# Predicting Coffee Ratings from Expert Reviews

Kate Meredith

November 2022

## The Problem

Online ratings and reviews are a valuable tool for helping consumers make decisions. For businesses, this means ratings and reviews matter a lot. This project examined coffee reviews and ratings provided by a small team of expert reviewers at Coffee Review. Specifically, this project asks “How might we better understand the relationship between review and rating, so that coffee roasters can achieve top scores and distinguish themselves in a highly competitive market?” Beyond helping specialty roasters stand out, concepts used in this project could be applied to other markets that rely on ratings.

## Background

The global coffee market is valued at over 100 billion USD and is growing (Businesswire, 2021). Increased concern about the social and environmental impact of products created a competitive market for high quality, responsibly sourced coffee (Research and Markets, 2022). This project primarily focuses on this subset of the industry, as the data is scraped from CoffeeReview.com. While reviews span 1997 to 2022, the site’s focus in the last 10 years has shifted to high quality coffee (Coffee Review, 2015).

At Coffee Review, the evaluation (cupping) is conducted by a small, trained team on unidentified coffee. The reviewer scores five subcomponents (aroma, acidity, body, flavor, and aftertaste) from 1-10 (higher is better). Evaluative and descriptive narrative is also provided. A final overall rating from 50-100 is given (higher is better). Top scores are in the 90s, with anything above 94 exceptional. More details are available here.

This project uses natural language processing (NLP) and takes advantage of existing NLP tools like Natural Language Toolkit (Bird, Loper, Klein, 2009). The underlying problem is broadly applicable to businesses that rely on reviews. While some other projects do look specifically at coffee reviews, this is a unique dataset.

## The DataSet

Data was scraped for this project by the author from Coffee Review using Beautiful Soup, with a total of 6,567 reviews scraped. The reviews contain a mix of data types. Descriptive data includes: roaster location, review date, coffee origin, roast-level, Agtron, and estimated price. Numeric evaluative data includes scores for aroma, acidity/structure, body, flavor, aftertaste, and with milk. Narrative evaluative data includes the ‘the blind assessment’. Additional paragraphs include a mix of evaluative and descriptive data.

**Cleaning and Preprocessing**

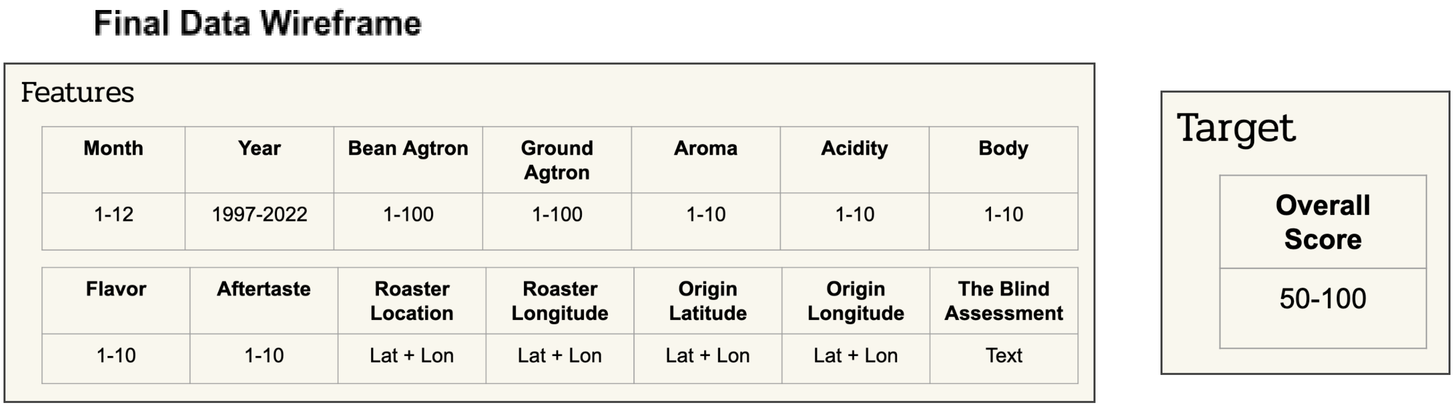
Some reviews were missing data, most frequently origin, price, flavor, aftertaste, and with milk. Cleaning therefore included sorting out which data belonged to which variable, and then addressing missing values. Where possible, missing values were filled in using correlation with other variables to determine what value to use. In a few instances the mean value was used.

All the data needed to become numeric to use in modeling. On some, the data type was simply updated, while others needed further transformation first (described below). Narrative text was transformed using NLP.

The roaster and origin locations were transformed to latitude and longitude using OpenStreetMap. Review date was changed to month and year. Roast-level was updated to an ordinal rating but later dropped due to overlap with Agtron, which measures roast on the bean and ground form. These measurements were split into two columns, one each for the bean and ground scores.

‘With Milk’ was dropped because there were too many missing values.  ‘Estimated Price’ was dropped because transforming it into meaningful data would have been too time consuming. Only the ‘blind assessment’ paragraph made it into the final model, as the other text was less effective at predicting outcome. Finally, the coffee and roaster names were dropped.

The final data wireframe (minus text vectorization) is shown along with pre-scaled data value ranges for each variable.



Final Data Wireframe

Features

| Month | Year | Bean Agtron | Ground Agtron | Aroma | Acidity | Body |
|---|---|---|---|---|---|---|
| 1-12 | 1997-2022 | 1-100 | 1-100 | 1-10 | 1-10 | 1-10 |

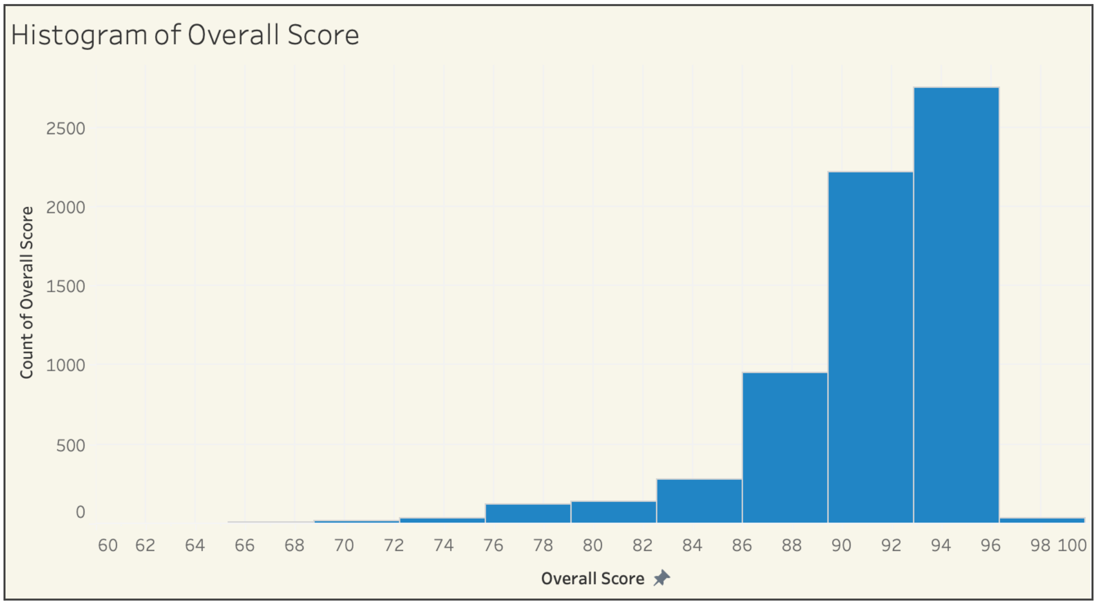
| Flavor | Aftertaste | Roaster Location | Roaster Longitude | Origin Latitude | Origin Longitude | The Blind Assessment |
|---|---|---|---|---|---|---|
| 1-10 | 1-10 | Lat + Lon | Lat + Lon | Lat + Lon | Lat + Lon | Text |

Target

| Overall Score |
|---|
| 50-100 |

**Exploratory Data Analysis (EDA)**

Histograms were created for each of the numeric variables, showing that the overall score tends to be between 90-96. This also revealed that the scores for aroma, acidity, flavor and aftertaste tend to be between 7-9. A correlation heatmap shows these same values tend to be highly correlated with the target value. When graphing the features against the target, these variables also had a fairly linear relationship. Given these findings, these features will likely be important when modeling. Regarding the text data, the top five most frequently occuring words from the final vectorizing method were “chocolate”, “sweet”, “toned”, “acidity”, and “fruit”.



## Modeling, and Results

Data was split into training (4,194 rows), validation (1,049 rows) and testing (1,311) sets. Model selection went through stages, using $R^2$ to assess model strength ($R^2$ closer to 1 is better).

First, XGBoost Regressor (XGBR), Random Forest, Linear Regression, Neural Network, KNN and Support Vector Regression (SVR) models were run on the numeric data to set a performance baseline. When testing with the validation data, XGBR performed best ($R^2$: 0.918), followed by Random Forest ($R^2$: 0.916) and then Linear Regression ($R^2$: 0.898).

Next, these NLP techniques were tried: CountVectorizer, TFIDF Vectorizer (non-stemmed and stemmed), and Word Embedding. Linear Regression was used to evaluate and optimize parameters. The chosen method was TFIDF without stemming ($R^2$: 0.744). The transformed text was then recombined with the rest of the data for final model optimization. The final dataset had 630 columns (features).

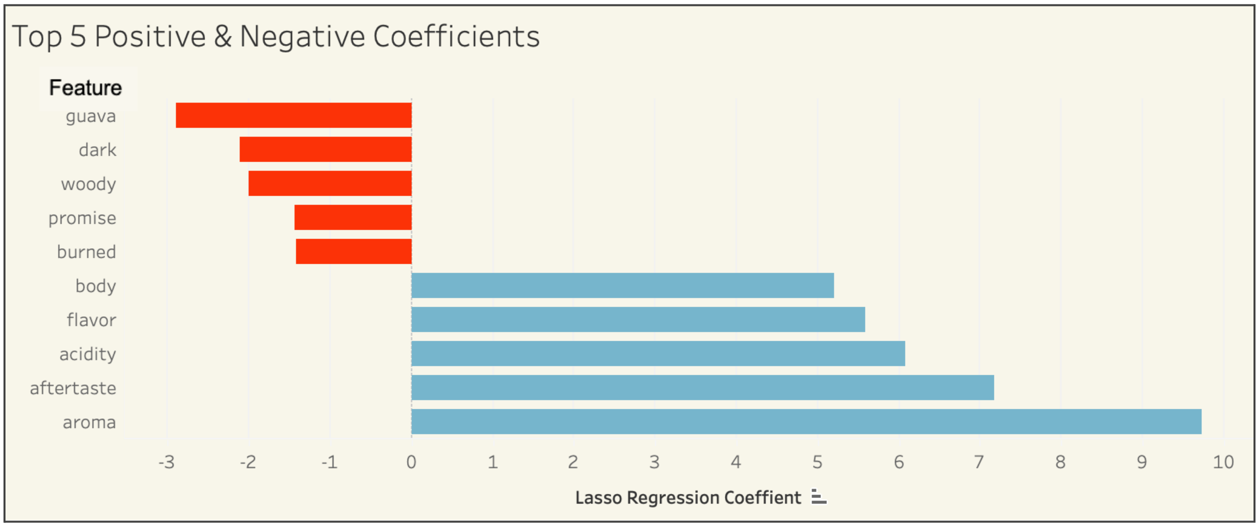
Model optimization on the full dataset used Linear Regression and XGBR. Principal Component Analysis (PCA) was applied to see how the models performed with fewer features. This reduced the features down to 280 while still representing 90% of the data variance. Parameters were optimized and tested using the validation data. In addition, the Mean Absolute Error (MAE) was also checked on each model to gauge accuracy (lower MAE is better).

Overall, PCA did not improve the models and obscured interpretability. However, when run on the full dataset, the models were very close in performance. In terms of $R^2$ the ElasticNet Regression on the non-PCA data performed best ($R^2$ of 0.906). Looking at the MAE, XGBR performed best (MAE of 0.593). The final selected model was the Lasso Regression model because it had strong results for both $R^2$ (0.904) and MAE (0.674) and retained interpretability (which is challenging with XGBR). Running the test data set on the Lasso Regression model yielded a $R^2$ of 0.910 and MAE of 0.691.

