An Analysis of Coffee Quality Institute Data

Tarunraj Amuthan - tamuthan@mail.sfsu.edu *

December 3, 2020

Contents

*Edited by Dr. Luella Fu, luella@sfsu.edu

1 Introduction

Coffee professionals are well aware of how much effort goes into providing a single cup of coffee to a consumer. Billions of dollars worth of unprocessed and unroasted coffee beans (green coffee) is imported just to the US every year. Each of these 26.2 million bags of coffee is usually picked, processed, and reviewed by hand. Aiding in the review process is a small society of licensed coffee-tasting experts called Q-Graders. These “coffee sommeliers" test and grade coffees on both physical and tasting factors. The Coffee Quality Institute, the issuer of Q-grading licenses, holds a repository containing thousands of coffees submitted for review. What insights can we gain from examining this data?

2 Data Description

Our first step would be to acquire data from the Coffee Quality Institute (CQI). The easiest way to do so was to use data scraped in January 2018 by the user jldbc on GitHub[4]. A cursory examination of the data finds some cleanliness issues. Many of our factors are inconsistent between coffees. For example, the altitude at which the selected coffee was grown was listed in some combination of numbers, text, feet, meters, English, Spanish, and Chinese. It is therefore clear that not all factors can or should be used without some cleaning.

2.1 Selecting our Factors

First, we list the 43 factors included in the dataset:

ID	Bag Weight	Balance	Expiration
Species	In Country Partner	Uniformity	Certification Body
Owner	Harvest Year	Clean Cup	Certification Address
Country	Grading Date	Sweetness	Certification Contact
Farm	Owner 1	Cupper Points	Unit of Measure
Lot	Variety	Total Cup Points	Altitude Low Meters
Mill	Processing Method	Moisture	Altitude High Meters
ICO	Aroma	Category One Defects	Altitude Mean Meters
Company	Flavor	Quakers	
Altitude	Aftertaste	Color	
Region	Acidity	Category Two Defects	
Producer	Body		
Number Bags			

To begin our considerations, we will want identifying information for the source and physical description of each coffee.

ID	Variety
Country	Processing Method

It would be useful to include more sourcing information like Farm, Owner, and Altitude, but the data is inconsistent. This includes Altitudes that are inconsistently measured in meters and feet, Farm names in languages other than English, and missing Owner names. To clean the data would take more time than we have, and simply dropping rows with inconsistencies in these features would not leave enough data to work with.

We also include as much of the coffee tasting data as possible. This will be the basis of much of our analysis, since they are all strictly numeric on a scale of 1-10 or 1-100.

Aroma	Acidity	Uniformity	Cupper Points
Flavor	Body	Clean Cup	Total Cup Points
Aftertaste	Balance	Sweetness	

Removing unusable factors cuts the columns to process in half.

2.2 Definitions of Factors

Many of these definitions, especially for the tasting attributes, are taken verbatim from the CQI website[2].

ID
Unique identifier for each coffee from 1 to 1312.

Country
Origin of Coffee. Like wine, the character of coffee is influenced by the terroir of its homeland. However, more important is usually the regionalized processing methods used to prepare the coffees.

Variety
Lists the subspecies, or cultivars, of Arabica. Most common are Bourbon and Caturra, but many new hybrids like Catimor have appeared to combat disease and pests like coffee rust. There are also rare local cultivars like the Ethiopian native Gesha.

Processing Method
The two major methods are natural and washed processing. There are subcategories for each, and methods are generally similar within regions.

Aroma
The aromatic aspects include Fragrance (defined as the smell of the ground coffee when still dry) and Aroma (the smell of the cof-

fee when infused with hot water). One can evaluate this at three distinct steps in the cupping process: (1) sniffing the grounds placed into the cup before pouring water onto the coffee; (2) sniffing the aromas released while breaking the crust; and (3) sniffing the aromas released as the coffee steeps. Specific aromas can be noted under "qualities" and the intensity of the dry, break, and wet aroma aspects noted on the 5-point vertical scales. The score finally given should reflect the preference of all three aspects of a sample's Fragrance/Aroma.

Flavor
Flavor represents the coffee's principal character, the "mid-range" notes, in between the first impressions given by the coffee's first aroma and acidity to its final aftertaste. It is a combined impression of all the gustatory (taste bud) sensations and retro-nasal aromas that go from the mouth to nose. The score given for Flavor should account for the intensity, quality and complexity of its combined taste and aroma, experienced when the coffee is slurped into the mouth vigorously so as to involve the entire palate in the evaluation.

Aftertaste
Aftertaste is the length of positive flavor (taste and aroma) qualities emanating from the back of the palate and remaining after the coffee is expectorated or swallowed. If the aftertaste were short or unpleasant, a lower score would be given.

Acidity
Acidity is often described as "brightness" when favorable or "sour" when unfavorable. At its best, acidity contributes to a coffee's liveliness, sweetness, and fresh-fruit character and is almost immediately experienced and evaluated when the coffee is first slurped into the mouth. Acidity that is overly intense or dominating may be unpleasant, however, and excessive acidity may not be appropriate to the flavor profile of the sample. The final score marked on the horizontal tick-mark scale should reflect the panelist's perceived quality for the Acidity relative to the expected flavor profile based on origin characteristics and/or other factors (degree of roast, intended use, etc.). Coffees expected to be high in Acidity, such as a Kenya coffee, or coffees expected to be low in Acidity, such as a Sumatra coffee, can receive equally high preference scores although their intensity rankings will be quite different.

Body
The quality of Body is based upon the tactile feeling of the liquid in the mouth, especially as perceived between the tongue and roof of the mouth. Most samples with heavy Body may also receive a high score in terms of quality due to the presence of brew colloids and sucrose. Some samples with lighter Body may also have a pleasant feeling in the mouth however. Coffees expected to be high in Body, such as a Sumatra coffee, or coffees expected to be low

in Body, such as a Mexican coffee, can receive equally high preference scores although their intensity rankings will be quite different.

Balance
How all the various aspects of Flavor, Aftertaste, Acidity and Body of the sample work together and complement or contrast to each other is Balance. If the sample is lacking in certain aroma or taste attributes or if some attributes are overpowering, the Balance score would be reduced.

Uniformity
Uniformity refers to consistency of flavor of the different cups of the sample tasted. If the cups taste different, the rating of this aspect would not be as high. 2 points are awarded for each cup displaying this attribute, with a maximum of 10 points if all 5 cups are the same.

Clean Cup
Clean Cup refers to a lack of interfering negative impressions from first ingestion to final aftertaste, the "transparency" of cup. In evaluating this attribute, notice the total flavor experience from the time of the initial ingestion to final swallowing or expectoration. Any non-coffee-like tastes or aromas will disqualify an individual cup. 2 points are awarded for each cup displaying the attribute of Clean Cup.

Sweetness
Sweetness refers to a pleasing fullness of flavor as well as any obvious sweetness and its perception is the result of the presence of certain carbohydrates. The opposite of sweetness in this context is sour, astringency or “green" flavors. This quality may not be directly perceived as it would be in sucrose-laden products such as soft drinks, but will affect other flavor attributes. 2 points are awarded for each cup displaying this attribute for a maximum score of 10 points.

