Final Project Report

Wen-shan, Liu

Coffee Review Analysis & Rating Prediction

Table of Contents

1. Introduction

Coffee is one of the most widely consumed beverages globally, with a growing market for specialty coffee. According to the research [1], over 2.25 billion cups of coffee are consumed daily in the world, underscoring coffee's widespread popularity. The market's value reached approximately $96.45 billion for at-home consumption, with out-of-home revenues, such as those from restaurants and cafes, contributing an additional $376.70 billion, bringing the total to around $473.10 billion.

However, assessing coffee quality remains a subjective task, often relying on expert cupping scores that evaluate attributes such as aroma, acidity, aftertaste, body, balance, and flavor. The complexity of these evaluations can introduce inconsistencies, making it challenging for coffee growers, roasters, and distributors to understand what drives high ratings.

This project aims to apply data science techniques to predict overall coffee ratings based on sensory and categorical attributes. By analyzing patterns in the data, we can identify the key factors influencing coffee quality and provide data-driven recommendations for coffee producers and industry stakeholders.

2. Problem Statement

The specialty coffee industry relies on quality ratings to determine pricing, marketing, and production decisions. However, the rating system can be subjective and inconsistent, influenced by human biases or varying standards among coffee graders. Additionally, coffee quality can be impacted by multiple factors, including:

- The origin of the beans (e.g., Ethiopia, Colombia, Brazil)
- Sensory features (e.g., acidity, aroma, body, flavor)
- Roast levels (e.g. Light, Medium-Light, Medium, Medium-Dark, Dark)

The primary objectives of this project are:

1. To develop a machine learning model capable of accurately predicting coffee ratings.
2. To determine the most influential attributes that contribute to high-quality coffee.
3. To explore the impact of coffee origin, sensory attributes, and roast levels on ratings.
4. To provide actionable recommendations to coffee producers on improving quality.

3. Data Sources & Preprocessing

The dataset used in this project consists of coffee reviews, including sensory scores, coffee origin, processing method, and roasting details. The coffee review data was obtained from *Coffee Review* [2], which using data scraping technique to visit each item of webpage and scrape down the data. In order to having a completed analysis, we also add geography data and the weather data of the coffee origin which by using geopy package[3] and OpenWeatherMap API[4].

Data Cleaning & Preprocessing

- Handling Missing Values: Missing values were removed if necessary.
- Encoding Categorical Variables: Features such as country of origin and processing method were encoded using one-hot encoding.
- Feature Scaling: Attributes were normalized using StandardScaler, which operates on the principle of normalization, where it transforms the distribution of each feature to have a mean of zero and a standard deviation of one. It ensures all features were on a similar scale.
- Outlier Detection: Outliers were identified and removed using interquartile range (IQR) methods to improve model stability. In this project, the definition of outliers of price is over 1.5* IQR.

Exploratory Data Analysis (EDA)

Before building predictive models, an EDA was performed to gain insights into the dataset. Key findings include:

- Correlation Analysis: Sensory features such as Aroma, Acidity, and Aftertaste had a strong positive correlation with overall coffee ratings.
- Geographical Influence: Coffee beans from Hawaii tended to receive higher median ratings compared to other regions.
- Coffee price Analysis: Coffee beans from Hawaii are priced higher on average than those from other regions.

Data Visualization

Visualizations, including histograms, scatter plots, and correlation heatmaps, were used to support these findings.

