Forecasting Startup Futures: Leveraging Data Science for Risk and Success Analysis

Aanchal Bafna, Prathmesh Mhatre, Vedant Wani

April 02, 2025

Abstract

Entrepreneurial ventures face substantial risks, with high failure rates due to factors such as market misalignment, funding challenges, and team dynamics. Traditional risk assessment methods often lack transparency, adaptability, and the ability to predict outcomes accurately. This research aims to address these limitations by developing interpretable machine learning and deep learning models that integrate data from multiple sources, including Crunchbase and SEMrush, to predict startup success and assess associated risks. Using advanced algorithms such as CatBoost, XGBoost and Neural Networks, the models are optimized through grid search and hyperparameter tuning. The study identifies key success metrics and risk factors while incorporating Explainable AI techniques like SHAP to ensure transparency and interpretability of the predictions. The model provides actionable insights that empower investors, entrepreneurs, and policymakers to make informed, data-driven decisions. This research contributes to improving decision-making processes and fostering innovation, ultimately supporting sustainable growth within the entrepreneurial ecosystem.

Keywords: Risk forecasting model, Explainable AI, Startup success forecasting, Dataset integration, Machine Learning, Deep Learning, Decision-making, Neural networks.

1. Introduction

The global startup ecosystem, valued at over $4 trillion, plays a crucial role in driving economic development, innovation, and job creation. Despite the rapid expansion of startups and a substantial inflow of venture capital$445 billion in 2023-2024 alone approximately 90% of startups fail within ten years. The primary reasons for failure include investment inefficiencies, poor product-market fit, operational bottlenecks, and scalability challenges. These failures not only lead to financial losses for investors but also hinder technological advancement and economic growth.

Traditional approaches to predicting startup success primarily focus on growth metrics such as revenue, user acquisition, and funding rounds. While these indicators provide valuable insights, they fail to capture the underlying risks that contribute to failure. Many predictive models rely on single-source datasets, limiting their ability to account for diverse factors affecting startups. Additionally, the lack of interpretability in existing models makes it difficult for entrepreneurs, investors, and policymakers to derive actionable insights. Furthermore, deep learning techniques, which have demonstrated superior performance in various predictive tasks, remain underutilized in startup forecasting. A significant challenge in using advanced AI models is their "black-box" nature, which reduces trust in their predictions. The application of Explainable AI (XAI) methods such as SHAP and Integrated Gradients can enhance interpretability, making predictions more transparent and actionable for stakeholders.

This study proposes a novel, explainable machine learning and deep learning framework for predicting startup success and assessing risk factors. By integrating Crunchbase and SEMrush datasets, the model aims to provide a more balanced and data-driven understanding of startup trajectories. The key contributions of this research include:

1. Building a predictive framework using machine learning and deep learning.
2. Identifying key risk factors contributing to startup failures.
3. Applying Explainable AI (XAI) techniques to enhance model transparency.
4. Providing data-driven insights to support investors and entrepreneurs.

This research advances startup analytics, aiding investors in funding decisions and helping entrepreneurs refine sustainable business models.

