Jeff Orcutt

DS 740

Final Project

Introduction: The Coffee Quality Institute is a non-profit organization that provides education and certification to coffee professionals that work in the production of coffee beans. They’ve developed an education process that helps objectively score coffee based on physical traits and aspects of taste, similar to a sommelier’s process in evaluating wine, providing downstream professionals a clearer picture of strengths and weaknesses of coffee beans. I approached this dataset with two questions in mind: Given information from trained graders, (1) to what extent is it possible to predict the country of origin of a coffee and (2) are there any useful clusters to provide suggestions of similar, or dissimilar, coffees to consumers.

Dataset: I obtained my data from James LeDoux’s [github](https://github.com/jldbc/coffee-quality-database/tree/master/data)[[1]](#footnote-1) where he has collected on 28 Robusta and 1318 Arabica coffee beans as tested by the Coffee Quality Institute with information on the bean’s provenance and flavor profile totaling 44 columns, 33 of which are meta data but eleven are detailing the flavor profile. This data is the result of data crawling in the review pages of the CQI’s website from 2018. There is no data dictionary provided – some assumptions will be made on the interpretation side of things like “acidity” and “sweetness” referring to qualified raters’ opinions rather than some sort of chemical testing. This dataset is not to be interpreted as a wholistic view of the coffee industry, rather a snapshot of available data from 2018. An updated dataset is not possible as the Coffee Quality Institute has made their forums private. Additionally, no assumptions are made about the consumption preferences of consumers, rather the data is producer centric and unfortunately limited to the data that exists.

Dataset Preparation: There are a large number of missing values in several columns. I did mean imputation for numeric variables and median imputation for categorical variables. The Country-of-Origin response variable was consolidated from 37 values down to 19 values through consolidating all countries that had lower than 20 observations in this dataset to regional areas. This was done to enable cross validation in caret, some of the outcome classes would have no representation in the sample. Only one record was removed from the dataset for not having the response variable identifiable. The harvest year column had values that were in many different formats and some with only months without years, but provided nearly a significant lift in accuracy in the overall model. Values that had a range of values were reduced to the earliest year in the range. Values that had no clear harvest year were imputed with the median. Lastly, a number of columns were removed that had data about the manufacturer or farm name or location, their certification numbers, information about the coffee bag’s lot number or weight, and observation id number.

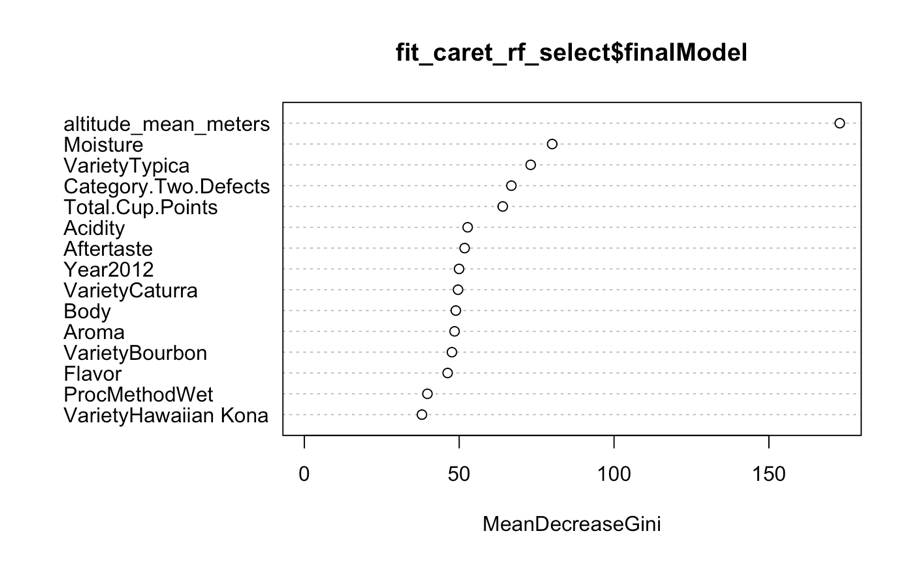
Research question 1: Given information from trained graders, to what extent is it possible to predict the country of origin of a coffee?

