International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Predicting Startup Survival Using Machine Learning

Sumit Kumar Dubey*, Abhilasha Singh*, Dr. Lakshmipathi KN**
*(Master of Computer Applications, Amity University Bengaluru
Email: sumitkumardubey1180@gmail.com, abhilashawork22@gmail.com)
**(Amity Business School, Amity University Bengaluru
Email: lakshmipathikani@gmail.com)

Abstract:
Understanding why startups succeed or fail remains a complex problem influenced by multiple uncertain factors. This paper presents a reproducible machine learning approach to predict startup survival using structured venture capital data derived from Crunchbase. A dataset containing over 17,000 startups is processed, engineered into meaningful features, and transformed into a balanced classification problem. Multiple models including Logistic Regression, Random Forest, and Gradient Boosting are evaluated. The results show that ensemble methods outperform linear models, achieving a maximum ROC-AUC of 0.74. Feature importance analysis reveals that company age and funding characteristics are the most influential predictors. The entire pipeline is designed to be reproducible, allowing future researchers to replicate and extend this work.

Keywords: Startup survival prediction, machine learning, venture capital, Random Forest, Gradient Boosting, Logistic Regression, Crunchbase, feature importance

I. I. Introduction

Startups are widely regarded as engines of innovation and economic growth. However, a large proportion of startups fail within a few years, making it difficult for investors and entrepreneurs to assess risk accurately.

Traditionally, startup evaluation has relied heavily on qualitative factors such as founder reputation, intuition, and market perception. While valuable, these methods lack consistency and scalability. With the availability of structured datasets and advances in machine learning, it is now possible to approach this problem from a data-driven perspective.

This paper explores whether machine learning models can effectively predict startup survival using historical venture capital data. The focus is not only on predictive performance but also on understanding which factors contribute most to survival.

II. II. Dataset and Problem Definition

The dataset is derived from Crunchbase and contains information on 17,727 startups. Key attributes include funding data, company category, location, and founding details.

The target variable is defined as: Survived (1) — Operating, acquired, or IPO; and Failed (0) — Closed. Since the dataset is highly imbalanced, downsampling is applied to create a balanced dataset of 1,106 samples.

III. III. Feature Engineering

The following features were constructed: Funding Total (total capital raised); Funding Rounds (number of investment rounds); Funding per Round (average capital per round); Company Age (time between founding and last funding); Founded Year (temporal indicator); and Category Code (industry classification, one-hot encoded). Missing values were handled using median imputation.

IV. IV. Methodology

Three classification models were implemented: Logistic Regression as the linear baseline, Random Forest as the ensemble tree-based method, and Gradient Boosting as the sequential boosting approach.

The dataset was split into training and testing sets using an 80/20 ratio. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. All experiments used Python with standard machine learning libraries under fixed random seeds to ensure reproducibility.

V. V. Results

A. A. Model Performance

Table I presents detailed performance metrics for all three models. Random Forest achieves the best overall performance, indicating the importance of non-linear relationships in the data.

TABLE I
Detailed Model Performance

Model	Accuracy	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression	0.62	0.64	0.62	0.62	0.62
Random Forest	0.66	0.74	0.67	0.66	0.66
Gradient Boosting	0.63	0.71	0.63	0.63	0.63

B. B. Model Comparison

Fig. 1 shows a visual comparison of Accuracy and ROC-AUC across all three models. Random Forest achieves the highest ROC-AUC of 0.74, outperforming Logistic Regression (0.64) and Gradient Boosting (0.71).

