

MGSC-690 MULTIVARIATE STATISTICS

# Predictive Modeling and Clustering Insights for Success on Shark Tank



*Abdulrahman AROWORAMIMO*

Professor  
Prof. Juan SERPA

December, 2023

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Data Description and Pre-Processing</b>	<b>1</b>
2.1	Data Description . . . . .	1
2.2	Data Pre-Processing and Feature Engineering . . . . .	1
<b>3</b>	<b>Exploratory Data Analysis (EDA)</b>	<b>2</b>
3.1	Business Categories Analysis . . . . .	2
3.2	Business Valuation . . . . .	2
3.2.1	Univariate Analysis: Business Valuation . . . . .	2
3.2.2	Bivariate Analysis: Business Valuation and Business Category . . . . .	3
3.3	Stake Offered by Entrepreneurs . . . . .	3
3.3.1	Univariate Analysis: Stake Offered by Entrepreneurs . . . . .	3
3.3.2	Bivariate Analysis: Stake Offered by Entrepreneurs and Business Category . . . . .	3
3.4	Clustering: Business Valuation and Stake Offered by Entrepreneurs . . . . .	3
3.4.1	Cluster Analysis . . . . .	4
3.4.2	Business Category Distribution by Clusters . . . . .	4
<b>4</b>	<b>Model Selection and Methodology</b>	<b>4</b>
4.1	Overview . . . . .	4
4.2	Class Distribution . . . . .	5
4.3	Hyperparameter Tuning . . . . .	5
<b>5</b>	<b>Model Results</b>	<b>5</b>
5.1	Model Performance Evaluation . . . . .	5
5.2	Features Used in the Final Model and Their Importance . . . . .	5
<b>6</b>	<b>Conclusions and Recommendations</b>	<b>6</b>
<b>A</b>	<b>Appendix</b>	<b>8</b>
A.1	Summary Table . . . . .	8
A.2	Frequency Plot of Business Category . . . . .	8
A.3	Summary Statistics of Business Valuation . . . . .	8
A.4	Boxplot of Business Valuation . . . . .	9
A.5	Summary Statistics of Business Valuation Grouped by Category . . . . .	9
A.6	Boxplots of Business Valuation Grouped by Category . . . . .	10
A.7	Summary Statistics of Stake Offered by Entrepreneurs . . . . .	10
A.8	Boxplot of Stake Offered by Entrepreneurs . . . . .	11
A.9	Summary Statistics of Staked Offered Grouped by Business Category . . . . .	11
A.10	Boxplots of Stake Offerings Grouped by Business Category . . . . .	12
A.11	Clustering Based on Business Valuation and Stake Offered . . . . .	13
A.12	Count of Observations in Each Cluster . . . . .	13
A.13	Cluster Characteristics . . . . .	13
A.14	Most Frequent Business Category in Each Cluster . . . . .	14
A.15	Class Distribution . . . . .	14
A.16	Grid Search Result . . . . .	15
A.17	LOOCV Test Result . . . . .	15

A.18 Confusion Matrix of Final Predictions . . . . .	16
A.19 Relative Importance of Features used in the Final Model . . . . .	17

## List of Figures

1 Boxplot of Business Valuation . . . . .	8
2 Boxplot of Business Valuation . . . . .	9
3 Boxplots of Business Valuation Grouped by Category . . . . .	10
4 Boxplot of Stake Offered by Entrepreneurs . . . . .	11
5 Boxplots of Stake Offerings Grouped by Business Category . . . . .	12
6 Clustering Based on Business Valuation and Stake Offerings . . . . .	13
7 Class Distribution . . . . .	14
8 Grid Search Result . . . . .	15
9 Confusion Matrix: Green indicates correct predictions. . . . .	16

## List of Tables

1 Summary Statistics of Numerical Features . . . . .	8
2 Summary Statistics of Business Valuation . . . . .	8
3 Summary Statistics of Business Valuation Grouped by Category . . . . .	9
4 Summary Statistics of Stake Offered by Entrepreneurs . . . . .	10
5 Summary Statistics of Stake Offered Grouped by Business Category . . . . .	11
6 Count of Observations in Each Cluster . . . . .	13
7 Cluster Characteristics . . . . .	13
8 Most Frequent Business Category in Each Cluster . . . . .	14
9 Random Forest Model Evaluation . . . . .	15
10 Variable Importance . . . . .	17

# 1 Introduction

Navigating the realm of innovative business ideas is a daunting task, with the elusive nature of groundbreaking concepts. This challenge is vividly illustrated on the popular platform, Shark Tank, where aspiring entrepreneurs present their ventures to a panel of investors, known as Sharks. The dynamic involves negotiating a deal, often involving relinquishing a percentage of their business in exchange for financial backing and the invaluable mentorship, connections, and expertise of the Sharks. The high-stakes nature of these negotiations compels the Sharks to meticulously scrutinize business profits, records, and performance, aiming to validate the purported valuation of the business. Recognizing the intricacies involved in this venture underscores the potential value of a predictive model capable of accurately forecasting whether a business can secure a deal. Moreover, the insights gleaned from previous pitches can empower entrepreneurs, offering them a strategic advantage in positioning themselves for success while optimizing their use of time and resources. This project aims to develop a classification model to predict business success on the platform, while also conducting exploratory data analysis and clustering to extract actionable insights from past pitches, thereby providing a valuable resource for aspiring entrepreneurs.

