MIT MANAGEMENT SLOAN SCHOOL

Predicting Startup Success

15.072: Advanced Analytics Edge

Master of Business Analytics

MIT Management Sloan School

Authors:

Anne-Castille Buisson
Sara Pasquino
Raghav Manoharan Jayanthi
Srikaran Reddy Boya

December 6, 2023

1 Problem Statement

In the dynamic landscape of business and innovation, the journey of startups is both thrilling and challenging. This project stands at the forefront, driven by the opportunity to help make critical investment decisions in the realm of startups that have received at least one round of funding. Using tools of data analytics and machine learning, the project's goal is to construct a predictive model that guides the assessment of the potential success of startups. The project's importance lies in the inherent complexity of startup trajectories. Understanding the effect of market trends, funding dynamics, team composition, and product-market alignment is vital for making informed decisions. By delving deep into the historical startup data, we aim to not only find these critical factors but also to determine whether a model with just these critical factors is better than a model than considers all the variable.

This model will be useful tool for investors, entrepreneurs, and accelerators navigating the ever-shifting terrain of startup ventures make better investment decisions. The ability to foresee challenges and opportunities can make all the difference, and this project attempts to provide the insights and foresight needed for astute decision-making in the unpredictable yet promising world of startups.

2 Data Exploration

The dataset was obtained from Kaggle [1]. This section covers the preliminary analysis of startup performance across multiple dimensions. Additional details are available in the Appendices (Section 7.1).

2.1 Early Funding as a Success Indicator

As seen in Figure 1, startups securing early funding demonstrate a higher likelihood of success. Emphasizing early financial support is key for our investment strategy.

2.2 Importance of Sustained Investment

Successful startups often receive later funding, highlighting the importance of continued investment (Figure 2). Considering additional funding rounds for promising startups can accelerate their growth.

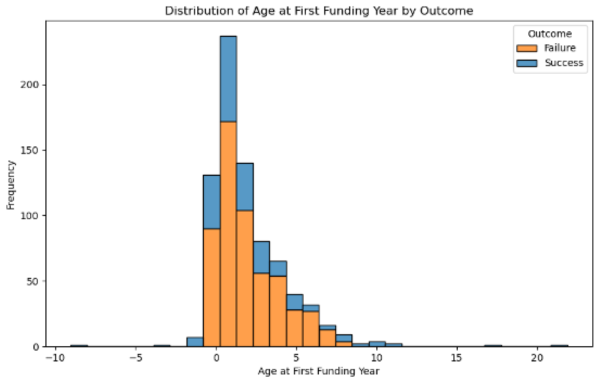

Figure 1: Age at First Funding By Outcome

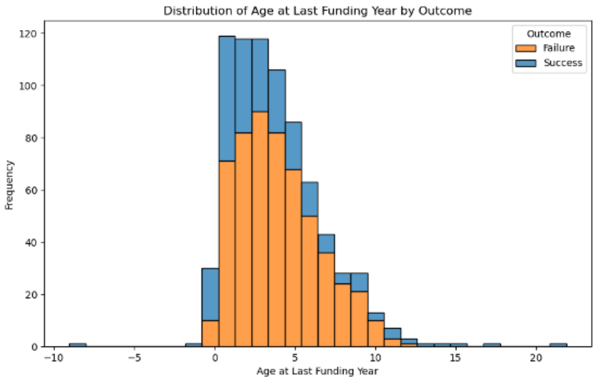

Figure 2: Age at Last funding by Outcome

2.3 Startup Density and Ecosystem Vibrancy

From Figure 3, we can observe that cities such as San Francisco and New York exhibit a high density of startups, indicating robust ecosystems that can offer networking and resource advantages. A presence in these hubs could be valuable for market penetration and talent acquisition. Nonetheless, the comparative success and failure rates highlight the varying risk profiles of different regions. This information could be pivotal for location-based risk assessment and investment decisions.

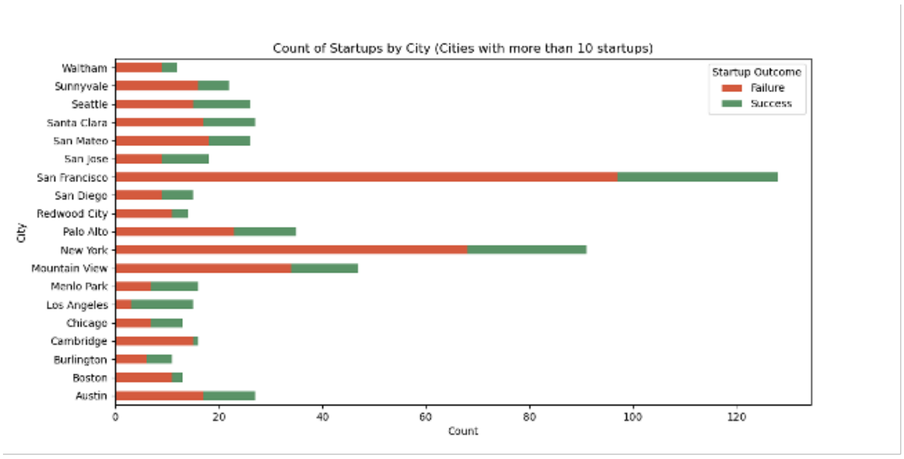

Figure 3: Count of Startups by City

2.4 Geographic Influence on Funding and Success

There is a stark gap between the list of cities with the highest number of startups, and that of cities with the highest average funding (see Figure 4 and Figure 5). Indeed, here regions like Loveland and Indianapolis stand out, potentially due to local economic conditions or investor activity. It may be advantageous to target these high-performing locations for future investments or business expansions. Nonetheless, in certain cities, the distribution is uneven, as indicated by the boxplot analysis. It is thus appropriate to remain cautious about average funding figures alone.

