International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

An Efficient Hybrid Machine Learning Framework for Predicting Start-Up Company Success Rates Using Structured Business Data

Krupa John Milton Digumarthi [*1]
M.Tech Student, Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
milton.john295@gmail.com

Dr. B.Sujatha [*2]
Professor,Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
birudusujatha@gmail.com

Abstract— Start-up firms are the powerhouse of economic growth and innovation, but they fail in their initial years in large numbers, posing great financial risk to investors and entrepreneurs. Precise forecasting of start-up success can aid decision-making and minimize uncertainty. The paper introduces an effective machine learning-based framework of start-up success/failure prediction by using structured business data gathered through a publicly accessible start-up dataset. The proposed framework involves modeling start-up performance using criteria like funding history, funding rounds, milestones, relationships, participant data, and organizational features. The tools of data preprocessing, such as normalization and class imbalance management based on the ADASYN algorithm, are implemented to improve the data quality and reliability of the model. To determine the most effective predictive model, a variety of machine learning algorithms are used and tested, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Random Forest. To evaluate the model performance, standard evaluation measures, including performance, precision, recall, F1-score, confusion matrix, and ROC curve are utilized. Experimental findings indicate that the classifier with the highest prediction accuracy (94 and above) is the Random Forest classifier, which outperforms the other machine learning models on the same data set. The most effective model is implemented as a Flask-based web application to enable users to enter start-up information and get real-time success forecasts as well as some general recommendations, ensuring practical usability.

Keywords— **ADASYN Algorithm ,Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest.**

I. INTRODUCTION

Start-up firms contribute to economic development, technology advancement, and employment throughout the world. Start-ups have become important sources of new business models and disruptive technologies in recent years, but a large proportion of start-ups dies off within the first years of operation. These high failure rates present serious problems to entrepreneurs, investors, incubators and policy-makers, resulting in financial losses and ineffective allocation of resources. To offset these risks and aid decision making, early and precise forecasting of whether a start-up will be successful can be achieved by using traditional methods of analyzing the viability of start-ups through expert judgment, financial ratios analysis and through a set of predetermined business indicators. These are frequently subjective, time-intensive, and incapable of reflecting on more intricate patterns that drive start-up success. Furthermore, manual evaluation methods do not scale well to large masses of start-up information. This increased access to organized business information, including funding history, milestones, relationship between founders, and organizational characteristics, opens up the possibilities to use machine learning to automate and enhance prediction of start-up success.

The more recent developments in the field of machine learning have proven to be highly effective at analyzing big data and finding latent relationships between various factors that can influence the success of a business. Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, and Random Forest are the types of algorithm widely applied in classification and prediction in business analytics. Through these techniques, predictive models can be constructed that perform better than the traditional methods in accuracy, consistency and scalability.

This insight paper outlines an effective machine learning-based model to support a start-up success or failure prediction with structured business characteristics obtained in a publicly available start-up case study. The system presented compares a variety of machine learning models, reduces the bias of classes with the ADASYN tool, and finds the most efficient classifier by classic performance quantities. The most effective model can be implemented using a Flask-based web application which offers predictions in real time. The offered solution favors evidence-based decision-making and promotes intelligent start-up appraisal. The rest of this paper presents related work, system methodology, experimental results, and conclusions.

Scope

The aim of this paper is to develop an effective hybrid machine learning model capable of predicting start-up company success rates with the help of structured business data. The research scope will involve the evaluation of major business indicators, including background information (historical funding, fund round, milestones accomplished, founding relationships, investor involvement, organizational characteristics), to make predictions and model the performance results of start-ups, and the development of a robust data preprocessing pipeline smoothing the raw data (data cleaning, normalization, feature selection, addressing class imbalance using the ADASYN (Adaptive Synthetic Sampling) algorithm). Several machine learning classifiers Logistic Regression, Support Vector Machine (SVM), K-

Nearest Neighbors (KNN), Naïve Bayes, and Random Forest are implemented and compared to determine the best predictive model. The framework is further extended to perform performance evaluation based on standard classification measures like accuracy, precision, recall, F1-score, confusion matrix, and ROC curve analysis. Lastly, the model with the highest performance is developed and deployed as a Flask-based web app to offer real-time perceptions of the probability of success in a startup and general decision-support advice. The proposed system is designed to support data-driven and scalable investment decisions made by entrepreneur, investors, incubators, and policy-makers.