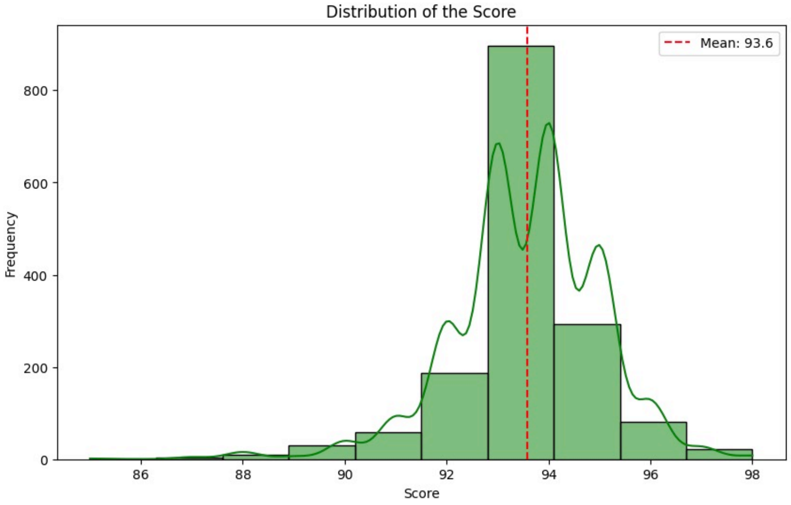


Figure1 : histogram shows the distribution of the rating scores

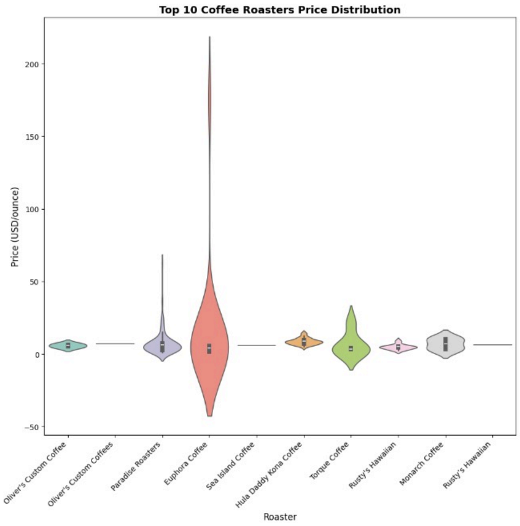


Figure2: violin plot shows top10 coffee roaster price distribution.

