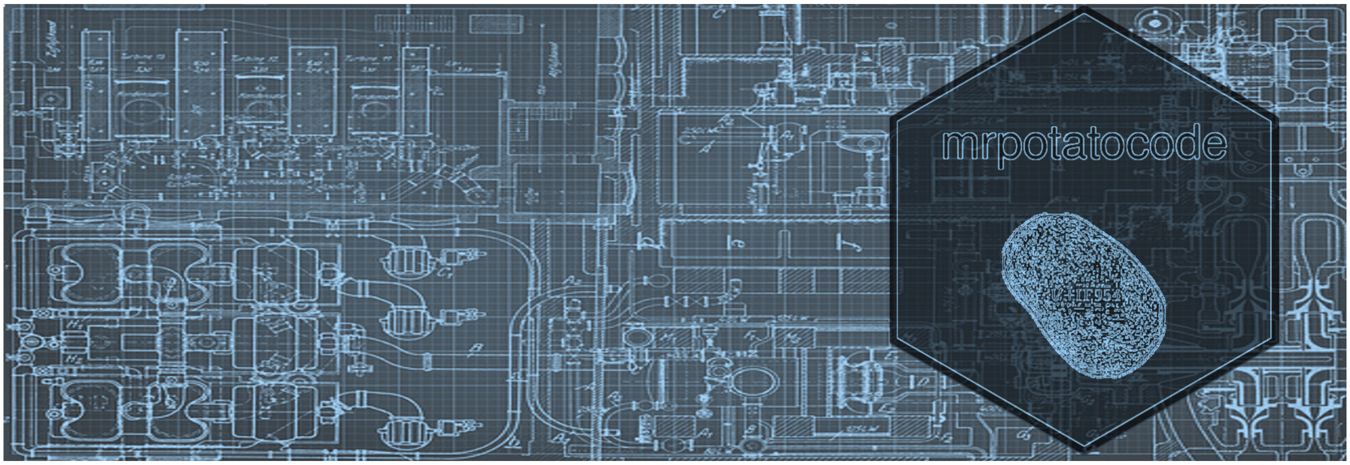


Coffee & Machine Learning: An Exploration into the Attributes that Define Single-Origin Coffees*

Thomas Rosenthal

April 24, 2021

Abstract

In this work, a boosted tree model explores whether combinations of unique coffee attributes can predict the tastes of specialty coffees. The attributes of 550 coffees, collected over a period of four months, include the coffee's country of origin, variety (subspecies), processing method, harvest characteristics (like altitude and season of harvest), and tastes generated through consistent cupping practices. These defining attributes are used as predictors for 22 distinct tasting groups and then compared to actual specialty coffees and found to often match two out of every three tasting groups provided by coffee roasters.

*Code and data are available at: github.com/mrpotatocode/COFFEE_COFFEE_COFFEE

Introduction

Coffee is consumed in every country, is the seventh most valuable agricultural product (Pendergrast, 2010), and supports the livelihood of 125 million people (Hoffmann, 2018) around the world. In many places, the daily coffee ritual signifies a start to the day and a nod to the productive buzz it imbues its consumer. Our appreciation for coffee has helped shape a complex landscape of horticulture, chemistry, food science, and globalised import-export economics. As a result, coffee has never tasted better. Our collective understanding of coffee information has led to phenomenal quality coffee beans and annual championships celebrating the labour and toil of farmers, producers, roasters, and baristas across the world.

Background

Combining coffee and data science, this exploratory research aims to examine whether a specialty coffee's Tasting Notes can be predicted by a gradient boosted algorithm based on the coffee's attributes (§Terminology). Several key research studies have explored the relationship between single-origin attributes and their Tasting Notes and thus sit as domain framework for technical applications made here. Exploring Tasting Notes is contingent firstly upon the acceptance of Tasting Notes as objective rather than subjective. This phenomenon has been well-noted by vintners and has likewise been researched with regard to specialty coffee (Croijmans & Majid, 2016). Research using blind taste and aroma tests generally supports consistency across highly trained panel tasters (Bhumiratana, Adhikari, & Chambers, 2011). The conclusions of Croijmans and Bhumiratana support the contention that Tasting Notes are a combined result of cupping experts' training and linguistic ability to identify tastes through plain nomenclature. For example, "A sour, sweet fruity aromatic that may be somewhat dark, musty and earthy, reminiscent of dark fruits and root vegetables such as beets and carrots [which] may also have an astringent mouthfeel" is identified as *Pomegranate* (World Coffee Research, 2017). Cupping, following protocols published by the Specialty Coffee Association (2003), serves as the only means for Tasting Note generation across all relevant studies. Standardized cupping routines were thoroughly explored as a case study in Rwanda (Goldstein, 2011) and found to be reliable.

As such, this project is firmly rooted in scientific consensus that Tasting Notes are consistent in their generation and thus measurable by machine learning models. From this assertion, two previous experiments have explored single-origin coffee Tasting Notes. Firstly, *Coffee Terroir: Cupping Description Profiles and Their Impact Upon Prices in Central American Coffees* (Conley & Wilson, 2018) combined Country of origin and Tasting Notes within a Multiclass Classification Neural Network and examined the attribute coefficients via regression analysis. Secondly, a research blog produced by Jonathan Gagné (2019) at the University of Montreal exemplifies Varietal:Tasting Note and Processing:Tasting Note relationships.

This work aims to reinforce the anecdotal assumption that curation of specialty coffees creates certain tastes. The industry cultivates (even without intention) a collection of expected Tasting Notes for known attribute combinations that are perceived to be desirable by consumers, thus reinforcing the industry's expectations and further cultivation of Tasting Notes. The results and implications of modelling are important within the coffee industry; if known attribute combinations directly contribute to desirable Tasting Notes, producers may know the value of a crop before buying and roasting. This in turn affects coffee farmers, as desirability has a noted effect on price; Traore et al. (2018) highlighted this phenomenon for Pacamara and Caturra varieties.

Terminology

As the world's appetite for coffee has grown, the coffee industry's "Third Wave" movement has flourished. The movement aims to treat coffee as an artisanal product that is carefully curated, where coffee quality is maximized at each stage: farming, producing, roasting, and selling (Rosenberg, Swilling, & Vermeulen, 2018). There is an "obsession" with coffee's taste and an often-altruistic approach to conducting business (Pendergrast, 2010). At its pinnacle, single-origin specialty coffee exists beyond a means to caffeinate and instead engages an increasingly discerning consumer.

Under the umbrella term “Third Wave”, coffees within this project are those that have been produced at a “single origin”: coffee that is sourced from a single producer, farm, or crop. These coffees are further graded as “specialty coffee”: coffee of the species *Coffea arabica* that is within the top 20% of graded coffee produced worldwide. Two grading systems are generally utilized: Q-grading, as defined by the Specialty Coffee Association, and sieve grading, where larger beans are generally preferred. Different countries utilize different grades (e.g. AA, 16+, Strictly High Grown, Extra Fancy) but top grades are well distinguished from middle- and commodity-grade coffees (Hoffmann, 2018).

Single-origin coffees are intended to be traceable, meaning consumers are privy to the production cycle before purchase. This data is represented by a core of common attributes: 1) Country and Region of production; 2) Variety; 3) Harvest characteristics; 4) Processing; 5) Taste; and 6) Roast. Full descriptions of these attributes and their definitions can be found in Appendix A.

Data

No publicly available coffee dataset was available for this project and so one was built. **Figure 1** shows the end-to-end process employed to collect, analyse, model, evaluate, and present this data. This comprised of several web scraping processes (§Data Collection) that were then automated and combined into a stable dataset. This dataset was placed alongside a conformed set of Tasting Notes (§Data Conformity) before modelling.

