Final Project: Shark Tank Dataset

Presented to
Professor Juan Serpa

By
Palummo, Julien - 260946408

MGSC 661 Multivariate Statistics - Section 076

McGill University - Desautels Faculty of Management
December 10th 2023

Table of Contents

1. Introduction

1. 1 Project Summary

In 2023, the "Shark Tank" dataset, a rich compilation of data from the popular TV show "Shark Tank," was analyzed to predict the likelihood of entrepreneurs securing deals from the Sharks. This dataset contains detailed information on 495 pitches made to the Sharks, encompassing aspects like the amount asked for, equity stakes offered, valuations, categories of products, and the involvement of multiple entrepreneurs. With diverse data ranging from categorical variables like the presence of specific Sharks and product categories to numerical ones such as episode numbers, valuations, and amounts requested, this dataset provides a comprehensive view of the factors influencing deal-making on the show.

1.2 Project Goals

The project begins with a thorough exploration of the "Shark Tank" dataset. We start by understanding the various aspects of the data, such as deal outcomes, entrepreneur details, and the Sharks' decisions. This involves analyzing each variable to comprehend their distributions, identifying anomalies, and understanding patterns within the data. This foundational analysis is crucial for setting the stage for more complex modeling and predictions.

Our next step delves into the development of predictive models. The first model we build is a Random Forest model, aiming to predict the likelihood of entrepreneurs securing deals based on a multitude of factors present in the dataset. This model considers various aspects, such as the amount of money asked for, equity stakes, company valuations, and which Sharks are present. By analyzing these factors, the model learns to identify key patterns that influence the Sharks' investment decisions.

In parallel, we worked on a second model focused on providing strategic advice to entrepreneurs. This model aims to guide entrepreneurs on which Sharks to target and the most opportune time within the season to make their pitches. It considers factors such as the product category, whether the entrepreneurs are solo or part of a team, and their geographic location. This model uses the insights gained from the Random Forest model and augments them with clustering techniques to analyze trends across different episodes and seasons.

Finally, we bring these models together to calculate the probability of deal success for an entrepreneur's pitch. This involves integrating the predictive power of the Random Forest model with the strategic insights from the clustering analysis. The result is a comprehensive probability score that provides entrepreneurs with a tangible metric to assess their chances of success in "Shark Tank."

Throughout the project, we emphasize the validation and testing of our models. This ensures that the models are not only accurate in predicting outcomes based on historical data but also reliable in providing actionable insights for future pitches. The ultimate goal of this project is to develop a robust

analytical framework that can accurately predict outcomes in "Shark Tank" and offer valuable guidance to entrepreneurs looking to navigate the complexities of securing venture funding.

2. Data Description

2.1 Introduction to the Shark Tank Dataset

Our dataset presents an intriguing compilation of 495 pitches from the television show "Shark Tank," with each entry meticulously detailing 24 attributes. These attributes encapsulate diverse aspects, ranging from deal status and episode specifics to entrepreneur characteristics and financial propositions.

In the preliminary stages of our analysis, we undertook a series of pre-processing steps to refine the Shark Tank dataset for a more coherent and focused study. These steps were critical in transforming the raw data into a format conducive for advanced analytical techniques.

Firstly, we addressed the 'deal' variable, which initially existed in a textual format. Recognizing its binary nature, we converted this variable into a numerical format where '0' represents the absence of a deal and '1' indicates a successful deal. This binary transformation is pivotal for our predictive modeling later on.

Similarly, the 'Multiple Entrepreneurs' variable underwent a transformation. Originally indicating the presence of multiple entrepreneurs in a pitch, it was converted to a binary format for uniformity and ease of analysis.

One of the unique aspects of our dataset was the information about the sharks (investors) present in each pitch. We extracted this information and created individual binary variables for each shark. These variables indicate whether a particular shark was involved in each pitch, thereby enabling us to analyze the impact of each shark's presence on the deal outcome.

The product categories presented in the dataset were diverse and numerous. To simplify our analysis, we grouped these categories into broader segments such as 'Technology', 'Food & Beverage', 'Fashion', ‘Health & Wellness’, ‘Entertainment’, ‘Home’, and ‘Others’. This categorization was followed by one-hot encoding, a technique that transforms categorical variables into a format that could be better utilized in our predictive models.

We also delved into the geographical aspect of the pitches by extracting the state information from the location data. This allowed us to identify any regional patterns or biases in the deal outcomes. We further simplified this information by grouping the states based on their frequency in the dataset, focusing on the top five states and categorizing the rest under 'Other'.

In our quest to ensure the dataset's robustness, we eliminated columns that were irrelevant to our analysis, such as 'description', 'entrepreneurs', and 'website'. This step was crucial in reducing the noise within the data and enhancing the focus on variables that potentially influence the deal outcome.

Lastly, we tackled the challenge of outliers in our dataset. Outliers can significantly skew the results of an analysis, especially in a dataset like ours with numerous numerical variables. We employed the Interquartile Range (IQR) method to identify and address these outliers, ensuring that our dataset provides a true representation of the typical scenarios encountered in Shark Tank pitches.

Through these pre-processing steps, we have transformed the Shark Tank dataset into a streamlined and analysis-ready format, setting the stage for a comprehensive exploration of factors that influence the success of pitches in Shark Tank.

2.2 Dependent Variable (Y)

The dependent variable in our analysis is the 'deal' status, indicating whether the entrepreneur was successful in securing a deal with the sharks. This binary variable is central to our study as it reflects the outcome of each pitch. A preliminary examination reveals a nearly balanced distribution between successful and unsuccessful pitches (see Appendix A).

