MGSC-661 Multivariate Statistical Analysis

Final Project Report

Name: Amr Maraqa

Student ID: 261101609

Date: 12/15/2022

Introduction

Shark Tank is a reality T.V. show produced by the American Broadcasting Company (ABC). The idea of the show was inspired by the Japanese show “Money Tigers” and has been a sensation in the commercial television industry since its debut in August, 2009 (Frater, 2016). In short, the show broadcasts live business pitches. After going through piles of paperwork and getting screened against hundreds of thousands of other applications, admitted entrepreneurs present their companies in front of a panel of investors (the “Sharks”) in hopes to gain their interest and raise funds. The format of the pitch is very similar to an elevator pitch, whereby the entrepreneur explains the idea of the product or business and informs the Sharks the amount of equity of the business he or she is offering for the amount of funding desired. Once the pitch is complete, the Sharks proceed to ask him/her questions about the business to assess its performance and decide if they would be interested in partnering with the entrepreneur. The questions asked pertain to sales, costs, scalability, future goals, and many more. Once the interrogation is over, two scenarios are possible. The first is the case in which a Shark is disinterested in the business and says that he or she is “out” or not placing an offer. The second situation is one in which the Shark is interested. In that case, offers are usually exchanged between one or more Shark and the entrepreneur in a negotiation process that ends in either an agreement happening or not. If an agreement is reached between any of the Sharks and the entrepreneur, a deal for a partnership is made. Otherwise, no deal is made and everyone leaves the pitch empty handed.

In this project, a dataset of Shark Tank pitches over the first six seasons of the show obtained from Kaggle is examined. The different characteristics of every company that has pitched in those seasons and the relationships between them are studied in an effort to understand how they contribute to the end result and to predict the status of the end of each pitch.

Data Description

General Data Description

Every company in the dataset is described by a collection of 19 variables. These variables and their descriptions can be found in Table.1. in the appendix. The project began with a data exploration process in an attempt to understand each variable representing the businesses. Upon preliminary search, it was found that 51% of the pitches in the dataset ended with a deal being made. This was good news as it demonstrated an even distribution in the data, which meant there was no bias in the response variable, the status of the deal. Thereafter, the distribution of the companies on the basis of other categorical variables was explored. It was found that over the course of six seasons, the founders of 489 companies that served a total of 54 markets had come to the ABC studios from 255 different American cities. Furthermore, it was discovered that 11 investors had taken part of the show up to the 6th season. Throughout the 122 Shark Tank episodes aired on U.S. television, 161 companies were pitched by two or more entrepreneurs, while the other 334 entrepreneurs presented their pitches alone.

The statistics also went beyond the surface to describe these pitches at a deeper level. During every pitch on the show, it can be clearly interpreted that the three most important features of the pitch are the company’s valuation, the funding the entrepreneur desires, and the stake of the company he/she is willing to forego. Sharks can always be seen sitting impatiently in their seats as they wait for the entrepreneur to unveil those details, and it's no secret why. That's where all the money is! Hence, these variables were examined in more depth to attain an understanding of the range of the asks entrepreneurs make, the average amount of stake offered, and much more. The discoveries of this process are shown in figures 1 to 3 in the appendix.

Variable Relationship Description

The final part of this phase of the project was to gain insight on the relationships present between variables. To accomplish this, a principle component analysis (PCA) was performed on the numerical variables in the data. This analysis was done to examine the relationship between the different numerical variables and the success of a pitch. The principle component plot is displayed in figure 4.

The goal from this assessment was to comprehend the trend in successful pitches in terms of money requested, valuation, stake exchanged, number of episode and the number of the season. It could be noted that, in general, more businesses didn't get a deal when their requests for funding and valuations were higher than average. This speaks to the economic sense of supply and demand, where the higher something costs, the lower the demand for it. Additionally, a very high concentration of successful projects was found in regions of low funding request and high stake or low values in both. This indicates that the Sharks are more attracted to companies that offer them a low barrier to entry, as it increases the potential of higher return on their investments.

