SharkTank: Deal or No Deal?

By: Albert Aung, Alexander Jensen, Benjamin Xue, & Maxwell Fang

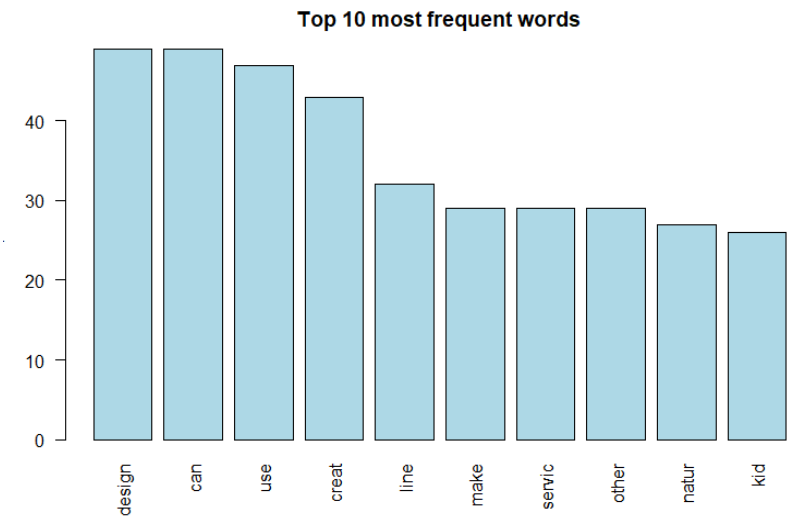
Introduction:

Shark Tank is a TV show that allows budding entrepreneurs to get a chance to turn their dreams into a reality. By presenting their ideas to the shark tank - five titans of the industry, they try to convince them to invest money in their idea.

One question is how exactly does the shark tank determine whether or not they should invest in someone's product. There are many factors that come into play when assessing someone's presentation. The background of the judges could come into play and other factors that involve risk also determine whether or not a product is worthy. However, out of all the factors, the most important one is the pitch. Therefore, we will be analyzing what exactly a pitch requires to be successful.

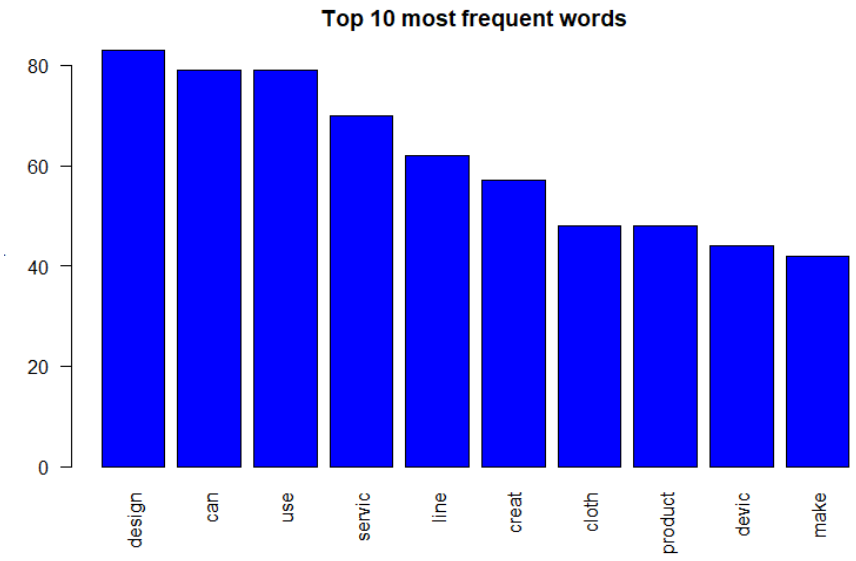
There are several variables that we took into consideration. One of the variables we explored was finding the frequency of words that appeared in deals that succeeded and ones that did not.

Word Visualization for Successful Offers:



Analysis: The most common words were "use", "design", "can", and "create". These words do not necessarily have a strong positive connotation, but are just general terms that people use to describe their product. However, there is one particular detail that is notable is that the appeal to kids might be important. By helping out kids or even making their lives better could potentially be a contributing factor.

Word Visualizations for Unsuccessful Offers:



Analysis: Similar to the top words in successful offers, words such as “design”, “can”, and “use” were very common. This signifies that these common words do not affect whether or not a pitch was offered a deal or not. It is crucial to point out that the word cloth appears quite often. This may show that many of the people who did not get offers were products that have cloth. Once again, the visualizations exhibit that the connotation of words does not affect the outcome.