2.3 Independent Variables x_i Overview

Our dataset contains a wide range of independent variables potentially influencing the likelihood of securing a deal. Appendix B shows the distribution and relation with the dependent variable deal for some of the independent variables. These variables are categorized into several groups:

- Episode Details: Variables like 'episode' and 'season' offer insights into the timing of the pitch. The 'episode' variable shows a relatively uniform distribution across the series.
- Financial Aspects: Attributes such as 'askedFor', 'exchangeForStake', and 'valuation' detail the financial propositions of each pitch. These variables exhibit a right-skewed distribution, indicating varied financial expectations from entrepreneurs (see Appendix B).
- Entrepreneur and Pitch Characteristics: This includes whether multiple entrepreneurs are pitching, the category of the product, and the state of the entrepreneur. These variables provide context to the pitch and may influence deal outcomes.
- Shark Presence: Each shark's participation in a deal is a critical variable, reflecting their individual investment patterns and preferences.

2.4 Correlations and Multicollinearity

In our exploration of variable relationships, the heatmap analysis revealed intriguing correlations, particularly in the realm of financial demands and their impact on deal success (see Appendix C). Our scrutiny through the lens of the Variance Inflation Factor (VIF) did not indicate significant concerns of multicollinearity, affirming the independence of our predictors. Naturally, the states, sharks, and

categories had some degree of correlation between them as they are binary variables closely related to each other. However, it is still important to keep all of them to make the predictions, and the VIF values are not alarming enough (all lower than 7) to remove one of them. We also thought about removing the valuation variable since it can be computed with the AskedFor and ExchangeForStake variables but the VIF value was lower than 7, so we kept all three of them.

2.5 Summary

In summarizing Section 2, we recognized the diverse nature of the Shark Tank dataset, capturing a wide array of factors influencing deal success. Our initial analysis shed light on the distribution patterns, potential predictive power of various variables, and data concerns such as collinearity, setting the stage for more complex modeling to forecast deal outcomes more accurately.

3. Model Selection

3.1 Methodology

In our project, we embarked on a comprehensive analysis of the 'shark_tank_data_processed' dataset, employing a multi-model approach to predict the likelihood of securing a deal in Shark Tank pitches. The dataset was initially divided into training and testing subsets, a crucial step for unbiased model evaluation. We began our modeling journey with the construction of a decision tree using the rpart package in R, setting a complexity parameter of 0.01. This tree served as a foundational model, providing a clear, visual representation of decision-making processes within the data (see Appendix D).

Recognizing the limitations of a single tree, we further developed an overfitted tree for comparison purposes. Through this, we aimed to identify and mitigate the risk of overfitting, leading to the creation of an optimal tree. This optimal decision tree was tailored to balance model complexity against the minimization of cross-validated error, ensuring both accuracy and generalizability.

To enhance the robustness of our predictions, we next implemented a random forest model, integrating 500 individual trees. The random forest methodology, known for its ensemble learning approach, offers an improvement in prediction accuracy and control over overfitting. Within this model, we conducted an in-depth variable importance analysis using metrics such as %IncMSE and IncNodePurity. This analysis was instrumental in highlighting the variables with the most substantial impact on the decision-making process, such as exchangeForStake, grouped_category_Other, and Multiple.Entreprenuers (see Appendix E).

Further refining our approach, we applied a generalized boosted model (GBM) with a Gaussian distribution, constituting an ensemble of 10,000 trees and an interaction depth of 4 (see Appendix F). The GBM was chosen for its proficiency in handling diverse data types and enhancing prediction accuracy. It

also enabled us to compute the relative influence of each variable, identifying significant predictors like episode, valuation, and askedFor.

Throughout this methodology, our focus was on leveraging the strengths of different modeling techniques, from the simplicity and clarity of decision trees to the robustness and predictive power of random forests and GBMs. Each model served a unique purpose, either in understanding the data better, controlling for overfitting, or enhancing predictive accuracy.

In the second model building section, we filtered the dataset to focus on 'Food & Beverages' pitches by solo entrepreneurs from California. This specificity allowed for a targeted analysis of trends and success rates within this subset. These values are only meant to serve as a foundational example.

We calculated the acceptance rates of individual sharks in this filtered dataset. The top five sharks were identified based on their acceptance rates, providing insights into which investors were most likely to make deals.

We also conducted a clustering analysis using the k-means algorithm, grouping the data into three clusters based on the 'episode' variable. This analysis helped us understand the distribution of deal successes and identify the cluster with the highest success rate (see Appendix G).

The clusters were visualized, and the data was further filtered to focus on the best-performing cluster. This subset's episode range was determined to identify the optimal time in the season for pitch presentations.

The methodology culminated in providing strategic insights, including identifying top sharks to target and the best time in the season for entrepreneurs to present their pitches for maximizing deal success. This comprehensive approach combined data filtering, acceptance rate calculations, clustering, and visualization to extract actionable insights from the Shark Tank dataset.

3.2 Predictor Inclusion/Exclusion Rationale

The selection of predictors was a critical step in our model development. We primarily focused on variables that directly influence investment decisions in Shark Tank, such as the amount requested by entrepreneurs, equity stakes, company valuations, and characteristics of the Sharks present. The rationale for including or excluding specific predictors was based on their statistical significance and practical relevance to the investment decision-making process in Shark Tank scenarios. Special attention was given to avoid multicollinearity, ensuring the independence of predictors for a more reliable model. We decided to keep all the predictors since the random forest model already efficiently uses them to provide the best accuracy.

3.3 Predictor Model Rationale

The construction of the Random Forest model was guided by its ability to effectively manage the complexity inherent in the Shark Tank data. The model was iteratively refined, with decisions driven by a balance of complexity and interpretability. The Random Forest model was chosen for its ensemble learning capability, which aggregates decisions from multiple decision trees to improve predictive accuracy and control overfitting. This model was complemented by a secondary model focused on strategic advice, employing clustering techniques to discern patterns and trends across episodes and seasons. The final model, therefore, combines the predictive power of Random Forest with the strategic insights derived from clustering analysis.

4. Results

4.1 Final Model

The final selected model was a random forest, chosen for its balance between accuracy and interpretability. The model incorporated a range of predictors, with emphasis on those that showed the highest importance scores, such as 'valuation', 'askedFor', and 'episode'. This model effectively captured the nuances of the Shark Tank deal-making process.