Cupper Points
The “overall" scoring aspect is meant to reflect the holistically integrated rating of the sample as perceived by the individual panelist. A sample with many highly pleasant aspects but not quite “measuring up" would receive a lower rating. A coffee that met expectations as to its character and reflected particular origin flavor qualities would receive a high score. An exemplary example of preferred characteristics not fully reflected in the individual score of the individual attributes might receive an even higher score. This is the step where the panelists make their personal appraisal.

Total Cup Points
A sum of the previous ten quality measures.

2.3 Cleaning Data

The first issue with our selected factors came in Country. We can examine a plot including all of the countries included over 30 countries, including many outliers in **Figure 2**. However, examining **Figure 4** we can see that the majority of the

countries have less than 30 samples coming from them. Therefore, we decided to select only countries with at least 30 samples.

3 Methods

3.1 One Hot/Dummy Encoding

Data sets with extensive categorical data require additional attention, since many models require numerical data to function properly. One Hot/Dummy encoding is one way to convert from categorical to numerical format. The name of the encoding comes from how it works: one represents the presence of a specific trait. This means that a set that looks like Table 1 can be encoded to have all numerical values like in Table 2

Flower	Color
1	Red
2	Blue
3	Purple

Table 1: Categorical Flower Data

However, if you're using linear regression it is important to account for collinearity by dropping the first categorical column. This absorbs that classification into the constant and provides independent features, an assumption when using linear regression.

3.2 Linear Regression

Linear Regression aims to find a linear relationship between inputs and a target. The basic formula looks like

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where each feature is represented by x_n and has an importance of θ_n. In other words, for each 1 unit increase of x_n, the feature affects the output $h_\theta(x)$ by θ_n amount. If we define our intercept by setting $x_0 = 1$, then this can be written concisely using matrices as

$$h_\theta(x) = \theta^T X$$

Flower	**Color_Red**	**Color_Blue**	**Color_Purple**
1	1	0	0
2	0	1	0
3	0	0	1

Table 2: Dummy Encoded Flower Data

Let's consider an example: there is a company spending some millions of dollars on marketing each year that results in a change in sales, also in millions. This is provided in table 4. Linear regressions can be solved analytically using the

Year	Marketing	Sales
1	23	651
2	30	856
3	48	1298

Table 3: Linear Regression Example

normal equation formula, where X is a matrix of the features and y is a matrix of the labels.

$$\theta = (X^T X)^{-1} X^T y :$$

When we plug in our values, we get

$$\theta = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 23 & 30 & 48 \end{bmatrix} \begin{bmatrix} 1 & 1 & 23 \\ 1 & 2 & 30 \\ 1 & 3 & 48 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 23 & 30 & 48 \end{bmatrix} \begin{bmatrix} 651 \\ 856 \\ 1298 \end{bmatrix},$$

which can be simplified to

$$\theta = \begin{bmatrix} 323 & 14 & 47 \end{bmatrix}$$

From this matrix we can write our hypothesized model for sales based on year and money spent on marketing.

$$h_\theta(x_{sales}) = 323 + 14x_{marketing} + 47x_{year}.$$

3.2.1 Assumptions

Linear regression operates on four main assumptions. We assume linearity since we are trying to fit a linear regression. We assume independence; correlation between features will cause issues. We assume that as the predictors increase, variance does not also increase - aka we have constant variance. Finally, we assume that errors are normally distributed.

3.3 Principal Component Analysis

Principal Component Analysis (PCA) is used when we want to reduce the number of features and are comfortable with losing interpretability. PCA finds the direction of greatest variance in the data. More technically put, we will calculate magnitude and direction of components using eigenvalues and eigenvectors respectively, and then transform our original data to align with these directions. The steps are as follows.

1. Separate out independent columns X and dependent column Y. Save this dependent column for later.

2. Prepare any categorical features by converting to numerical values.

3. Center the features by subtracting the mean of each column from each entry so each feature has a mean of zero.

4. Finish standardizing by dividing each entry by the column's standard deviation; call this new matrix Z.

5. Find the covariance matrix of Z (up to a constant) by multiplying by its transposed matrix: $Z^T Z$

6. Calculate eigenvectors P and eigenvalues D of $Z^T Z$, aka PDP^{-1}.

7. Sort D from largest to smallest and sort the according P. We call this sorted set of independent eigenvectors P^*.

8. Calculate $Z^* = ZP^*$, where Z is the centered and standardized version of X where each observation in $Z^*$is a combination of the original variables weighted by the eigenvectors.

9. Calculate the proportion of variance explained by dividing the sum of the kept eigenvalues by the sum of all the eigenvalues. Decide how many eigenvalues to keep using one of three methods:

   (a) Arbitrarily select how many to keep! This is useful when you know you want to make a 2d/3d graph.

   (b) Calculate proportion of variance explained and add components that have the largest explained proportion of variance until you reach a desired proportion.

   (c) Plot the scree plot, the cumulative proportion of variance explained as you add more features. Identifying the point where there is a significant drop in variance gives you an idea of how many features to keep.

3.3.1 Assumptions

There are no assumptions with PCA. PCA is a nonparametric method. It can take whatever data you input and use matrix transformations to map it into a linear subspace.

3.4 Random Forest

Random Forest is an ensemble classifier that uses bootstrapped and aggregated (bagged) decision trees to predict a classification, but adds an additional step where it selects a random subsection of the features for each split [3]. We start by creating a bootstrapped sample. Bootstrapping is the process of random sampling with replacement. We then attempt to split our sample into the purest classifications using a decision tree. Random Forest de-correlates decision trees

by considering a subset of the features at each node, generating more unique trees[3], improving on simply bagging decision trees. The criteria for splitting differs depending on the type of Random Forest being constructed.

3.4.1 Regression and MSE/MAE

When building a Random Forest Regression, Mean Squared Error (MSE) or Mean Absolute Error (MAE) is generally used to find the best possible splits for a dataset.[5] The formulas are fairly similar:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

The choice of which to use depends on how much you care about outliers. MSE is more easily affected by outliers than MAE, so it makes more sense to use MSE when we definitely do not want any large errors. For example, if being off by four was just twice as bad as being off by two, we would use MAE. If for some reason it was more than twice as bad, we would consider MSE to better represent that fact.

3.4.2 Classification and Gini Impurity

To find the correct splits when using a Random Forest classifier, we minimize the Gini Impurity at each node of the tree. The Gini Impurity is the chance we have classified our point incorrectly and is represented as:

$$I_G(n) = 1 - \sum_{i=1}^{J}(p_i)^2$$

where n is some node, J is the number of possible classifications, and p_i is the fraction of examples in each classification. This process is repeated recursively until each node contains samples from a single classification or we reach a specified depth. The output of multiple decision trees is finally aggregated by averaging to get a final prediction.

Let's take this example from our project. Figure 1 is a shortened version of the tree used to predict coffee score.