Global Startup Ecosystem Report 2023 (GSER 2023) Link

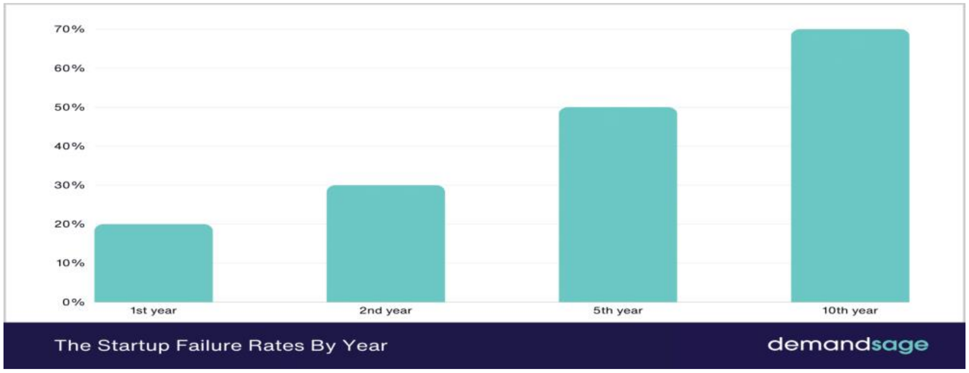


Figure 1: Shocking Startup Failure Rates Link

2. Literature Review

1] Potsulin, A., Sergeeva, I., Aleksandrova, A., Kuporov, Y., & Shik, I. (2024). Developing a model for forecasting risks of innovative entrepreneurial projects with machine learning tools. *Pakistan Journal of Life and Social Sciences*, *22*(2), 2549-2561.The paper aim to forecast the risks associated with innovative entrepreneurial projects, focusing on the factors influencing the success or failure of these projects. The authors analyze historical data from startup projects to predict risks and assess the effectiveness of these projects for economic growth. The study applies **machine learning** techniques to predict the risks of innovative projects, using historical data sourced from a startup graveyard website. The exact models or techniques used are not specified. Limited by the lack of data analysis for successful and unsuccessful innovative business projects. The study relies on data from projects that have already failed, which might not fully capture successful entrepreneurial project dynamics.

2] Pandya, D. D., Patel, A. K., Purohit, J. M., Bhuptani, M. N., Degadwala, S., & Vyas, D. (2023, April). Forecasting number of Indian startups using supervised learning regression models. In *2023 International Conference on Inventive Computation Technologies (ICICT)* (pp. 948-952).IEEE. This paper forecasts the future growth of startups in India, focusing on key variables such as financing, market demand, and competition. The goal is to provide insights into the factors influencing the growth of Indian startups and help policymakers and investors understand the startup ecosystem. The study uses **Supervised Learning Regression models**, particularly **linear regression models**, to predict the number of startups in India. Data from the Startup Database, official papers, and scholarly journals inform the analysis. The study is limited to linear regression models, which may not fully capture non-linear trends in startup growth. It may not account for unexpected external factors or market shifts, limiting the accuracy of predictions over time.

3] Bhattacharya, D. (2024). Utilizing Base Machine Learning Models to Determine Key Factors of Success on an Indian Tech Startup. The paper aims to determine the key factors contributing to the success of Indian tech startups. It focuses on the differences between the Indian and American startup ecosystems and seeks to provide insights for entrepreneurs on how to successfully raise and scale their companies. The study uses **base machine learning models** such as **Random Forest**, **XGBoost**, **LightGBM**, and **CatBoost** to predict the success of startups. It also employs **feature importance tools** provided by these models to identify the most influential factors for startup success. The study only analyzes startups that are either acquired, have acquired other companies, or went public, which may not reflect the full spectrum of startup success. The analysis is based on data from Crunchbase, which may have inherent biases or incomplete data for certain startups.

4] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023) [6]. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33. This paper provides an in-depth review of **Explainable AI (XAI)**, exploring its core ideas, techniques, and solutions. The authors examine the need for explainability in AI systems and propose various techniques to enhance the interpretability and transparency of machine learning models, particularly in high-stakes areas like healthcare, finance, and autonomous systems.The study reviews multiple **XAI techniques** such as **Model-Agnostic Methods** (e.g., LIME, SHAP), **Interpretable Models**, **Post-Hoc Explanations** . They explore both **local** and **global** explainability approaches and highlight **transparent models** like decision trees and linear regression. The study is more of a review and doesn't focus on implementing or testing these XAI methods in real-world applications. There is no specific case study or empirical analysis provided, which may limit the practical application of the findings. The paper does not extensively cover the challenges and limitations of adopting XAI in complex deep learning models, especially for non-experts.

3. Methodology

3.1 Datasets

The process begins with gathering a variety of datasets that include important risk and success metrics, providing a comprehensive view of startup dynamics:

3.1.1 Crunchbase Dataset: This dataset offers key features that are critical for assessing startup performance, including detailed information on funding rounds, team size, industry type, market reach, and acquisition status. By examining these metrics, we can gain insights into the financial health, growth potential, and strategic positioning of startups within their respective industries.