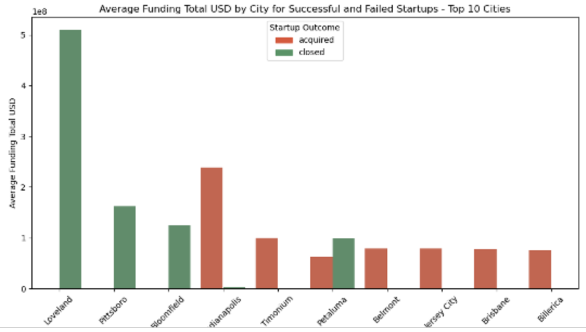

Figure 4: Average Funding by City and Outcome

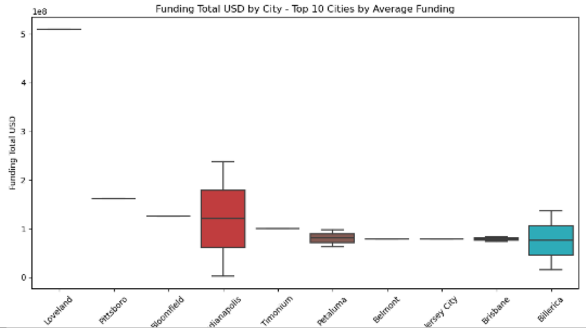

Figure 5: Average Funding by City Boxplots

2.5 Recommendations

Following this analysis, we can make several recommendations. Firstly, we advise supporting early-stage startups, which involves sustained financial backing for current investments. Moreover, it would be beneficial to prioritize investments in regions with a history of high performance, and to engage in strategic operations in startup-dense cities to effectively tap into the ecosystem benefits.

3 Methods

3.1 Data Preparation

In the data cleaning process, unnecessary columns were removed. The 'status' column was transformed into a binary format, distinguishing successful and failed startups, our predictor variable. One-hot encoding was then employed for the 'city' and 'zip_code' columns, effectively converting categorical data into a numerical format. Date-related columns ('founded_at', 'first_funding_at', and 'last_funding_at') were converted into datetime objects to facilitate the creation of new, insightful features such as the number of days between the founding and the first funding. Missing data was also dealt with by dropping rows that contained any NA values, ensuring data quality and consistency. This removed 16.47% of the data (152 rows).

3.2 Train, Validation and Test Splits

One challenge was the imbalanced nature of the training data, where approximately 75% of instances were labeled as 1 (successful startups) and the remaining 25% as 0 (unsuccessful startups). To address this imbalance and enhance model learning across both outcomes, a strategic down-sampling approach was implemented ,removing a substantial number of training samples labeled 1, resulting in a more balanced distribution with approximately 50.1% labeled 1 and 49.9% labeled 0 post down-sampling.

After addressing the imbalanced dataset, model was evaluated using iterative cross-validation procedures. This involved five distinct runs. Four-fifths of the data were utilized as training sets, and the remaining one-fifth served as the test set in each iteration. This methodology aimed to robustly assess model generalizability and ensure consistency in performance across various partitions of the dataset.

Furthermore, to foster a fair comparison of model performance, the same five datasets were consistently used to evaluate the three distinct models during cross-validation testing. This standardized approach enhances the reliability of conclusions drawn regarding the relative efficacy of the models, allowing for a deeper understanding of their performance characteristics. Accuracies and AUC values from cross validation are preferred over those obtained from running the model on a single train and test set because it decreases reliance on a single train-test split and enhances generalization.

3.3 Lasso Regression for Feature Selection

The first step of the analysis was that of identifying the most relevant features, as our initial dataset was significantly large (600+ features). Therefore, we developed a predictive model using sparse logistic regression with Lasso (L1) regularization to reduce the dataset dimension and identify key factors influencing startup success. Lasso regression assists by penalizing the absolute size of the coefficients, effectively reducing less important feature coefficients to zero, and producing more interpretable results.

Lasso regression selected 35 features out of the original 635, significantly reducing the complexity of the model. It can be observed that factors like age at first funding and age at milestones, as well as the operational industries of the startups, are significant features for our prediction. For here on, we decided to run each predictive model on both the original dataset, as well as on the Lasso-reduced one, to ensure accuracies where maintained whilst enhancing interpretability.