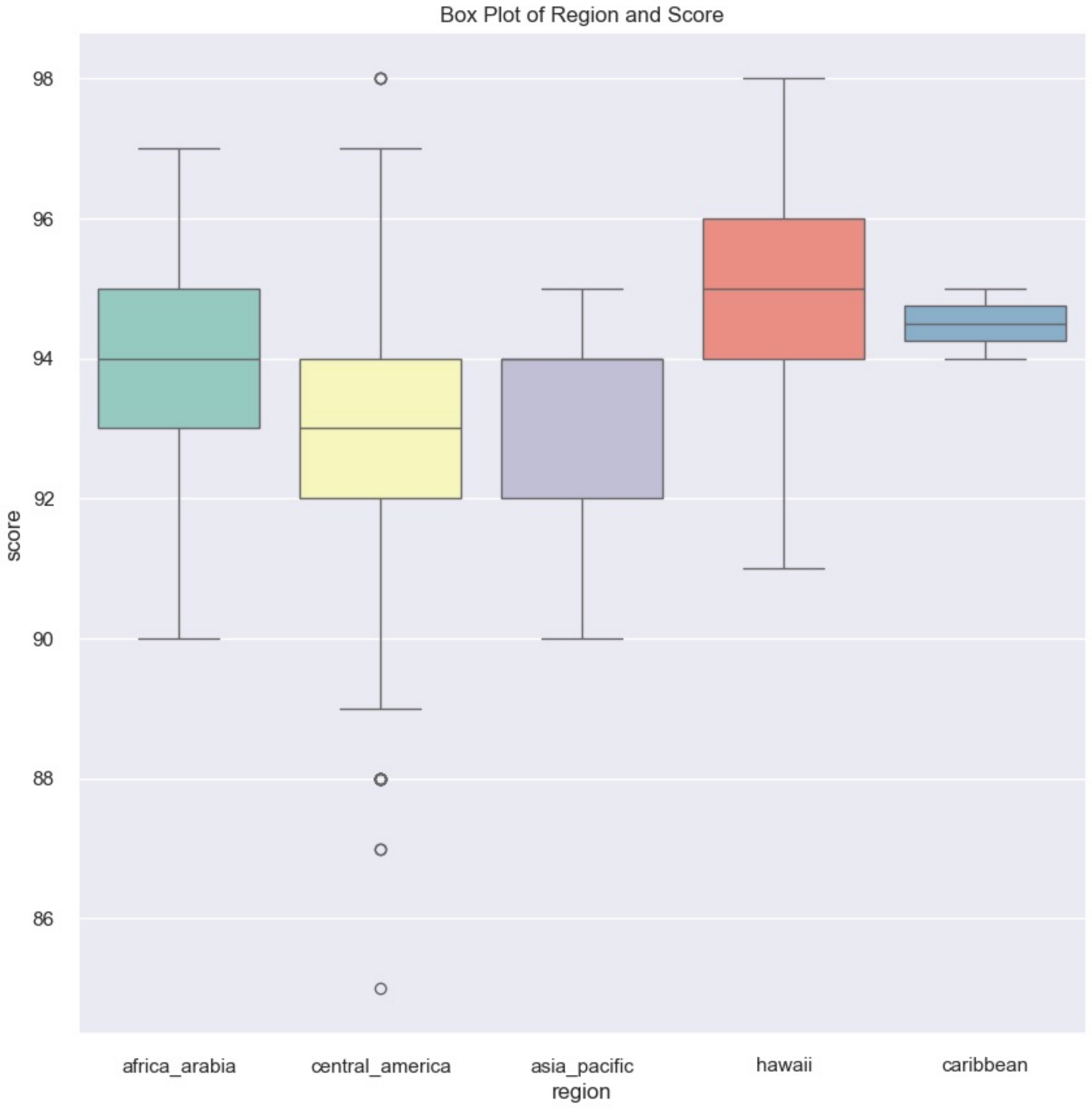


Figure3: box plot shows the distribution of coffee regions and their rating score

