Milestone 2 Submission: Acquire and Understand the Data

Data imbalance

The kaggle website contains 3 dataset: one for Arabica coffee, one for Robusta coffee and a merged dataset containing the rows from the previous ones. We are going to use the merged dataset for our study. Some data issues in the merged dataset include missing values, particularly with regard to geographic location of the bean source. Robusta beans consistently have missing values for “altitude_high_meters”, “altitude_low_meters”, and “altitude_mean_meters”, therefore we will drop these columns and use “Altitude” instead. Also, we noticed that the column “Lot.Number” mostly contains missing values. In addition, we noticed that there is a farm with “altitude_high_meters” greater than 8,000 meters. Further analysis revealed that it was most likely due to input error as the same farm has an altitude that is scaled down by 100. We also noticed some duplicated columns : “Owner” and “Owner.1”.

We also notice several class imbalances: notably, in the ‘Species’ column, there are only 28 ‘Robusta’ observations, while the rest are ‘Arabica’. Another source of class imbalance is in the ‘Country’ column, where there are a few highly represented countries, four with greater than 100 observations, and many other countries with fewer than 20 observations. We can address this by applying stratified sampling during our train-test split; we can also create an “other” category to gather the countries who are the least represented.

Data Skewness

After standardizing the data and visualizing the distributions of the numeric features, we notice that many of the features have a few dramatic outliers that are more than 15 standard deviations from the mean. These outliers have skewed initial plots that we have made to further visualize the data, and a visualization of the dataset’s second principal component reveals most of the data clumped together with only a few points demonstrating the variance of the PC. When visualizing the summary statistics of the scaled numeric features in boxplots (without outliers), we notice that there are a few features with almost no variance (“Uniformity”, “Clean.Cup”, “Sweetness”, “Category.One.Defects”, and “Quakers”).

Collinearity

Because we are primarily interested in some of the taste-related features to predict coffee score, we created a pairplot between these features, to gain an understanding of

any interactions between the predictors. Upon visual inspection, the ‘Aroma’, ‘Flavor’, ‘Aftertaste’, ‘Body’, ’Acidity’ and ‘Balance’ appear to have a strong positive correlation with each other. This correlation may violate some statistical assumptions of our model, so we will have to consider strategies to address this such as reducing dimensionality using PCA or dropping correlated features.

Data interpretation:

- Quality descriptions: https://database.coffeeinstitute.org/coffee/654175/grade
- Defects descriptions: https://database.coffeeinstitute.org/coffee/654175/green