Introduction

Shark Tank is a popular TV series where entrepreneurs pitch their business ideas to a panel of investors, known as "sharks," in the hope of securing funding. Entrepreneurs aim to convince the sharks to invest in their businesses while negotiating the valuation and funding amount. Often, entrepreneurs enter the negotiations with high valuations to maximize their funding potential. However, the sharks, relying on their assessment of the business's worth and growth potential, typically counter with lower valuations. This negotiation process leads to intense discussions as both sides strive to strike a deal that aligns with their goals and expectations.

As a fan of Shark Tank, I often find myself curious about the paths of the startups featured on the show. What fascinates me most is the decision-making process of the sharks and the aspects of a pitch that lead them to close a deal. Additionally, the dynamics among the sharks—how their interactions influence one another's decisions—are particularly intriguing. Unfortunately, this dataset does not include details on which shark made the offer or the conversations during negotiations.

A key area of interest is predicting whether an entrepreneur is likely to secure a deal based on attributes of their pitch, such as the funding amount requested, equity offered, and the business category. This report focuses on predicting the likelihood of securing a deal, identifying patterns in deal-making decisions, and providing insights into the factors that drive successful outcomes on Shark Tank.

Data Description

The dataset captures details about pitches on Shark Tank, focusing on deal outcomes, categories, valuations, and the involvement of the sharks. Each pitch is categorized into specific industries, with a cleaned-up category column for easier analysis. The deal outcome indicates whether the pitch resulted in an investment (True/False), and the valuation reflects the estimated worth of the company at the time of the pitch, based on the amount requested and the equity offered. Additionally, the dataset includes information on the sharks present during the pitch (Shark1 to Shark5), providing insight into their involvement and potential influence on the final decision. The dataset also includes information on whether the team consisted of one or multiple entrepreneurs. This file contains information from all episodes across six seasons of Shark Tank.

Data Preprocessing and Feauture Engineering

To prepare the data for analysis, multiple preprocessing steps were undertaken:

1. After looking at distribution of all columns, we noticed that "category" and "location" were highly scattered.
    - Category Generalization: Categories were grouped based on conceptual similarity. For example, all categories related to babies were merged into one. Notably, pitches in categories like *Food and Beverages* and *Men and Women Clothes* were the most frequent (Appendix 1 and 2).

- Location Simplification: For location, the city (last two letters) was extracted, and cities that appeared in fewer than 2% of the data points were excluded. Most startups originated from California, New York, Texas, and Florida.

2. The order of sharks in the dataset had no significance; only their presence mattered. Instead of creating separate dummy variables for each shark's position (e.g., “Shark1_Barbara”, “Shark2_Barbara”), 11 dummy variables were created—one for each shark (e.g., “shark_Barbara_Corcoran”). This eliminated redundancy while preserving the necessary information.
3. Monetary columns like “valuation” and “askedFor” were highly skewed, with many small values and a few large outliers. To address this, after removing outliers, a log transformation was applied to normalize their distributions, ensuring that the data was more evenly spread and suitable for analysis.

These preprocessing steps ensured that the dataset was cleaned, simplified, and ready for further modeling and analysis.

Exploratory Descriptive Analysis

Based on our research questions, we investigated whether the probability of securing a deal varies depending on factors such as category, location, valuation, and other relevant variables.

Location

The success rate in different cities vary from 80% (in New Jersey) to 33% in Colorado and North Carolina. California as a traditional startup hub stands in the middle, probably because it has the greatest number of pitches (141) followed by other (99) so the success rate is lower.

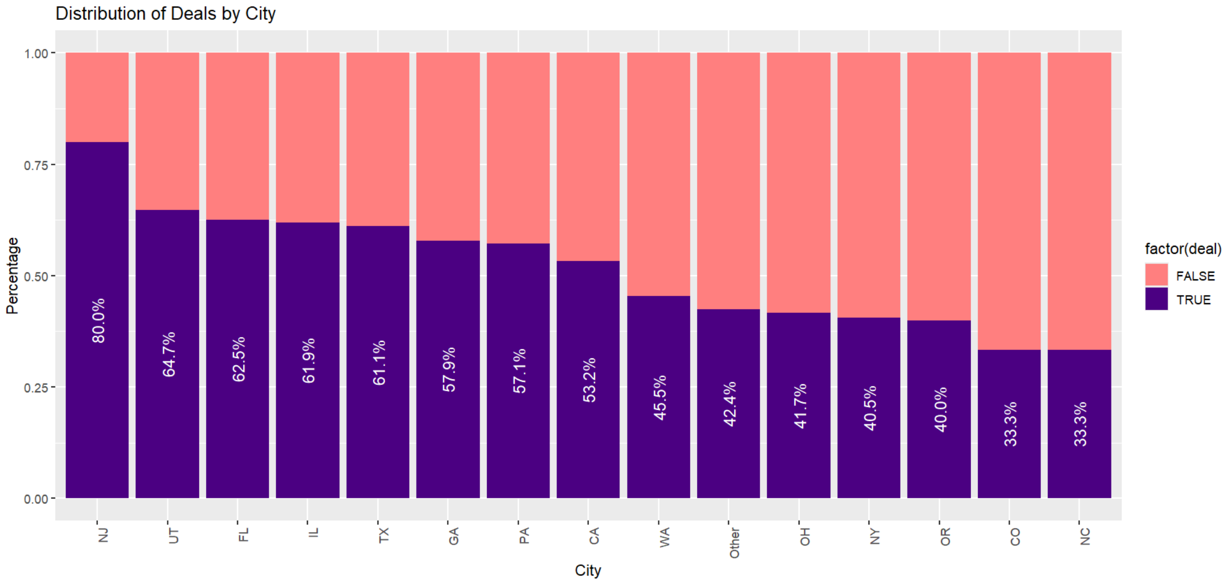


Figure 1. Success rates of deals across cities

Category

Below, we can see the success rate in different categories. Education and holiday Cheer have the highest success rate but they contain less than 10 pitches. The best performing categories with enough data in terms of ratio of closed deals are Home and Garden and Kitchen Tools.