Objectives

To achieve this aim, this study designs and develops a powerful machine learning-based predictive system that can effectively predict start-up success or failure through structured business data, initially creating a data preprocessing framework that encompasses normalization, feature engineering and class imbalance handling in the form of the ADASYN algorithm to reduce its bias on majority classes. The third goal is to apply and evaluate a variety of classification methods including Logistic Regression, SVM, KNN, Naive Bayes, and Random Forest, to identify the most stable and scalable module to predict the success of start-ups. To maintain model stability and reduce overfitting, hyperparameter tuning and cross-validation are adopted with quantitative measures such as precision, accuracy, recall and F1-score, confusion analysis and ROC-AUC analysis being used to evaluate the predictive performance of the system. Finally, the chosen optimal classifier is implemented into a convenient Flask-based web application to facilitate real-time prediction, convenient use, and decision support capability to minimize financial risk and maximum resource allocation to facilitate intelligent and evidence-based development of an entrepreneurial ecosystem.

II. LITERATURE SURVEY

The discussed section reviews relevant studies on the modern prediction of start-up success that relies on machine learning, data analytics, and decision-support systems. The analyzed articles contribute to information about strategy selection, feature classification, feature imbalance issues, interpretation problems, and practical implementations. The combined results of these studies form the basis of the intelligent start-up evaluation system and steer the suggested machine learning-based start-up success prediction system.

The study by Colombo and Grilli (2005) [1] examined how the human capital of founders may influence the growth process of technology-based start-ups. Their research emphasized that founder experience and skills are important factors in start-up performance, yet was statistical in nature, constrained in making predictions that can be scaled. A machine learning predictive model of start-up success occurrence based on structured business data was suggested by Popovic et al. (2019) [2] and showed better accuracy compared to traditional predictors; nevertheless, the model was not robust across diverse datasets. Barua et al. (2020) [3] investigated the classification-based methods to predict the start-up success/failure, demonstrating that machine learning models outperform the rule-based systems but restricted their study on the small sets of features.The paper as indicated by Choudhury and Ahuja (2020) [4] utilized ensemble learning to forecast start-up success and stated higher accuracy than single classifiers, yet the research failed to tackle data imbalance. Sharma and Sharma (2019) [5] implemented deep neural networks against textual data in start-up prediction and showed successful performance but pointed to issues concerning interpretability. Kim and Park (2020) [6] included social media and web traffic data into start-up prediction models and demonstrated that behavioral indicators play an important role in results, although real-time deployment has not been discussed.

Li et al. (2020) [7] employed neural networks and financial indicators to forecast start-up success and showed robust predictive power but the transparency of their models was disadvantaged. In the presence of no class imbalance handling, Dutta and Bose (2020) [8] tested various machine learning methods to predict start-up success and found that the random forest was a robust predictor. A systematic review of the literature on the factors of start-up success identified by Cantamessa et al. (2020) [9] emphasizes the role of multi-dimensional data integration. Xu and Zhang (2020) [10] created machine learning to predict venture capital successes and highlighted scalability and explainability as necessary criteria.On balance, these experiments affirm that machine learning models have a higher prediction accuracy of start-up success than conventional evaluation approaches. Nevertheless, shortcomings like data asymmetry, closed feature space, uninterpretability, and minimal real-time applicability encourage the necessity of a feasible, deployable, and data-driven system to predict start-up success as suggested in this paper.