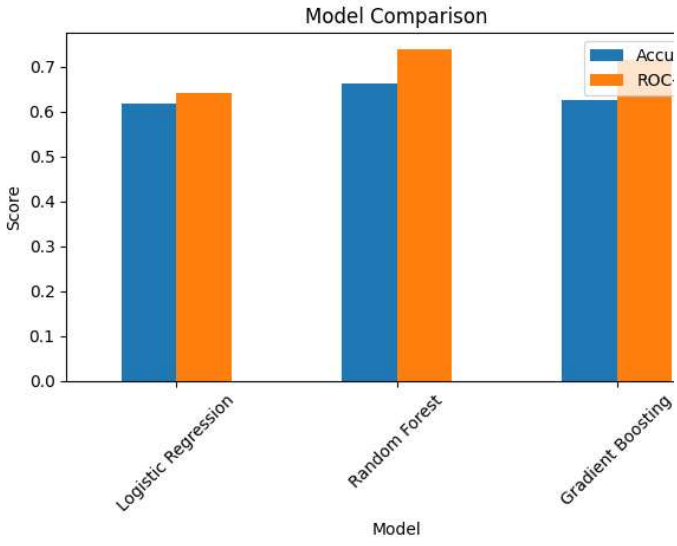

Fig. 1. Comparison of Accuracy and ROC-AUC

C. C. ROC Curves

Figs. 2-4 display the Receiver Operating Characteristic (ROC) curves for all three classifiers. The Random Forest curve shows the greatest separation from the diagonal baseline, confirming its superior discriminative power. Gradient Boosting shows moderate performance, while Logistic Regression tracks closest to the diagonal.

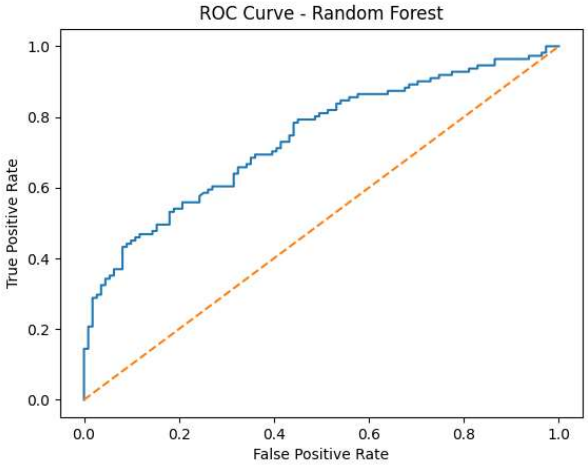

Fig. 2. ROC Curve - Random Forest

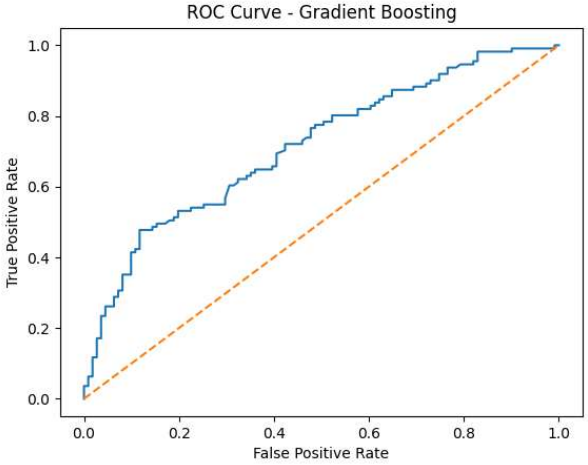

Fig. 3. ROC Curve - Gradient Boosting

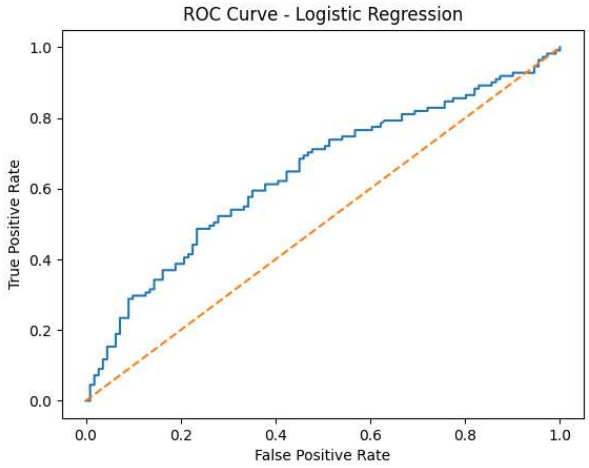

Fig. 4. ROC Curve - Logistic Regression

D. D. Feature Importance

Table II and Fig. 5 summarise the top features identified by the Random Forest model. Company Age is the single most predictive feature (0.221), followed by Funding Total (0.172), Funding per Round (0.164), Founded Year (0.155), and Funding Rounds (0.052).

