Startup Success Prediction and Feature Selection

Avi Sahal
Mechanical and Automation Engineering, IGDTUW
Indira Gandhi Delhi Technical university for Women
Delhi, India
avisahal020202@gmail.com

Srishti Dwivedi
Mechanical and Automation Engineering, IGDTUW
Indira Gandhi Delhi Technical University for Women
Delhi, India
srishtidwivedi231@gmail.com

Abstract— ***Predicting the success of a startup has always been a struggle for both practitioners and researchers. In this study, we use data obtained from one of the largest machine learning dataset platforms – Kaggle. This work aims to create a predictive model based on machine learning for the purpose of forecasting a startup company's success. Many similar attempts have been made in recent years. Plenty of those experiments, often conducted with the use of data gathered from several different sources, reported promising results. However, we found that very often they were significantly biased by their use of data containing information that was a direct consequence of a company reaching some level of success (or failure). Such an approach is a classic example of the look ahead bias. We designed our experiments in a way that would prevent the leaking of any information unavailable at the decision moment to the training set. We compared six algorithms on the training dataset. Despite the conscious decision to limit the number of predictors, we reached very promising results in terms of precision, recall, and F1 scores which, for the best model were 81.60%. The best outcomes were obtained with the adaboost classifier. We give detailed information about the importance of different features, with the top two being startup category and its funding rounds.Our model can be applied directly as a decision support system for different types of venture capital funds.***

I. Introduction

A startup is a technology-based company that offers a new product or service using the added value of the incorporated technology. In recent years, advances in new technologies have meant that companies have adopted new business models that incorporate globalization and using the Internet as a promotion tool for products and services. With the evolution of technologies since the first decade of the 21st century, these business models have been adapting to include new processes and social changes, as well as the new demands of consumers, who are increasingly supported in this new digital era where the use of new technologies has become a habit in both professional and personal worlds Startups use scalable business models, that is, startups make investments for the improvement of the technology on which they base their project, and once the technology has been improved, the product or service is created.

Startups operate in a very risky and extremely vulnerable environment, therefore the prediction of startup success is a crucial task for various stakeholders, including investors, entrepreneurs, and policymaker, as it has significant implications for resource allocation and decision-making. Predicting startup success presents a formidable challenge due to the inherently volatile landscape of the entrepreneurial ecosystem. It is estimated that approximately 90% of startups fail within their first five years, a failure rate that has remained relatively constant over the past few decades, despite considerable advancements in technology and business practices. Consequently, the accurate prediction of startup success can assist investors in more effectively allocating their resources and enable entrepreneurs to make better-informed decisions. There are very few startups that achieve a status of unicorn-startups.This is possible largely by identifying new opportunities which differentiate them, and getting strong competitive advantages through innovation in a niche market.

The startups must have knowledge of their success factors that are advantageous in their business, since they want to develop successful and profitable business models over time.

Consequently, this study aimed to identify the key factors that make a startup successful .We analyze the dataset from crunchbase which contained various factors such as categories, founding dates, funding rounds, types of startup etc. Among this we have used different algorithms to find which of the above factors play an important role in success of a startup.

This research applies a naive Bayes classifier to analysis of Facebook comments on agri-tech startups, in Thailand. This research applies a naive Bayes classifier to analysis of Facebook comments on agri-tech. The researchers start analyzing positive and negative comments and eliminated those which couldn't be analyzed in data cleaning and selection process to avoid errors in the process further[1]. Limitation to Kewsuwun et al. research is that small and concise data available since it is a niche market and too specific[1]. Here in this research they have predicted success of startup in their early stages using machine learning algorithms like logistic regression and neural networks by analyzing the dataset from crunchbase[2]. The study only included company's presence in the web space based on the mentions for prediction. Therefore full prediction is not made. Prediction quality doesn't saturate when a full set amount of incorporated mention approach. Only domain level mentions were considered therefore restricting the data.

