Department of Economics, University of Warwick

Monash Business School, Monash University

as part of

Monash Warwick Alliance

Predicting Specialty Coffee Auction Prices Using Machine Learning

Zoltan Aldott

Warwick-Monash Economics Student Papers

September 2021 No: 2021-15

ISSN 2754-3129 (Online)

https://warwick.ac.uk/fac/soc/economics/research/wmesp/

The Warwick Monash Economics Student Papers (WM-ESP) gather the best Undergraduate and Masters dissertations by Economics students from the University of Warwick and Monash University. This bi-annual paper series showcases research undertaken by our students on a varied range of topics. Papers range in length from 5,000 to 8,000 words depending on whether the student is an undergraduate or postgraduate, and the university they attend. The papers included in the series are carefully selected based on their quality and originality. WM-ESP aims to disseminate research in Economics as well as acknowledge the students for their exemplary work, contributing to the research environment in both departments.

"*We are very happy to introduce the Warwick Monash Economics Student Papers (WM-ESP). The Department of Economics of the University of Warwick and the Economics Department at Monash University are very proud of their long history of collaboration with international partner universities, and the Monash Warwick Alliance reflects the belief in both Universities that the future will rely on strong links between peer Universities, reflected in faculty, student, and research linkages. This paper series reflects the first step in allowing our Undergraduate, Honours, and Masters students to learn from and interact with peers within the Alliance.*"

Jeremy Smith (Head of the Department of Economics, University of Warwick) and Michael Ward (Head of the Department of Economics, Monash University)

Recommended citation: Aldott, Z. (2021). Predicting specialty coffee auction prices using machine learning. *Warwick Monash Economics Student Papers* 2021/15

[1] Warwick Economics would like to thank Lory Barile, Gianna Boero, and Caroline Elliott for their contributions towards the selection process.

Predicting Specialty Coffee Auction Prices Using Machine Learning

Zoltan Aldott*

Abstract

This paper aims to contribute to the coffee pricing literature pertaining to the Cup of Excellence (CoE) competitions by revising the feature set used and extending the modelling approach using machine learning. The specific dataset used is merged from data provided by the Alliance for Coffee Excellence and information collected through scraping public information from the Cup of Excellence website. The paper compares popular supervised learning algorithms exploring multiple interpretations of tasting notes to attain an efficient predictive model of prices. The algorithms compared include OLS, regularised linear algorithms, the decision tree, as well as, bagging and gradient-boosting ensemble methods. The best-performing models are further optimised using hyperparameter tuning and the most efficient one is selected. Based on a gradient-boosting regression, the final model is analysed to find the key relationships driving model predictions. Permutation feature importance and accumulated local effects analyses are used to provide insights into the non-linearities present in the data generating process.

JEL codes: C53, C81, D44, Q11

Keywords: specialty coffee, machine learning, prediction, Coffee Taster's Flavor Wheel, Cup of Excellence

*Contact**:** zoltan.aldott.1@gmail.com

Link to online appendix and Python/R code: https://github.com/zozi0406/Predicting-Specialty-Coffee-Auction-Prices-Using-Machine-Learning

Note: The dataset is not public and cannot be found in the GitHub repository. .

The author is grateful to Prof. Wiji Arulampalam for her supervision and continued support, and to the Alliance for Coffee Excellence (ACE) for making the Cup of Excellence dataset available for use in this research.

1. Introduction

As per capita annual consumption levels reach 5 kg in the EU (CBI, 2020), coffee has become integral to our lifestyle. In recent decades coffee prices (ICO, 2020) have seen numerous fluctuations, and the mainstream coffee market has become highly competitive. One of the main ways to attain a price premium is to join the specialty market (Wollni and Zeller, 2007). To classify as specialty coffee, a product needs a Specialty Coffee Association quality score of 80 out of 100. Recently, this segment has been outgrowing the mainstream coffee market substantially and is expected to continue growing (CBI, 2020). Understanding the specialty coffee premium can help market actors stay competitive in the face of global economic slowdown following the Covid-19 pandemic.

Predictions generated by the model described in this paper can help potential buyers enter the market by providing an insight into the expected valuation of coffees. Additionally, non-linear supervised learning algorithms can provide insight into the underlying non-linear relationships governing pricing, giving landowners information on maximising sale prices.

The paper follows a three-step methodology to estimate a predictive model for specialty coffee prices within the Cup of Excellence competitions (Alliance For Coffee Excellence, 2020a), comparing (1) supervised learning algorithms and (2) dataset definitions, (3) optimising select models to achieve peak performance.

To ensure clarity, the following terminology applies throughout the paper:

- "Machine learning algorithm" – An algorithm from the supervised learning toolkit of machine learning used to create a predictive model.
- "Baseline dataset" – The dataset defined in Section 3.4 to serve as a comparison.
- "Dataset permutation" – A definition of the dataset via the choice of binary/cumulative encoding and a level of granularity for the tasting note variables.
- "Model" – An object produced by training an algorithm on a given dataset permutation. It can be used to predict prices by supplying covariates in the appropriate format.

Web scraping, data cleaning and the interpretation of tasting notes were carried out in R (R Core Team, 2020), while the data exploration and modelling were done in Python (Van Rossum and Drake, 2009). Specific packages and references can be found in Appendix C.

2. Literature Review

The literature has seen authors apply the hedonic pricing model to understand the pricing of specialty coffee (Donnet and Weatherspoon, 2006; Donnet et al., 2008; Teuber and Herrmann, 2012; Wilson and Wilson, 2014; Traore et al., 2018). The Cup of Excellence dataset has been used extensively to determine the importance of different factors during online auctions. The literature distinguishes three groups of attributes when determining auction prices: Sensory/material attributes refer to the specific properties of the coffee observed during the tasting process; Reputation/Symbolic factors include growing country, competition ranking, coffee variety, processing method, (Organic/Rainforest Alliance) certifications, and farm altitude; Additional control variables include dummies for different years, the auction winner's location, and the International Coffee Organisation composite price.

Regarding the sensory attributes, earlier papers (Donnet and Weatherspoon, 2006; Donnet et al., 2008) assume that all aspects of taste are captured in the quality score of the coffee and only use the score variable in their analyses using linear specifications. Wilson and Wilson (2014) move to a quadratic functional form, finding that controlling for the rank achieved in a competition the marginal quality score premium decreases for higher scores. However, this effect only appears using the truncated regression method instead of OLS, also introduced by Wilson and Wilson (2014). The actual impact of quality score is likely non-linear and dependent on the rank (Wilson, 2014).

Traore et al. (2018) relax the assumption that quality score appropriately represents all sensory attributes. Their paper evaluates a model replacing quality score with a set of dummies for different taste profiles. However, given the substantial heterogeneity in taste profiles, this reduction of the tasting notes to 20 dummies could cause valuable information to be lost. Additionally, the number of descriptors recorded for a coffee was also added as a predictor with significant effect.

Regarding symbolic attributes, the general agreement is that they form a significant part of the pricing process. Specifically, the rank achieved is found to be crucial in every study. Recent papers (Traore et al., 2018; Wilson and Wilson, 2014) add dummy variables for the buyers' location, arguing that different buyers may be looking for different types of coffees, explaining the heterogeneity present in the market. Although this may be true, the buyer is hardly interpretable as an attribute of the product. Additionally, this is not determined before the auction, so it cannot be included in a pre-auction predictive model.