TABLE II
TOP FEATURE IMPORTANCE

Feature	Importance
Company Age	0.221
Funding Total	0.172
Funding per Round	0.164
Founded Year	0.155
Funding Rounds	0.052

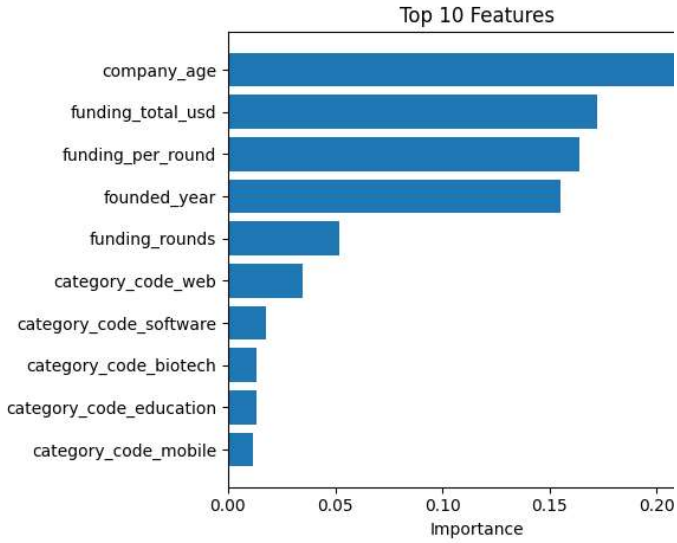

Fig. 5. Top 10 Feature Importance

E. E. Data Distribution

Figs. 6 and 7 illustrate the distributions of key input features. Funding is highly right-skewed, confirming that most startups raise modest amounts. Company age also skews right, with a concentration of younger companies reflecting typical venture-backed startup lifecycles.

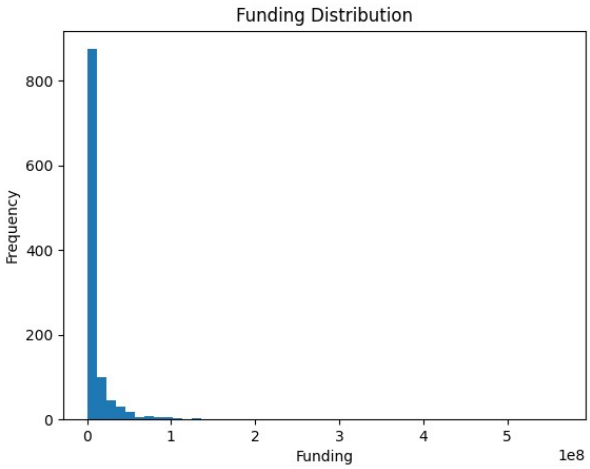

Fig. 6. Funding Distribution

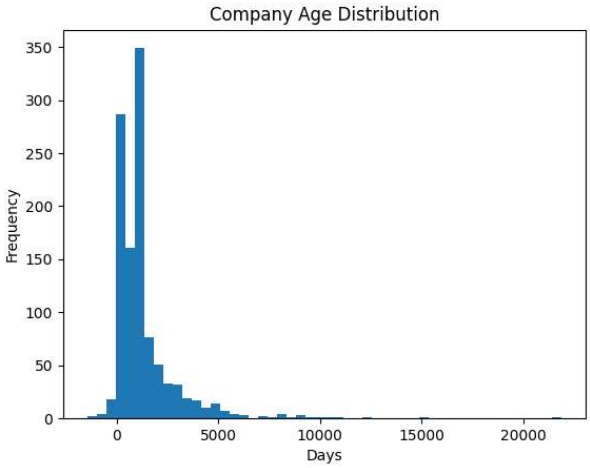

Fig. 7. Company Age Distribution

VI. VI. DISCUSSION

The results show that startup survival is strongly influenced by financial and temporal factors. Company age emerges as the most important feature, suggesting that longevity itself is a key indicator of stability. Companies that remain active for longer periods demonstrate resilience, operational maturity, and the ability to adapt to market conditions.