## 2 Data Description and Pre-Processing

### 2.1 Data Description

With the project's foundation set, the methodology involves leveraging a dataset comprising data from 495 business pitches across six seasons of the show. Each of the 495 rows corresponds to a pitch, and the dataset consists of 19 columns, each representing a characteristic or feature of the respective pitch. Included features encompass a binary variable indicating whether the entrepreneur secured a deal, the business description, pitch episode and season, business category, entrepreneur names and locations, proposed business valuation, stake offered, amount requested, sharks present (identified as shark 1, 2, 3, 4, and 5), pitch title, and a binary variable indicating the presence of multiple entrepreneurs.

All the features in the dataset except business valuation, stake offered, and amount requested are categorical, which significantly influences the type of exploratory data analysis that can be conducted. The summary statistics for the aforementioned numerical features are provided in [Appendix A.1](#). Each of these variables underwent thorough exploration during the exploratory data analysis (EDA) phase after the data was pre-processed.

### 2.2 Data Pre-Processing and Feature Engineering

All features underwent meticulous examination to identify inconsistencies and potential issues. First, a check for null values was conducted, and none was identified in the dataset. Subsequently, certain columns were deemed irrelevant for this project's scope and were excluded. These columns include company website, business description, entrepreneur names (considered as mere identifiers), entrepreneur locations, pitch episodes, seasons, and titles. While leveraging project descriptions through natural language processing could be explored, it falls beyond this project's current scope. Moreover, the limited dataset size limits the applicability of such technique.

Further, the locations feature was dropped due to an abundance of unique values, hindering its ability to capture meaningful variations in the data. A similar challenge was identified in the business category feature; however, given its crucial role in business assessment, strategic steps were taken to address the issue while preserving its utility.

After ensuring data consistency, the observations in the business category column were strategically regrouped. Initially consisting of 60 unique categories, the business categories were systematically combined to form seven broader groups, namely **beverages**, **apparel and accessories**, **health and wellness**, **services**, **home products**, **baby products**, and **others**. This restructuring was carried out with the aim of retaining all pertinent information and variation captured by the feature while simplifying the categorization for enhanced analysis and utility for predictive modeling.

### 3 Exploratory Data Analysis (EDA)

After the data pre-processing steps were undertaken to prime the data for advanced analysis and modeling, both univariate and bivariate analyses were conducted. The extensive prevalence of categorical features in the dataset constrained the scope of EDA. Nevertheless, univariate analysis was performed on two of the numerical features, business valuation and stake offered. In parallel, bivariate analysis was conducted to understand how these numerical features vary across business categories. Amount requested was omitted from the analysis as it essentially duplicates information captured by stake offered. Since stake offered is represented as a percentage of business valuation, the amount offered column is the absolute amount offered, which lacks standardization for comparative analysis across businesses with varying valuations. Further, clustering was employed on the same two numerical features to unveil additional insights into diverse groups or types of businesses that have historically appeared on the show. This multi-faceted approach aims to extract valuable patterns and relationships within the dataset, contributing to a comprehensive understanding of the factors influencing business success on the show.

#### 3.1 Business Categories Analysis

The conducted bivariate analyses hinge on business categories, making it crucial to comprehend the distribution of these categories across the six seasons of Shark Tank under review. Refer to the frequency plot in [Appendix A.2](#) for a visual representation of the distribution. It is evident from the plot that the most frequently occurring category is the broad group labeled as **other**, which limits the scope for meaningful inference. Nonetheless, among the identifiable categories, **apparel and accessories** emerges as the most popular, while **electronics** appears to be the least favored. This insight provides a foundational understanding of how business categories are distributed throughout the six seasons, setting the stage for more in-depth bivariate analyses.

#### 3.2 Business Valuation

##### 3.2.1 Univariate Analysis: Business Valuation

The summary statistics for business valuation, detailed in [Appendix A.3](#), reveal that the average valuation for businesses featured on the show across six seasons exceeds \$2.1 million. However, a significantly lower median valuation of \$1 million suggests the presence of outliers in the dataset, indicating that a few projects exhibit exceptionally high valuations, as depicted in the boxplot in [Appendix A.4](#). Other notable insights include a minimum valuation of \$40,000 and a maximum of \$30 million, emphasizing the existence of atypical business valuations. This observation aligns with reality, where certain businesses or categories may be associated with higher-than-average valuations. Bivariate analysis was conducted to delve deeper into understanding the factors influencing these valuations.

### 3.2.2 Bivariate Analysis: Business Valuation and Business Category

The **electronics** category stands out prominently in terms of business valuations, as illustrated by the summary statistics table in [Appendix A.5](#). Businesses within this category exhibit the highest average valuation, significantly influencing the overall disparity between mean and median valuations across the six seasons of Shark Tank in the dataset. Notably, the electronics businesses also boast the highest median valuation, indicating that substantial valuations are the norm in this category. This reflects the sector's characteristic high costs, profits, and revenues.

Moreover, examining outliers across all business categories, as depicted in [Appendix A.6](#), reveals a concentration of outliers in the "other" category. Intriguingly, the electronics category not only contains outliers but also demonstrates a more substantial gap between median and maximum valuations. This observation leads to the inference that businesses in the **electronics** category significantly contribute to elevating the average valuation across all businesses in the dataset.