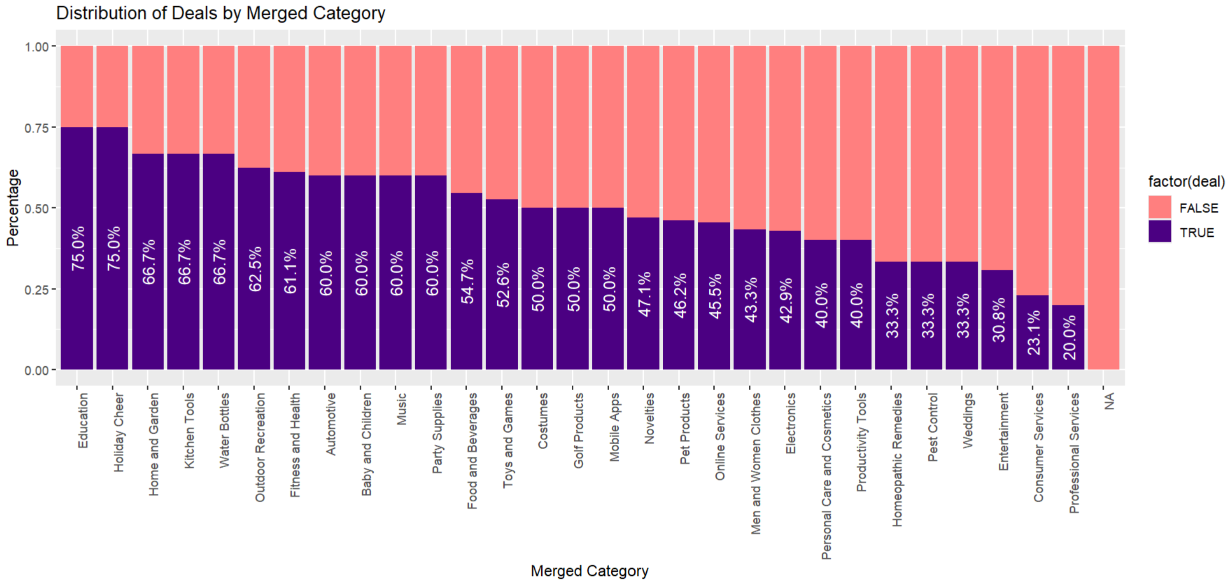


Figure 2. Success rates of deals across categories

Features related to Funding

Additionally, we examined the relationship between the outcome of the pitch (deal or no deal) and the financial funding metrics, including the amount of money requested, the percentage of equity exchanged, and the company's valuation.

The violin plot (Appendix 5) highlights that the majority of deals were closed when entrepreneurs offered around 10% of their equity. Conversely, pitch rejections were most frequent when entrepreneurs offered approximately 20% of their equity. However, this observation is influenced by a bias: a larger number of entrepreneurs offered equity in the 15% to 25% range. To address this, we will now analyze the success rates to provide a clearer understanding of the relationship between equity offered and the likelihood of securing a deal.

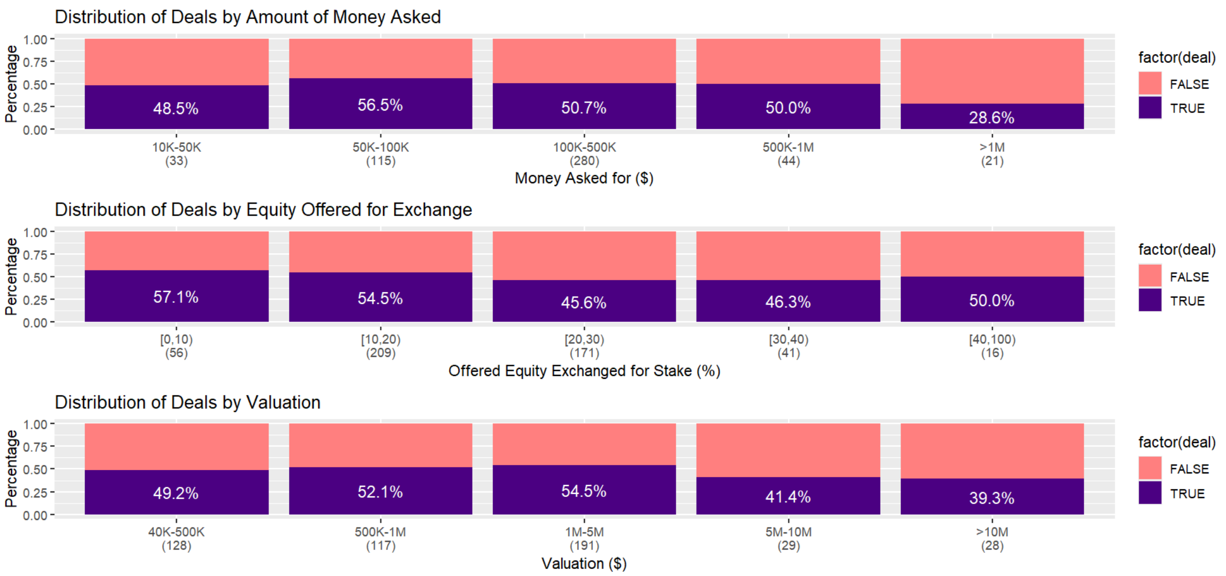


Figure 3. Success rates of deals based on funding metrics (valuation, money asked for and equity exchanged)

We see a relationship between success rate (% of pitches that resulted in a deal) with ranged of valuation and the amount of money asked for. If the entrepreneur asked for money in the range of 50,000 USD up to 1 million USD they had a success rate of 50% and above. Moreover, highest percentage of deals happen when valuation is between 1 million and 5 million and for startups with valuation above the percentage of deals considerably drop.

