International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

STARTUP SUCCESS RATE PREDICTION

Aachal Funde, Hemangi Chopde, Vaishnavi Thombre, Sahil Bhure

Dr. Mrs.Punam U.Chati (Guide)

Department of Electronics & Telecommunication Engineering

Priyadarshini College of Engineering, Nagpur

1. Introduction

We're using the 50-startups dataset for this problem statement and we will be using the concept of Multiple linear regression to predict the profit of startups companies.

How do startups work?

Well, we can say that startups pipeline operates on the same principles which are similar to other MNCs the major difference between both of them is that on the one hand startups work to make products that are beneficial for the customers on a small scale while other established companies do that work on a large scale by re-doing something which is already being done.

How startups are being funded?

As I mentioned above the startups are not such economical balanced company that has covered a path from an idea to a product so for the same reason no established investor will be going to come forward for those

companies which don't really have their market value hence, startups allow the early investors to start supporting in the format of seed funding which would help them to make a product out of their idea. In a nutshell, we can see that it's hard to manage and analyze the investments and to make a profit out of them. We need a way by which we can analyze our expenditure on the startups and then know a profit put of them!

How this model can help here?

This machine learning model will be quite helpful in such a situation where we need to find a profit based on how much we are spending in the market and for the market. In a nutshell, this machine learning model will help to find out the profit based on the amount which we spend from the 50 startups dataset.

2. Literature survey

There are existing works that have modelled the opportunity evaluation process in the context of entrepreneurship. Many of the existing works evaluate the success of business opportunities by utilizing social, cultural and personal factors. The outcomes of the opportunity evaluation are then used by nascent entrepreneurs to decide whether or not to proceed. The objective of the elaboration is to understand the opportunity in a way to reduce the uncertainties by identifying the anticipated problems and to maximize the potential benefits (Van der Veen and Wakkee, 2006). Despite the rich literature in opportunity evaluation in entrepreneurship, there are only few studies associated with the impact of uncertainty factors in entrepreneurship. De Koning and Muzyka (De Koning and Muzyka, 1999) identify a socio-cognitive framework of opportunity recognition by considering three cognitive activities such as information gathering, thinking through talking and resource assessing through the interacting with the extensive network of people. The success of young firms plays a crucial role in our economy since these firms often act as net creator of new jobs (Henrekson and Johansson, 2010) and push, through their product and process innovations, the societal frontier of technology. Success stories of Schumpeterian entrepreneurs that reshaped entire industries are very salient, yet from a probabilistic point of view it is estimated that only 10% of startups stay in business long-term (Griffith, 2014;Krishna etal., 2016). Hence, many researchers define startups based on the available information in their data set (Luger and Koo, 2005). In this paper, the definition of startups is based on the

available data. Hence, companies are considered as startups, which are active in industries defined by S&P500 with years in business not more than 10 years. Due to lack of information, it is not possible to identify and exclude spin-offs and startups that are founded by larger corporations.

The following table summarizes the various papers and highlights the different methods that can be employed in order to use machine learning for predicting the success of start-ups .

Sr. No.	Paper Title	Methodology	ML algorithm used	Results
1	Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method [3]	Hybrid intelligence method: Machine and collective intelligence	Logistic Regression, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Random Forest	Mathew correlation Coefficient is used as an evaluation metric and it was shown that hybrid approach gives better results than a machine or human only prediction
2	Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments[5]	Multi-class approach, focus on early-stage companies, Time-aware analysis and Multi-class prediction problem	Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Extremely Randomized Trees (ERT) and Gradient Tree Boosting (GTB).	The results of the study show a global accuracy of around 82% of the best algorithm, Gradient Tree Boosting.

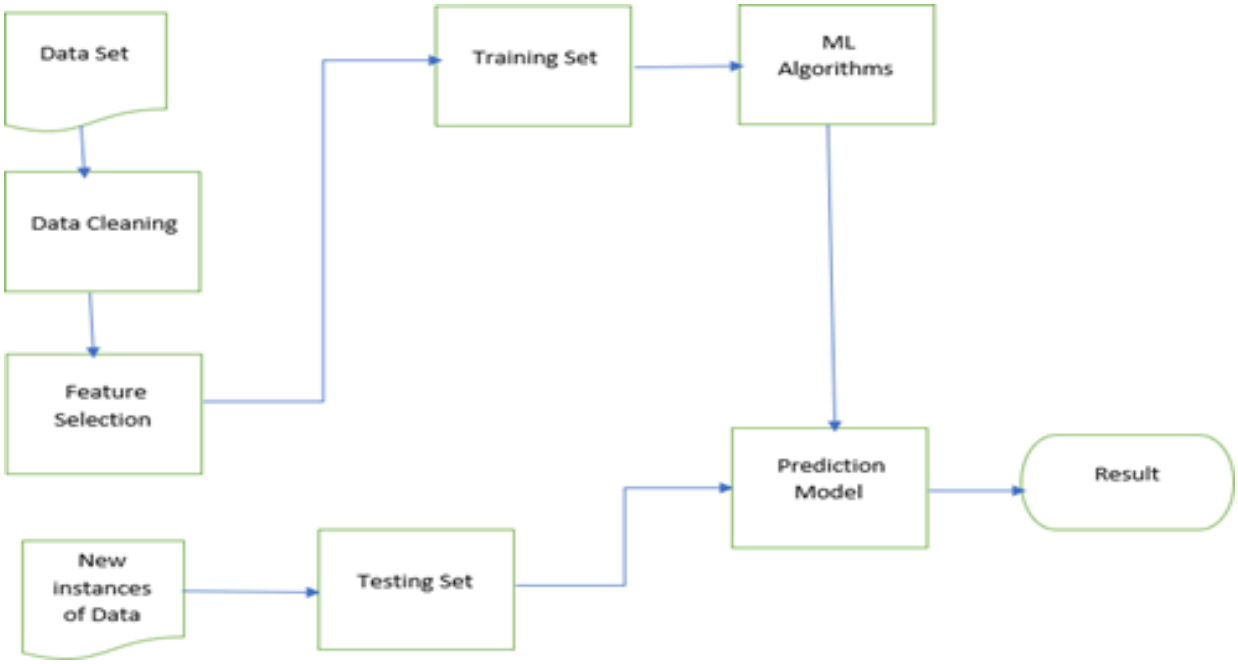
3	Creating Classification Models from Textual Descriptions of Companies Using Crunchbase[2]	Multilabel text classification based on textual description of a company.	Multinomial Naive Bayes, Support Vector Machines, and Fuzzy Fingerprints	Accuracy is above 65% with multiclass approach
4	Machine Learning Prediction of Companies' Business Success[4]	Classification using Supervised models	Logistic Regression, Random Forests, KNN	F1 scores are used as primary metric, KNN gave the highest F1 score

3. Problem Definition

Startup is a business that has just been established and grown supported by digital services and has also become an important element of innovation systems and economies around the world. The Startup ecosystem is growing very rapidly and still needs a lot of funding to operate with a minimalist working group. So it is very important for VC to monitor the performance and performance of Startup, so that it can be used as a consideration to decide whether to fund a Startup to drive its growth or refuse to take part in funding. To monitor startup performance, it is important to analyze what makes a Startup successful and how to determine its success.

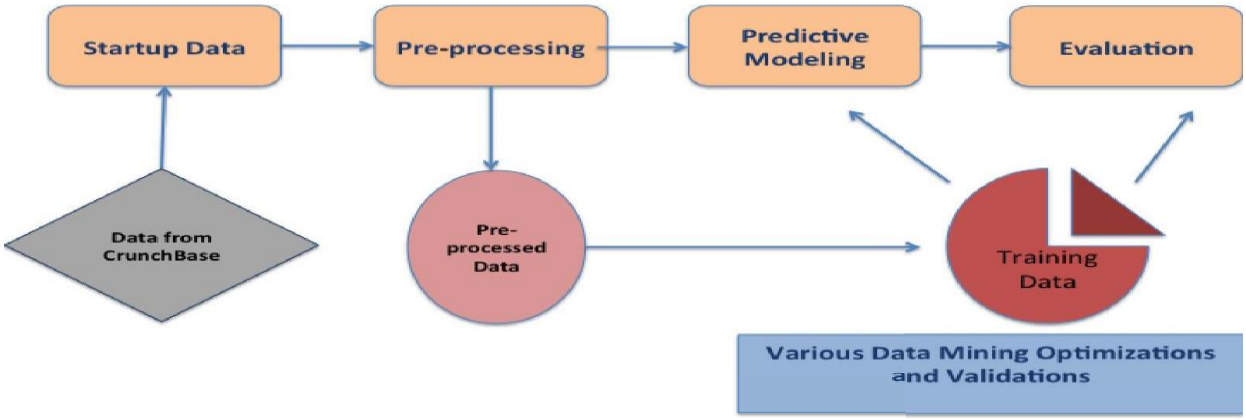
There are numerous startups every year, but start-ups fail to make it big many a times or even to survive for that matter. Huge amounts of funding are invested in many start-ups, but the question remains as to which start up a venture capitalist fund into. It is of primary importance to understand what makes businesses successful and predict the success of a company. Machine Learning based approaches have been used lately for this task. But still human intelligence cannot be questioned as humans are still considered as the "gold standard" when it comes to predicting in presence of risk factors. There are various machine learning approaches which are used to help the venture capitalists in selecting the right start-ups.

