Machine Learning Models for Startup Investment Prediction

G. Vijaya Lakshmi [1], D. Narendra Varma[2]
[1,2] *Department of Computer Science and Engineering, Sanketika Vidya Parishad Engineering College, Visakhapatnam, AP, India.*

***Abstract*—The startup investments gained much importance as most of the people are investing in the stock market. In the present study, predicting of startup investments is considered for identifying the selection of companies using classification and regression tasks. The selected classification models for the startup status prediction are Linear Model, Support Vector Machine (SVM), and Random Forest. Similarly for Regression tasks, Linear Regression, Decision Trees and Random Forest are employed for forecasting investments. The optimal models are identified from the forementioned popular models. In the prediction task, the Random Forest model outperforms the Linear Model and SVM. However, in regression tasks, Decision Tree Model outperforms Linear Regression and Random Forest models. A correlation matrix is applied for data segmentation and feature selection to identify the influential attributes. A detailed analysis is performed to access the correlations and the overall impact on models by eliminating the redundant sub-attributes. The methodology aims to predict the startup investment and guides the investors in identifying the valuable startups.**

***Index Terms*—Startup investments, Machine Learning, SVM, Random Forests, Decision Trees.**

1. INTRODUCTION

The rapid emergence of startups as key contributors to innovation and economic growth has been accompanied by a high level of uncertainty and risk. While new ventures can revolutionize industries through disruptive products and services, they also face intense market competition, limited resources, and operational challenges. A significant proportion of startups fail within their initial years of operation, highlighting the necessity for reliable tools to forecast their potential for success or failure. Accurate prediction not only benefits investors in optimizing funding decisions but also enables entrepreneurs to refine their strategies and improve sustainability.

In recent years, machine learning techniques have emerged as powerful tools for startup success prediction. By analysing diverse datasets including financial performance, founder characteristics, digital engagement, and market indicators, these algorithms can uncover patterns that are often invisible to conventional analysis. Classification models such as Support Vector Machines, Random Forests, and Logistic Regression, along with regression models like Decision Trees and Linear Regression, have demonstrated strong capabilities in modelling complex and non-linear relationships between input features and business outcomes. The integration of feature selection methods such as correlation matrix analysis further enhances prediction accuracy by eliminating redundant or noisy attributes.

The present paper is organized as follows. Section 2 presents the Literature Study, reviewing prior works in startup success prediction and identifying research gaps. Section 3 describes the Methodology, detailing dataset preprocessing, feature selection, and model implementation for both classification and regression tasks. Section 4 discusses the Results and Analysis, including model performance comparisons and feature importance insights. Section 5 outlines the Conclusions and Future Scope, summarizing the key findings and suggesting directions for extending this research toward more robust and industry-adoptable decision support systems.

2. LITERATURE STUDY

Startup success prediction has emerged as a significant research area due to the inherently high risks and uncertainties in early-stage investments. Although startups drive innovation and economic growth, over 75% fail within the first five years (Bednár &

Tarišková, 2017). This high attrition rate has motivated the adoption of machine learning (ML) techniques to assist investors in making data-driven decisions that minimize risk and identify promising ventures.

Early studies approached startup success prediction through statistical methods, such as logistic regression and heuristic models, which achieved moderate accuracy levels. In contrast, machine learning approaches have demonstrated superior performance. Vasquez et al. (2023) developed a systematic ML-based method for predicting information technology startup (SIT) success, integrating factor selection, preprocessing, and hybrid models. They reported up to 89% accuracy using Extreme Gradient Boosting (XGBoost) and k-nearest neighbors (KNN), emphasizing the importance of relevant feature selection and hybrid ensemble strategies.

Several works have leveraged large datasets from platforms like Crunchbase to evaluate the predictive power of various models. Razaghzadeh Bidgoli et al. (2024) applied Random Forest, Gradient Boosting, and Support Vector Machines to predict startup outcomes, with Random Forest achieving 82% accuracy. Key predictors included founder experience, funding amount, and digital engagement metrics. Park et al. (2024) addressed predictor bias by limiting features to pre-success data and using GAN-based oversampling to balance datasets, improving generalizability.

