Prediction of Ratings Using BeerAdvocate Dataset

Zhicheng Wang, Mingyang Yao, Hantian Lin, Yiyu Weng[a]

[a]University of California San Diego, 9500 Gilman Dr, La Jolla, 92093, CA, United States

Abstract

In the digital age, reviews have become an indispensable part of consumer's decision-making, particularly when it comes to purchasing products and services. In this project, we focus on the prediction of the overall rating in beer reviews. We analyze the dataset to understand its characteristics and correlations with ratings thoroughly. We explore and implement a variety of predictive models, evaluating each to identify features that are the most significant. During optimization, we utilized a validation set to tune our model's parameters, ensuring that it reaches its maximum performance when compared to other alternatives. As a result, random forest regression outperforms all other models with well-tuned parameters. With this in mind, we would assume that the relationship between features and overall rating is nonlinear.

Keywords: Text Mining, Machine Learning, Optimization

1. Introduction

1.1. Dataset Overview

In this study, we present a comprehensive analysis of beer ratings using the BeerAdvocate Dataset, which comes from the BeerAdvocate website. We plan to build five models to predict the overall rating of each review with component ratings and analysis of the review text using this large dataset.

1.2. Predictive Task

The predictive task for this dataset, which includes beer reviews with multiple rated dimensions, is to predict a beer's overall rating (review/overall) based on a set of numerical, categorical as well as text features.

1.3. Preprocessing and Encoding

Our first approach involves a detailed examination of different aspects of beer: alcohol by volume(ABV), style, aroma, appearance, palate, and taste. The sensory aspects are given a separate rating by users and through correlation checks, we figured that these aspects are more significant than others as they are highly correlated to the overall rating of the beer. We employed regression models like linear regression, random forest regression, and gradient boosting regression which we will discuss in depth in later sections. These models allow us to not only predict the overall ratings based on these features but also understand the relative importance of each aspect in determining the overall rating. It also offers insights into which aspects are most critical to consumers.

We then leverage the textual content of reviews. Utilizing Term Frequency - Inverse Document Frequency (TF-IDF) Vectorization, we extracted features from the review text, focusing on the importance of specific words in the reviews. This is a technique that underscores the significance of words within the text. After using this technique our model improved by a significant amount.

In the final approach of our study, we explored user-item interaction. Unlike the other approaches, this method steps outside the review text and focuses on modeling user experiences, item popularity, and user-item preferences. This attempt considers a user's rating history, the overall rating of products, and the alignment of user preferences with product features. This shall provide a more dynamic understanding of interactions between users and products.

1.4. Model Selection

We select models to evaluate and use generally based on two sources. In the first place, we attempted representative recommender models and text mining techniques covered in the class, such as the latent factor model and TF-IDF measure on review text. Further, we pick random forests because according to previous empirical studies, random forest is a high-performance model overall.(3)

2. Related Work

Learning Attitudes and Attributes from Multi-Aspect Reviews by Julian McAuley, Jure Leskovec, and Dan Jurafsky, presented at the International Conference on Data Mining (ICDM) in 2012, explores the domain of aspect-based sentiment analysis using beer reviews as a case study. The authors propose a novel machine learning model, PALE LAGER, capable of predicting aspect labels, summarizing reviews, and recovering missing ratings.

The study utilizes two primary datasets: BeerAdvocate and RateBeer, both of which employ a five-aspect rating system encompassing four sensory aspects (feel, look, smell, and taste) and an overall rating. Similar datasets have been investigated in the past, including Amazon product reviews, where users rate items based on overall quality, with the Toys & Games category allowing further feedback through ratings on fun, durability, and educational value. Additionally, the audiobook rating platform Audiblecom allows users to rate audiobooks on a five-star scale while providing written reviews.

State-of-the-art methods for analyzing this type of data include latent Dirichlet allocation (LDA) and support vector machines (SVMs). LDA, a generative probabilistic model, assumes that documents are generated from a mixture of topics. LDA has been applied to beer reviews to identify the prevailing topics discussed. For instance, one study employed LDA to uncover the dominant topics in BeerAdvocate reviews, finding that taste, aroma, and appearance were the most prevalent.