A strong parallel can be drawn between the specialty coffee and the wine market. A new theme in the wine literature has been the use of machine learning in the creation of models to predict

scores/prices of wines based on review texts (Ramirez, 2010; Chen et al., 2014; Flanagan et al., 2015; Hendricks et al., 2016; Flanagan and Hirokawa, 2016; Chen et al., 2018). Chen et al. (2014) propose that tasting notes need to be categorised into common taste profiles and translated into a one-hot-encoded data matrix. Consequently, the papers use different, usually black-box, models to form their predictions.

3. Data

The Cup of Excellence (CoE) (Alliance For Coffee Excellence, 2020a) program is a series of competitions and auctions of specialty coffees. Samples are submitted to an international jury for tasting, in which judges assign scores to different aspects of the coffees (e.g., aroma, flavour, and acidity), adding up to a final quality score. Notes including a wide variety of flavour and non-flavour descriptors are also recorded. Coffees are ranked according to the given quality score and auctioned off to international buyers.

3.1 Data sources

The dataset used in this paper was aggregated from two different sources:

1. The Cup of Excellence website (Alliance For Coffee Excellence, 2020a) contains extensive data, especially regarding the tasting notes, on separate web pages for 3812 coffees. The data was scraped and structured into a database format.
2. The dataset provided by the Alliance for Coffee Excellence (Alliance For Coffee Excellence, 2020b) contains 4152 observations with complete data on auction prices, but roughly 50% of the data on other covariates is missing.

These two datasets were joined and filtered to observations that contain data on tasting notes. The timeframe was restricted to 2008 onward, as few observations were complete prior to that year. The resulting data set contains 2889 observations. To avoid sample selection issues, correlation coefficients (Fig. 1) were calculated for critical continuous variables and the selection variable based on missing tasting notes.

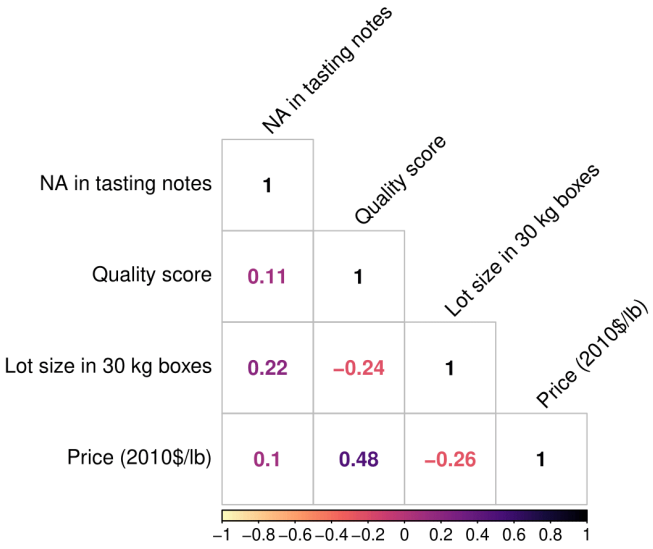

Figure 1: Correlations of missing values and key continuous variables

The data contains several variables on market characteristics, symbolic attributes, and a set of tasting notes to interpreted in Section 3.3. A table of descriptive statistics can be found in Appendix A.

3.2 Outcome variable: Price

The primary variable of interest is the final auction price attained. Coffee prices have undergone significant changes over the 2008-2020 period, so they have been scaled via the US Producer Price Index (PPI) for Coffee and Tea (Fig. 2A). Figure 2B shows that recent years have seen a substantial increase in specialty coffee prices, pointing to a divergent trend from the rest of the coffee and tea industry.

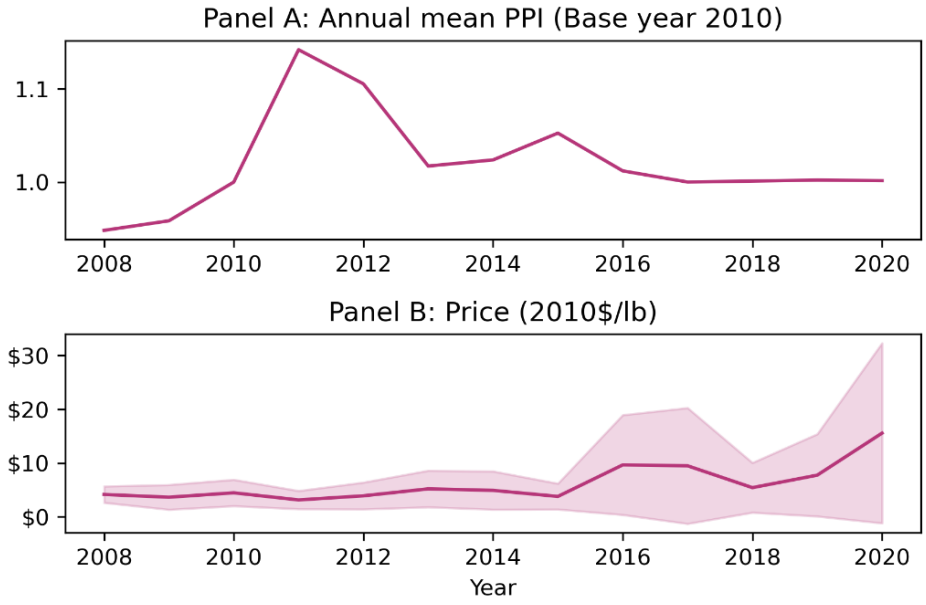

Figure 2: Panel A: Coffee & Tea Producer Price Index (PPI) for the sample period,

Panel B: PPI adjusted annual mean Cup of Excellence prices and standard deviations

Examining the overall distribution of prices (Fig. 3), it is apparent that the distribution is strongly left-skewed with 95% of coffees valued at less than $16. It is also clear that several outliers are present with very high prices. These are included in the sample as exceptionally high premia can contain valuable information on what combinations of traits yield maximum prices. They also set a challenge for testing the predictive capability of the chosen model. To ensure that outliers do not appear due to recording errors, manual verification confirms that only coffees ranked first or second attained prices above $30.

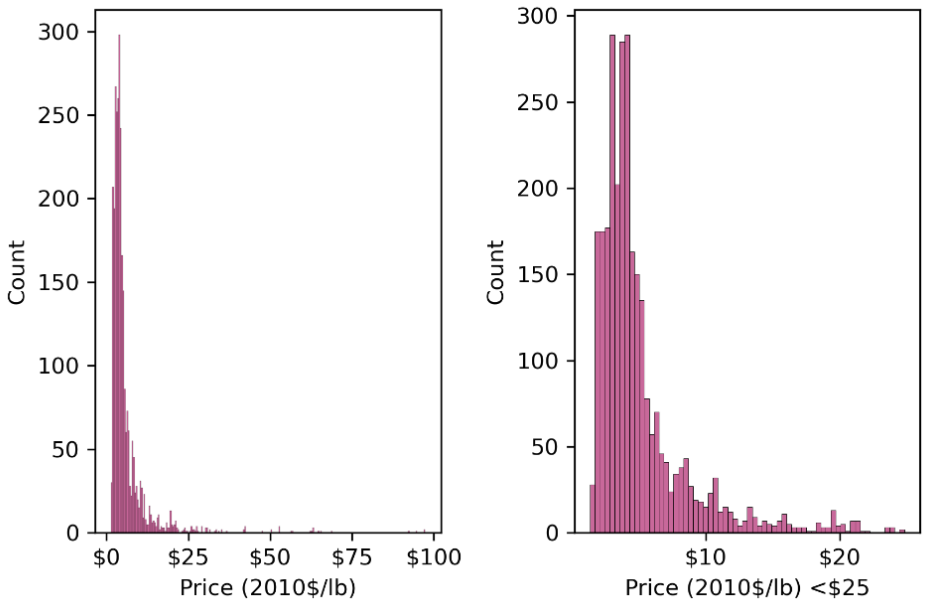

Figure 3: Distribution of prices, left: Whole domain, right: Lower domain.