4.2 Model Performance Metrics

The performance of the random forest model was primarily evaluated using accuracy and the F1 score. These metrics provided insight into the model's ability to accurately predict deal outcomes. The accuracy of the model was found to be around 53.19%, with an F1 score of approximately 54.17%. These measures of predictive accuracy suggest that while the model is reasonably effective in identifying successful deals (slightly more effective than random which would be 50%), there is room for improvement. The moderate F1 score also points to a balance between precision and recall, but it underscores the potential to refine the model further for better performance. We also computed and got an MSE of 0.2717731, which might be less relevant for a binary variable outcome but still showcases a positive result.

4.3 Predictor Significance

In our analysis, the significance of each predictor was assessed based on its impact on the model's decision-making process. Key predictors that emerged were 'valuation', 'askedFor', and 'episode', each playing a crucial role in influencing the likelihood of a deal. The presence of certain sharks, as indicated by their respective variables (e.g., 'shark_Lori.Greiner', 'shark_Mark.Cuban'), also significantly affected the deal outcome, reflecting the individual preferences and investment styles of the sharks.

4.4 Predictions for the Shark Tank Deals

Using the random forest model, we made predictions on the likelihood of deals being made in various scenarios. These predictions are instrumental in understanding the dynamics of the Shark Tank investment process. They can provide valuable insights for future participants on how different factors might sway the sharks' decision-making. Predictions were also made for entrepreneurs to maximize their chances of getting a deal based on their specific characteristics and products. For example, we concluded that a single entrepreneur from California pitching a product in the Food & Beverages category would have the highest probability of getting a deal if the pitch is made towards the end of the season and in front of sharks Barbara Corcoran, Lori Greiner, Mark Cuban, Robert Herjavec, and Daymond John. We then computed that the probability of this entrepreneur getting a deal with these conditions would be 0.66.

4.5 Conclusion of Results

The analysis and modeling of the Shark Tank dataset offer several business and strategic insights. For entrepreneurs, understanding key factors that influence deal success can guide their pitch preparation. The model's insights into the importance of valuation, capital requested, and the type of product or service can be leveraged to increase the likelihood of securing a deal. For the sharks, the model could aid in decision-making, helping to identify potentially successful pitches and investment opportunities. The second model can also help entrepreneurs to decide when to pitch their product and to whom if they are given the choice in order to maximize their chances of getting a deal.

5. Limitations

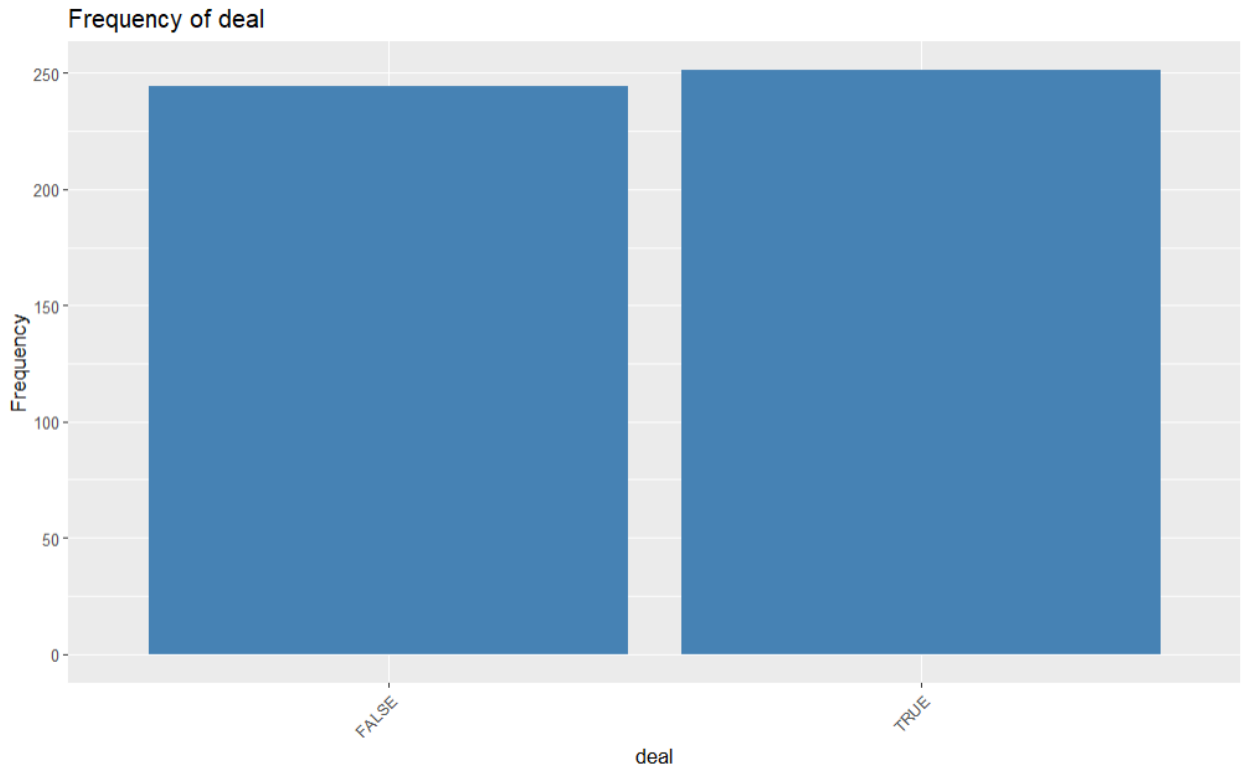
5.1 Limitations of the model

While our model provides a foundational understanding of the factors influencing deal outcomes in Shark Tank, it is essential to note the inherent limitations of predictive modeling. The dynamic nature of business pitches, the unique qualities of products or services, and the individual decision-making styles of the sharks mean that each case might have unique elements not fully captured by the model. Therefore, while the model offers a general guide, each pitch and negotiation in Shark Tank remains nuanced and multifaceted which explains our accuracy score of 0.54. Moreover, we can note that the fact that the dataset only has around 490 rows highly diminishes the predictive power of the random forest model. This code is however scalable and could provide more accurate results if feeded with more data.

