Summer
2025

White Paper: Coffee Rating Predictions

DSC 680-T301

LUKE SYVERSON

Topic

The Coffee Rating Predictions project aims to inform the coffee processing and reselling industry of user preferences based on key words found within their reviews of different blends. By identifying the most influential words in a review, resellers may hone their procurement strategies to maximize value to the consumer and increase revenue.

Business Problem

With so many beans to choose from, it may seem impossible to determine the tastes of a given market, especially the discerning coffee consumer market. Providing an analysis of coffee preferences in the form of a prediction assists resellers in assessing the relative tastes profiles within their target demographics and adjusting their sourcing and refining practices to better suit the consumer and subsequently increase sales. Understanding consumer preferences is an arduous task as objective evaluations of product are rare and often noisy representations of the subset of the consumer base willing to submit feedback. While the feedback limitations are present, this analysis leverages a database of existing coffee reviews to better quantify the impact of subjective consumer evaluations on an objective, aggregate score.

Background

The data are pulled from the since-inactive coffeereview.com website and preprocessed in a Kaggle dataset. The feature set of 12 denotes the origin information of a given reviewed blend, along with the date of the review and text itself. With a light record count of 1,267, opportunities to expand the analysis for a productionalized process exist, but the premise of this project still holds water. Or coffee.

Luke Syverson
DSC 550-T301

Preliminary Analysis

The dataset used leverages 1,246 reviews from 2017 onwards to rate the same number of distinct blends. Ratings in this sample are left-skewed, normally distributed, and range from 84/100 to 97/100.

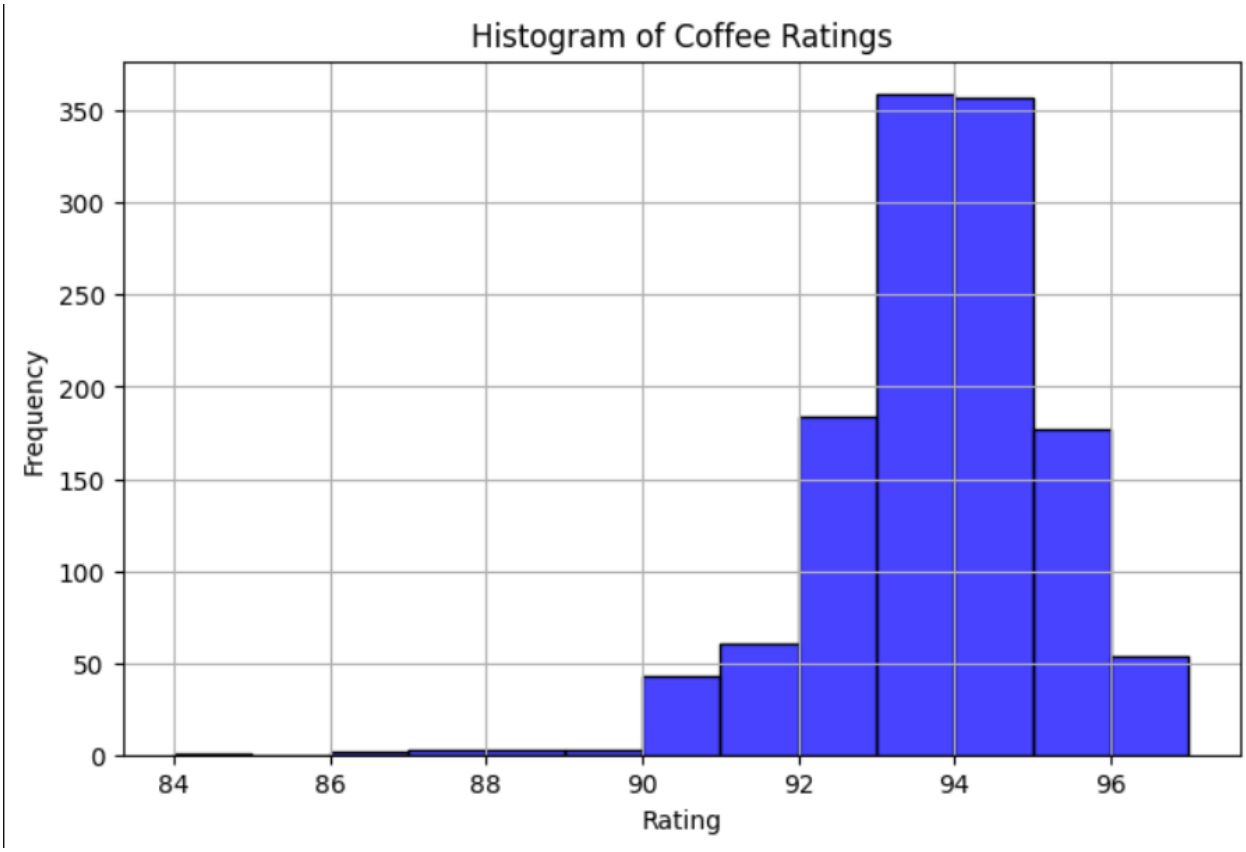

Figure 1: Coffee Rating Histogram

The presence of minimal outliers is later revealed to have an insignificant effect on our model's performance, which is expected to be per the small sample size. The small range of ratings implicates potential unsuitability of the dataset to generalize to other models of the same scale, but it seems the unimodal distribution may fit the population rating distribution well.