3.3 Transforming tasting notes into predictors

Traore et al. (2018) introduced tasting notes as predictors for prices and quality scores. Their basis for identifying categories has been the "SCAA Coffee Taster's Flavor Wheel " (Appendix B) (Spencer et al., 2016), henceforth called the flavour wheel. Additionally, several extra, non-flavour descriptors (e.g., "Big body", "Creamy mouthfeel", etc.) were selected.

This paper structures the use of tasting notes by disassembling the flavour wheel to levels of granularity. The first level is the inner-most circle and includes general flavour groups (e.g., fruity, sweet, etc.). This is also the level of granularity used by Traore et al. (2018). The middle ring contains more specific categories but generally not independent tastes. The third level contains the set of individual flavours without any grouping.

Given that low-granularity variables are constructed based on the more granular ones, including all three levels as predictors simultaneously is not feasible as issues of multicollinearity would arise. To explore which granularity is optimal for predictions, different definitions of the dataset need to be considered. The exact methodology is explored in Section 4.2.

Tasting notes are recorded manually and divided by commas into 1-3 words long descriptors. To retain interpretability, lexicons are used to slot the descriptors into categories defined by levels of the coffee wheel. The lexicons and the matching process aim to account for synonyms, minor typos and compound words. However, with the sheer number of tasting notes, errors are bound to occur.

As descriptors are recorded sequentially from different tastings, some descriptors appear repeatedly for a given coffee, which brings up variable encoding. Similar to the level of granularity, a cumulative and a binary encoding of tasting note variables are compared in the modelling process.

Despite the availability of the flavour wheel, tasting notes are not limited to its categories. Additional descriptors include non-flavour attributes, such as the ones used by Traore et al. (2018) and flavour attributes outside of the flavour wheel. The only categories external to the coffee wheel included in this analysis are the ones used by previous research. Additionally, the number of descriptors listed for a coffee, including unused ones, is also recorded as a predictor.

3.4 Baseline dataset

The baseline definition used for comparing different versions of the dataset recreates the same variables used by Traore et al. (2018), except for the nationality of the final buyer and the competition year. To make continuous variables more suitable for estimation, they are scaled to z-scores.

3.5 Data Caveats

Inspecting the descriptive statistics (Appendix A), a few observations can be made that have consequences for the final model's predictive capability.

Given the nature of the competition and the dataset, certain countries are favoured less for holding competitions throughout the sample period. For example, Ethiopia and Bolivia only have one competition featured in the sample each, suggesting that these countries' future predictions are less reliable.

Certain tasting note variables may not have any variance at all. For example, the "Fruity" level-1 feature is present for 98% of the coffees. Predictions for coffees with these descriptors could end up being overly sensitive to them.

Lastly, the variable for farm altitude contained a significant proportion of missing values, which were replaced by the mean, thereby reducing the variable's variance and predictive capability.

4. Methodology

The model selection process follows the following three-step methodology:

1. Compare multiple machine learning algorithms based on their predictive performance when trained on the baseline dataset and choose a handful of algorithms to continue.
2. Train every chosen algorithm on every dataset permutation and select a subset of models to continue with.
3. Optimise selected models by choosing appropriate hyperparameters and choose the final model based on predictive performance.

In an ideal setting, every restrictive step would be avoided, and every algorithm would be optimised for every dataset permutation to find the global minimum within the search space. However, due to a lack of computational power, this is not feasible.

4.1 Evaluation criteria

Measuring out-of-sample prediction power is done by training algorithms on one subsample of the data and evaluating them by generating predictions on another, the test sample. Based on the predictions, the errors can be aggregated into continuous evaluation metrics. As the test sample is exogenous to the model estimation, these metrics measure how well the model can generalise to unseen data. This paper chooses the root mean squared error (RMSE) metric as the objective criterion, as it penalises large errors more than minor errors. This way, models that perform well at predicting outliers are favoured. The root transformation reverts the unit of the metric to USD. Throughout parts of the process, the mean absolute error (MAE) is also calculated as it carries an intuitive interpretation.

To make efficient use of the data, this paper uses k-fold cross-validation (Refaeilzadeh et al., 2009) to split the dataset with k=5. A test sample (20%) is set aside for unbiased final evaluation. The rest of the sample is divided into five smaller subsamples. The algorithms are trained on four subsamples at a time and evaluated on the 5^{th}. This is done five times to produce five different RMSE values. These are then averaged to attain the cross-validated RMSE for the model.

4.2 Baseline model comparison

The supervised learning toolkit contains various linear and non-linear algorithms that can be used to approach regression tasks. Each of them has different strengths and weaknesses, so choosing between them is non-trivial for any use case.

As the first step of the estimation strategy, a subset of the most popular machine learning models is compared with each other by cross-validation using the baseline dataset definition. Overall, the comparison features nine algorithms categorised into two main groups, linear and tree-based. Besides the decision tree, the latter includes gradient-boosting and bagging ensemble methods. The Scikit-learn (Pedregosa et al., 2011) implementations were used for every algorithm except for XGBoost (Chen and Guestrin, 2016).

Linear methods

The linear algorithms included are OLS, Lasso, Ridge regression, and Elastic-net regression (Hastie et al., 2009, pp.43–99). Although OLS is the best in-sample linear predictor, it is prone to overfitting. The other methods feature regularisation, meaning that they penalise non-zero coefficients. As these algorithms are outperformed by non-linear methods in terms of predictive power, they are not discussed individually.

Decision tree

The simplest tree-based algorithm is the decision tree (Hastie et al., 2009, pp.295–336). A tree consists of decision nodes and leaves. Every decision node operates on a simple yes-or-no basis of a single condition on a single covariate (e.g. Is quality score above 90?). At every decision node, optimal thresholds are calculated for every covariate choosing the one optimising the objective function, which in this case is the within-sample MSE. At the end of every decision path, a leaf is reached, which contains the predicted value. A covariate can be used in decision nodes repeatedly, allowing for various non-linearities. The larger the tree, the more nodes there are, and the better the model fits the training sample. This also means that a tree can be "grown" large enough to perfectly fit the training sample, reducing its generalisation ability. Hence, to prevent overfitting, hyperparameters can be chosen to limit the tree's size, generally either by limiting the maximum depth or the minimum number of training observations associated with a leaf.

Ensemble methods

Tree-based ensemble methods extend the idea of decision trees by using a multitude of them. The two classes of ensemble methods compared here are the bagging and the gradient-boosting approaches. While a concise decision tree can be interpreted visually, ensemble methods can be classified as black-box models, meaning they cannot be interpreted by examining the model itself. Therefore, the models' behaviour can only be approximately understood by using permutation importances (Breiman, 2001), accumulated local effects (Molnar, 2020, pp.161–183) or other methods (Molnar, 2020, pp.143–236).

