Seminar Report

On

STARTUP SUCCESS PREDICTION USING MACHINE LEARNING

By

Akshay Lokhande

PRN: 1032192238

Under the guidance of

Prof. Sarika Bobde

MIT-World Peace University (MIT-WPU)
Faculty of Engineering
School of Computer Engineering & Technology

*** 2021-2022 ***

MIT-World Peace University (MIT-WPU)

Faculty of Engineering

School of Computer Engineering & Technology

CERTIFICATE

This is to certify that Ms. Akshay Vinod Lokhande of B.Tech., School of Computer Engineering & Technology, Trimester – IX, PRN. No. 1032192238,
has successfully completed a seminar on

STARTUP SUCCESS PREDICTION USING MACHINE LEARNING

To my satisfaction and submitted the same during the academic year 2021-2022 towards the partial fulfillment of the degree of Bachelor of Technology in School of Computer Engineering & Technology under Dr. Vishwanath Karad MIT- World Peace University, Pune.

Prof. Sarika Bobde
Seminar Guide

Prof. Dr. V.Y. Kulkarni
Head
School of Computer Engineering & Technology

ACKNOWLEDGEMENT

I wish to express my gratitude and thank my research guide Prof. Sarika Bobde from the school of CET from the bottom of my heart. She helped me throughout the duration of seminar preparation and the completion of this work with the documentation.

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 the 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 decisive 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

Keywords: SVM, M&A, Lightgbm,RF(Random Forest),DT(Decision Treee)

INDEX

List of Tables

List of Figures **Page No.**

1. 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

a 90% probability of failure, which means a lot of investments without proper returns

2. Literature Survey

Table No. 1

Sr. No.	Paper Name	Authors' Name	Description	Limitations
1.	Predicting Startup Success with Machine Learning	Paper presented as requirement for obtaining the Master's degree in Information Management Published in Nov 2017	**1**)Merge and Acquisition approach 2)IPO	Sparse dataset,
2.	Startup Success prediction in Dutch Ecosystem	Delft University of Technology Conference Published in Year 2019	158 variables studied for predicting dependent variables	variables are not grouped into any particular categories in the model

3.	A machine learning model for startup selection and exit prediction	Venhound Inc., DE, USA	Two ML ensemble, One for exit and one for funding	difficulty in the IPO vs acquisition outcome
4.	Information Processing and Managemen t	Warsaw University of Technology	How is the business Market regarding to idea	
5.	Web- based start-up success prediction	Michael Roizner and also with university of amsterdam	WBSSP approach (Web-Based Startup Success Prediction)	in addition to tracking only the sources of startup mentions
6.	A machine learning, bias-free approach for predicting business success using Crunchbase data	Kamil Bukowski	1)Support vector machine 2)XGBoost	Increase recall of the model
7.	A Machine Learning Approach Towards Startup Success Prediction	IRTG 1792 Paper, No. 2019-022	1. Adaptive Synthetic 2. Sampling Approach (ADASYN)	

8.	Predicting the Outcome of Startups: Less Failure, More Success .	IEEE 16th International Conference on Data Mining Workshops	Predictive Modeling	Depends on data quality

3. Proposed methodology/algorithm

3.1 Dataset Description

We've obtained this dataset from GitHub and converted it into a CSV file in order 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 split the dataset into 80-20 % for training and testing the data respectively.

