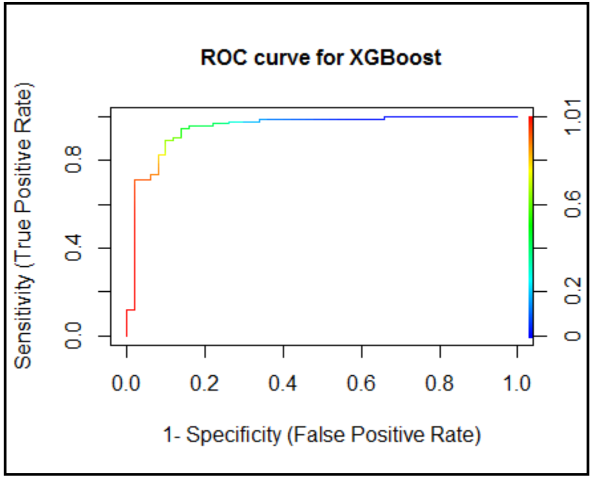

**Fig 6.1:ROC curve for XGBoost model**

## Conclusion

Having discussed the results from various modeling techniques applied in the study we see that data mining techniques tend to improve the predictive accuracy and gives us interesting insights concerning the success or failure of the startups.

When internet activity score is high (~ 115), which means when the company is very much active in social media, the chance of success is higher than companies having lower (~ 3) internet activity score. Hence Online presence of company given by the Internet Activity Score plays a big role in determining the success. When the percent skill of entrepreneurship of the co-founders is high, the company has more chance to be successful.

When the startups don't focus on any of the public or private data, the proportion of failure is high compared to other options. Similarly focusing on consumer data significantly increases the chance of success. The same result arises when the startup uses Big data. Companies belonging to the Analytics industry tend to be more successful.

Skills of Founders and Co-Founders in Data Science, Engineering, Business Strategy, Entrepreneurship, Sales and Marketing were found to be significant.

## References

1. Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984). Classification and Regression Trees. Belmont, California: Wadsworth..
2. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232.
3. Hothorn, T., P. Buhlmann, T. Kneib, M. Schmid and B. Hofner (2010): Model-based boosting 2.0. Journal of ¨ Machine Learning Research 11, 2109-2113.
4. Hothorn, T., P. Buhlmann, T. Kneib, M. Schmid and B. Hofner (2011): mboost: Model-Based Boosting. R ¨ package version 2.1-0. https://r-forge.r-project.org/projects/mboost/
5. L. Breiman. Random forests. Machine Learning, 45(1): 5–32, 2001.
6. Peng, C. J., & So, T. H. (2002). Logistic regression analysis and reporting: A primer. Understanding Statistics, 1(1), 31-70.
7. Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The Journal of Educational Research,96(1), 3-14.
8. Startup Compass Inc. (2014). The Tech Salary Guide.
9. Startup outlook report 2016 By Innoven Capital.