Bagging algorithms

Bagging algorithms, such as the random forest (Breiman, 2001) and the Extratrees (Geurts et al., 2006) methods, grow many trees simultaneously and take their average predictions. Different trees are grown by running the tree algorithm on random subsamples of the data. They are fundamentally similar, but their behaviour diverges when it comes to choosing decision conditions. Random forests choose the optimal threshold for a given covariate at a decision node, while the Extratrees algorithm chooses them randomly. In other words, while random forests use decision trees, Extratrees uses “extremely randomised trees”. Both algorithms are expected to perform better than a singular decision tree.

Gradient-boosting algorithms

Gradient-boosting regressions (Hastie et al., 2009, pp.337–387) do not grow trees simultaneously but sequentially. They apply the fundamental idea that learning from mistakes helps form better decisions in the future. Starting with a simple tree, the error can be iteratively minimised by fitting new trees on the gradient calculated from the criterion (in-sample MSE). Both the gradient-boosting regression (shortened GBReg) and XGBoost belong in this category. XGBoost includes a regularisation term in its objective function, making it more robust to overfitting.

A subset of these algorithms is selected based on cross-validation performance to use in the next step.

4.3 Comparing dataset permutations

As mentioned in Section 3.3, choosing the level of granularity for the tasting note variables is essential in optimising predictions. Therefore, every combination of dataset permutations and algorithms (selected in the previous step) is evaluated via cross-validation to provide a clear picture of performance. This is done separately for both binary and cumulative encodings of the variables to compare these methods.

Dataset permutations are defined by the level of granularity and the inclusion of non-flavour extra descriptors. Additionally, principal components (Hastie et al., 2009, pp.534–552) of these sets of variables are calculated and added in certain dataset permutations. The number of components included is chosen via MLE (Minka, 2000) as implemented in Scikit-learn (Pedregosa et al., 2011). The definitions of dataset permutations can be found in Appendix D.

After evaluating every combination, they are ranked according to their performance, and the top five are chosen for further optimisation. Additionally, four other combinations that scored in the upper half of the evaluation are also included in the next step. This is necessary to

understand how the rest of the models would perform when optimised and assess whether the default hyperparameters could be driving the performance difference in this second stage. The four extra combinations are evaluated with both encoding types to reveal whether the extra information from cumulative encoding makes a difference when hyperparameters are chosen optimally.

4.4 Hyperparameter optimisation for select models

As described in Section 4.2, tree-based algorithms take several hyperparameters as inputs for their training process. These are mainly used to control overfitting and optimise the model for generalisation. Although the regularised linear models discussed above also have hyperparameters, these are optimally chosen in the baseline model comparison as they are inexpensive to find. There are multiple ways to find ideal hyperparameters, manually or automatically. Many modern applications use a search algorithm to choose from the multi-dimensional parameter space.

A full grid search would require the entire model to be evaluated for every combination of parameters, which is computationally expensive – especially when using cross-validation as the model would need to be fit five times for every combination of parameters.

As an alternative to a full grid search, this paper uses the Scikit-optimize (Head et al., 2020) implementation of sequential hyperparameter optimisation using the Extratrees algorithm as a surrogate. This method avoids having to estimate the full, expensive model for every combination and allows continuous variables without restrictions. The optimisation algorithm starts by choosing ten sets of random starting hyperparameters, evaluating each via cross-validation. It then estimates a small, inexpensive surrogate model on the ten observations of the objective function, using the hyperparameters' values as the covariates. Based on this, the parameters that minimise the surrogate predictions are used to evaluate the full model. This is then added to the previous results, the surrogate is re-estimated, and the process repeated for N iterations.

The value of the parameter N is chosen for the specific algorithm optimised, depending on how susceptible it is to overfitting on the cross-validation sample via hyperparameters. This is done by manually inspecting how the chosen parameters' test performance changes when increasing the number of iterations. The optimal stopping point may vary based on the dataset, so the baseline definition is used as it only aims to serve as a comparison. This way, the optimisation procedure is not specifically suited to find the best test score for any given dataset permutation.

When the optimised models are attained, they are sorted by cross-validated RMSEs, and the model with the minimum score is chosen as the final model. To evaluate generalisation performance, a comparison featuring the baseline models and the final model is conducted, where the algorithm is trained on the whole cross-validation sample and evaluated based on predictions on the test set.

5. Results

5.1 Baseline model comparison

The first of the three steps outlined in Section 4 is computationally inexpensive and yields a convincing comparison (Fig. 4) of performance across the algorithms. It is important to reiterate that Cross-validation scores measure out-of-sample prediction performance, so the scores below are a good proxy for the performance expected on the test set.

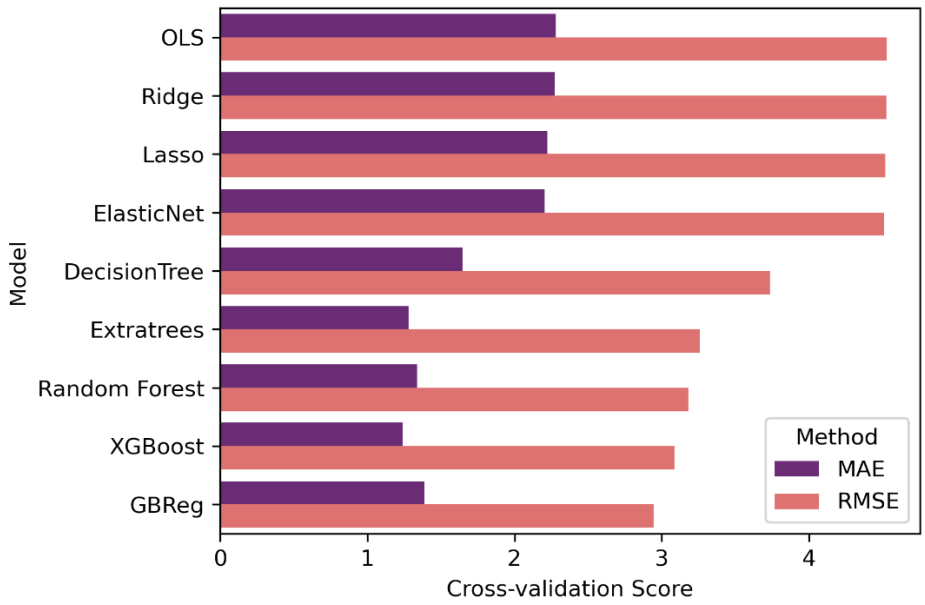

Figure 4: Baseline model comparison

Based on Figure 4, It is apparent that the tree-based methods outperform the linear models. The latter seem to perform similarly to each other, indicating that the variables included constitute a well-specified OLS model that does not overfit. Ensemble methods deliver better results than a single tree, with gradient-boosting having an extra edge above bagging models.

Based on this analysis, the forthcoming only features the Extratrees, Random Forest, XGBoost, and GBReg algorithms.

5.2 Comparing dataset permutations

In the second step, the four algorithms are evaluated on every dataset permutation using both binary and cumulative encodings.

