Most important features in a startup's success

Gaby Levis, Ariel Siman Tov, Zohar Attar, Tal Klein, Yarden Tzaraf

July 7, 2024

Abstract

In the dynamic and competitive landscape of startups, identifying the key factors that drive success is essential. This study aims to uncover these critical features by analyzing a comprehensive dataset that includes geographical locations, financial metrics, industry sectors, funding history, and outcomes of various startups. Through extensive data cleaning, feature engineering, and exploratory data analysis (EDA), we identified significant patterns and correlations that highlight the elements influencing startup success.

We employed Random Forest (RF) and Neural Network (NN) models to predict startup outcomes, finding that RF outperformed NN across multiple metrics, including AUC, accuracy, precision, recall, and F1 score. The results underscored the importance of financial health and growth trajectories, with metrics such as the Funding to Milestones Ratio and Company Age at Last Funding Year emerging as critical predictors. Additionally, SHAP values provided valuable insights into feature impacts, emphasizing the significance of relationships, financial metrics, and strategic milestones.

Overall, this research provides a detailed and actionable understanding of the factors contributing to startup success, offering practical recommendations for entrepreneurs and investors. Future work will focus on refining the predictive models, enhancing interpretability, and addressing any data imbalances to further improve the robustness and applicability of our findings.

1 Introduction

In the rapidly evolving world of startups, understanding the key factors that influence a startup's success is crucial. Our project aims to identify these critical features by analyzing a comprehensive dataset of various startups, detailing their geographical locations, financial metrics, industry sectors, funding history, and eventual outcomes—whether they were acquired or not.

2 Related Work

2.1 Econometric Estimation and Founder Personalities

The study "Econometric Estimation of the Factors That Influence Startup Success" identifies key success indicators such as significant revenue and obtaining financing. Factors like location, dedication of promoting partners, commercial ability, age of the startup, presence of investor partners, and founders' technological training are highlighted as critical. On the other hand, "The Impact of Founder Personalities on Startup Success" examines startup success through liquidity events and identifies distinct personality traits (adventurousness, modesty, activity level) and diverse founder teams as significant success factors. Both studies emphasize the importance of founder characteristics and strategic positioning.

2.2 Predictive Models for Startup Success

The research "Predicting Startup Success in the U.S." develops and compares several models (Logistic Regression, Linear Discriminant Analysis, Extreme Gradient Boosting, and Support Vector Machine) to predict startup success, with Logistic Regression showing reliable and conservative results. The study emphasizes the importance of capturing additional risk factors to enhance prediction accuracy.

Complementarily, the Kaggle project "Startup Success Prediction" used data preprocessing, Multinomial Logit Model, and Random Forest Model to evaluate features impacting success. The project highlights the foundation year, milestones, relationships, and total received funds as significant variables, with Random Forest providing better classification accuracy but more complexity.

2.3 Model Interpretability

The article "Using SHAP Values to Explain How Your Machine Learning Model Works" introduces SHAP values for model interpretability, emphasizing their importance in understanding feature impact on predictions. SHAP's ability to provide local and global explanations makes it a critical tool for ensuring model transparency. Similarly, "Understanding Random Forest Algorithms with Examples" explains the mechanics and applications of Random Forests, highlighting their robustness, ease of use, and feature importance ranking.

2.4 Differences in Our Project

Our project aims to investigate factors influencing startup success using a combination of advanced machine learning models and interpretability tools like SHAP. Unlike previous studies, we will integrate a broader range of features and possibly new data sources to capture nuanced factors. By combining insights from related work and focusing on both prediction and interpretability, our project will provide practical recommendations for entrepreneurs and investors, offering a comprehensive analysis of geographical, financial, team dynamics, and technological aspects. This approach aims to build on existing knowledge while addressing gaps identified in prior research, providing a detailed and actionable understanding of startup success factors.

3 Methodology

3.1 About the Data

3.1.1 The dataset encompasses a wide range of attributes for each startup such as:

- **Geographical Information**: Including state code, latitude, longitude, zip code, and city.
- **Identifiers**: Unique IDs for each startup and record, along with the startup's name.
- **Temporal Data**: Key dates such as founding date, first and last funding dates, and milestone achievements.
- **Financial Metrics**: Funding rounds, total funding received, and the number of significant business relationships or partnerships.
- **Industry and Domain**: Industry sector, presence in specific domains like software, web, mobile, etc.
- **Investment Data**: Indicators of venture capital funding, angel investments, and specific funding rounds.
- **Outcome**: The current status of the startup (active, acquired, or closed).