Here the authors used several algorithms of supervised learning which are NaiveBayes, ADTrees, BayesNet, Lazylb1, RandomForest, SimpleLogistics to predict success and failure of startup at an early stage. Further values of roc curve and recall were studied[3]. Żbikowski et al[4]. decided to use 3 simple model whose implementation is easily available in machine learning. These models are logistic regression, Support Vector Machine and XGBoost. Recall metrics for all three classifiers is very low; i.e models have misclassified 70% of successful startups as unsuccessful. In future they want algorithms to discover more successful startup based on different pattern not only

based on popular ones[4].The authors have used MLP(multi-layer perceptron), random forest, XGboost, KNN and essemble performance to predict success of startups[5]. Potanin et al. [6] investigated MLP, Random Forest, XGBoost and used mostly numerical features from the dataset. Authors conducted a review on existing machine learning techniques that are recently contributed to understanding the need of start-ups, trends of business and can provide recommendations to plan their future strategies to deal with the business problems. Limitation to this is The current strict filters used to determine successful companies (IPO/ACQ/UNICORN) could also be loosened to potentially capture more companies in the "gray area" between success and failure.

The authors used various machine learning algorithms which included logistic regression, Naïve Bayes, KNN and SVM(Support Vector Machine) these studies relied heavily on either financial data provided by VCs, which is not accessible to the broader research community, or on qualitative data collected through questionnaires, which is very time-consuming and limited[7]. During this analyzes ,first, a Latent Dirichlet Allocation (LDA) model was used, which is a state-of-the-art thematic modeling tool that works in Python and determines the database topic by analyzing tweets for the #Startups hashtag on Twitter (n = 35.401 tweets). Secondly, a Sentiment Analysis was performed with a Supervised Vector Machine (SVM) algorithm that works with Machine Learning in Python to divide the identified topics into negative, positive, and neutral for the key factors that make a startup business successful. Thirdly, a Textual Analysis was performed on the results with Text Data Mining techniques using the Nvivo qualitative analysis software[8]. The limitations of this study are due to the size of the sample, the topic chosen for the study, and the methodological approach taken to reach the conclusions and implications presented. Bento et al. used logistic regression, Support Vector Machine and Random Forest algorithms to predict startup success.It should be noted that their best performances were achieved in categories with a higher number of observations while the ones in the present study didn't always follow that behavior[9]. Stahl et al. used Gradient Recurrent Units(GRU), Gradient Tree Boosting and CNN. One major drawback of the GRU model is the lack of explainability when it comes to determining the exact effects of signals on predictions[10].

To avoid the drawbacks of their work some alternatives or better work are suggested for the above researches. Kewsuwun et al. to avoid the drawback suggested future scope For further research , researcher advise on using sentiment analysis and use in-depth analyzes of interviews [1]. To avoid including company's presence from just webspace, further work can include sentiment analysis for discovering various mention pages. Discovering various indirect mentions must be included. Further distinguishing between these sections may provide us with a more fine-grained signal for predictive modeling[2]. To increase accuracy Future work must involve increasing accuracy and precision values incorporating more severity factors. Development web- based tool for current approach is also suggested[3]. Future works could increase the recall of the models by enriching the dataset. More detailed data about the founder's prior experience and the company's product or service could improve the performance of models. This increases the recall metrics for all three classifiers Żbikowski et al[4].

The future machine learning models can provide exit predictions in real time along with a feature analysis that identifies aspects of the company that make it a good investment or, on the other hand, raise red flags. Thus, the modern venture capitalist can peer into the black box and make investment decisions faster and more reliably [5]. Recent advances in learning interpretable models will be explored, which will pave the way for learning fair models and representations that are invariant to sensitive attributes such as gender, race, etc. Causal models aim to capture the underlying mechanism driving the decision-making process while ignoring other domain-specific factors. In the future, we aim to train independent models and detect anomalies and strong deviations from the model that may indicate new trends [7]. To overcome the drawbacks Future lines of research could improve the methodological process of text mining and increase the sample size to try to find new indicators for startups[8]. Bento et al [9]suggests the application of different algorithms to the same data source and simpler transformations to the dataset than those applied here to achieve similar results. Also, by providing an easy to use API, CrunchBase database could be turned into an operations tool which could be of use to funds, investors and all the other players operating in this space. Insights on this data as predictive models or segmentation are all explorations possible through the available data. Stahl et al. suggest for future work a thorough correlation study to deeply understand the impact of specific features on the prediction of early- and late-stage rounds. Further, we propose an extension of the feature subset to include extensive team and hiring information. Also, payment data could be of value here as a proxy for revenue. To provide even more value to investors, the model should ideally do predictions on smaller granularities than 24 months. . Also, for future work the expansion of the decision-support system into deal selection could be of great value. Here, we could predict which startup funding round results in a high return of investment [10].