3.4 Predictive Models

In our analysis, we explored the performance of various classification models, including CART, Random Forest, and XGBoost, using full and reduced variable sets. Initially, we employed grid search techniques for hyperparameter tuning to identify the optimal model configuration. Upon finalizing the model, we conducted a comprehensive evaluation that included generating confusion matrices and ROC AUC curves, calculating AUC scores, and identifying the top 10 most significant features for each model. Furthermore, we performed extensive cross-validation, assessing each model's performance in terms of AUC, Precision for the positive class (Precision_1) and negative class (Precision_0), and Accuracy, to ensure robustness and reliability of our results.

4 Results

Table 1 presents the results of the models tested. Additional results including the Confusion Matrix, ROC Curve and Feature Importance Charts are included in the Appendices (Section 7.2).

Table 1: Model Performance Metrics

Model	Variables	AUC	Precision (1)	CV AUC	CV Precision (1)	CV Precision (0)
CART	All	0.572	0.57	0.670	0.81	0.56
Random Forest	All	0.697	0.67	0.749	0.78	0.77
XGBoost	All	0.655	0.56	0.843	0.78	0.74
CART	Lasso selected	0.654	0.63	0.702	0.82	0.63
Random Forest	Lasso selected	0.691	0.61	0.766	0.74	0.95
XGBoost	Lasso selected	0.658	0.56	0.817	0.76	0.78

5 Discussion

The Lasso-reduced models demonstrated comparable performance to the models incorporating all the variables. Consequently, for enhanced interpretability, we advocate for the adoption of simpler models. Among these, the XGBoost algorithm exhibited the most robust cross-validation AUC score of 0.817, a crucial metric for cases with class imbalance, albeit with a slight reduction in precision.

Keeping business impact in mind, the success rate of startups in the US is around 10%, but for startups with a single funding round, it ranges between 30-40% [2]. Consequently, random investment in a company with such a baseline precision falls significantly short of our targeted strategy, which offers a precision of 75%, ensuring that 75% of predicted "successes" are true positives.

Our ultimate model recommendation is contingent upon the specific client profile we are addressing:

- **High-Risk Investors:** These clients do not want to miss out on potentially successful startups. Therefore, we want to minimise the "false negatives", where the model predicts a startup will fail when it could potentially succeed. In essence, the objective is to increase Precision_0, making the Random Forest algorithm the ideal choice for achieving this goal.

- **Low-Risk Investors:** These clients are more risk-averse and prioritize the preservation of capital. They avoid investments that may result in financial losses. To minimise the "false positives", where the model predicts a startup will succeed (class 1) when it actually fails (class 0), we maximise Precision_1, a goal best achieved through the use of the CART model.

Across the models we found that the following features are significant in predicting startup success:

- **Average Number of Participants in Funding Rounds:** This metric reflects the engagement level of individuals involved in the startup's journey, encompassing founders, investors, and mentors or advisors who actively participate in its development and investment stages.

- **Total Funding in USD:** The cumulative amount of capital raised serves as a crucial indicator of the financial backing and resources available to the startup.

- **Time Since Last Milestone:** This tracks the duration since the startup's most recent significant achievement or development milestone, offering insights into its progress and momentum.

- **Is Top 500 Ranking:** This factor considers whether the startup is listed among the top 500 entities (based on revenue, growth, or other significant metrics). Being part of such a list can be indicative of the startup's standing and recognition in its field.

6 References

[1] https://www.kaggle.com/datasets/manishkc06/startup-success-prediction

[2] https://spdload.com/blog/startup-success-rate/

7 Appendices

7.1 Additional Preliminary Data Analysis

7.1.1 Optimal Participant Range in Funding Rounds

A correlation exists between the number of participants in funding rounds and startup success. This finding could provide insight into potential strategy regarding the number of partners to involve in funding activities. As a general trend, it seems that the more partners involved, the better performance, but further analysis is required in order to claim a causal effect. (see Figure 6)

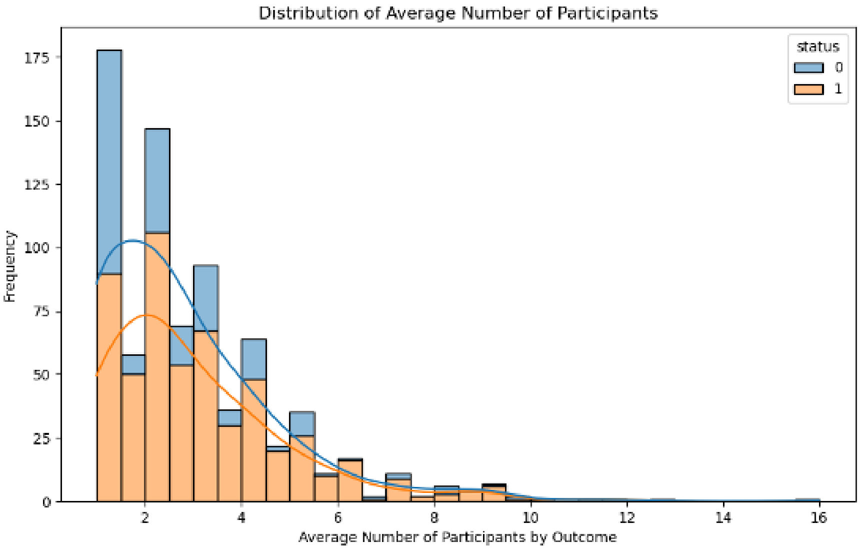

Figure 6: Average Number of Participants by Outcome

7.1.2 Founding Month Correlations

The analysis on seasonal trend of startup success rate displayed no such a trend in the likelihood of a startup succeeding. (see Figure 7)