Piskunova et al. (2021) examined the Ukrainian startup ecosystem, comparing Logistic Regression, Decision Trees, and Random Forests. Decision Trees excelled in interpretability, while Random Forests achieved higher AUC values, suggesting a trade-off between transparency and predictive power. Similar findings by Arroyo et al. (2019) highlighted the significance of social media presence, founder diversity, and industry convergence.

Gangwani and Zhu (2023) present a comprehensive survey of business success modelling and prediction methods, covering over a decade of advancements in data-driven decision-making. The authors classify existing approaches into statistical, machine learning, and hybrid modelling categories, noting that the choice of method is often dependent on data availability, domain specificity, and interpretability requirements. Their review highlights the increasing adoption of ensemble learning techniques such as Random Forest, Gradient Boosting, and stacking models, which consistently outperform single learners in predictive accuracy. They also emphasize the integration of multi-modal data, including financial records, operational metrics, social media activity, and managerial attributes, into unified predictive frameworks. Importantly, the survey identifies a research trend towards explainable AI (XAI) in business prediction, reflecting the need for interpretability in investor and managerial decision-making. The authors conclude that while predictive accuracy has improved significantly, challenges remain in handling imbalanced datasets, mitigating bias in feature selection, and developing models that generalize across industries and geographies. These insights are directly relevant to startup investment prediction, where heterogeneous data sources and the trade-off between performance and interpretability are critical considerations.

The integration of qualitative factors into ML models has been another advancement. Bai and Zhao (2021) demonstrated that non-financial metrics such as “planning strategy” and “team management” strongly influenced venture capital (VC) investment decisions. Using a balanced scorecard and multiple classifiers, they achieved 78% accuracy with attribute-selected Logistic Regression and SVM, underscoring the weight of qualitative criteria over purely quantitative metrics.

Further, early-stage investment frameworks have been explored by Shi et al. (2023), who compared standard ML models (Random Forest, Gradient Boosting, Logistic Regression) for classifying early-stage startups as investable or non-investable. Their work confirmed that Random Forest consistently outperformed others in both accuracy and feature stability, with top predictors being funding history, market size, and founder background.

Despite these advancements, several gaps remain in the literature:

1. Most research focuses on classification (success/failure) rather than regression for investment amount forecasting.

2. Limited studies combine both tasks into a single decision-support pipeline.
3. Although Random Forest and Gradient Boosting dominate in performance, Decision Trees offer interpretability, which is critical for investor trust.
4. The role of feature selection, using correlation analysis, SHAP values, and permutation importance, remains underexplored in regression contexts.

In a nutshell, current research establishes ML, particularly ensemble methods like Random Forest and Gradient Boosting, as effective tools for predicting startup success. However, combining classification and regression approaches, integrating qualitative and quantitative data, and addressing feature bias represent promising directions, aligning with the present study's goal of building optimal models for both startup status classification and investment forecasting.

3. METHODOLOGY

The methodology for the present study follows a structured workflow to ensure accurate prediction of startup status and investment amounts. The process begins with importing essential Python libraries and loading the dataset obtained from Crunchbase and Mattermark. Initial data pre-processing is carried out to remove empty rows, eliminate unwanted spaces in column names, and convert relevant financial data (such as funding_total_usd) from string to numeric format. Missing values are handled through mean or mode imputation, depending on the feature type. Duplicate entries are removed, and categorical variables (e.g., status, country code, market) are transformed into numerical form using encoding techniques. Numerical features are scaled using the Min–Max Scaler to normalize data within a 0–1 range.

Feature selection and reduction is then performed to identify the most relevant attributes from the original 39 features. By eliminating redundant sub-attributes (e.g., date fields linked to founded year, or subcategories of total funding), the dataset is reduced to eight key features: status, country_code, market, seed, venture, funding_rounds, founded_year, and funding_total_usd. These features form the input variables for the machine learning models (Table 1).