3.1.2 SEMrush Data: SEMrush provides valuable digital performance metrics, such as website traffic, SEO rankings, and rank growth. These data points offer context on a startup's online presence and market reach, which are increasingly important for evaluating a startup's competitive positioning and visibility. SEMrush data also helps to assess a startup's digital footprint and its effectiveness in engaging and retaining customers, factors that directly influence long-term sustainability and success. SEMrush data fills in the gaps in conventional financial indicators by offering additional insights into a startup's market presence.

Together, these datasets are integrated to provide a well-rounded understanding of a startup's performance, allowing for a more robust and nuanced analysis of both risk and success factors. The combination of these sources enables a holistic approach to prediction, factoring in not only financial and team-related metrics but also digital and competitive elements, thus enhancing the accuracy and depth of the model's predictions.

3.2 Model Building

To develop an effective and interpretable predictive model for startup success and failure, we leverage **CatBoost, XGBoost, and Neural Networks**. Each of these models is chosen based on its strengths in handling structured data, feature importance analysis, and predictive accuracy. Below, we justify their use in this study, explain their specific roles, and highlight their effectiveness in addressing the research problem.

3.2.1 CatBoost

CatBoost (Categorical Boosting) is particularly effective for datasets with categorical variables, such as industry type, funding stages, and acquisition status, which are prevalent in Crunchbase and SEMrush data. Unlike traditional tree-based models, CatBoost efficiently encodes categorical features without requiring extensive preprocessing, reducing data leakage risks and improving performance. CatBoost is employed to analyze structured startup data, capturing nonlinear relationships between success indicators and risk factors. It excels in ranking tasks, making it suitable for evaluating the likelihood of startup success based on multiple features.

How does it address the research problem?

- **Better handling of categorical data:** Startups operate in diverse industries, and categorical features such as sector, location, and funding rounds and web traffics are crucial in predicting success.
- **Robust performance with small datasets:** It mitigates overfitting issues while maintaining high accuracy, even with limited labeled data.
- **Interpretability:** CatBoost provides built-in feature importance, allowing insights into which factors most influence startup success or failure.

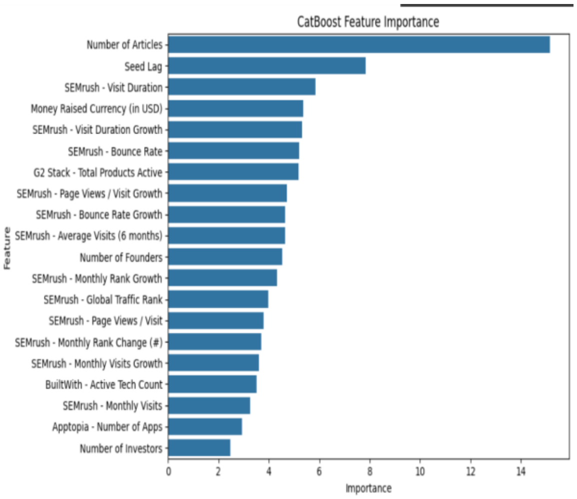


Figure 2: CatBoost Feature Importance on CrunchBase.

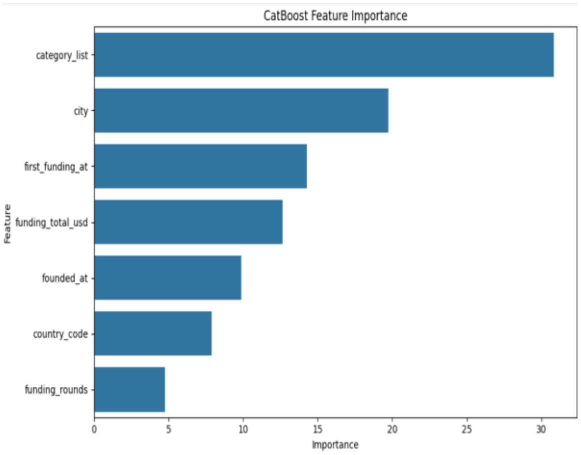


Figure 3: CatBoost Feature Importance on SEMrush Dataset

3.2.2 XGBoost

XGBoost (Extreme Gradient Boosting) is one of the most powerful gradient boosting algorithms for structured data. It is known for its efficiency, scalability, and ability to handle missing values, making it a strong choice for large and diverse startup datasets. XGBoost is used for predictive modeling, identifying key success and risk factors that influence startup outcomes. By iteratively refining decision trees, it captures complex interactions between features such as funding, market performance, and team composition.