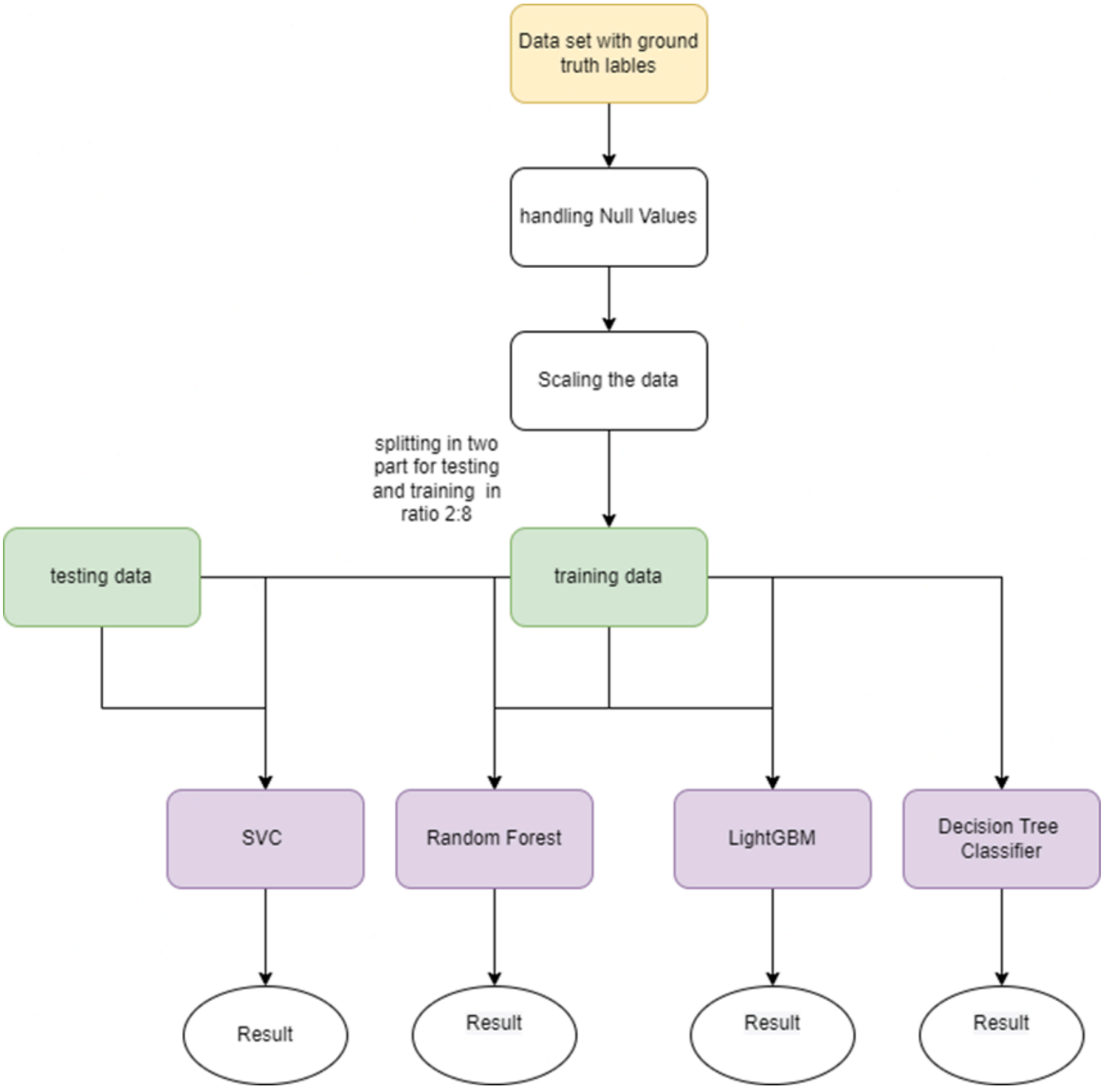


fig.1 System Architecture

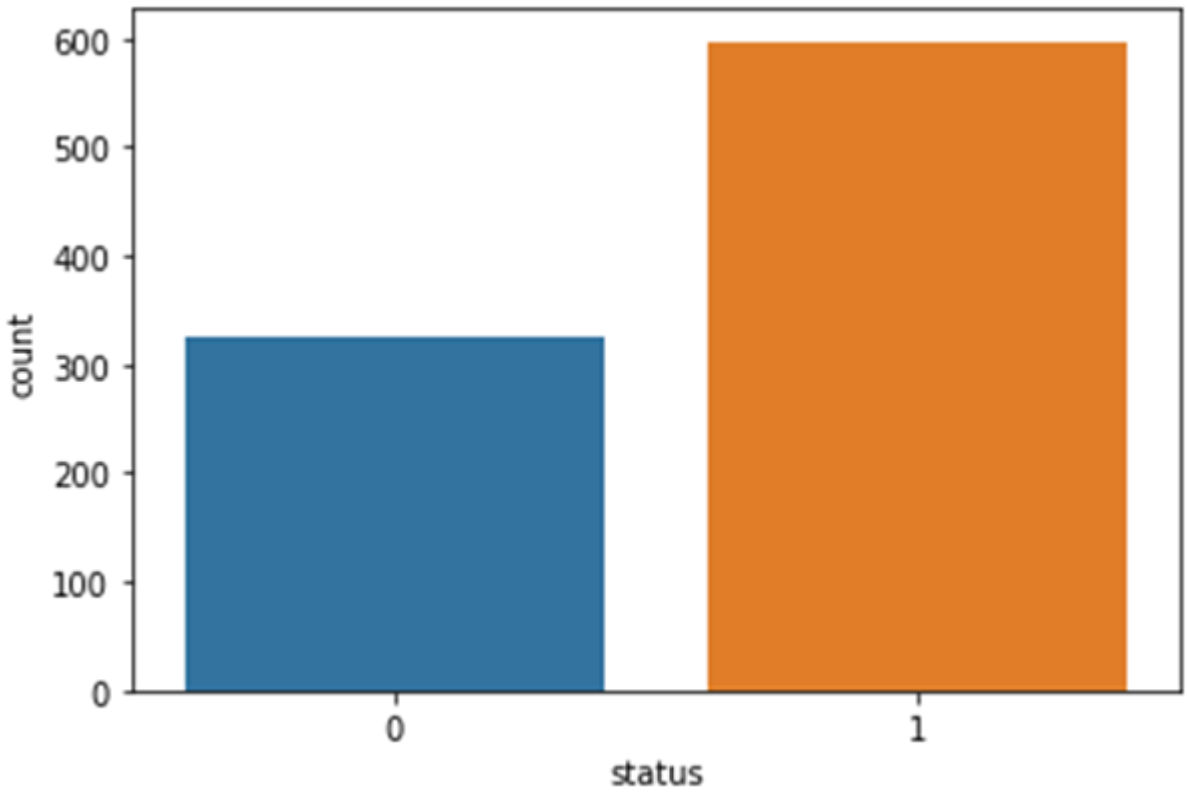


Fig. 2 Plot showing acquisition status