This PCA didn't only allow for deductions to be made between the numerical variables and the categorical ones, but also amongst the numerical variables themselves. It can be noted from the graph that the Eigen vectors of the variables "askedFor" and "valuation" move closely in the same direction. The same thing was observed for "episode" and "season". This is a sign of possible collinearity between the variables that will be explored in more detail in the next section.

Model Selection and Methodology

Data Pre-processing

Having obtained a rough idea of the nature of the variables and they are related, the process preparing the data for the classification of the pitch status commenced. First, the columns that had no effect on the outcome of the pitch, like the link to the company's website were dropped. Variables that couldn't be analyzed directly, such as the business description that needs natural language processing (NLP) to produce quantitative results, were dropped as well. A total of five variables were discarded during this process. After that, the remaining data was scanned to remove null values to make sure classification models would face no issues when operating on the data.

Next, the data was tested for outliers. To do this, a logistic regression between the target and all the variables was performed, and a Benferroni outlier test was conducted. Six observations in the data were found to deviate from the mean by more than 4 standard deviations, and so, were flagged as outliers. Figure 5 shows the output of this outlier test.

The last problem to check was collinearity. After the observations made from the examination of the principle component plot, it had to be assured that no collinear variables were present simultaneously in the data in order not to skew predictions. Therefore, a correlation matrix between the quantitative variables was produced. The findings from the matrix confirmed our observation of collinearity between “askedFor” and “valuation”, while no significant correlation was detected between “episode” and “season”. The next step was, thus, to remove one of the collinear variables. The variable that was discarded was “valuation”.

Variable and Model Selection

In order to gain an understanding of which predictors affect the target variable and to what extent, a Random Forest (RF) classification model was run. The RF classification was chosen for the variable importance feature that delivers insights on the contribution of each predictor to the prediction. Three RF models were performed. The first was operated on all the variables remaining from the pre-processing stage. The results of this RF can be seen in figure 7 in the appendix. The feature importance plot of this RF showed that all the sharks were more or less insignificant. That raised the question whether it was the number of the shark or the shark him/herself that didn’t matter. So, a variable was made for the presence of each shark, and a second RF model was run. From the RF plot of the second, it could be deduced that the sharks themselves also didn’t contribute to the fate of the pitch. Therefore, from these two plots, it was concluded that all the sharks were insignificant variables and should be left out of the classification model.

From the examination of the first two variable importance plots, it was realized that “askedFor”, “exchangeForStake”, “location”, “category”, and “episode” were consistently significant. They were, hence, taken for further examination. It was now desired to know whether all locations and categories were equally important. So, the location and category variables were split into of four and nine groups, respectively. In the case of location, the 255 cities in the dataset were mapped to four U.S. regions, namely the North East, the Mid West, the West, and the South. As for the categories, the nine categories formed were Media and Entertainment, Fitness, Wellness, Fashion, Services, Home, Children, Nutrition and Occasion. Once the data was ready, a third RF model was performed on the data, the results of which can be found in Figure 9. At the end, “episode”, “askedFor”, “exchangeForStake”, “services”, “media_entertainment”, and “south” were kept.

With all the important variables defined, the logistic regression model and the Gradient Boosting Model (GBM) were selected to be tested. For both models, the data was split into a train and test that consisted of 70% and 30% of the data, respectively. The models were then built, trained, and performed on the test set. Subsequently, the probability of predictions of both models were converted to a binary form. A pitch was considered successful if the predicted probability was higher than chance (50%). 50% was chosen as the determining proportion as that was the original split between successful and failed pitches. Then, the accuracy scores of both models were calculated. The results are discussed in further detailed in the following section.

Results

The accuracy scores of the two classification models tested were calculated by determining the number of correct predictions and dividing it by the total number of predictions made. The results obtained showed a 55% prediction accuracy for the logistic regression model and a 14.7% accuracy for the GBM model. Even though the predictive power of the GBM is weak, the model provided a plot of relative importance, Figure 12, that reaffirmed previous findings: “askedFor” and “exchangeForStake” are some of the most important features of a business on Shark Tank. It remains unclear, however, why “episode” was ranked higher than both these variables, but the fact that there seems to be a higher probability of success the closer the show nears its end is apparent.