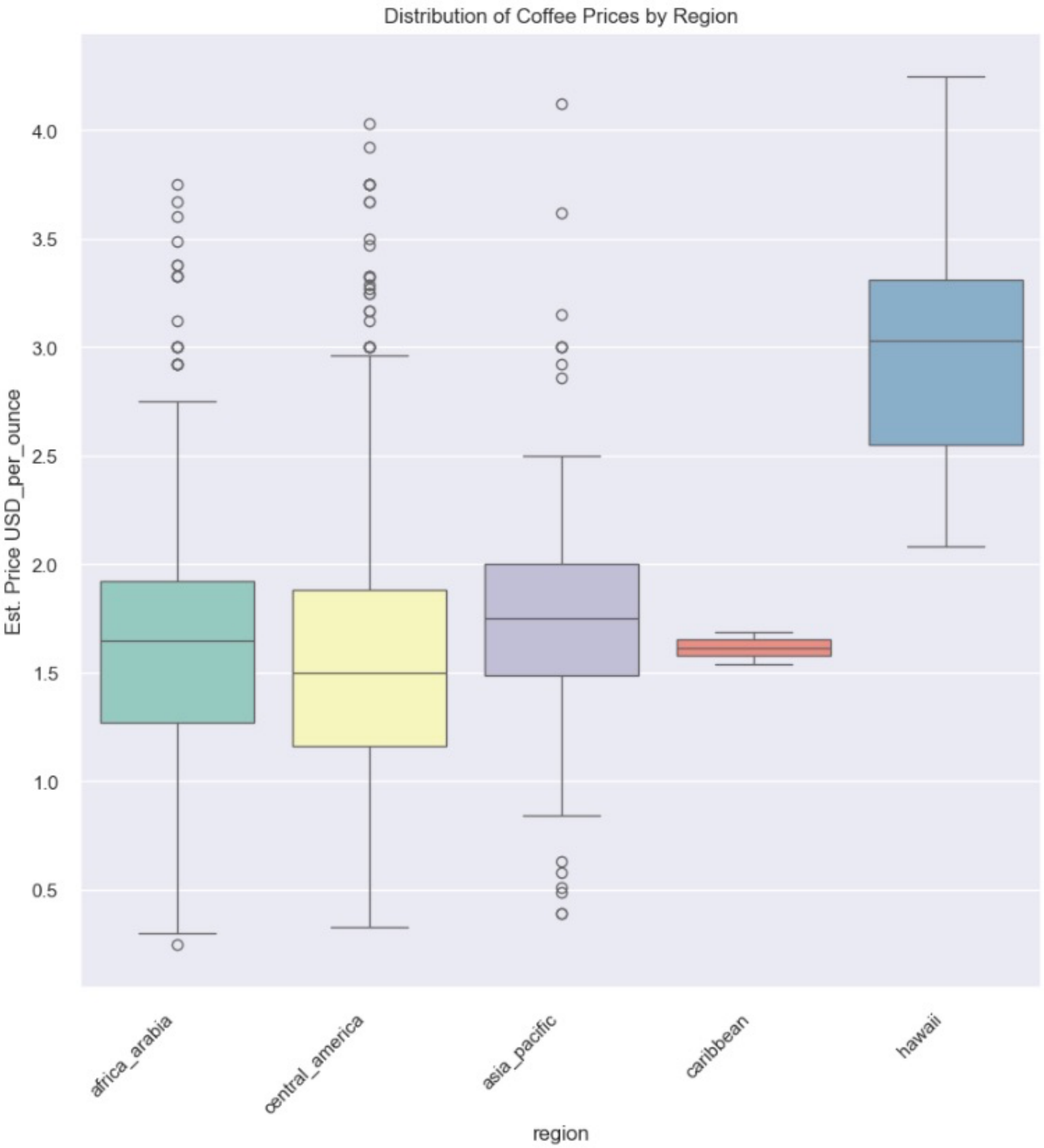


Figure4: box plot shows the distribution of coffee regions and their coffee price

