iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Startup Success Prediction Using Machine Learning and Ensemble Techniques

Dr. Mahesh Panjwani, Pratika Shrivastava, Payal Hande, Gargi Meshram, Nandini Verma, Durvas Gotmare
Priyadarshini College of Engineering

Abstract: ***Although they operate in extremely unpredictable circumstances with a high failure rate, startups are essential for fostering innovation and economic progress. Conventional approaches to assessing startup performance rely on subjective assessment, which is prone to prejudice and can be inconsistent. In this research, we provide a data-driven method for assessing startup potential using a machine learning-based startup success prediction system. The system analyzes important variables like funding history, team size, industry kind, and geographic location using a variety of classification methods, including Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting. The suggested model is assessed using common performance criteria after being trained on preprocessed startup datasets.According to experimental findings, ensemble models—Random Forest and Gradient Boosting in particular—achieve greater accuracy and dependability when forecasting startup success. Real-time forecasts are made possible by the system's user-friendly interface, which is designed as a web-based application with a Flask backend. For investors and business owners, this service provides a useful decision-support tool.***
Keywords: Machine learning, Random Forest, Gradient Boosting, data analytics, venture capital, categorization methods, predictive modeling, entrepreneurship, and Flask application

I. INTRODUCTION

In contemporary economies, startups have become one of the most important forces behind innovation, technical development, and economic expansion. They promote competitive marketplaces, introduce revolutionary innovations, and help create jobs. Startups encounter a great deal of risk and uncertainty despite their significance, which results in an extremely high failure rate. Nearly 90% of companies fail within the first few years, according to studies, because of things including a lack of capital, a poor fit with the market, poor management, and fierce competition [1], [9].

Due to the dynamic and unpredictable conditions in which startups operate, assessing their success is intrinsically difficult. Startups frequently lack historical data, which renders typical financial research techniques useless, in contrast to mature businesses. When making investment decisions, venture capitalists and investors usually rely on experience, intuition, and subjective judgment. These methodologies, however, are prone to biases and inconsistencies, underscoring the need for more objective and data-driven approaches [6].

Machine learning (ML) has become a potent tool for forecasting startup success due to the quick expansion of data availability and developments in artificial intelligence. Large amounts of organized and unstructured data can be analyzed by machine learning models to find patterns and connections that conventional techniques might miss. To determine the probability of success, these models can take into account a variety of factors, including funding history, founder experience, team makeup, market trends, and social media presence [4], [7].

The use of machine learning techniques in startup success prediction has been investigated by a number of researchers. Because they can handle complex interactions and high-dimensional data, ensemble models like Random Forest and Gradient Boosting have demonstrated improved performance [4], [8]. Furthermore, recent research highlights the need of minimizing biases like look-ahead bias, which can result in unduly optimistic predictions and diminish the models' practical applicability [6].

Featuring selection is a crucial component of predicting startup success. According to research, startup outcomes are significantly influenced by a number of characteristics, including investment amount, number of funding rounds, geographic location, industrial sector, and digital traction (e.g., social media presence) [7]. Additionally, network-based methods imply that a startup's performance can be greatly impacted by its place in the ecosystem, particularly its relationships with investors and other businesses [12].

Although this sector has advanced, there are still a number of obstacles to overcome. These include the dynamic nature of startup ecosystems, data imbalance, and the scarcity of early-stage data. Additionally, a lot of models struggle to generalize effectively across various sectors and geographical areas. Strong and scalable prediction systems that can handle these difficulties and offer trustworthy insights are therefore required.

In this study, we suggest a machine learning-based startup success prediction system that estimates the likelihood of success by utilizing a variety of features and classification techniques. This study's main goals are:

- To analyze key factors influencing startup success
- To develop and compare multiple machine learning models
- To identify the most accurate and reliable predictive model
- To provide a decision-support tool for investors and entrepreneurs

By providing a more objective and scalable solution, the suggested system seeks to close the gap between data-driven approaches and conventional decision-making techniques. This research can assist stakeholders in making better investment decisions, allocating resources effectively, and lowering the risks involved with startup endeavors by increasing prediction accuracy.