3.2 Dataset Visualization

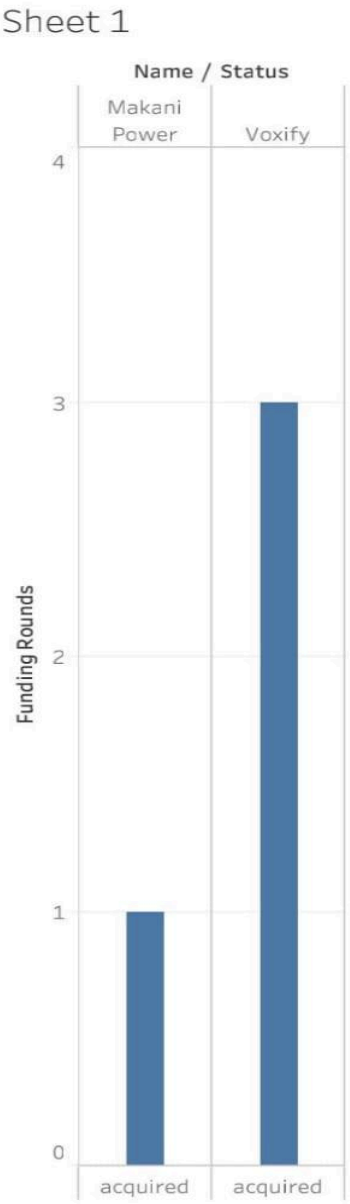


fig.3 Funding round status of selected companies.