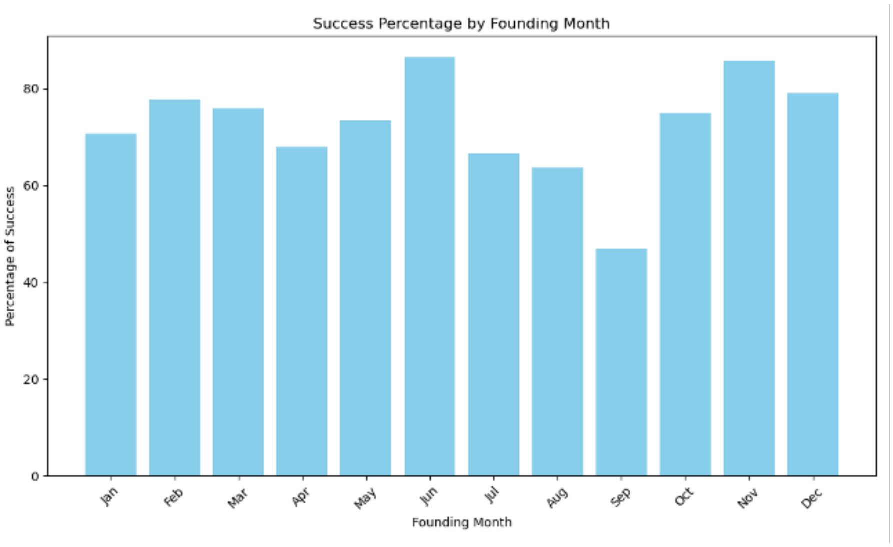

Figure 7: Success Rate by Funding Month

7.2 Additional Results

7.2.1 CART Model Utilizing All Variables

Table 2: Confusion Matrix for the CART Model Utilizing All Variables

	Actual 0	**Actual 1**
Predicted 0	28	17
Predicted 1	21	23

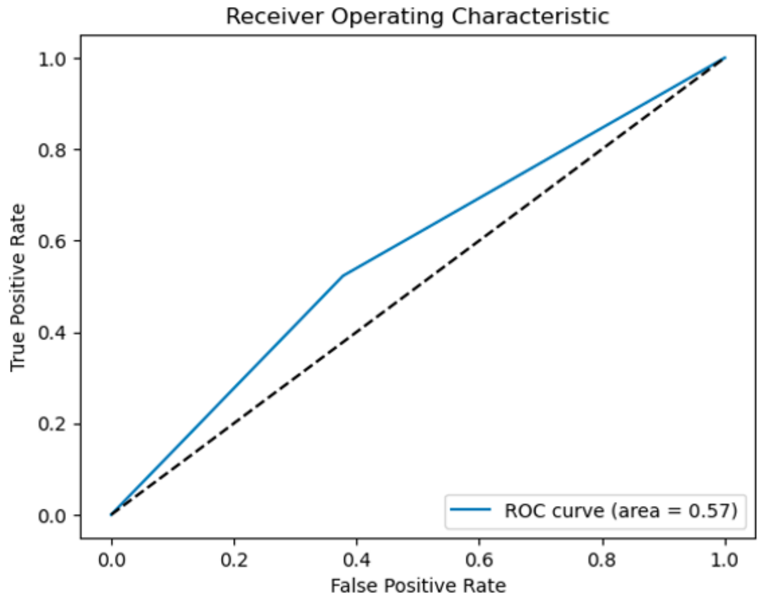

Figure 8: ROC Curve for the CART Model Utilizing All Variables

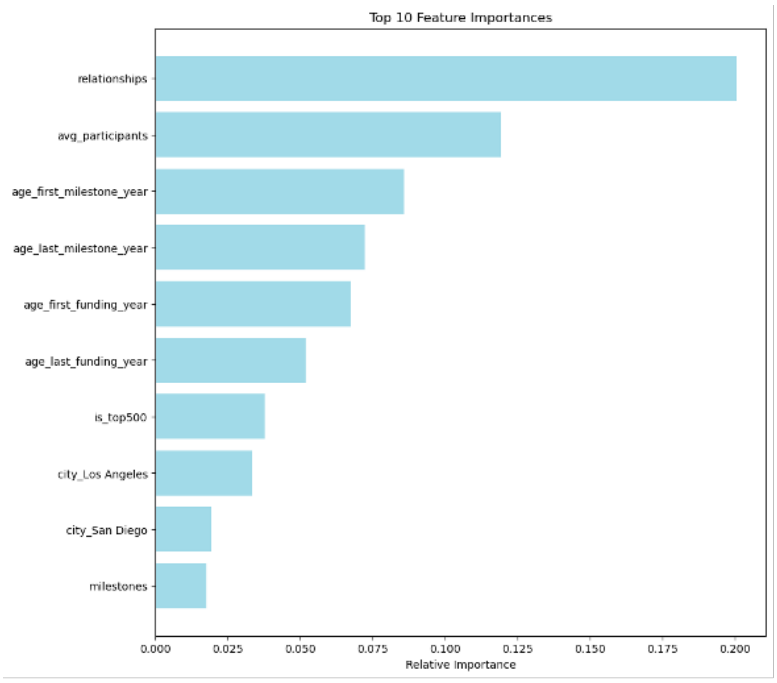