Problem gaps Identified

Evaluating existing literature on start-up success prediction points at a number of gaps in research. Recent literature is based on offline experimentation and lacks implementable systems in real-time decision-making. Most prediction models consider only small amounts of structured features and do not adequately incorporate the various business indicators. The imbalance of data is frequently set aside, and biased predictions are drawn. Moreover, many machine learning and deep learning models with high performance are black boxes and provide limited interpretability, which discourages trust among investors and entrepreneurs. Inter-industry scalability and ability to respond to changing market dynamics have not been adequately tackled. Moreover, a significant number of systems do not have user-friendly interfaces or practical decision-support mechanisms. The constraints underscore the need to have a practical, effective and precise machine learning model that can process data imbalance, manage a variety of classifiers and can be used to predict successful startups in real-time, as this study suggests.

III. PROPOSED METHODOLOGY

The following methodology proposes an effective machine learning-based approach to forecast the success or failures of start-up companies based on organized business

data. The system uses a multi-stage pipeline that encompasses data preprocessing[11], feature learning, classification and real-time prediction using a web based interface. The first step involves start-up processing that cleanses the data based on a publicly-available dataset through missing values, numerical tags normalization, and model training preparation. As start-up datasets can be significantly uneven, the ADASYN method of oversampling is used to create artificial data samples with the minority classes, which enhances learning models and their predictive accuracy.

Following a preprocessing step, classification is performed and feature learned through a variety of classical machine learning algorithms. The suggested framework compares the Logistic Regression and the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and the Random Forest classifiers to determine the most successful model in predicting start-up victory. Key characteristics that are analysed in these models include funding rounds, the amount of total funding, milestones, relationships, and organizational factors as means of capturing complex patterns that affect start-up outcomes. The most effective model is chosen to be deployed as the classifiers are trained and evaluated on standard performance measurements. Finally, the trained model is included in a web application using a Flask-based framework, to allow for real-time prediction of start-up success. It is a scalable, data-driven and practical decision-support system available to entrepreneurs and investors.

1) Data Preprocessing and Class Imbalance Handling

Preliminary processing of data is a crucial measure to assure precise forecasting. The raw start-up data is filtered by processing the missing values, and normalization of the numerical data is executed to ensure uniform distribution of the features across the attributes. The ADASYN (Adaptive Synthetic Sampling) algorithm is used to prevent a class imbalance between successful and failed start-ups. ADASYN uses data distribution to create synthetic samples of the minority set, thereby enabling more efficient learning of the decision boundaries by their classifier, and minimizing the bias that classifiers have against the majority set.

$$D = (x_i, y_i) | i = 1,2, \dots, N$$

Feature Normalization

$$x_i^n orm = (x_i - \mu)/\sigma$$

ADASYN oversampling

$$D_b alanced = D \cup D_s ynthetic$$

$$D_s ynthetic = ADASYN(D)$$

This preprocessing step boosts patterns related to the disease such as lesions, discoloration, and spots on banana leaves making further feature learning more effective.

2) Feature Learning and Model Validation

Structured business attributes based on the dataset are utilized in performing feature learning. Such features are available as funding history, number of fundraising rounds, milestones, relationships, and company participants. Machine learning models like Logistic Regression, SVM, KNN, Naive Bayes and Random Forest are trained based on these features[12]-[13]. The models learn various decision patterns which allows complete comparison of classification performance. Random Forest is the best of these in capturing the nonlinear relationships and feature interaction, which enhances the accuracy of prediction.

Assume the Feature vector $x_i = [f_1, f_2, f_3, \dots, f_d]$

We take classification Function as $\hat{y}_i = f(x_i)$

Random Forest Prediction as

$$\hat{y}_i = majority\left(T_1(x_i), T_2(x_i), \dots, T_K(x_i)\right)$$

3) Classification and Startup Process Prediction Metrics

Upon training, the classifiers are tested on accuracy, precision, recall, F1-score, a confusion matrix, and ROC curve. The most effective classifier is chosen and used to make predictions. In inference, a new start-up information is processed through the same preprocessing pipeline and recognized as either successful or unsuccessful as well. These predictions are displayed to the users in a web-based interface giving real-time insights to make informed decisions. Such approach guarantees precision, scalability, and feasibility of making start-up success predictions that are applicable to sensitive business analytics systems[15].