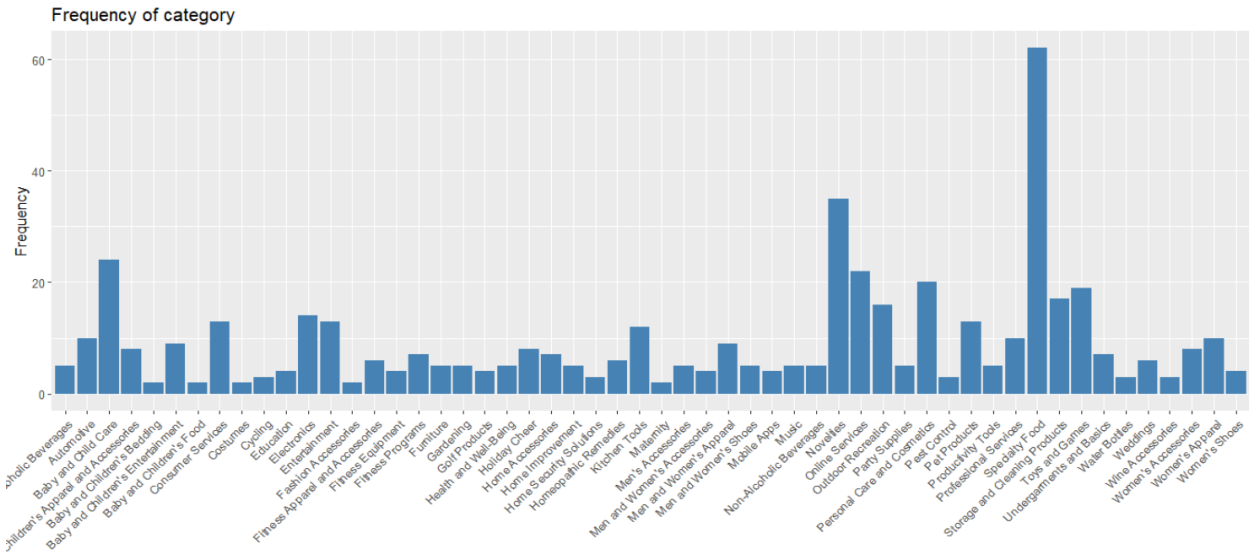
6. Appendices

Appendix A : Deal and Category Frequency Distribution

Graph 1 - Frequency of deals - Histogram



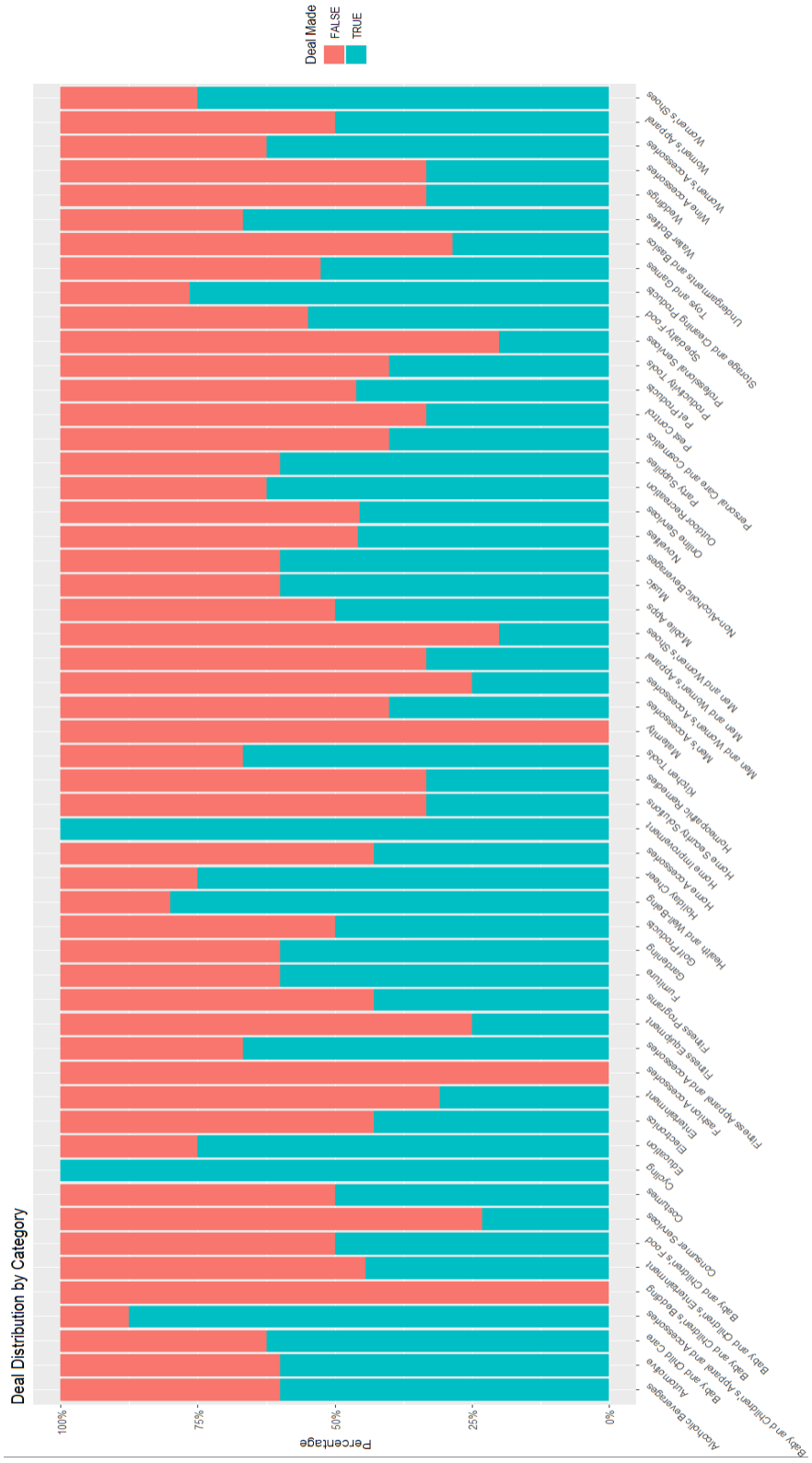
Graph 2 - Frequency of category- Histogram