3.2 Data Cleaning and Feature Engineering:

To ensure the accuracy and relevance of our analysis, we conducted extensive data cleaning and feature engineering:

3.2.1 Data Cleaning:

Addressed missing values, removed duplicate and redundant features, and converted relevant columns into appropriate formats.

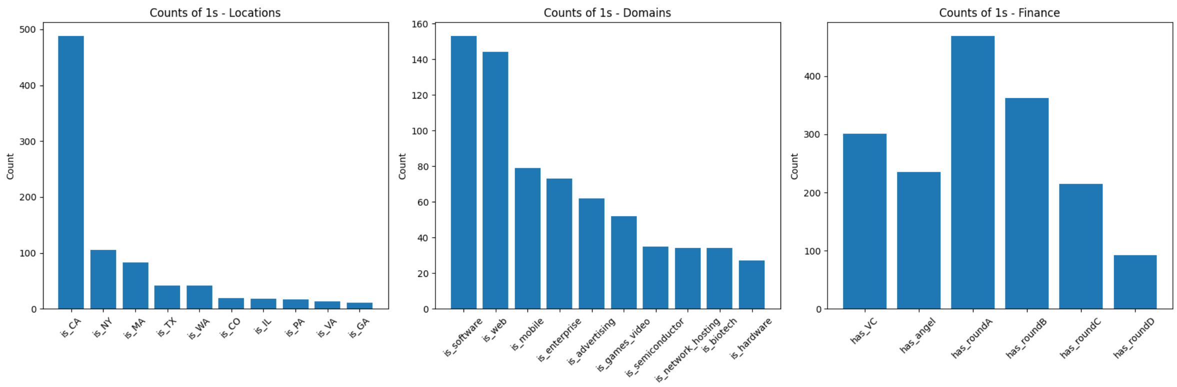


Figure 1: General Statistics about the data

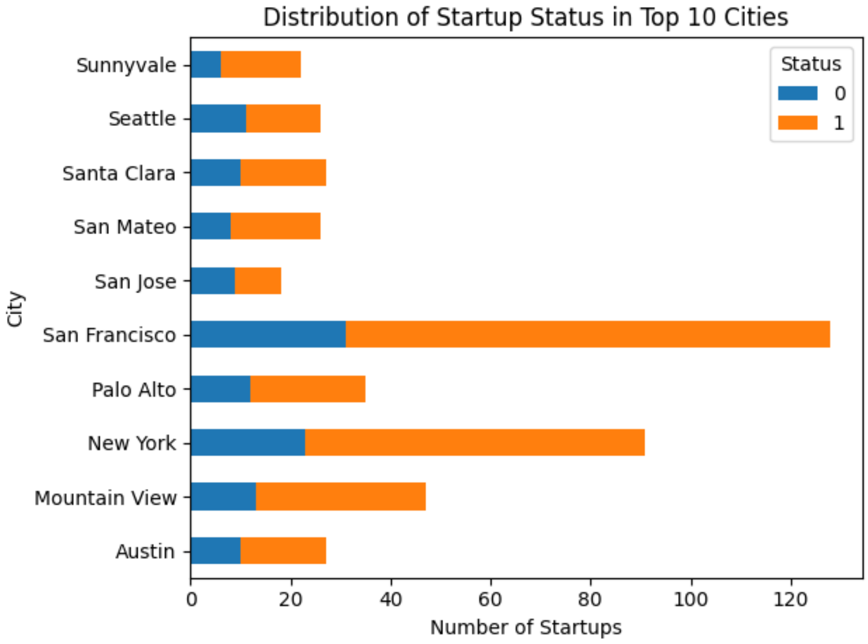


Figure 2: Status represents weather the startup was sold or shut down

3.2.2 Feature Engineering:

– **Company Age**: Calculated from the difference between founding year and last funding year.

– **Funding Frequency**: Number of funding rounds divided by the age of the company.

– **Funding to Milestones Ratio**: Total funding received to the number of milestones achieved.

3.3 Insights from Exploratory Data Analysis (EDA)

3.3.1 Our EDA revealed significant patterns and correlations:

– **Geographical Insights**: A high concentration of startups in California, especially in San Francisco.