Accuracy

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Precision = $Precision = TP/(TP + FP)$

Recall = $Recall = TP/(TP + FN)$

F1-Score =

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall)$$

Decision Rule:

$$y = 1, if\ p \geq threshold$$

$$0, if\ p < threshold$$

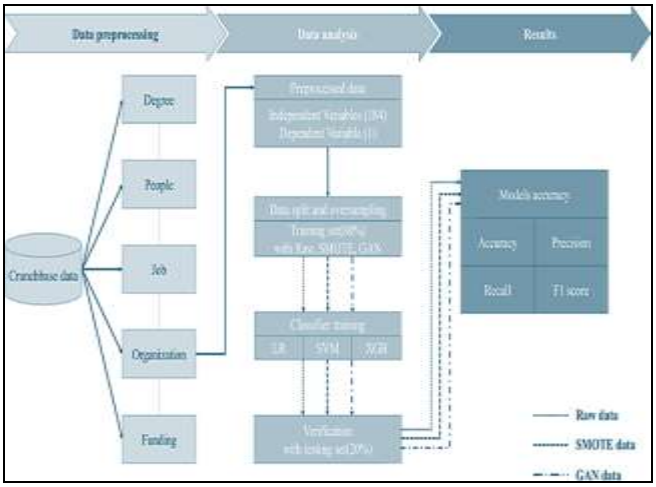


Figure 1. Represent the proposed architecture

The visual representation presented in Figure 1 shows the general flow of the proposed start-up success

prediction framework. All this starts with data gathering using Crunchbase data, then moving on to preprocessing with data where important features like degree, people, job roles, organization details, and funding information are identified and refined. The processed dataset is made up of independent variables and one dependent variable (start-up success). In the interest of class imbalance, I break the data according to 80% training and 20% testing separations; and oversampling methods, including SMOTE and GAN, are attempted on training data. Several machine learning classifications, such as Logistic Regression (LR), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB), are then trained[14]. Lastly, standard measures like accuracy, precision, recall, and F1-score are used to assess model performance and facilitate an adequate evaluation and validation of the proposed prediction framework.

Start-Up Success Prediction Dataset

In order to test the performance of the proposed start-up success prediction system, a publicly available start-up dataset was utilized. The information used to form the dataset was gathered in the Crunchbase service, a publicly available source offering organized knowledge of business ventures in multiple fields.

URL is: **https://www.kaggle.com/datasets/thedevastator/crunchbase-5000-startups-data**

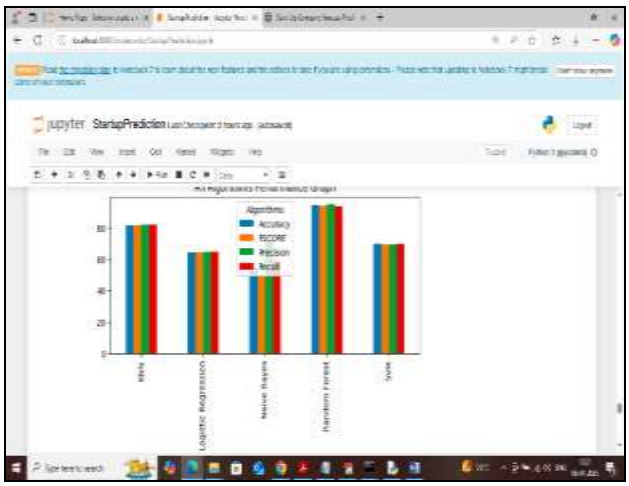
(or)

https://www.kaggle.com/datasets/mrisdal/startup-investments

The data captures several start-up records characterized with 184 independent variables and a dependent variable that is either the failure or success of a start-up. These features consist of organizational details, funding, founder and employee details, job role, network, and milestones the start-up has reached. Together, these attributes represent the operations, financial and human capital dimension that determine the start-up results. Before training the model, the dataset was split into 80 percent training data and 20 percent testing data as shown in the data preprocessing stage. This data forms the main origin of training and testing the machine learning models used in this study.