Data cleaning/Pre-processing

In the dataset provided, we are given five properties to work with:

1) Episode in which the business was featured in
2) The company name or product name
3) The description of the business pitches for each product/company
4) The status of whether the deal was reached between the investors and the entrepreneur (Label)
5) The sharks which closed the deal if the deal was closed (Label)

The initial dataset overall was pretty clean but included some noise in the data, more specifically in the “Pitched_Business_Desc” section which stood for the description for the business pitches. We thought we could make it better to process the data if we got rid of non-alphanumeric words and stop words which are common words used in our everyday conversations (e.g. “he”, “she”, “the”, “because”, “but”, etc…). To improve fluency when running the machine models later, we took the liberty to transform all the words in the “Pitched_Business_Desc” column into lowercase letters and words and got rid of symbols such as “/”, “@” and “\\|” which were present in the data. Afterwards, to aid with visualization, we divided the training dataset into two separate datasets: a training dataset only including offers and a training dataset only including non-offers. This helped us to draw up visualizations of the

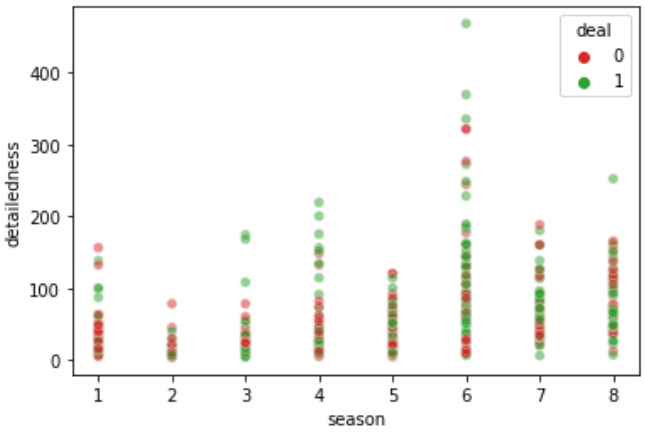
data which has been already addressed in the report. To run our models, we further looked to remove all types of punctuations from “Pitched_Business_Desc” and replaced symbols such and “&” with the word “and” to improve data readability.

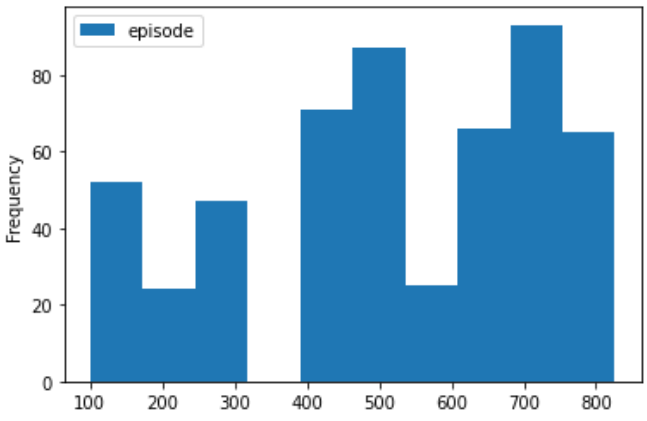
One peculiarity with the episode count is that the episode number isn’t successive, which indicated that it wasn’t the exact number of the episode, but instead the episode within the given season, in which the first digit represented the season, and the rest of the digits represented the episode number within the respective season.

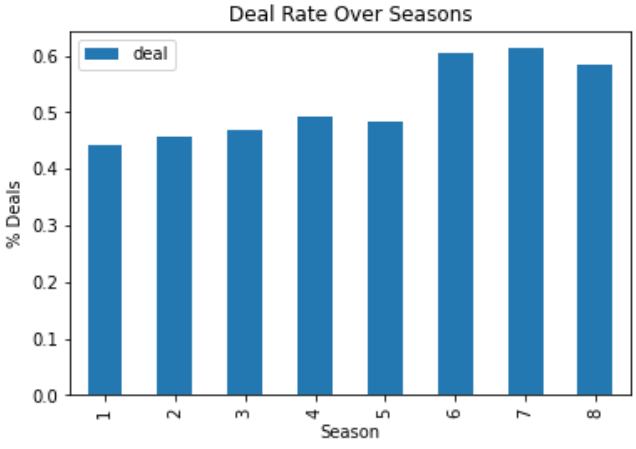
A feature that we looked into was lengthiness, usually referred to as detailedness in the code. The word count from the company name and description are multiplied together in order to quantify how detailed the description of the company itself is.

Visualizations

One visualization we looked into was the correlation with lengthiness (also called detailedness in our code) and deal rate. Our visualization gave us the insight that the more length a company has, the more likely they may get a deal. This relation isn’t as clear in pitches with less lengthiness. Another trend we looked at was the correlation between seasons and the deal rate. In our visualization, we saw that there was a slight hike in deal rates after season 5. A possible explanation is the increasing popularity of the show, thus possibly increasing the quality of the pitches due to higher demand. Another possible explanation could be that to increase TV show ratings, the screening process of the show made it tougher for lesser known and less popular products.