## Findings

The final model (Lasso Regression) performed quite well, both in terms of accounting for variance ($R^2$: 0.904) and accuracy (MAE: 0.674). However, the numeric data was crucial to achieving these scores, particularly (as hypothesized), the subscores for aroma, aftertaste, acidity, flavor and body. The text did not actually improve the model, but it does provide useful additional insights.

Regarding insights, the top positive coefficients (meaning, as these increase, the overall score tends to increase) were in this order: aroma, aftertaste, acidity, flavor and body. These were also the strongest absolute coefficients. The top negative coefficients (meaning, as these increase, overall score tends to decrease) were in this order: guava, dark, woody, promise, and burned. Perhaps most interesting was learning that aroma is the most important subscore, followed by aftertaste. This implies that first and last impressions are very important. It would be useful to talk to an expert about how these subscores usually translate to the final overall score, to see if this finding is expected. As for additional insights for roasters, ‘winey’, ‘almond’ and ‘hibiscus’ all surfaced as positive associations with the overall score and may be flavor profiles to explore. Roasters should stay away from ‘guava’, which has the most negative relationship with overall score. In addition, roasters should exercise caution with darker roasts, as ‘dark’ and ‘burned’ also had negative relationships with overall score.



Overall, using text for predictions was less effective than the rest of the data. It’d be beneficial to spend more time trying other methods to see if performance from the text alone could be improved.

Finally, other areas for improvement or future work include:

- Remove the approximately 300 score values found in the blind assessment earlier
- Include more of the text (2nd and 3rd paragraphs) when optimizing the models
- Try other methods for dealing with missing values
- Explore topic modeling

### Conclusion

Roasters looking to standout in a competitive market could benefit from the above insights. Further, the methods applied here could be applied to other markets reliant on ratings and reviews.

## APPENDIX

### Data Dictionary

The following provides a data dictionary for the final variables

*Features/Independent Variables:*

- **month:** 1-10, indicates the month the review was conducted
- **year:** 1997-2022,  indicates the year the review was conducted
- **bean_agtron:** 1-100, measurement taken using a special tool on the whole beans, lower numbers indicate a darker roast
- **ground_agtron:** 1-100, measurement taken using a special tool on the ground coffee, lower numbers indicate a darker roast
- **aroma:** 1-10, higher is better
- **acidity:** 1-10, higher is better
- **body:** 1-10, higher is better
- **flavor:** 1-10, higher is better
- **aftertaste:** 1-10, higher is better
- **roaster_lat:** latitude coordinate of roaster location
- **roaster_lon:** longitude coordinate of roaster location
- **origin_lat:** latitude coordinate of origin location
- **origin_lon:** longitude coordinate of origin location
- **p1:** "the blind assessment"; constitutes the narrative review portion; portion transformed with natural language processing

*Target/Dependent Variable*

- **overall_score:** 50-100, higher is better, more info on interpretation can be found here

## References

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

"Coffee Market Report 2021 - Global $100+ Billion Market Growth, Trends, COVID-19 Impacts, and Forecasts to 2026 - ResearchAndMarkets.com". Business Wire (2021). https://www.businesswire.com/news/home/20210601005832/en/Coffee-Market-Report-2021---Global-100-Billion-Market-Growth-Trends-COVID-19-Impacts-and-Forecasts-to-2026---ResearchAndMarkets.com

"Coffee Market - Growth, Trends, COVID-19 Impact, and Forecasts (2022 - 2027)". Research and Markets (2022). https://www.researchandmarkets.com/reports/5165416/coffee-market-growth-trends-covid-19-impact?utm_source=BW&utm_medium=PressRelease&utm_code=9qbbjh&utm_campaign=1545146+-+Coffee+Market+Report+2021+-+Global+%24100%2b+Billion+Market+Growth%2c+Trends%2c+COVID-19+Impacts%2c+and+Forecasts+to+2026&utm_exec=chdo54prd

David, Kenneth. "How Coffee Review Works". CoffeeReview (2015). https://www.coffeereview.com/how-coffee-review-works/

"Interpreting Reviews". Coffee Review. https://www.coffeereview.com/interpret-coffee/