# Prediction of Coffee Review Ratings Using TF-IDF and Regularization Models

**Qiuli Lai**

**Data Science Initiative, Brown University**

## 1. INTRODUCTION

Coffee is among the most widely consumed beverages globally (67% of U.S. adults drank coffee in a given day) and its specialty sector continues to grow rapidly, with the global market projected at $101.6 billion in 2024 and rising to $183 billion by 2030[1-2]. The sensory profile of a coffee is shaped by stages including botanical variety, altitude, climate, processing, roasting and brewing, these elements defined the attributes aroma, flavor, acidity, body. During cupping events, coffees are brewed and served in a standardized manner and evaluated by certified Q-graders using the Specialty Coffee Taster Wheel (see Appendix A). Final scores are assigned on a 0–100 scale following Specialty Coffee Association (SCA) guidelines, with qualitative descriptors corresponding to numerical bands[3].

While this human-centered evaluation remains an industry standard, it is subjectivity and limited, additionally the process is time-consuming and demands significant training to ensure consistency. Hence, we aim to use the supervised machine learning to predict coffee ratings, aiming to make the evaluation process more consistent and scalable. The dataset we use can be publicly accessed from Kaggle, which is derived from *CoffeeReview*, it covers specialty coffee reviewed between 2017 and 2022[4]. The dataset contains 1,246 entries with fields including coffee name, roaster, country of origin, roast level, review date, review text, retail price per 100g, and rating which is our target variable.

## 2. EXPLORATORY DATA ANALYSIS

By inspecting the whole dataset, we found 12 missing values in the attribute 'roast', and repeated coffee indicating the non-iid structure of the dataset.

Then, from figure 1 we discovered most 85 to 90 point coffees have lower prices, while a cluster of 92.5 to 95 point coffees spans a wider price range. Coffees above 96 points form a premium tier, likely representing rare or exceptional lots. Figure 2 shows that the average price shows a strong upward trajectory from 2017 to 2021, then a sharp decline to 2022, likely driven by factor like the economic pressure, however, the rating after during 2020 and

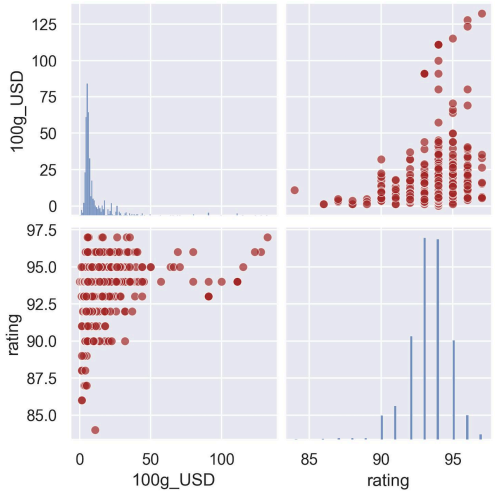
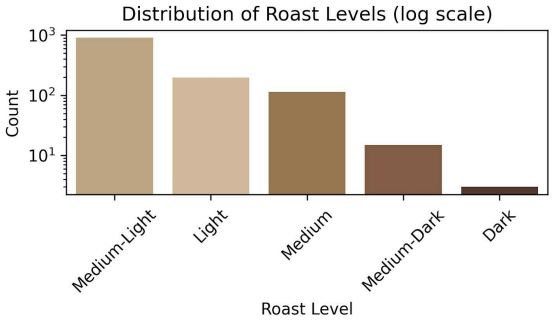



*Figure 1.* Pair Plot of Price and Rating


*Figure 2.* Plot of Average Price and Rating Trends Over time in dual axes. The left Y-axis tracks average price, and the right Y-axis tracks average expert ratings.

2022 is relatively stably increasing.