Figure 9: Feature Importance Chart for the CART Model Utilizing All Variables

7.2.2 CART Model with Lasso-Selected Variables

Table 3: Confusion Matrix for CART Model with Lasso-Selected Variables

	Actual 0	**Actual 1**
Predicted 0	28	17
Predicted 1	15	29

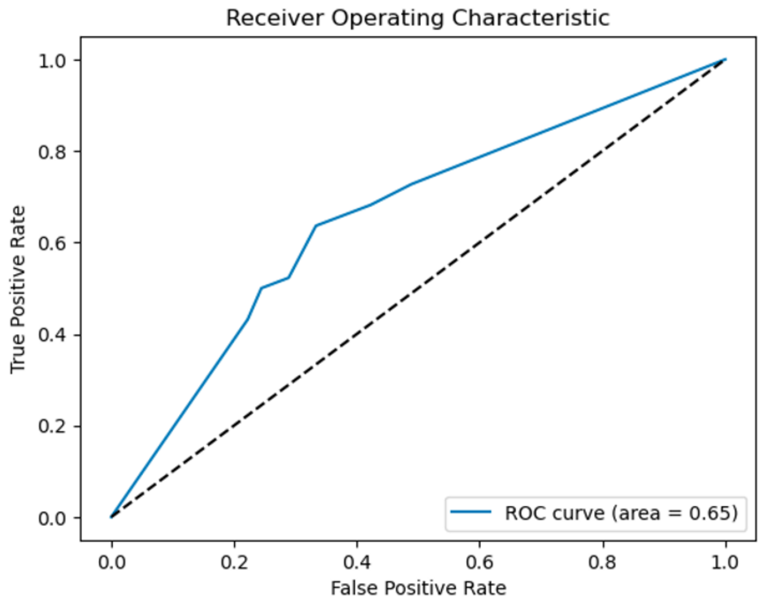

Figure 10: ROC Curve for CART Model with Lasso-Selected Variables

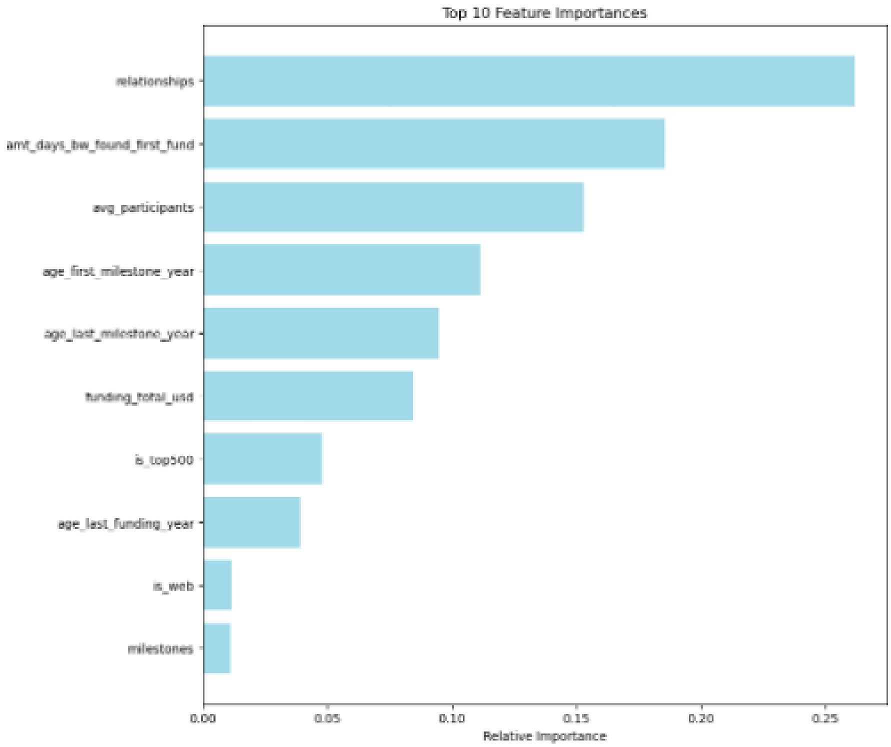

Figure 11: Feature Importance Chart for the CART Model with Lasso-Selected Variables