Appendix B : Variable Analysis Plots

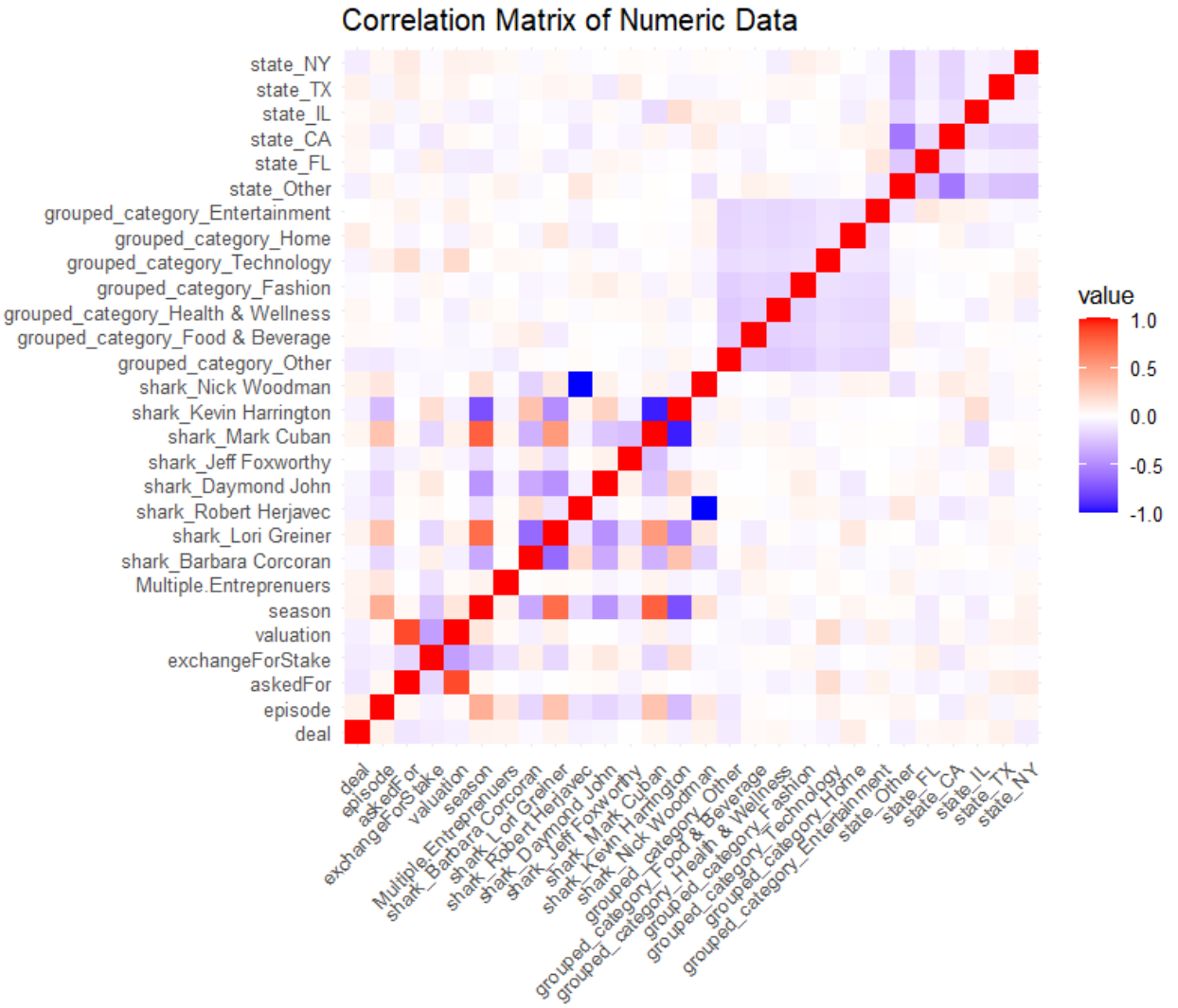
Graphs 3-10 : Scatter Plots/Boxplots of Deal vs 4 variables and histograms for some distributions