The remainder of this paper is organized as follows: Section 2 is the literature review which reviews related works in the area of startup success prediction and machine learning. Section 3 describes dataset collection, preprocessing, and feature selection. Section 4 presents the experimental results of the supervised approach. Section 5 describes some other ideas about company and investor scoring. Finally, Sections 6 and 7 provide the conclusion of the study and discuss prospective research avenues in this domain.

II. LITRATURE REVIEW

[1] Uses Naïve Bayes to study and analyze opinion and categories them into positive and negative comments by using the classifier. It classifies the sentiments and attitude of people and investors. In future work it is advised to use sentiment analysis so as to overcome the drawback to less precise data and for further in-depth analyzes. [2] Here prediction of a startup's success is done by Logistic Regression, Neural Network, GBODT(CatBoost).The is data is divided into 4 broad categories based on the information collected from the dataset. Future work must involve inclusion of sentiment analysis so as to include all the related data present at other web-space too. [3] Success and failure of startups are predicted using NaiveBayes, ADTrees, BayesNet, Lazylb1, RandomForest, SimpleLogistics algorithms.Further stydy of recall and roc curve is done.

[4] Use of models are logistic regression,Support Vector Machine and XGBoost is done for prediction. Use of algorithms which increase the recall value is suggested for future work. [5] MLP(multi-layer perceptron), random forest, XGboost, KNN and essemble performance to predict success of startups. Machine learning models can provide exit predictions in real time along with a feature analysis that identifies aspects of the company that make it a good investment or, on the other hand, raise red flags.Pre-processing involves 2 steps, first importing relational database before use of sql and secondly use of python programming. [6] In order to identify successful companies i.c IP0/ACQ/UNICORN, use of MLP, Random Forest, XGBoost algorithms were used. [7] Various machine learning algorithms which included logistic regression**,** Naïve Bayes, KNN and SVM(Support Vector Machine) are used for analyzing data that relied heavily on either financial data provided by VCs, which is not accessible to the broader research community, or on qualitative data collected through questionnaires, which is very time-consuming and limited. [8] Identification of sentiment and categorizing them to positive and negative and neutral sentiments. This is done by Latent Dirichlet Allocation (LDA), Supervised Vector Machine (SVM), Text Data Mining techniques. Future line of work must include increase sample size so as to help find researchers new indicators. [9] Logistic regression (LR), Support Vector Machines (SVM) and Random Forests (RF) are used for determining companies acquisions.[10] proposed model has been optimized to handle time-series signals by using Gated Recurrent Units (GRU). Gradient Tree Boosting (GTB) , CNN. It displays good performance across early & late stage funding rounds, supporting a multi-stage strategy.