– **Temporal Trends**: Most startups were founded between 1995 and 2010, with peak funding activities around 2005-2015.

- **Financial Insights**: Startups typically received 2 to 4 funding rounds, with substantial variations in total funding.
- **Status and Attributes**: Many startups did not acquire, but there was significant participation in venture capital and angel investments.
- **Industry Dominance**: A strong presence in the software and web sectors.

3.4 Model Selection

When selecting a model to effectively address our research question of **identifying the key factors influencing startup success**, we opted to evaluate the performance of two classification models: **Random Forest and Neural Network.**

Random Forest (RF): RF is chosen for its ability to handle complex datasets with diverse predictors effectively. By aggregating predictions from multiple decision trees, RF can capture non-linear relationships and feature interactions crucial for understanding startup outcomes. Its capability to rank feature importance provides valuable insights into which factors drive success or failure in startup ventures.

Neural Network (NN): NNs were selected due to their capacity to learn intricate patterns and representations from data through layers of neurons. This model excels in scenarios where startup success depends on nuanced relationships across various factors.

These models were chosen for their ability to discern complex relationships within data and provide insights into which variables most significantly impact the outcome of startup ventures.

By comparing and analyzing the results from RF and NN models, we aim to gain comprehensive insights into the factors most influential in determining the success of startups.

3.5 Model Training and Testing

We dropped non-feature columns such as name, city, status, and others. The dataset was then split into training (80%) and testing (20%) sets. The models were trained on the training set, and their performance was evaluated on the testing set using metrics including ROC, AUC, accuracy, precision, recall, and F1 score.