SVMs, on the other hand, are discriminative classifiers trained to distinguish between different data classes. SVMs have been employed in beer review analysis to classify reviews as positive or negative. For example, a study utilized SVMs to classify RateBeer reviews, achieving an accuracy of 80

The paper introduces PALE LAGER, a novel machine learning model designed to predict aspect labels, summarize reviews, and recover missing ratings. The authors evaluated PALE LAGER on five million reviews from BeerAdvocate, RateBeer, Amazon, and Audible. The model demonstrated superior performance compared to existing methods across all three tasks.

The study by McAuley, Leskovec, and Jurafsky makes significant contributions to the field of aspect-based sentiment analysis by introducing a powerful machine learning model, PALE LAGER, and demonstrating its effectiveness on a large-scale dataset of beer reviews. The findings highlight the potential of PALE LAGER in various applications, including product recommendation, summarization, and sentiment analysis.

3. Data Descriptive Analysis

In our analysis, we first converted the 'review/time' field in our dataset to a datetime format, enabling us to extract the year and month from each review. This step was critical for understanding trends over time. As illustrated in Figure 1, there is a noticeable trend in the average beer ratings over the years. We observed a decrease in average ratings over time. This trend could be indicative of various factors, such as evolving user preferences, changes in the quality of beers reviewed, or even shifts in the demographic of the website's user base. The declining trend warrants a deeper investigation into the factors influencing user ratings and how these have evolved over the years.

We then explored the relationship between beer styles and

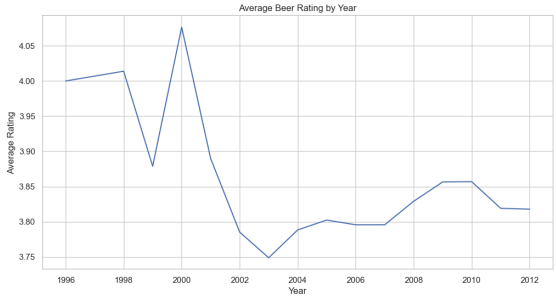

Figure 1: Rating over the years

their overall ratings, focusing specifically on the most popular styles within our dataset. After converting 'review/overall' to numeric values and addressing missing data, we identified the top 10 most prevalent beer styles for a more concentrated analysis. Among these styles, Russian Imperial Stout and American Double/Imperial Stout emerged as the most popular. As demonstrated in our box plot (Figure 2), we analyzed the distribution of ratings for these leading beer styles. The visualization offers a comprehensive view of how each style fares in terms of overall ratings, including aspects like median rating, range, and any potential outliers.

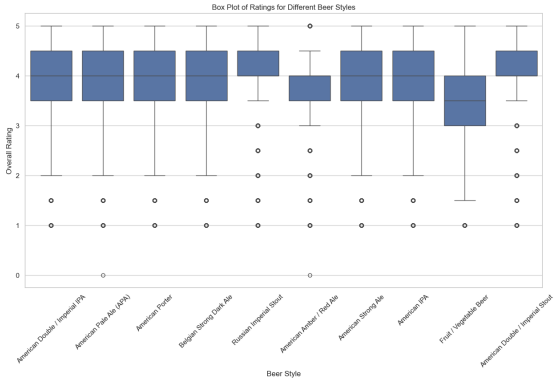

Figure 2: Top Beer Styles

For subsequent analysis, we conducted a correlation assessment to explore the relationships between various beer characteristics and the overall rating. Specifically, we examined 'beer/ABV', 'review/appearance', 'review/aroma', 'review/palate', and 'review/taste'. The results of this correlation check were quite revealing. We found that 'review/taste' and 'review/palate' exhibited the strongest correlations with the overall rating. This indicates that the taste and mouthfeel of the beer are paramount factors in determining how a beer is rated by users. In contrast, 'beer/ABV', which denotes the alcohol content, showed a relatively weaker correlation with overall ratings. This suggests that while ABV is an important characteristic of beer, it does not significantly influence how consumers rate the overall quality of the beer.

