Coffee Selection

White Paper

Kinnick Fox

Bellevue University – DSC680

9/13/2023

In this paper, I will be going over the Coffee Selection project, its execution, and an analysis of the results and viability. Topics of deployment and ethics will also be covered.

# Business Problem

New businesses that deal with the sale of coffee need to understand which selection will be the most reliable for sustaining a growing business. Partnering with roasters or growers that produce coffee with less mass appeal could lead to a decrease in customer base.

# Background

Coffee is an everyday luxury for many with “Two billion cups of coffee [being] consumed every single day, with 1 billion coffee drinkers worldwide” (O'Connor, 2023). There are differing methods in growing, processing, ageing, roasting, grinding, and preparing that can all change the coffee drinking experience for the consumer. The goal of this project is to identify key features that hold influence over the enjoyability of coffee.

# Data Explanation

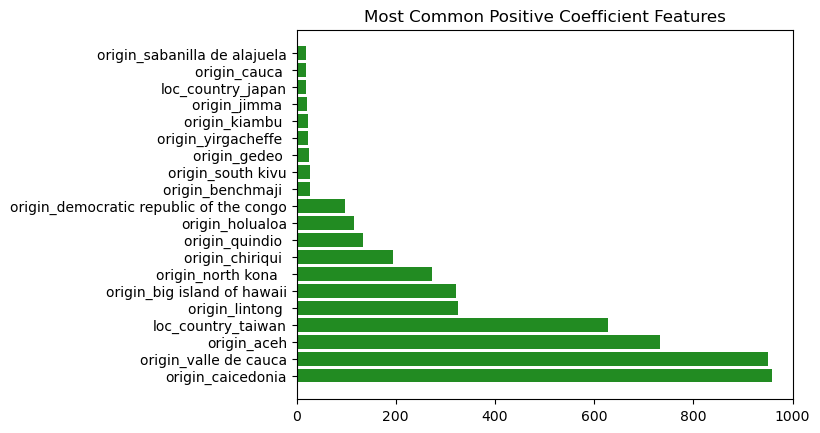
The data used for this project is public domain and is hosted on Kaggle. Within the dataset, features pertaining to name, roaster, origin, price, rating/reviews, and description. Roaster and roast style data proved to be not very useful so it was excluded form the model. Consideration for using sentiment analysis on the description data was made and ultimately not pursued due to words that could be considered positive for a coffee description, such as bitterness and astringency, could be seen as negative in a sentiment analysis. No major transformations needed to occur within the data before modeling, the main issues were found with differing spellings for the same origins. After attempting to isolate the correct origin name with code, it was decided that these changes would be made by hand within the CSV file. Dummy variables were then created or all object data and the two origin columns were combined as to not dilute the presence of a specific origin that appeared in both features. Finally, outliers with fewer than ten occurrences were removed.

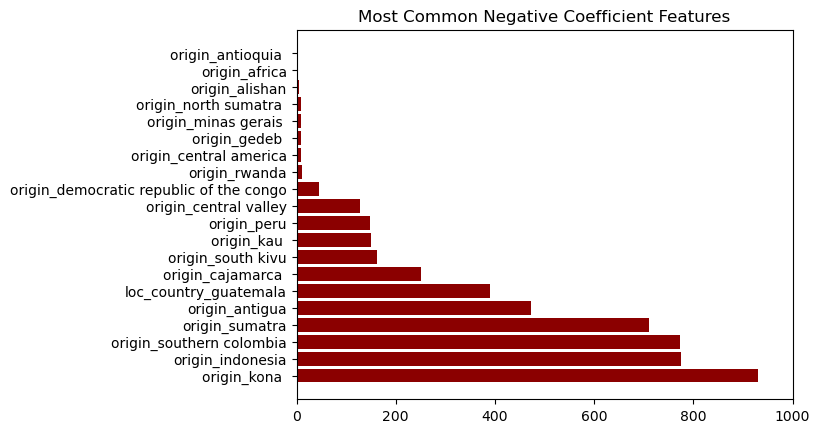
# Method

Linear regression was the method used for this project. This was chosen due to the target variable, rating, being an integer. A train/test split was also included to assess accuracy of the resulting model.