7.2.3 Random Forest Model Utilizing All Variables

Table 4: Confusion Matrix for Random Forest Utilizing All Variables

	Actual 0	**Actual 1**
Predicted 0	29	16
Predicted 1	12	32

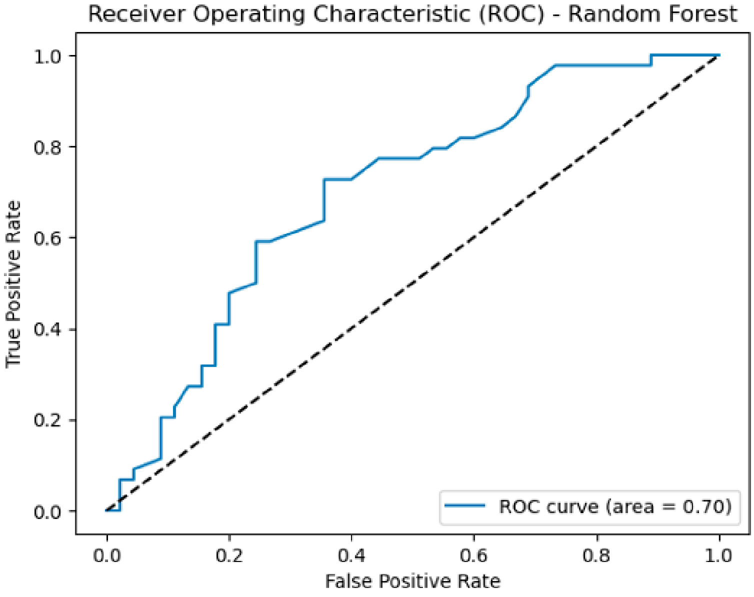

Figure 12: ROC Curve for Random Forest Model Utilizing All Variables

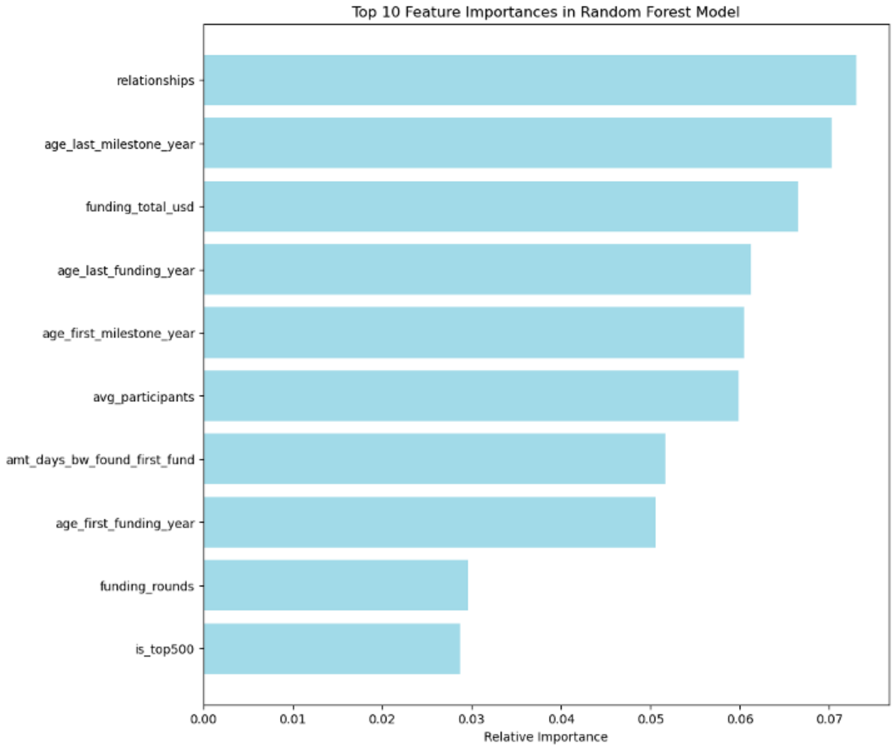

Figure 13: Feature Importance Chart for the Random Forest Model Utilizing All Variables