How does it address the research problem?

- **Handles imbalanced datasets effectively:** Since startup failures significantly outnumber successful cases, XGBoost's ability to balance predictions is crucial.
- **Feature importance analysis:** It ranks key predictors, helping to identify which factors (funding, team size, digital presence, SEMrush Monthly Rank Growth) contribute most to startup success.
- **Optimized computational efficiency:** Unlike traditional boosting algorithms, XGBoost reduces training time while maintaining accuracy, making it suitable for large datasets.

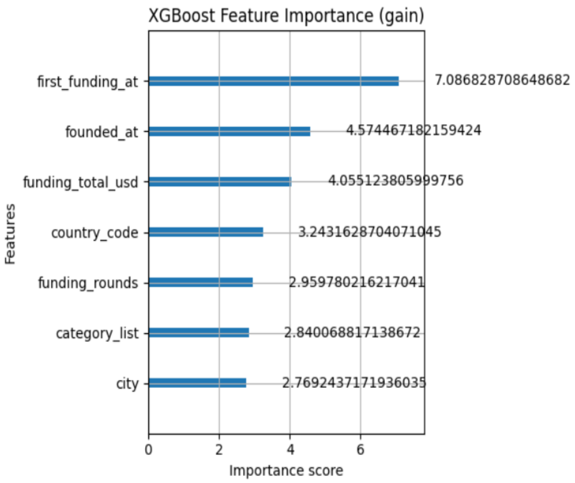


Figure 4: XGBoost Feature Importance on CrunchBase.

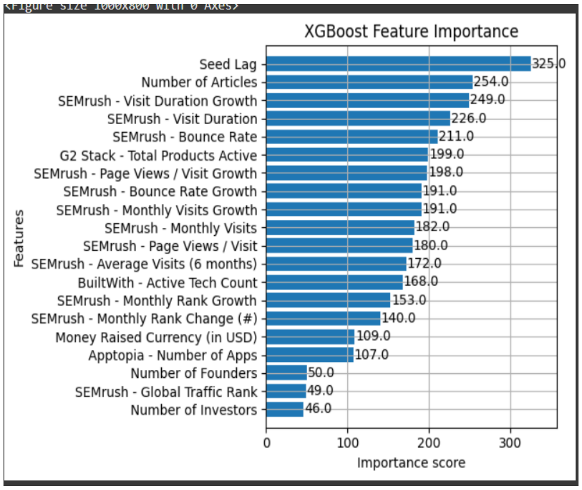

```
<Figure size 1000x800 with 0 Axes>
```

Figure 5: XGBoost Feature Importance on SEMrush Dataset

3.2.3 Neural Network

While tree-based models like CatBoost and XGBoost excel at structured data, Neural Networks provide additional advantages in capturing highly nonlinear and deep feature interactions. They are particularly useful when incorporating high-dimensional data, such as SEMrush's web traffic metrics and time-series trends. Neural Networks are employed to model complex relationships between financial, operational, and digital success factors. By leveraging multiple hidden layers, they uncover intricate patterns that may not be evident with traditional models.

How does it address the research problem?

- **Learning from unstructured and high-dimensional data:** Neural Networks process web traffic trends, funding history, and team compositions more effectively than traditional models.
- **Time-series prediction capability:** Helps forecast startup growth based on evolving funding rounds, traffic, and other progressive variables.
- **Robust deep feature extraction:** Identifies subtle correlations between digital presence and startup survival, contributing to more informed predictions.

4. XAI-Driven Training and Validation

We split the dataset into **training, validation, and test sets** to ensure robust performance evaluation. The split follows:

Training Set (70%): Used to fit the models.

Validation Set (15%): Used for hyperparameter tuning and early stopping.

Test Set (15%): Used for final evaluation and generalization assessment.