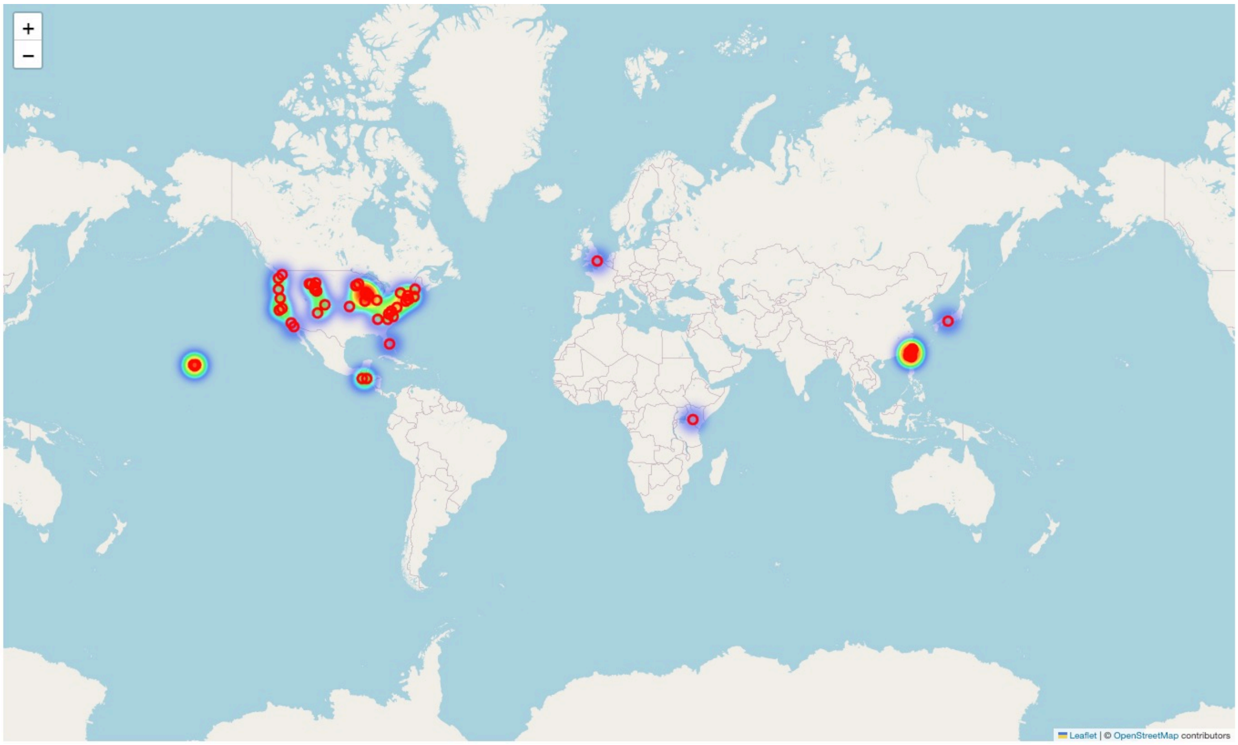


Figure5: heatmap in the map of the coffee roast location

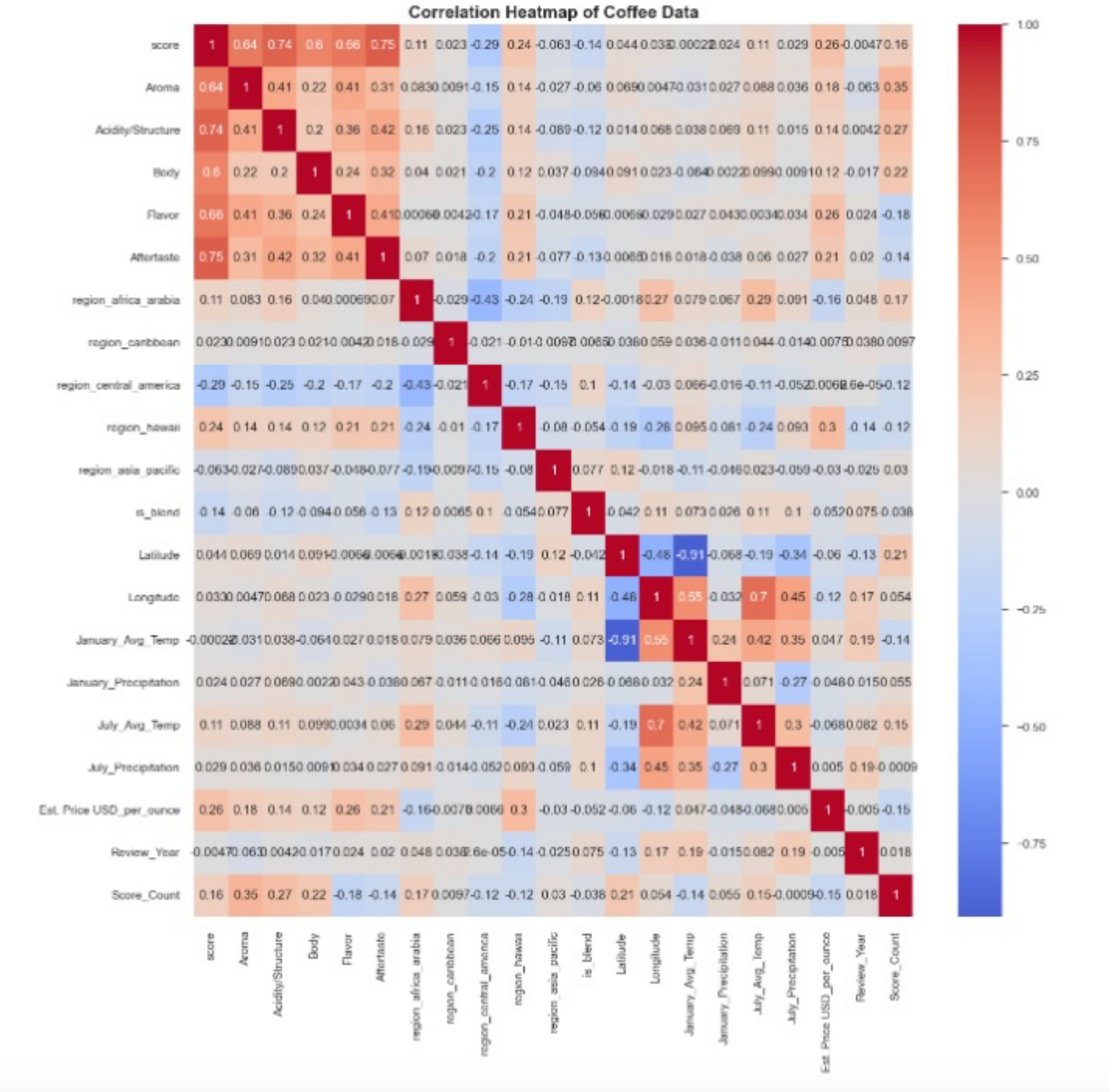


Figure 6: Correlation heatmap of each feature in coffee review dataset

Data Inferential Statistic

In order to examinate hypotheses and assumptions, the statistic approach is crucial for provide reliable conclusions from sample data. In this project, we used:

- ANOVA test: to examinate the relation between categorical feature and continuous numeric feature.
- T-test: to examinate the relation between categorical feature (which has two groups) and continuous numeric feature.

Below are the tests which be implied:

- Use ANOVA test to examine the relationship between the coffee origin(categorical) and the score(continuous numeric).

- Use t-Test to examine the relationship between the is_blend(categorical) and the score(continuous numeric).

- Use ANOVA test to examine the relationship between the roaster(categorical) and the price(continuous numeric).

- Use ANOVA test to examine the relationship between the roast level(categorical) and the score(continuous numeric) .

All those results show that p-value is less than 0.05 which means features have significant different. In addition, from the post- host test, we found that the region in Central America and Asia Pacific consistently show higher coffee scores. Also, by using post-host test, for different roast level, medium roast coffees demonstrated the highest average scores.

4. Machine Learning Methodology