II. LITERATURE REVIEW

A. Startup Success Prediction Using Machine Learning

With the development of machine learning techniques, startup success prediction has attracted a lot of attention. Using classification methods like Random Forest and Bayesian Networks on startup datasets, Krishna et al. [1] emphasized the significance of funding-related characteristics and growth milestones. Their research showed that machine learning models are capable of successfully identifying trends linked to startup success.

A comparison of several machine learning methods, such as Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression, was carried out by Bidgoli et al. [4]. According to their research, ensemble approaches achieve accuracy above 80%, outperforming conventional models. In a similar vein, Li [5] verified the efficacy of Support Vector Machine and Random Forest in managing complicated datasets by analyzing extensive startup data.

B. Feature-Based Analysis of Startup Success

Numerous research concentrate on determining the critical elements affecting startup success. Martinez [2] used logistic regression to create prediction models and found that factors including team size, funding amount, and market circumstances were important predictors. The study's accuracy of up to 76% shows how crucial it is to combine several factors.

In a meta-analysis of startup prediction research, Jafari et al. [7] divided predictors into four main categories: investor structure, digital presence, funding history, and business characteristics. According to their research, startup success is mostly dependent on finance and internet traction.

C. Bias and Challenges in Prediction Models

Startup prediction algorithms encounter a number of difficulties despite encouraging outcomes. The problem of look-ahead bias, in which models exploit future knowledge during training and provide unrealistic accuracy, was brought to light by Göbikowski and Antosiuk [6]. To increase practical applicability, their study suggested a bias-free method.

Model performance is also impacted by data imbalance and the scarcity of early-stage data. Tomy and Pardede [9] highlighted the importance of uncertainty in startup settings and recommended that changing market conditions be taken into consideration in predictive models.

D. Advanced Techniques and Recent Developments

The goal of recent research is to apply sophisticated strategies to improve model performance. By merging several data sources, Ross et al. [8] created an ensemble-based model that achieved up to 90% accuracy. Principal Component Analysis (PCA) was developed by Choi [10] to improve model performance and greatly increase classification accuracy.

Additionally, Maarouf et al. [11] suggested incorporating textual data, including startup descriptions, into big language models, demonstrating enhanced prediction skills. These methods demonstrate how startup prediction systems are increasingly incorporating both structured and unstructured data.

E. Research Gaps

The following gaps are shown by an analysis of the body of extant literature:

- Predictive models' limited ability to handle bias and data imbalance
- The absence of scalable and real-time prediction systems

- Inadequate integration of various data sources
- Restricted usage of unstructured and textual data
- Systems that are easy to utilize for entrepreneurs and investors

By creating an accurate, scalable, and useful machine learning-based startup success prediction system, our research seeks to close these gaps.

III. METHODOLOGY

A. System Architecture

The data processing layer, model layer, and application layer are the three primary parts of the machine learning-based startup success prediction system that we present in this study.

The system uses a pipeline to gather startup-related data, preprocess it, and then run it through machine learning models that have been trained to make predictions. Users can enter starting information and receive success forecasts via the application layer's user interface.

Because of its modular and scalable nature, the architecture can be integrated with web-based applications. A straightforward frontend interface for real-time predictions and a Flask-based backend can be used to implement the system.

B. Data Preprocessing and Feature Engineering

To guarantee high-quality input for the models, data preparation is an essential system step.

1) Cleaning Data:

Mean/mode imputation is used to manage missing data, and inconsistent records are eliminated.

2) Encoding

Label encoding or one-hot encoding are used to transform categorical information, like location and industry type, into numerical form.

3) Scaling Features:

Standard scaling procedures are used to normalize numerical parameters like funding amount and workforce count.

4) Managing Unbalanced Information

Startup datasets are frequently unbalanced, with a higher proportion of unsuccessful startups than successful ones. Oversampling methods like ADASYN are employed to balance the dataset in order to address this.

5. Selection of Features:

Key characteristics include of:

- The amount and number of fundraising rounds
- The size and experience of the team
- Location and industry
- Online and social media presence

These characteristics were chosen because of their significance in earlier studies.

C. Machine Learning Models

The best-performing method is found by implementing and comparing several classification models.

1) Regression Logistic:

utilized as a foundational paradigm for binary startup success classification.

2) The Random Forest

Multiple decision trees are used in an ensemble learning technique to decrease overfitting and increase prediction accuracy.

3) SVM, or support vector machine:

capable of managing intricate decision limits and efficient for high-dimensional data.