Lastly, we examined whether there was a trend in success rates across the seasons (Appendix 8). While the graph suggests a slight upward trend, the differences are not substantial.

Methodology and Results

The goal of our analysis is to identify the factors that influence the likelihood of securing a deal and leverage these insights to inform entrepreneurial strategies. Building on the ideas derived from our exploratory data analysis (EDA), we proceed with developing predictive models and conducting deeper analyses to gain actionable insights.

By focusing more specifically on funding strategies, we conducted a deeper analysis to understand how three key features—'Amount of Money Asked For' and 'Offered Equity in Exchange' (And valuation which is derived by these two)—affect the deal outcome. To analyze these features, we used Linear Discriminant Analysis (LDA) to classify deal outcomes based on these variables. One limitation of LDA and QDA is that they assume the features are normally distributed. While this assumption holds reasonably well for Offered Equity in Exchange, it does not hold for 'Amount of Money Asked For' and 'Valuation'. So, we used log transformation to approximate a normal distribution But the partition graphs did not show any pretty classification (Appendix 6 is an example). The classification error was 45.8% and clearly there is not a way to separate True and False deals by drawing lines between them (So no need for QDA results.)

Decision Tree

We grew a decision tree based on having a website and funding metrics. We tested different combinations of feautures but the best one was this with optimal cp=0.01.

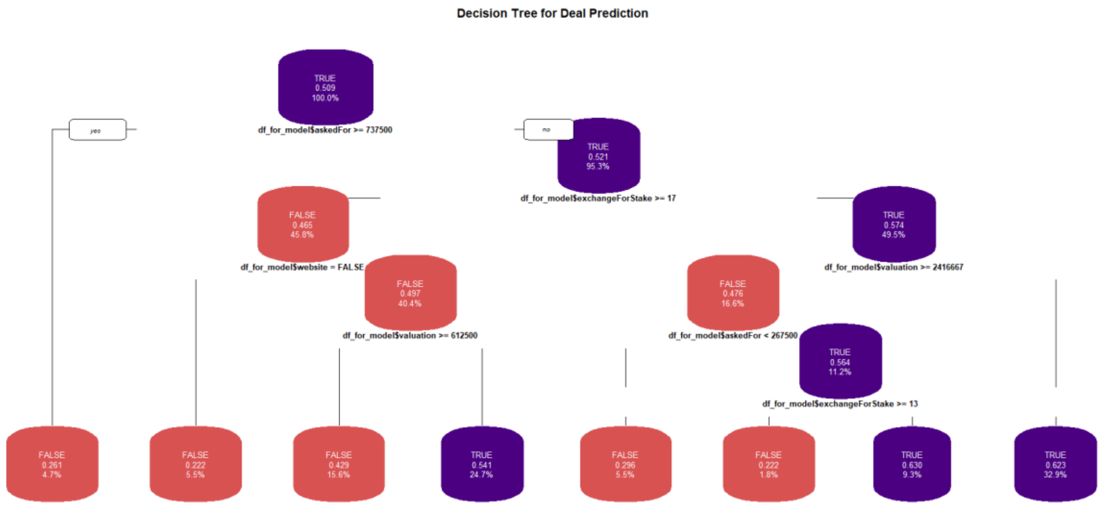

Decision Tree for Deal Prediction
TRUE
0.509
100.0%
df_for_model$askedFor >= 737500
yes
no
TRUE
0.521
95.3%
df_for_model$exchangeForStake >= 17
FALSE
0.465
45.8%
df_for_model$website = FALSE
TRUE
0.574
49.5%
df_for_model$valuation >= 2416667
FALSE
0.497
40.4%
df_for_model$valuation >= 612500
FALSE
0.476
16.6%
df_for_model$askedFor < 267500
TRUE
0.564
11.2%
df_for_model$exchangeForStake >= 13
FALSE
0.261
4.7%
FALSE
0.222
5.5%
FALSE
0.429
15.6%
TRUE
0.541
24.7%
FALSE
0.296
5.5%
FALSE
0.222
1.8%
TRUE
0.630
9.3%
TRUE
0.623
32.9%

Figure 5. Decision tree for interpreting shark tank deals

Observed insights from the tree were:

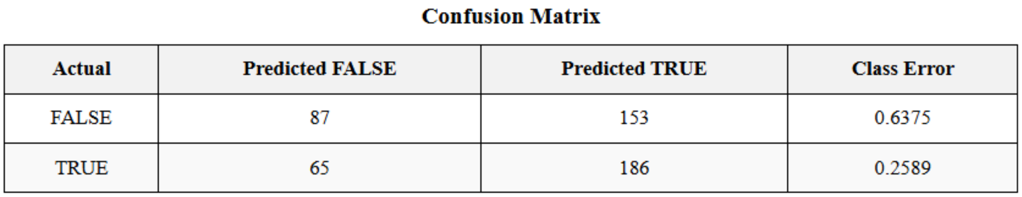
- Entrepreneurs asking for higher than $737,500 have a low chance of securing a deal. On the other hand, at lower level of money asked, if the business have a website and a reasonable valuation the chance of having a deal is high.
- When the equity offered is less than 17%, having a website improves the probability of securing a deal significantly. (In other branches, the effect was insignificant.)
- If entrepreneurs offer ≥ 17% equity while they have a valuation of ≥ $2.4M, the likelihood of securing a deal increases to 62.3%.
- Most of the data ends up at two nodes which predict securing a deal. First Node from right contains 32.9% of the datapoints and fifth node has 24.7% of the datapoints.