The predictive power of the logistic regression was also low. A 55% accuracy indicates a very low improvement on predictions being made as a random guess. The weakness of this model is also solidified by Figure 12, which displays an R^2 of 8.3%. This means that only 8.3% of the change in the status of a pitch is caused by the current combination of predictors. These results coupled with the low accuracy score of the GBM and the high error rates in the RF models is a clear indication of a massive gap in the dataset that should be filled by the acquisition of more variables related to a business in a Shark Tank pitch.

Conclusions

The aim of this project was to perform a comprehensive analysis on the Shark Tank dataset to extract valuable business insights from past data. In order to achieve this goal, each variable in the dataset was first studied singularly to comprehend how it behaved on its own, what this behavior meant, and what were, if any, its hidden trends. The following step was to study the relationships that lied between the variables. The initial technique used was the PCA, which uncovered significant information about the correlation between the amount of funding requested and valuation that was later confirmed by the collinearity test. Once the data was ready, RF models were conducted as preliminary class prediction models. Despite the high error rates in the predictions of these models, they provided an idea of the importance of the features in the data, which allowed the hierarchy of the variables in the data to be understood more clearly. The RF variable importance feature also enabled the preparation of a set of variables that was used in the construction of other classification models.

The other classification models tested on the data were the logistic regression and GBM models. Both of these models were performed to verify the validity of the list of predictors that resulted from the RF models and as a ground of comparison for the performance of the classification models built. The performance of both of these models was very poor, indicating that the data needed augmentation in order to enable the production of more accurate predictions.

The success of this project lies in the detection of a major flaw in the data. The process a Shark undergoes in deciding whether or not to invest in a company is not duly covered by the variables in the dataset. An addition to the data already present in the dataset is the addition of quantitative variables where possible. For example, there are instances in the show when Sharks expressed their appreciation for the name of a business. This stands as proof that there is invaluable information in the description and the title variables that can be deciphered by NLP techniques to obtain an estimate of the power of these words on a Shark's decision. Other more direct factors that are accounted for in the Shark decision making include business sales, customer acquisition costs, customer churn, business models, profit, return on investment (ROI), scalability and entrepreneurial vision. It is, therefore, advisable to ameliorate the dataset in a way that reflects the investment decision making process more closely to be able to make better predictions about the fate of the Shark Tank businesses.

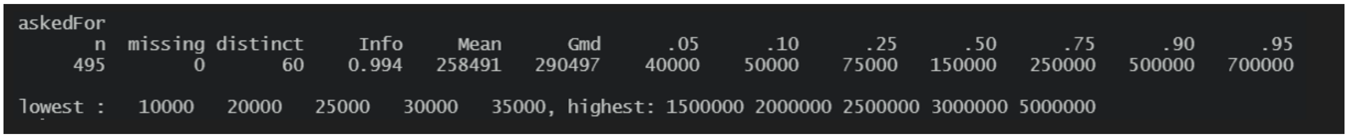
References

Frater, Patrick. (April, 2016). MipTV: 'Dragon's Den' Business TV Show to Be Adapted in China. https://variety.com/2016/tv/asia/dragons-den-shark-tank-in-china-1201744544/

Appendix

Variable Name	Variable Type	Variable Description
Deal	Boolean	Whether a deal with a Shark was made
Description	String	Description of the business product/service
Episode	Integer	Number of the episode
Category	String	Industry/Market the business is in
Entrepreneur	String	Name of the entrepreneur(s)
Location	String	Location of the business in the U.S.A.
Website	String	Link to the business website
Asked For	Integer	Amount of funding requested
Exchange For Stake	Integer	Percentage of the company stock offered for the investment
Valuation	Integer	Company's worth (Asked For * Exchange For Stake)
Season	Integer	The number of the show season
Shark 1	String	Name of the first Shark
Shark 2	String	Name of the second Shark
Shark 3	String	Name of the third Shark
Shark 4	String	Name of the fourth Shark
Shark 5	String	Name of the fifth Shark
Title	String	Name of the company
Episode-Season	String	Episode and season number
Multiple Entrepreneurs	Boolean	Whether more than one entrepreneur was present at the pitch