## 3.3 Stake Offered by Entrepreneurs

### 3.3.1 Univariate Analysis: Stake Offered by Entrepreneurs

The average stake offered by businesses was 17.54%, closely followed by a median value of 15.00%, as illustrated in [Appendix A.7](#). This suggests that the majority of entrepreneurs tend to offer values around these figures, with minimal anomalies based on the data under review. Interestingly, one entrepreneur was willing to relinquish 100% of their business, yet they did not secure a deal. The decision to relinquish the entirety of their business might indicate a strategic move to expand their network capitalizing on the sharks' influence even if it means letting go of everything. However, this move might have been perceived as a red flag by the sharks, deterring them from making a deal. Few other notable outliers were observed, as depicted in [Appendix A.8](#). For instance, the minimum stake offered was 3%, which is emphatically atypical; this entrepreneur also did not land a successful deal.

### 3.3.2 Bivariate Analysis: Stake Offered by Entrepreneurs and Business Category

After gaining insights into the distribution of stake offered and its typical values across all businesses, understanding how it varies across different business categories is imperative. With an average value of 19.09%, novelties seem to offer the highest stake percentage. However, the median value of 15.00% suggests otherwise, as shown in [Appendix A.9](#). The business category that offered the highest median percentage of stakes is a tie between **health and wellness** and **apparel and accessories**. The lowest median stake offered is observed in baby products. As mentioned earlier, there aren't many outliers relating to stake offered by businesses. This is evident in [Appendix A.10](#), where all categories showcase similar stake percentages, with some categories even completely devoid of outliers.

## 3.4 Clustering: Business Valuation and Stake Offered by Entrepreneurs

The relationship between business valuation and stake offered was assessed to identify patterns. The curiosity was to see if higher-valued businesses tend to offer higher or lower stakes and vice versa. However, after reviewing the scatter plot of the relationship, as shown in the clusters in [Appendix A.11](#), no discernible pattern was observed in their relationship. In that vein, k-means clustering was employed to cluster these two data points, aiming to create segments of the different kinds of businesses that have graced the show and conduct more in-depth analysis into these groups to uncover valuable insights.

To determine the optimal number of clusters, a value that marks the end of a drastic reduction in within-cluster variation and a drastic increase in between-cluster variation, as normally observed with an

increasing number of clusters, was chosen. This value was found to be five. In other words, by creating five clusters, businesses in each cluster are very similar, and businesses in a cluster are very different from the ones in other clusters. The resulting clusters are shown in [Appendix A.11](#).

### 3.4.1 Cluster Analysis

After creating the five clusters, each project in the dataset was labeled with its respective cluster. [Appendix A.12](#) provides an overview of the distribution of projects across different clusters. Cluster 1 appears to be formed by all the aforementioned outliers with only eight businesses with exceedingly high valuations, while Cluster 4 is the most populated with 313 observations. To glean further insights from these clusters, their centroids (average values of features) were assessed, as depicted in [Appendix A.13](#), showcasing the centroids of each cluster. Cluster 4, the most popular, has an average valuation of approximately \$579,000, which is the lowest and tends to offer the highest stake, averaging 20.87%. Cluster 1, with the highest average valuation of \$23.1 million, tends to offer a low percentage of their business (8.75%), second only to Cluster 3. Businesses in Cluster 3, with an average valuation of about \$5.5 million, tend to offer 8.37% stake. Cluster 2 has the second-highest average valuation (\$11.3 million) and offers a moderate 11% stake. Cluster 5, with an average valuation of about \$2.1 million, offers a moderate average stake of 13.22%.

### 3.4.2 Business Category Distribution by Clusters

The table in [Appendix A.14](#) highlights the prominent business categories within each cluster. Notably, Cluster 1 is predominantly characterized by the electronics and mobile apps business categories, while Cluster 3 is marked by a prevalence of specialty food. Clusters 4 and 5, particularly the former being the most popular, also exhibit a significant dominance of specialty food. In Cluster 2, consumer goods take the lead, along with six other distinct business categories.

It's essential to recognize that this analysis was conducted before aggregating business categories into broader groups, allowing a more granular understanding. As anticipated, the electronics and mobile apps category prominently dominates Cluster 1, the cluster with the highest average valuation. This aligns with the observed atypical valuations of businesses in these categories, as illustrated in the scatter plot in [Appendix A.11](#). Notably, businesses in the electronics and mobile apps categories stand out with the most extreme valuations in the dataset, and they are the ones that make up cluster 1.

## 4 Model Selection and Methodology

### 4.1 Overview

The objective of developing the predictive model is to forecast whether a business will secure a deal. The dataset's limitations in terms of both observations and features underscore the necessity for a classification algorithm capable of capturing complex patterns. In that vein, two powerful algorithms, Random Forest and Gradient Boosted Trees, were selected. These algorithms are capable of discerning complex patterns. Additionally, they are robust to outliers, which are notably present in the dataset. Models were developed using the algorithms, and the resulting models were then compared to determine the superior one. Several critical steps were taken to enhance the final model's performance. These steps include addressing class imbalance (if present), conducting hyperparameter tuning, and implementing cross-validation to ensure the model's generalization.

## 4.2 Class Distribution

The class distribution between businesses that secured a deal and those that did not is even, as depicted in the count plot in [Appendix A.15](#). This balance makes it easy for the model to perform well in classifying both classes accurately, eliminating the need to address class imbalance. Further, accuracy will be an acceptable metric to evaluate the model's performance given the balance between the negative and positive classes.