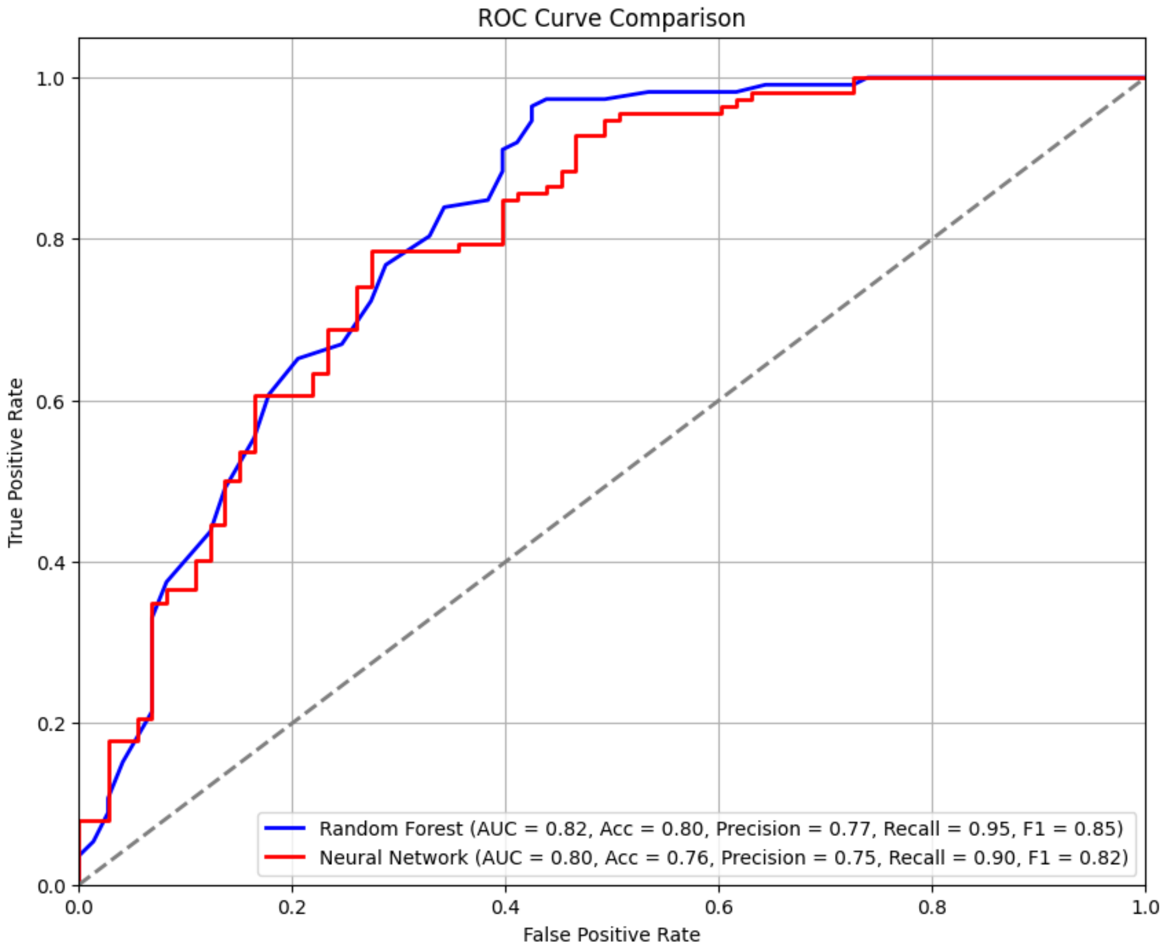


Figure 3: ROC Curve

- **AUC:** Random Forest (RF) has a slightly higher AUC (0.82 vs 0.80), indicating better overall performance in distinguishing between positive and negative cases across all thresholds.
- **Accuracy**: RF is slightly more accurate (0.80 vs 0.76), meaning it makes fewer incorrect predictions overall.
- **Precision:** RF has slightly higher precision (0.77 vs 0.75), indicating it has a lower false positive rate.
- **Recall:** RF has higher recall (0.95 vs 0.90), indicating it's better at identifying true positives.
- **F1 Score**: RF has a slightly higher F1 score (0.85 vs 0.82), suggesting a better balance between precision and recall.

Overall, Random Forest outperforms the Neural Network (NN) across all these metrics. The higher recall values for both models indicate their strength in identifying positive cases, with RF showing a slight edge in precision and overall performance. This comprehensive evaluation leads to the conclusion that **Random Forest is the preferred model for predicting startup outcomes** in this scenario, given its superior performance across multiple key metrics.

4 Results

The dataset provided comprehensive insights into various aspects of startup companies, including geographic locations, financial metrics, industry sectors, funding history, and outcomes (acquisition or closure). Through exploratory data analysis (EDA), significant patterns and correlations were identified that shed light on factors influencing startup success.

Key findings from the analysis include:

4.0.1 Financial and Operational Insights:

Startups exhibited a wide range in total funding received, with outliers receiving substantial investments. Metrics such as Funding to Milestones Ratio and Company Age at Last Funding Year emerged as critical predictors of startup outcomes.

4.0.2 Geographic and Temporal Trends:

California, particularly San Francisco, emerged as a dominant location for startups in the dataset. Most startups were founded between 1995 and 2010, with significant funding activities observed around 2005-2015.

4.0.3 Model Performance:

Random Forest (RF) outperformed Neural Networks (NN) across multiple metrics (AUC, accuracy, precision, recall, F1 score), indicating its robustness in predicting whether a startup would be acquired or closed.