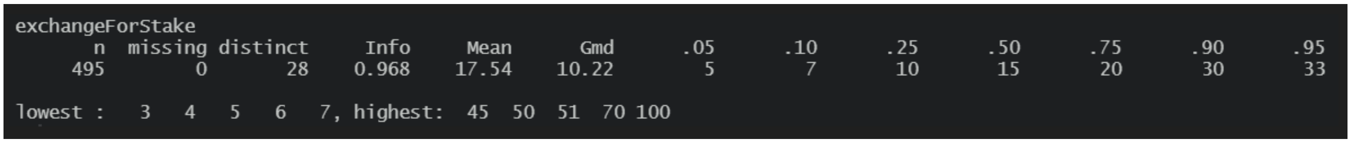
Table 1: askedFor Statistical Description Results



```
askedFor
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     495        0       60    0.994   258491   290497    40000    50000    75000   150000   250000   500000   700000

lowest :   10000   20000   25000   30000   35000, highest: 1500000 2000000 2500000 3000000 5000000
```

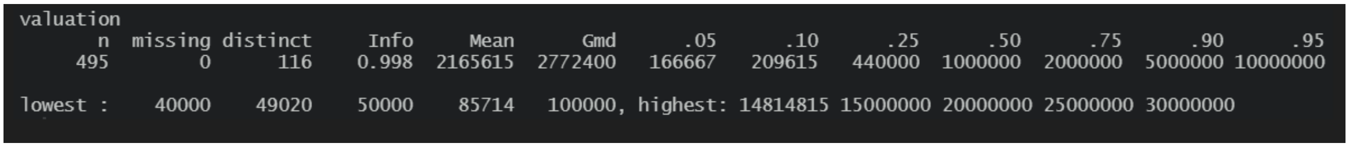
Figure 1: askedFor Statistical Description Results



```
exchangeForStake
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     495        0       28    0.968    17.54    10.22        5        7       10       15       20       30       33

lowest :   3   4   5   6   7, highest:  45  50  51  70 100
```

Figure 2: exchangedForStake Statistical Description Results



```
valuation
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     495        0      116    0.998  2165615  2772400   166667   209615   440000  1000000  2000000  5000000 10000000

lowest :    40000    49020    50000    85714   100000, highest: 14814815 15000000 20000000 25000000 30000000
```

