**Final Report– Shark Tank**

By the Zucc Team

Mariam Souleyman, Usama Khan, Edwin Bwambale, and Wasif Zawad Talukder

**Table of Contents**

[**Executive Summary 2**](#_Toc153060981)

[**Introduction 2**](#_Toc153060982)

[**Data 3**](#_Toc153060983)

[Data Cleaning 3](#_Toc153060984)

[**I. Analysis 4**](#_Toc153060985)

[**1. Got Deal: 4**](#_Toc153060986)

[a. Logistic Regression 4](#_Toc153060987)

[b. Classification 5](#_Toc153060988)

[c. Comparing the Models 5](#_Toc153060989)

[**2. Total Deal Amount 6**](#_Toc153060990)

[a. Multiple Linear Regression 6](#_Toc153060991)

[b. Regression Tree 7](#_Toc153060992)

[c. KNN Regression 7](#_Toc153060993)

[d. Comparing the Models 8](#_Toc153060994)

[**3. Total Deal Equity 8**](#_Toc153060995)

[a. Multiple Linear Regression (MLR) 8](#_Toc153060996)

[b. KNN Regression 9](#_Toc153060997)

[c. Regression Tree 10](#_Toc153060998)

[d. Comparing the models 11](#_Toc153060999)

[**II. Discussion and Conclusion 11**](#_Toc153061000)

[**Appendix 12**](#_Toc153061001)

Executive Summary

This is a project paper on a Data.world dataset on the "Shark Tank" reality TV show. The primary focus is to understand the driving factors behind successful Shark Tank deals and provide practical insights for entrepreneurs. The dataset was partitioned into 60% training set and 40% test data. Various predictive models were built for each of the three outcome variables "Got Deal," "Total Deal Amount," and "Total Deal Equity" using the training data, and the best models were chosen based on the prediction accuracy on the test data. The winning models can be used by entrepreneurs to build their optimal pitches by using their data in the predictor variables and considering the model outputs holistically.

Logistic Regression and Classification Tree models were developed for predicting successful deals. The more successful Logistic Regression model highlighted the significance of the original ask amount and original offered equity in determining deal success.

Multiple Linear Regression, Regression Tree, and k-Nearest Neighbor Regression models were employed to predict both numeric variables Total Deal Amounts and Total Deal Equity. The Regression Tree model exhibited the lowest RMSE for both numeric variables. The findings underscore the importance of considering factors such as original ask amount, offered equity and the number of sharks. Additionally, the tradeoff between immediate financial gain (royalties) and long-term ownership (equity) should be carefully evaluated.

Continued validation and refinement of models are crucial, involving regular updates with new data and enhancing model interpretability. The model should also be refined to include crucial sales data like revenue, gross profit and net profit over the business’s operational period. This analysis serves as a valuable first step for stakeholders in the Shark Tank ecosystem, empowering entrepreneurs to refine their pitches and providing investors with nuanced insights into deal dynamics.

# Introduction

In this project, we undertook a comprehensive analysis of the “Shark Tank” reality TV show. Entrepreneurs (Pitchers) present their business concepts to potential investors, known as “Sharks”, with the aim of securing funding and mentorship. The show also serves as a platform for enhancing brand visibility and driving sales, even without formal deal agreements.

Our analysis covered approximately 13-14 years of the show's history, spanning from its first-airing episode to the present day. Our primary objective was to understand the specific challenges present in the Shark Tank environment and come up with practical insights for potential participants and interested observers.

To achieve this, we outlined specific goals:

1. Identify patterns characterizing successful Shark Tank deals
2. Formulate a practical guideline for crafting effective pitches
3. Investigate factors contributing to unsuccessful deal outcomes
4. Identify key factors influencing the magnitude of successful deal amounts

We adjusted our analysis methods to match the type of dependent variable we deemed notable. For binary dependent variables, we employed classification tree and logistic regression. For numeric dependent variables, our methodologies included K Nearest Neighbor, Regression Tree, and Multiple Linear Regression.

Through these analytical approaches, our project aimed to offer a comprehensive understanding of the dynamics within the Shark Tank arena, providing practical insights for entrepreneurs, investors, and enthusiasts alike.