4. Proposed System Block Diagram



Flow of predicting start-up success using Machine Learning based approach

Working of system



Block Diagram of Success/Failure Prediction System

4.1 Data Collection

Data can be collected from the most widely used source. There are other sources too from where data could be collected for experimentation and study purpose.

The data set is split into 2 sets, namely training set and testing set. As the name implies, training set is used to train the model. The model is then tested using the testing data set.

4.2 Data Cleaning & Preprocessing

Data is cleaned to remove all redundant, irrelevant, duplicate information, missing values are cleaned, unused fields are discarded and outliers are removed. If the study models the categories of a company based on the text available in the description field of the company, one can make use of NTK to remove the punctuation marks, to remove the stop-words.

4.3 Feature Selection:

The most essential features to companies' business need to be identified and used. The most widely used features are:

- Category of the company
- Funding total used
- Funding rounds
- Funding duration
- Number of Unique Investors
- Known Investor Count
- First Funding at UTC

- Last Funding at UTC

The context of the study should be kept in mind in order to define which data to be included into the final data set or features essential for the study.

4.4 Models and Algorithms

Supervised learning algorithms make predictions based on the training data fed to it. In supervised learning, if x represents the input variables and y represents the output variable, the algorithm learns to map the function (y=f(x)) and then can correctly predict or classify after getting new input data x. So, the model is trained using the training data first.

Some of the approaches that have been used in the study of predicting business' success are:

Logistic Regression: In LR, the relationship between dependent and independent variables is found out wherein the dependent variable can take binary values – "0" or "1" or "not successful" or "successful". In ML, LR is one of the simplest and fastest algorithms and therefore it is used as a starting point for many classification problems.

Support Vector Machines: It is a vector-space-based machine learning method where the goal is to find a decision boundary between two classes that is far from any point in the training data, outliers could be discarded."

Random Forest:
Random Forest is a collection of Decision Trees. The goal of Random Forest is to prevent overfitting which it does by creating the random subsets of features and building shallower trees using the subsets. In RF, at each split point in the decision tree, only a subset of features is selected to take into consideration by the algorithm. The candidate features are generated using bootstrap.

Naive Bayes : A Naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. Depending on the precise nature of the probability model, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood .

4.5 Training the model

The model is trained by applying the training set to the model. The model is then tested using the testing data set.

4.6 Evaluation

Most often to evaluate the classification techniques, metrics such as Accuracy and F-score are used.
Accuracy is defined as the rate of correct classification. F1 score is the harmonic average of Precision and recall. Precision estimates how many positively identified samples are correct, and Recall estimates what proportion of positive samples was correctly identified.

Accuracy= (TP+TN)/ (TP+TN+FP+FN)
Precision = TP/(TP+FP)
Recall = TP/ (TP+FN)(3) F1= 2*(Precision*Recall)/(Precision+ Recall)

Confusion matrix is used to describe the performance of a classification model.
TP (true positive): an outcome where the model correctly predicts the positive class.
TN (true negative): an outcome where the model correctly predicts the negative class.
FP (false positive): an outcome where the model incorrectly predicts the positive class.
FN (false negative): an outcome where the model incorrectly predicts the negative class.

	0(Predicted Negative)	**1, (Predicted Positive)**
0 (Actual Negative)	True Negative (TN), company classified as not successful and it is not successful	False Positive (FP), company classified as successful and it is not successful
1(Actual Positive)	False Negative (FN), company classified as not successful and it is successful	True Positive (TP), company classified as successful and it is successful

5. Applications

1) Statistical analysis provides accurate and reliable predictions even for financial decisions
2) it comes to new ventures, and predicting their successes and failures
3) Collective intelligence and Machine Intelligence are aggregated and evaluated. It was proved that hybrid approach gives better results than a machine or human only prediction

References

[1] H. Janáková, "The Success Prediction of the Technological Start–up Projects in Slovak Conditions", Procedia Economics and Finance 34 (2015) 73 – 80
[2] G. Ross, S. Das, D. Sciro, H. Raza, "CapitalVX: A machine learning model for startup selection and exit prediction", The Journal of Finance and Data Science 7 (2021) 94-114
[3] U. Kaiser, J. M. Kuhn, "The value of publicly available, textual and non-textual information for startup performance prediction", Journal of Business Venturing Insights 14 (2020) e00179
[4] Q. Zhang, T. Ye, M. Essaidi, S. Agarwal, V. Liu, B. T. Loo, "Predicting Startup Crowd-funding Success through Longitudinal Social Engagement Analysis", CIKM'17, November 6–10, 2017, Singapore, ACM ISBN 978-1-4503-4918-5/17/11
[5] A. Prohorovs, J. Bistrova, D. Ten, "Startup Success Factors in the Capital Attraction Stage: Founders' Perspective", Journal of East-West Business, DOI: 10.1080/10669868.2018.1503211