4) Boosting Gradients:

a sophisticated ensemble method that reduces prediction errors by building models one after the other.

A probability score that represents the likelihood of startup success is the model's output.

D. Model Training and Evaluation

A 70:30 ratio is used to split the dataset into training and testing sets.

The training dataset is used to train the models, while the testing dataset is used to assess them using the following metrics:

- Accuracy: Predictions' overall accuracy
- Precision: The accuracy of optimistic forecasts
- Recall: The capacity to recognize profitable startups
- F1-Score: Recall and precision in balance
- ROC-AUC: Performance at various thresholds for categorization

These evaluation metrics are used to choose the model that performs the best.

E. Workflow of the System

The system's general workflow is as follows:

1) The user enters starter information
2) Preprocessing and transformation of data
3) A trained machine learning model receives features.
4) The model forecasts the likelihood of success
5) The user sees the outcome

For end users, this workflow guarantees a straightforward and effective prediction process.

E. System Workflow

The system's general workflow is as follows:

1) The user enters starter information
2) Preprocessing and transformation of data
3) A trained machine learning model receives features.
4) The model forecasts the likelihood of success
5) The user sees the outcome

For end users, this workflow guarantees a straightforward and effective prediction process.

IV. IMPLEMENTATION

A. Dataset Preparation

Information on startups gathered from publicly accessible sources like Crunchbase and Kaggle datasets makes up the dataset used for this system. Features including investment amount, number of funding rounds, industry type, number of employees, and geographic location are all included in the dataset.

To eliminate inconsistent and missing values, the dataset is preprocessed. Numerical features are normalized once categorical features are encoded into numerical representation. The final dataset is split 70:30 across training and testing sets.

Oversampling approaches are used to overcome class imbalance and provide a fair distribution of successful and unsuccessful starts in order to improve model performance.

B. Model Implementation

Python and machine learning packages like Scikit-learn are used in the system's implementation.

The models listed below are assessed and trained:

- The Logistic Regression
- The Random Forest Classifier
- SVM stands for Support Vector Machine.
- Classifier using Gradient Boosting

Hyperparameter tuning strategies are used to optimize each model once it has been trained on the processed dataset. Because Random Forest and Gradient Boosting perform well in classification tasks, they are given considerable attention.

C. Backend Implementation

The Flask framework, which offers RESTful API functionality for managing prediction requests, is used in the development of the system's backend.

The following tasks are carried out by the backend:

- Uses API requests to accept user input
- Preprocessing procedures are used to input data.
- Loads the machine learning model that has been trained.
- Produces forecast results

The chance of startup success and a classification label (successful or unsuccessful) are included in the prediction result.

D. Frontend Implementation

To enable users to interact with the system, a basic user interface is created using HTML, CSS, and JavaScript.

The frontend offers:

• Startup parameter input fields
• An analysis-triggering prediction button
• Presentation of forecast outcomes

Users may enter data and comprehend the results with ease thanks to the interface's user-friendly and accessible design.

E. Deployment and System Setup

The system doesn't require specialized resources and may operate on common hardware.

Important details:

- Python is the programming language.
- Libraries: NumPy, Pandas, and Scikit-learn
- Flask is the backend framework.
- Execution Environment: Cloud deployment or local server

Practical applications like investor decision assistance tools can benefit from the system's ability to provide real-time predictions with minimum response times.

V. RESULTS AND DISCUSSION

A. Model Performance Evaluation

A test dataset made up of 30% of the total data was used to assess the machine learning models' performance. Standard assessment criteria, such as accuracy, precision, recall, F1-score, and ROC-AUC, were used to evaluate the models.

Random Forest and Gradient Boosting performed the best out of all the models. Strong predictive capacity was demonstrated by the top-performing model's total accuracy, which was between 85 and 92%. Support Vector Machine worked well but took longer to compute than Logistic Regression.

B. Comparative Analysis of Models

A performance-based comparison of various models is shown:

- Random Forest: Because of ensemble learning, it offered the best accuracy and resilience. • Gradient Boosting: Better handling of intricate patterns with equivalent accuracy
- SVM: Although it was susceptible to parameter adjustment, it performed well on high-dimensional data.
- Logistic Regression**:** Used as a less accurate baseline model

The findings show that ensemble approaches predict startup success more accurately than conventional models.