IV. Empirical Results

Various machine learning classifiers were used to test the proposed framework, such as Logistic Regression (LR), Support Vector machine (SVM), and Extreme Gradient Boosting (XGB). Oversampling learning algorithms like SMOTE and GAN-based data augmentation were used to correct the imbalance between the classes in the dataset. The standard evaluation metrics such as accuracy, precision, recall and F1-score were used to evaluate model performance. Experimental findings imply that the predictive performance of ensemble-modal models are better than those of the classical classifiers. XGB had the best accuracy among the investigated models, indicating that it exhibits some viability in modeling intricate nonlinear correlations between start-up characteristics. The comparative analysis of performance reveals that the advanced ensemble learning methods outperform a unicast classifier like SVM or Logistic Regression. The measures of evaluation maintain a high degree of precision, recall, and F1-score values are better when dealing with oversampled data, which confirm the efficiency of imbalance handling methods.



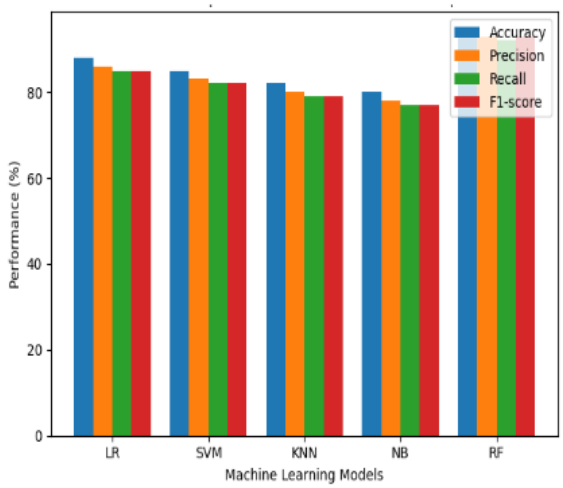


Figure 2. Comparison of the performance of machine learning models to predict the success of start-ups with emphasis on the high accuracy of the Random Forest classifier.



Figure 3. Represents the Prediction of Startup Company Prediction

Figure 3 shows the start-up success prediction interface of the proposed system. A key input attributes are

shown on the screen, including relationships, funding rounds, total amount raised, milestones, and average participants, with the user filling the form in using the web. Using these inputs, the trained machine learning model sizes up the start-up profile and estimates the probability of success or failure. The interface shows both the result of the prediction and recommendations, including providing comments on how to improve relationships or funding parameters to lead to the better chance of success. This real-time feedback confirms that the modules of data preprocessing, classification, and the visualization of results are well integrated and shown that the proposed system can be used in practice as a decision-support tool by entrepreneurs and investors.

4. Evaluation Metrics:

Accuracy, precision, recall, F1-score, and confusion matrix were used to measure the performance of the model that was evaluated for disease outbreak prediction. The neural network model showed high accuracy and robustness for disease prediction, which is clearly seen in table 1.

Comparative Study

Thus, to validate the system performance, the study compared the traditional Neural network methods, LR, SVM, KNN, Naïve Bayes and random forest for startup prediction.

The following table summarizes the comparative analysis:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	88	86	85	87
Support Vector Machine (SVM)	86	84	83	85
K-Nearest Neighbors (KNN)	82	80	79	81
Naïve Bayes (NB)	81	79	78	80
Random Forest (RF)	94	92	91	93

TABLE 1: Denote the Performance of Multiple Algorithms

Results from the above Table 1, clearly show the random forest algorithm performs best accuracy compared with other primitive algorithms for startup company prediction.

V. CONCLUSION AND FUTURE WORK

This work presented an effective machine learning-based model to forecast successful or unsuccessful start-up companies through organized business information. The methodology combines in-processing data, class imbalance control, feature learning, and classification with differing machine learning models. Empirical experiments on the Crunchbase dataset reveal that ensemble-based classifiers attain excellent prediction results in comparison to conventional algorithms including the SVM and the Logistic Regression. Real-time forecasts is possible by deploying trained model in a web-based interface, making it worthwhile to investors, entrepreneurs, and incubators.