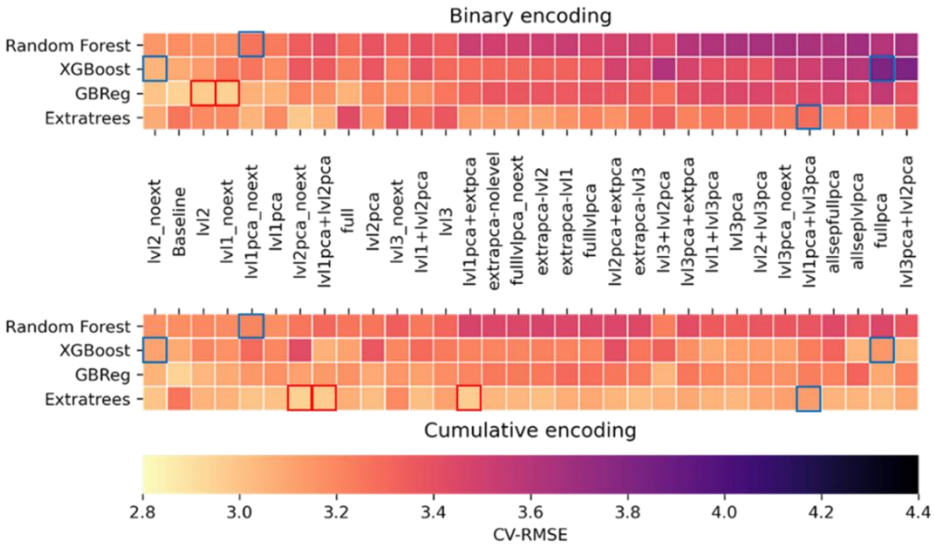

Figure 5: Comparison of dataset permutation across algorithms and encodings,

Red squares: Top five models, Blue squares: Extra combinations (see Section 4.3)

Looking at Figure 5, the baseline dataset performs well, suggesting that extra granularity might not increase performance. Binary encoding seems to yield lower performance on average, however two of the best models also use this encoding. Performance with cumulative encoding seems to be mostly uniform. In contrast, binary encoding works better with sparser dataset permutations.

When comparing the algorithms, the random forest falls behind the other methods. In contrast, the related Extratrees algorithm seems to deliver the most consistent performance, especially with cumulative encoding and when using principal components. The gradient boosting algorithms seem to be performing equally well, but GBReg seems to have a slight edge with default parameters.

The models marked in Figure 5 and the baseline models are carried over to the next step.

5.3 Hyperparameter optimisation for select models

The third step applies the optimisation algorithm discussed in Section 4.4 to find the optimal hyperparameters for the models selected above. Before starting the process, the number of optimisation iterations needs to be selected. At every iteration, the chosen hyperparameters are used to evaluate the test-sample RMSE, determining whether the model is being overfitted on the cross-validation subsample or not.

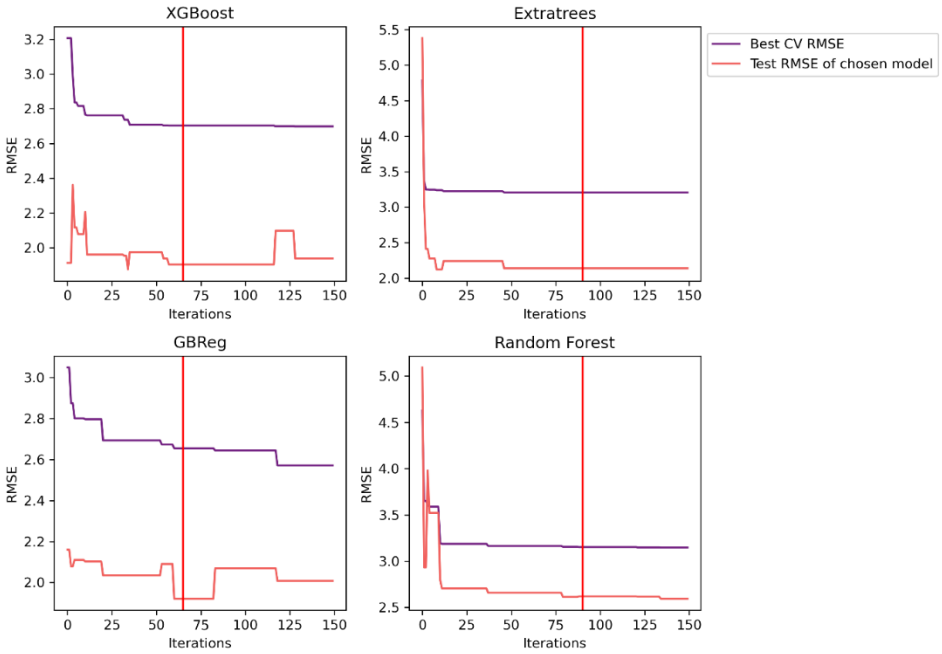

Figure 6: Choosing the number of optimisation iterations by running the optimisation algorithm for 150 periods on the baseline models (see Section 4.4), Red lines: Cut-off points.

Based on Figure 6, the gradient-boosting algorithms achieve their minimum test scores after around 65 iterations, starting to overfit slightly after. In contrast, the bagging algorithms do not appear to be overfitting across the 150 iterations, but early stopping after 90 iterations reduces computing time.

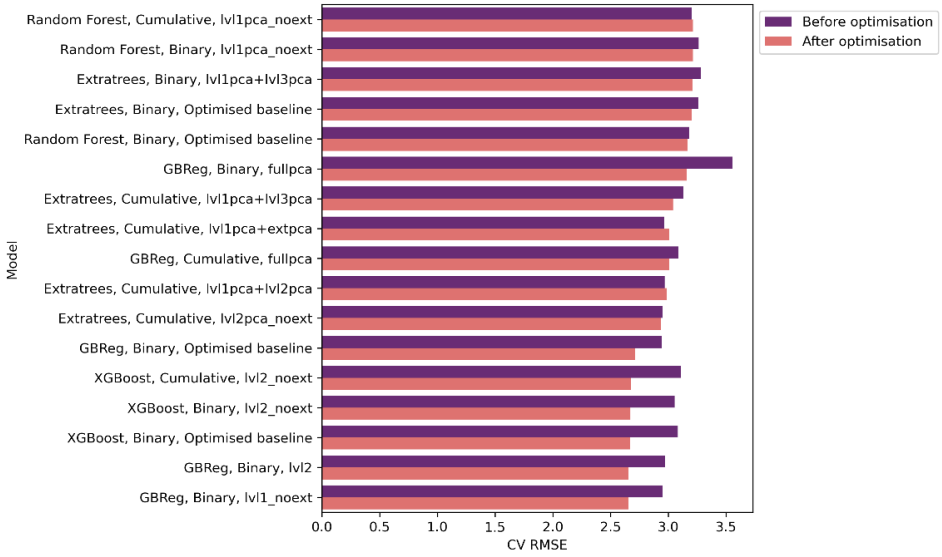

Figure 7: Hyperparameter optimisation results

Figure 7 reveals the cross-validation results attained from optimising the selected models. It shows that hyperparameter optimisation has a more significant effect on the gradient-boosting algorithms. As the bagging methods show minimal or negative improvement, the default

settings may be ideal for this use case. The result shows that bagging methods prefer cumulative, while gradient-boosting algorithms prefer binary encoding in this setting.

Nonetheless, the top spots are taken by gradient-boosting models with negligible score differences between the best five. The best dataset permutations do not include non-flavour descriptors, with level-2 and level-1 granularities performing head-to-head.

XGBoost was only included as an extra combination but optimising its hyperparameters yielded high performance. This could mean either that its default settings are inappropriate for this dataset or that its scope for improvement via optimisation is larger. Either way, this implies that other XGBoost models from the previous Section could be more potent than the ones found here. However, given computational constraints and that the performance differences at the top are minuscule, this is not pursued further.