Table.1: Data set features description

Feature	Description	Type
Status	Position of the startup (closed, operating, acquired).	Nominal
Country_code	Country which the startup is based in.	Nominal
Market	Market or the category a startup belongs to.	Nominal
Seed	Initial funding a business receives.	Numeric
Venture	Funding received from the Venture Capitalist.	Numeric
Funding_rounds	Number of rounds in which a startup gets funding.	Numeric
Founded_year	The year in which startup was started.	Numeric
Funding_total_usd	Total funding a startup receives.	Numeric

For classification, three supervised learning algorithms are trained, Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier, to predict startup status (closed, operating, acquired). The dataset is split into training and testing subsets, and model performance is evaluated using accuracy, MCC score, F1-score, classification reports, and confusion matrices. For regression, Linear Regression, Decision Tree Regression, and Random Forest Regression are implemented to forecast the total funding amount. The models are compared based on their R^2 scores to determine the most effective prediction approach.

The present layered methodology ensures separation between data handling, feature engineering, and model evaluation, enabling easier debugging and improvement. The architecture diagrams (Fig.1 for classification and Fig.2 for regression) illustrate the overall process from raw data input through to final predictions.

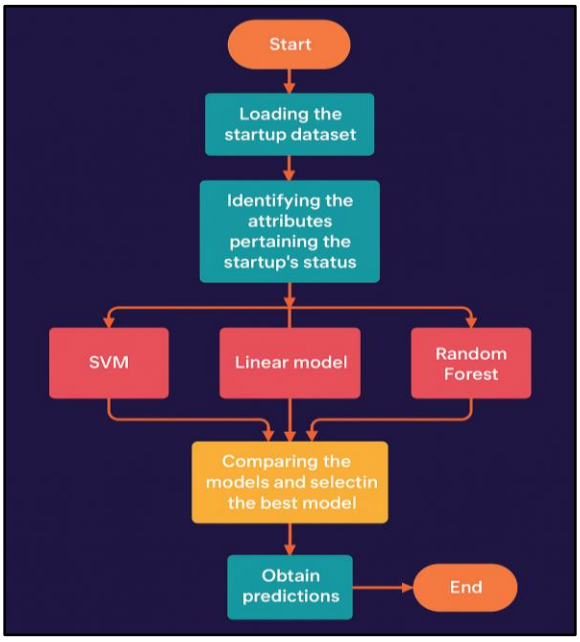

Fig.1: Classification problem architecture

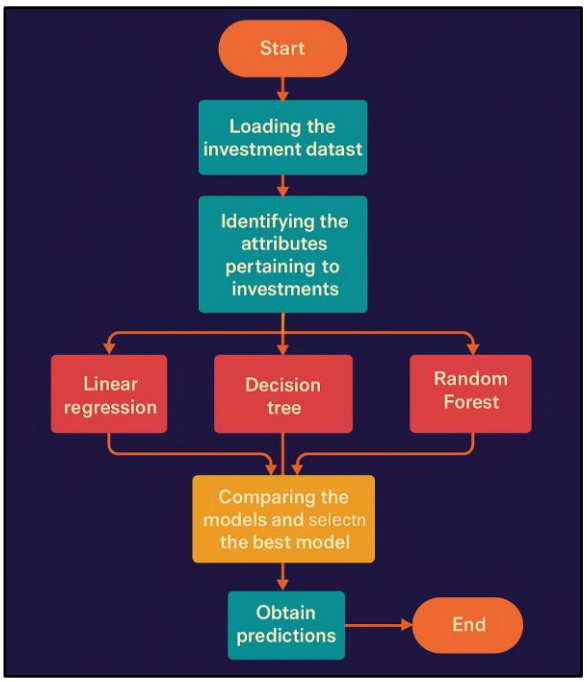

Fig.2: Regression problem architecture

4. RESULTS AND DISCUSSION

4.1 Testing and Evaluation Approach

The developed models were trained on 80% of the dataset and evaluated on the remaining 20% reserved for testing. This split ensured that model performance was measured on unseen data, thereby providing a reliable estimate of generalization capability. For the classification task, performance metrics (Table 2) included accuracy, F1-score, and Matthews Correlation Coefficient (MCC), with visual inspection through confusion matrices (Figures 3 - 5). These matrices display the distribution of correct and incorrect predictions for each startup status category (operating, acquired, and closed). For the regression task, model evaluation relied on the coefficient of determination (R^2) (Table 3) to assess how closely the models' predictions aligned with true funding values.