S no.	Title	Algorithm	Accuracy	Pre- Processing	Evaluation Parameters
1.	A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier	Naive Bayes	F-measure=75.00%, precision= 80.00% recall= 75.00%, accuracy= 61.00%	Study analyzes opinions and categorizes positive and negative comments by using naive Bayes classifier to examine the sentiments and attitudes of people and investors.	Recall,F-measure
2	Web-based Startup Success Prediction	Logistic Regression, Neural Network, GBODT(CatBoost)	ROC-AUC = 0.854 , P@100 = 0.626 , F@100 = 0.383 , P@200 = 0.535 , F@200 = 0.439	Data is preprocessed based on features classified into four broad categories according to the information sources that they capture: general, investor, people, and mentions	ROC-AUC , Precision , Fβ scores (β = 0.1)
3	Predicting the Outcome of Startups: Less Failure, More Success	NaiveBayes, ADTrees, BayesNet, Lazylb1, RandomForest, SimpleLogistics	Precision accuracies of 73.3%, 86.3%, 88.1 %, 87.5%, 86.7%, 88.4%, 86.4%, 87.9%, 87%, 96.3%	Data is preprocessed on the basis of Key factors likeSeed funding, Series A funding, Series B funding,Severity scores	ROC-AUC, Precision, Accuracy
4	A machine learning, bias-free approach for predicting business success using Crunchbase data	logistic regression, Support Vector Machine and XGBoost	Accuracy=0.86, Precision= 0.60, Recall= 0.33, F1= 0.43	Data is pre- processed by cleaning analyzing the enriching data, combining numerical features with data extracted from textual information crawled from the Internet.	Accuracy,Precision,Recall, F1
5	CapitalVX:A machine learning model for startup selection and exit prediction	MLP(multi-layer perceptron), random forest, XGboost, KNN and essemble performance	overall accuracy=88%, IPO precision=84%, Acquisition precision =82.5%, recall (IPO =77.5%, acquisition= 92.5%)	Pre-processing includes two steps : the first involves importing to a relational database before using SQL to transform and derive new features the second utilizes the Python programming language for fine-grained conditioning and additional derivation. The steps are outlined below and are applied before training the exit models as well as for the follow-on funding modeling, i.e., the same set of features are used for both ensembles.	Accurcy, Recall, Precision
6	Startup Success Prediction and VC Portfolio Simulation using Crunchbase Data	MLP, Random Forest, XGBoost	Precision=0.92, Recall=0.64, ROC-AUC=0.86, PR-AUC=0.65	Data pre-processing was done in such a way so as to divide the data into Initial Public Offering (IPO), Acquisition (ACQ), or Unicorn status (UNIC)	Precision,Recall, ROC-AUC, PR-AUC

7	Making it into a successful series A funding: An analysis of Crunchbase and LinkedIn data	Logistic Regression, Naïve Bayes, KNN and SVM(Support Vector Machine)	AUC=0.6 (±0.01) Acc = 0.64(±0.02) Prec=0.63(±0.02) Sens=0.64 (±0.02) Spec=0.64 (±0.02)	Descriptive data analysis was done by dividing the data into organization, people and investment data. Further data cleaning and analysis is done.	AUC, ROC
8	Detecting Indicators for Startup Business Success:Sentiment Analysis Using Text Data Mining	Latent Dirichlet Allocation (LDA), Supervised Vector Machine (SVM), Text Data Mining techniques		Identified the sentiments of a sample collected from social networks and divided them into negative, positive, and neutral sentiments.	
9	Predicting Startup Success with Machine Learning	Logistic regression (LR), Support Vector Machines (SVM) and Random Forests (RF)	True Positive Rate (TPR) =94.1% False Positive Rate=7.8% Precision=92.2% AUC= 93.2%	The process will follow general changes (as transversal to all thirteen tables in-use) and changes made to the organizations table as it is where all relevant information converges, ultimately becoming the training dataset of the task at hands. Due to the nature of the data and problem the priority is understanding its interdependence and not minimizing correlations.	AUC, Precision
10	Leveraging Time-Series Signals for Multi-Stage Startup Success Prediction	Gated Recurrent Units (GRU). Gradient Tree Boosting (GTB) ,cnn	Precision=85 %,	terms of pre-processing utilization three methods: Min-max scaling, log-transformation (clipped) and dense embedding. Missing values are imputed by either a median value for the datapoint class(funding round) or by interpolation (growth proxies).	Precision

III. METHODOLOGY

An experimental approach is taken where we determine which class a startup company is predicted to be in given its current state as represented by its features. In a broad sense the class a company can be in is either "successful" or "unsuccessful" depending on whether it exits the funding stage successfully or not.

Given failure can mean different things to companies and investors we define failure as (i) closed or (ii) acquired or (iii) ipo. This requires a robust model for classification while feature selection, extraction and model evaluation is continued iteratively to maximize accuracy, precision and AUC-ROC curve before formatting.

A. DATASET

The process of training the models begins by first identifying the data sources. We used Kaggle database export (CSV Export) as the primary data source, which is also supported by a well-documented API.