## 4.3 Hyperparameter Tuning

Initially, a Random Forest model was built, but the performance was unsatisfactory with a predictive accuracy on unseen data of 51%. Considering Gradient Boosted Trees' ability to generalize better than Random Forest, Gradient Boosting was employed. A grid of hyperparameters, including the maximum depth of each tree (`interaction.depth`), number of trees (`n.trees`), learning rate (`shrinkage`), and minimum samples required in a node to stop splitting (`n.minobsinnode`), was explored and tuned. The optimal hyperparameter combination was found to be 100 trees, shrinkage of 0.1, `n.minobsinnode` of 10, and `interaction.depth` of 2, as depicted in [Appendix A.16](#). These low numbers are likely due to the limited observations in the dataset. Therefore, the best model was complex enough but not overly complex, ensuring model parsimony. Additionally, the relatively low learning rate enhances model generalization.

To evaluate the model's generalization, a Leave-One-Out Cross-Validation (LOOCV) was employed during the grid search to maximize accuracy. Due to the limited number of observations in the dataset, this cross-validation method was considered most appropriate. It ensured that each observation had the opportunity to participate in the training process, offering a reliable assessment of the model's performance. After obtaining the optimal set of hyperparameters, the final model was utilized to make predictions on the entire dataset.

# 5 Model Results

## 5.1 Model Performance Evaluation

The performance of the final model is not satisfactory, which is expected given the limited number of features and observations. The LOOCV test result indicates an out-of-sample accuracy of 52% for the best model, as shown in [Appendix A.17](#). When the model was applied to the entire dataset as the test set, the accuracy slightly improved to 67%, as depicted in the confusion matrix in [Appendix A.18](#). This suggests that there is room for improvement in the generalization of the model. As mentioned earlier, given the dataset's class balance, accuracy suffices as an evaluation metric, representing the ratio of total predictions to correct predictions.

## 5.2 Features Used in the Final Model and Their Importance

The dataset used in this project initially included non-feature columns, some of which were removed. It is crucial to pinpoint columns that possess potential predictive power and utilize only those in constructing the model. However, tree-based ensemble algorithms excel at handling high-dimensional datasets, as features with greater predictive power contribute more significantly to the ensemble's overall predictions compared to those with lower predictive power. Consequently, a feature importance value is assigned to each feature in the model to quantify its contribution to the model's predictions.

The determination of feature importance is based on the mean decrease in impurity in each leaf. A relative importance value is assigned to each feature, indicating its contribution to increasing the purity



of each leaf. Essentially, this measures how effectively a feature can group observations, ensuring that each observation is as similar or homogeneous as possible and distinct from others. The specific set of features utilized in the final predictive model and their relative importance is outlined in [Appendix A.19](#).

Notably, both stake offered and amount requested were included as features. This decision was made based on the robustness of tree-based models to collinearity, and it was observed that the model's performance improved with the inclusion of both features. Tree-based models, such as the random forest and gradient boosted trees utilized in this analysis, are known for their resilience to issues related to multicollinearity. Including both stake offered and amount requested allows the model to capture nuanced patterns and interactions between these variables, contributing to the overall predictive capability.

The relative importance of features, as depicted in [Appendix A.19](#), demonstrates the significance of valuation, amount requested, and stake offered in determining the success of a business pitch. Additionally, the business category plays a role, albeit to a lesser extent. Contrary to common belief, the presence of particular sharks is not as crucial as the aforementioned factors, with most of the sharks carrying zero relative influence. This inherent feature selection strategy in tree-based ensemble models not only provides actionable insights but also enhances the model's ability to make accurate predictions on new data.

## 6 Conclusions and Recommendations

In navigating the complex landscape of innovative business ideas, this project set out to develop a predictive model for business success on Shark Tank. The analysis of 495 business pitches across six seasons of the show unearthed valuable insights into the factors influencing entrepreneurs' ability to secure deals. While the predictive model's performance indicates room for improvement, the drawn insights significantly contribute to our understanding of the dynamics at play.

A critical finding highlights the paramount role of fair business valuation in securing a deal. The Sharks' decisions appear to be markedly influenced by the perceived value of a business, emphasizing the need for entrepreneurs to meticulously substantiate their claims. Additionally, entrepreneurs must ensure that their businesses are fairly valued from the outset. Surprisingly, the presence of specific Sharks exhibited minimal impact compared to the influential trio of business valuation, amount requested, and stake offered.

Additionally, the cluster analysis revealed intriguing patterns, categorizing businesses into distinct clusters with varying average valuations and stake offerings. According to the analysis, the most prevalent group comprises the least valued businesses, suggesting a preference for ventures with lower financial risk. Despite the model's limitations, these clusters offer entrepreneurs nuanced perspectives for strategic positioning.

Furthermore, distinct clusters, representing specific groups of businesses, reveal varying average stake offerings among entrepreneurs. Despite this diversity, a common thread emerges – entrepreneurs, irrespective of their cluster affiliation, tend to offer around 15%. This observation, coupled with the nuanced differences within specific clusters, suggests that entrepreneurs can strategically leverage this insight. By determining the cluster to which their business belongs, entrepreneurs can align their strategies with the prevalent norms within that identified group. It's crucial for entrepreneurs to recognize the cluster dynamics, as, on average, businesses adhere to the 15% norm; however, certain clusters exhibit substantially different norms. For instance, businesses in Cluster 3 tend to offer around 8.37%, and the ones in Cluster 4 tend to offer around 20.87%. These values are substantially different from the overall average. This discrepancy underpins the need for entrepreneurs to discern the unique characteristics of their respective business groupings and adapt their business models accordingly.