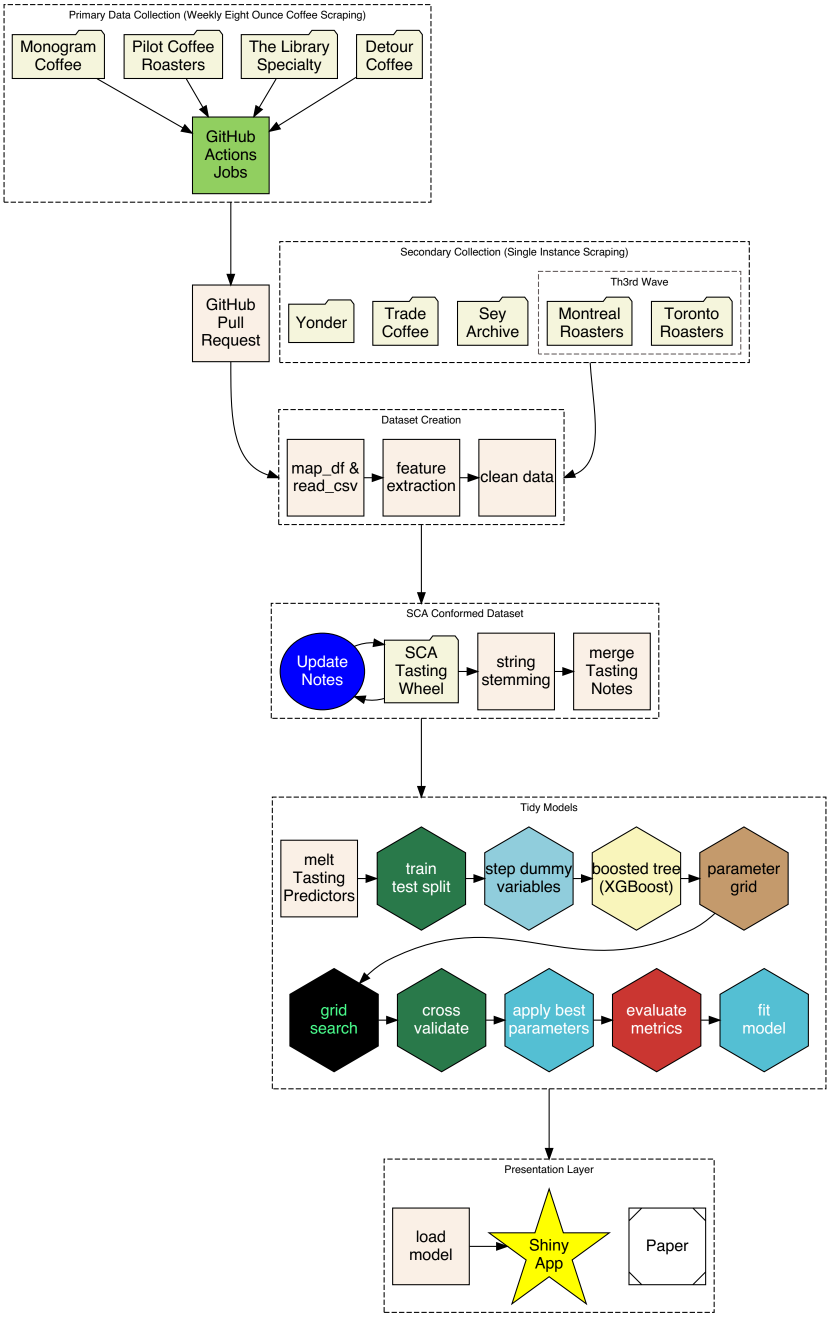


Figure 1: Coffee Model Workflow

Data Collection

Coffee data was scraped from five websites over the course of four months, comprising of 550 unique coffees from 106 roasters. Data scraping varied by website format: some websites organized each coffee by roaster, so several scrapers were built to scrape data from each roaster with consistent formatting across weeks; others were large collections of coffees that were scraped in entirety from top to bottom. Websites were selected by ease of scraping and quality of coffee offerings. Scrapers were generally run once a week, and duplicate coffees were discarded if there were any. Attributes varied by websites and roaster, but core attributes were generally present. Coffees marked as blends, espressos, and/or decaf were removed.

These scrapers were automated with GitHub Actions to run on a weekly or monthly basis, depending on the frequency at which the coffees were updated on their respective websites. Each scraper produced a .csv file per roaster or site. These .csv files were then added to the existing data and any duplicates removed. One website, not suitable for scraping but of excellent data richness, had coffee data manually collected from it. Coffee models became significantly more reliable around 500 coffees, so this manual collection was necessary to help reach this threshold within the time frame allowed.

Data collection proved a major limitation to this project. Coffee data collection should continue throughout the year to accommodate distinct coffee harvesting seasons throughout the world. For example, Brazil (the world's largest producer of arabica grade beans) is significantly underrepresented due to the timing of data collection. Furthermore, scraped websites were inconsistent in their coffee listings. Coffee attributes present in one week were not guaranteed for the next, and the order in which the attributes were listed varied from week to week. Data that was unable to conform to the established standards was discarded before modelling in order to prevent erroneous variable relationships (such as a Country being listed as a Process).

Data Conformity

To facilitate the primary focus of this work (predicting Tasting Notes), a conformed table based on the Specialty Coffee Association Taster's Flavor Wheel (2016) (**Figure 2**) and World Coffee Research Sensory Lexicon (2017) created a hierarchical categorization for Tasting Notes. Tasting Notes were placed into larger Tasting Groups. Tasting Groups were placed into larger Tasting Traits. **Table 1** provides a sample of this relationship for six Tasting Notes within the "Fruity" Tasting Trait.

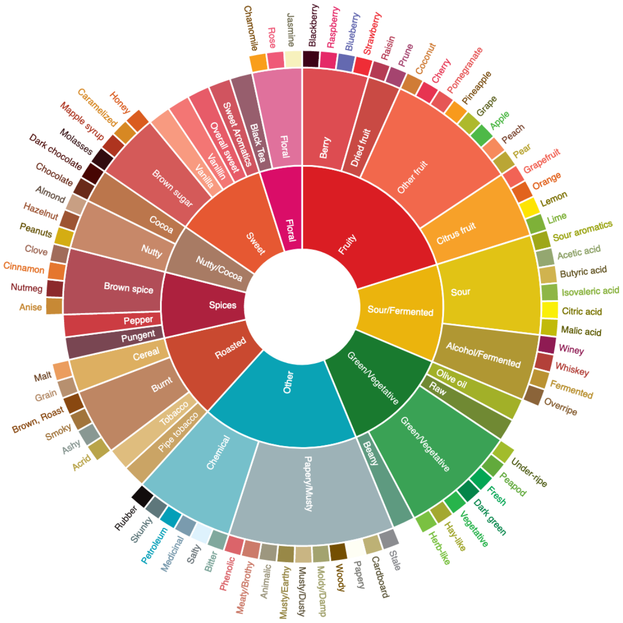


Figure 2: Specialty Coffee Association Taster's Flavor Wheel (Courtesy SCA and WCR, Creative Commons Licensing)

Table 1: Sample SCA Tasting Wheel Rows

Tasting Trait	Tasting Group	Tasting Note
Fruity	Citrus Fruit	Lemon Sorbet
Fruity	Citrus Fruit	Lemon Cookies
Fruity	Citrus Fruit	Pink Grapefruit
Fruity	Other Fruit	Fruity
Fruity	Other Fruit	Fruit
Fruity	Berry Fruit	Strawberry Jam

New Tasting Notes were added each time new data was added to the dataset. A total of 577 Tasting Notes exist within 31 Tasting Groups and 10 Tasting Traits. **Figure 3** shows the number of Tasting Notes in each Tasting Group, where Tasting Traits are represented by each coloured dot. Not all Tasting Notes were observed within the dataset: some were listed by the Specialty Coffee Association Taster's Flavor Wheel, others were added in anticipation of future coffees (for example, adding Raspberry Jam after Strawberry Jam was observed in the dataset). 427 distinct Tasting Notes were observed in the dataset.