The main goal of this research was to collect a labelled dataset for training a machine learning model to classify companies as either successful or unsuccessful.

B. PREPROCESSING

The NULL values in columns like "name", "state_code", "city", "region" and "homepage_url" are replaced by a constant string value.

For columns with <5% NULL values,we have dropped the rows like "category_list". For columns with >5% and <60% NULL values, we have replaced them with mode like "country_code". For columns with >60% NULL values, we have dropped the columns. Columns like "homepage_url", "permalink", "name", "state_code", "region", "city" are dropped due to their insignificant role in influencing the success of a startup.

New columns like "diff_funding", "diff_funding_months", "diff_first_funding_months" are created to drop columns like "founded_at", "first_funding_at", "last_funding_at" because of their numerical complexity.

Finally, "status" is chosen as the deciding column for the machine learning model.

TABLE I.

Classification Algorithm	Evaluation Parameters			
	Precision	***Recall***	***F1 score***	***Overall Accuracy(%)***
Logistic Regression	0.82	1.00	0.90	81.55
Decision Tree	0.83	0.88	0.85	75.07
Random Forest	0.82	0.94	0.88	78.82
Support Vector Machine	0.82	1.00	0.90	81.60
Naive Bayes	0.83	0.93	0.87	78.20
K-Nearest Neighbors	0.82	0.94	0.88	78.67

Fig. 1. Classification Report of Various Machine Learning Algorithms

C. MODEL

Once the data is pre-proccesed, we segregated the data into training and test sets. We chose 20% of the total dataset to be used for testing. The rest of the dataset is used for training the model.

Various machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes and K-Nearest Neighbors are then applied on the training dataset with highest accuracy by Support Vector Machine algorithm.

Based on the above results, AUC-ROC curve is plotted. Thereafter Adaboost classifier is applied on the test dataset and the importance of each feature is plotted against the prediction of success of a startup.

IV.RESULT

We now describe the metrics that we have used to evaluate the quality of our predictions. First, we have used ROC-AUC, a standard classification metric. Second, for a clear measure of performance quality from a business perspective, we analyze the accuracy percentage of each machine learning algorithm. In a practical scenario an investor will only be able to fund a very small fraction of startups, so our interest lies with the algorithms that give us the highest accuracy.