The aforementioned insights and recommendations are invaluable; however, it is essential to acknowledge the project's limitations. The dataset's constraints in terms of both features and observations necessitate caution in drawing overarching conclusions. Insights gleaned from this project are heavily dependent on the dataset used, which may not fully reflect the complexity of reality and account for intangibles. Recommendations for future analyses include integrating additional data encompassing business profit, performance, cost, revenue, and more for a comprehensive and robust assessment. This broader dataset will provide a more nuanced and accurate understanding of the factors influencing business success on platforms like Shark Tank.

In essence, while the predictive model's current accuracy suggests refinement opportunities, the insights gleaned form a robust foundation for subsequent analyses. Entrepreneurs are advised to focus on justifying their business valuations, as this emerges as a linchpin in the Sharks' decision-making process. As the Shark Tank journey continues for aspiring entrepreneurs, the lessons learned from this endeavor provide invaluable roadmap for navigating the complex realm of innovative business ventures.

## A Appendix

### A.1 Summary Table

Table 1: Summary Statistics of Numerical Features

Statistic	N	Mean	St. Dev.	Min	Max
Amount Requested(\$)	495	258,490.90	461,599.90	10,000	5,000,000
Stake Offered	495	17.54	10.06	3.00	100.00
Business Valuation	495	2,165,615.00	3,761,971.00	40,000.00	30,000,000.00

Click [here](#) to go back to Data Description and Pre-Processing

### A.2 Frequency Plot of Business Category

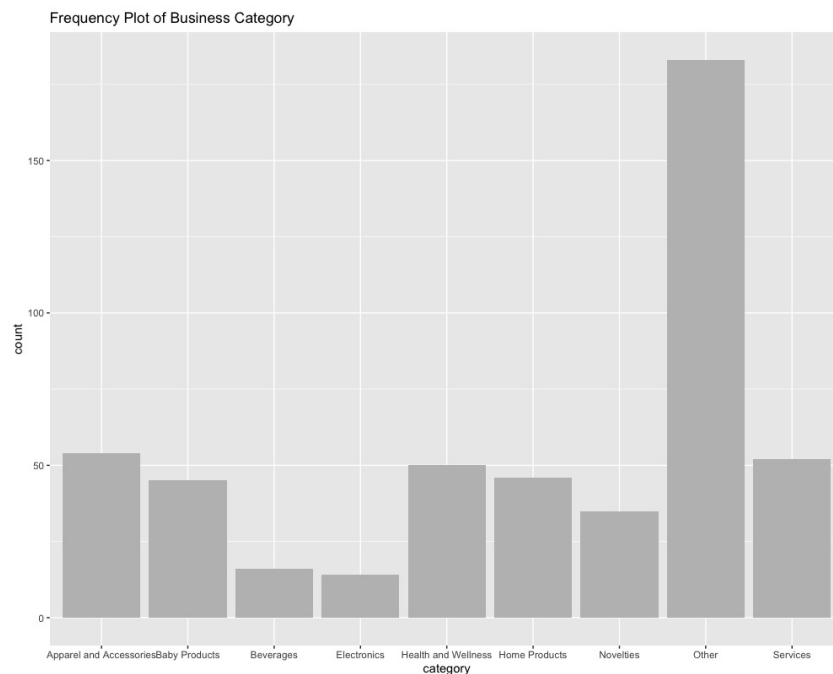


Figure 1: Boxplot of Business Valuation

Click [here](#) to go back to Business Categories Analysis

### A.3 Summary Statistics of Business Valuation

Table 2: Summary Statistics of Business Valuation

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Valuation (\$)	40,000	440,000	1,000,000	2,165,615	2,000,000	30,000,000