Figure 3: valuation Statistical Description Results

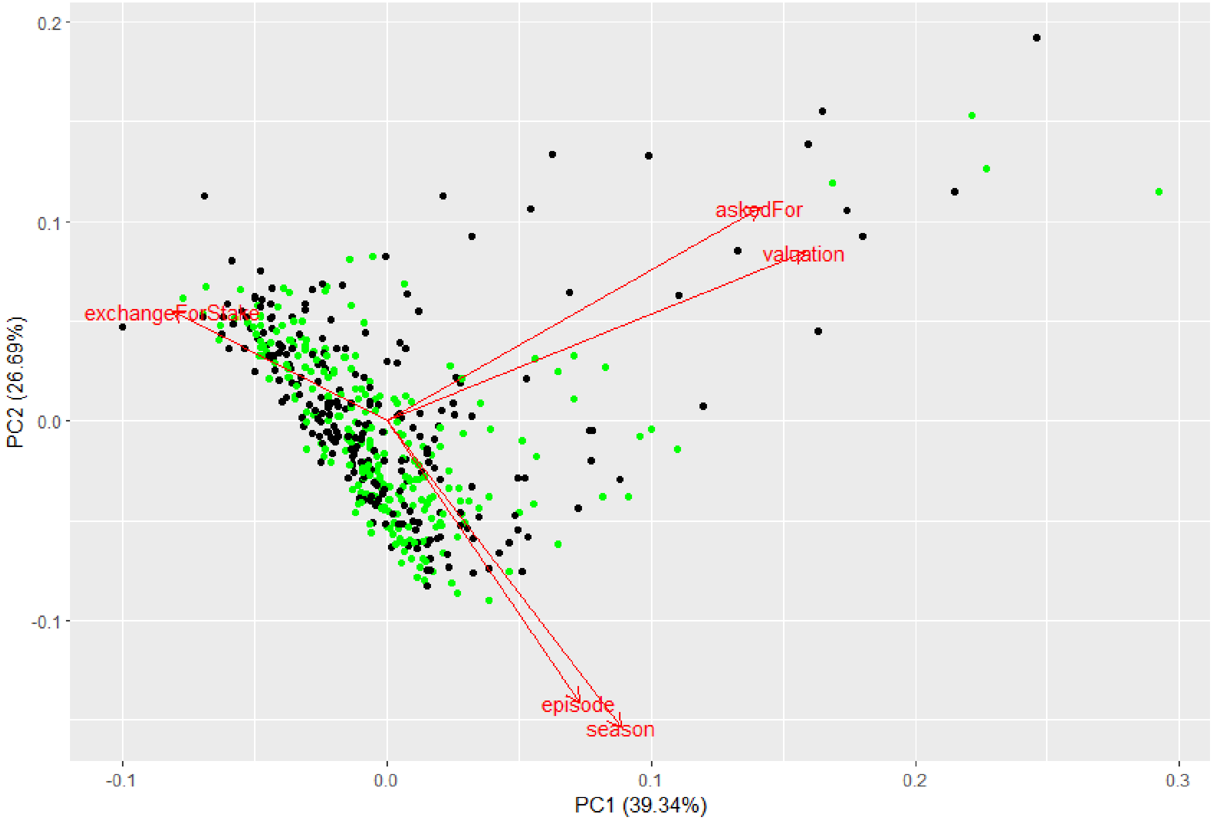
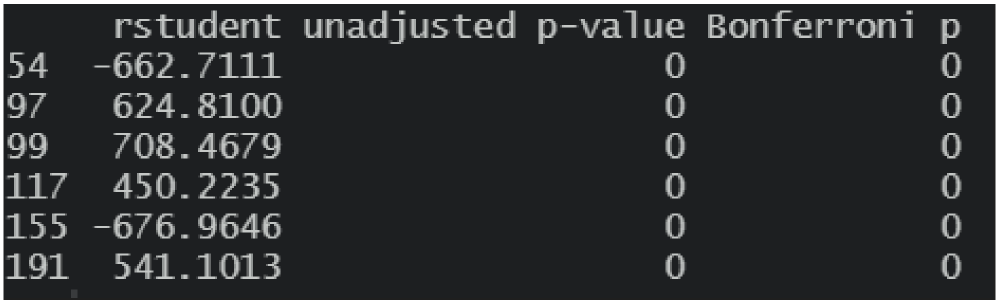
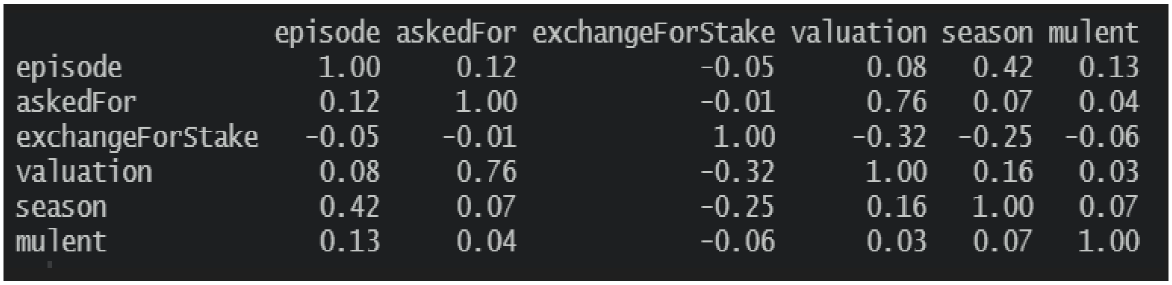


