# Startup Success Prediction using Machine Learning

## MENTOR- SHAKTI KINGER

Muhammad Arab
MIT-WPU,PUNE

Shreyash Memane
MIT-WPU,PUNE

Akshay Lokhande
MIT-WPU,PUNE

Nihaal Shetty
MIT-WPU,PUNE

**Abstract**

Predicting the success of a business venture has always been a struggle for both practitioners and researchers. However, thanks to companies that aggregate data about other firms, it has become possible to create and validate predictive models based on an unprecedented amount of real-world examples. This work aims to create a predictive model based on machine learning to forecast a company's success. Plenty of those experiments, often conducted with the use of data gathered from several sources, reported promising results. However, we found that very often their use of data containing information was a direct consequence of a company reaching some level of success (or failure significantly biased them). Such an approach is a classic example of the look-ahead bias. We designed our experiments to prevent the leaking of any information unavailable at the decision moment of the training set. These analyses will provide investors and venture capital companies with effective methods, reduce their large human resources input for prediction, and improve the efficiency of their analysis of startup companies.

## I. INTRODUCTION

Start-ups are booming everywhere as more colleges, governments and private companies invest and stimulate people to pursue their ideas throughout these ventures. Companies are raising millions with ease and achieving unicorn status (i.e., a one-billion-dollar valuation) in a matter of years. Slack, a messaging app, achieved it after operating for 1.25 years (Kim, 2015). Examples like Uber and Airbnb are changing societies in such impactful ways that regulation had to keep pace with a new reality. Start-ups are having such an impact that, ultimately it becomes every investor's ambition to be part of a large acquisition such as Facebook acquiring WhatsApp (another messaging app) for nineteen billion dollars which allowed Sequoia (a Venture Capital fund) to have a 50x return on investment (Neal, 2014). But there is a catch, start-ups are companies with about 90% probability of failure, which means a lot of investments without proper returns

.

## II. MOTIVATION

In order to properly formulate the startup success prediction problem, the notion of "success" has to be formalized meaningfully. The definition should satisfy two main conditions: First, it should translate to real profitability. Second, success defined that way should be both available for evaluation (that is, it should be determinable from publicly available data) and should not require us to forecast into the distant future, in order to maintain tractability.

1. *Revenue*: A proper success metric would be revenue. Making revenue is the ultimate financial objective of a business, and this is what investors actually hope when giving funding. Unfortunately, this is a challenging target for prediction.

2. *M&A*: One such interaction is an M & A (Merger and Acquisition) event. Instead of predicting Revenue, the choice we make in this paper is to focus on predicting whether the company is acquired or not. The fact of a particular company being acquired usually demonstrates the acquiring party's high regard of the company's business.

## III. PROBLEM DEFINITION

The success of a business venture is a reason for founders and investors to feel proud. It is also strongly connected with a financial reward. Both founders and investors are actively looking for tools, methods, and advice that can give them an advantage over their competitors. It is debatable whether being a successful entrepreneur is associated with some intrinsic skills or whether those skills can be acquired (e.g. through formal business education). It is also very difficult to measure the significance of exogenous factors, such as the industry that the company operates in, the area where the headquarters is located, or the level of competition in a particular sector and its sub-sectors.

## IV. OBJECTIVES

To solve this problem, we're employing some machine learning techniques that will assist us in predicting the success of a company based on the features in the algorithm.

## V. DATASET DESCRIPTION

We've obtained this dataset from github and converted it into a csv file inorder to model's requirement. In this model we've used 2 datasets i.e. train.csv and test.csv for respective purposes . While training the model we've split the dataset into 80-20 % for training and testing the data respectively.

## VI. ALGORITHMS

*Support Vector Classifier*: The aim of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

Random Forest: Random forest classifier is a meta-estimator that fits several decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size, but they drew the samples with replacement.

LightGBM: LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.

Decision Tree: Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

## ALGORITHM PERFORMANCE

**SVC**

roc_aoc = 0.87047455968

| | Precision | Recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.79 | 0.85 | 73 |
| 1 | 0.88 | 0.95 | 0.91 | 112 |
| macro_avg | 0.89 | 0.87 | 0.88 | 185 |
| wght_avg | 0.89 | 0.89 | 0.88 | 185 |

**RANDOM FOREST**

roc_aoc = 0.7287181996

| | Precision | Recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.49 | 0.64 | 73 |
| 1 | 0.74 | 0.96 | 0.84 | 112 |
| macro_avg | 0.82 | 0.73 | 0.74 | 185 |
| wght_avg | 0.81 | 0.78 | 0.76 | 185 |