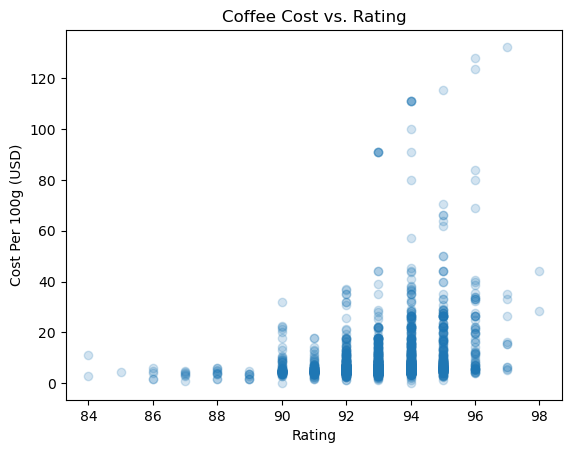
# Analysis

Initial analysis of the model’s accuracy concluded that this would be a poor prediction model due to it possessing a 15-30% accuracy depending of train/test split. Moving on from this, it was noticed that the same or similar variables would have the highest and lowest correlation coefficients. The model was created and tested 1000 times in order to record features with the highest and lowest correlation coefficients. Of the eighty-six features, two were within the top five highest correlation coefficients 95% of the time and one was in the top five lowest correlation coefficients 93% of the time.





Thankfully, cost was not a major factor in any correlation coefficients although it does appear that some of the more expensive products do trend towards a higher rating.



# Conclusion

Analysis shows that coffee that originates from Valle De Cauca or Caicedonia has an increased likelihood of receiving a better rating. A similar argument could be made for coffee originating from Aceh although the frequency of this feature appearing in the highest correlation coefficient was lower at 73%. Analysis also shows that coffee originating from Kona as a decreased likelihood of receiving a better rating. Coffee from Columbia and Indonesia may have a similar effect although the frequency of these features appearing in the lowest correlation coefficient was 77%.

# Assumptions

The major assumption that is being made for this project is that the ratings included within the dataset are impartial and universally accepted. Another assumption is that coffee from varying growing regions is included for each origin as opposed to all coffee coming from a single farm.

# Limitations

As mentioned previously, this model is limited to reference use only as it performs poorly when judged by accuracy of prediction. That is not to say that the previous conclusion is not viable as the trends found across the model pool are undeniable.

# Challenges

The main challenge for this project was the data normalization within object features. This soaked up most of the time but was required to ensure accuracy of the model. The difficulty in implementing an automating cleaning process for this data also added to this challenge.

# Future Uses

The intended use of these results was to inform business owners and ensure they plan to stock coffee that is enjoyable for a broad range of customers. The results indicate that the model is still applicable for the desired use case and can provide insight to business owners.

# Recommendations

Based on the results, I would recommend businesses purchase and sell coffee from Valle De Cauca or Caicedonia if they do not already. It is also recommended to avoid coffee from Kona unless they are willing to accept a little more risk or have plenty of other selections and wish to satisfy a niche consumer base.

# Implementation Plan

These results lend themselves to a business planning meeting more than an implementation into everyday operations. That said, data collected by businesses could be used to grow this dataset and stay up to date on possible trends.

# Ethical Assessment

All business decisions based on insights derived from data science have the potential to take possible business from a group or company. In this case, it could hurt the coffee production of Kona or any other origin that is not Valle De Cauca or Caicedonia. This exact ramifications of this are unknown but hurting the Kona coffee industry, for instance, could lead to turmoil for the workers of that industry.

# References

O'Connor, A. (2023, July 14). *How Many People In The World Drink Coffee?* Retrieved from Full Coffee Roast: https://fullcoffeeroast.com/how-many-people-in-the-world-drink-coffee

Schmoyote. (2023). *Coffee Review Dataset*. Retrieved from Kaggle: https://www.kaggle.com/datasets/schmoyote/coffee-reviews-dataset

Appendix

The number of people in the world that drink coffee is roughly 1 billion. Two billion cups of coffee are consumed every single day, with 1 billion coffee drinkers worldwide. Interestingly enough, around 3.7 billion cups of tea are consumed daily, with approximately 2 billion people drinking tea every morning. (O'Connor, 2023)