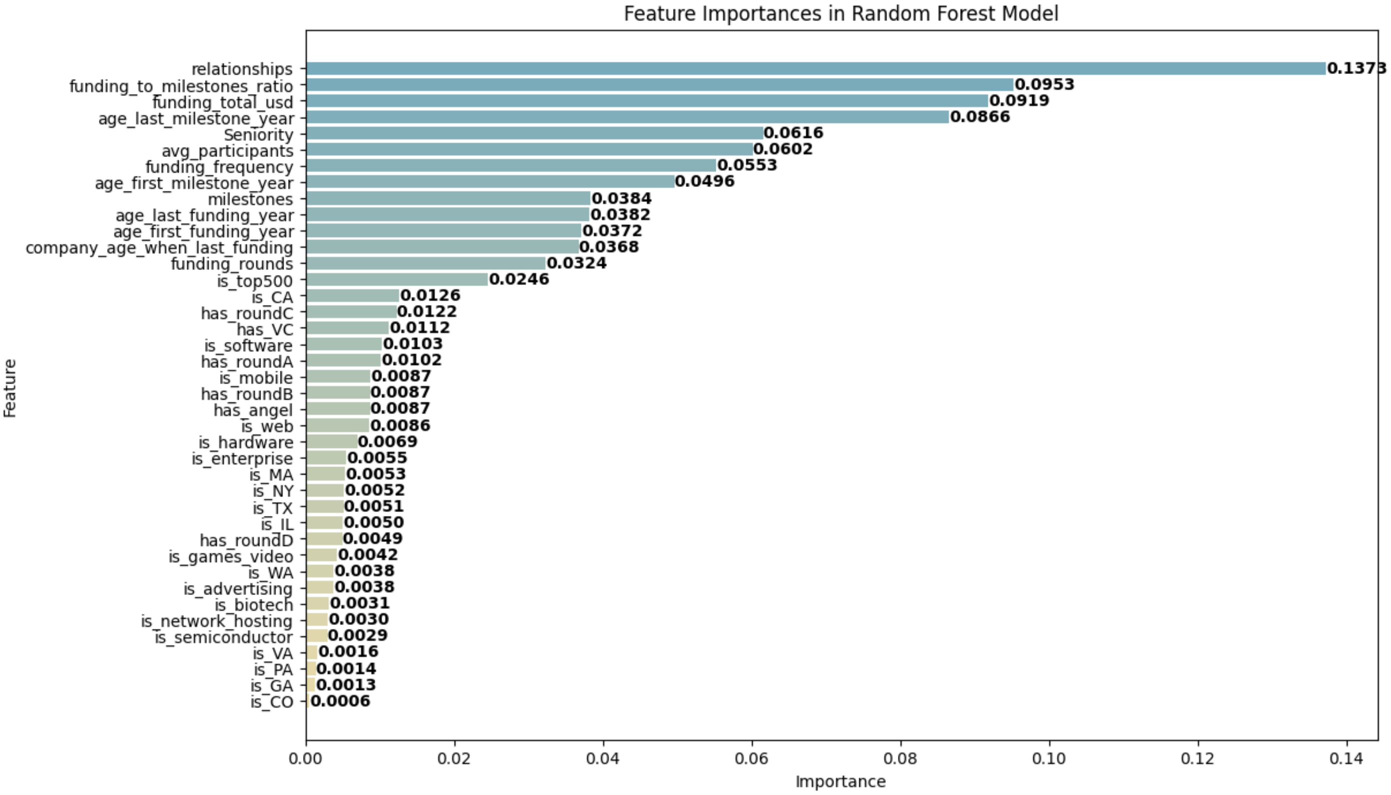
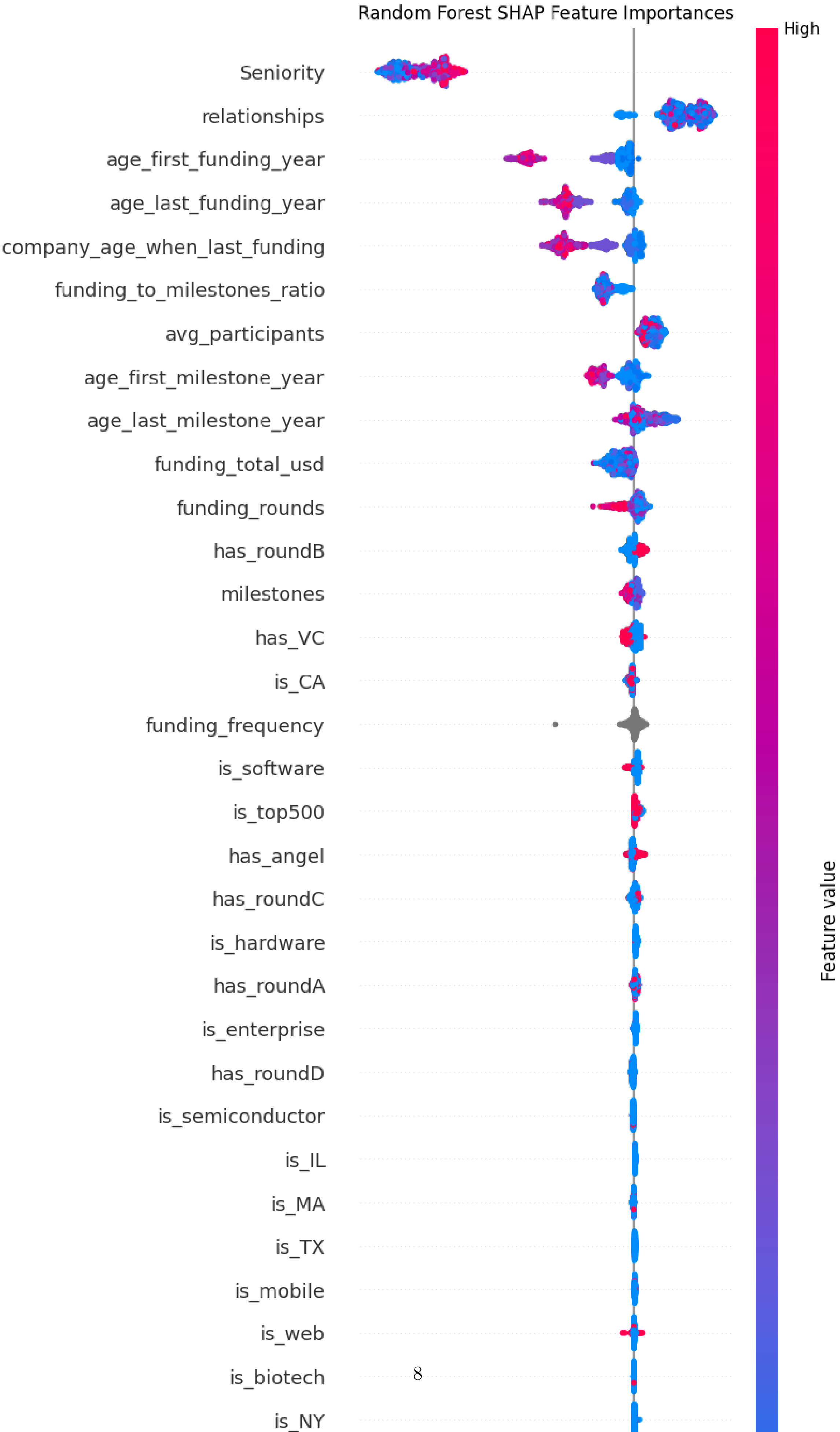