Table 2: Classification Results

MODEL	ACCURACY	MCC SCORE	F1 SCORE
Linear Model	0.867	-0.008	0.807
SVM	0.869	0.013	0.808
Random Forest Classifier	**0.982**	**0.924**	**0.982**

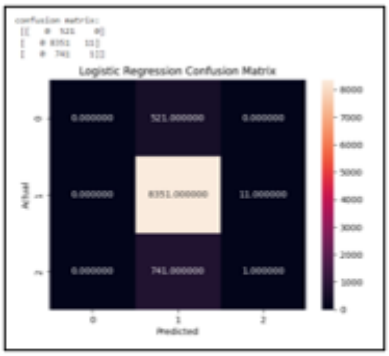

Fig. 3: Logistic Regression Confusion Matrix

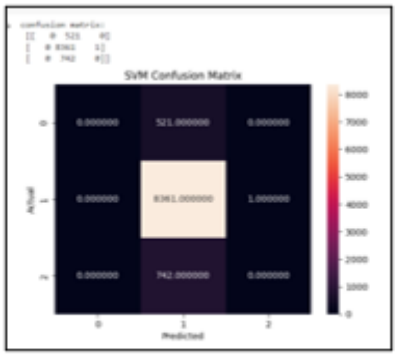

Fig. 4: SVM Confusion Matrix

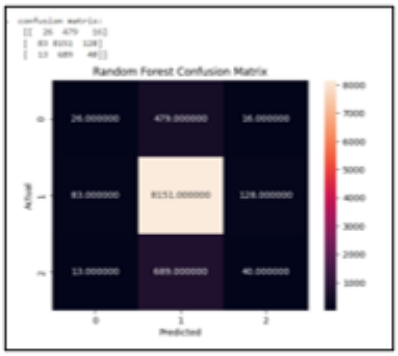

Fig. 5: Random Forest Classifier Confusion Matrix

Table 3: Regression Results

Model	R2 SCORE
Linear Model	0.411
Decision Tree	**0.969**
Random Forest	0.836

4.2 Classification Results: Prediction of Startup Status

Three supervised learning models i.e., Support Vector Machine (SVM), Logistic Regression (Linear Model), and Random Forest Classifier were implemented to predict startup status. The confusion matrix for the Random Forest Classifier revealed a strong concentration of predictions along the diagonal, indicating high accuracy across all classes. This model achieved the best results among the three, with the highest accuracy, F1-score, and MCC. SVM delivered moderate performance but required extensive parameter tuning, while the Linear Model performed the fastest but exhibited lower accuracy, suggesting that the underlying relationships are not strictly linear.

4.3 Regression Results: Forecasting Investment Amounts

For investment amount prediction, Linear Regression, Decision Tree Regression, and Random Forest Regression were compared. Decision Tree Regression produced the highest R^2 value, as evident from the clustering of points near the ideal prediction line and well in predictions (Figure 6). Random Forest Regression performed competitively but tended to average predictions, reducing its sensitivity to extreme funding amounts. Linear Regression showed weaker performance, confirming that a purely linear approach cannot fully capture the complex patterns in investment data.