Graph 11 : Deal distribution by category



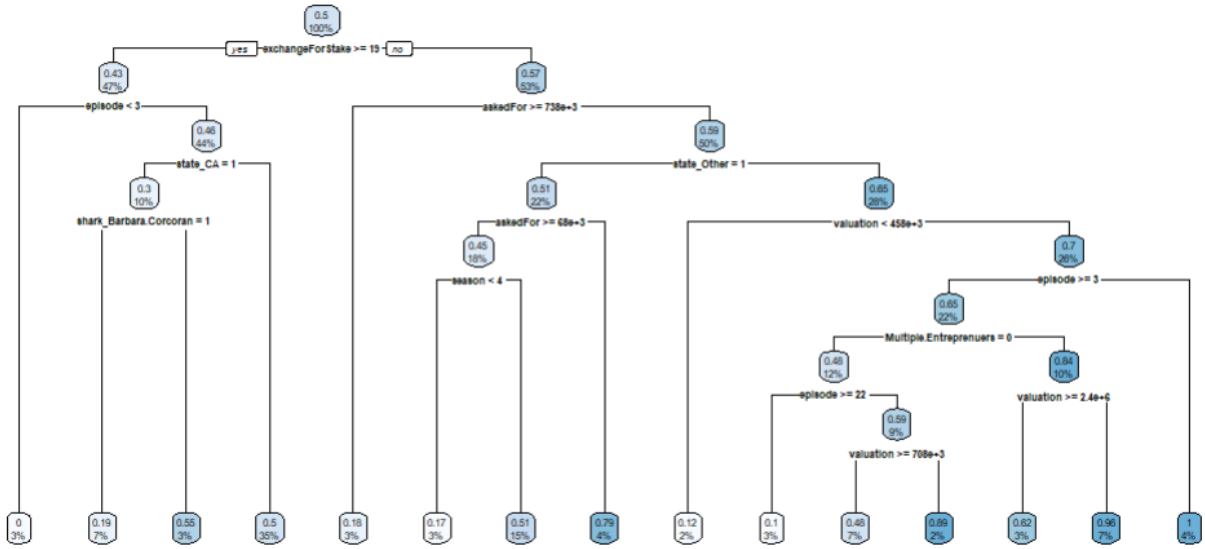
Appendix C : Correlation Heatmap

Graph 12 : Correlation Heatmap of SharkTank data

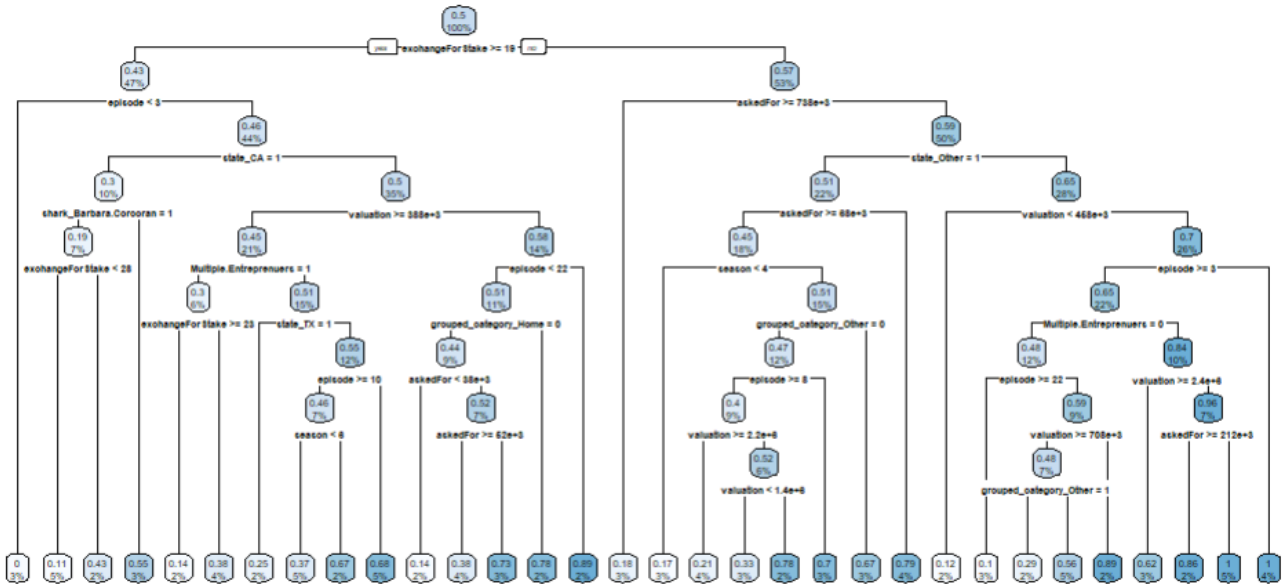


Appendix D : Trees Building

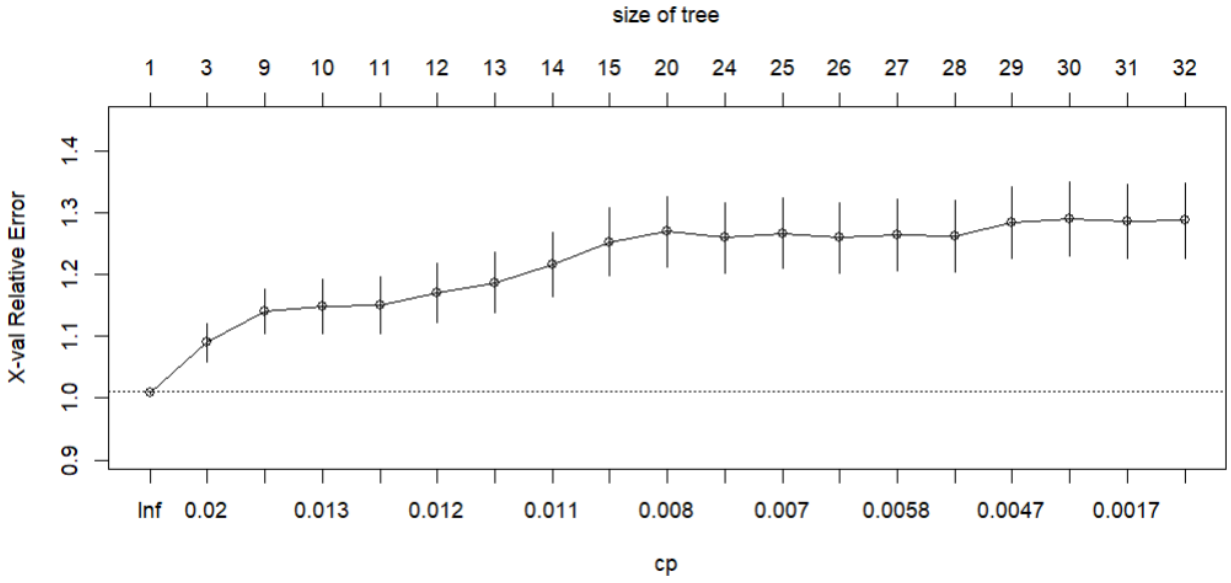
Graph 13 : Initial Tree



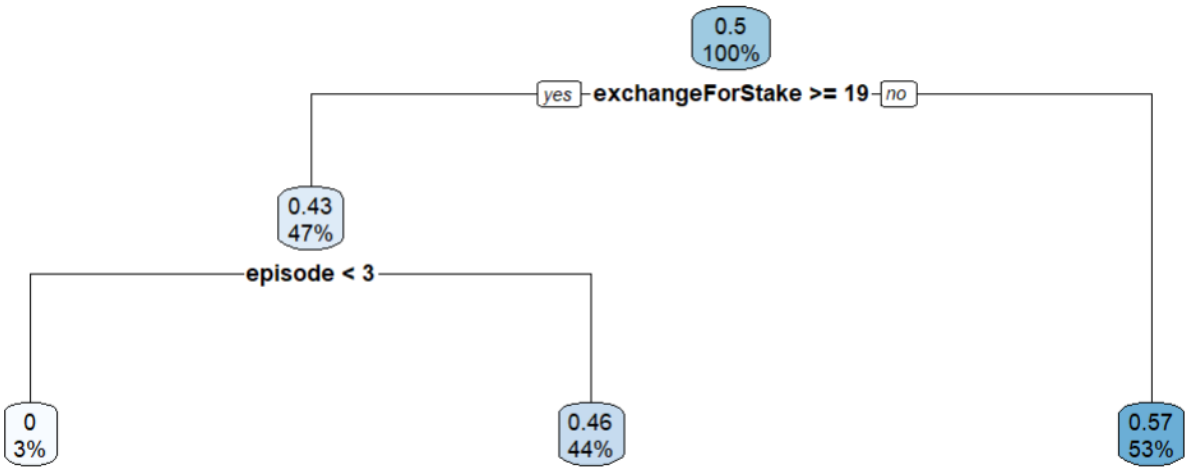
Graph 14 : Overfitted tree



Graph 15 : Relative error for cp and size of tree for overfitted tree

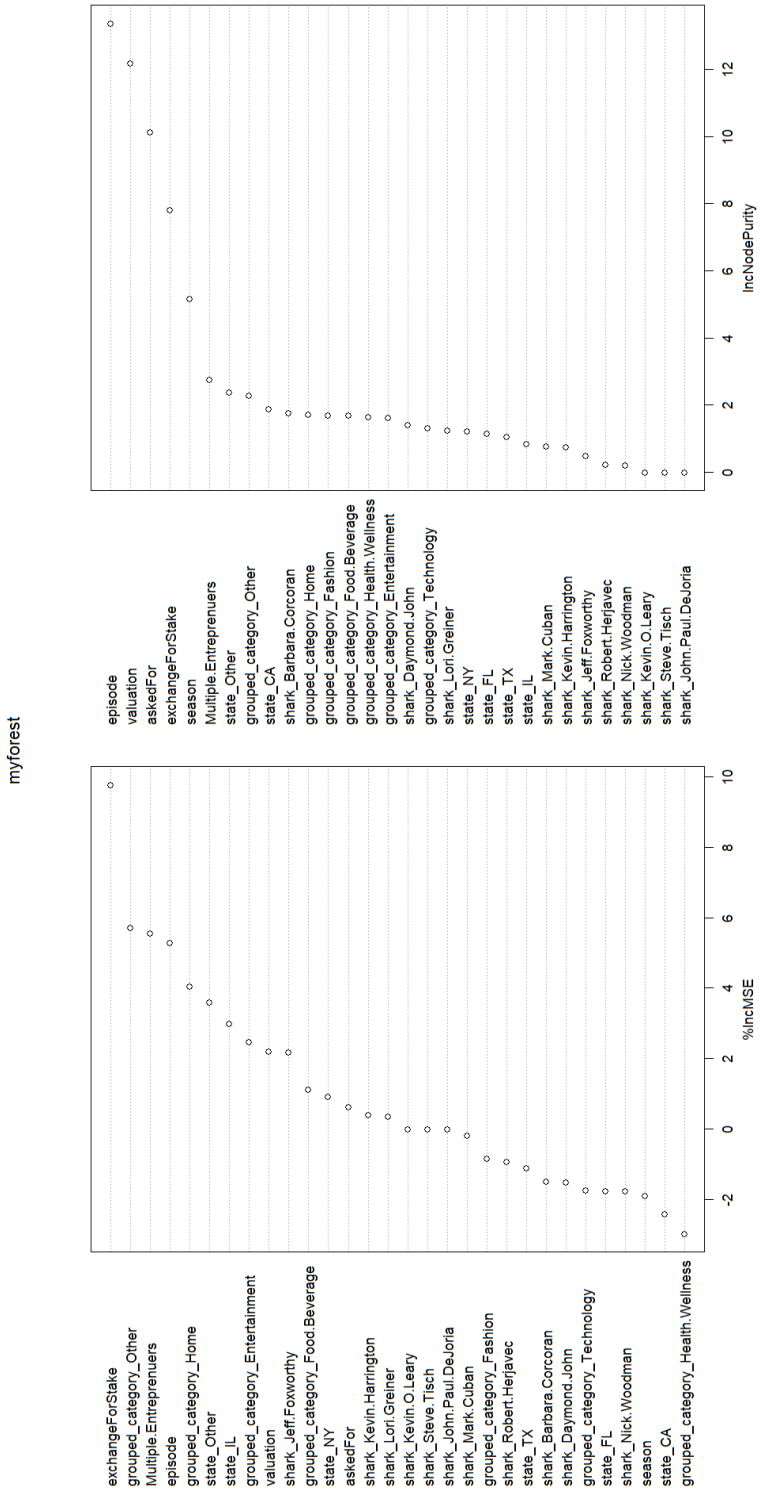


Graph 16 : Optimal tree



Appendix E :Random Forest

Graph 17 :Variable importance in the random forest model



Appendix F : Generalized Boosted Model