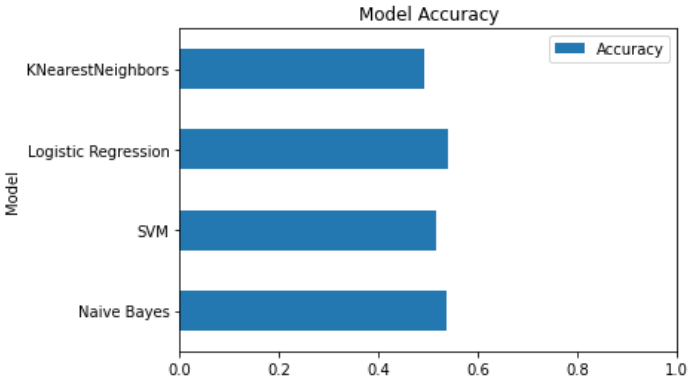


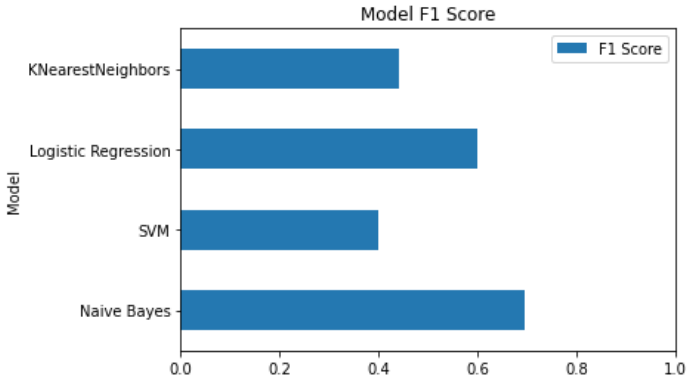
Modeling & Results

Models that we explored: Naive Bayes, SVM, Logistic Regression, and KNN

We used TF-IDF to vectorize both the company name and company description for our features. Alongside these vectors, we also included the detailedness feature described in the visualization section, which helped boost the accuracy difference between the models. After averaging out the accuracy and f1 scores of 100 different training sets, we got the following results:

	Accuracy	F1 Score
Naive Bayes	0.536792	0.694986
SVM	0.514623	0.399911
Logistic Regression	0.538208	0.599842
KNearestNeighbors	0.490472	0.439574





From the visualizations, we can see that there isn't much difference in accuracy, but there is a huge difference in f1 score between the models. Naive Bayes has good accuracy and competes with Logistic Regression in terms of accuracy, but Naive Bayes has a marginal increase in f1 score, which offsets the slight inaccuracy relative to Logistic Regression. Because of this decision, we decided to use Naive Bayes for our final model.

Conclusion

Using our model, we predicted 173 companies out of 176 companies would strike a deal, and 3 companies out of 176 companies that would not reach a deal from the testing dataset. One observation of the 3 companies that failed to reach a deal was that the descriptions provided were generally short and lacked detail. For example, we predicted that the company “iReTron” would not reach a deal, and the description associated with the company was “an electronics recycling service . electronics recycling service”. This description is shorter than most of the other company descriptions and also did not contain any of the 10 most frequently used words in descriptions that landed a deal. Similarly, “Origaudio” was another company that we predicted to not reach a deal, with their description being “innovative portable speakers . innovative portable speakers”. This description also does not contain any of the 10 most frequently used words in descriptions that landed a deal, and there are only 3 unique words in the description, which is significantly less than most other companies.

Having included lengthiness (described as detailedness in the code) as one of the factors in determining the success of a pitch, it would come to no surprise that descriptions which weren’t lengthy and have less significant information in them in terms of their products would not perform as well as other pitches which contained more relevant information of their products rather than just offering a quick two to three words summary of the product. We could agree that there was definitely biases in our results since our models were based on the lengthiness feature that we added so pitches could be penalized for being concise and effective. To combat this, we could add more features and assign grading scales to the added features so that more significant features will play a bigger role in predicting the success of a pitch.

In the future, we can look to add more features to help us extract the most out of the given dataset since it is evident in our data that improvements can be made in terms of model accuracy and f-1 score. One feature that comes to mind are the sharks’ areas of interest which could be one hot encoded to add more depth to the data. To dive deeper into the shark's deterministic ability, we could also explore which sharks are more willing to link up with one another for a deal. From watching a few episodes of Shark Tank, it can be quite visible how certain sharks are more biased towards forming teams with certain other sharks when working towards a deal and this feature would help capture that to offset the biases in the data. Another feature we can look into is whether or not a product is connected to success or will perform better than other products that already exist in the market.