FINAL PROJECT

MULTIVARIATE STATISTICS

MMA - MCGILL UNIVERSITY

What is
Shark Tank ?

Shark Tank is an American Business reality show released in 2009 on ABC. This program show entrepreneurs making business presentations to five investors called "Sharks" who based on the presentation decide whether to invest in the company.
As of 2020, the US franchise has had 11 seasons with 222 episodes. 894 pitches, 499 deals, and $143m worth of invested capital.

In this project, I will be analyzing 6 seasons to figure out how to succeed in Shark Tank. To find what describes a successful idea, I will be using different statistical tools to show the main characteristics of the best projects. Moreover, I will use machine learning algorithms to be able to tell you if your idea will be successful or not.

INTRODUCTION

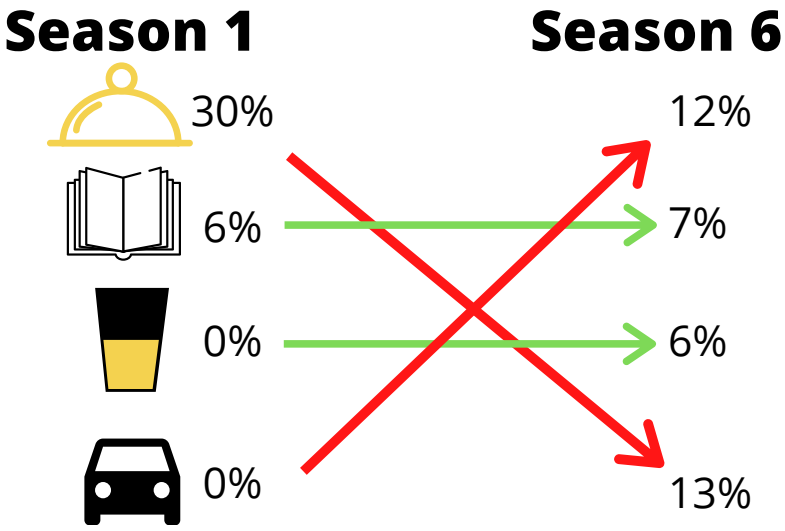
DATA DESCRIPTION

Categories Overview
Shark Tank has 50 categories which the most important in terms of money invested are Specialty Food, Automotive, and Education. It is important to know that the businesses categorized as Specialty Food are the ones that are valued the lowest.

Which category gets the most rejected?
Online services are the category that has the most fail deals, followed by Personal Care and Cosmetics. Novelties and Baby and Child Care.

Which categories are the Sharks investing in the most?

Categories	RH	MC	LG	KOL	DJ	BC
Specialty Food	5%	17%	9%	21%	1%	33%
Automotive	45%	5%	4%	0%	6%	0%
Education	27%	0%	0%	0%	0%	4%
Alcoholic Beverages	0%	4%	0%	33%	0%	0%
Entertainment	0%	24%	0%	0%	2%	0%



Trending Categories
Right now Sharks are investing in businesses focused on Automotive, Education and Alcoholic Beverages. This shift started during season 5 and keeps growing.
If you are a Specialty Food Business, you better have a very good pitch as this category has been decreasing.

51%
Average Deal Success Rate

17%
Average Exchange for Stake

MODEL SELECTION AND METHODOLOGY

To be able to predict if an entrepreneur will close a deal the following methodology has been approached to develop a machine learning model that will determine the probability of success.

Methodology

Data Preprocessing

In this step, I deleted variables like the number of the episode, the name of the business, website, description. As these features do not add value to the model.

Then, I created dummy variables for the categorical features such as State and Categories.

OutlierTest

Outliers can drastically change the accuracy of every model, therefore I performed a Bonferroni test to evaluate which data points were outliers.

The results of the test found that only one point was considered an outlier.

Next, this point was removed from the data frame that was used for the feature selection.

Feature Selection

Feature Selection is a core activity to find the variables that best predict if someone is going to close a deal or not.

Two algorithms were used to evaluate these variables: Random Forest and Gradient Boosting Model.

In the random forest, I took into consideration the variables that if removed would affect the accuracy of the model. Then using the Gradient Boosting Method (GBM), I evaluated all the variables and I deleted the ones that the GBM showed me had no relevance.

Finally, taking into account both results I evaluated which features increased the overall accuracy of the model.

Model Selection

Classification

To select the model I compared the Random Forest Classification and the Gradient Boosting Classification. I found that the accuracy of the GBM was approximately 20% more than the Random Forest.

Clustering

For the clustering, the K-means algorithm was chosen to perform the analysis. In this model, I dropped all the categorical variables and I kept the numerical only. Then I scaled all the data to get the best results.

The results of the model are the .

Results

Classification

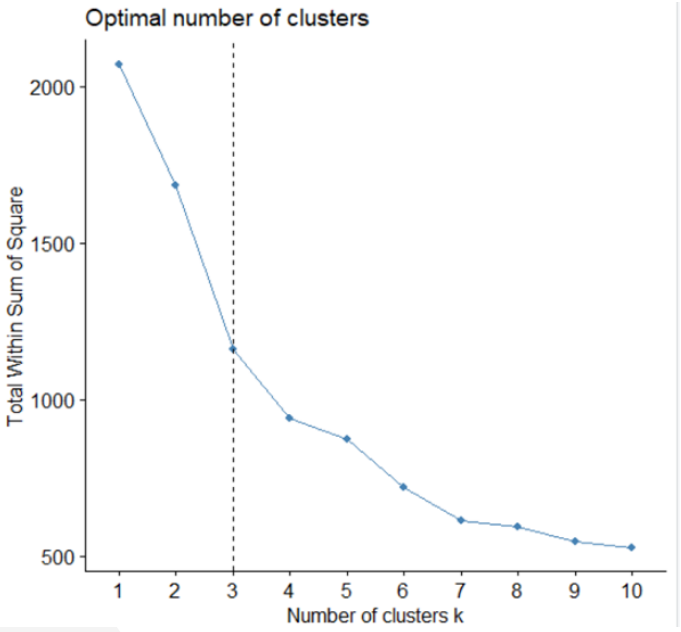
Using the Gradient Boosting Method. I achieved an accuracy score of 89% based on variables as quantity to be asked, the value of the company, category, how much % are you willing to give up, length of the name of your business, US state (founding), and if you have a website.

The following factors represent 80% of the likelihood to close a successfull deal:

-Valuation 25%
-Length title of the business 23%
-Asked For 18%
-ExchangeForStake 11%
-Business from California 3%
-Size of team entrepreneurs 2%

Clustering

To perform the clustering the K-means algorithwas used and to find the optimal number of cluster the "Elbow method" was used. I found that 3 clusters were the optimal for this dataset.



Furthermore, I found the characteristics of these 3 clusters:

Cluster	# Members	Asked For	% Exchange	Valuation	Length Name
1	1	179k	17	1,602k	12
2	1	2,515k	15	1,896k	11
3	2	198k	16	1,854k	12

RESULTS

Shark Tank is a contest where a lot of factors are in the game. Luckily in this analysis, we have found which ones determine if you are likely to get the investment of your life or not.

always remember the following recommendation if you are going to be part of this contest:

Businesses from California, with a high exchange for the stake, names composed of two words, highly valuated, and teams made up by 2+ are more likely to be successful.

Sink or Swim

CONCLUSION