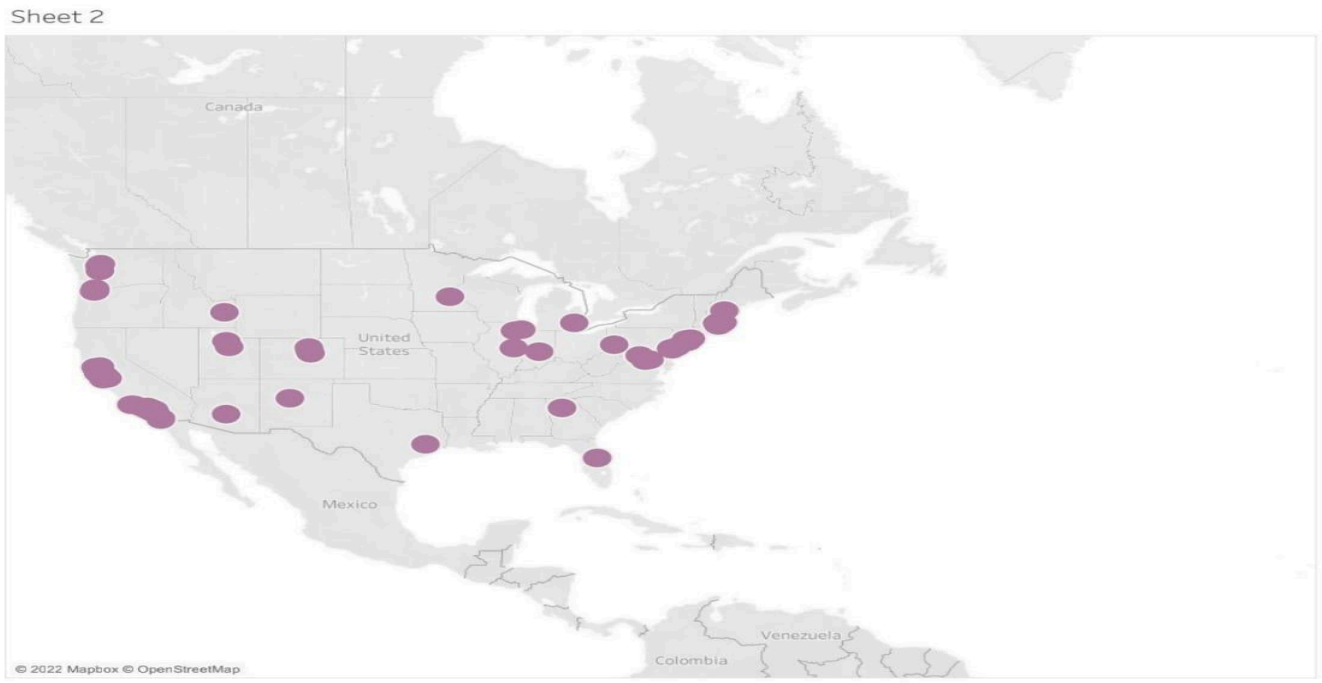


Fig.4 Location map of startups

According to the above location-based map we can conclude that Startups that are founded in urban areas are most likely to succeed.