**LGBM CLASSIFIER**

roc_aoc = 0.947590

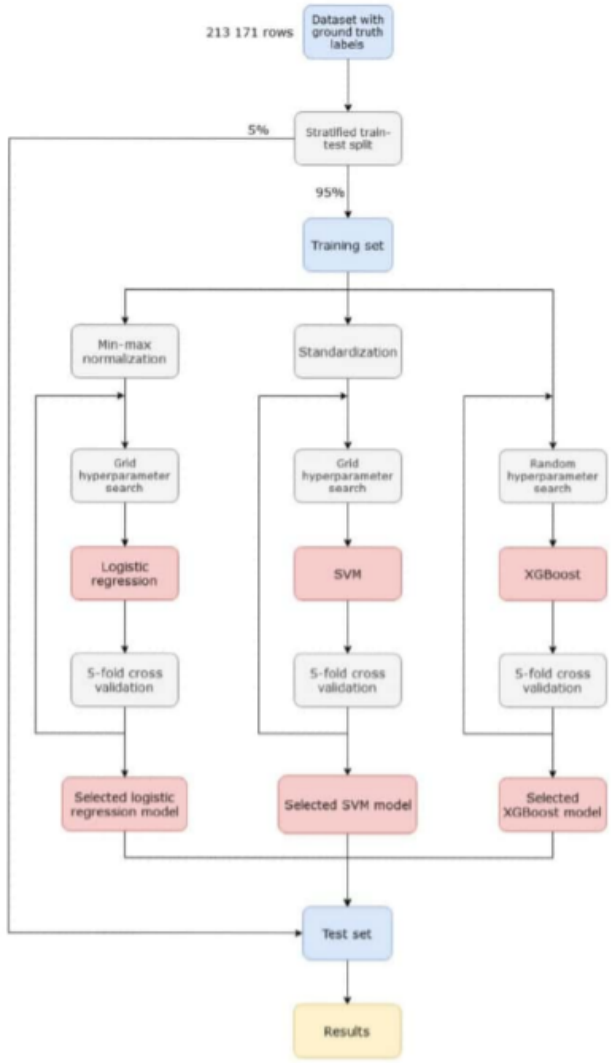
| | Precision | Recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.90 | 0.94 | 73 |
| 1 | 0.94 | 0.99 | 0.97 | 112 |
| macro_avg | 0.96 | 0.95 | 0.95 | 185 |
| wght_avg | 0.96 | 0.96 | 0.96 | 185 |

**DECISION TREE**

roc_aoc = 0.77837

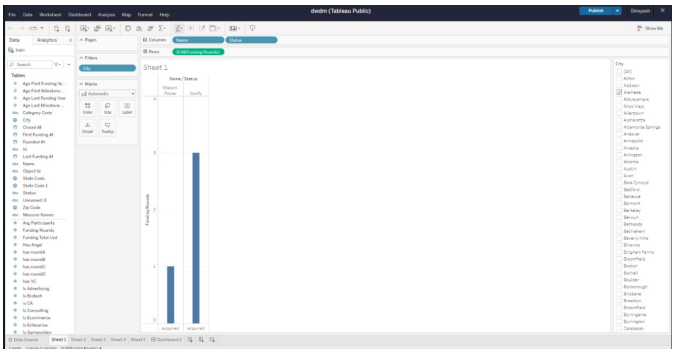
| | Precision | Recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.59 | 0.68 | 73 |
| 1 | 0.77 | 0.90 | 0.83 | 112 |
| macro_avg | 0.78 | 0.75 | 0.75 | 185 |
| wght_avg | 0.78 | 0.78 | 0.77 | 185 |