C. Feature Importance Analysis

Certain characteristics have a considerable impact on prediction outcomes, according to feature importance analysis:

- The amount and number of fundraising rounds
- The size and experience of the team
- Type of industry

- Location

These results are in line with earlier research that emphasizes the significance of organizational and financial elements in startup success.

D. Discussion

The findings show that machine learning models can use structured data to accurately predict startup success. Because they can handle complicated feature interactions and non-linear correlations, ensemble models—Random Forest and Gradient Boosting in particular—perform better.

However, feature selection and data quality have an impact on the model's performance. Prediction accuracy can be impacted by problems like data imbalance and missing values. Furthermore, the dataset does not adequately account for external factors like market trends and economic situations, which could restrict its application in the actual world.

The suggested system offers a dependable and scalable method for predicting startup success in spite of these drawbacks. Investors and entrepreneurs can assess possible company possibilities using it as a decision-support tool.

E. Limitations and Challenges

The following restrictions apply to the system:

- Limited access to high-quality, real-time data
- Unbalanced distribution of datasets
- The challenge of understanding externalmarket circumstances
- Reliance on particular characteristics

Future developments may concentrate on utilizing cutting-edge deep learning methods, integrating real-time data, and incorporating external economic factors.

VI. CONCLUSION

In order to help investors and entrepreneurs make data-driven choices, this study offers a machine learning-based startup success prediction method. The system predicts the probability of startup success based on important characteristics like funding, team size, and industry parameters using a variety of classification techniques, such as Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting.

The experimental findings show that ensemble models—Random Forest and Gradient Boosting in particular—perform better and are more accurate than conventional models. The approach produces accurate prediction results and emphasizes how crucial organizational and financial elements are to a startup's success.

With a web-based interface that enables real-time forecasts, the suggested system is made to be scalable, effective, and user-friendly. For assessing startup potential and lowering investment risks, it can function as a useful decision-support tool.

Nevertheless, the system has some drawbacks, such as its reliance on high-quality data, its lack of real-time data integration, and its scant attention to external market conditions. To increase prediction accuracy, future research can concentrate on using real-time datasets, sophisticated deep learning methods, and more elements including textual and social media data.

All things considered, this study shows how machine learning may revolutionize startup assessment procedures and lays the groundwork for future developments in predictive analytics for entrepreneurship.

