**U.S. COFFEE PRODUCTION**

Daniel Kaupa

Dr. Gregory Tanner

DATA 318 : Data Mining

Final Project Report

4/30/20

**I. Introduction**

How do you prefer your morning cup of coffee? Do you take it black, with cream, with sugar or with cream and sugar? Do you brew it in a coffee machine, use a pour-over, or a French press? However you get your morning fix of caffeine, it all comes back to the beans. As the world’s biggest consumer of coffee, America produces a surprisingly minimal amount of the coffee that it drinks (Haqqi). In fact, 90% of the coffee that America consumes is imported. This is perhaps largely due to climate limitations combined with competing crops and livestock already established in areas where coffee could be grown. As of February, 2019, Hawaii and California were the only states in the United States with well-established coffee growing businesses (Misachi). Even with the climate restrictions, however, it is interesting to consider the production of coffee in the United States.

The purpose of this project was to attempt to gain a better understanding of the production and price of coffee in the United States. Additionally, this project attempted to create a model that would be able to accurately model and predict the price of coffee in the United States based on the prices of other fruit and nut variety agricultural products (sometimes requiring similar climate requirements to grow as coffee). Personally, this is an interesting topic for a two reasons. First, as an avid coffee drinker myself, I found it interesting to learn more about the trade and production dynamics of coffee. Second, a good friend of mine is planning on going into the international coffee trade and would enjoy hearing more about the United States coffee trade. In the model I created to predict the price of coffee, data on the price of five fruits were used. These variables were the price of apples, cranberries, grapefruit, sweet cherries, and tangerines. As stated above, the importance of these variables could stem from the climates in which they are best produced and their similarities to coffee. However, cranberries, apples, and cherries are grown all around the United States in climates vastly different from the more tropic weather that coffee plants grow. This is interesting (as detailed later in this report) because these fruits are highly correlated with coffee Any particular changes in demand and thus price of these products could be correlated to price fluctuations in coffee, but is unknown.

**II. Data Source**

The dataset used for this project was found on the United States Department of Agriculture’s website through the National Agriculture Statistics Service. This database contains official published aggregate estimates related to U.S. agricultural production, collecting data through sample surveys as well as the census. For more information this service, one can visit the website at <https://quickstats.nass.usda.gov/> (“USDA/NASS QuickStats Ad-Hoc Query Tool”). Since this dataset is so large, there is an online interface where users can select A screenshot of a computer screen

Description automatically generatedsubsets of the data to extract as needed. For this project, I extracted survey information for the price received for fruit and tree nut products. To simplify analysis, I used data for the United States as a whole rather than on a state or county level. Although the exact details of how the data was collected were not explicitly documented on the website, it can be assumed that the data is reliable as it comes from a government agency and is expected to be accurate. A close up of a map

Description automatically generatedWithout knowing more about how the data was collected, it is difficult to speculate on possible sources of error. One point of confusion, however, may arise when one considers the production versus the processing of food items. Especially with coffee, which although the United States doesn’t grow a lot of, it certainly process a lot of it, the exact details of the data is unclear. For the purposes of this project, however, it was assumed that production was in reference to strictly growth of crops and not processing.

IMAGE 2

**III. Exploratory Analysis**

The predictor variables that were used in the model were the result of several filtering processes/rounds as well as their correlation to the target variable. In order to have a substantial number of years worth of records, many possible variables were ruled out by default. Assumed to be attributed to the boom of the information era, the amount of data available before the 1990’s was rather small in comparison to the amount of data by the early 2000’s. The predictors of price received for apples, cranberries, grapefruit, cherries (sweet), and tangerines were chosen from a subset of nine possible predictors. As shown in image 1, these variables were all more than 40% (positively or negatively) correlated with the target variable.

Image 2 above shows the correlation between the price received for apples and the price received for coffee. As is expected given the calculated correlation values, the graph shows a respectable degree of correlation between the two variables, suggesting that apples will be an important part of any model.

IMAGE 1

On the opposite end of the spectrum, the image 3 illustrates the correlation between the price received for oranges and the price received for coffee. Again, as is expected given the calculated correlation values, the graph shows next to no correlation (positive or negative). This suggests that price received for oranges will not be an important part of any model created. In the final model that was created, one can clearly see the influence of the predictors in relation to the target variable. In image 4, a plot of the decision tree model used shows the role of each predictor used in A screenshot of a social media post

Description automatically generatedthe model to determine values for the target variable.