Click [here](#) to go back to Univariate Analysis: Business Valuation

## A.4 Boxplot of Business Valuation

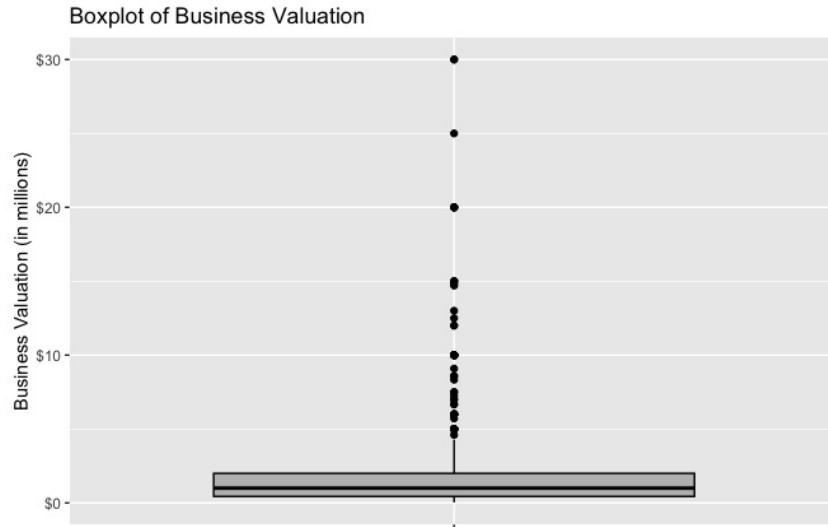


Figure 2: Boxplot of Business Valuation

Click [here](#) to go back to Univariate Analysis: Business Valuation

## A.5 Summary Statistics of Business Valuation Grouped by Category

Table 3: Summary Statistics of Business Valuation Grouped by Category

Category	Mean (\$)	SD (\$)	Median (\$)	Min (\$)	Max (\$)
Apparel and Accessories	1,693,485.00	2,342,475.00	657,327.50	85,714.00	12,000,000.00
Baby Products	1,307,844.00	1,926,066.00	625,000.00	106,061.00	10,000,000.00
Beverages	3,179,911.00	6,112,350.00	900,000.00	133,333.00	25,000,000.00
Electronics	6,064,727.00	9,303,031.00	1,257,576.00	100,000.00	30,000,000.00
Health and Wellness	1,838,767.00	3,253,221.00	666,667.00	100,000.00	20,000,000.00
Home Products	1,409,975.00	1,467,734.00	1,000,000.00	100,000.00	7,500,000.00
Novelties	1,586,883.00	2,165,487.00	800,000.00	50,000.00	10,000,000.00
Other	2,356,332.00	3,705,287.00	1,000,000.00	49,020.00	20,000,000.00
Services	2,737,432.00	4,924,174.00	1,000,000.00	40,000.00	30,000,000.00

Click [here](#) to go back to Bivariate Analysis: Business Valuation and Business Category

## A.6 Boxplots of Business Valuation Grouped by Category

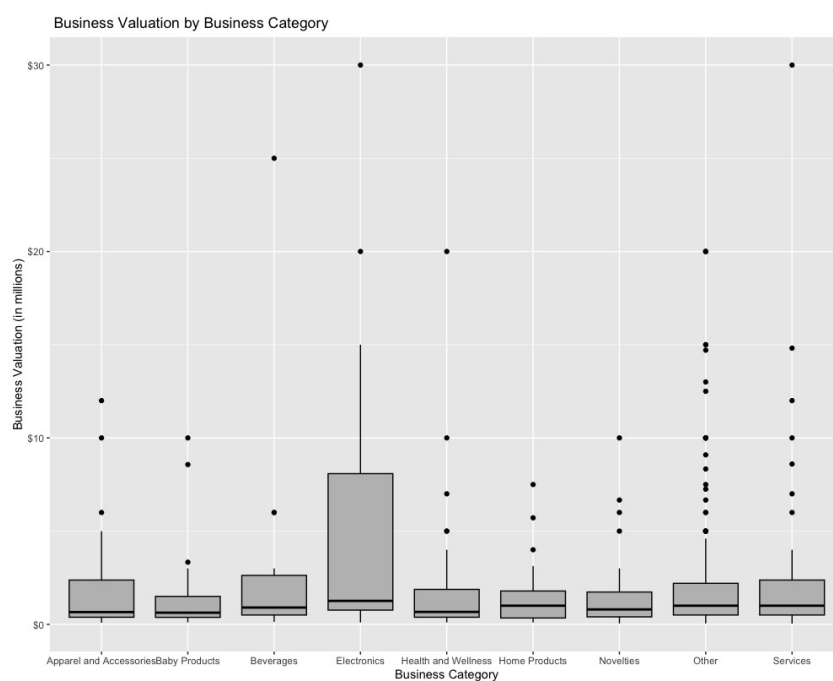


Figure 3: Boxplots of Business Valuation Grouped by Category  
Click [here](#) to go back to Bivariate Analysis: Business Valuation and Business Category

## A.7 Summary Statistics of Stake Offered by Entrepreneurs

Table 4: Summary Statistics of Stake Offered by Entrepreneurs

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Stake Offered (%)	3.00	10.00	15.00	17.54	20.00	100.00

Click [here](#) to go back to Univariate Analysis: Stake Offered

## A.8 Boxplot of Stake Offered by Entrepreneurs

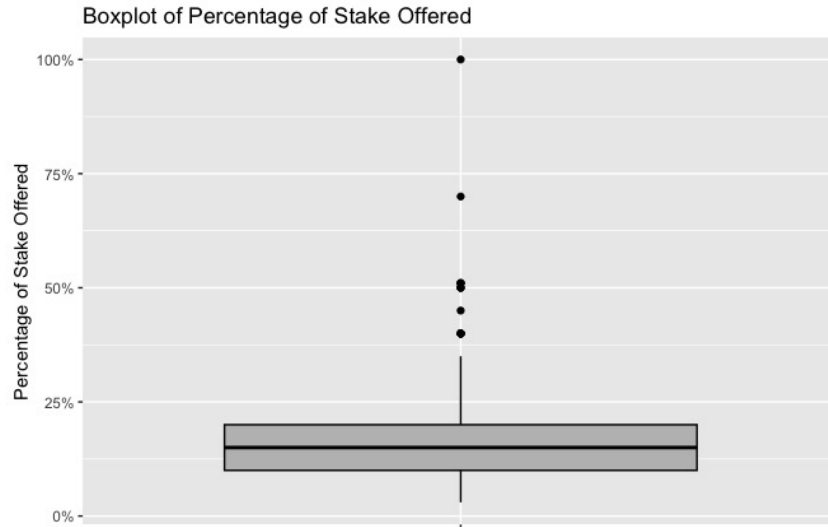


Figure 4: Boxplot of Stake Offered by Entrepreneurs  
Click [here](#) to go back to Univariate Analysis: Stake Offered

## A.9 Summary Statistics of Staked Offered Grouped by Business Category

Table 5: Summary Statistics of Stake Offered Grouped by Business Category

Category	Mean	SD	Median	Min	Max
Apparel and Accessories	18.76	9.85	20.00	4.00	51.00
Baby Products	17.04	7.32	15.00	5.00	33.00
Beverages	16.25	10.41	10.00	5.00	35.00
Electronics	18.86	12.13	17.50	5.00	50.00
Health and Wellness	17.90	10.29	20.00	5.00	70.00
Home Products	17.13	9.67	15.00	5.00	50.00
Novelties	19.09	16.51	15.00	5.00	100.00
Other	17.69	9.61	15.00	3.00	51.00
Services	15.19	7.61	14.50	3.00	40.00