Random Forest Classifier

We try to predict outcome of a deal by utilizing tree-based classification algorithms to predict the deal column. Tree-based classifiers, such as the Random Forest classifier and Gradient Boosting, are effective due to the fact that they handle categorical variables well and capture non-linear relationships in the data. Furthermore, these methods are interpretable, providing insights into feature importance.

Since we have a small dataset here, it is crucial to use as much data as possible for training while still evaluating the model's generalization. So, we decided to not split the data in test train and do cross validation instead. A single train-test split can lead to variability in performance metrics depending on how the data is split. Cross-validation provides a more robust estimate of model performance by averaging across multiple splits.

For the Random Forest function, first we went with most feautures we had, and this is the confusion matrix. It shows us the model is performing True values better. We also should have this in mind that the dataset is quite balanced in terms of number of True and False.

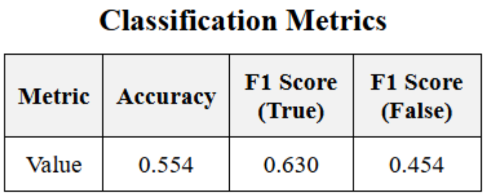


Confusion Matrix

Actual	Predicted FALSE	Predicted TRUE	Class Error
FALSE	87	153	0.6375
TRUE	65	186	0.2589

Table 1. Confusion matrix for deal success prediction

The accuracy we are getting before doing cross validation is 55.4%. The F1 score, a harmonic mean of precision and recall, provides a more balanced evaluation by considering both false positives and false negatives. In this case, the F1 score for predicting True values is significantly better than for predicting False values.



Classification Metrics

Metric	Accuracy	F1 Score (True)	F1 Score (False)
Value	0.554	0.630	0.454

Table 2. Performance metrics for deal success prediction

We drew a graph of the power of prediction of different feautures (Appendix 7). The graph showed us how much each predictor has an effect in decreasing accuracy if it is randomly permuted while other feautures stay the same. Features with a higher Mean Decrease in Accuracy like equity offered for exchange are more important because their absence (or randomization) has a larger negative effect on the model's predictive performance. As a result, we decided to remove the features that had a negative mean decrease in accuracy, as these features do not contribute positively to the model's predictive performance.

Next, we ran a Cross validation using random forest after deleting those feautures. For the hyper tuning we checked different level s of mtry (Number of variables randomly selected for splitting at each node.) and 27 was slightly better. With this mtry we got accuracy of %55.4 which is not much better than guessing randomly. Below, you can see the top 20 important feautures of Random Forest with Cross validation.

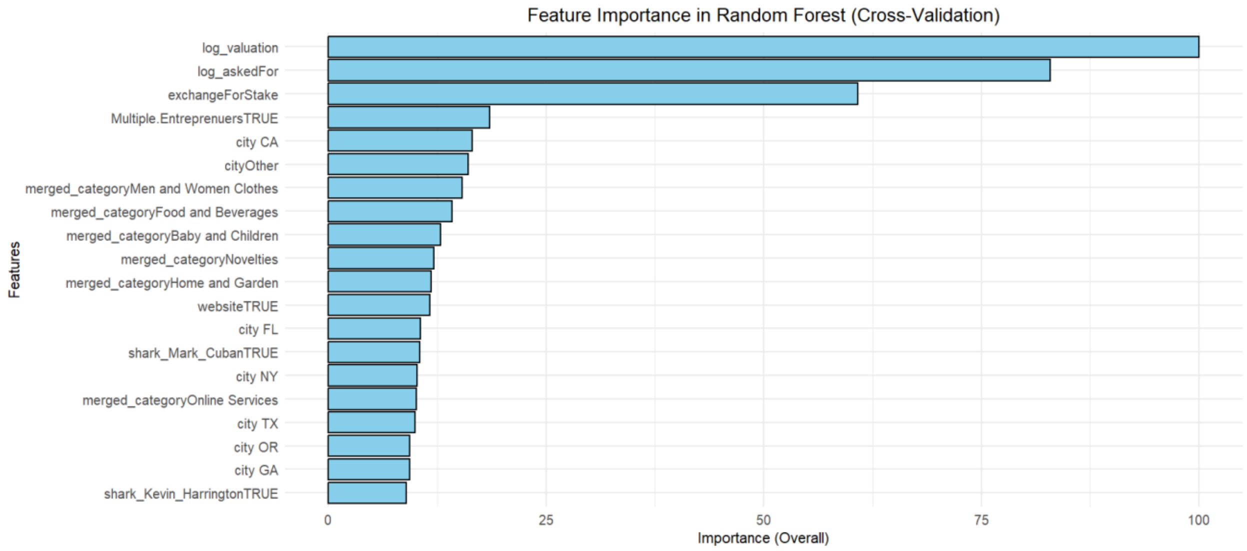


Figure 6. Feauture importance in random forest classifier

From the chart, we can conclude that features related to funding strategy—such as log(Valuation), log(Money Asked For), and Offered Equity for Exchange—are the most influential predictors in the model. Additionally, the model was better able to predict outcomes for businesses with multiple entrepreneurs, those based in California, or those categorized under Men and Women's Clothing, Food and Beverages, or Baby and Children's Products.