IMAGE 4

**IV. Model Discussion & Interpretation**

A close up of a map

Description automatically generatedSince the problem of predicting the price received for coffee is a quantitative and not qualitative problem, regression models were best suited to this. When beginning to create models, I first began by creating a model including all variables as predictors of price received for coffee. I created a model in this fashion for each of the types of models I had chosen to use: k-nearest-neighbor (knn) regression models, linear regression models, decision tree models, and random forest models. In the second phase of creating models, I then used the correlation data to create models (of each type) using one to five of the variables most correlated with price received for coffee. It is important to note that in the consideration of the top five most correlated variables, the year variable was omitted from consideration. This was done because the high correlation could have been largely caused by inflation over time as well as the growing popularity of coffee. While interesting trends to observe, this could have drawn the data away from the investigation into the dynamics between price received for coffee relative to other agricultural crops. For the linear regression models, there was the additional consideration of collinearity as well as near-zero-variance predictors, and so I built additional linear models to accommodate these concerns. For the knn regression models, I also created models with traditional cross validation as well as leave-one-out-cross-validation (LOOCV). This was done in case the small amount of data available proved to be a problem as LOOCV can sometimes be an aid in these cases. The reason that the top one through five most correlated variables (again, excluding year) were chosen was somewhat arbitrary. The top five most correlated variables all fell above a 0.40 correlation with the price of coffee, and while merely above 0.40 is certainly not an ideal low range to have in a model, it was thought to be the best balance between high correlation and number of predictor variables.

IMAGE 3

A close up of a map

Description automatically generatedAlthough variants of linear regression were considered, such as ridge regression or smoothing splines, they were not pursued. As variants of linear regression, it was assumed that they would at best improve the models slightly. This would’ve perhaps been worth considering, but as the general linear regression model performed the worst of all the models, and the decision tree model performed so much better than linear regression, it was thought to be unnecessary.

A close up of a map

Description automatically generatedRegarding the performance of the models, the decision tree models overall produced the highest R-squared values, as well as some of the lowest low root mean squared error (RMSE) and mean average error (MAE) values. Random forest models were not far behind the decision tree models in their performance, but were slightly worse. It was surprising that the KNN regression and linear regression models performed so poorly to the decision tree and random forest models. It was particularly surprising that KNN outperformed the linear regression models. Ultimately, the model named ‘TREEModeltrim5’ performed the best, producing an R-squared of 0.7895, an RMSE of 544.04, and an MAE of 444.08. For reference, the best linear model produced an R-squared of 0.5964, an RMSE of 709.77, and an MAE of 547.02, the best knn model produced an R-squared of 0.6729, an RMSE of 637.90, and an MAE of 494.75, and the best random forest model produced an R-squared of 0.7313, an RMSE of 604.58, and an MAE of 464.20. As illustrated in image 5 and 6, the decision tree model appears to perform quite well on the test set.

IMAGE 6

IMAGE 5

Regarding the typical error, it is unknown what the typical error in the context of this problem would be. In a expert’s analysis of this problem, more information and data regarding the production of coffee worldwide as well as consumption trends would likely be known and included in a model. Additionally, data regarding climate conditions and perhaps setbacks due to extraordinary events would likely be included. Overall, an expert’s model would likely be far more accurate, be able to explain levels of inaccuracy, or both. If given the opportunity to enhance the model, I would look into including data on the climate conditions in which coffee can grow in the United States. Also, if it were possible to include data about more crops, that would be ideal. After such a strong filtering process, there was very little left to work with when building the model. Including data on more crops would hopefully widen the range of possible predictors and allow more significant predictors to arise.

**V. Bibliography**

Foster, James. “Caffeine (Coffee) Consumption By Country.” *Caffeine Informer*, 15 Oct. 2018, [www.caffeineinformer.com/caffeine-what-the-world-drinks](http://www.caffeineinformer.com/caffeine-what-the-world-drinks).

Haqqi, Ty. “10 Countries with the Highest Coffee Consumption in the World.” *Yahoo! Finance*, Yahoo!, 13 Dec. 2019, finance.yahoo.com/news/10-countries-highest-coffee-consumption-045444069.html.

Kenner, Bart, and Jaclyn Kramer. “Fruit and Tree Nut Yearbook Tables.” *USDA ERS - Fruit and Tree Nut Yearbook Tables*, United States Department of Agriculture, 8 Jan. 2020, [www.ers.usda.gov/data-products/fruit-and-tree-nut-data/fruit-and-tree-nut-yearbook-tables/#All%20Yearbook%20tables](http://www.ers.usda.gov/data-products/fruit-and-tree-nut-data/fruit-and-tree-nut-yearbook-tables/#All%20Yearbook%20tables).

Kunst, Alexander. “Americans' Daily Coffee Consumption at Home during the Week 2019.” *Statista*, 11 Feb. 2020, [www.statista.com/statistics/250230/americans-daily-coffee-consumption/](http://www.statista.com/statistics/250230/americans-daily-coffee-consumption/).

Misachi, John. “Which States Grow Coffee?” *WorldAtlas*, WorldAtlas, 25 Feb. 2019, [www.worldatlas.com/articles/which-states-grow-coffee.html](http://www.worldatlas.com/articles/which-states-grow-coffee.html).

“USDA/NASS QuickStats Ad-Hoc Query Tool.” *USDA/NASS QuickStats Ad-Hoc Query Tool*, USDA, quickstats.nass.usda.gov/.

“Where Coffee Grows.” *Coffee and Health*, 24 Nov. 2014, [www.coffeeandhealth.org/all-about-coffee/where-coffee-grows/](http://www.coffeeandhealth.org/all-about-coffee/where-coffee-grows/).