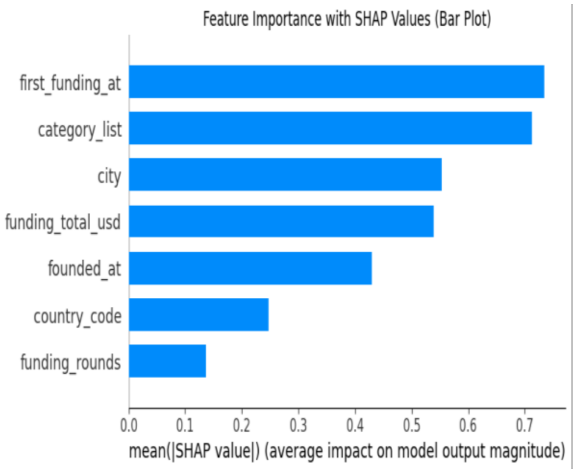
To prevent overfitting:

- **CatBoost and XGBoost** use **grid search and early stopping** to optimize parameters.
- **Neural Networks** use **dropout regularization, batch normalization, and Adam optimizer**.

By combining **SHAP with training and validation**, we ensure not only accurate predictions but also transparent, interpretable, and actionable insights for decision-makers in the startup ecosystem.

4.1 Feature Importance and Feature Distribution with SHAP Values

A] CatBoost:



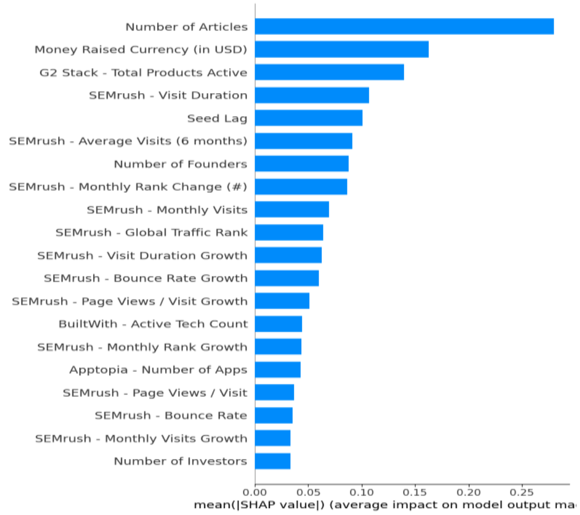
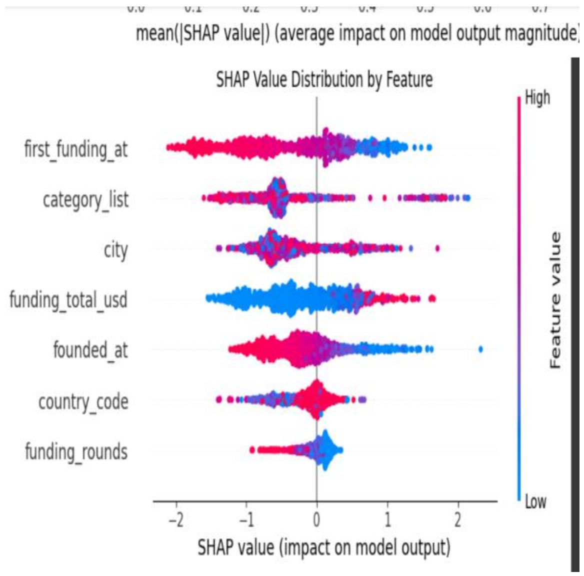


Figure 6: Feature Importance on CrunchBase with SHAP Values.

Figure7: Feature Importance on SEMrush with SHAP Values.