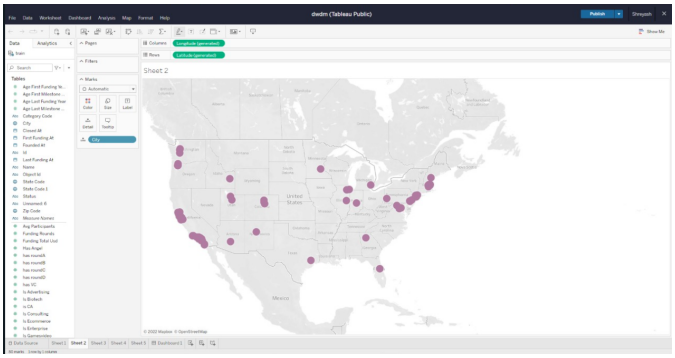
## VII. SYSTEM ARCHITECTURE

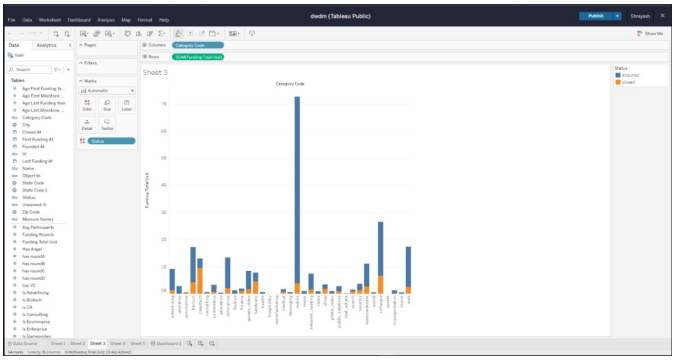


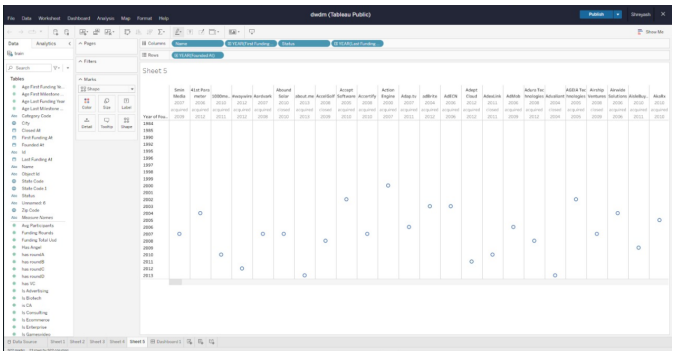

## VIII. RESULTS

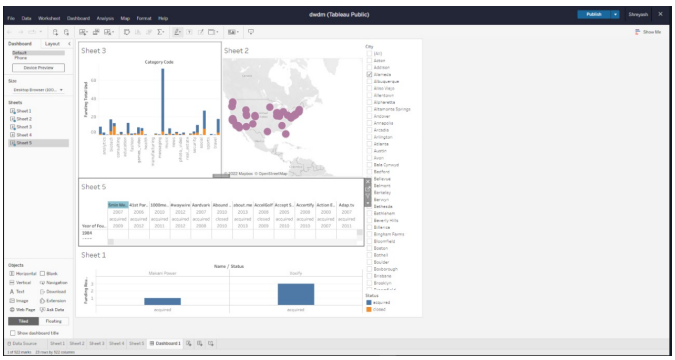
We read past few research papers based on this topic we came across a very interesting observation that meaning of the research paper included the approach like revenue generated age first funding your age last year but I thought that this warrant that robust approaches so we went for merge and acquire approach that is whenever the company is win merged or acquired by another Angel investor or venture capitalist it is likely to be success and we have trained our model based on the strategy











## IX. CONCLUSION

The main objective of the present study was to generate a model to classify successful companies or start-ups. In this paper, we used a few machine learning algorithms to construct models for predicting success of early stage startups. Precision accuracies of *87.05%, 72.87%, 94.76%* and *77.84%* for models trained using *SVC, Random Forest, LGBMClassifier* and *Decision Tree* respectively. 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.

## X. REFERENCES


- [1]Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory - COLT '92 (pp. 144–152). http://doi.org/10.1145/130385.130401
- [2]Artificial Intelligence and Machine Learning: Top 100 Influencers and Brands. (2016). Retrieved January 31, 2017, from http://www.onalytica.com/blog/posts/artificial-intelligence-machine-learning-top-100- influencers-and-brands/
- [3]Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. Retrieved from https://www.jair.org/media/953/live-953-2037-jair.pdf
- [4]Farrar, C. R., & Worden, K. (2012). Structural Health Monitoring: A Machine Learning Perspective - Charles R. Farrar, Keith Worden - Google Livros. Wiley. Retrieved from https://books.google.pt/books?hl=ptPT&lr=&id=2w_sp6lersUC&oi=fnd&pg=PP11&dq=machine+learning+health&ots=E1vmyBFsvo&sig=Mavuhd4Aq5DqiafMeP8nhHmyPOg&redir_esc=y#v=onepage&q=machine learning health&f=false
- [5]Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News. Retrieved from https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-byRandomForest.pdf
- [6]Lennon, M. (2014). CrunchBase Data Export Now Includes International Startups, Investors -. Retrieved October 20, 2017, from https://about.crunchbase.com/blog/crunchbase-data-export-now-includes-internationalstartups-investors/


DATASET-
https://raw.githubusercontent.com/dphi-official/Datasets/master/startupdata/training_set_label.csv
https://raw.githubusercontent.com/dphi-official/Datasets/master/startupdata/testing_set_label.csv