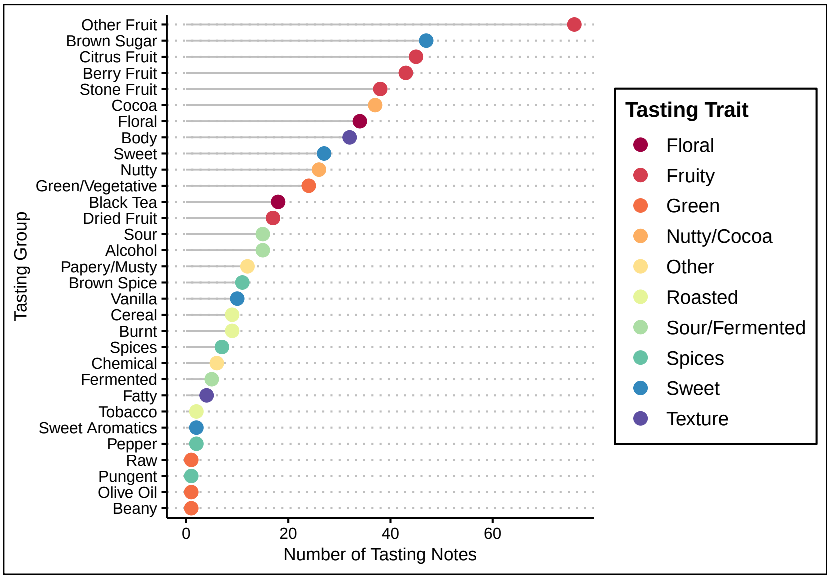


Figure 3: Hierarchy of Tasting Notes, Groups, and Traits

Adding Tasting Notes to the conformed table introduces some subjectivity. While many Tasting Notes were obvious (e.g. Black Currant is a Berry Fruit), others were difficult to place within a single category (e.g. Chocolate Orange). Attempts were made to avoid adding any Tasting Traits or Tasting Groups, instead finding space within existing Tasting Groups. Furthermore, some coffee Tasting Notes were less standard, occasionally showing regionality (e.g. *Pouding Chomeur* from a Montreal roaster) or brand names (e.g. Kit-Kat). This limitation had less impact in the model of this study, which predicted Tasting Groups rather than individual Tasting Notes, but would need to be addressed for future models (§Next Steps) by larger dataset volume and/or stricter Tasting Note filters (i.e. only accepting certain standard values).

Additionally, the World Coffee Research program's Arabica Coffee Varieties (2019) catalog was used to help standardize coffee Varieties. Because the catalog does not cover all Countries within the dataset, the most frequent Varieties from those Countries were included in modelling. Less frequent Varieties, when valid, should be included as the dataset grows. A total of 15 distinct Varieties were used in the modelling process.

Dataset

Scraped .csv files were flexible enough to allow for significant variance in coffees. By nature, coffees are curated by humans; their coffee cherries harvested at various Ripeness points, their Varieties occasionally blended to offset flavours, their Processing sometimes including a collection of multiple nearby farms. Producers and roasters make an effort to make these variables as transparent as possible. To create a standardized approach to the information rich data, data extraction focused on specific and recurring features. In total 25 different attributes were collected, but many of these were extremely sparse (less than 1%) and were not used within modelling.

First, Tasting Notes in both the dataset and the conformed table were stemmed using the `SnowballC` package (Bouchet-Valat, 2019). This removed the need to explicitly list both plural and nonplural forms of Tasting Notes (e.g. Strawberry, Strawberries). Tasting Notes were also matched using ASCII/TRANSLIT so that accented characters were not matched indiscriminately to non-accented characters (e.g. Rosé, Rose).

Second, the variable Ripeness was extracted from both Variety and Process columns. Ripeness is expressed by colouration (Pink, Red, Yellow, Orange, White, Black) as a modifier to either column (e.g. Pink Bourbon, Black Honey).

Third, Altitude, which was typically expressed in Metres Above Sea Level (MASL), was converted to a numeric value. Values that were presented as a range were averaged (e.g. 2000-2100 = 2050).

Finally, both Variety and Process columns were parsed in cases where more than one value was presented. This was fairly common for Variety, where single-origin farms blend small quantities of other Varieties with a primary Variety (e.g. Caturra + Colombia + Castillo becomes three columns: Variety1, Variety2, Variety3). It is assumed these values are listed from greatest to least, as occasionally the percentage of each Variety is specified, and in these cases the primary Variety has always been listed first. These coffees are generally not considered blends when produced by the same farm or producer. Process columns similarly can describe secondary Processes performed after main Processes (e.g. Washed + Patio Dried becomes two columns: Processing1, Processing2). In these instances, the secondary Process describes the drying method, rather than the Processing method. In other instances, the secondary Process was a synonym for the primary (e.g. Natural + Dry Processed, where, by definition, Natural is a Dry Processing method). The model used only the primary values for both Variety and Process.

Future models aim to incorporate Regions within Countries using a conformed table to create additional nuance; regional differences are well-noted by Hoffmann (2018). Similarly, in instances where Country was not explicitly listed, but rather presumed by Region, a conformed process could provide missing data (e.g. Huehuetenango would indicate Guatemala).

Figure 4 shows coffee frequency by Country prior to frequency filters required by modelling. Ethiopia and Columbia are the most frequent Countries in the dataset; this is to be expected considering when the data was collected, as Ethiopia and Columbia both have multiple growing seasons that coincided with the data collection period. A total of 25 Countries were represented in the data, out of the 35 coffee-growing Countries discussed by Hoffmann (2018). Hoffmann speaks to several of these countries having extremely limited production, especially for specialty-grade arabica beans, and thus their omission is unsurprising (e.g. Vietnam).

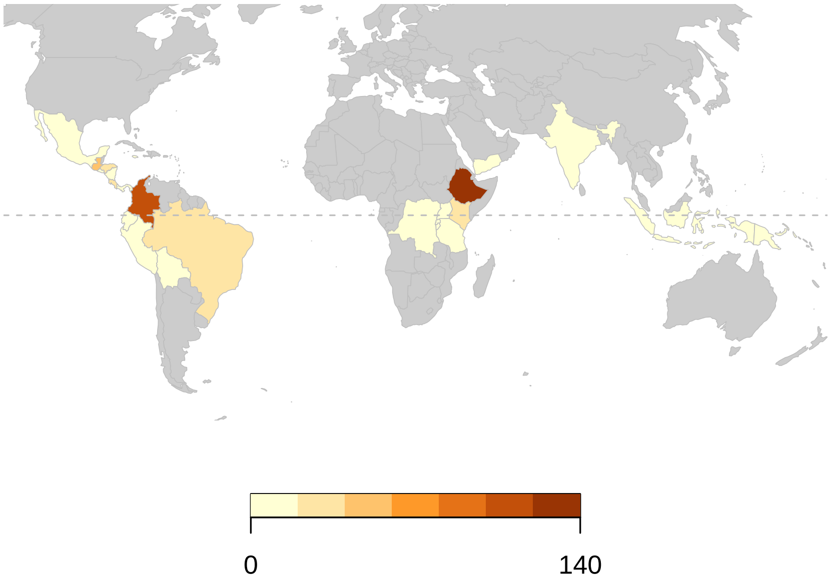


Figure 4: Map of Coffees from Each Country in the Dataset

In order to produce higher accuracy within models, extremely infrequent Processes, Countries, and Varieties were excluded. This threshold was set at five individual occurrences for each variable. This had a significant effect on the number of Ethiopian coffees included in the model dataset. Ethiopia has a long history of unique Varieties, some of which are endemic to Ethiopia, and the Variety lineage is less established. **Table 2** shows the most frequent Countries, Varieties, and Processes following this filtering process.

Table 2: Frequent Varieties, Countries, and Processes

Country	n	Variety1	n	Processing1	n
Colombia	94	Bourbon	90	Washed	293
Ethiopia	74	Caturra	88	Natural	63
Guatemala	40	Heirloom	50	Honey	25
Brazil	26	Catuai	41	Fully Washed	15
Costa Rica	24	Sl28	25	Pulped Natural	6
Kenya	23	Typica	25		

When Country, Variety, and Process are combined, **Table 3** shows that Columbia + Caturra + Washed coffees are the most frequent. Four of the six most frequent Countries and Varieties are present in the most frequent combinations of attributes; only two distinct Processes are present.