Dimension Reduction by Principal Discriminative Analysis

To understand how the deal outcome is related to the presence of each shark or their combinations, we applied PCA to reduce the dimensionality of the shark-related features. With 11 sharks and 9 unique combinations of them, PCA allowed us to visualize these relationships effectively. In the PCA charts below, the size of the circles represents the number of pitches that had those specific combinations of shark characteristics.

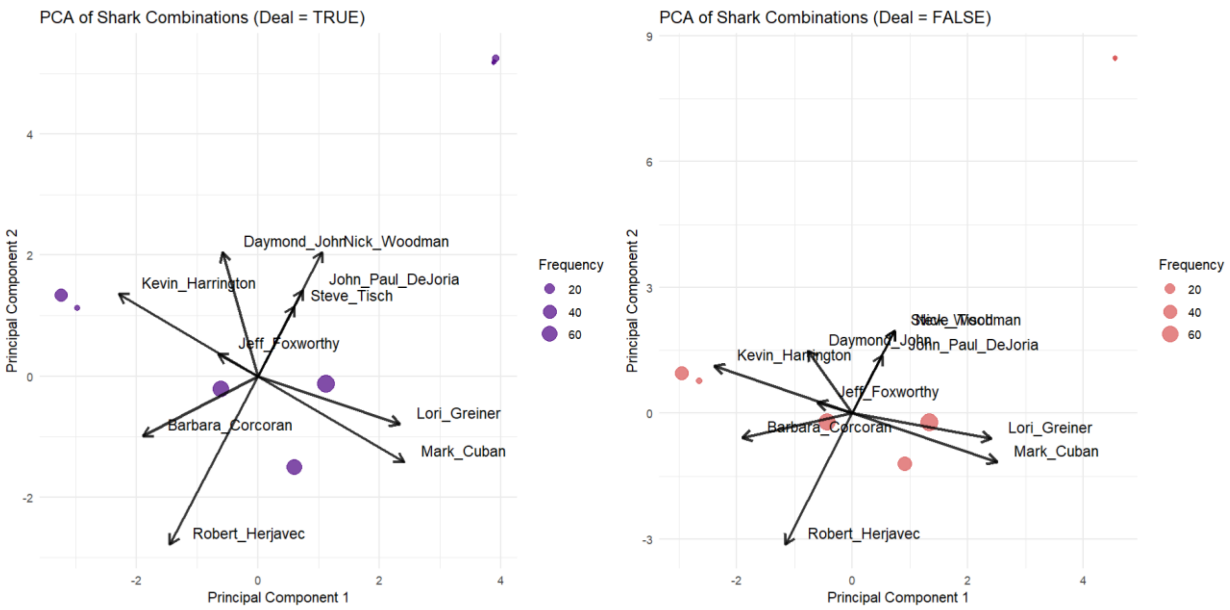


Figure 7. PCA plot of deal outcomes by shark presence

- Sharks Play a Greater Role in Successful Deals: In the "Deal = True" chart, arrows are generally longer, indicating sharks have a more significant role in explaining variance for successful deals compared to unsuccessful ones.

- Prominent Sharks in Deals: Sharks like Mark Cuban, Kevin Harrington, and Robert Herjavec contribute more strongly to successful deals, as shown by their longer arrows.

- Frequent Combinations: Sharks Nick Woodman, John Paul DeJoria, Daymond John, and Steve Tisch are closely aligned, suggesting they frequently appeared together in pitches.

Conclusion

This analysis looked at the factors that influence whether entrepreneurs secure a deal on Shark Tank. While we gathered useful insights into funding strategies, categories, and sharks, the performance of the predictive models was not very strong. This suggests that the data we had was not enough to fully explain what drives deal success. We could have more interesting analysis if the dataset had information investor feedback, and negotiation details.

Although gradient boosting was also explored, its performance was not significantly better. The best prediction accuracy achieved (55.4%) came from the random forest classifier after feature selection.

Below are the key insights on the success of a deal, aggregated from our EDA, decision tree analysis, random forest feature importance, and PCA analysis on the sharks:

1. What Works in Securing Deals:

- Reasonable Funding Requests: Asking for $50,000–$1 million and offering equity in the 10%-17% range increases success chances. Valuations of $1 million–$5 million are most favorable.
- Having a Website: Businesses with an online presence, especially those offering less equity, are more likely to succeed.
- Industry Trends: Categories like Food and Beverages, Men and Women's Clothing, and Baby Products tend to perform better.
- Team Advantage: Pitches with multiple entrepreneurs are often more successful, likely due to perceived teamwork and expertise.