To predict coffee ratings, multiple machine learning models were explored, ranging from simple linear models to more complex ensemble techniques.

Baseline Models

- Linear Regression: Used as a benchmark model to understand linear relationships between features and ratings.
- Decision Tree Regressor: Provided insights into feature importance but suffered from overfitting.

Advanced Models

- XGBoost: Achieved the best overall results with optimized hyperparameters.
- Multi-Layer Perceptron Regression: A deep learning model which consists of multiple layers of neurons that transform input data into an output through weighted connections and activation functions.

Hyperparameter Tuning

For optimal performance, models were fine-tuned using GridSearchCV and optimizing key parameters such as:

- Number of trees in Decision Forest and XGBoost
- Maximum depth of trees
- Learning rate for XGboost and MLP model

- Hidden layer sizes for MLP model

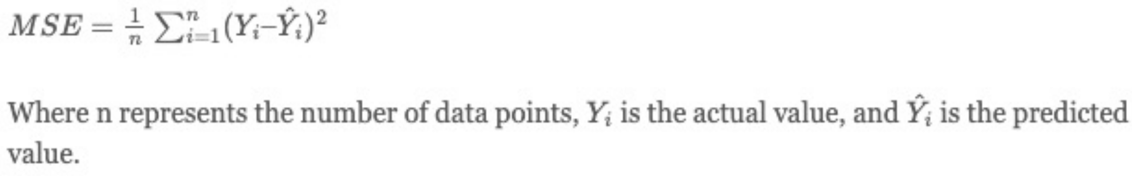
5. Model Evaluation

The models were evaluated based on key performance metrics:

- R² Score (Coefficient of Determination): it is used to measure the goodness of fit or best-fit line

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

- Mean Squared Error (MSE): it is the average of the squares of the errors or deviations. The error is the amount by which the actual values differ from the predicted values.



$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i – \hat{Y}_i)^2$$

Where n represents the number of data points, Y_i is the actual value, and $\hat{Y}_i$ is the predicted value.