Additionally, the distribution in figure 3 shows that most specialty coffee in our dataset are medium-light, light, and medium, and only three are dark roasts. Lighter roasted beans tend to preserve more nuanced flavor, while darker roasting often mutes the original characteristics of the beans, resulting in a more homogeneous taste profile, so



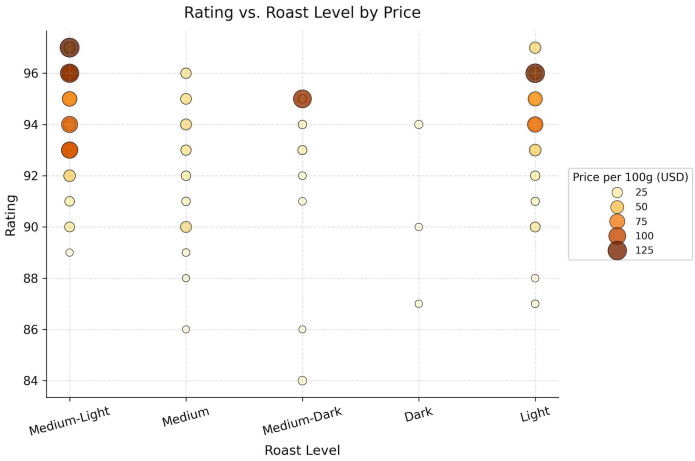
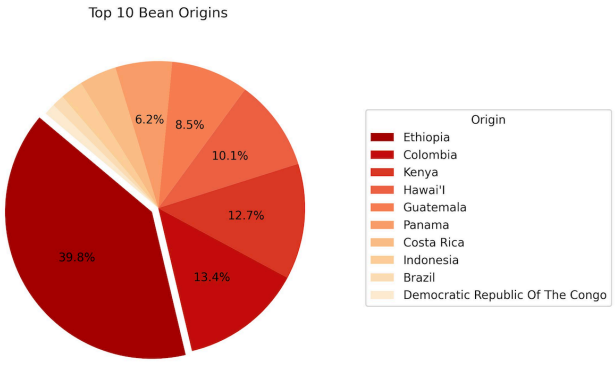




*Figure 3.* Histogram of Distribution of Roast Levels(log scale); *Figure 4.* Rating v.s. Roast Level by Price

rating is complicated. Interestingly, from figure 4 we notice that roasting level does not exhibit a monotonic relationship with the rating. The light roast group has a noticeable left tail, one reason is that some are more likely to be underdeveloped. We also found some evidences:

- Medium-light: rating are relatively high, those above 92 points are over $75/100g(nonlinear), this kind of bean contributes the highest score with most expensive price.
- Medium: wide rating range with consistent price under $50/100g.
- Medium-dark: rating is sparse with large price discrepancy, either under $25 or above $125 per 100g, contribute the main outlier.
- Dark: few and inexpensive comparatively.
- Light: similar to the medium light outcome, but surprisedly the highest rated coffee in this category has the lowest price.

There are 27 origins in the dataset, figure 5 and figure 6 summarize origin prevalence and rating patterns. Ethiopia contributes nearly 40% of the beans, and together with Colombia and Kenya exceeds 60%. High-performing origins such as Ethiopia, Hawaii, Kenya Panama, Yemen and Taiwan show consistently elevated ratings with higher IQRs, while lower-frequency origins like Thailand and Nepal cluster at the lower end of the rating range.



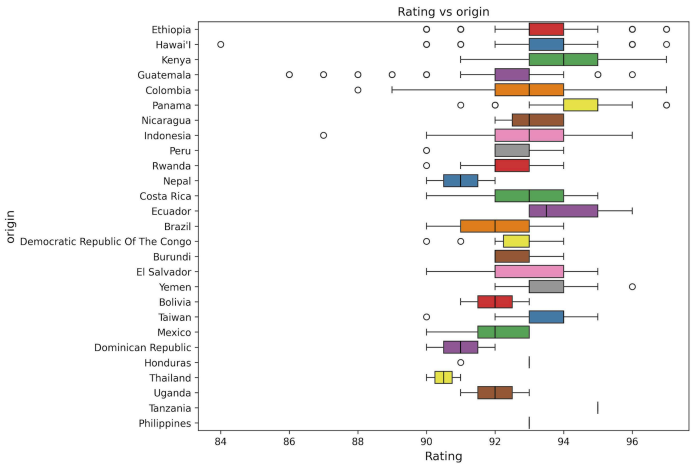




*Figure 5.* Pie Plot of Bean Origins; F*igure 6.* Boxplot of Rating and Origins

## 1. METHODOLOGY

### 3.1 Data Engineering

To refine for our ML algorithm and reduce noise, we did data diagnosis for the unstructured attribute ‘review’, we transfer all the text into lowercase, and keep only letters in english. Then, we use tokenizaiton and lemmtization from the NLTK library to split the text in to tokens and to reduce different word

variations to a single, canonical root[5]. Moreover, we remove the irrelavent stop words (eg.'and','or' etc.).