2. Sharks' Role:

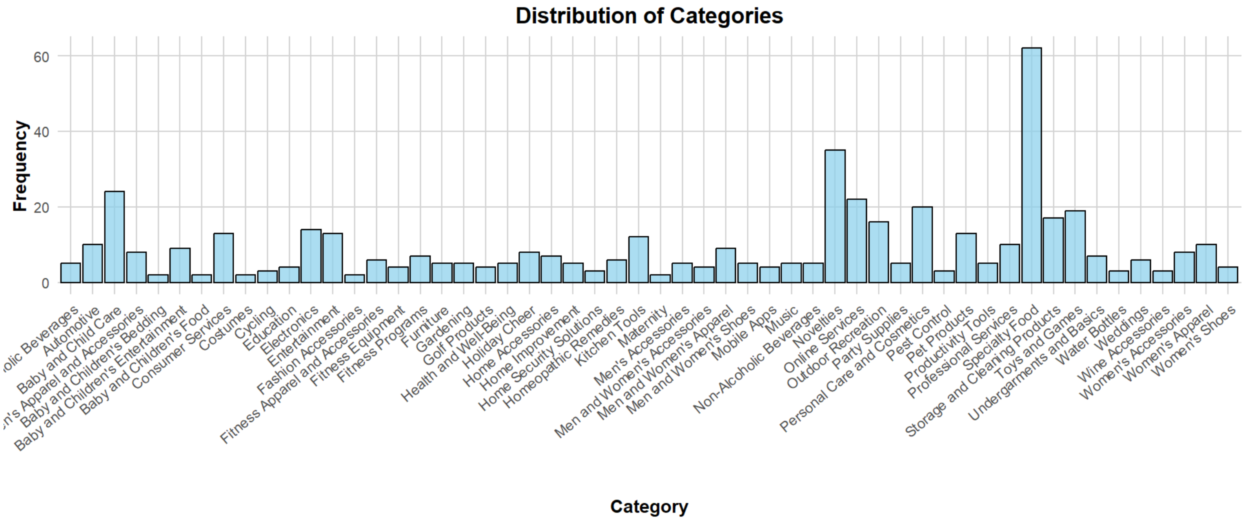
- Sharks like Mark Cuban and Kevin Harrington have a stronger influence on successful deals. Some sharks also appear to work better as a team in securing investments.

References

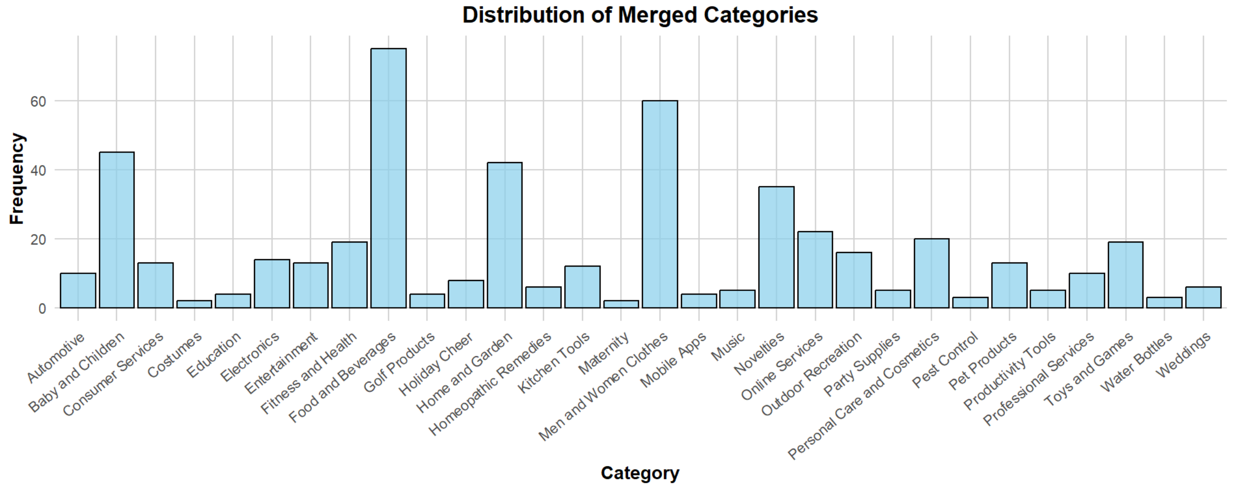
https://www.investopedia.com/articles/company-insights/092116/how-business-valued-shark-tank.asp