Click [here](#) to go back to Bivariate Analysis: Stake Offerings and Business Category

## A.10 Boxplots of Stake Offerings Grouped by Business Category

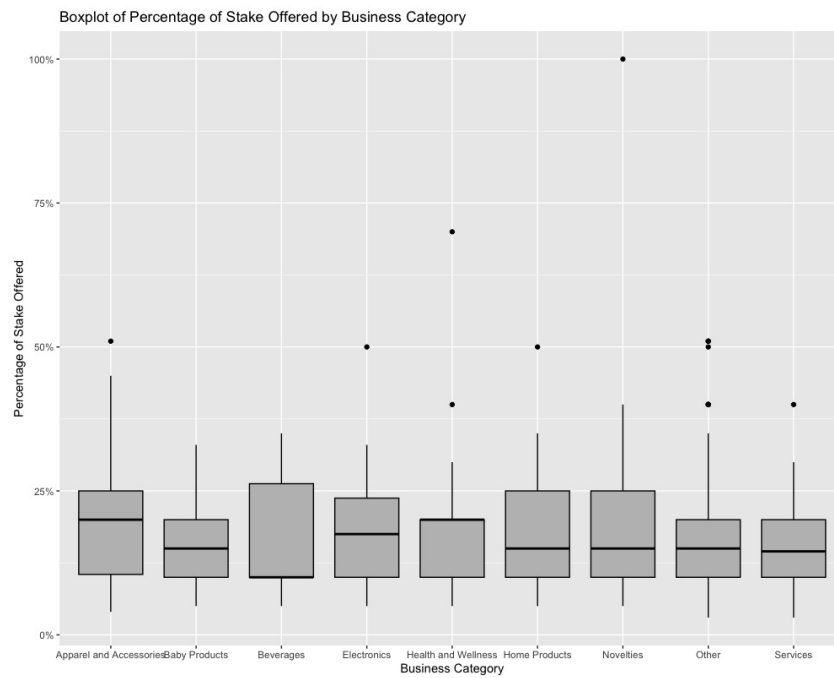


Figure 5: Boxplots of Stake Offerings Grouped by Business Category  
Click [here](#) to go back to Bivariate Analysis: Stake Offerings and Business Category

## A.11 Clustering Based on Business Valuation and Stake Offered

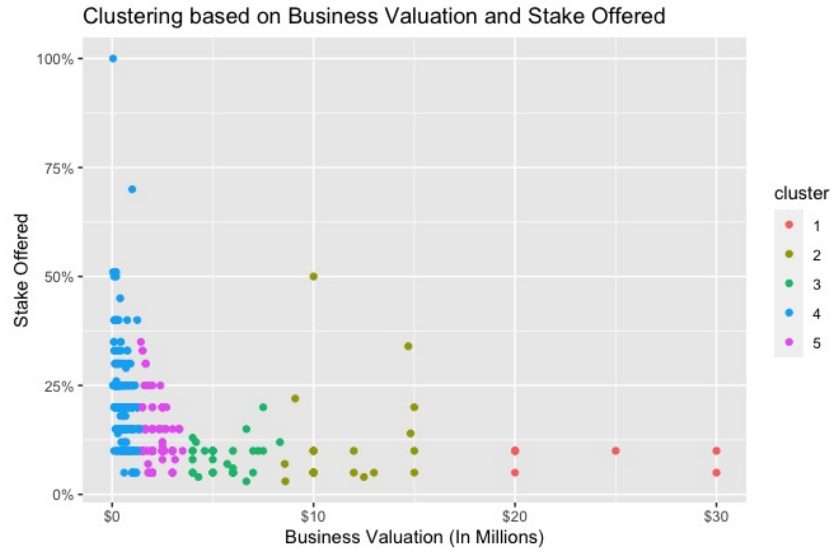


Figure 6: Clustering Based on Business Valuation and Stake Offerings

Click [here](#) to go back to Clustering: Business Valuation and Stake Offered by Entrepreneurs

Click [here](#) to go back to Business Category Distribution by Clusters

## A.12 Count of Observations in Each Cluster

Table 6: Count of Observations in Each Cluster

Cluster	Count
1	8
2	23
3	35
4	313
5	116

Click [here](#) to go back to Cluster Analysis

## A.13 Cluster Characteristics

Table 7: Cluster Characteristics

Cluster	Average Stake (%)	Average Valuation (\$)
1	8.75	23,125,000.00
2	11.04	11,316,653.70
3	8.37	5,505,238.10
4	20.87	578,998.80
5	13.22	2,179,191.90

Click [here](#) to go back to Cluster Analysis



## A.14 Most Frequent Business Category in Each Cluster

Table 8: Most Frequent Business Category in Each Cluster

Cluster	Category	Count
1	Electronics alongside Mobile Apps	2
2	Consumer Services alongside 6 other categories	2
3	Specialty Food	43
4	Specialty Food	11
5	Specialty Food	6