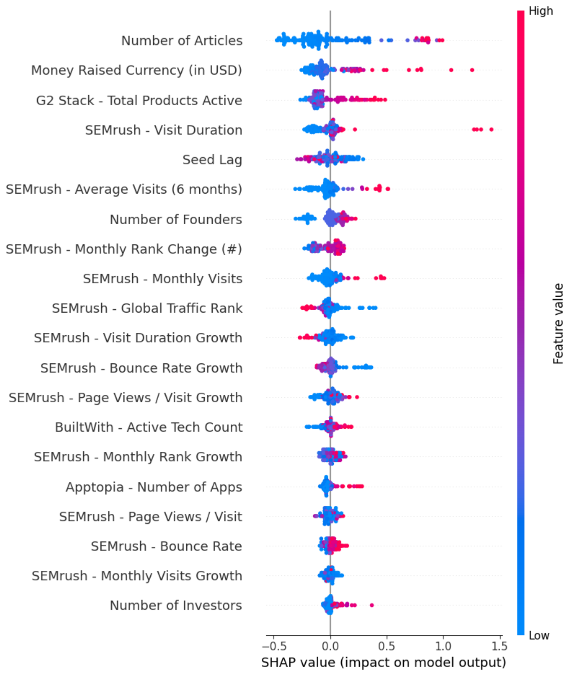


Figure8: SHAP Value Distribution by Feature on Crunchbase Dataset.

Figure9: SHAP Value Distribution by Feature on SEMrush Dataset

B] XGBoost:

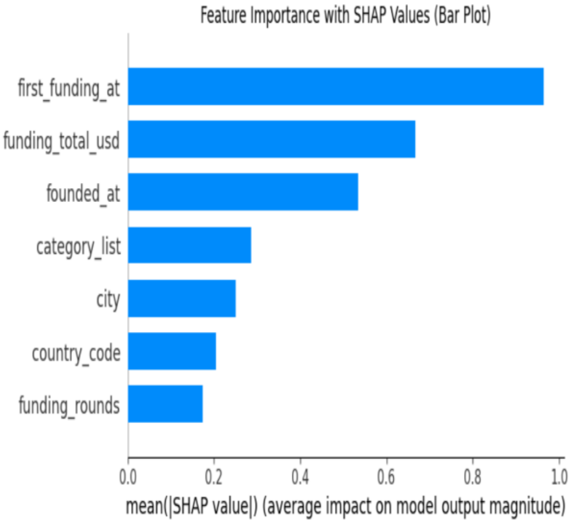


Figure10: Feature Importance on CrunchBase with SHAP Values

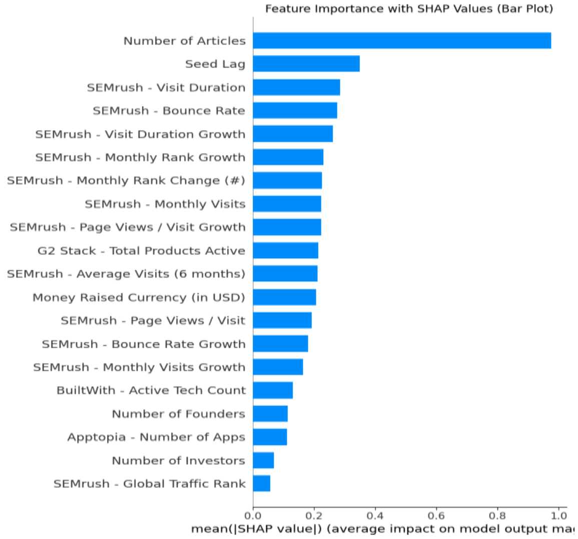


Figure11: Feature Importance on SEMrush with SHAP Values

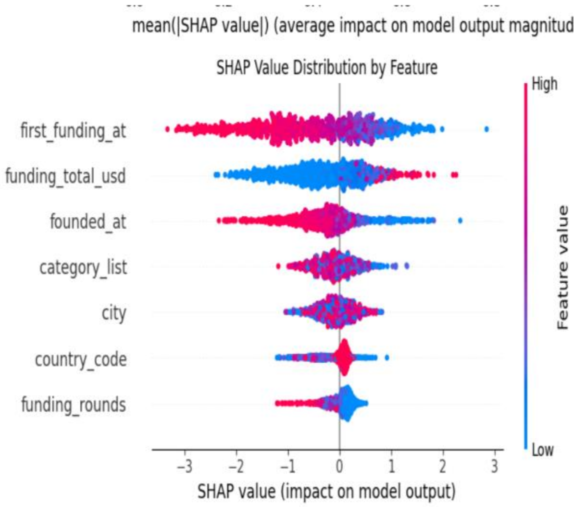


Figure12: SHAP Value Distribution by Feature on Crunchbase Dataset

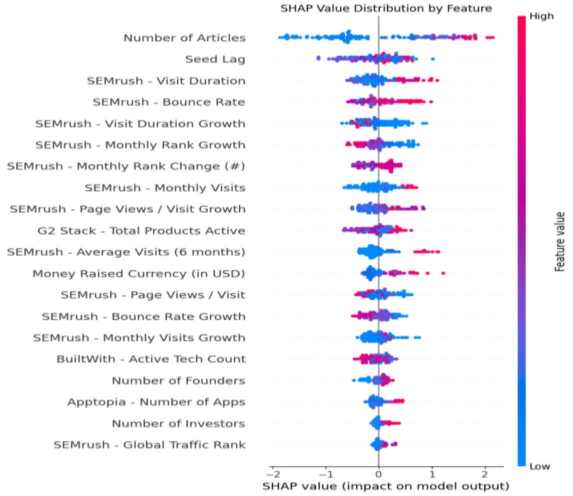


Figure13: SHAP Value Distribution by Feature on SEMrush Dataset

5. Results:

The model architectures were trained with various different combinations of layers, dropout rates, number of layers and other hyper-parameters. The results of the trained models are summarized in the table 1. All models were tested on the test set, calculating accuracy, recall, precision and F1 score.

Accuracy: Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. It is a measure of the overall correctness of the model.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Recall (Sensitivity or True Positive Rate): Recall is the ratio of correctly predicted positive observations to the all observations in actual class. It gives the completeness of the model, i.e., the proportion of actual positive cases which are correctly identified.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the accuracy of the positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

F1 Score: F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is a better measure than accuracy for imbalanced datasets.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 1: Classification Report

Models with Datasets	**Class**	**Precision**	**Recall**	**F-1Score**	**Accuracy**
CatBoost_Crunchbase	0	0.98	0.97	0.97	0.95
	1	0.77	0.83	0.80	
CatBoost_SEMrush	0	0.96	0.99	0.97	0.96
	1	0.97	0.87	0.92	
XGBoost_Crunchbase	0	0.97	0.89	0.93	0.88
	1	0.49	0.77	0.60	
XGBoost_SEMrush	0	0.79	0.93	0.85	0.76
	1	0.58	0.29	0.39	
NeuralNetwork_Crunchbase	0	0.94	0.88	0.91	0.84
	1	0.38	0.53	0.44	
NeuralNetwork_SEMrush	0	0.81	0.97	0.88	0.81
	1	0.81	0.34	0.48	

CatBoost (SEMrush) is the best-performing model, **excelling in failure detection and overall accuracy**.

CatBoost (Crunchbase) is also strong, but slightly weaker for failure prediction.

XGBoost and Neural Networks need improvements in detecting **failing startups (Class 1)**. For startup risk prediction, **CatBoost (SEMrush) is the optimal choice**, offering a **balanced, explainable, and highly accurate model**.