Table 3: Greatest Combination Frequencies

ItemSet	n
Colombia + Caturra + Washed	41
Ethiopia + Heirloom + Washed	31
Guatemala + Bourbon + Washed	24
Kenya + Sl28 + Washed	20
Colombia + Castillo + Washed	17
Ethiopia + Heirloom + Natural	15
Rwanda + Bourbon + Washed	15

Processing has a dramatic effect on Tasting Notes (Hoffmann, 2018), and despite large imbalances in unique Processing frequency, the variable is essential to include within modelling. This effect can be observed in the dataset when a roaster offers two distinct Processes for a coffee that is otherwise the same and presents two distinct sets of Tasting Notes (**Table 4** as an example).

Table 4: Two Different Processed Coffees

Roaster	Coffee Name	Country	Region	Variety	Processing	Tasting Notes
Pista	Fugi Ikizere Naturel	Rwanda	Nyaruguru	Bourbon	Natural	Guava, Honeydew, Hibiscus
Pista	Fugi Ikizere Lave	Rwanda	Nyaruguru	Bourbon	Washed	Apricot, Green Apple, Jasmine

As expressed previously, adding Region to models was desirable. However, combination frequencies (**Table 5**) demonstrate the limitations of a dataset this size, where very few Regions had more than five occurrences when combined with Variety and Process.

Table 5: Combination Frequencies with Region > 5

ItemSet	n
Ethiopia + Guji + Heirloom + Washed	13
Kenya + Nyeri + Sl28 + Washed	12
Guatemala + Huehuetenango + Bourbon + Washed	10
Colombia + Huila + Caturra + Washed	6
Ethiopia + Guji + Heirloom + Natural	6

This sparseness in data also affected the model's ability to predict Tasting Notes. Instead, Tasting Groups were determined to be the next best candidate for prediction, having considerably more nuance than Tasting Traits but still frequently occurring for nearly all class labels. As such, for all coffees in the dataset, Tasting Groups were stacked so that a single Tasting Group column was the predictor. This process is demonstrated as **Table 6** transforms into **Table 7**.

Table 6: Coffee Sample before Stacking

Country	Variety1	Processing1	TastingGroup1	TastingGroup2	TastingGroup3
Colombia	Castillo	Natural	Stone Fruit	Cocoa	Fermented

Table 7: Coffee Sample after Stacking

Country	Variety1	Processing1	TastingGroup
Colombia	Castillo	Natural	Stone Fruit
Colombia	Castillo	Natural	Cocoa
Colombia	Castillo	Natural	Fermented

Figure 5 shows the frequency of all Tasting Groups following this transformation prior to modelling. Seven Tasting Groups—Other Fruit, Brown Sugar, Stone Fruit, Berry Fruit, Cocoa, and Floral—are much more prominent than other Tasting Groups. This imbalance has a strong influence on the model (§Results), as very few coffees are predicted with Tasting Groups aside from these seven. However, some Tasting Notes are rare for coffees as a whole, but highly frequent for specific coffee attribute combinations. For example, Black Tea is an often-curated Tasting Note for Ethiopian coffees (Hoffmann, 2018). As such, the lack of presence within the overall dataset does not indicate that a given Tasting Group will not be predicted.

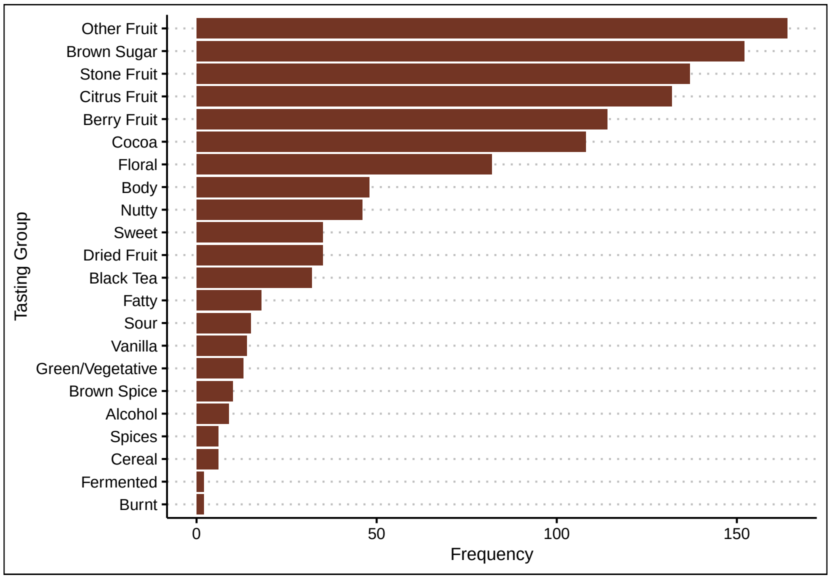


Figure 5: Tasting Group Frequency Prior to Modelling

Model

Using `tidymodels` (Kuhn & Wickham, 2020), a workflow was developed wherein a Gradient Boosted Tree model with an `XGBoost` (Chen et al., 2021) engine predicted Tasting Groups using Country, first Varietal, and first Process for all coffees (Predictors: *Variety1*, *Processing1*, *Country*; Outcome: *TastingGroup*). The dataset was split into training and testing sets using a 75/25 non-stratified split. All predictor values were converted to dummy values (as they were all nominal). The Gradient Boosted Tree was tuned using grid search hyperparameterization on the number of trees (trees), the splitting criteria for each node (min_n), the maxiumum depth of trees (tree_depth), and the learning rate between iterations (learn_rate). 256 total parameter combinations were run (four parameters with four selections each, 4^4). During grid search, 10-fold cross validation was performed. The model was then evaluated on the testing dataset before the best parameter set was selected based on the highest ROC AUC (accuracy under curve) value. The finalized model was then fit to the entire dataset.

XGBoost (Chen & Guestrin, 2016) uses the same principals as other Gradient Boosted Trees: loss minimization through gradient descent. XGBoost balances the need to explore all base learners (training loss) and the performance requirements of calculating a loss function for each point of gradient descent (regularization). This is achieved by calculating a residual similarity score after each leaf splitting criteria is evaluated. The node with the greatest information gain (maximum loss reduction) is selected greedily (so not all trees or learning rates are explored). As an ensemble model, each coefficient term is added to the previous to adjust error rates created by previous model iterations.

The final prediction, achieved by fitting base learners to the minimum loss by gradient descent, is as follows:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i)$$

- $\hat{y_i}$ is the predicted value
- K is the number of K additive functions used to predict the output
- f_k is the space of trees (structure: depth, leaf weight)
- x_i is the set of attributes defining the tree (the dataset).

Trees are defined by their splitting criteria (leaf weight) and structure (tree depth), such that:

$$f_t(x_i) = w_{q(x)}$$

- w is the leaf weight
- $q(x)$ is the tree structure.

XGBoost seeks to minimize its regularized function:

$$L(\phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k)$$

- $l(\hat{y_i}, y_i)$ is the difference between prediction and actual

plus a loss coefficient (learning rate) as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

- $\Omega(f_t)$ is the loss coefficient applied to the tree
- γT is the number of leaves in the tree
- $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is the Euclidean norm for leaf scores.

The ensemble method is then calculated through boosting:

- $\hat{y_i}^{(0)} = 0$
- $\hat{y_i}^{(1)} = \hat{y_i}^{(0)} + f_1(x_i)$
- $\hat{y_i}^{(2)} = \hat{y_i}^{(1)} + f_2(x_i)$

until finalized terms are added to the model:

$$\hat{y}^{(t)} = \sum_{i=1} \lambda(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

- $\hat{y}^{(t)}$ is the model at training

- $\sum_{i=1} \lambda(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$ adds each boosting function, including the functions added in previous iterations ($\hat{y}_i^{(t-1)}$)
- $f_t(x_i)$ as the new function
- $\Omega(f_t)$ as the loss coefficient previous calculated.