3.2 Data Splitting

Considering the non-iid structure of this dataset, GroupShuffleSplit (n_splits = 4, train_size = 0.8) is used to generate training sets, with the remaining 20% held out as test sets."

3.3 Preprocessing

The numeric feature 100g_USD is scaled using StandardScaler to normalize price values. Categorical feature like 'origin', 'roast' etc. are encoded via OneHotEncoder. For the textual feature 'review', we use TfidfVectorizer to convey free-form text into a weighted feature matrix, this is a widely used method in text mining to represent the importance of words, allow machine learning to focus on distinctive linguistic signals[6]. Missing values in the roast column are handled using SimpleImputer(strategy = 'constant', fill_value='Unknown') to ensure consistent categorical encoding and preserve missing values as an informative signal.

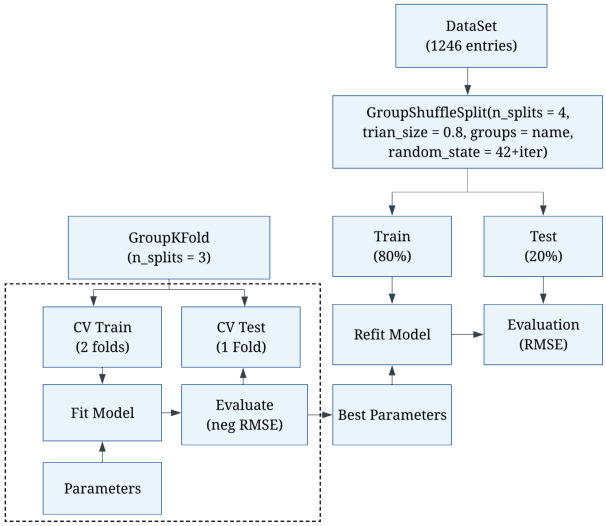



*Figure 7.* Flow Chart of the Model Development in Section 3.1-3.3

3.4 Modeling

We perform hyperparameter tuning using GroupKFold cross-validation on the training set generated by GroupShuffleSplit, ensuring no group leakage across folds. Five regression models with tuned hyper parameters can be found in Table1. Each model is tuned using GridSearchCV with the GroupKFold object as the cross-validator. To address randomness from data splitting and model stochasticity, the entire process from GroupShuffleSplit in Section 3.2 to GroupKFold is repeated with 10 different random states. Moreover, a baseline RMSE is calculated under each random test set. Figure 7 provides an illustration of one random state's pipeline.

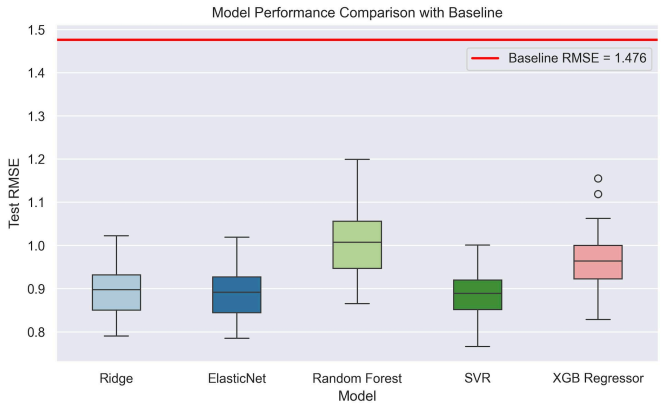
**Table 1. Models and Tuned Hyperparameters**