GBReg on the "lvl1_noext" dataset permutation (see Appendix D for definition) with binary encoding is selected as the final model. This permutation of the dataset uses the same set of tasting note variables as the baseline definition but does not include non-flavour variables.

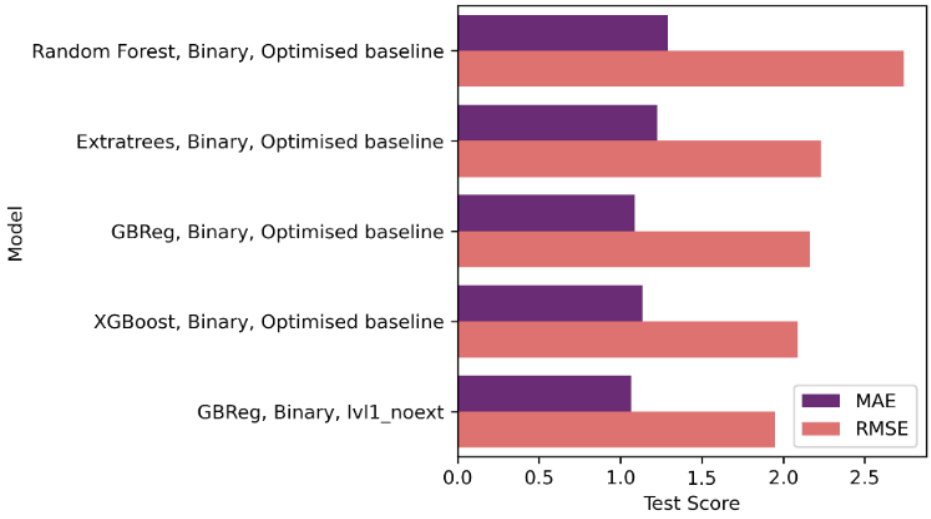

Figure 8: Test scores of baseline models and final model

To compare the model's generalisation performance against the baseline, it is retrained on the whole training sample, and test-sample scores are calculated (Fig. 8). Although by a small margin, the final model outperforms every optimised baseline model in test performance. The distribution of predictions (Fig. 9) seems to be aligned with the actual prices, recreating the left-skewed distribution with the occasional outliers. The following Section further examines the behaviour of this model.

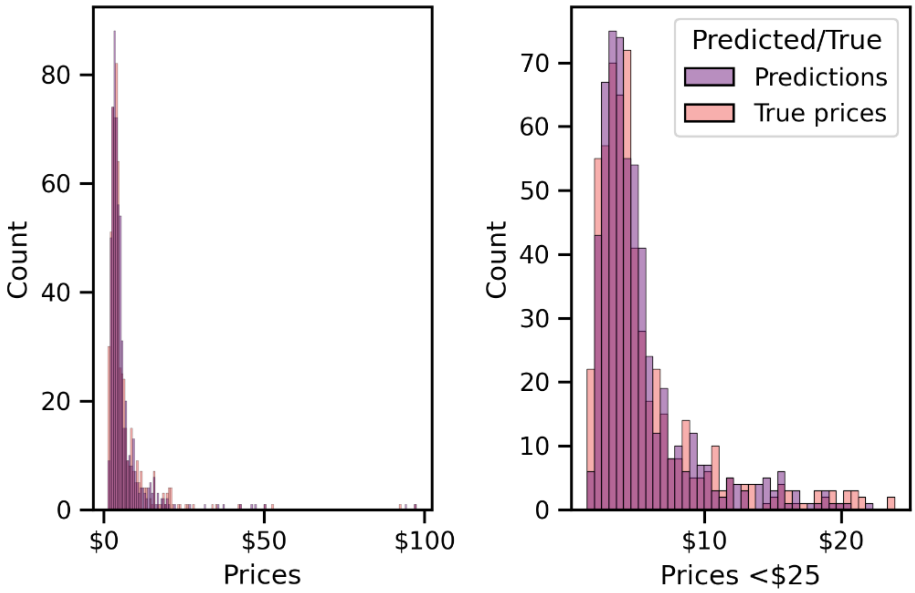

Figure 9: Distribution of Test predictions and prices

6. Discussion

6.1 Feature importance

To understand how vital a feature's impact is on the model's predictions, feature permutation importances (Breiman, 2001) are calculated. The test-sample RMSEs are used to serve as a baseline score, and then, one variable at a time, the values are permuted, and the test RMSEs recalculated. The more significant the deterioration compared to the baseline shows the contribution of a variable to predictions. This process is repeated 15 times to have multiple observations for every variable.

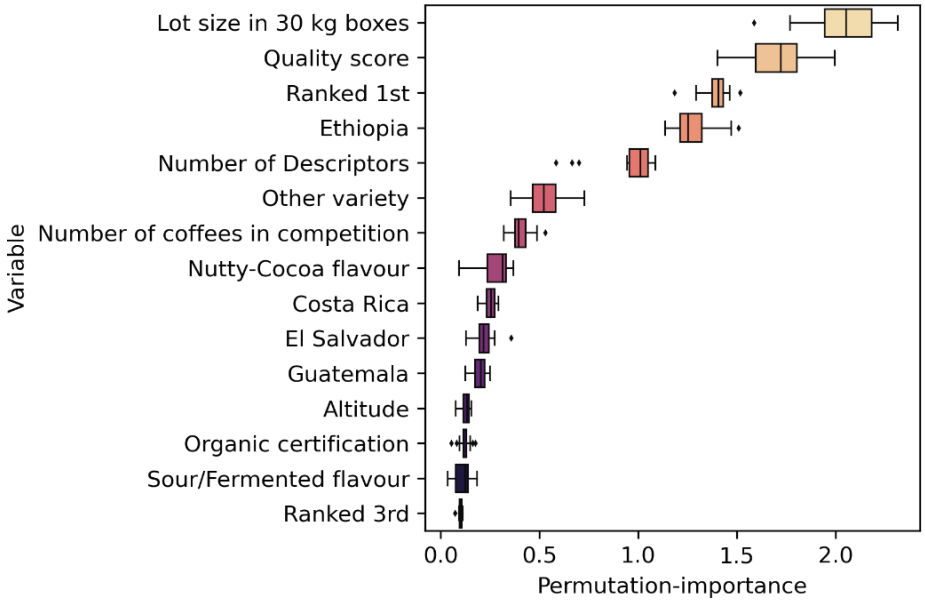

Figure 10: Permutation importance plot of the top 15 features

Inspecting Figure 10, the most significant information sources are the same variables previous research (Wilson and Wilson, 2014; Wilson, 2014; Traore et al., 2018) found to be critical predictors: Lot size, Quality score, ranking 1st, and the number of descriptors. Ethiopia showing up as a top predictor suggests that Ethiopian coffee has either a different base price or different pricing mechanics. However, this is possibly driven by the low number of observations in the sample. Other countries also show up, however, with lower importance. The only two flavour descriptors that show up in the top 15 are the Nutty-Cocoa and the Sour/fermented flavours. This implies that taste descriptors individually do not seem to contribute to the accuracy of predictions significantly.