The Gini values for each leaf are listed, but we can calculate them by hand. Let's start with the root node. First we'll find our values for each classification at the node:

- Below Average = 211
- Average = 395

- Above Average = 208

Each node has a sample size listed, but that sample is the number of unique features represented by the bootstrapped sample. For the Gini calculation we instead need to add up the total number of values at the node:

$$211 + 395 + 208 = 814$$

Then we'll use our formula to find our Gini value at the root node.

$$I_G(root) = 1 - \left(\frac{211}{814}^2 + \frac{395}{814}^2 + \frac{208}{814}^2\right) = 0.63$$

We can repeat this for every following leaf - take the second row.

$$I_G(left) = 1 - \left(\frac{211}{814}^2 + \frac{394}{814}^2 + \frac{205}{814}^2\right) = 0.63$$

$$I_G(right) = 1 - \left(\frac{0}{814}^2 + \frac{1}{814}^2 + \frac{3}{814}^2\right) = 0.38$$

By minimizing the Gini value at each node, we build trees that have the most useful splits of data.

Although the official website for Random Forest claims it does not overfit, we can balance our bias-variance tradeoff with a few techniques. We can increase the number of trees, decrease the depth of each tree, decrease the number of features to decrease variance, or increase the number of folds in our cross-validation. Scikit-learn offers a few implementations of K-folds validation, a form of cross-validation where the data is split into K number of folds and one of these subsets is withheld at a time as a testing group while the remaining folds are used for training. The final score is an average of the chosen score for each fold.[3] In the case of the Random Forest Classifier in Sklearn, that score is the mean accuracy of the prediction.

3.4.3 Assumptions

Random Forest, like simple decision trees, is nonparametric, so we do not need to assume the data follows any specific probability relationship [3]. We do still need to assume that our sample is representative of the larger population; this assumption will carry over to bootstrapped samples. Specifically to using scikit-learn, we are also assuming that our encoded categorical features are independent and do not result in collinearity. It is possible to use other methods to construct random forests without encoding, which negates this assumption.

4 Results

4.1 Linear Regression

For our linear regression we tested the relationship between the numerical data in our dataset. Doing so showed us that they were indeed linearly related, with coefficients of nearly one, with an R^2 value of 0.99998.

Feature	Coefficient
Aroma	1.003527222
Flavor	0.996804239
Aftertaste	1.006090577
Acidity	0.997914463
Body	1.002493035
Balance	1.003518367
Uniformity	1.005217219
Clean.Cup	1.00086604
Sweetness	0.998234109

Table 4: Results of Linear Regression

See figure 3 to see nearly perfect accuracy. This makes sense; our target, Total Cupping Score, is a simple sum of the other tasting scores.

Our next round of linear regression models combined encoded categorical data with our numerical data. Doing so gave us an unreasonable MSE greater than $5 \cdot 10^{21}$ and an R^2 value less than $-6 \cdot 10^{20}$. This suggests that there is likely not a linear relationship between these features. For a full list of features and their coefficients, view table 7. However, it can be noted that the largest coefficients were for countries and processing methods, which suggests they may have the most effect on coffee score. This is supported by our PCA and Random Forest results, discussed below, as well.

4.2 PCA

Initially, we chose to conduct PCA on our numeric data with the intention of lowering the number of features to a level that could be easily visualized. Doing so gave the plot in figure 5. This tells us that principal component 1 has the most effect on the score. The problem with the graph in figure 5 is that our target is a limited range of scores, so the actual target result is not easily represented in shades. However, we know from summary statistics that the mean for Total Cup Score is 82.1, the minimum is 59.8, the maximum is 90.6, and the 25 and 75 quartiles are 81.1 and 83.6. For a plot of the normally-distributed tasting subscores see figure 9. It seems that only extraordinary coffees break far from the mean score of 82. Therefore, let us consider three scores: Below Average (less than 81.1), Average, and Above Average(greater than 83.6). With that addition our graph looks like figure 6. However, this still uses the numeric score as the target; we wanted to see how PCA would handle categorical data as the target, so we re-ran it with that output. This resulted in figure 7.

We decided to examine how many variables we could remove while maintaining the most variance. With examination of the pca.n_components_ variable and the scree plot in figure 8 we find that to maintain 80% variance we would need to keep 3 of the 7 total principle components.

Next we wanted to include our categorical data as features. PCA cannot

handle non-numeric data as features, so we encoded our features using One Hot Encoding (OHE). This results in figure 10. In this we see that the target, Total Cupping Score represented by colors, varies most with principal component 1.

We created another scree plot, figure 11. The number of features necessary to maintain a reasonable amount of variance suggests our data does not have clear delineations and cannot be easily visualized. We did start with 75 columns though, so columns can be combined into 55 components to maintain 90% variance. Therefore, the first 50 eigenvectors should probably be used to construct the dimensions for a new feature space, if we need to use PCA.

To separate our eigenvectors into their component features we examined the pca.explained_variance_ratio_ variable. Doing so suggested that there were some feature classifications that had significantly more influence on the final score than others. The first three components explained 45% of the variance. Here is the makeup of the first component, which explains 23% of the variance.

Feature	**Coefficient**
Country_ Brazil	0.29
Country_ Guatemala	-0.13
Country_ Mexico	-0.21
Country_ United States Hawaii)	0.1
Variety_ Caturra	-0.08
Variety_ Hawaiian Kona	0.1
Variety_ Typica	-0.12
Variety_ Yellow Bourbon	0.08
Processing.Method_ Natural / Dry	0.57
Processing.Method_ Semi-washed / Semi-pulped	0.07
Processing.Method_ Washed / Wet	-0.68

Table 5: Coefficients for each feature

Separating our eigenvectors has shown us that the features that affect the total cupping score most are the Processing Method and Country. This proves consistent with our findings from Linear Regression; we will see that it also proves consistent with Random Forest.

4.3 Random Forest

To begin, we ran a Random Forest regression model using the encoded categorical data from earlier with the total cupping score as the target. This model assigned each feature's importance as seen in figure 12, suggesting that whether a coffee was from Mexico and whether it was a washed-process coffee were the two most important factors in determining cupping score. This proves consistent with Linear Regression and PCA, suggesting that Country and Processing method matter more than Species. This model gave us a mean absolute error of 1.68, only marginally better than the 1.69 MAE of guessing the average. The next question was whether these features better predicted one of the other

numerical values, but table 6 shows that there is negligible improvement when using a random forest model in comparison to the base error of guessing the average. Examining our full tree structure in figure 15 suggests an issue: the tree is unequally growing towards the left. This results in a 'sparse' looking tree, which means our regression model may be splitting too finely because of our encoded features.

To try to combat this, we then used a Random Forest classifier model on our data. To do so, we prepared our target by categorizing scores into Below Average, Average, and Above Average. This produced more readable trees like in figure 14, and more similarly scaled importances as seen in figure 13. However, our performance took a hit. The Random Forest classifier guessed correctly 54% of the time, while guessing that all are average resulted in a hit rate of 61%. Examining our tree in figure 16 we see a slightly less, but still sparse, decision tree. Improving this model will likely require using a Random Forest model that accepts categorical features without encoding.

target	**error improvement**	**error**	**base error**
Aroma	0.01	0.22	0.23
Flavor	0.01	0.23	0.24
Aftertaste	0.03	0.22	0.25
Acidity	0.01	0.21	0.22
Body	0.01	0.19	0.2
Balance	0.03	0.23	0.26
Uniformity	-0.01	0.24	0.23
Clean.Cup	-0.04	0.28	0.24
Sweetness	-0.05	0.12	0.07
Cupper.Points	0.02	0.26	0.28

Table 6: Error comparisons between guessing the average and using our random forest model.