Figure 4: Feature Importance


Random Forest SHAP Feature Importances
Seniority
relationships
age_first_funding_year
age_last_funding_year
company_age_when_last_funding
funding_to_milestones_ratio
avg_participants
age_first_milestone_year
age_last_milestone_year
funding_total_usd
funding_rounds
has_roundB
milestones
has_VC
is_CA
funding_frequency
is_software
is_top500
has_angel
has_roundC
is_hardware
has_roundA
is_enterprise
has_roundD
is_semiconductor
is_IL
is_MA
is_TX
is_mobile
is_web
is_biotech
is_NY
High
Feature value

5 Discussion

The results underscored the importance of financial health and growth trajectories in determining startup success. Metrics like Funding to Milestones Ratio and Company Age at Last Funding Year were identified as significant predictors, highlighting the critical role of financial stability and strategic milestones in a startup's journey.

The SHAP values provided additional insights, emphasizing the impact of features such as ”relationships,” ”funding_to_milestones_ratio,” and ”funding_total_usd” on predictions. These findings suggest that maintaining strong financial metrics and cultivating strategic partnerships are crucial for startup longevity and success.

6 Future Work

Future research will focus on fine-tuning the Random Forest model through hyperparameter optimization to further enhance predictive accuracy. Advanced feature selection techniques will be explored to improve model interpretability and efficiency, providing clearer insights into the drivers of startup success.

Enhancing model interpretability using SHAP values will be pivotal in providing stakeholders with transparent explanations of predictions, aiding in strategic decision-making within startup ecosystems.

Addressing any class distribution imbalance and refining model evaluation techniques will ensure robust performance across diverse startup scenarios, ultimately advancing the predictive capabilities necessary for informed decision-making and sustainable growth.

References

Carlos Díaz-Santamaría and Jacques Bulchand-Gidumal ”Econometric Estimation of the Factors That Influence Startup Success” https://www.mdpi.com/2071-1050/13/4/2242

Paul X. McCarthy, Xian Gong, Fabian Braesemann, Fabian Stephany, Marian-Andrei Rizoiu Margaret L. Kern ”The impact of founder personalities on startup success” https://www.nature.com/articles/s41598-023-41980-y

Felipe Veloso, The University of North Carolina at Charlotte ”Predicting Startup Success in U.S” https://www.proquest.com/docview/2407618502?

Vinicius Trevisan ”Using SHAP Values to Explain How Your Machine Learning Model Works” https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

Sruthi E R ”Summary of Understand Random Forest Algorithms With Examples (Updated 2024)” https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Fabio Polcari ”Startup Success Prediction” https://www.kaggle.com/code/fpolcari/startup-success-prediction/script