https://www.stat.cmu.edu/capstoneresearch/spring2023/315files_s23/team16.html

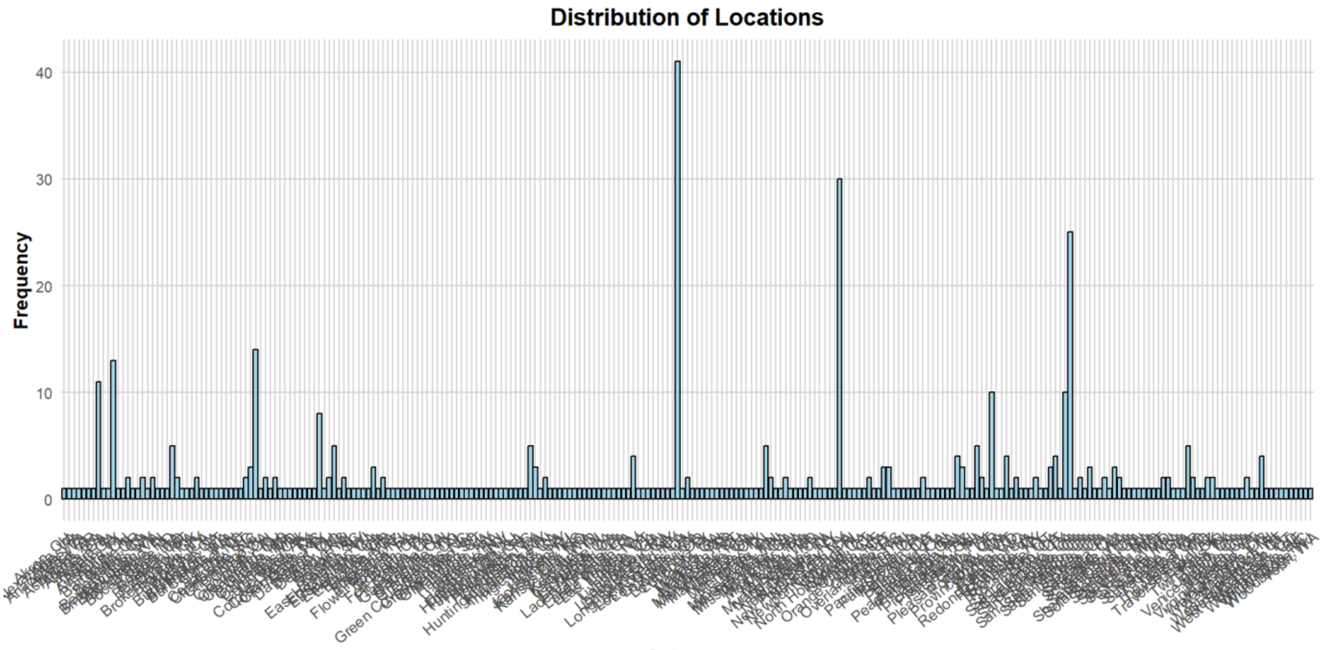
Appendix



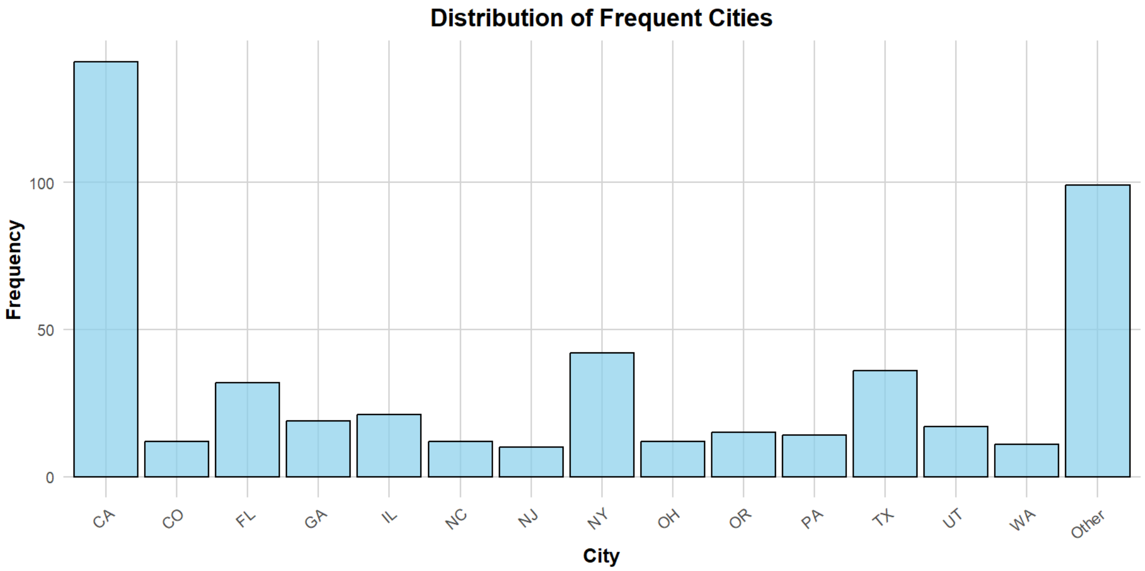
A.1. Distribution of category column before generalizing



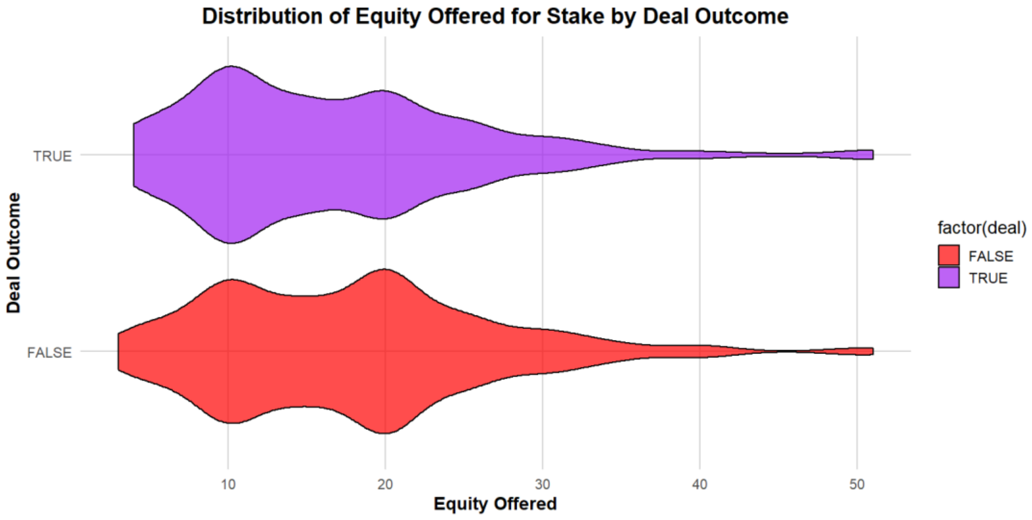
A.2 Distribution of merged category after generalizing