# Data

The dataset used by our group contained 1274 pitches made by entrepreneurs on the show Shark Tank.   
We looked at three independent variables:

1. Got Deal: a binary variable that indicated the success of securing the deal i.e., whether the pitch was able to receive funding from any of the sharks for their business
2. Total Deal Amount: a numeric variable that stated the final amount that both entrepreneur(s) and Shark(s) agreed upon
3. Total Deal Equity: numeric variable representing the percentage of equity given up by the entrepreneur(s) to the Shark(s) in exchange for the investment.

# Data Cleaning

Before commencing any analysis on our dataset, we encountered a challenge involving numerous vacant rows and columns. To address these issues, we conducted data cleaning to ensure that our analytical models wouldn't be influenced by any missing values. Variables where the results are only known when a deal is successful were excluded from predictive models to avoid overfitting with training data. Total Deal valuation was omitted from our models to prevent multicollinearity because valuation is simply a calculation between equity stake and investment amount, both of which are used in our model as predictors.

Variables such as Number of Sharks in a Deal, Royalty Deal, Loan, Total Deal amount, Total Deal Equity, and Multiple Entrepreneurs contained numerous missing values. For numeric variables like Number of Sharks in a Deal, Total Deal amount, and Total Deal Equity, we observed that empty rows were due to the absence of a secured deal. Consequently, missing values for these variables were substituted with 0.

Binary variables like Royalty Deal and Loan had over 95% of missing values indicating that no deal was secured in those cases. Hence, all empty cells for these two variables were replaced with 0.

Regarding the Multiple Entrepreneurs variable, also binary, we examined the pitcher's gender column. For empty values where the gender was stated as Mixed (a team containing both male and female members), empty cells were filled with 1. For rows with a male or female in the gender column, the Multiple Entrepreneur variable was populated as 0.

Categorical variables were converted into dummy variables. Pitches gender and the industry to which the business belonged were transformed into dummy columns, while the state variable was discarded as it held no utility for our analysis.

After addressing all empty rows and columns, we partitioned our dataset, allocating 60% to the training set and 40% to the test set. This partitioning proportion was kept consistent across all our models.

# Analysis

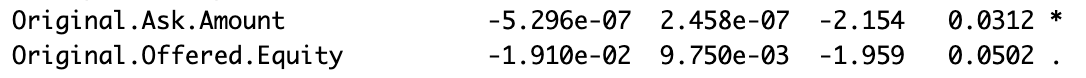
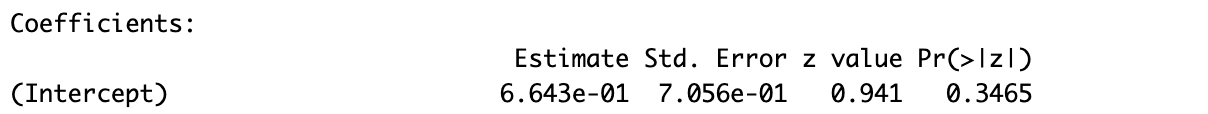
Multiple models were tested for each outcome variable of interest and the best model was chosen based on the lowest value of Root Mean Square Error (RMSE) or Error Rate as applicable which would result in the highest prediction accuracy for the test data.

# Got Deal:

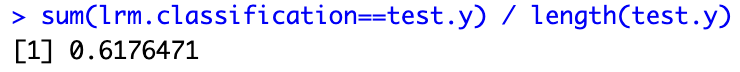
For the “Got Deal” variable, we built Logistic Regression and Classification Tree models, to predict whether entrepreneurs will successfully secure a deal with investors given their respective values of predictor variables. Given the binary nature of this variable, our models were tailored to address binary outcomes.

# Logistic Regression

Concerning the Logistic Regression on all eligible variables in our dataset, p-values revealed that the only significant variables in these particular models were the original ask amount and the original offered equity. This means that the amount entrepreneurs ask and the equity they offer in exchange for a partnership with sharks are the most important factors when determining whether or not they will be offered a deal.



The predicted accuracy for our logistic regression model was **61.76%.** While we acknowledge that the accuracy falls below optimal levels, it is important to consider the limitations imposed by the absence of key variables like Revenue, Gross Profit, and Net Profit in our dataset. These variables seem to hold significant weight which the sharks take into account before evaluating potential deals.



The confusion matrix for the model is shown below:

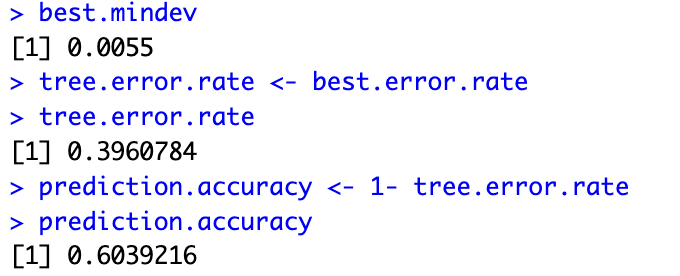
**Actual Results**

|  |  |  |
| --- | --- | --- |
| **Predicted Results** | FALSE | TRUE |
| FALSE | **55** | **45** |
| TRUE | **150** | **260** |

# Classification Tree

The top branches for the classification tree were Original Asked Amount and Original Equity Offered, which were also significant variables in our logistic regression. From the classification tree, we noticed that entrepreneurs who asked for less than $355,000 and offered equity of around 17.75 % tended to secure deals more frequently. This finding holds practical implications for aspiring pitchers presenting their business ventures and ideas on Shark Tank.

The best mindev (minimum deviation) for our model was found at 0.0055 after looping over many different values and finding the lowest error rate. The classification had a prediction accuracy of **60.39%.**  Again, this accuracy is on the low end as our dataset did not have some key variables.



# Comparing the Models

Based on the results, our Logistic Regression model (**61.76%** prediction accuracy**)** outperformed the Classification Tree (**60.39%** prediction accuracy) by a small margin.

# Total Deal Amount

The second dependent variable we decided to investigate was “Total Deal Amount” which reflects the monetary value of the deals made between pitchers and sharks. We ran 3 models, a multiple-linear regression, a regression tree, and a k-nearest neighbor regression. The goal for these was to determine a predictive model with the best prediction accuracy or lowest Root Mean Squared Error (RMSE) for a deal amount that would have the maximum likelihood of securing a deal.

# Multiple Linear Regression

Using R's lm() function, we conducted a thorough “Multiple Linear Regression”. Our variable selection was guided by practical considerations.

> #MLR Interaction

> MLRInteractionmodel <- lm(Total.Deal.Amount ~ . + Original.Ask.Amount\*Original.Offered.Equity -Got.Deal -Valuation.Requested, data=training)

**Got.Deal:** This variable was deliberately omitted to address concerns of potential overfitting associated with actual deal instances. This decision aligns with practical scenarios where users seek to predict “Total Deal Amount” before the confirmation of a deal.

**Valuation.Requested:** Valuation = Ask Amount / Offered Equity. Since we have included both variables that are used to calculate Valuation in our model, driving forces of valuation are presumed to be captured.

**Original.Ask.Amount \* Original.Offered.Equity:** Our investigation aimed to uncover any meaningful impact of the interaction between the asked amount and offered equity. However, the analysis revealed that this interaction term failed to achieve statistical significance in influencing Total Deal Amount.

The findings from our output (figure 2.1) revealed that, among all considered variables, only 8 demonstrated statistical significance. Consequently, we refined our model to concentrate solely on these significant variables.

> # With only the significant independent variables

> MLRInteractionmodel <- lm(Total.Deal.Amount ~ Original.Ask.Amount + Original.Offered.Equity + Total.Deal.Equity + Number.of.sharks.in.deal + Loan + Industry\_Media.Entertainment + Industry\_Travel + Original.Ask.Amount\*Original.Offered.Equity, data=training)

The ensuing summary (figure 2.2) shows us that our model’s statistical strength is confirmed by its large F-statistic of 100.8 and low p-value (< 2.2e-16). The multiple R-squared of 0.5166 indicates that approximately 51.66% of the variability in the dependent variable is explained by the model. Overall, the model seems to provide a reasonable fit to the data.

# Regression Tree

For Regression Tree, we used the “tree()” function with the formula “Total.Deal.Amount ~ . -Got.Deal -Valuation.Requested” to indicate that 'Total.Deal.Amount' served as the dependent variable, and all variables, except 'Got.Deal' and 'Valuation.Requested', were to be utilized as predictors, using the 'training' dataset.

Following this, we used the “predict()” function to get predictions from the tree model on the test data. In addition, we calculated the Root Mean Squared Error (RMSE) based on the predicted outcome values and actual outcome values of test data. The RMSE result yielded was 137039.8. We also plotted our tree and added labels to the nodes.

Our regression tree output (figure 2.3) uncovered a clear correlation between ask amounts and average deal amounts. Starting at the tree's top, the primary decision node, "Total.Deal.Equity < 1.25" indicated that the initial factor in prediction depended on whether the total deal equity was less than 1.25. If so, the prediction led to a terminal node with the value "4421" suggesting a potential average deal amount. Since the average deal amount for a very low equity offer is very low, this indicates a rare deal success scenario.

If the total deal equity was more than 1.25, the analysis shifted to the next decision node: "Original.Ask.Amount < 1.45e+06" (or less than $1,450,000). A branch was established, directing the analysis left if the original asked amount fell below this threshold.

Another decision node, "Original.Ask.Amount < 325,000," was encountered. For original ask amounts less than $325,000, a subsequent check determined if it was less than $140,000. If affirmative, the analysis concluded at a terminal node with a value of "97970." Conversely, if the amount fell between $140,000 and $325,000, the terminal node displayed a value of "258700."

For original ask amounts between $325,000 and $1,450,000, another check assessed whether it was less than $900,000. This model could be used to guide pitchers in setting realistic investment expectations.

# KNN Regression

Our team conducted a KNN regression model by first setting some parameters to check the best value of RMSE by looping for different values of k. We then set our seed and scaled our dataset. Following this, we looped from 1 to our maximum k, performing a k-nearest neighbor regression for each value of k. In each iteration, a KNN regression output was computed using the “knn.reg” function on the training and test datasets.

The RMSE was then computed based on the predicted values (“full.knn$pred”) and the test values, representing the average deviation between predicted and actual values. The best k value was stored in “best.k”, representing the number of neighbors that yielded the minimum RMSE. Simultaneously, the corresponding RMSE was stored in “knn.rmse”. Our findings revealed that the optimal k value was 18, with an associated RMSE of 219,411.4.

The following lines of code can also be used to predict Total Deal Amount by replacing test.x with new values of predictor variables.

> full.knn <- knn.reg(training.x, test.x, training.y, k=best.k)

> full.knn$pred

# Comparing the Models

Among the three models considered, the regression tree emerged as the most accurate, boasting the lowest Root Mean Square Error (RMSE) at 137039.8 for predicting total deals amounts. While the RMSE may appear relatively high, it is important to acknowledge that deals are transacted in thousands of dollars, rendering this level of RMSE acceptable for practical applications.

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| MLR | 240800 |
| KNN | 219411.4 |
| Regression Tree | 137039.8 |

# Total Deal Equity

Three models were run on the outcome variable “Total Deal Equity”. The models used were Multiple Linear Regression, Regression tree, and k-nearest neighbors regression. The goal for these was to determine a predictive model with the best prediction accuracy or lowest Root Mean Squared Error (RMSE) for a deal equity stake that would have the maximum likelihood of securing a deal.

# Multiple Linear Regression (MLR)

A screenshot of a computer program

Description automatically generatedThe lm() function in R was used to run the Multiple Linear Regression. Some variables were left out of the model, and some were included:

**Got Deal:** This variable would inform the model of an actual successful deal instance, which would overfit the model with training data and lead to a higher error rate with new data. In practical use case, the user would use our model to predict their Total Deal Equity before knowing whether they have a deal.

**Valuation Requested:** Valuation = Ask Amount / Offered Equity

Since we have included both of the variables that are used to calculate Valuation in our model, driving forces of valuation are presumed to be captured.

**Original Ask Amount \* Original Offered Equity:** We also decided to observe whether the interaction between Ask Amount and Offered Equity had any significant effect on the Total Deal Equity by using an interaction term. This interaction term turned out to not be very significant.

The output showed some of the variables were not significant (figure 3.1), and so we created a smaller model with only the significant variables. The model’s summary (figure 3.2) reveals the following:

The Multiple R-squared value indicates that the model explains 55.50% of the variations of Total Deal Equity and this is relatively acceptable for a linear regression model.

**Original Ask Amount:** A 1 dollar increase in the Original Ask Amount results in an average 0.0000064% reduction in the Total Deal Equity considering all other independent variables to be constant. This means that a higher asking amount tends to be associated to lower equity deals.

**Original Offered Equity**: A 1% increase in the Original offered causes a 0.52% average increase in the Total Deal Equity keeping other independent variables constant. This means that higher offered equity stakes tend to be associated with higher equity deals.

**Total Deal Amount: The** findings show that for every additional dollar invested, the average equity stake increases by 0.000009924% considering other factors equal.

**Number of Sharks in a deal:** An additional increase in the number of sharks by one shark results in a 12.95% average increase in the Total Deal Equity. This indicates that deals with more sharks tend to be associated with higher equity deals, indicating the positive impact of collaborative investor engagement.

**Royalty Deal:** On average deals with Royalty agreements have a Total Deal equity stake of 7.31% lower than the deals without royalty agreements.

**Loan:** A 0ne dollar increase in the loan amount leads to an average 0.00001% decrease in the Total Deal Equity. This is consistent with the expectation that a higher loan amount would lead to a lower total deal equity. The Sharks are willing to offer less equity in deals with higher loan amounts because they are taking more risk when they make large loan offers since the Pitchers might default hence losses.

# KNN Regression

We divided our dataset into 60% as training to train the model set and 40% as test set to evaluate the performance of our model. We scaled the data to normalize the data to ensure that variables in larger scales do not have larger influences on the model. We then built the KNN model. Set.seed function was used to set the starting point for random number generation since it was useful for reproducing our results. To find the optimal value of k and the RMSE, a loop was used since it captured each interaction and checked to see if the current RMSE beats the best RMSE so far. The model’s performance was assessed using the test set which included calculating the RMSE for different values of k and identifying the optimal k that minimizes the RMSE and then used the trained model to make predictions for new Total Deal Equity.

Entrepreneurs can use the information from the model to understand how their proposed valuation impacts predicted Total Deal Equity. With k of 7, they know that considering the total equity deals for seven nearest entrepreneurs on the show is optimal for prediction. The RMSE of 11.02 can guide Sharks in their negotiation strategies since it helps them understand the level of confidence they can place in a model’s predictions and make informed decisions during predictions.

The following lines of code can also be used to predict Total Deal Equity by replacing test.x with new values of predictor variables.

> full.knn <- knn.reg(training.x, test.x, training.y, k=best.k)

> full.knn$pred

# Regression Tree

The regression tree (figure 3.3) showed that the Total Deal Equity at Shark Tank is split by 5 variables as interpreted below:

**Total Deal Amount:** The first split shows that deals with a total deal amount less than $5000 do not yield any total deal equity while deals with total deal amount above $5000 is further split into other independent variables to predict average Total deal equity of 27.35%.

**Original offered Equity:** The second split in the tree was based on the Original offered equity with a threshold of 19%. For deals with original offered equity below 19%, the regression tree further analyzes factors like Original Ask amount, number of sharks and royalty deal to refine its predictions. If the Original Offered Equity is 19% or higher, the tree proceeds to additional splits, specifically isolating cases where the Original Offered Equity exceeds 28%. In these instances, the model predicts a Total Deal Equity of 47.58%. This indicates a notable impact of the Original Offered Equity level on the final prediction, emphasizing its importance in determining the outcome.

**Original Ask Amount:** Deals with an original ask amount less than threshold of $175000 have an average total deal equity of 23.18% whereas deals with an original ask amount less than the threshold have an average 11.25% total deal equity.

**Number of Sharks:** Deals with a high number of sharks are likely to attract a higher total deal equity of on average 37.19%. This observation highlights the importance of competition and validation on the Shark Tank Show. A high number of Sharks interested in a deal signifies that they believe in the potential of the business, and this provides valuable validation and confidence for entrepreneurs especially during the early stages of their ventures.

**Royalty Deal:** Deals that do not involve a royalty deal tend to have a total deal equity of 23.53% while deal involving a royalty deal resulted in a significantly lower total deal equity of 12.07%. This suggests that royalty deals often come at the expense of equity ownership for the entrepreneurs.

Overall, if an entrepreneur wants to make a prediction for his Total Deal equity using this regression tree, one must consider the fact that deals with a low Original Ask Amount, high Original Offered Equity, high number of Sharks, and absence of a Royalty deal is likely to yield averagely a higher Total Deal Equity.

# Comparing the models

The Regression Tree model had the lowest RMSE. The RMSE comparison is as follows:

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| MLR | 11.12 |
| KNN | 11.02 |
| Regression Tree | 9.85 |

# Discussion and Conclusion

Entrepreneurs seeking deals on Shark Tank are advised to carefully consider their Original Asked Amount and Original Offered Equity. The analysis suggests that asking for amounts below $355,000 and offering equity around 17.75% increases the likelihood of securing a deal.

However, the entrepreneurs should also carefully consider the tradeoff between immediate financial gain (Royalties) and long-term ownership (equity). Accepting a royalty deal may limit their potential profit and control over the business in the long run. Our findings show that royalty deals provide Sharks with a way to participate in a company’s success without taking on significant financial risk since they receive a percentage of sales, regardless of the company’s profitability. The observation about the number of Sharks and its impact is crucial for understanding the dynamics of the Shark Tank Show. It reveals how competition and validation play a significant role in influencing investment decisions and how collaboration can be a powerful tool for success.

Continuous validation and refinement of these models are vital for maximizing predictive accuracy. This can be achieved through the following:

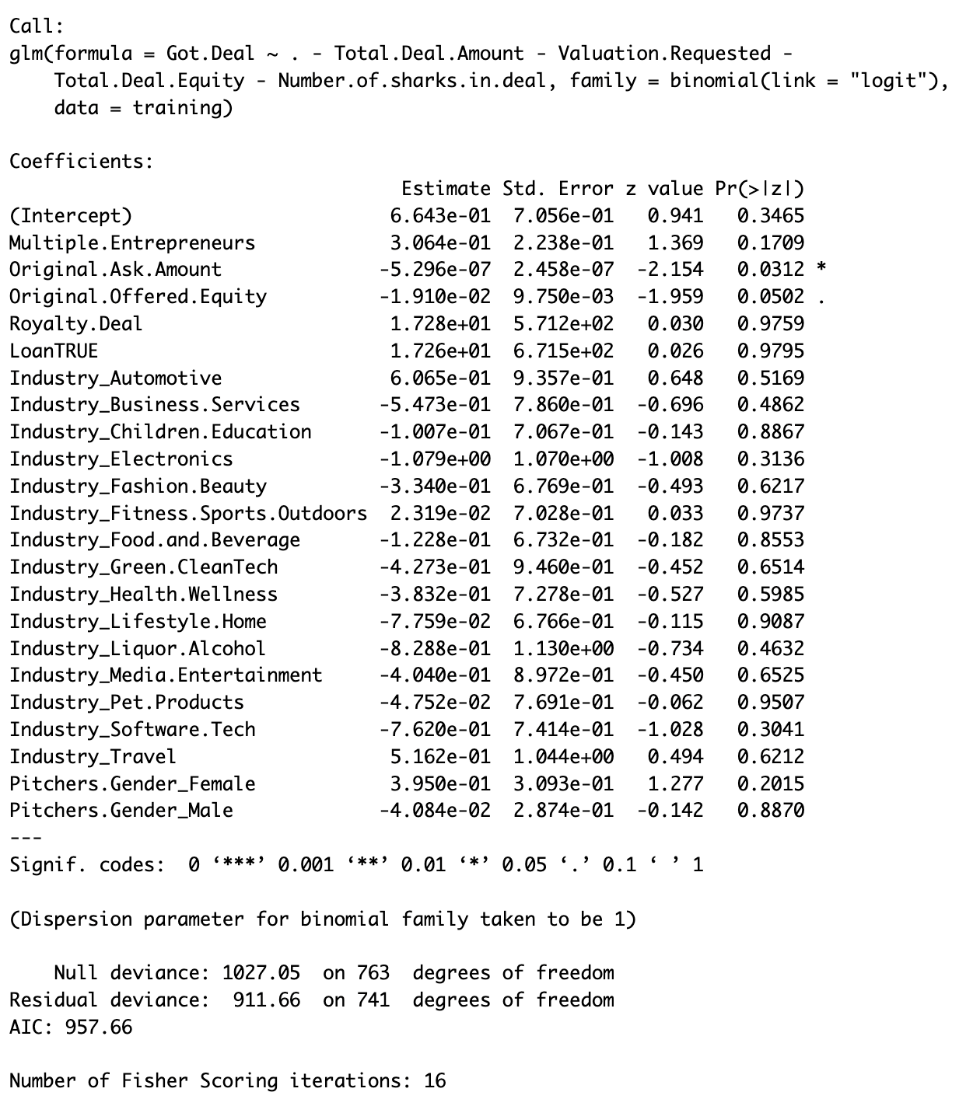
* Regular updates of the new episodes and outcomes. Including additional relevant variables such as Revenue, Profit margins and industry specific metrics.
* Enhancing interpretability of the models, especially the Regression Tree by refining decision nodes or simplifying complex structures. Decision nodes should align with the practical insights from Shark Tank context.
* The model should also be refined to include crucial sales data like revenue, gross profit and net profit over the business’s operational period.

Our analysis revealed valuable insights into the dynamics of Shark Tank deals. While the Logistic Regression highlighted key factors considered for securing a deal, Multiple Linear regression and Regression trees models unveiled better insights into Total Deal amount and Total Deal Equity. These findings empower entrepreneurs to refine their pitches and offer investors a more nuanced understanding of deal dynamics. For predicting Total Deal Amount, the Regression Tree model demonstrated superior performance. Entrepreneurs can utilize this model for more accurate estimations of the monetary value of potential deals. This analysis serves as a valuable first step for stakeholders in the Shark Tank ecosystem, empowering entrepreneurs to refine their pitches and providing investors with nuanced insights into deal dynamics.

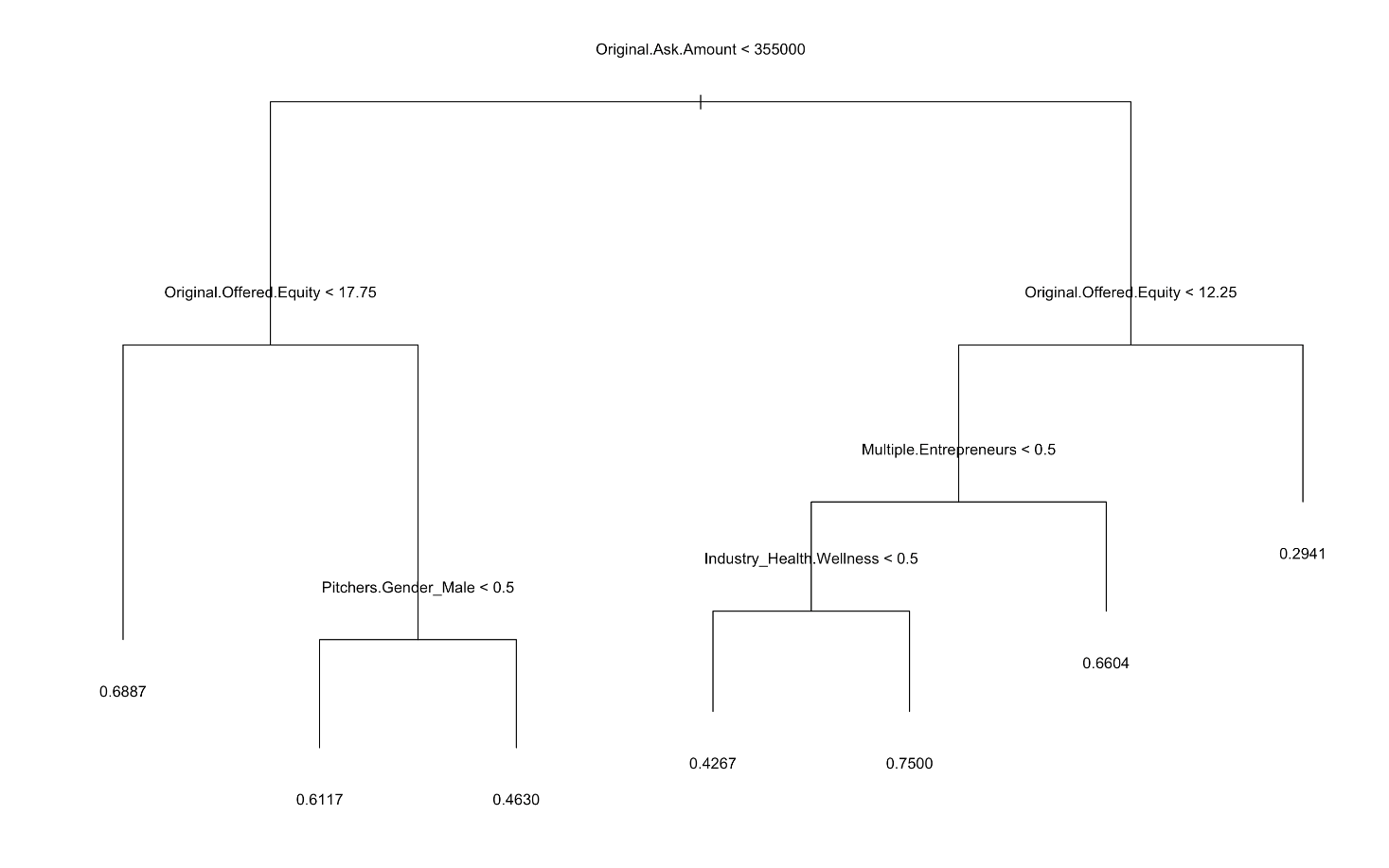
# Appendix

1. **Got Deal:**