| Model | Hyperparameters Tuned | Search Space |
|---|---|---|
| **Ridge Regression** | alpha, max_iter | np.logspace(-3, 3, 7); [1000] |
| **Elastic Net** | alpha, l1_ratio, max_iter | np.logspace(-3, 3, 5); [0.1, 0.5, 0.9]; [2000] |
| **Random Forest Regressor** | n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features | [200, 300, 400]; [10, 20, None]; [2, 5, 10]; [1, 3, 5]; ['sqrt', 'log2'] |
| **Support Vector Regression** | gamma, C, epsilon, kernel | np.logspace(-3, 3, 7); [0.1, 1, 10]; [0.1, 0.2, 0.5]; ['linear', 'rbf'] |
| **XGB Regressor** | n_estimators, learning_rate, max_depth, subsample, colsample_bytree, reg_alpha, early_stopping_rounds | [300, 500]; [0.01, 0.1]; [3, 5, 10]; [0.8], [0.9], [0.1, 1]; [50] |

## 4. RESULT

The performances of the models discussed in section 3.3 are revealed in figure 8-9. SVR achieved the lowest average test RMSE values and highest model improvement percentage. Regarding the best SVR model corresponding to the best test set, we then provide below interpretations:

- Permutation feature importance: model agnostic and measures how shuffling each feature affects predictive performance, though it can be computationally expensive and sensitive to multicollinearity.



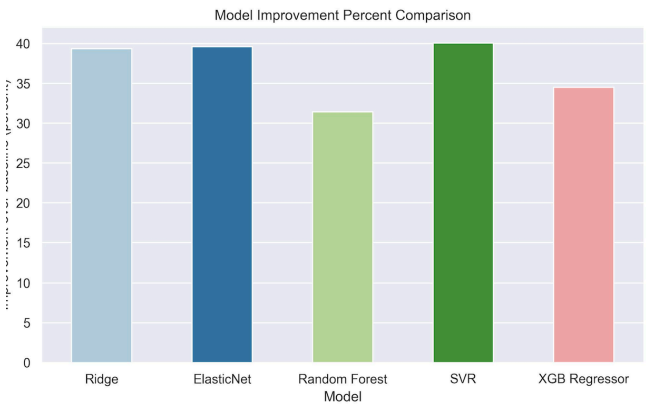




Figure 8. Model Test RMSE compared with baseline RMSE; Figure 9. Model improvement Percent Comparison using (baseline_rmse-mean_test_rmses)/baseline_rmse. SVR performs the best, with lowest test rmse and highest improvement percentage.

- SHAP: finer insight by attributing prediction impact to each feature globally and locally in consistent, interpretable units.
- Partial dependence: shows how the model's average prediction changes as one feature varies while all others are held fixed. It isolates the marginal effect of a feature and reveals whether its relationship with the target.

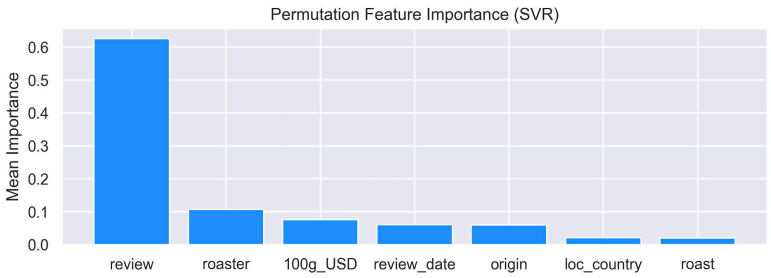



*Figure 10.* Histogram of permutation feature importance using the best SVR model evaluated on the corresponding test set.

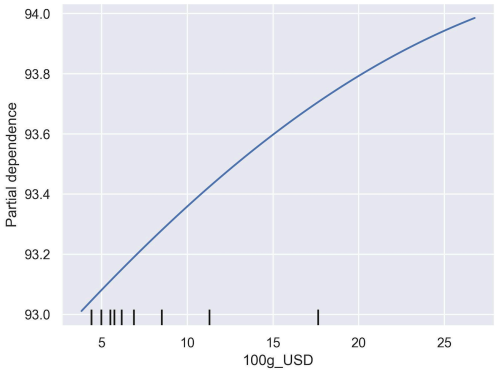



*Figure 11.* Partial dependence of the feature 100g_USD for the best SVR model, evaluated on the corresponding X_test

Result from figure 10 shows that the most dominant predictor is 'review', followed with the second importance 'roast' and then '100g_USD'. By inspecting the partial dependence of '100g_USD' we see a clear positive relationship between price and rating score by coding all other features at their observed values and varying only '100g_USD'.