7.2.4 Random Forest Model with Lasso-Selected Variables

Table 5: Confusion Matrix for Random Forest with Lasso-Selected Variables

	Actual 0	**Actual 1**
Predicted 0	26	19
Predicted 1	14	30

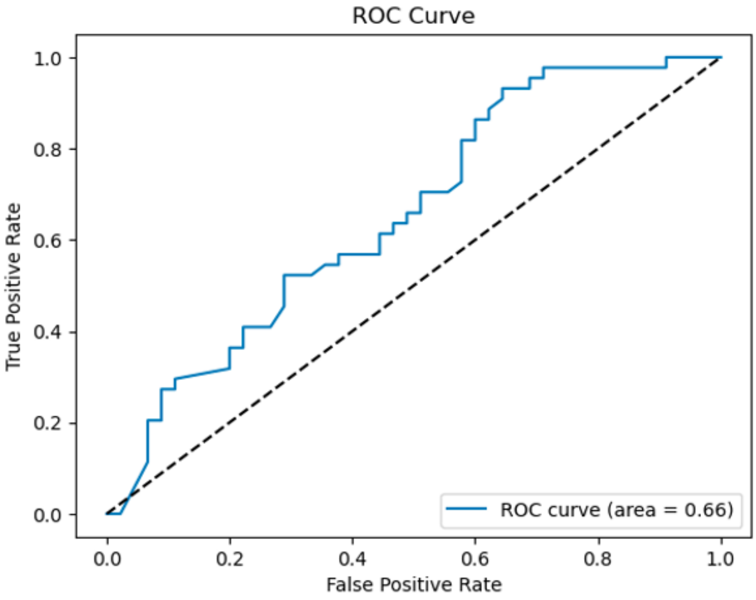

Figure 14: ROC Curve for Random Forest Model with Lasso-Selected Variables

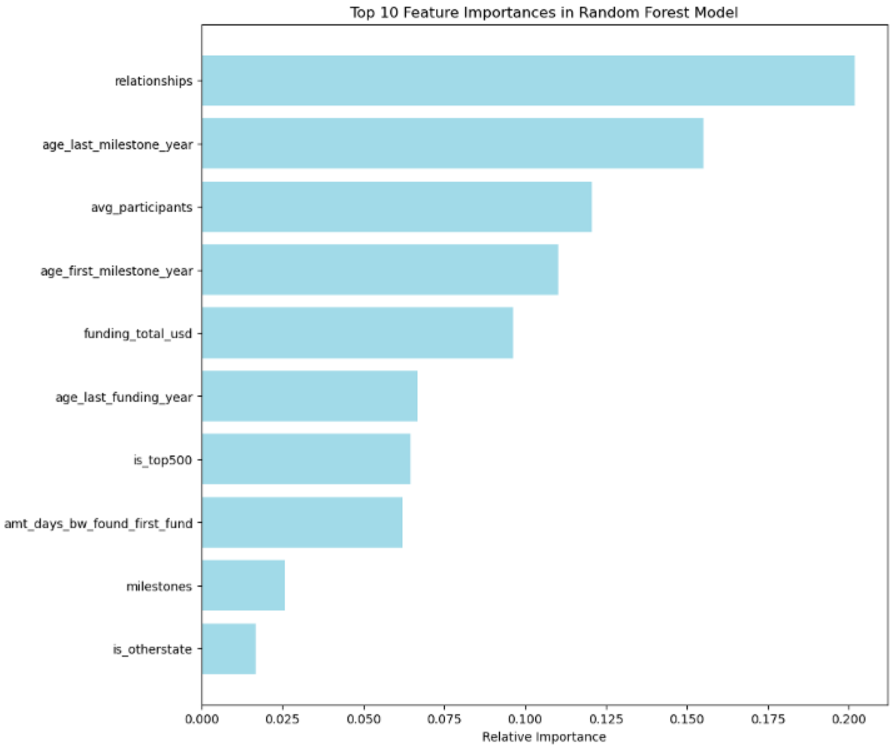

Figure 15: Feature Importance Chart for the Random Forest Model with Lasso-Selected Variables