6.2 Accumulated local effects (ALE) analysis

Although the predictions are formed via complex non-linear relationships, the effect of individual variables on predictions can be approximated by using accumulated local effects (ALE) plots (Molnar, 2020, pp.161–183), implemented in the PyALE package (Jomar, n.d.). In the case of continuous variables, it is defined as the average deviation from the mean prediction conditional on the predictor's value. This is calculated as a per-quantile average for 20 quantiles. For this analysis, the final model is re-estimated for the whole sample. As Molnar et al. (2020) point out, the estimation uncertainty of ALE is understudied, and 95% confidence intervals are included only as visual cues to reflect intervals with low data density. Note, these relationships describe how the model forms predictions; they cannot be interpreted causally in a generalised way.

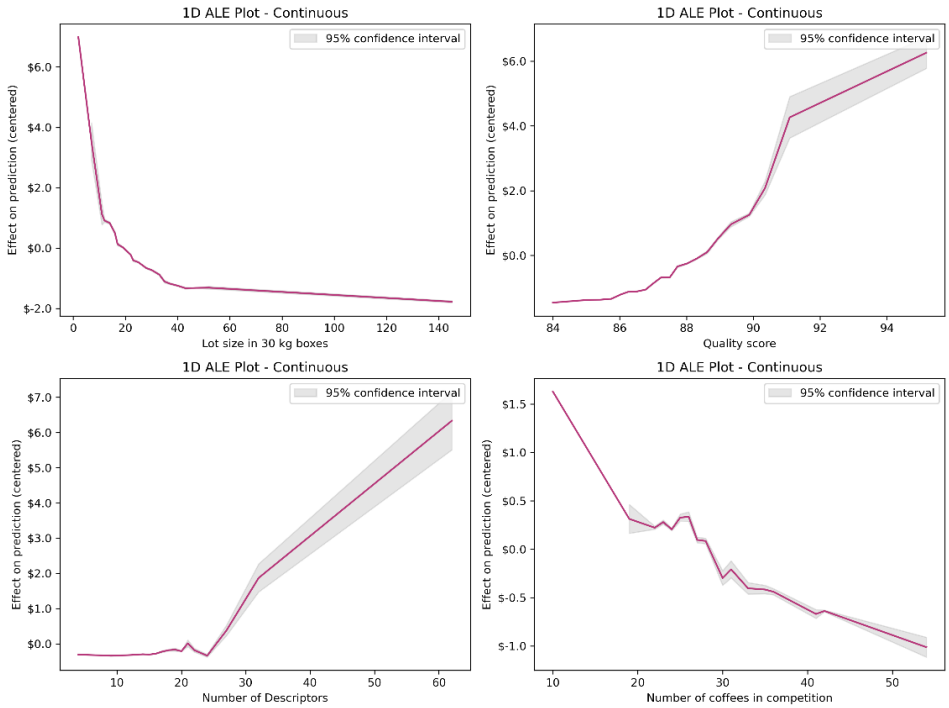

Figure 11: Accumulated local effects plots for continuous variables

The ALE plots (Fig. 11) show that high predictions are driven by scarcity. Predictions increase drastically as lot size declines below 40 boxes. Similarly, competitions with a small number of coffees also seem to have higher priced coffees. The question of functional form for modelling the relationship between quality score and prices has been a question in the literature. Wilson and Wilson (2014) estimate a quadratic model, expecting decreasing marginal returns from the quality score. The ALE plot indicates increasing returns to quality, with a sudden change around 91, after which marginal returns decline. Wilson (2014) finds a cubic relationship fitting better than a quadratic; however, given Figure 11, an inverse cubic relationship could be

explored. A high number of descriptors also has a strong positive effect on predictions, but only a few coffees in the sample have above 30, making these high predictions less reliable.

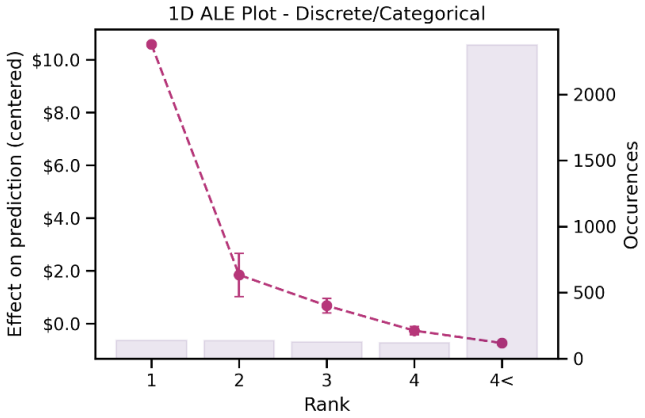

Figure 12: Categorical ALE with respect to rank achieved

The model seems to allocate a significant premium to coffees ranking first (Fig. 12), which coincides with previous research results(Traore et al., 2018; Wilson and Wilson, 2014). While achieving second and third place also yields higher predictions, the differences are less significant.

The relationships driving model predictions all fit into the framework of previous research and economics and, as such, are likely not caused by statistical anomalies or overfitting. Future research could further open the black-box by looking at bivariate interactions.

The machine learning toolkit still has a lot to offer to researchers attempting to understand the pricing of specialty coffee. With access to more computational power, the analysis in this paper could be further extended to fully optimise the model space presented in Sections 4.3 and 5.2. Furthermore, expanding this methodology to other datasets could be used to find more externally valid models. As to the best of the author's knowledge, no generalised specialty coffee datasets exist, and the construction of one with sufficient granularity is a challenge itself. With respect to methodology, the use of word-embedding could be a tool to replace lexicons when interpreting tasting notes. Finally, regarding the modelling aspect, embracing the opportunities provided by deep learning could be used to further improve price predictions to a point where commercial use could become more viable.

7. Conclusion

This paper aims to explore how supervised learning can be applied in predicting specialty coffee auction prices. Algorithms with different strengths and weaknesses were compared to find that a gradient-boosting regression is the most appropriate tool within the comparison scope. The methodology also involved searching for the appropriate encoding and granularity in interpreting tasting notes. This resulted in a finding that predictions do not improve by using more granular data than previous research, the inner-most circle of the SCAA coffee wheel (Spencer et al., 2016). After optimising top-performing candidates' hyperparameters, the final model predicts coffee prices with an RMSE of ~$1.95 and an MAE of ~$1.07.

The high accuracy makes the model viable for predicting auction prices after the ranking of coffees has been published. This enables new buyers entering the market to approximate the prices the coffees are likely to attain. The discussion in Section 6 attempts to untangle the model's non-linearities, uncovering information about the formation of predictions. Although these cannot be interpreted causally outside of the model, they seem to make sense from an economic perspective and can yield insights for landowners attempting to maximise their profits.

8. References

Alliance For Coffee Excellence, 2020a. *Cup of Excellence*. [online] Cup of Excellence. Available at: <https://cupofexcellence.org/> [Accessed 24 Oct. 2020].

Alliance For Coffee Excellence, 2020b. *Cup of Excellence Dataset*.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1), pp.5–32. https://doi.org/10.1023/A:1010933404324.

CBI, 2020. *What is the demand for coffee on the European market?* | *CBI - Centre for the Promotion of Imports from developing countries*. [online] Available at: <https://www.cbi.eu/market-information/coffee/trade-statistics> [Accessed 21 Nov. 2020].

Chen, B., Rhodes, C., Crawford, A. and Hambuchen, L., 2014. Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel. In: *2014 IEEE International Conference on Data Mining Workshop*. 2014 IEEE International Conference on Data Mining Workshop. pp.142–149. https://doi.org/10.1109/ICDMW.2014.149.