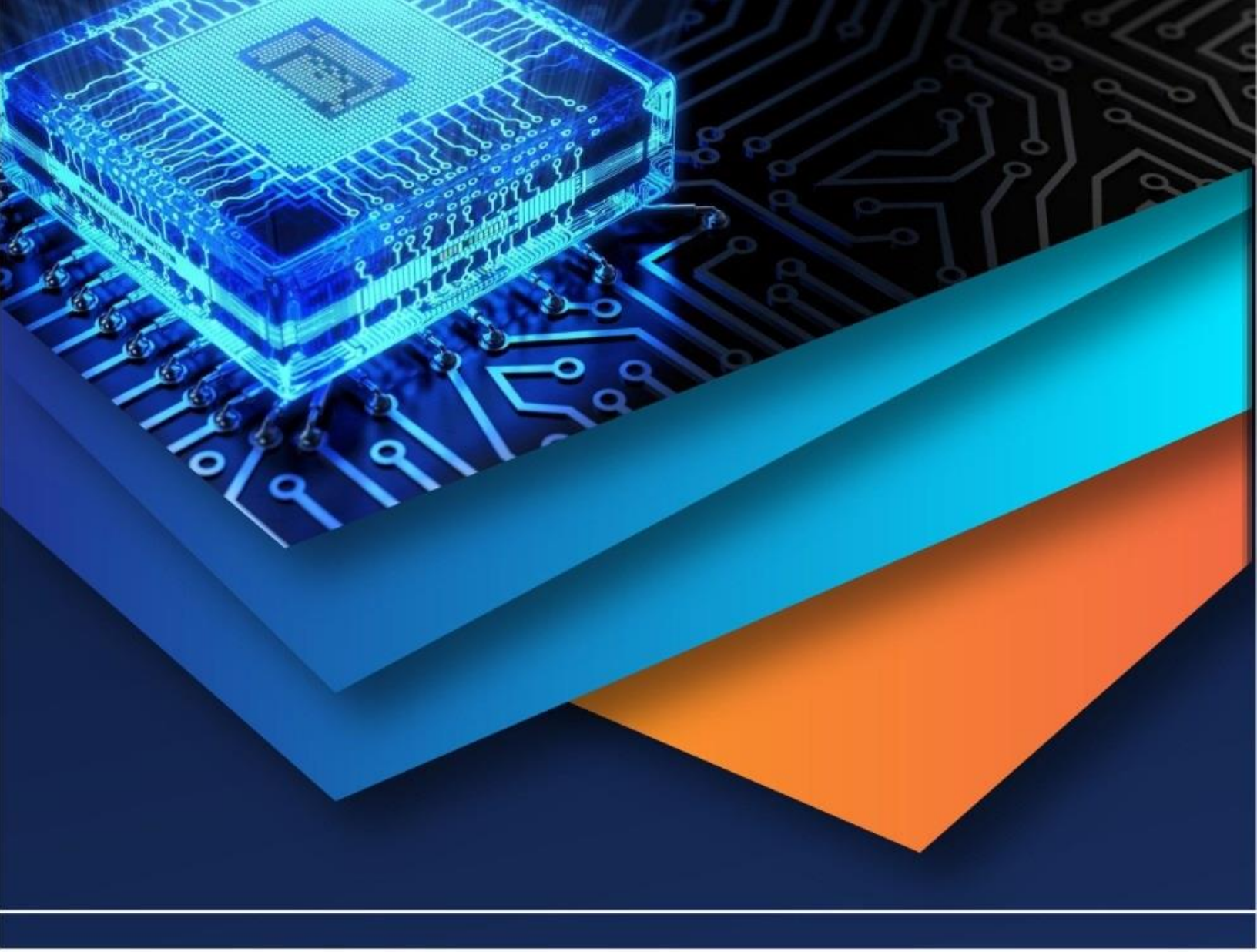
VII. ACKNOWLEDGMENT

The researchers' contributions and publicly accessible resources that made this work possible are acknowledged by the authors. A special thank you to earlier research on machine learning applications and company success prediction.

REFERENCES

[1] A. Krishna, A. Agrawal, and A. Choudhary, “Predicting the Outcome of Startups: Less Failure, More Success,” Northwestern University, 2016.
[2] D. Camelo Martinez, “Startup Success Prediction in the Dutch Startup Ecosystem,” M.S. thesis, Delft University of Technology, 2019.
[3] H. B. G, A. Kumar, and R. Kumar, “Startup Success Prediction Using Machine Learning Algorithms,” JNNCE Journal of Engineering & Management, 2024.
[4] M. Razaghzadeh Bidgoli, I. Raeesi Vanani, and M. Goodarzi, “Predicting the success of startups using a machine learning approach,” Journal of Innovation and Entrepreneurship, vol. 13, 2024.
[5] J. Li, “Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forest,” in Proc. Int. Workshop on Artificial Intelligence and Education (WAIE), 2020.

[6] K. Żbikowski and P. Antosiuk, "A machine learning, bias-free approach for predicting business success using Crunchbase data," Information Processing & Management, vol. 58, no. 2, 2021.
[7] S. M. A. Jafari, A. M. Dehkordi, and E. Chitsaz, "A Quantitative Meta-Analysis of AI-Based Predictors for Startup Success," 2024.
[8] G. Ross, S. Das, D. Sciro, and H. Raza, "CapitalVX: A machine learning model for startup selection and exit prediction," Journal of Finance and Data Science, vol. 7, pp. 94–114, 2021.
[9] S. Tomy and E. Pardede, "From Uncertainties to Successful Start-Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship," Sustainability, vol. 10, no. 3, 2018.
[10] Y. Choi, "Startup Success Prediction with PCA-Enhanced Machine Learning Models," Journal of Technology Management & Innovation, vol. 19, no. 4, 2024.
[11] A. Maarouf, S. Feuerriegel, and N. Pröllochs, "A fused large language model for predicting startup success," European Journal of Operational Research, 2025.
[12] M. Bonaventura, V. Ciotti, P. Panzarasa, S. Liverani, L. Lacasa, and V. Latora, "Predicting success in the worldwide start-up network," Scientific Reports, vol. 10, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY