Machine Learning based Outcome Prediction of New Ventures: A review

Swati Varma,
Department of Information Technology,
Vidyavardhini's College of Engineering and Technology,
Vasai,India

Abstract- - **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. This paper aims to provide a survey of different approaches with their advantages and disadvantage.**

Index Terms* – *Classification, Start-ups, Ventures, Machine Learning, Prediction, Venture capitals

I. INTRODUCTION

Start-ups/new ventures define a country's future. They play a major role orienting the economic revolution of a country. A country invests and motivates new ventures to help support the economic and industrial revolution. But helping a right start-up is equally important, especially for venture capitalists. The prowess of machine learning can be exploited to aid the venture capitalists in deciding which start-ups to finance so that funds can be invested in right start-ups.

While entrepreneurship research faces many challenges, the analytical capabilities of computational science may come to our rescue. Many studies and even VCs focus on predicting whether a company will sustain in the market, let alone finding the potential unicorn. Human intelligence can never be questioned but that too comes after spending years in the industry. Machine learning techniques can be useful in collating, processing, identifying and then predicting the success of new companies. Some work has already been started in this field. This paper summarizes and compares work by different authors.

The rest of the paper is organized as follows: section II outlines the related work in this field, Section III gives the flow of Machine Learning based approach. Section IV compares and summarizes various approaches that have been used in this field and in Section V conclusion is presented.

II. RELATED WORK

Predicting success of start-ups is a difficult task, mainly because of the complexity and uncertainty involved. Machine learning capabilities can be exploited to make things easy and reliable for us. Statistical analysis provides accurate and reliable predictions even for financial decisions [1]. Machine learning has become a popular tool for researchers to use in the domain of finance and investment [4]. The most important element of any ML based approach is the data set. When it comes to new ventures, and predicting their successes and failures, Crunhbase database is widely used in statistical analysis[2]. Many authors and researchers have used Crunchbase for their studies.

The work presented by [2] describes multi-label text classification experiments over a dataset extracted from Crunchbase. Three classification approaches namely, multinomial Naive Bayes, SVM, and Fuzzy fingerprints, have been used along with different combinations of text representation features. The experiment yields precision of 70% and recall of 42%. When multi-class approach is being used, the accuracy is above 65%[2].

The work presented in [3] highlights the importance of human decision making and proposes a design science research approach which is a Hybrid Intelligence method that combines the strength of both machine and collective intelligence to demonstrate its utility for predictions under extreme uncertainty. 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 [3].

The work presented in [4] is a pure ML based approach and the data set has been used from crunchbase. The primary goal is to predict the status of the start-ups, whether they have gone through M&A(i.e. Merger & Acquisition) or IPO (Initial public offering). The ML approaches that have been used are Logistic Regression, Random Forests, and K Nearest Neighbors. The study compares different ML approaches stated above and identifies the best algorithm with the said dataset, F1 scores are used as primary metric, and it was noted that KNN gave the highest F1 score[4].

The work presented in [5] focusses not on the conventional two categories of classification, namely M&A and IPO but also on more possible outcomes subsequent funding round or the closure of the company. In their study a VC investor would be provided with more information to set up a portfolio with lower risk that may eventually achieve higher returns and not just finding the unicorns. Again, the popular database, crunchbase has been used in this study. The main features of their approach are focus on early-stage companies, Time-aware analysis and Multi-class prediction problem. ML algorithms that are used in this study are 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[5].

III. MACHINE LEARNING BASED APPROACH

Machine learning has evolved from the efforts of scientists like Arthur L. Samuel exploring whether machines could learn to play games like checkers to cover a broad spectrum of applications. Machine learning approaches could also be successfully used in the problem of predicting the success of new ventures. There are various studies available to support the same. Though the accuracy can be improved, these studies have been a significant achievement in the field.

The detailed explanation of the diagram in “Fig 1” is as follows:

3.1 Data Collection

Data can be collected from the most widely used source Crunchbase. There are other sources too from where data could be collected like TechCrunch[2], Mattermark, Dealroom[3] 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.

3.2 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. Data could also be transformed if needed.

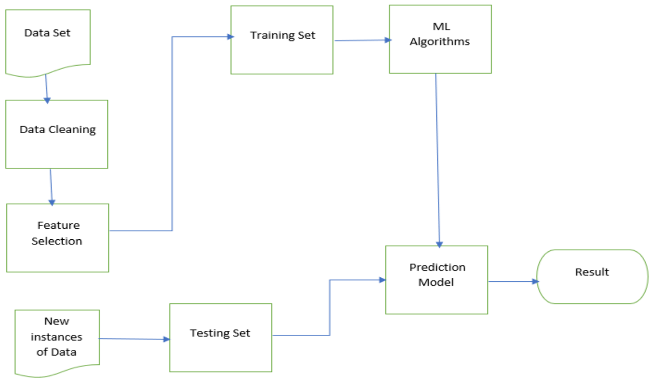

Fig 1:- Flow of predicting start-up success using Machine Learning based approach

Various NLP tools could also be used to pre-process the dataset before it is advanced to feature selection and training usage. 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.

3.3 Feature Selection:

If the corpora includes textual features, representation techniques like bag-of-words, Word2Vec could be used in order to deal with them.

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.

3.4 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 represesnts 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[6].

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.

NaiveBayes : 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[8].

Artificial Neural Network:

The concept is borrowed from the biological neural network. They are efficien computing systems, consisting of large collection of units which are interconnected in some pattern. These units, called as neurons, operates in parallel.

The multilayer perceptron (MLP) is an ANN model that uses a back-propagation algorithm in the training process. MLP has three components: an input layer, hidden layers, and an output layer. Information is carried from one neuron to another by the weight value. the inputs (independent variables) propagate forward using the sigmoid or logistic-activation function, thus producing output values (dependent variables) for each hidden layer. After that, the error is propagated backward by updating the weights and biases. Errors are computed and then weights and biases are updated.The steps are repeated until the overall error is minimized[7].

3.5 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.

3.6 Evaluation

Most often to evaluate the classification techniques, metrics such as Accuracy and F-score are used.
Accuracy(1) is defined as the rate of correct classification. F1 score(4) is the harmonic average of Precision and recall. Precision(2) estimates how many positively identified samples are correct, and Recall(3) estimates what proportion of positive samples was correctly identified.

$$Accuracy= (TP+TN)/ (TP+TN+FP+FN) \quad (1)$$
$$Precision = TP/(TP+FP) \quad (2)$$
$$Recall = TP/ (TP+FN) \quad (3)$$
$$F1= 2*(Precision*Recall)/(Precision+ Recall) \quad (4)$$

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.

TABLE I. CONFUSION MATRIX

	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

IV. EXISTING MACHINE LEARNING BASED LITERATURE SURVEY

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 .

TABLE II. SUMMARY OF DIFFERENT METHODS OF START-UP SUCCESS PREDICTION

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)	The results of the study show a global accuracy of around 82% of the best algorithm, Gradient Tree

			and Gradient Tree Boosting (GTB).	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

V .CONCLUSION

Predicting the success of startups is a tough task and also the costs of misclassification is high which can lead to wrong funding decisions. The literature on predicting start-up success is recent and the work done is still exploratory. Machine learning can help a great deal to venture capitalists in basic screening of start-ups. The accuracy given by the earlier studies is good enough, but the study shows that there is still a chance of improvement and the values of the metrics could be improved further. Although machine learning has proven to be a beneficial and usable solution, techniques for handling extreme uncertainties should also be considered. An hybrid approach appears to be a promising solution.

VI .REFERENCES

[1] HuiYuan., Lau, Raymond Y.K.Lau, Wei Xu,, " The determinants of crowdfunding success: A semantic text analytics approach", Decision Support Systems (91), pp. 67–76.

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

[3] Dominik Dellermann, Nikolaus Lipusch, Philipp Ebel, Karl Michael Popp, Jan Marco Leimeister," Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method", Thirty Eighth International Conference on Information Systems, South Korea 2017.

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

[5] 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

[6] F. R. da Silva Ribeiro Bento, ``Predicting start-up success with machine learning," M.S. thesis, Dept. Inf. Manage., Universidade Nova do Lisboa, Lisbon, Portugal, 2018.

[7] Serpil Kılıç Depren, Öyküm Esra Aşkın, Ersoy Öz," Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS", Educational Sciences: Theory & Practice, 17, 1605–1623

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