5 Conclusions

5.1 Limitations

Our conclusions are limited by our choice of tool and the amount of data. Increasing our data-set would result in more accurate models. Random Forest especially performs better with more data [3]. It would be reasonable to say that with a magnitude greater number of rows we may have been able to get more concrete results. We also chose early on to use scikit-learn (sklearn) in Python as our modeling package. Sklearn performs admirably in basic modeling situations and offers more transparent results than a package like TensorFlow. Sklearn, however, falters when it comes to creating Random Forest models with qualitative data. Sklearn requires categorical data to be encoded, which intro-

duces unnecessary assumptions with the data. By one-hot encoding a categorical variable, we create many binary variables, and from the splitting algorithm's point of view, they are all supposed to be independent [1]. In the future, a platform like H2O.ai, which natively accepts categorical data for Random Forest models, would give us results with less noise. One Hot Encoding also affects the decision trees in random forest. In our study, we find that decision trees look "sparse", splitting extensively in one direction. This occurs because decision trees tend to split on variables with higher cardinality, which in this case are Country and Variety. This could be avoided with tools that do not presuppose one hot encoding, like H2O.ai, but scikit-learn requires categorical data to be encoded [6].

5.2 Discussion

Our general finding is that coffee is complicated to model. None of our models outperformed simply guessing the average, with the best alternative being random forest. With that in mind, there are some takeaways. All methods suggest that processing method and origin country seem to be more important than coffee variety when it comes to the final cupping score. It makes sense for these two to be similarly important: processing methods are usually based on regional traditions. Anecdotally this is supported by the fact that coffee packaging will commonly list the country of origin and processing method but rarely list the subspecies of coffee.

References and Figures

[1] Nick Dingwall and Chris Potts. *Are categorical variables getting lost in your random forests?* URL: https://roamanalytics.com/2016/10/28/are-categorical-variables-getting-lost-in-your-random-forests/. (accessed: 04.29.2020).

[2] Coffee Quality Institute. *Coffee Quality Institute Database*. URL: https://database.coffeeinstitute.org/. (accessed: 04.29.2020).

[3] Gareth James et al. "An Introduction to Statistical Learning, Seventh Edition". In: Springer, 2013. Chap. 1.2.

[4] James LeDoux. *coffee-quality-database*. 2018. URL: https://github.com/jldbc/coffee-quality-database. (accessed: 04.29.2020).

[5] *Performance and Prediction*. URL: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/performance-and-prediction.html#model-performance. (accessed: 04.29.2020).

[6] Rakesh Ravi. *One-Hot Encoding is making your Tree-Based Ensembles worse, here's why?* URL: https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769. (accessed: 04.29.2020).

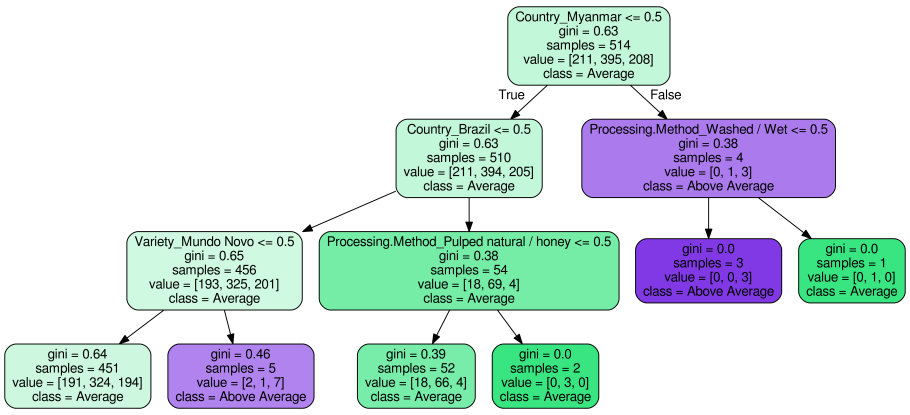

Figure 1: A Short Tree

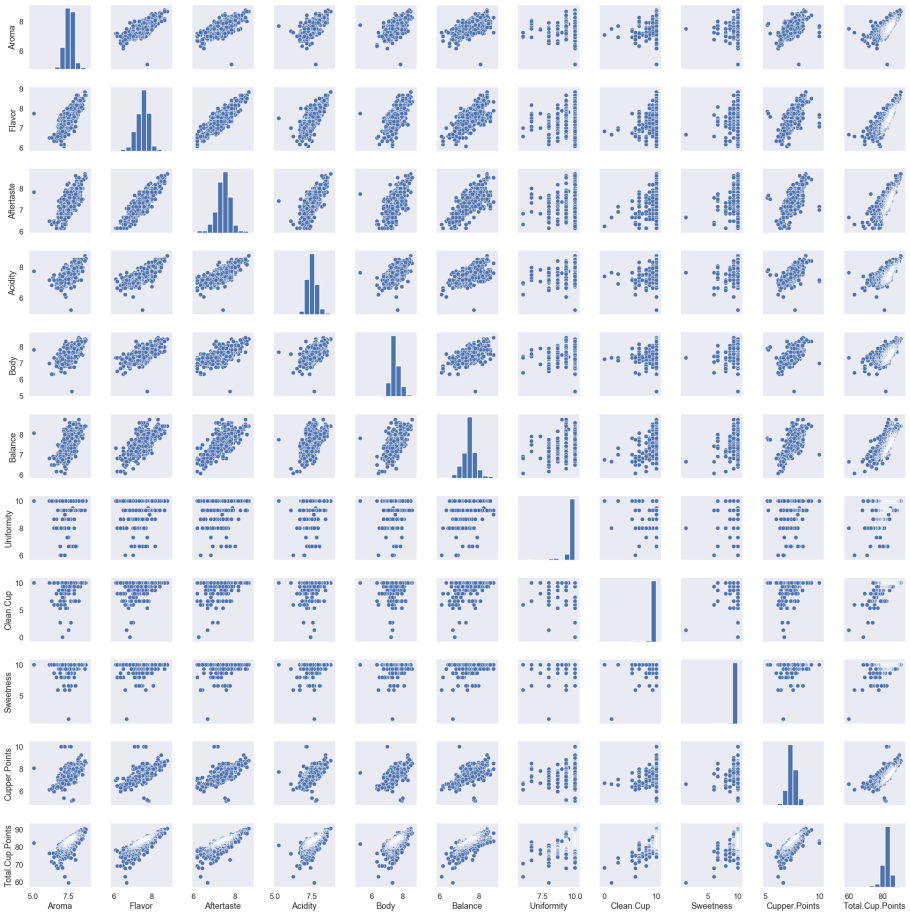

Figure 2: A pairplot examining all coffees by Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Cupper Points, Total Cup Points