As such, **Table 8** shows the best selected parameters for the model following hyperparametization through grid search and optimized for ROC AUC.

Table 8: XGBoost Model Parameters

trees	min_n	tree_depth	learn_rate
2000	14	1	0.1

Results

The Gradient Boosted Tree model produces ranked predictions (from most to least probable) for 22 Tasting Groups. Because most coffees have three Tasting Notes, the model's three most probable predictions are treated as the overall prediction for any given combination of coffee attributes. Not all Tasting Groups were predicted, generally due to sparsity or lack of ability to differentiate coffees with unique Tasting Groups from coffees with the same attributes and more typical Tasting Groups. **Table 9** shows the most frequent predictions compared to their actual frequencies. Other Fruit, a catch-all category for the myriad of fruits that present themselves in coffees, is both the most frequent Tasting Group in the dataset and the most predicted by the model. The model overestimates the top four Tasting Groups (Other Fruit, Brown Sugar, Stone Fruit, Citrus Fruit) compared to their actual frequencies in the dataset.

Table 9: Model Prediction by Tasting Group

TastingGroup	Predictions	PredictionFreq	ActualFreq
Other Fruit	290	0.244	0.139
Brown Sugar	269	0.226	0.129
Stone Fruit	253	0.213	0.116
Citrus Fruit	166	0.140	0.112
Cocoa	83	0.070	0.092
Berry Fruit	67	0.056	0.097
Floral	52	0.044	0.069
Nutty	8	0.007	0.039

A confusion matrix (**Figure 6**) run on the conserved test data appears to show that only seven Tasting Groups were predicted. These Tasting Groups coincided with the most frequently occurring Tasting Groups (as shown by **Figure 5**). Like **Table 9**, Brown Sugar and Other Fruit classes were the most highly predicted results. But confusion matrices only show the most probable prediction for each coffee within the test data. The overall accuracy metrics that can be derived from this confusion matrix are thus quite low. Berry Fruit, for example, appeared 22 times within the test data, six of which were predicted correctly out of 29 total predictions: a precision rate of 0.207 and a recall rate of 0.273. Some classes were never predicted correctly as the most probable prediction (e.g. Dried Fruit).

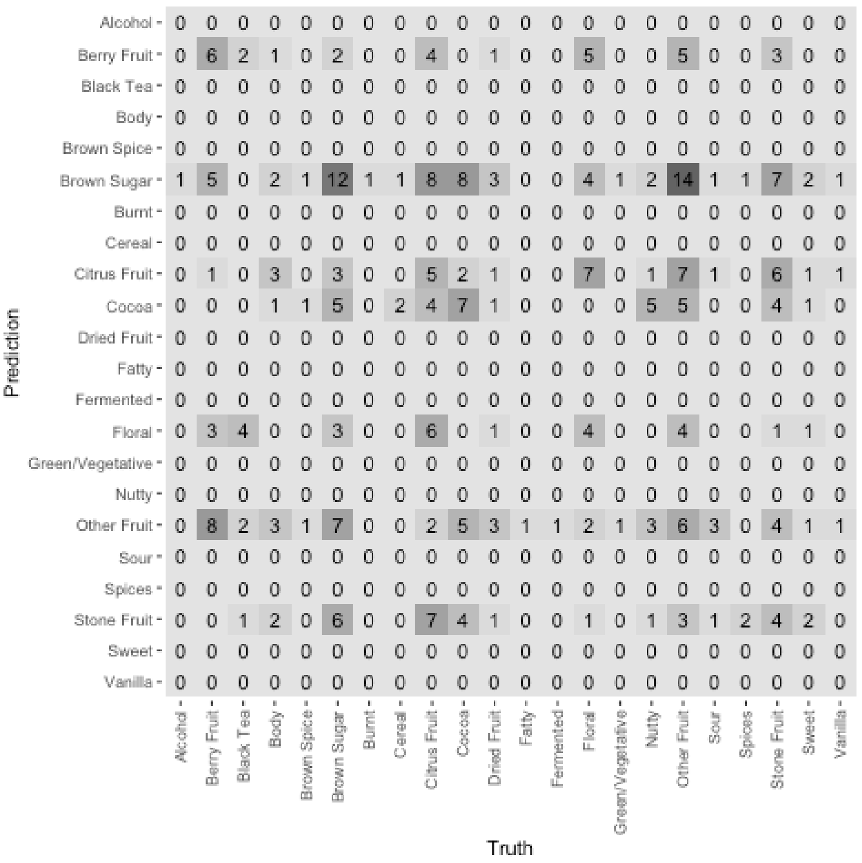


Figure 6: Confusion Matrix

However, both **Table 9** and the confusion matrix are poor representations of the model's performance, and the accuracy score for the model cannot be conventionally calculated. Instead, a better accuracy metric must measure the three most probable predictions while simultaneously deemphasizing the order in which Tasting Groups appear. These needs are specific to coffee: Tasting Group order is irrelevant because there is neither standardization (e.g. fruits are not listed before sugars) nor hierarchy (i.e. the first note is not more prevalent than the second). Therefore, the most probable prediction does not need to coincide with the first of three Tasting Groups.

All possible unique inputs were provided to the model, and the top three most probable predictions were tabulated alongside their actual (truth) values. A Boolean flag checked whether the first predicted value was in any of the three Tasting Groups. This process was repeated for the second and third predicted values. This method also correctly predicts Tasting Groups for coffees that do not have three unique Tasting Groups. Tasting Groups do not need to be unique because they represent Tasting Notes that are unique (e.g. Black Currant and Raspberry are both Berry Fruits, therefore both the first and second Tasting Group would be Berry Fruit). Because the model cannot predict the same Tasting Group twice, the Boolean flag allowed duplicate Tasting Groups found within a given coffee to be counted for each actual occurrence of a Tasting Group, as if the model had predicted duplicate values. An example coffee (**Table 10**) demonstrates the recalculated accuracy metric. For a coffee with given Country + Variety + Process attributes: Rwanda + Bourbon + Honey, predictions were Brown Sugar, Other Fruit, and Stone Fruit. Actual values were Brown Sugar, Other Fruit, and Other Fruit. Because the model has predicted both Brown Sugar and Other Fruit, the accuracy is 100% for the given set of coffee attributes.

Table 10: Rwanda + Bourbon + Honey Accuracy Measurement

Prediction	Probability	TG_Actual1	TG_Actual2	TG_Actual3	TG_Correct1	TG_Correct2	TG_Correct3
Brown Sugar	0.259	Brown Sugar	Other Fruit	Other Fruit	TRUE	FALSE	FALSE
Other Fruit	0.233	Brown Sugar	Other Fruit	Other Fruit	FALSE	TRUE	TRUE
Stone Fruit	0.106	Brown Sugar	Other Fruit	Other Fruit	FALSE	FALSE	FALSE

Thus, the total number of true predictions was 534 of 1188 for a model accuracy score of 44.949% (**Table 11**).