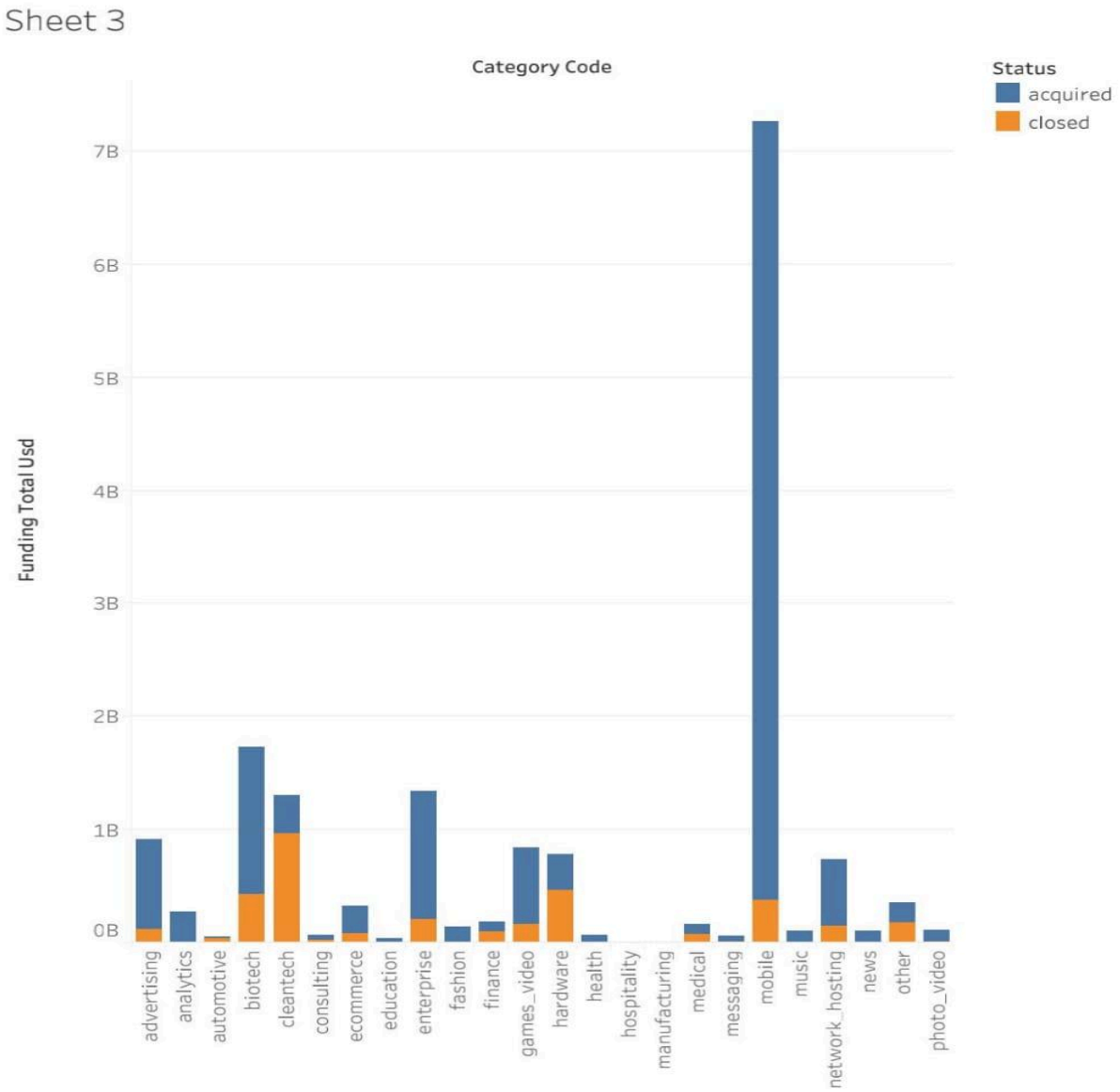


fig.5 Diversified investment in various startup categories

3.3 Algorithms

3.3.1 Support Vector Machine(SVM)

The aim of a Linear SVC (Support Vector Classifier) is to fit 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.

3.3.2 Random Forest

Random forest classifier is a meta-estimator that fits several decision trees on various sub-samples of datasets and uses an 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.

3.3.3 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 parallel and GPU learning.
- Capable of handling large-scale data.

3.3.4 Decision Tree Classifier

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.

- Following is the graph plotted for Decision Tree Classifier.