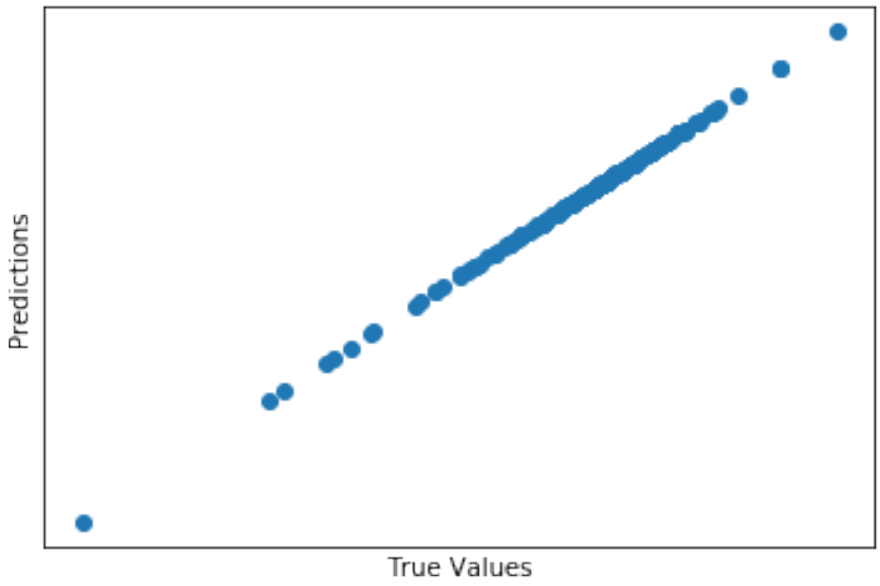

Figure 3: Results comparing predictions to actual values of a linear regression, where the target is total coffee score.