TABLE -II

ACCURACY	ROC AUC
0.8056	0.522

Fig. 3. Evaluation Metrics

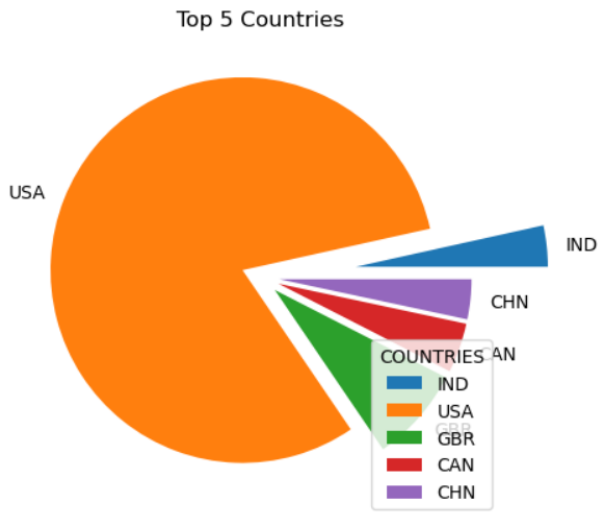

Fig. 2. Top five countries in terms of startup culture

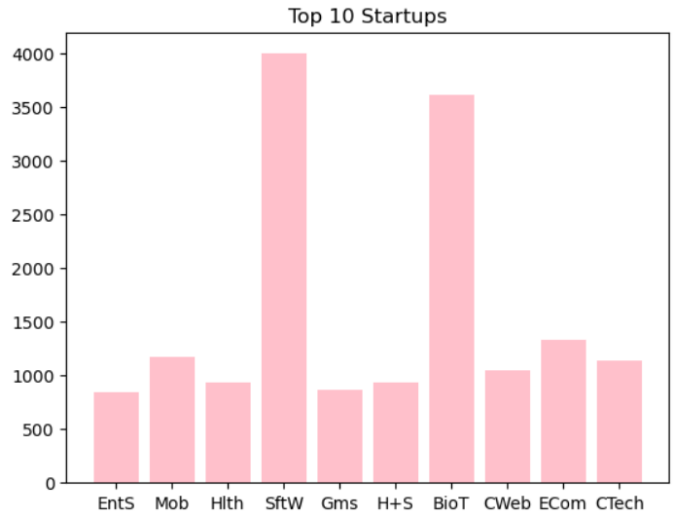

Fig. 4. Top ten domains of startups globally

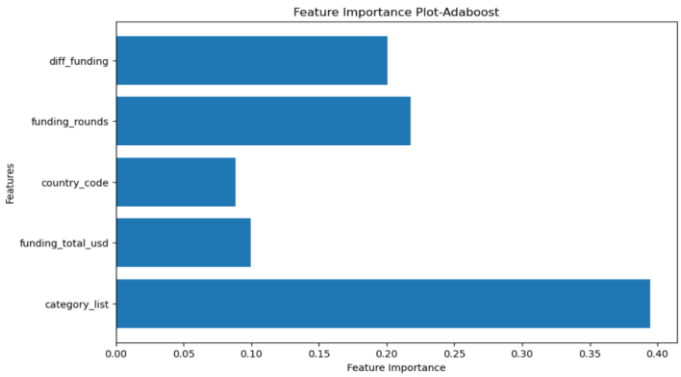

Fig. 5. Feature Importance Plot-Adaboost

V.CONCLUSION AND FUTURE SCOPE

In this paper, we addressed the problem of predicting the success of startup companies during their early stages of development. We utilized a rich and heterogeneous set of signals including data from Kaggle. We also developed a robust and diversified prediction pipeline based on a combination of several machine learning models. Besides building a predictive model, we contributed by providing a thorough analysis of this model and obtained results. Quite unexpectedly, structured company data such as category came out to be the most important feature for determining the success of a startup.

Despite the fact that we have addressed various limitations of previous research into startup success prediction, our work highlights several opportunities for improvement. First, the dataset should not have more than thirty percent of null values in any of the columns that can be a deciding factor in the prediction of whether a startup will succeed or fail.

Second, further work should be done to increase the AUC Area as a low value indicates presence of large number of false positive values in the dataset.

Third, the dataset should include columns that also contain information regarding the type of funding and the amount of funding received in various funding rounds.

REFERENCES

[1] Kewsuwun, N., & Kajornkasirat, S. (2022). A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier. International Journal of Electrical & Computer Engineering (2088-8708), 12(3)

[2] Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018, October). Web-based startup success prediction. In Proceedings of the 27th ACM international conference on information and knowledge management (pp. 2283-2291).

[3] Krishna, A., Agrawal, A., & Choudhary, A. (2016, December). Predicting the outcome of startups: less failure, more success. In 2016 IEEE 16th international conference on data mining workshops (ICDMW) (pp. 798-805). IEEE.

[4] Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management, 58(4), 102555.

[5] Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. The Journal of Finance and Data Science, 7, 94-114.

[6] Potanin, M., Chertok, A., Zorin, K., & Shtabtsovsky, C. (2023). Startup success prediction and VC portfolio simulation using CrunchBase data. arXiv preprint arXiv:2309.15552.

[7] Te, Y. F., Wieland, M., Frey, M., Pyatigorskaya, A., Schiffer, P., & Grabner, H. (2023). Making it into a successful series a funding: An analysis of crunchbase and linkedin data. The Journal of Finance and Data Science, 9, 100099.

[8] Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. Sustainability, 11(3), 917..

[9] Bento, F. R. D. S. R. (2017). Predicting start-up success with machine learning (Master's thesis, Universidade NOVA de Lisboa (Portugal)).

[10] Stahl, R. H. A. (2021). Leveraging time-series signals for multi-stage startup success prediction (Doctoral dissertation, Master's thesis, ETH Zurich & EQT Partners).