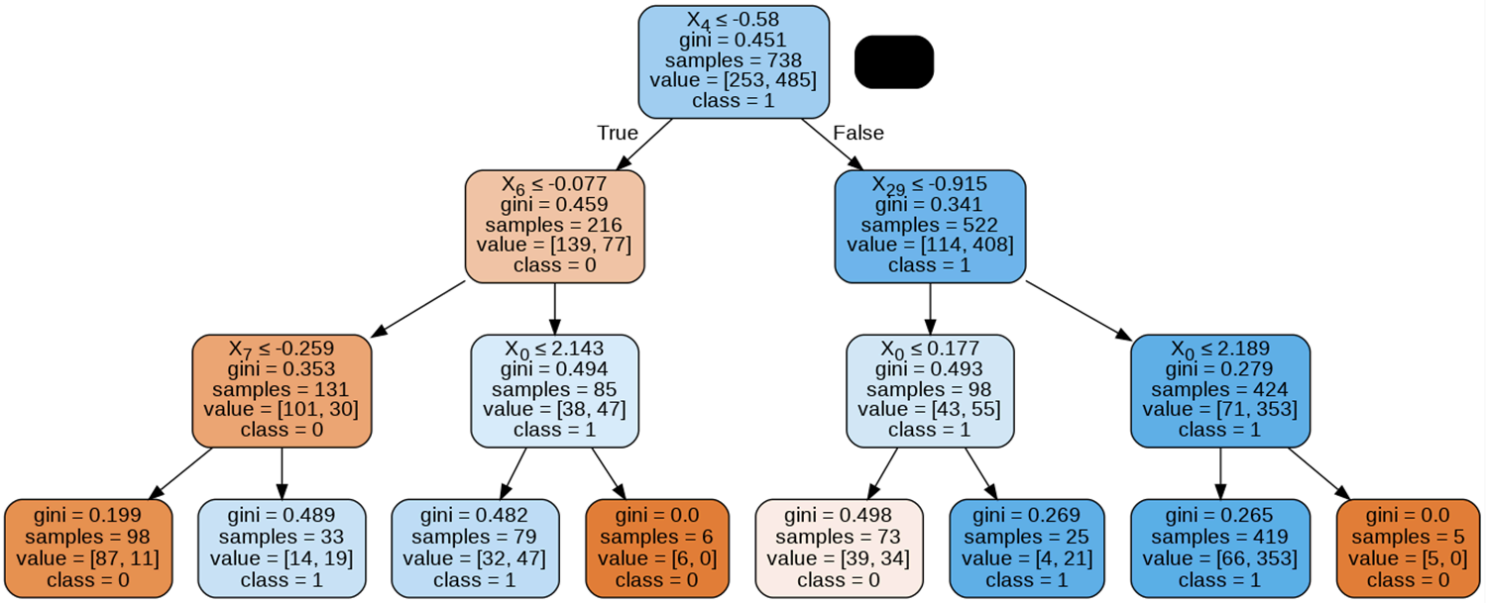


fig.6 Decision Tree

4. Experimental Work:

SVC

	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

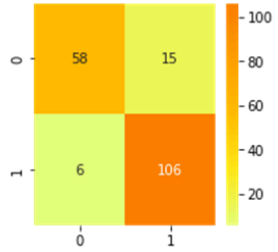
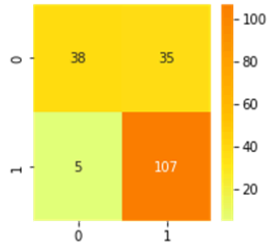


fig.7 Confusion matrix for SVC

RANDOM FOREST

	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



16

fig.8 Confusion matrix for RF

LGBM CLASSIFIER

	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

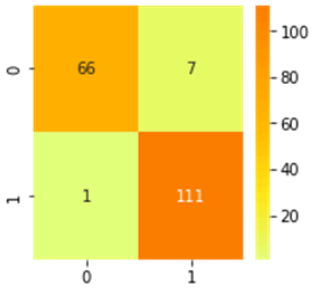


fig.9 Confusion matrix for LGBM

DECISION TREE

	Precision	Recall	f1-score	support
0	0.75	0.45	0.56	73
1	0.72	0.90	0.80	112
macro_avg	0.73	0.68	0..68	185
wght_avg	0.73	0.72	0.71	185

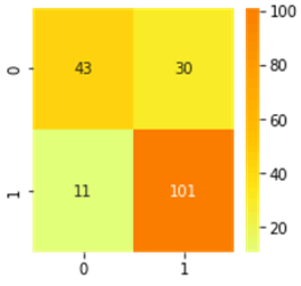


fig10 Confusion matrix for DT

4.1 Performance Achieved:

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

4.2 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.

5. References

[1] [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] [2]Artificial Intelligence and Machine Learning: Top 100 Influencers and Brands. (2016). Retrieved January 31, 2017, from

[3] http://www.onalytica.com/blog/posts/artificial-intelligence-m achine-learning-top-100-influencers-and-brands/

[4] [3]Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer,

[5] 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

[6] [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_sp6lers

UC&oi=fnd&pg=PP11&dq=machine+learning+health&ots= E1vmyBFsvo&sig=Mavu

hd4Aq5DqiafMeP8nhHmyPOg&redir_esc=y#v=onepage&q

[7] =machine learning health&f=false

[8] [5]Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News. Retrieved from https://www.researchgate.net/profile/Andy_Liaw/publication

[9] /228451484_Classification_and_Regressio

n_by_RandomForest/links/53fb24cc0cf20a45497047ab/Cla

ssification-and-Regression-byRandomForest.pdf

[10] •[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

[11] -now-includes-internationalstartups-investors/

[12]

6. Plagiarism Check report