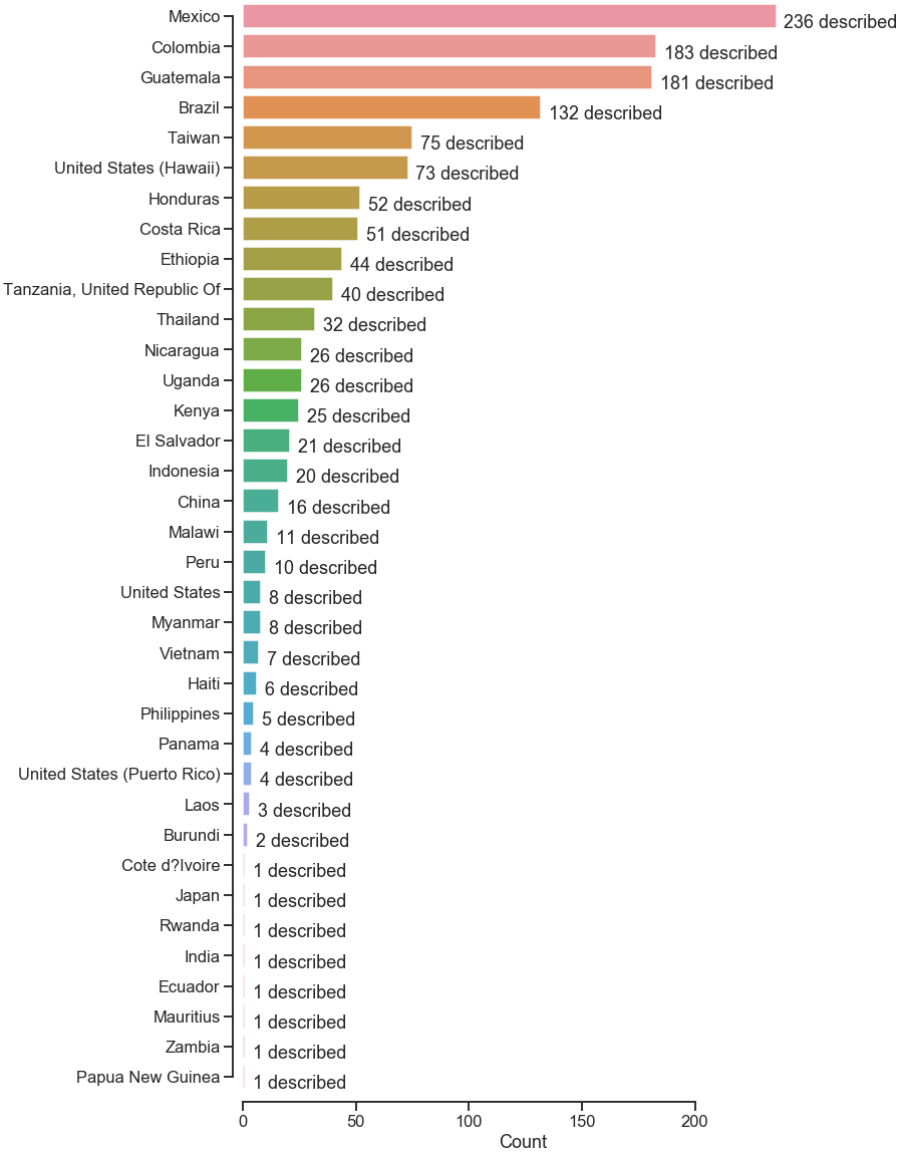

Figure 4: Number of Coffees per Country

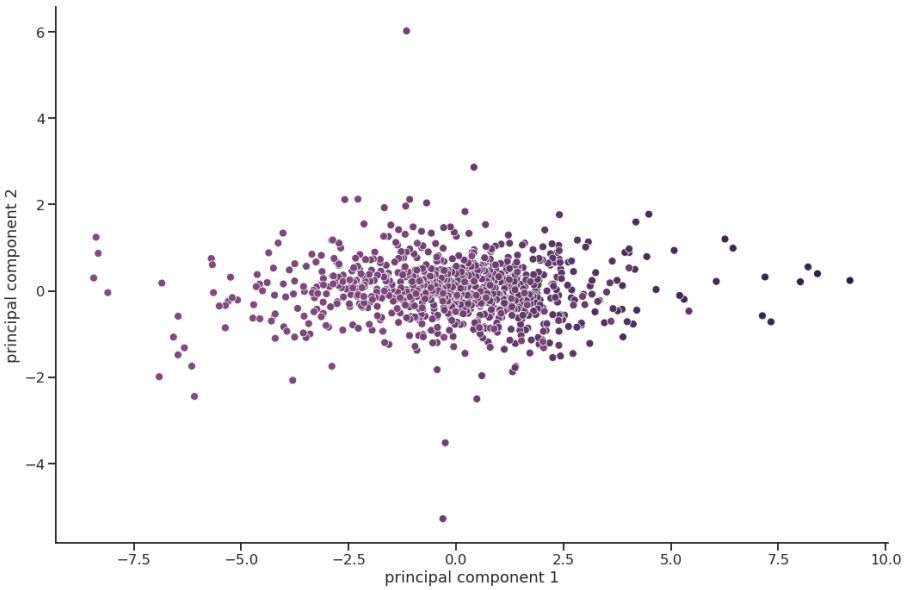

Figure 5: Number of Coffees per Country

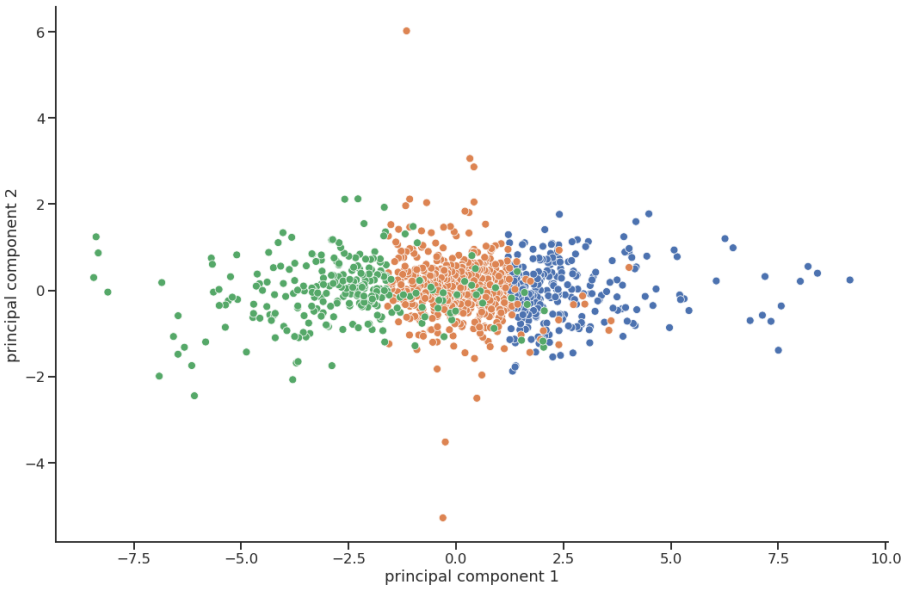

Figure 6: Comparing PC 1 and 2 vs Target

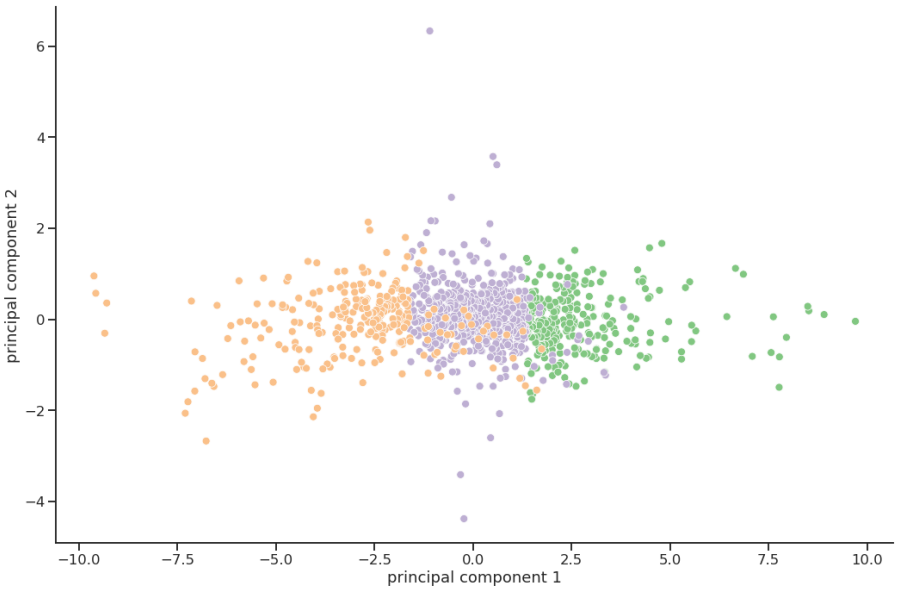

Figure 7: Comparing PC 1 and 2 vs Categorized Target

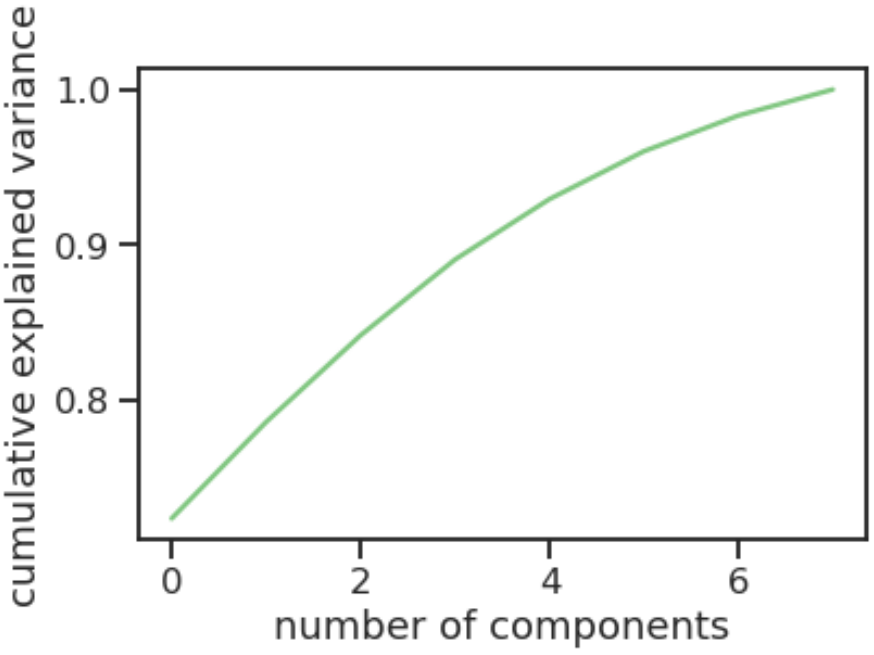

Figure 8: Scree Plot

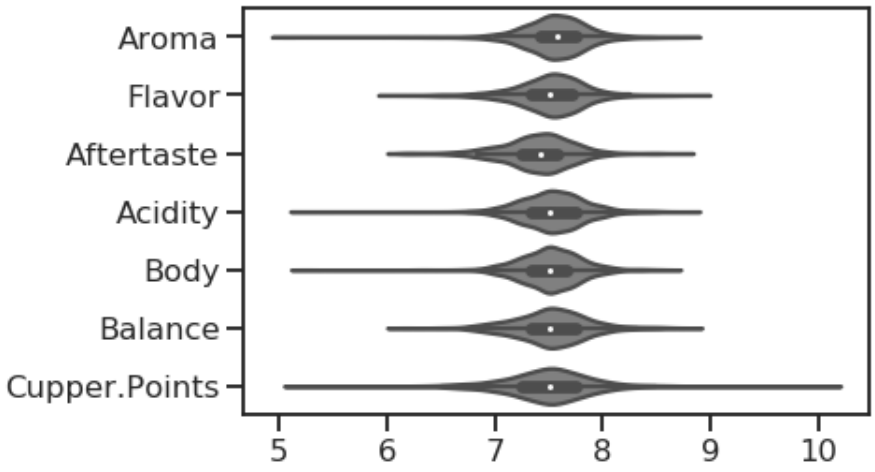

Figure 9: Violin plot for normally distributed numerical features

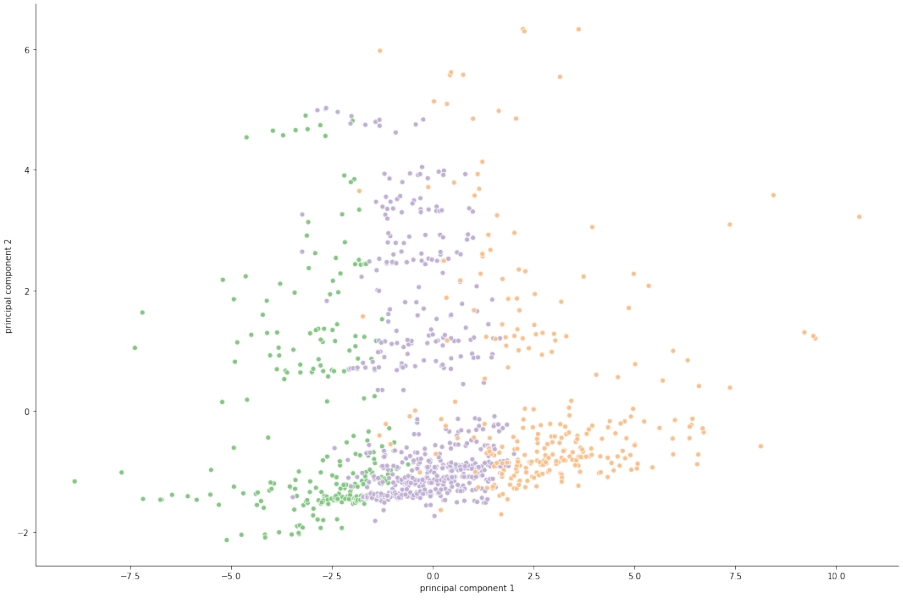

Figure 10: Comparing PC 1 and 2 vs Classified Target with Categorical Data

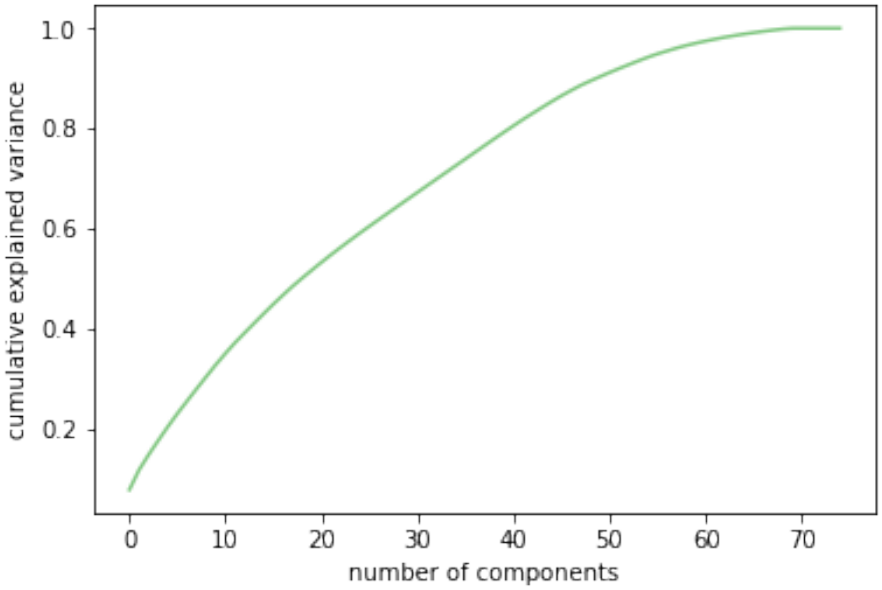

Figure 11: Scree Plot

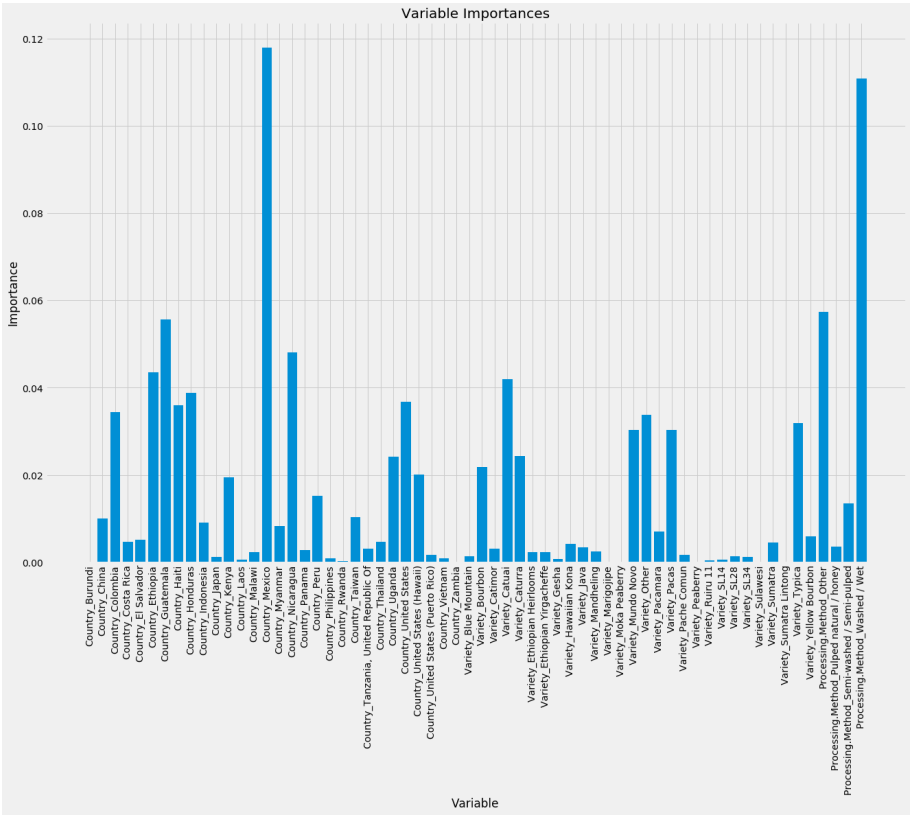

Figure 12: Importances from Random Forest Regression

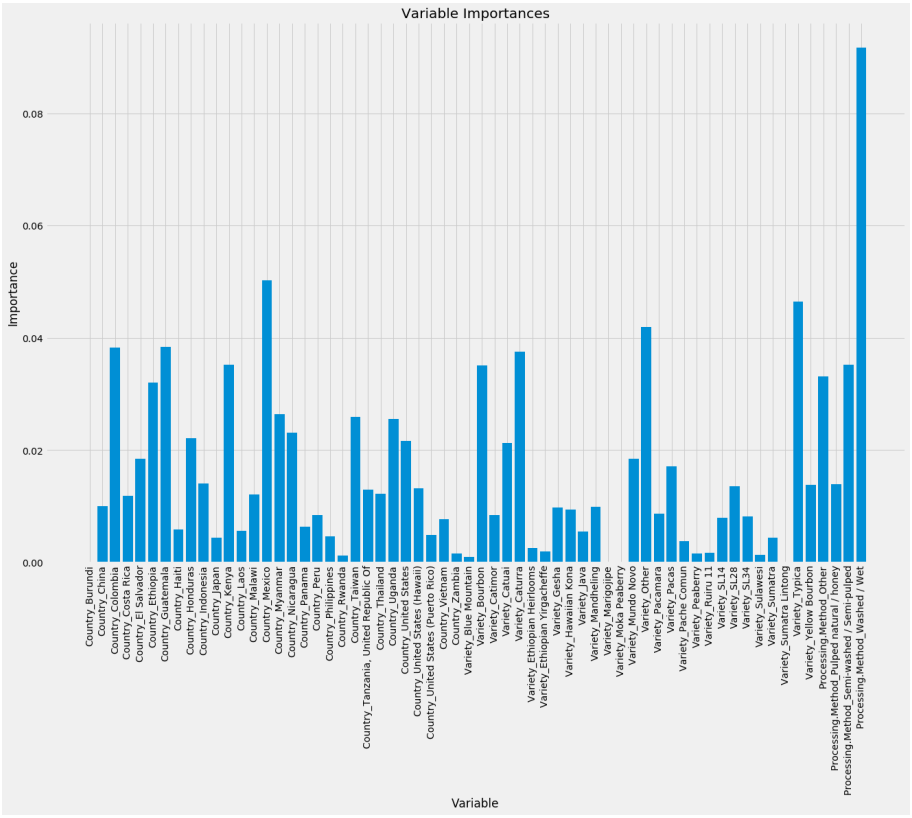

Figure 13: Importances from Random Forest Classifier

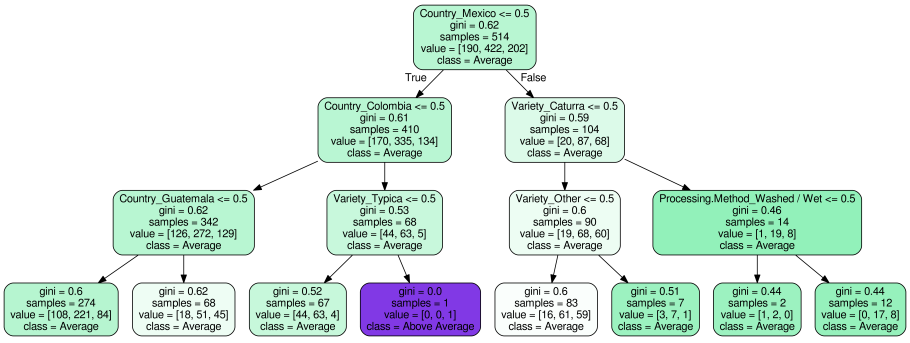

Figure 14: Truncated tree structure

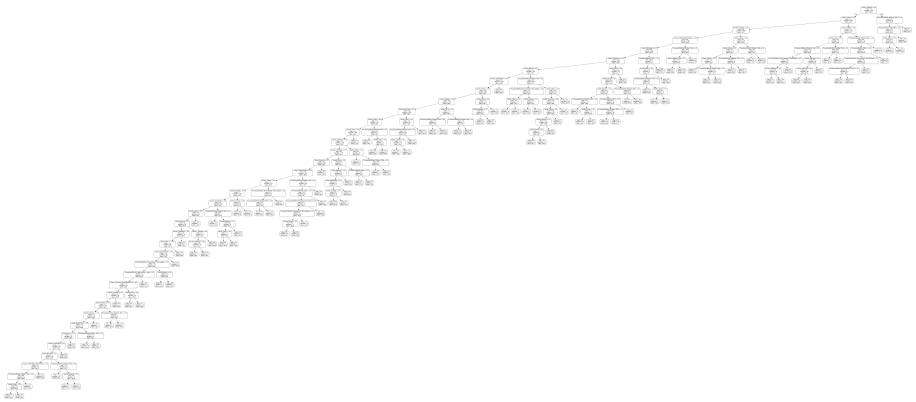

Figure 15: Full tree structure for Random Forest Regression

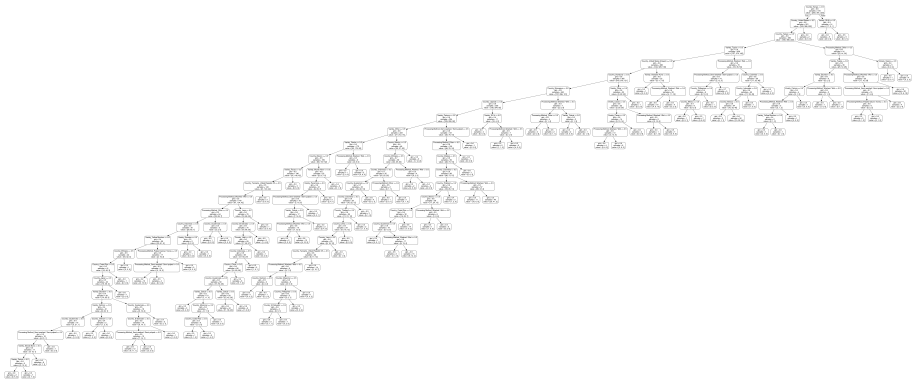

Figure 16: Full tree structure for Random Forest Classification