Graph 18 :Relative influence of variables for the GBM Graph

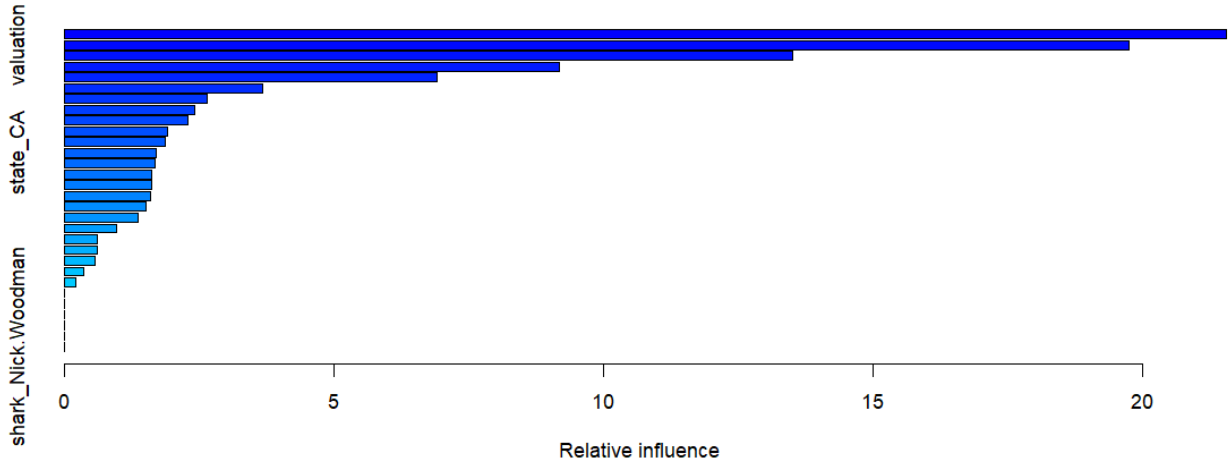


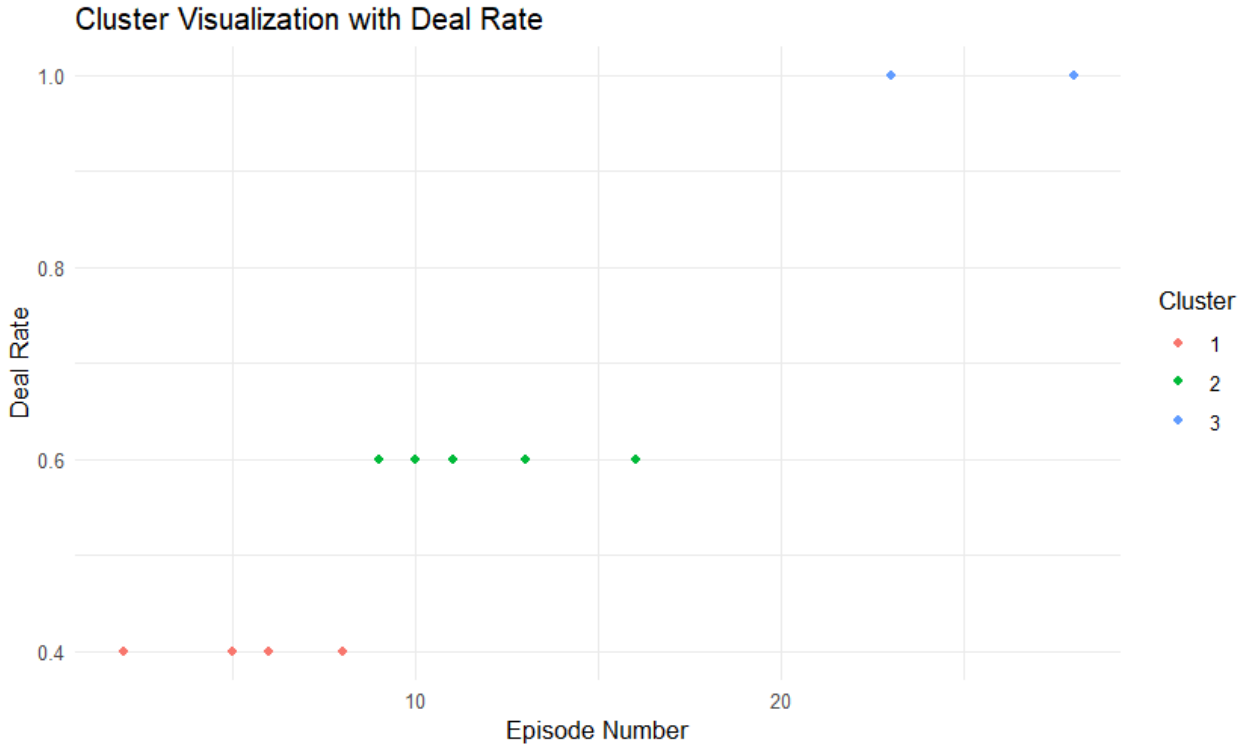
Table 1 :Relative influence of variables for the GBM Table

Table: Summary of Generalized Boosted Model

Variable	RelativeInfluence
episode	21.5593323
valuation	19.7448793
askedFor	13.5145671
exchangeForStake	9.1762258
season	6.8994012
state_Other	3.6793542
shark_Barbara.Corcoran	2.6346201
Multiple.Entreprenuers	2.4138544
grouped_category_Other	2.2862727
grouped_category_Home	1.8968071
shark_Lori.Greiner	1.8653465
state_CA	1.6893675
grouped_category_Food.Beverage	1.6840116
grouped_category_Health.Wellness	1.6084552
grouped_category_Entertainment	1.6051649
grouped_category_Fashion	1.5842428
grouped_category_Technology	1.5018153
shark_Daymond.John	1.3654692
state_NY	0.9548347
shark_Kevin.Harrington	0.6014441
shark_Mark.Cuban	0.5983239
state_FL	0.5695098
state_TX	0.3566185
state_IL	0.2100817
shark_Robert.Herjavec	0.0000000
shark_Kevin.O.Leary	0.0000000
shark_Steve.Tisch	0.0000000
shark_Jeff.Foxworthy	0.0000000
shark_John.Paul.DeJoria	0.0000000
shark_Nick.Woodman	0.0000000

Appendix G : Episode Clustering

Graph 19 : Episode/deal frequency clustering graph for prediction example



7. Code

Please view the attached R file with this submission in order to view the code.