Funding-related variables also play a dominant role, confirming that access to capital significantly increases survival probability. Industry categories contribute less, indicating that while sector matters, financial strength is more critical.

The moderate performance of all models reflects the inherent uncertainty of startup success, which is influenced by factors not captured in the dataset, such as founder quality, market timing, and competitive dynamics.

VII. VII. Reproducibility

To ensure reproducibility, all steps of the machine learning pipeline are systematically defined, including data preprocessing, feature engineering, model training, and evaluation. The dataset is derived from Crunchbase and processed using consistent data cleaning and transformation procedures. Missing values are handled using median imputation, and categorical variables are encoded using one-hot encoding.

The experiments are conducted using Python with standard machine learning libraries. Model training is performed using fixed random seeds to ensure consistent results across runs. Hyperparameters for each model are explicitly defined and kept constant during evaluation. All figures and results presented in this paper are generated programmatically from the processed dataset. The complete implementation, including code and instructions for data preparation, can be made available to enable other researchers to replicate and extend this work.

VIII. VIII. Limitations

Several limitations affect this study. The dataset exhibits a bias toward successful startups since companies that never receive formal venture funding are unlikely to appear in Crunchbase. The absence of qualitative variables such as team composition, product quality, and competitive positioning limits the predictive ceiling of any purely data-driven approach. Additionally, the dataset has limited temporal depth, preventing long-horizon survival analysis across economic cycles.

IX. IX. Conclusion

This paper demonstrates that machine learning can provide useful insights into startup survival. While prediction performance is moderate, the analysis highlights key factors such as funding and company age as primary determinants of survival. Ensemble methods, particularly Random Forest, offer meaningful improvements over linear baselines. Future work can focus on incorporating richer datasets including qualitative signals, social network data, and market indicators, as well as exploring more advanced modeling techniques such as deep learning and survival analysis to improve predictive accuracy.

X. References

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.
[4] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[7] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[8] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," The Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.
[9] S. Shane, The Illusions of Entrepreneurship: The Costly Myths That Entrepreneurs, Investors, and Policy Makers Live By. Yale University Press, 2009.
[10] P. Gompers, A. Kovner, J. Lerner, and D. Scharfstein, "Venture capital investment cycles: The impact of public markets," Journal of Financial Economics, vol. 87, no. 1, pp. 1–23, 2008.
[11] CB Insights, "The Top 20 Reasons Startups Fail," 2021. [Online]. Available: https://www.cbinsights.com/research/startup-failure-reasons-top/
[12] OECD, "Entrepreneurship at a Glance 2020," OECD Publishing, 2020.
[13] World Bank, "Entrepreneurship Database," 2020. [Online]. Available: https://www.worldbank.org/
[14] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.
[15] N. N. Taleb, The Black Swan: The Impact of the Highly Improbable. Random House, 2007.
[16] M. Marmer et al., "Startup Genome Report: A new framework for understanding why startups succeed," Startup Genome, 2011.
[17] E. Mollick, "The dynamics of crowdfunding: An exploratory study," Journal of Business Venturing, vol. 29, no. 1, pp. 1–16, 2014.
[18] M. E. Porter, On Competition. Harvard Business School Publishing, 2008.
[19] R. Brown and S. Mawson, "Entrepreneurial ecosystems and public policy in action," Regional Studies, vol. 53, no. 5, pp. 689–701, 2019.