Figure 4: PCA Numerical Variables Plot



```
     rstudent unadjusted p-value Bonferroni p
54   -662.7111                 0            0
97    624.8100                 0            0
99    708.4679                 0            0
117   450.2235                 0            0
155  -676.9646                 0            0
191   541.1013                 0            0
```

Figure 5: Outlier Test



```
                 episode askedFor exchangeForStake valuation season mulent
episode             1.00     0.12            -0.05      0.08   0.42   0.13
askedFor            0.12     1.00            -0.01      0.76   0.07   0.04
exchangeForStake   -0.05    -0.01             1.00     -0.32  -0.25  -0.06
valuation           0.08     0.76            -0.32      1.00   0.16   0.03
season              0.42     0.07            -0.25      0.16   1.00   0.07
mulent              0.13     0.04            -0.06      0.03   0.07   1.00
```

Figure 6: Correlation Matrix of Numerical Variables

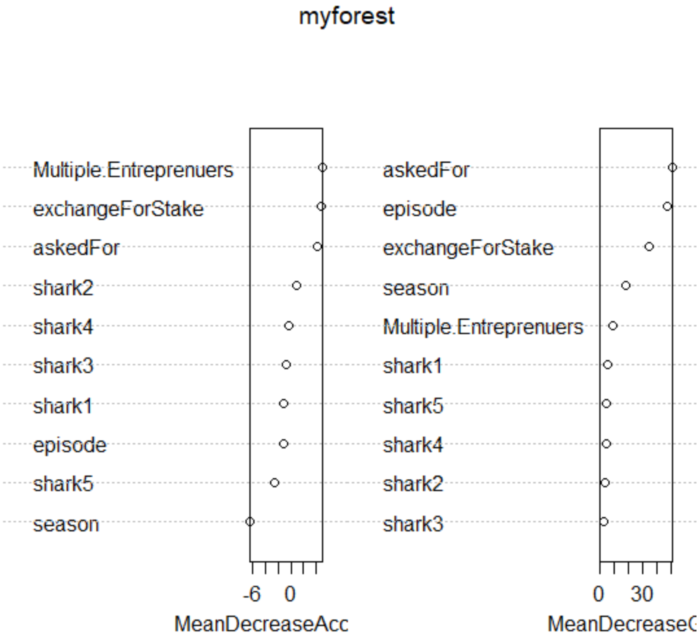


Figure 7: 1st Random Forest Importance Plot

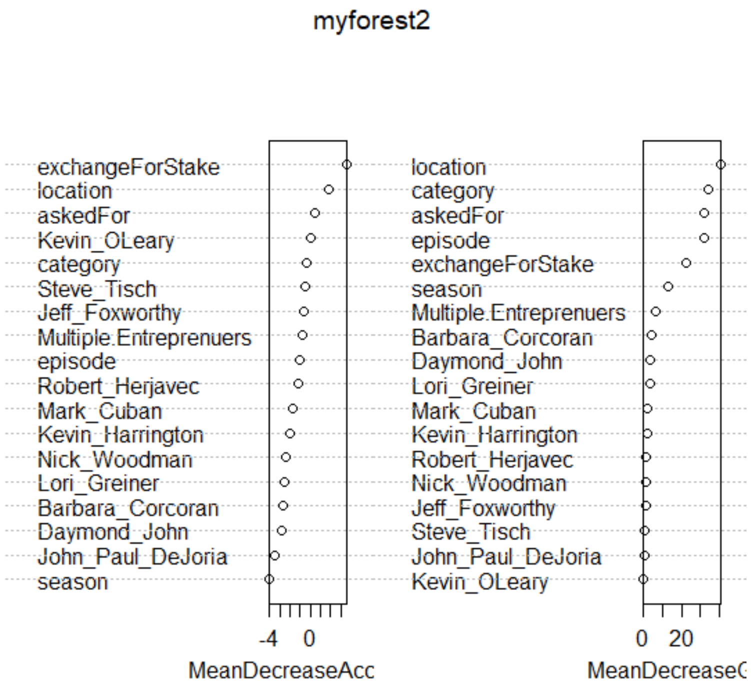


Figure 8: 2nd Random Forest Importance Plot

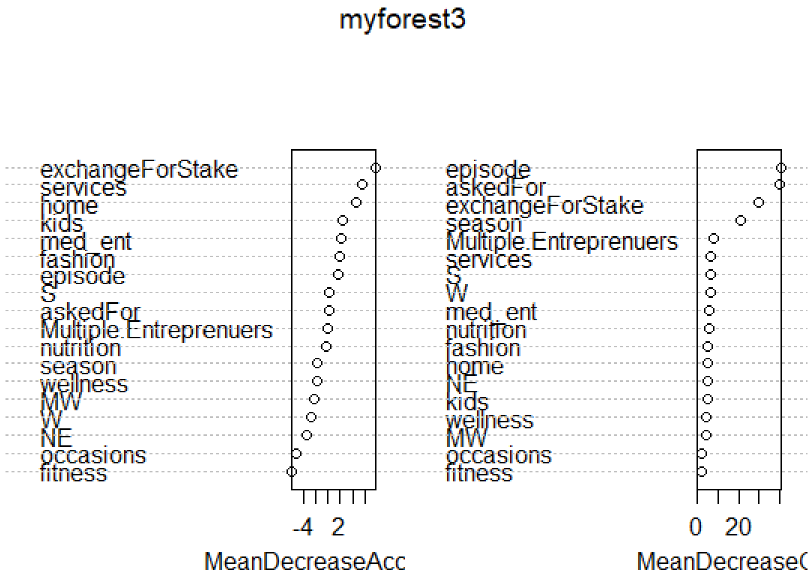
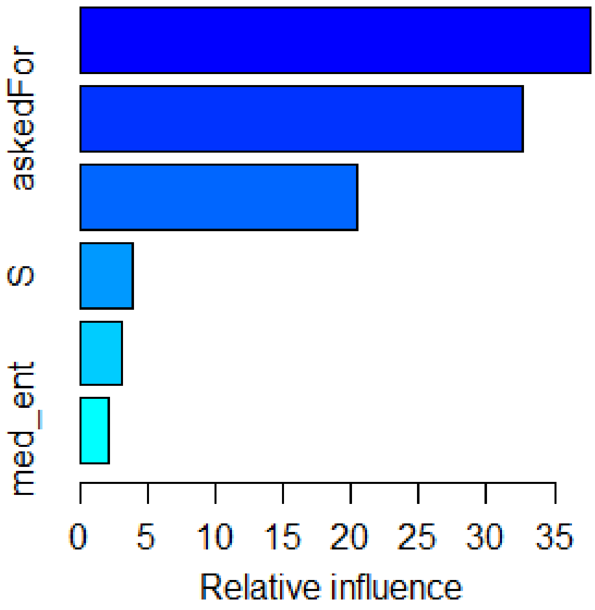
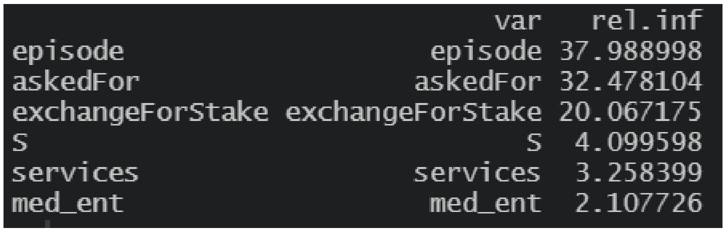