Click [here](#) to go back to Business Category Distribution by Clusters

## A.15 Class Distribution

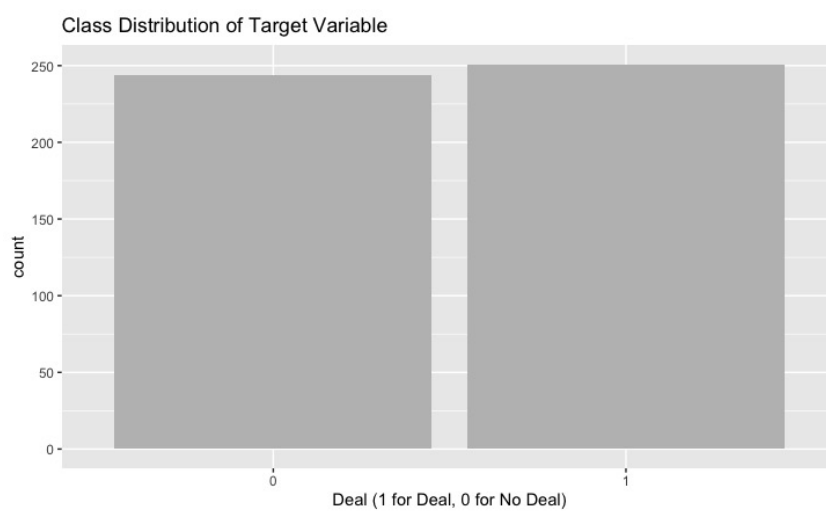


Figure 7: Class Distribution  
Click [here](#) to go back to Class Distribution

## A.16 Grid Search Result

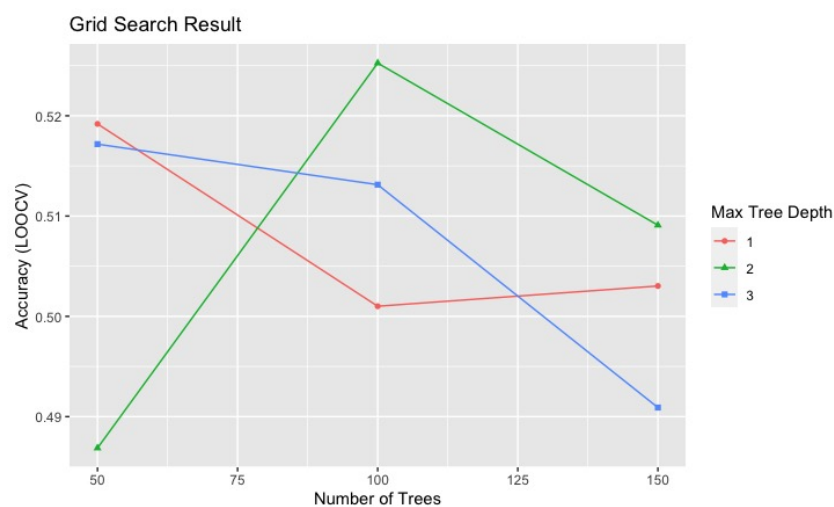


Figure 8: Grid Search Result  
Click [here](#) to go back to Hyperparameter Tuning

## A.17 LOOCV Test Result

n.trees	interaction.depth	Accuracy	Kappa
50	1	0.5192	0.0378
50	2	0.4869	-0.0266
50	3	0.5172	0.0345
100	1	0.5010	0.0011
100	2	<b>0.5253</b>	0.0506
100	3	0.5131	0.0266
150	1	0.5030	0.0054
150	2	0.5091	0.0178
150	3	0.4909	-0.0187

Click [here](#) to go back to Model Performance Evaluation

## A.18 Confusion Matrix of Final Predictions

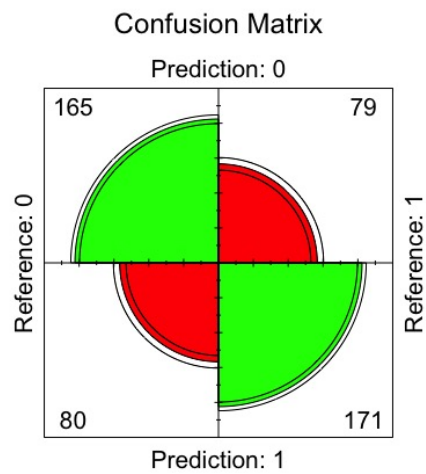


Figure 9: Confusion Matrix: Green indicates correct predictions.

Accuracy = correct predictions divided by total predictions.

$$\text{Accuracy} = (165+171)/(165+79+80+171)=67\%$$

Click [here](#) to go back to Model Performance Evaluation

## A.19 Relative Importance of Features used in the Final Model

Table 10: Variable Importance

Variable	Relative Importance
Business valuation	29.65%
Amount requested	25.44%
Stake offered	13.14%
Category: Services	4.96%
Category: Others	4.84%
Category: Home Products	4.68%
Category: Health and Wellness	2.76%
Shark1: Lori Greiner	2.68%
Presence of multiple entrepreneurs	2.63%
Category: Baby Products	2.51%
Shark5: Kevin Harrington	1.78%
Shark5: Mark Cuban	1.44%
Shark2: Robert Herjavec	1.23%
Category: Novelties	0.95%
Shark3: Kevin O'Leary	0.82%
Shark3: Robert Herjavec	0.48%
Shark2: Kevin O'Leary	0.00%
Shark2: Steve Tisch	0.00%
Shark4: Jeff Foxworthy	0.00%
Shark4: Kevin O'Leary	0.00%
Shark4: Mark Cuban	0.00%
Shark5: John Paul DeJoria	0.00%
Shark5: Nick Woodman	0.00%
Category: Beverages	0.00%
Category: Electronics	0.00%

Click [here](#) to go back to Features Used in the Final Model and Their Importance