**Figure 1.1: Coefficient Summary**



**Figure 1.2: Classification Tree**



1. **Total Deal Amount:**

**Figure 2.1: Initial MLR Analysis**

A screenshot of a computer program

Description automatically generated

**Figure 2.2: MLR with Most Significant Variables Overview**

A computer screen shot of a black screen

Description automatically generated

**Figure 2.3: Deal Amount Regression Tree**

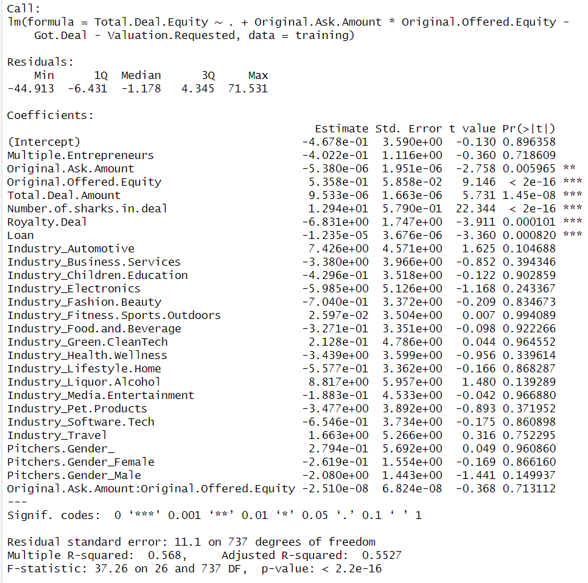
A white background with black lines

Description automatically generated

RMSE = **137039.8**

1. **Total Deal Equity**

**Figure 3.1: Initial Multiple Linear Regression Analysis**

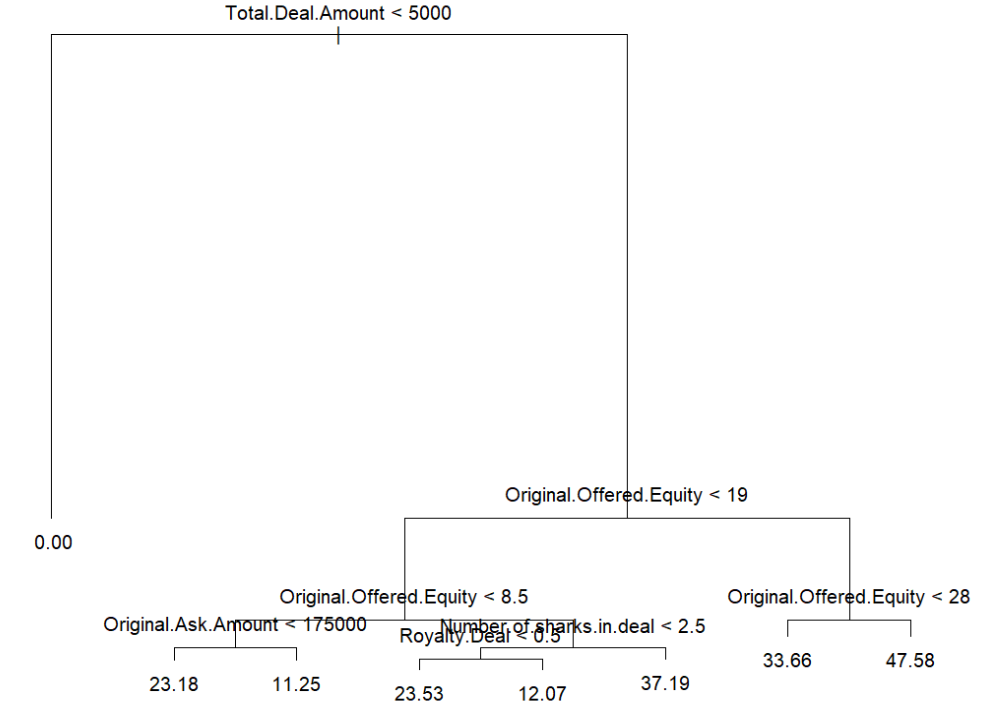


**Figure 3.2: Multiple Linear Regression with Most Significant Variables Summary**

A screenshot of a calculator

Description automatically generated

**Figure 3.3: Total Deal Equity Regression Tree**



RMSE 9.848957