Figure 9: 3rd Random Forest Importance Plot

	Dependent variable:
	Shark Tank Deal Status
Episode Number	0.023
	(0.017)
Requested Funding ()	-0.00000***
	(0.00000)
Company Stake Offered	-0.027**
	(0.012)
Product Category: Services	-0.578*
	(0.312)
Product Category: Media and Entertainment	0.003
	(0.332)
Company Location: Southern U.S. Region	0.335
	(0.253)
Constant	0.502
	(0.331)
Observations	342
Log Likelihood	-225.983
Akaike Inf. Crit.	465.965
Note:	*p<0.1; **p<0.05; ***p<0.01

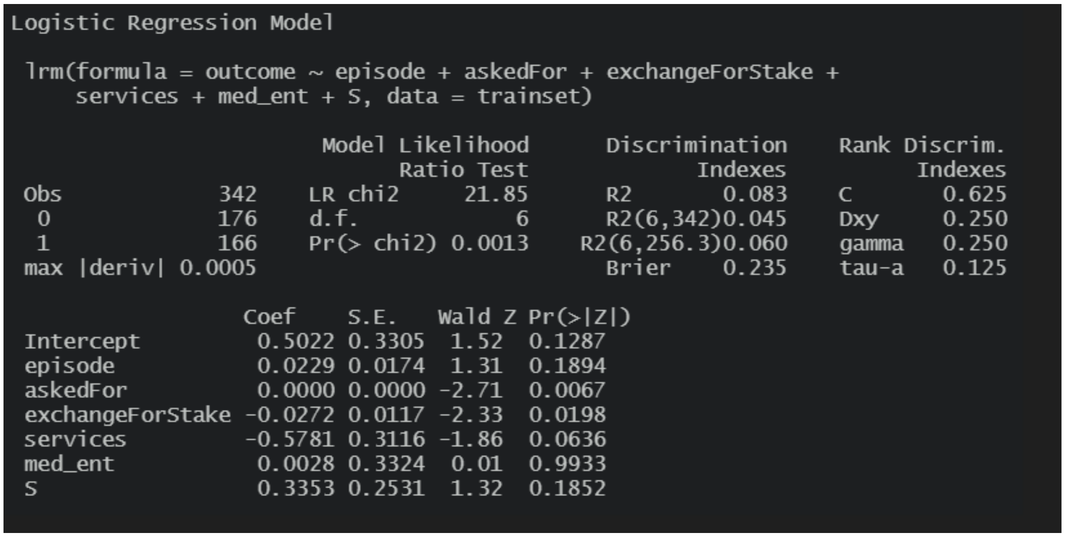
Figure 10: Logistic Regression Output





```
                                var   rel.inf
episode                     episode 37.988998
askedFor                   askedFor 32.478104
exchangeForStake exchangeForStake 20.067175
S                                 S  4.099598
services                   services  3.258399
med_ent                     med_ent  2.107726
```

Figure 11: GBM Relative Influence Plot (top) and Values (bottom)



```
Logistic Regression Model

 lrm(formula = outcome ~ episode + askedFor + exchangeForStake +
     services + med_ent + S, data = trainset)

                        Model Likelihood      Discrimination    Rank Discrim.
                              Ratio Test             Indexes          Indexes
 Obs            342    LR chi2     21.85      R2       0.083    C       0.625
  0             176    d.f.            6      R2(6,342)0.045    Dxy     0.250
  1             166    Pr(> chi2) 0.0013    R2(6,256.3)0.060    gamma   0.250
 max |deriv| 0.0005                           Brier    0.235    tau-a   0.125

                  Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept         0.5022 0.3305  1.52  0.1287
 episode           0.0229 0.0174  1.31  0.1894
 askedFor          0.0000 0.0000 -2.71  0.0067
 exchangeForStake -0.0272 0.0117 -2.33  0.0198
 services         -0.5781 0.3116 -1.86  0.0636
 med_ent           0.0028 0.3324  0.01  0.9933
 S                 0.3353 0.2531  1.32  0.1852
```

Figure 12: Logistic Regression LRM Output