7.2.5 XGBoost Model Utilizing All Variables

Table 6: Confusion Matrix for XGBoost Model Utilizing All Variables

	Actual 0	**Actual 1**
Predicted 0	22	23
Predicted 1	15	29

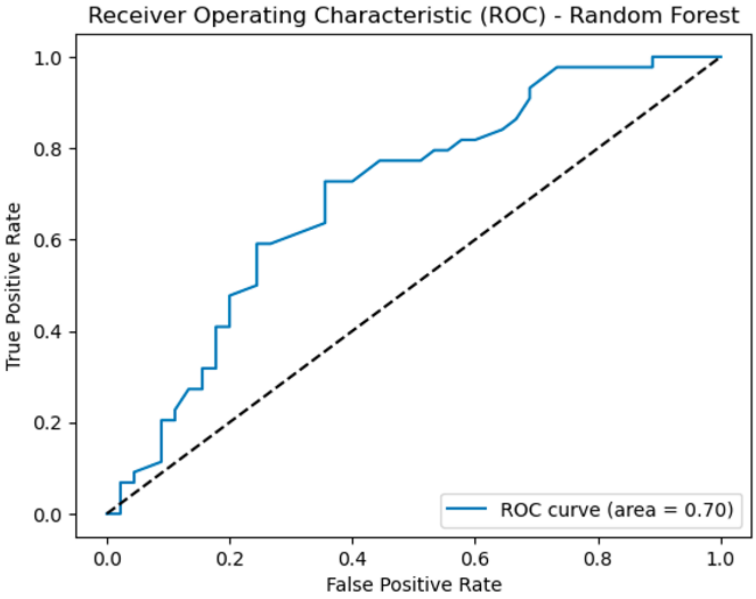

Figure 16: ROC Curve for XGBoost Model Utilizing All Variables

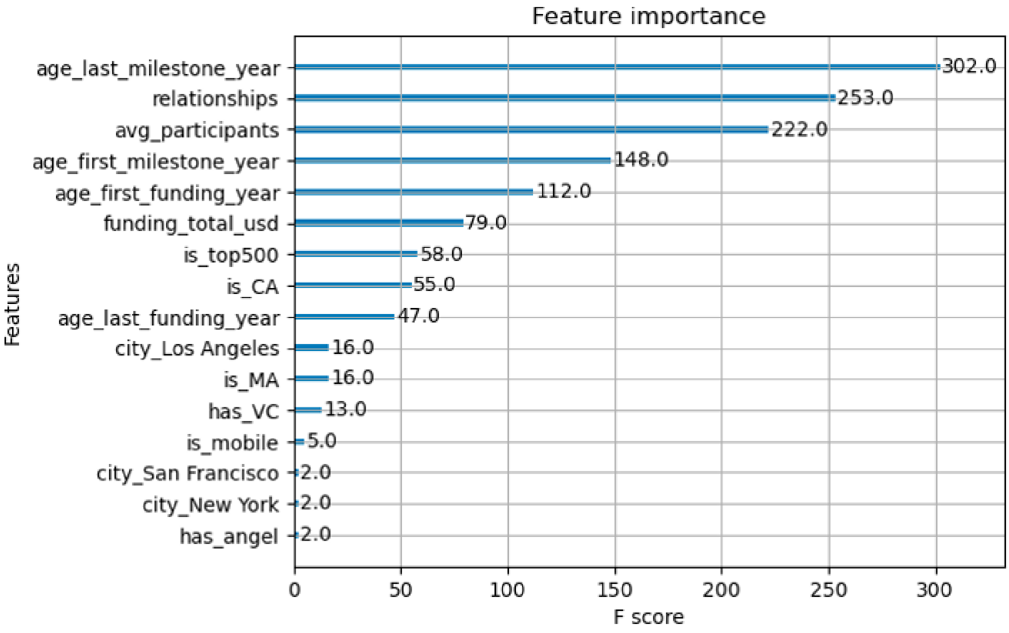

Figure 17: Feature Importance Chart for the XGBoost Model Utilizing All Variables

7.3 XGBoost Model with Lasso-Selected Variables

Table 7: Confusion Matrix for XGBoost Model with Lasso-Selected Variables

	Actual 0	**Actual 1**
Predicted 0	22	23
Predicted 1	15	29

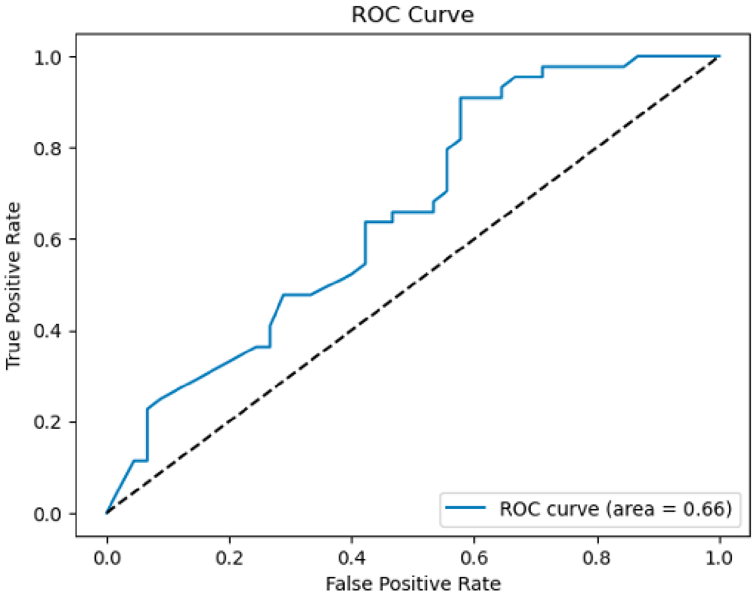

Figure 18: ROC Curve for XGBoost Model with Lasso-Selected Variables

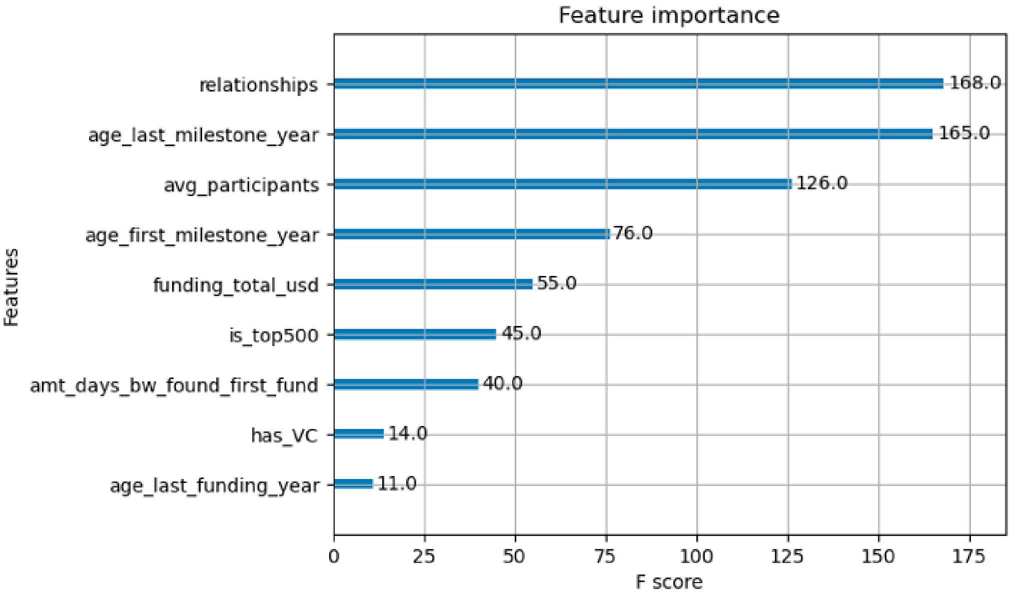

Figure 19: Feature Importance Chart for the XGBoost Model with Lasso-Selected Variables