Interestingly, unlike the result from permutation feature importance in figure 10, the best SVR relies most heavily on num_100g_USD and some categorical feature from the global interpretation of SHAP in figure 12. Feature num_100g_USD is the widest, the higher priced coffee systematically push the predicted rating upwards. Some origin values like Ethiopia, Kenya, Panama have predominately positive SHAP, which is consistent with our EDA in Figure 5. However the textual feature 'structure brisk' push the SHAP value negatively.

For local interpretation, we randomly inspected two instances as shown in Figure 12. For instance 79, the predicted rating is 92.29, which is below the base value. The feature num_100g_USD is the main positive contributor, increasing the predicted score. In contrast, the review date, the roaster, and several textual descriptors such as brisk, magnolia, and finish notes contribute negatively and lower the prediction.

For instance 215, the predicted rating is 95.18, which is significantly above the base value. The textual descriptor magnolia again acts as a strong negative contributor, followed by fruity and the categorical feature origin_Hawai'i.

On the other hand, roaster_Hula Daddy Kona Coffee and num_100g_USD are major positive influences, supported by the review date and the textual feature flavor.

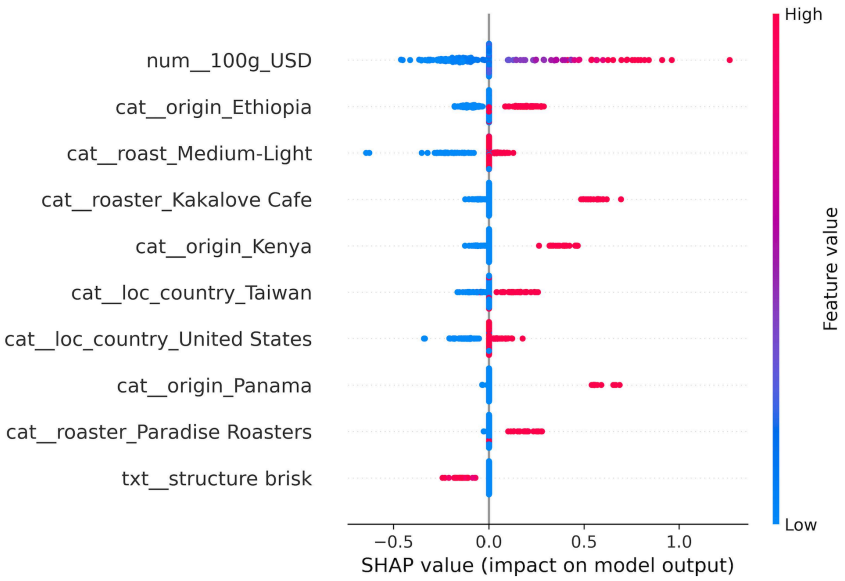



*Figure 12.* Top 10 SHAP values for the global interpretation using the best SVR model evaluated on the corresponding test set.