I approached this question with two models, one that examined all of the remaining columns, and one that only looks at the numerical testing results, mean harvest altitude, and the variety of coffee, information you’d find in a typical coffee shop’s coffee display, in random forest models. This method uses an ensemble of decision trees to make a prediction of the value of the response variable, which in this case is the country of origin of the coffee. One of the things that makes random forests a reasonably stable model is a concept known as bagging, which subsets the data for each tree modeled. Additionally, the model can be tuned in a number of ways: the number of trees tried, the number of random predictors to consider in each tree, and the minimum number of observations that will result in a new node split. The hyperparameters that were made available to be tuned in the model selection process for this modeling technique were the number of predictors to be tried in each tree – from one to all predictors, the minimum node size - from one to ten, and the number of trees in each forest’s model – 100, 250, 500, 1000, or 2000.

When evaluating which of the two modeling techniques provides the best accuracy in predicting the origination country of the coffee, the models go through a process of cross validation. This process of cross validation randomly selects the training and validation datasets from a given dataset through the Caret package in R resulting in a comparison that allows the selection of the best model. The best models are compared head-to-head to see which is the most accurate after the hyperparameters are optimized. When combined a further segmentation of the dataset known as double cross-validation, this creates a very consistent way of assessing the performance of the selected model.

When building the models, I looked at two sets of predictors from the dataset – one in which all predictors were used, and one in which only the numerical tasting values, mean growing altitude and variety of coffee was used. In my analysis, the random forest model was shown to be the best model in the process described above with all of the predictors included. The model was tuned to have the following characteristics: three variables were selected for each tree, 500 trees per forest, and a minimum node size of one. The selected model showed an accuracy of 75.3%, which performed better than the estimated 73.3% by the assessed double cross validation process and decisively better than the naïve 17.7% no-information-rate. This model produces a wide array of classification errors ranging from Colombia with 11.4% misclassified of its 183 values misclassified to Indonesia with 80.0% of its 20 values being misclassified. This suggests that within this dataset some countries produce very identifiable coffee, while others may not be as easily identifiable.

The top three most important variables are the mean altitude, moisture content, whether the coffee was of typica variety. The chart below shows the top fifteen predictors in the model sorted in descending importance.



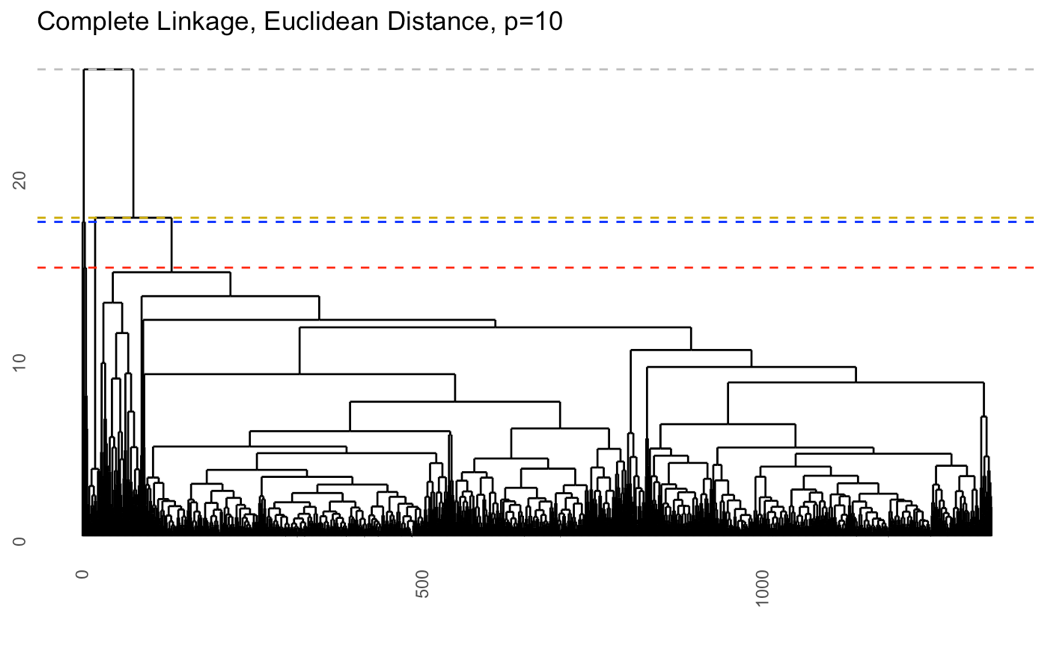
Overall, the random forest model performed much better than I expected it would in predicting the country of origin considering how many countries are in the response variable. I also found it very interesting that the height at which the coffee is grown is indicative of the country to such a strong extent.