Chen, B., Velchev, V., Palmer, J. and Atkison, T., 2018. Wineinformatics: A Quantitative Analysis of Wine Reviewers. *Fermentation*, 4(4), p.82. https://doi.org/10.3390/fermentation4040082.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794. https://doi.org/10.1145/2939672.2939785.

Donnet, M.L. and Weatherspoon, D., 2006. *Effect of Sensory and Reputation Quality Attributes on Specialty Coffee Prices*. [2006 Annual meeting, July 23-26, Long Beach, CA] American Agricultural Economics Association (New Name 2008: Agricultural and Applied Economics Association). Available at: <https://econpapers.repec.org/paper/agsaaea06/21388.htm> [Accessed 5 Nov. 2020].

Donnet, M.L., Weatherspoon, D.D. and Hoehn, J.P., 2008. Price determinants in top-quality e-auctioned specialty coffees. *Agricultural Economics*, 38(3), pp.267–276. https://doi.org/10.1111/j.1574-0862.2008.00298.x.

Flanagan, B. and Hirokawa, S., 2016. Support Vector Mind Map of Wine Speak. In: *Human Interface and the Management of Information: Information, Design and Interaction*. [online] International Conference on Human Interface and the Management of Information. Springer, Cham.pp.127–135. https://doi.org/10.1007/978-3-319-40349-6_13.

Flanagan, B., Wariishi, N., Suzuki, T. and Hirokawa, S., 2015. Predicting and Visualizing Wine Characteristics Through Analysis of Tasting Notes from Viewpoints. In: C. Stephanidis, ed. *HCI International 2015 - Posters' Extended Abstracts*, Communications in Computer and Information Science. Cham: Springer International Publishing.pp.613–619. https://doi.org/10.1007/978-3-319-21380-4_104.

Garnier, S., 2018. *viridis: Default Color Maps from 'matplotlib'*. [online] Available at: <https://CRAN.R-project.org/package=viridis>.

Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning*, 63(1), pp.3–42. https://doi.org/10.1007/s10994-006-6226-1.

Harris, C.R., Millman, K.J., Walt, S.J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M.H. van, Brett, M., Haldane, A., Río, J.F. del, Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T.E., 2020. Array programming with NumPy. *Nature*, 585(7825), pp.357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer Series in Statistics. [online] New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7.

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. and Shcherbatyi, I., 2020. *scikit-optimize/scikit-optimize*. [online] Zenodo. https://doi.org/10.5281/zenodo.4014775.

Hendricks, I., Lefever, E., Croijmans, I., Majid, A. and Van den Bosch, A., 2016. Very quaffable and great fun: Applying NLP to wine reviews. [online] 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.pp.306–312. Available at: <https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2301532> [Accessed 10 Nov. 2020].

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), pp.90–95. https://doi.org/10.1109/MCSE.2007.55.

ICO, 2020. *International Coffee Organization - Historical Data on the Global Coffee Trade*. [online] Available at: <http://www.ico.org/new_historical.asp> [Accessed 21 Nov. 2020].

Izrailev, S., 2014. *tictoc: Functions for timing R scripts, as well as implementations of Stack and List structures.* [online] Available at: <https://CRAN.R-project.org/package=tictoc>.

Joblib Development Team, 2020. *Joblib: running Python functions as pipeline jobs*. [online] Available at: <https://joblib.readthedocs.io/>.

Jomar, D., n.d. *PyALE: ALE plots with python*. [Python] Available at: <https://github.com/DanaJomar/PyALE> [Accessed 2 Apr. 2021].

Khalil, S. and Fakir, M., 2017. RCrawler: An R package for parallel web crawling and scraping. *SoftwareX*, 6, pp.98–106. https://doi.org/10.1016/j.softx.2017.04.004.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. In: S. van der Walt and J. Millman, eds. *Proceedings of the 9th Python in Science Conference*. pp.56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Minka, T.P., 2000. Automatic choice of dimensionality for PCA. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00. Cambridge, MA, USA: MIT Press.pp.577–583.

Molnar, C., 2020. *Interpretable Machine Learning*. [online] Available at: <https://christophm.github.io/interpretable-ml-book/> [Accessed 19 Nov. 2020].

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M. and Bischl, B., 2020. Pitfalls to Avoid when Interpreting Machine Learning Models. *arXiv e-prints*, 2007, p.arXiv:2007.04131.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), pp.2825–2830.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. [online] Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Ramirez, C.D., 2010. Do Tasting Notes Add Value? Evidence from Napa Wines*. *Journal of Wine Economics*, 5(1), pp.143–163. https://doi.org/10.1017/S1931436100001425.

Refaeilzadeh, P., Tang, L. and Liu, H., 2009. Cross-Validation. In: L. LIU and M.T. ÖZSU, eds. *Encyclopedia of Database Systems*. [online] Boston, MA: Springer US.pp.532–538. https://doi.org/10.1007/978-0-387-39940-9_565.

Rinker, T.W., 2015. *qdapTools: Tools to Accompany the qdap Package*. [online] Buffalo, New York: University at Buffalo/SUNY. Available at: <http://github.com/trinker/qdapTools>.

Spencer, M., Sage, E., Velez, M. and Guinard, J.-X., 2016. Using Single Free Sorting and Multivariate Exploratory Methods to Design a New Coffee Taster's Flavor Wheel. *Journal of Food Science*, 81(12), pp.S2997–S3005. https://doi.org/10.1111/1750-3841.13555.

Teuber, R. and Herrmann, R., 2012. Towards a differentiated modeling of origin effects in hedonic analysis: An application to auction prices of specialty coffee. *Food Policy*, 37(6), pp.732–740. https://doi.org/10.1016/j.foodpol.2012.08.001.

The pandas development team, 2020. *pandas-dev/pandas: Pandas*. [online] Zenodo. https://doi.org/10.5281/zenodo.3509134.

Tierney, N., 2017. visdat: Visualising Whole Data Frames. *The Journal of Open Source Software*, 2. https://doi.org/10.21105/joss.00355.

Tierney, N., Cook, D., McBain, M. and Fay, C., 2020. *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. [online] Available at: <https://CRAN.R-project.org/package=naniar>.

Traore, T.M., Wilson, N.L.W. and 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), pp.349–368. https://doi.org/10.1017/aae.2018.5.

Van Rossum, G. and Drake, F.L., 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Waskom, M.L., 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), p.3021. https://doi.org/10.21105/joss.03021.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y. and Zemla, J., 2013. corrplot: Visualization of a correlation matrix. *R package version 0.84*, 230(231), p.11.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H., 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), p.1686. https://doi.org/10.21105/joss.01686.

Wilson, A.P. and Wilson, N.L.W., 2014. The economics of quality in the specialty coffee industry: insights from the Cup of Excellence auction programs. *Agricultural Economics*, 45(S1), pp.91–105. https://doi.org/10.1111/agec.12132.

Wilson, N., 2014. *When Higher Quality Does Not Translate to Higher Prices: A Case of Quality and Specialty Coffees from the Cup of Excellence Auctions*. [online] AgEcon Search. https://doi.org/10.22004/ag.econ.170701.

Wollni, M. and Zeller, M., 2007. Do farmers benefit from participating in specialty markets and cooperatives? The case of coffee marketing in Costa Rica. *Agricultural Economics*, 37(2-3), pp.243–248. https://doi.org/10.1111/j.1574-0862.2007.00270.x.