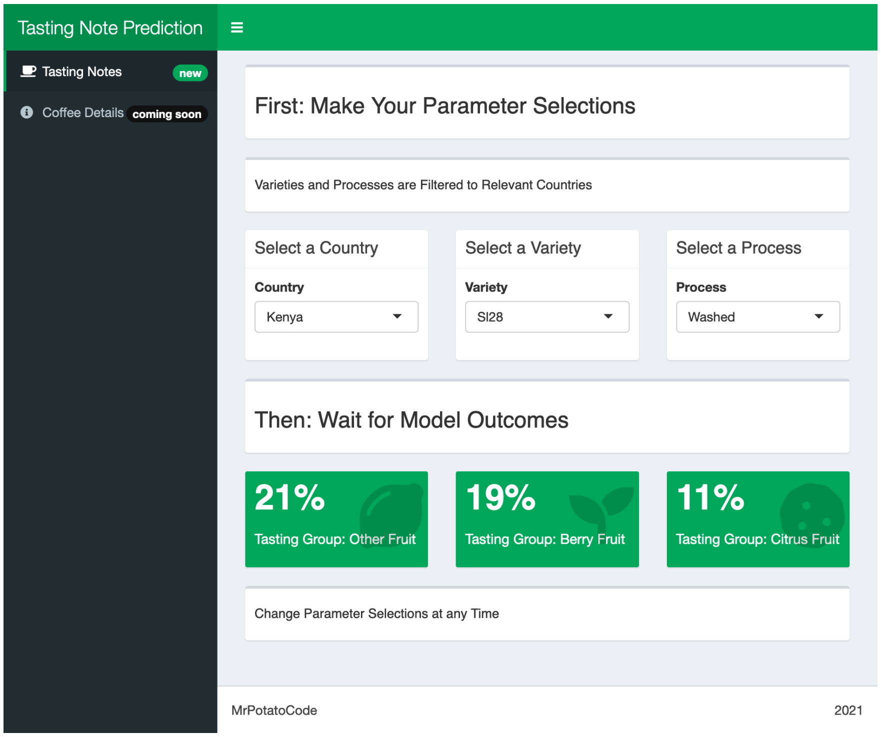
Table 11: Adjusted Model Accuracy

TG_Correct1	TG_Correct2	TG_Correct3	TotalCorrect	TotalTastingGroups	Accuracy
213	172	149	534	1188	44.949%

Shiny App

The model was placed in a Shiny App (**Figure 7**) to allow exploration of any given attribute combination. 98 distinct selections are possible, consisting of 20 Countries, 15 Varieties, and 5 Processes. The Shiny App requires a user to first select a Country from a dropdown and then provides the list of possible Varieties within that Country. Processes are then filtered based on Variety selection. Combinations that did not exist within the dataset could not be passed to the model (as many of these would not occur in the real world, and thus predictions would be irrelevant).

The top three model predictions were listed alongside their probability for occurring. For example, a Kenya + SL28 + Washed selection shows a prediction of Other Fruit (21%), Berry Fruit (19%), and Citrus Fruit (11% probability). A new set of parameters could be selected at any time. The Shiny App does not show model accuracy or the degree that Tasting Groups vary within the given set of attributes (§Findings and Implications) in its current design; this should be added in the future in order to give a better sense of confidence to predictions beyond probability.



First: Make Your Parameter Selections

Varieties and Processes are Filtered to Relevant Countries

Select a Country

Country: Kenya

Select a Variety

Variety: Sl28

Select a Process

Process: Washed

Then: Wait for Model Outcomes

21% Tasting Group: Other Fruit

19% Tasting Group: Berry Fruit

11% Tasting Group: Citrus Fruit

Change Parameter Selections at any Time

MrPotatoCode 2021

Figure 7: Tasting Note Predictions Shiny App

The Shiny app was deployed using shinyapps.io by RStudio and is available at: https://mrpotatocode.shinyapps.io/TastingNotePredictions/.

Discussion

While the results of this study were limited by the size of the dataset, the data scrapers and models developed will provide a solid foundation for future research. Evaluating the significance of this work is difficult, but models for certain combinations appear to meet known expectations and are comparable to previous work by Gagné (Varietal + Process to predict Tasting Notes) and Conley and Wilson (Country to predict Tasting Notes).

Gagné produced flavor wheels for five common Varieties (Bourbon, SL28, Heirlooms, Caturra, Geisha) of Washed coffees, and word clouds for 14 Varieties and 11 Processes (five of which were Ripeness + Honey, and two other infrequent but well-known processes - carbonic and anaerobic). Gagné's work did not include prediction, but rather frequency visualizations of 1500 coffees. When comparing these flavor wheels to the model, two matches stand out: Kenya + SL28 + Washed is predominantly Berry Fruit with Other Fruit and Citrus Fruit Tasting Groups (see **Figure** 7 for model output); similarly, Ethiopia + Heirloom + Washed is predominantly Floral with Stone Fruit and Citrus Fruit, though Black Tea was not predicted as the dominant Tasting Group. It is difficult to compare results beyond SL28 and Heirloom Varieties, which appear almost exclusively in Kenya and Ethiopia, respectively. Varieties that appear throughout the world, such as Bourbon or Caturra, have much more variety in their predictions than Gagné's work, which generalizes Bourbon regardless of Country of origin. Gagné's work cannot be compared to model predictions for Processes other than Washed.

Conley and Wilson's models classified Costa Rican Tasting Notes as Stone Fruit (Cherry), Sweet/Brown Sugar (Sugar Cane, Brown Sugar, Syrup, Sweet), and Vanilla. Without knowing either Variety or Process, the best way to draw comparisons with the model was to find the most commonly occurring Costa Rican coffees in the dataset: Caturra + Natural, Catuai + Honey, Catuai + Washed, Caturra + Washed, Caturra

+ Honey. Of these, four of five predictions produced Brown Sugar as a Tasting Group, and four of five produced Stone Fruit (**Table 12**). None produced predictions of Vanilla. Because Conley and Wilson have used a very small subset of Tasting Notes, placing them within Tasting Groups potentially eliminates the nuance of their analysis, while substituting the complexity of this model. This may suggest that coffees are easier to overgeneralize than specifically predict, but without a greater number of Tasting Notes and more specificity of Variety and Process, it is hard to differentiate the model results from one another.

Table 12: Most Frequently Occurring Costa Rican Coffees and their Predictions

ItemSet	TG_Pred1	TG_Pred2	TG_Pred3
Costa Rica + Catuai + Honey	Other Fruit	Brown Sugar	Stone Fruit
Costa Rica + Catuai + Washed	Brown Sugar	Other Fruit	Stone Fruit
Costa Rica + Caturra + Honey	Other Fruit	Brown Sugar	Stone Fruit
Costa Rica + Caturra + Natural	Other Fruit	Berry Fruit	Nutty
Costa Rica + Caturra + Washed	Other Fruit	Brown Sugar	Stone Fruit

Since the combination of these two bodies of work has not yet been established elsewhere, the model provides a rudimentary framework for more multi-class predictions. Comparisons of the model alongside both Gagné's (2019) and Conley and Wilson's (2018) work, as well as personal coffee journals, show that predictive results are reproducible. Thus, despite what appears to be a low overall accuracy rate, model performance is not so prohibitive to suggest that the combination of attributes cannot produce a reliable result.

Findings and Implications

Contextualizing the model's results within the dataset is a complex endeavour. While measuring the model's accuracy against individual coffees is useful to quantify the model's quality on the whole, the prevalence of three Tasting Groups is not necessarily a good determinate of any set of coffee attributes. To draw any valid conclusions, a set of coffee attributes needs to have a low diversity of Tasting Groups *and* the model needs to predict the most frequent (or majority) of Tasting Groups as they appear within the dataset.

Two coffees within **Figure 8** show a simplified version of this problem. The first, Costa Rica + Caturra + Natural, is predicted by the model to taste of Other Fruit, Berry Fruit, and Nutty (see **Table 12** for all Costa Rican predictions). The first two Tasting Groups are the most frequent (and thus largest squares within the figure), but Body, Brown Sugar, Dried Fruit, and Vanilla are equally as frequent as Berry Fruit (all coloured as light green), and the third model prediction, Nutty, does not appear at all. As such, the predicted Tasting Groups represent the minority of occurring Tasting Groups (the overall square describing the coffee's attributes is less than half of the predicted Tasting Groups). This suggests that the model does not represent these coffees with sufficient nuance; simply put, there are too many other valid Tasting Groups for the itemset.

This can be contrasted with Costa Rica + Catuai + Honey, where Other Fruit, Brown Sugar, and Stone Fruit (the model predictions) represent the majority of occurring Tasting Groups (only Dried Fruit occurs more frequently). Conclusions made based on the model's outcomes for this attribute set will be more reliable as the coffees show less Tasting Group diversity.

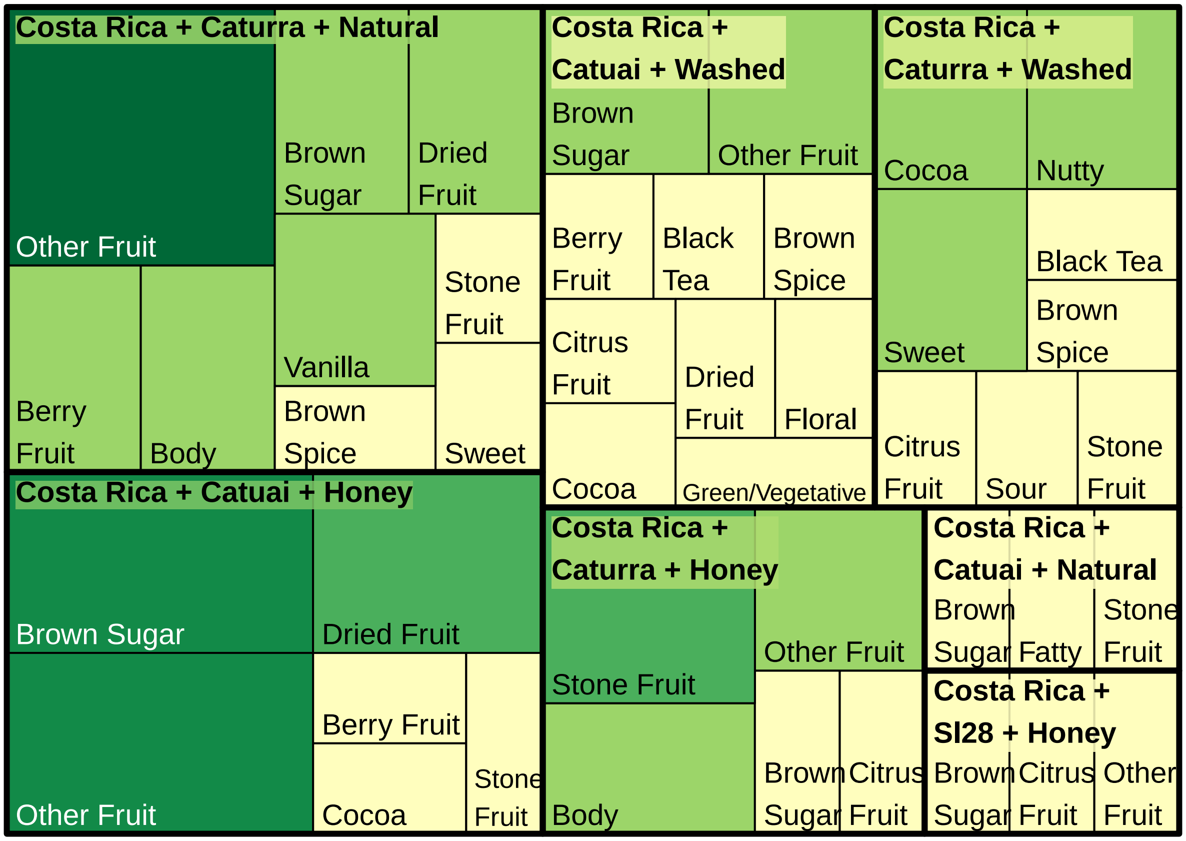


Figure 8: Treemap of Costa Rican Coffee's Tasting Groups as they Appear within the Dataset

This effect is problematic. When examining the model, several well-known Tasting Group combinations for coffee attribute sets appeared to be correct (e.g, Kenya + SL28 + Washed). However, it is difficult to say whether this demonstrates consistency or confirmation bias. For lesser-known, more interesting coffee combinations, or highly varying coffee (such as Costa Rica + Catuai + Washed in **Figure** 8), there is insufficient knowledge (and data) to evaluate predicted values.

Broader implications, such as using this model within a commercial environment, or using it to inform farming decisions, should be avoided at this time. Accurately describing a coffee's taste alongside its attributes can logically be applied with reinforcement algorithms when customer preferences are known and collected. This dataset is not suitable for this application. In general, if conclusions are to be made at all, they should be made cautiously, with considerable domain knowledge, and with awareness that the model requires a much larger dataset (§Next Steps).

There is, however, some novelty in the model's creation, and with some optimism, it begins to unravel some of the complications of specialty coffees. Conley and Wilson highlight the importance of accurate models as, "enabl[ing] the development of formal appellations to confirm each country's unique coffee profile" (2018). The model should not stifle creativity, but instead reaffirm the anecdotes of well-trained coffee tasters, and can serve as evidence when describing coffee complexities to skeptical audiences.

Next Steps

As mentioned previously, continued data collection is paramount to develop this model further. Additional sources of weekly data should be produced much in the same way as the original coffee scrapers. One-off entire site scraping is not required, though if suitable and easy-to-scrape sites are identified, the model will improve from greater coffee data depth. Weekly scrapers generally focused on a single roaster, though this was not necessarily intentional. Using a small collection of single roasters was intended to control the quality of coffee, but as specialty coffee has become much more commonplace, new data sources for the same coffees can be identified. A partnership with a group of roasters who could reliably provide data, rather than scraping, is another avenue worth considering. Regardless of source, data richness should be of utmost importance. As discussed, Region is the next most logical variable to be added to the model. Altitude should follow shortly thereafter, but correlation between variables should be considered.

Region also allows the introduction of conformed GIS data describing a coffee's terroir (Mighty, 2015). Though regional climates are likely generalizable for all coffee growing Regions, small differences may vastly improve Tasting Group or Tasting Note prediction. Should more coffee roasters begin to regularly provide latitude and longitude pairs for farm locations, the precision of terroir details can substantially improve, and further aid the model's ability to differentiate like-coffees from one another.

In addition to capturing an entire year's coffee harvest and production, the dataset should be considered nascent until at least 2000 coffees have been collected. This strikes a balance between sparsity and zero variance, allowing for the Boosted Tree model to explore relationships between some of these variables. This quantity will also allow for other well-recognized Varieties and Processes to become prevalent within the dataset (e.g. Batian and SL34, Anaerobic and Carbonic). Some filtering to reduce extremely experimental Varieties or Processes will still be required. It may also prove interesting to model coffees beyond their first Variety, especially for frequently occurring itemsets. For example, SL34 is often paired with SL28 in Kenyan Washed coffees. Understanding how slight modifiers, like Ripeness or a second Variety, alter Tasting Groups/Notes could help to delineate coffees that are currently overgeneralized.

All data collection efforts should aspire to achieve the objective of this study at its conception: predict Tasting *Notes* rather than Tasting Groups. Tasting Groups, while essential for the success of this project, are still too reductive. Specialty coffees are interesting because of their nuance: a coffee tasting of Apple is quite different than a coffee tasting of Banana, yet both of these are categorized as Other Fruit. For the model to be used for any utilitarian purposes, this nuance must be preserved.

In order to help models predict Tasting Notes, it will likely prove prudent to add a fourth layer to the conformed tasting table to reduce the overall complexity of Tasting Notes. Some Tasting Notes are overly specific: for example, Bosc Pear, White Pear, Baked Pear, Bartlett Pear, Yellow Pear, Asian Pear, and Pear Tart all describe Pear. This has two effects: firstly, it will help the model make nuanced predictions within slightly broader categories; secondly, the increased granularity will provide a clearer sense of how model predictions for coffees with the same Tasting Group should be evaluated (such as the example coffee in **Table 10**).