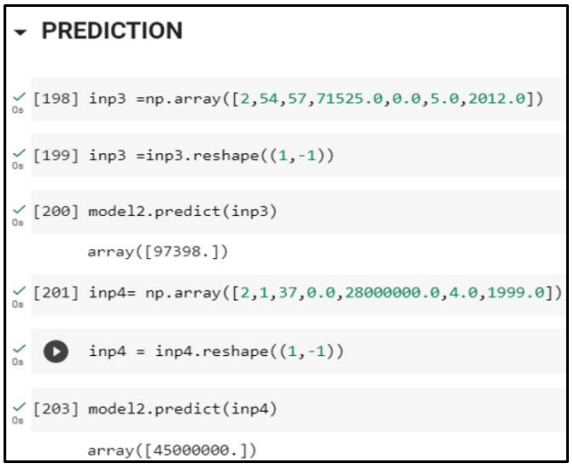

Fig. 6: Sample Prediction of Investment

4.4 Feature Influence and Interpretation

Feature importance analysis indicated that funding_rounds, market, and venture investments had the greatest impact on startup status classification, while funding_total_usd, seed investments, and founded_year were most influential in predicting funding amounts. The feature selection process, which removed redundant sub-attributes, contributed to improved model efficiency without reducing predictive accuracy.

The results highlight the advantage of tree-based and ensemble models for both classification and regression tasks. Random Forest Classifier offered the most accurate classification results due to its ability to handle complex, non-linear patterns while mitigating overfitting. Decision Tree Regression emerged as the best regression model, offering high accuracy with transparent, interpretable decision rules. The findings suggest that while ensemble methods deliver strong performance, single-tree models can be advantageous when interpretability is prioritized by decision-makers.

5. CONCLUSIONS AND FUTURE SCOPE

The present study implemented and evaluated multiple machine learning models for predicting startup status and forecasting investment amounts. For classification, Random Forest Classifier achieved the highest accuracy, F1-score, and MCC, significantly outperforming the Linear Model and SVM. For regression, Decision Tree Regression produced the best results with the highest R^2 value and lowest error, demonstrating its ability to model complex patterns in investment data. The feature selection process, which removed redundant sub-attributes, improved efficiency without compromising predictive accuracy. These findings confirm that tree-based and ensemble models are highly effective in addressing both classification and regression tasks for startup investment prediction.

The developed framework provides a valuable decision-support tool for investors, entrepreneurs, and policymakers. By identifying promising startups early and estimating potential funding needs, it can enhance investment strategies, reduce financial risks, and promote more sustainable business growth. The visual outputs, such as confusion matrices and predicted-versus-actual plots, further aid interpretability and decision-making.

Future work can focus on expanding the dataset with more diverse and real-time sources, including social media sentiment, founder network analysis, and market trend indicators. Advanced ensemble learning and deep learning techniques, combined with explainable AI methods, could further improve both predictive accuracy and transparency. Additionally, deploying the model as a web-based or cloud application would allow broader accessibility for real-world investment decision-making.

REFERENCES

[1] Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. IEEE Access, 7, 124233–124243. https://doi.org/10.1109/access.2019.2938659

[2] Bai, S., & Zhao, Y. (2021). Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches. Systems, 9(3), 55. https://doi.org/10.3390/systems9030055

[3] Bednár, R., & Tarišková, N. (2017). Indicators of startup failure. *Industry 4.0, 2*(5), 238–240.

[4] Gangwani, D., & Zhu, X. (2024). Modeling and prediction of business success: a survey. Artificial Intelligence Review, 57(2). https://doi.org/10.1007/s10462-023-10664-4

[5] Park, J., Choi, S., & Feng, Y. (2024). Predicting startup success using two bias-free machine learning: resolving data imbalance using generative adversarial networks. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00993-8

[6] Piskunova, O., Ligonenko, L., Klochko, R., Frolova, T., & Bilyk, T. (2022). Applying Machine Learning Approach to Start-up Success Prediction. Scientific Horizons, 24(11), 72–84. https://doi.org/10.48077/scihor.24(11).2021.72-84

[7] Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. Journal of Innovation and Entrepreneurship, 13(1). https://doi.org/10.1186/s13731-024-00436-x

[8] Shi, Y., Eremina, E., & Long, W. (2023). Machine learning models for early-stage investment decision making in startups. Managerial and Decision Economics, 45(3), 1259–1279. https://doi.org/10.1002/mde.4072

[9] Vasquez, E., Santisteban, J., & Mauricio, D. (2023). Predicting the Success of a Startup in Information Technology Through Machine Learning. International Journal of Information

Technology and Web Engineering, 18(1), 1–17.
https://doi.org/10.4018/ijitwe.323657