Table 7: Features' Coefficients after Linear Regression

Feature	**Coefficient**
Aroma	0.3040548
Flavor	0.326631645
Aftertaste	0.338905644
Acidity	0.300910812
Body	0.274147768
Balance	0.341402733
Uniformity	0.477705834
Clean.Cup	0.751469313
Sweetness	0.449878798
Cupper.Points	0.387829176
Country_Brazil	85099294908
Country_Burundi	23230423232
Country_China	33982598978
Country_Colombia	93939538483
Country_Costa Rica	56212944142
Country_El Salvador	35012088989
Country_Ethiopia	31817504987
Country_Guatemala	1.02489E+11
Country_Haiti	19094248022
Country_Honduras	57393328617
Country_Indonesia	35012088989
Country_Japan	8554991442
Country_Kenya	3.21353E+11
Country_Laos	14804016689
Country_Malawi	28242639661
Country_Mexico	1.14501E+11
Country_Myanmar	24118988156
Country_Nicaragua	31817504987
Country_Panama	17086312156
Country_Peru	14804016689
Country_Philippines	17086312156
Country_Rwanda	8554991442
Country_Taiwan	68800204981
Country_Tanzania, United Republic Of	51829515604
Country_Thailand	28242639661
Country_Uganda	42299224635
Country_United States	24118988156
Country_United States (Hawaii)	56212944142
Country_United States (Puerto Rico)	17086312156