With these steps taken, it may become possible to examine whether a specialty coffee's Tasting Notes can be predicted with both nuance and precision based on the coffee's attributes. The possible applications of this work can benefit producers and farmers by shifting purchases from a one-time transaction to a long-term investment at each stage of the coffee lifecycle, and also to academics exploring coffee composition at the molecular level. Above all else, predictable coffee can benefit consumers, who, with a data-rich cup of coffee, have already begun to promote a seed-to-cup approach and usher in the next wave of coffee: *ethical consumption.*

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Iannone, R. (2020). *Rmarkdown: Dynamic documents for r*. Retrieved from https://github.com/rstudio/rmarkdown

Arnold, J. B. (2019). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggthemes

Bengtsson, H. (2020). *DoFuture: A unifying framework for parallel and distributed processing in r using futures*. Retrieved from https://arxiv.org/abs/2008.00553

Bhumiratana, N., Adhikari, K., & Chambers, E. (2011). Evolution of sensory aroma attributes from coffee beans to brewed coffee. *LWT - Food Science and Technology*, *44*(10), 2185–2192. https://doi.org/10.1016/j.lwt.2011.07.001

Bion, R. (2019). *Ggradar: Create radar charts using ggplot2*. Retrieved from https://github.com/ricardo-bion/ggradar

Bivand, R., Keitt, T., & Rowlingson, B. (2021). *Rgdal: Bindings for the 'geospatial' data abstraction library*. Retrieved from https://CRAN.R-project.org/package=rgdal

Bouchet-Valat, M. (2019). *SnowballC: Snowball stemmers based on the c 'libstemmer' utf-8 library*. Retrieved from https://CRAN.R-project.org/package=SnowballC

Chang, W., & Borges Ribeiro, B. (2018). *Shinydashboard: Create dashboards with 'shiny'*. Retrieved from https://CRAN.R-project.org/package=shinydashboard

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). *Shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *CoRR*, *abs/1603.02754*. Retrieved from http://arxiv.org/abs/1603.02754

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Li, Y. (2021). *Xgboost: Extreme gradient boosting*. Retrieved from https://CRAN.R-project.org/package=xgboost

Conley, J., & Wilson, B. (2018). Coffee terroir: Cupping description profiles and their impact upon prices in central american coffees. *GeoJournal*, (85), 67–79. https://doi.org/http://dx.doi.org/10.1007/s10708-018-9949-1

Croijmans, I., & Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *PLOS ONE*, *11*(6), e0155845. https://doi.org/10.1371/journal.pone.0155845

Dowle, M., & Srinivasan, A. (2019). *Data.table: Extension of 'data.frame'*. Retrieved from https://CRAN.R-project.org/package=data.table

Gagné, J. (2019). How coffee varietals and processing affect taste. Retrieved from https://coffeeadastra.com/2019/07/23/how-coffee-varietals-and-processing-affect-taste-2/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Hal Daumé III, & Crawford, K. (2020). *Datasheets for datasets*.

Goldstein, J. E. (2011). The "coffee doctors": The language of taste and the rise of rwanda's specialty bean value. *Food and Foodways*, *19*(1), 135–159. https://doi.org/10.1080/07409710.2011.544226

Granjon, D. (2021). *ShinydashboardPlus: Add more 'adminlte2' components to 'shinydashboard'*. Retrieved from https://CRAN.R-project.org/package=shinydashboardPlus

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Grothendieck, G. (2017). *Sqldf: Manipulate r data frames using sql*. Retrieved from https://CRAN.R-project.org/package=sqldf

Hoffmann, J. (2018). *The world atlas of coffee: From beans to brewing – coffees explored, explained and enjoyed*. Firefly Books, Inc.

Iannone, R. (2020). *DiagrammeR: Graph/network visualization*. Retrieved from https://CRAN.R-project.org/package=DiagrammeR

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. Retrieved from https://www.tidymodels.org

Mighty, M. A. (2015). Site suitability and the analytic hierarchy process: How gis analysis can improve the competitive advantage of the jamaican coffee industry. *Applied Geography*, *58*, 84–93. https://doi.org/https://doi.org/10.1016/j.apgeog.2015.01.010

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287596

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Pendergrast, M. (2010). *Uncommon grounds: The history of coffee and how it transformed our world*. Basic Books.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rosenberg, L., Swilling, M., & Vermeulen, W. J. V. (2018). Practices of third wave coffee: A burundian producer's perspective. *Business Strategy and the Environment*, *27*(2), 199–214. https://doi.org/10.1002/bse.2010

Schmidt, D. (2017). *Introducing the float package: 32-bit floats for R*. Retrieved from https://cran.r-project.org/package=float

South, A. (2011). Rworldmap: A new r package for mapping global data. *The R Journal*, *3*(1), 35–43. Retrieved from http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf

Specialty Coffee Association. (2003). *Protocols & best practices*. Retrieved from https://sca.coffee/research/protocols-best-practices

Specialty Coffee Association. (2016). *The coffee taster's flavor wheel*. Retrieved from https://sca.coffee/research/coffee-tasters-flavor-wheel

Tennekes, M. (2017). *Treemap: Treemap visualization*. Retrieved from https://CRAN.R-project.org/package=treemap

Traore, T. M., Wilson, N., & Fields, D. (2018). What explains specialty coffee quality scores and prices: A case study from the cup of excellence program. *Journal of Agricultural and Applied Economics*, *50*(3), 349–368. https://doi.org/http://dx.doi.org/10.1017/aae.2018.5

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Rvest: Easily harvest (scrape) web pages*. Retrieved from https://CRAN.R-project.org/package=rvest

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., Hester, J., & Ooms, J. (2020). *Xml2: Parse xml*. Retrieved from https://CRAN.R-project.org/package=xml2

World Coffee Research. (2017). *World coffee research sensory lexicon 2.0*. Retrieved from https://worldcoffeeresearch.org/media/documents/20170622_WCR_Sensory_Lexicon_2-0.pdf

World Coffee Research. (2019). *Arabica coffee varieties*. Retrieved from https://varieties.worldcoffeeresearch.org/varieties

Xie, Y. (2020a). *Bookdown: Authoring books and technical documents with r markdown*. Retrieved from https://github.com/rstudio/bookdown

Xie, Y. (2020b). *Knitr: A general-purpose package for dynamic report generation in r*. Retrieved from https://yihui.org/knitr/

Zhu, H. (2020). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Appendix A: Single-Origin Coffee Terminology

Terms within this project have been capitalized to aid in recognition.

1) Country and Region of production, where the coffee was cultivated.

2) Variety, the subspecies cultivated through selection (e.g. Typica, Bourbon, Caturra, SL28, Geisha, Heirloom). Variety may also be referred to as varietal, a single instance of a Variety, for example within a crop or a farm, rather than the subspecies. There are an unknown number of Varieties in the world; however, there are a handful of popular, well-identified Varieties whose morphologies are well-documented.

3) Harvest characteristics, including the following: Ripeness colouration (e.g. Black, Red, Orange, Yellow, White—from most to least ripened), farm's location (latitude and longitude) and terroir (i.e. temperature, rainfall, soil composition, days of sunshine, Altitude [MASL]), season of harvest, and picking methods (e.g. hand, stripping, mechanical). Farmers (sometimes called estate owners) control these factors.

4) Processing (Natural, Washed, Honey, Pulped) and drying (patio, raised, mechanical) methods. Farmers or producers determine these factors, depending on available infrastructure. In the case where farmers are not able to Process and dry their crop, producers (often within co-ops) handle the Processing of beans and are thus sometimes considered the "origin" when multiple lots are collected and Processed simultaneously.

5) Roast, where all beans have been roasted (i.e. have undergone the Maillard Reaction). Roasters determine duration and temperature. Roasters are occasionally referred to as producers, especially within relationship-coffee purchases. For the sake of simplicity and with due consideration of the effect of roast, the scope of this project will endeavour to exclude coffees roasted beyond the "first crack".

6) Taste, as identified by cupping, to identify Tasting Traits (e.g. Sweet, Fruity, Texture, Roasted, Nutty/Cocoa), Tasting Groups (e.g. Citrus, Berry, Stone Fruits, Brown Sugar, Floral), and Tasting Notes (e.g. Chocolate, Raspberry, Lemon, Black Tea, Hazelnut, Velvety).

Appendix B: Datasheet for Dataset, v0.1

Available here: https://github.com/mrpotatocode/COFFEE_COFFEE_COFFEE/blob/main/journal/Week8/DataSheet-0.1.md

Appendix C: Model Card, v0.1

Available here: https://github.com/mrpotatocode/COFFEE_COFFEE_COFFEE/blob/main/journal/Week12/ModelCard.md