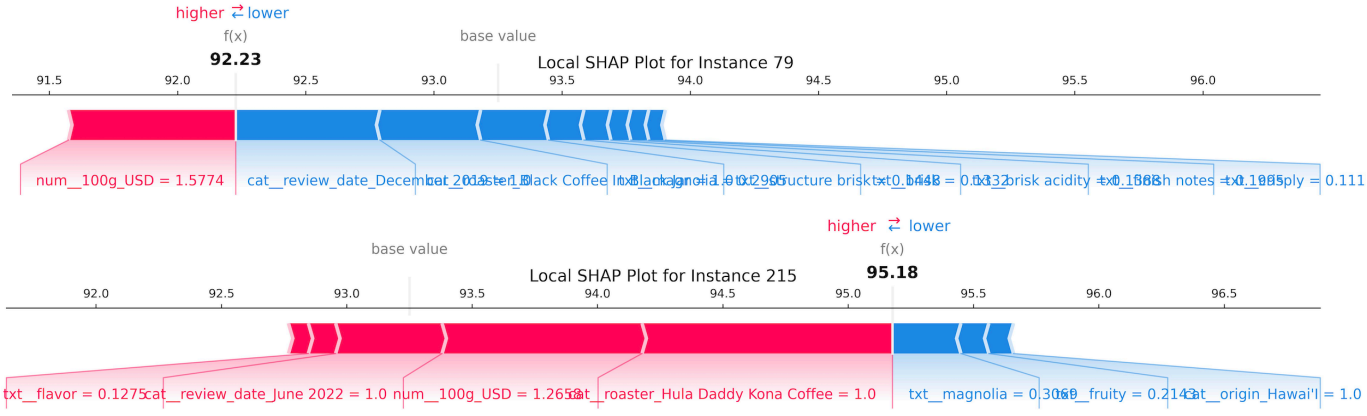



*Figure 13-14. Local SHAP plots for the best SVR model evaluated on the corresponding test instance 79 and 215.*

## 5. OUTCOMES

The SVR model outperforms the baseline and the other tested models, yet several aspects of the project can be strengthened.

- The SHAP analysis highlights textual features such as 'structure brisk' is informative. Since TF-IDF cannot capture word order or contextual nuance, incorporating more advanced NLP representations, such as contextual embeddings, may better capture sentiment and improve model performance.
- Most samples cluster in the 85–90 rating range, and high-priced items dominate. More data from lower rating and lower price segments would help the model learn the full price quality.
- The dataset lacks key coffee quality predictors. Future data collection could include processing methods known to affect taste, botanical varieties tied to quality and price, altitude which is standard quality indicators in coffee cupping. These extensions would improve model predictability, enhance generalization to previously unseen or less processed data, and enable a more robust investigation of advanced machine learning and deep learning applications in the coffee industry.

- One hot encoding of categorical feature like loc_country, roaster and origin leads to high dimensional sparse features. Possible improvements include target encoding, feature hashing or dimensionality reduction, while monitoring interpretability.

## 6. REFERENCES

[1]Specialty Coffee Association and National Coffee Association, “2024National Coffee Data Trends (NCDT)Specialty Coffee Breakout Report,”Jun.122024. [Online]. Available: https://sca.coffee/sca-news/2024-national-coffee-data-trends-specialty-coffee-breakout-report-now-available. [Accessed: Nov.162025].

[2]Grand View Research, Inc., “Specialty Coffee Market Size, Share & TrendsAnalysisReport by Product, by Distribution Channel, by Region, and Segment Forecasts, 2025–2030,” MarketResearch.com Marketplace, last accessed Aug.2025. [Online]. Available: https://www.marketresearch.com/Grand-View-Research-v4060/Specialty-Coffee-Size-Share-Trends-38710766/. [Accessed: Nov.162025].

[3]Specialty Coffee Association, “Coffee Value Assessment (CVA) – A System to Assess Coffee Value,” Oct.12025. [Online]. Available: https://sca.coffee/value-assessment. [Accessed: Nov.162025].

[4] S. Schmoyote, “Coffee Reviews Dataset,” Kaggle, 2021. Available: https://www.kaggle.com/datasets/schmoyote/coffee-reviews-dataset. [Accessed: 19-Nov-2025].

[5]“TF-IDFVectorizer,” scikit-learn.org, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. [Accessed: Nov.162025].

[6]“NLTK — Natural Language Toolkit,” nltk.org, [Online]. Available: https://www.nltk.org/. [Accessed: Nov.162025].

**7. Github Repository:** https://github.com/kyulli/data1030_final_project_coffee_review_prediction_of_rating.git

Appendix A

Flavour Wheel


Speciality Coffee Taster Flavour Wheel

Appendix B

Coffee Rating

Coffee Rating Interpretation

| Score Range | Quality Description |
|---|---|
| 97–100 | Exceptional and rare |
| 93–96 | Distinctive, sweet-toned, structurally flawless |
| 89–92 | High quality with personality or small quirks |
| 85–88 | Solid and drinkable, but conventional |
| <85 | Commercial grade or unbalanced in structure/flavor |