(To be continued)

Country_Vietnam	22571709524
Country_Zambia	67696610215
Variety_Arusha	45346664582
Variety_Blue Mountain	2.11413E+11
Variety_Bourbon	2.70586E+11
Variety_Catimor	90061898911
Variety_Catuai	1.68793E+11
Variety_Caturra	2.81585E+11
Variety_Ethiopian Heirlooms	20317130424
Variety_Ethiopian Yirgacheffe	28719517438
Variety_Gesha	70022927413
Variety_Hawaiian Kona	1.32071E+11
Variety_Java	28719517438
Variety_Mandheling	35157853740
Variety_Marigojipe	20317130424
Variety_Moka Peaberry	20317130424
Variety_Mundo Novo	1.11546E+11
Variety_Other	2.0046E+11
Variety_Pacamara	53605239632
Variety_Pacas	72848235439
Variety_Pache Comun	20317130424
Variety_Peaberry	38824488349
Variety_Ruiru 11	28719517438
Variety_SL14	83149724139
Variety_SL28	1.49393E+11
Variety_SL34	1.09457E+11
Variety_Sulawesi	20317130424
Variety_Sumatra	35157853740
Variety_Sumatra Lintong	20317130424
Variety_Typica	2.63588E+11
Variety_Yellow Bourbon	1.13277E+11
Processing.Method_Natural / Dry	8.76977E+11
Processing.Method_Other	3.30777E+11
Processing.Method_Pulped natural / honey	2.35324E+11
Processing.Method_Semi-washed/Semi-Pulped	4.74467E+11
Processing.Method_Washed / Wet	9.85572E+11