Here is the table show results of all algorithms.

Algorithm	Model accuracy score(R2)	Mean Squared Error(MSE)
Ridge Regression	0.994493	0.012673
Decision Tree Regressor	0.983122	0.038837
XGBoost	0.986694	0.030617
MLP	0.563319	1.004845

Ridge outperformed other models, achieving the highest R^2 and lowest MSE which make it the most reliable predictor.

Feature Importance

Compared to Ride Regression and XGboost, we found that their top 5 feature importance in both models are almost same, though they are in a different order. In addition, sensory features have high score.

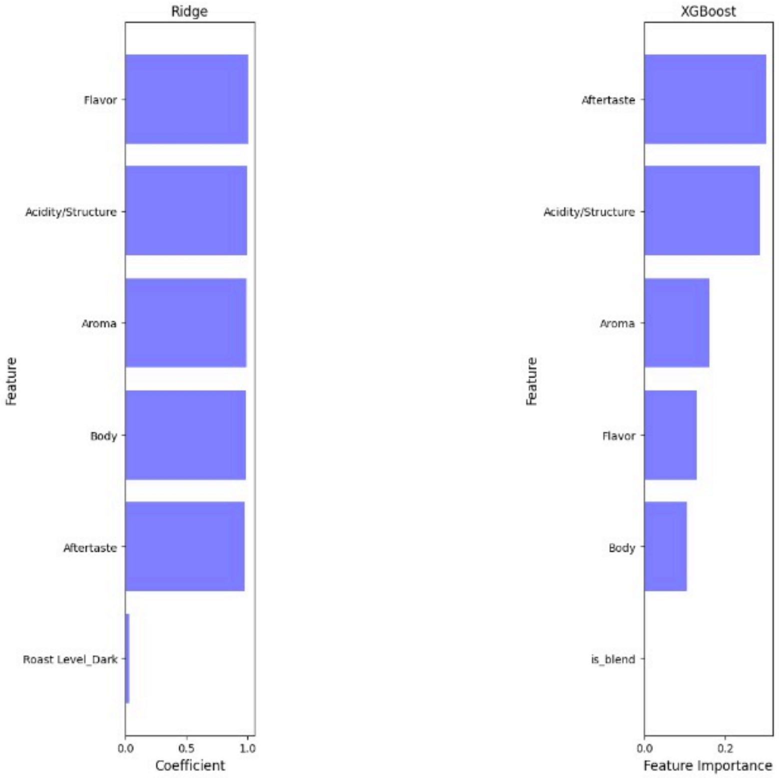


Figure7: Top5 feature importance in Ridge regression and XGboost

6. Future Work

To further improve the predictive model and derive more insights, future research could include:

- Sentiment Analysis: Using Natural Language Processing (NLP) to analyze customer reviews alongside numerical ratings.
- Market Pricing Analysis: Investigating whether coffee price correlates with ratings.

- Deep Learning Models: Exploring Convolutional Neural Networks (CNNs) to analyze coffee bean images for quality assessment.

7. Conclusion

This project successfully demonstrated that machine learning can predict coffee ratings with high accuracy. Ridge regression emerged as the best-performing model, highlighting the significance of sensory attributes, coffee origin, and processing methods in determining quality.

By implementing data-driven strategies, coffee producers and retailers can optimize production processes, enhance marketing efforts, and ultimately improve the overall quality of specialty coffee.

Reference

[1] Rodgers, E. (2025, February 25). *Coffee Statistics: Consumption, Preferences, & Spending*. Drive Research. https://www.driveresearch.com/market-research-company-blog/coffee-survey/

[2] *Coffee Review - the world's leading coffee guide*. (2025, February 13). Coffee Review. https://www.coffeereview.com/

[3] *geopy*. (2023, November 23). PyPI. https://pypi.org/project/geopy/

[4] OpenWeatherMap.org. (n.d.). *Weather API - OpenWeatherMap*. https://openweathermap.org/api