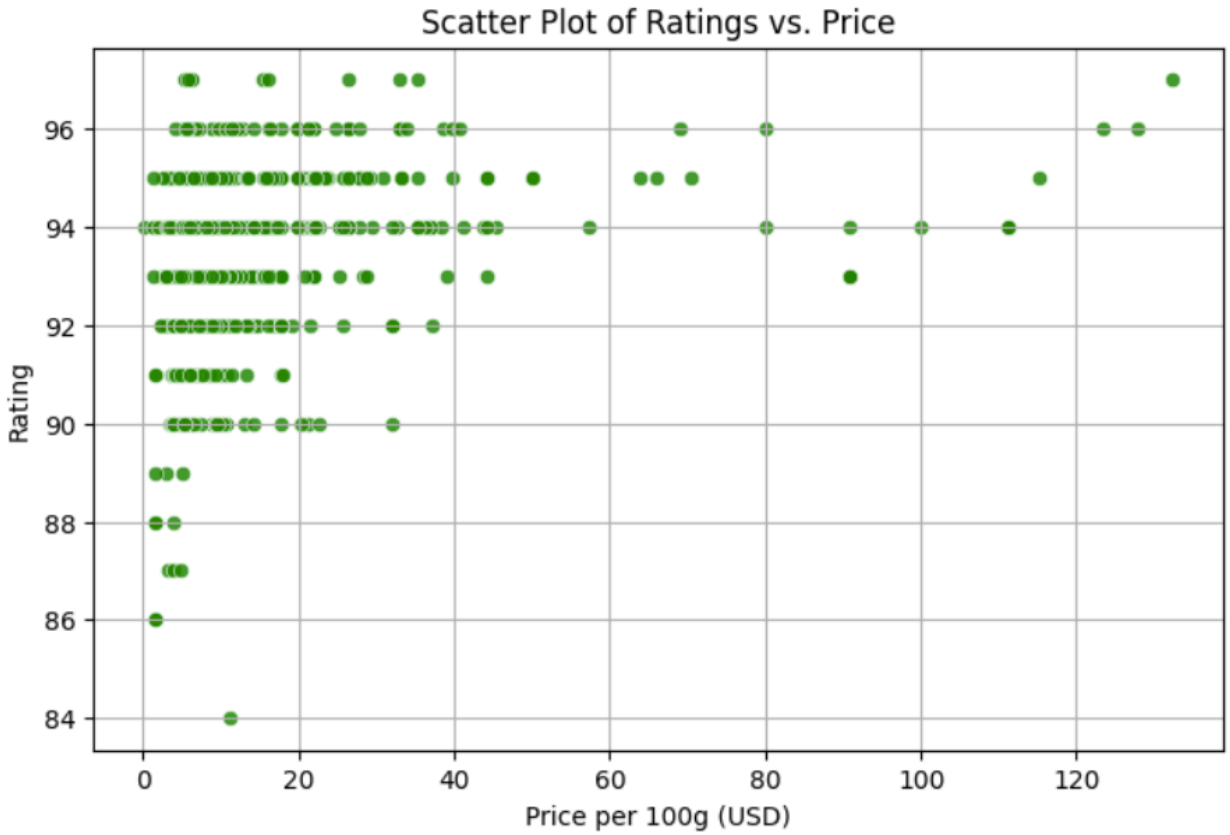

Figure 2: Ratings vs. Price Plot

Observations near the right edge are more costly, but it appears that price isn't enough to qualify for a higher rating. Furthermore, the tendency of price is to stay below the $40 per 100-gram rank, which indicates the popularity of relatively expensive blends is exceptional. Coffee consumers are price conscious, since the amount and frequency of consumption is high, and 100 grams of coffee isn't enough to sustain consistent consumption over a few months for the average consumer.

Methods

For the first-pass analysis, lightweight data munging was performed, including an imputation of "N/A" for null roast types and the encoding of categorical features. A simple TF-IDF matrix is then created and encoded to capture review "sentiment" by means of the top 100 terms. Lastly, an "origin price average" feature was engineered to quantify the potential interactive relationship between price and blend origin, possibly in service to coffee firms seeking to expand procurement practices in a region. The potential for multicollinearity is accounted for with different regression models, and a slim feature set is chosen for initial modeling purposes.

The initial feature set:

- Roast
- Loc_country (blend location)
- Origin
- 100g_USD (price (USD) per 100 grams)
- Review (in the form of tf-idf terms)
- Origin_price_avg (engineered feature)

The quantitative ratings warranted a regression approach. Four regressions are employed to provide for a variety of data behaviors: Linear, Lasso, Ridge, Elastic Net. A Random Forest regression model was also included for feature importance and baseline control against possible regression assumption violations.