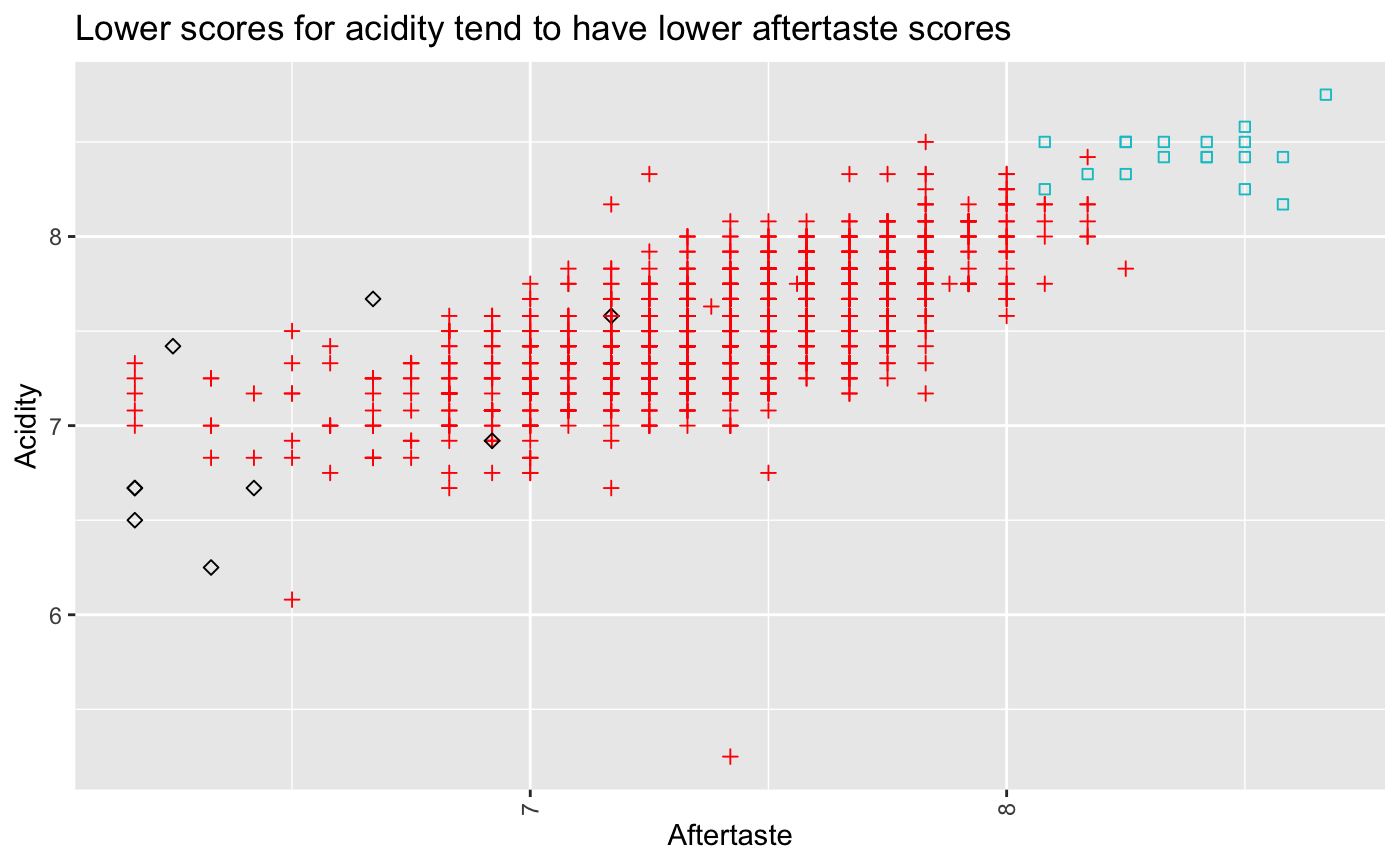
Research Question 2: Are there any meaningful clusters in this dataset to provide suggestions for similar, or dissimilar, coffee beans?