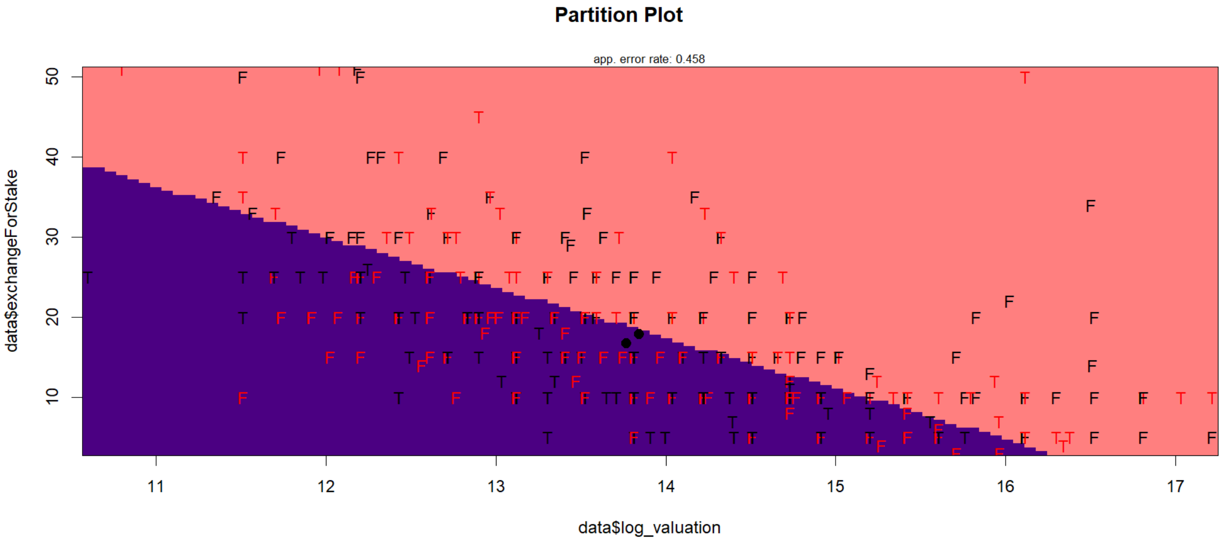
A.3 Distribution of location (no information)



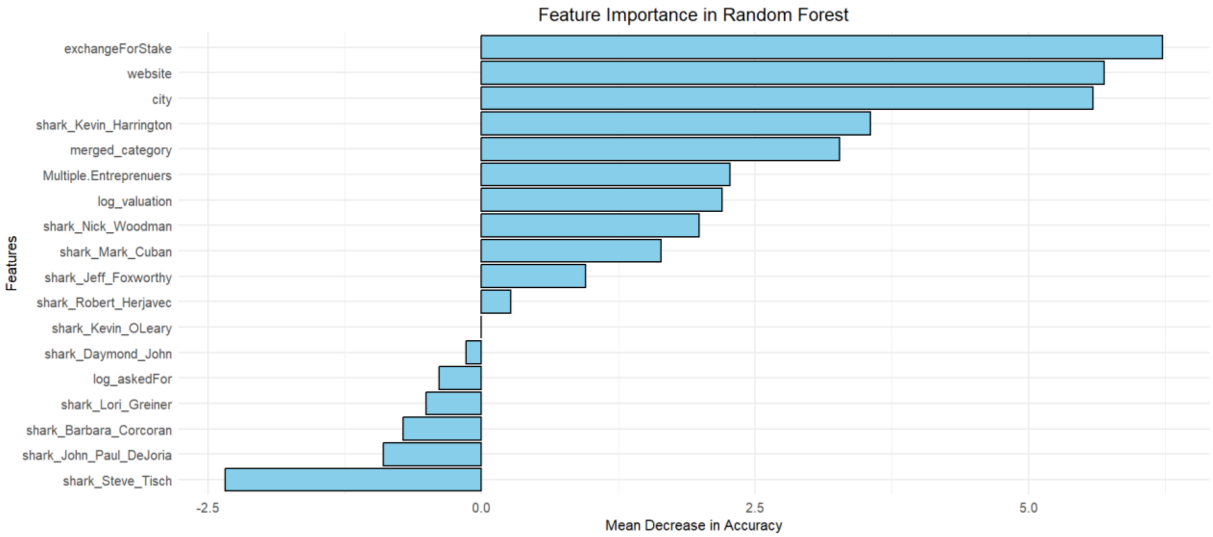
A.4 Distribution of more frequent cities



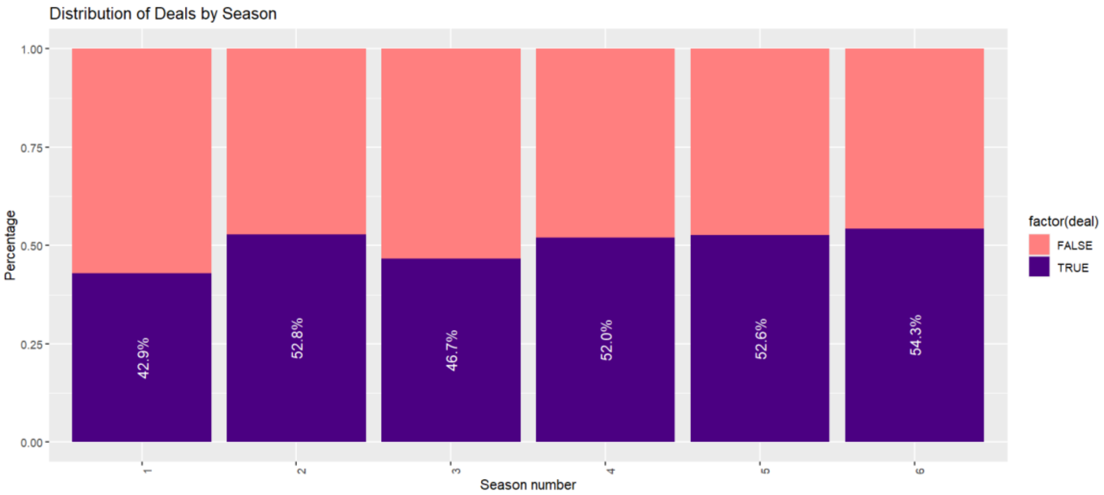
A.5 Violon graph of percentage of equity offered based on deal outcome



A.6 Partition plot of deal outcome based on log(Valuation) and percentage of equity offered



A.7. Feature importance of Random Forest predicting deal outcome



A.8. Success rates of deals through six seasons of the show