Analysis (Model Evaluation)

All models performed accurately, given the possible range of 13. Prediction variance didn't exceed more than 2.25 for MSE in the worst models, and 0.9 for MSE in the best.

The best performing model in this dataset was the linear regression model. Given the normal nature of the data with minimal complications, the simplest model achieved the best accuracy, with prediction errors (MAE, MSE, RMSE) all less than one, and an explained variance of 62%. The Ridge regression model performed closely but didn't benefit from the additional complexity imposed on the data. This discrepancy may not apply to other datasets exhibiting different central tendencies and is thus included in the model with the rest of the regression models.

The Random Forest outperformed the Lasso and Elastic Net regression models but underperformed relative to the Linear and Ridge regression models. Presumably, the additional complexity of these three models wasn't serviced by the data and led to lesser accuracy. The presence of many features due to dummy variable initialization was more significant than the standard-tuned Lasso & Elastic Net models could compress and functionally added noise to the Random Forest model's predictions.

High rating accuracy +/- 1 is certainly an acceptable accuracy for most use cases. Achieving such accuracy while only explaining ~ 60% of the variance indicates a generalizable model, which may be applied to other datasets with different variance profiles and still enjoy some degree of predictive accuracy.

The presence of review terms in the Random Forest feature importance reveals the significance of review text as a predictor. Notably, price is the second-most important feature, with origins and locations rounding out the bottom.

	Feature	Importance
95	juicy	0.116664
0	100g_USD	0.095447
80	flavor	0.067593
139	velvety	0.047020
60	chocolate	0.034066
..	...	...
7	loc_country_Canada	0.000005
23	origin_Dominican Republic	0.000003
34	origin_Nicaragua	0.000000
14	loc_country_Kenya	0.000000
15	loc_country_New Taiwan	0.000000

Figure 3: RF Feature Importance

Conclusion

Given the relatively low importance of some origin and location dummy variables and the of multicollinearity within the feature set, further revisions may benefit from including only one or neither of the geographic features. More interactions between price and other features may be insightful, as the predictive power of price could reveal other interactive relationships within predictors. Of course, testing on a larger dataset would presumably aid the model's accuracy and generalizability, but in its current state, the model performs well and exhibits prediction potential for larger samples.

Assumptions & Limitations

Derived value from coffee (as an abstraction of rating) isn't just determined by price, and reviewers are likely to select and review blends that are of a higher standard than unremarkable or common blends. This potential selection bias trains the model to favor prediction on more specialty blends, which will service the business seeking to understand coffee consumption in their consumers as more than just a formality, but an extension of value via hospitality or sales revenue.

The lack of redundancy among observations means that the profile and rating of each blend is a single-pass operation, which can't be construed as an objective evaluation of the blend. Expanding the dataset to include more reviews of the same blends may be helpful, but adequate normalization should be in place to control the more popular blends on the review site.

Challenges

The slim feature set might not hold prediction potential required to achieve adequate performance, so feature engineering may be a prerequisite for the model. The small sample size also might lend to overfitting, so methods to hedge against model overperformance should be considered depending on the outcomes of the first pass modeling.

Future Usage & Implementation

Care should be taken to engineer features that align with the insights gathered from this modeling exercise. The price calculation requires a physical weight in grams, and the review text varies in length and verbosity, which means some configuration may be necessary to derive the greatest accuracy when generalizing to other samples. Null values in the categorical variables should be resisted by the model, provided they're not excessive.

Recommendations

Since the dataset holds volunteer-only reviews, future experimental design should strive for a more diverse review sampling by implementing collection practices to increase the relevance of the ratings to the broader coffee-consuming community.

Ethical Assessment

The prediction of ratings carries a subjective component that can't be verifiably generalized to a different segment of the coffee industry; specifically, blends reviewed on this site are predominantly sold to a non-retailer demographic of enthusiasts, specifically the subset population vocal enough to seek out a review site and submit reviews. The suitability of this sample to represent the overall coffee community is debatable, but the ease of access to the data model lowers the barrier to entry and permits a straightforward proof-of-concept project.

Sample Questions

1. How could coffee ratings follow a normal distribution?
2. What influence does price have on blend ratings?
3. Why purposefully include multicollinearity in your models?
4. What review terms are the most impactful to the ratings?
5. How accurate is the model, really?
6. Is there a selection bias in the dataset?
7. Can you generalize the results of this model?
8. How can you improve the experimental design of a coffee-rating prediction project?
9. What can the coffee industry do with this information?
10. What's the next iteration of this project?

Luke Syverson
DSC 550-T301

References

Schmoyote. (2023, January 19). *Coffee Reviews Dataset*. Kaggle. https://www.kaggle.com/datasets/schmoyote/coffee-reviews-dataset

Luke Syverson
DSC 550-T301