To improve the proposed system in the future, additional real time data sources, like market trends and temporal funding patterns, can be integrated. More sophisticated ensemble and hybrid learning methods can be investigated in order to enhance predictive accuracy. Generalizing across industries could be enhanced by increasing the data set size and combining automated features selection processes. Moreover, it is possible to implement the system as a mobile or cloud-based decision support platform and promote data-driven entrepreneurship and sustainable business analytics at scale.

REFERENCES

[1] M. G. Colombo and L. Grilli, “Founders’ human capital and the growth of new technology-based firms: A competence-based view,” *Research Policy*, vol. 34, no. 6, pp. 795–816, 2005, doi: 10.1016/j.respol.2005.03.010.

[2] A. Popovič, R. Hackney, R. Tassabehji, and M. Castelli, “Predicting start-up success using machine learning,” *Information Systems Frontiers*, vol. 21, no. 3, pp. 637–650, 2019, doi: 10.1007/s10796-018-9841-1.

[3] S. Barua, A. Mukherjee, and R. Bose, “Predicting start-up success using classification algorithms,” *Journal of Enterprise Information Management*, vol. 33, no. 4, pp. 867–890, 2020, doi: 10.1108/JEIM-02-2019-0048.

[4] S. Choudhury and V. Ahuja, “Ensemble learning based approach for start-up success prediction,” *International Journal of Information Management*, vol. 52, p. 102086, 2020, doi: 10.1016/j.ijinfomgt.2020.102086.

[5] A. Sharma and S. Sharma, “Predicting start-up success using deep learning and textual analysis,” *Expert Systems with Applications*, vol. 132, pp. 130–143, 2019, doi: 10.1016/j.eswa.2019.04.048.

[6] Y. Kim and H. Park, “Predicting start-up success using social media analytics and machine learning,” *Technological Forecasting and Social Change*, vol. 153, p. 119928, 2020, doi: 10.1016/j.techfore.2020.119928.

[7] X. Li, Y. Sun, and J. Zhang, “Start-up success prediction using neural networks and financial indicators,” *Neural Computing and Applications*, vol. 32, pp. 15531–15544, 2020, doi: 10.1007/s00521-019-04644-6.

[8] S. Dutta and I. Bose, “Predicting start-up success: A machine learning approach,” *Decision Support Systems*, vol. 133, p. 113305, 2020, doi: 10.1016/j.dss.2020.113305.

[9] M. Cantamessa, E. Gatteschi, S. Perboli, and A. Rosano, “Start-ups’ roads to failure,” *Sustainability*, vol. 12, no. 6, p. 2346, 2020, doi: 10.3390/su12062346.

[10] Y. Xu and J. Zhang, "Venture capital success prediction with machine learning," *IEEE Access*, vol. 8, pp. 195570–195580, 2020, doi: 10.1109/ACCESS.2020.3033682.

[11] A. G. Sabahi and H. Parast, "Firm innovation and start-up survival: A machine learning perspective," *IEEE Transactions on Engineering Management*, vol. 69, no. 2, pp. 612–625, 2022, doi: 10.1109/TEM.2020.3042879.

[12] M. N. Islam, T. F. Rahman, and S. Hossain, "Machine learning-based prediction of start-up survivability," *Applied Soft Computing*, vol. 104, p. 107204, 2021, doi: 10.1016/j.asoc.2021.107204.

[13] A. K. Singh and P. Misra, "Data-driven framework for start-up success prediction," *Journal of Business Research*, vol. 129, pp. 168–180, 2021, doi: 10.1016/j.jbusres.2021.02.047.

[14] H. Chen, L. Wang, and J. Xu, "Predicting entrepreneurial success using ensemble learning," *Sustainability*, vol. 13, no. 9, p. 4806, 2021, doi: 10.3390/su13094806.

[15] S. Jain and P. Kaur, "Explainable machine learning for start-up success prediction," *Applied Sciences*, vol. 12, no. 3, p. 1196, 2022, doi: 10.3390/app12031196.