I approached this question using hierarchical clustering with single, complete, and average linkage methods to a subset of the dataset that includes only the tasting scores from the trained tasters. The goal is of clustering is to group the dataset into similar groups in order to provide suggestions for similar or dissimilar coffees to consumers. The difference between the methods is how similarity clusters are measured and thus clustered. I chose the complete linkage method for my final clustering method. When looking for the optimal number of clusters, I used a package called NbClust that looks at the scaled dataset, creates the Hubert and D indices, then looks for the point at which the most significant decline/increase, in difference in slope from the previous point to the next point occurs. The package then applies a majority rule to establish which cluster to suggest is best. I tested cluster numbers from 2 to 30 and was presented with the following two plots. You can see in the first panel of the two plots where the plot first starts to have diminishing returns is located at the 3-cluster level, which is known as the significant knee.

|  |  |
| --- | --- |
|  |  |

After establishing that 3 was an appropriate number of clusters for this dataset, I created a dendrogram to visually depict the clustering decisions of the clustering model. The dendrogram below depicts the three clusters of sizes 18, 1310, and 9. As you can see below, the dendrogram produced a tree that has many observations in the third branch, with the lines indicating the cluster cut size, starting from the top of the diagram with 1 cluster to the red line representing 4 clusters.





Looking at the data another way, I found that showing the relationship between Acidity and Aftertaste scores by their hierarchical group showed some separation, as shown in the chart above. This indicates that while not wholly responsible for the clustering choices, they’re at least partially related. If you were to apply this information to a customer who has previously enjoyed coffees with a higher, more pleasing, aftertaste, one could recommend coffees in the 18 blue square coffees. Other relationships and clustering patterns likely exist with the clustering dataset pairs. Given the size of the largest cluster, 97.98% of the overall dataset, and small number of clusters presents difficulties in differentiating this dataset’s coffee selections without considering other predictor variables.

Conclusion: In conclusion, this study has demonstrated the potential of using machine learning techniques, specifically random forest models, to predict the country of origin of coffee beans based on their physical traits and taste profiles. The model achieved an accuracy of 75.3%, significantly outperforming the no-information-rate. This suggests that certain characteristics, such as the mean altitude at which the coffee is grown, the moisture content, and the variety of the coffee, are strongly indicative of the coffee's country of origin.

However, the model's performance varied across countries, suggesting that some countries produce more identifiable coffee than others. Future research could explore this variation in more detail, perhaps by incorporating additional data or using different modeling techniques.

The study also explored the use of hierarchical clustering to group similar coffees together. While the clustering did reveal some patterns, the large size of the largest cluster and the small number of clusters overall suggest that the dataset's coffee selections are not easily differentiated based on the variables considered in this study. Future research could explore other variables or clustering techniques to better differentiate between coffees.

Overall, this study has provided valuable insights into the characteristics that distinguish coffees from different countries and the potential for using these characteristics to recommend similar coffees to consumers. However, more research is needed to fully realize this potential and to overcome the limitations identified in this study.

1. LeDoux, J. (2018, January). *JLDBC/coffee-quality-database: Building The Coffee Quality Institute Database*. GitHub. https://github.com/jldbc/coffee-quality-database/ [↑](#footnote-ref-1)