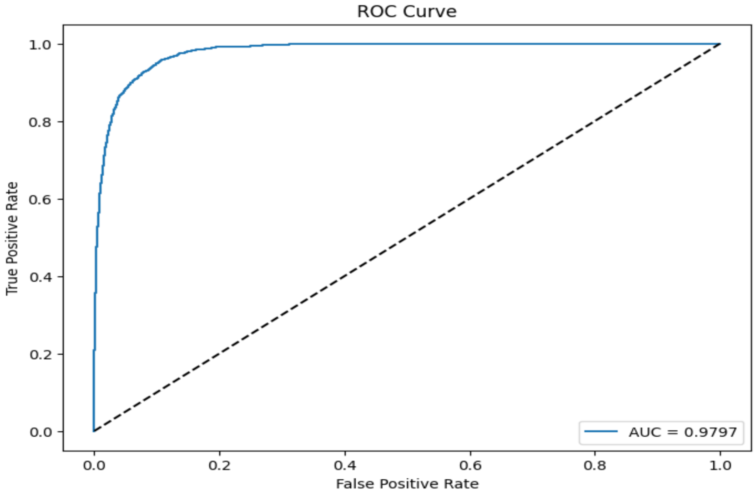


Figure 14: ROC Curve (CatBoost) of Crunchbase dataset

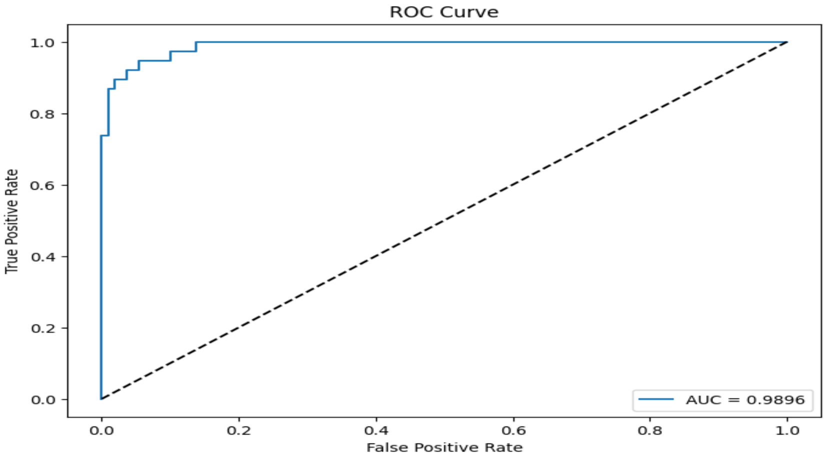


Figure 14: ROC Curve (CatBoost) of SEMrush dataset

6. Future Scope & Recommendations

Future research should focus on integrating real-time financial, market sentiment, and investor network data to improve predictive accuracy. A multimodal approach, combining structured and unstructured data, can enhance insights. Developing dynamic models that adapt to changing market conditions and leveraging advanced Explainable AI (XAI) techniques will improve transparency and stakeholder trust. Expanding the framework to other industries, such as healthcare and finance, will increase its applicability. Real-time forecasting through cloud-based solutions and user-friendly dashboards can empower entrepreneurs and investors, while ethical AI considerations must be prioritized to ensure fairness and bias mitigation.

7. Conclusion

This study presents a **comprehensive, data-driven framework** for predicting startup success and assessing risk factors using **machine learning and deep learning techniques**. By integrating **multi-source datasets** from **Crunchbase** and **SEMrush**, the model provides a holistic view of both **success indicators** (funding, team size, digital presence) and **failure risks** (market misalignment, operational inefficiencies, and financial instability). The implementation of **CatBoost, XGBoost, and Neural Networks** allows for capturing complex, non-linear relationships within startup data. Additionally, leveraging **Explainable AI (XAI) techniques** such as **SHAP** ensures that predictions remain transparent, interpretable, and actionable for investors, entrepreneurs, and policymakers. For startup risk prediction, **CatBoost (SEMrush) is the optimal choice**, offering a **balanced, explainable, and highly accurate model**.

8. References

Potsulin, A., Sergeeva, I., Aleksandrova, A., Kuporov, Y., & Shik, I. (2024) [1]. Developing a model for forecasting risks of innovative entrepreneurial projects with machine learning tools. *Pakistan Journal of Life and Social Sciences*, *22*(2), 2549-2561.

Pandya, D. D., Patel, A. K., Purohit, J. M., Bhuptani, M. N., Degadwala, S., & Vyas, D. (2023, April) [2]. Forecasting number of Indian startups using supervised learning regression models. In *2023 International Conference on Inventive Computation Technologies (ICICT)* (pp. 948-952).

Bhattacharya, D. (2024) [3]. Utilizing Base Machine Learning Models to Determine Key Factors of Success on an Indian Tech Startup.

Ünal, C. (2019) [4]. *Searching for a unicorn: A machine learning approach towards startup success prediction* (master's thesis, Humboldt-Universität zu Berlin).

Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024) [5]. Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, *13*(1), 80.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023) [6]. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

Żbikowski, K., & Antosiuk, P. (2021) [7]. A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), 102555.