In the final part of our analysis, we delved into text analysis to gain insights from the linguistic content of the reviews. Utilizing a TF-IDF (Term Frequency-Inverse Document Fre-

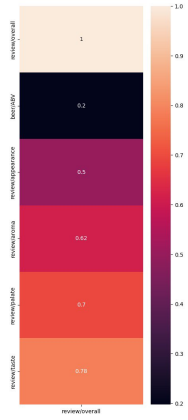

Figure 3: Feature Correlations

quency) approach, we first extracted feature names representing key words from the reviews. We then applied a linear model to determine the weights or coefficients of these features, which reflect their importance in the context of the reviews.

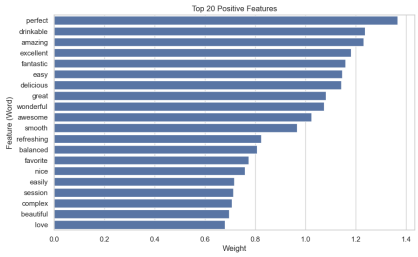

Figure 4: Positive Features

The top 20 features with the highest positive weights, as displayed in our first bar plot, revealed that words like 'perfect', 'drinkable', and 'amazing' were among the most influential in positive reviews. These words, carrying strong positive connotations, suggest that aspects such as the drinkability and overall excellence of a beer significantly contribute to higher ratings.

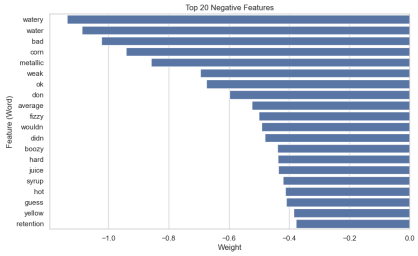

Figure 5: Negative Features

On the other hand, The frequent appearance of 'watery' and 'water' in less favorable reviews suggests that reviewers often associate these terms with a lack of flavor, body, or complexity in the beer. Such descriptions are typically indicative of beers that fail to meet expectations in terms of taste, richness, or overall quality.

4. Models and Evaluations

4.1. Baseline Model

For the baseline model, we calculate the average of overall ratings for each beer style and predict all overall ratings as the average of their corresponding beer style.

4.2. Random Forest Regressor

We chose random forest to predict the overall rating. Random forest combines the output of multiple decision trees to reach a single result. It works well with a mixture of numerical and categorical features, making them suitable for datasets with diverse types of information. Our dataset is a combination of numerical, categorical features, and textual data. Based on these characteristics, we believe that random forest will show good performance on the predictive task. We optimized the model through two main hyperparameters, the number of estimators (n_estimators) and the maximum depth of the trees (max_depth). Through our experiment, we found that when taking n_estimators=100 and max_depth=10, the model outperformed.

When taking n_estimators=1000, its MSE is much worse than that when n_estimators=100, since a large number of estimators causes the overfitting.

During experiments, we found that the feature 'beer/ABV' is not highly correlated with the overall rating. Including this feature increases the MSE. Thus, we decided to exclude 'beer/ABV' from our model.

Advantage: Handles non-linear relationships well, and provides feature importance.

Disadvantage: May not capture intricate patterns in highly complex data.

4.3. Latent-Factor Model

We considered using a latent-factor model for predicting the overall rating but we ran into several issues. Since this model is taking user-item interaction as the feature, we had to organize the dataset into a user set and an item set. However, not every review provides a user profile name, resulting in us having to cut off data that are missing the profile name attribute, which could exclude some crucial data. Also, this model cannot include review text as the feature since it is not numerically based. It would disregard other categorical ratings such as appearance rating and aroma rating as well. Another problem is that the latent factor has too many hyperparameters that need to be experimented with, including gradient descent optimizer learning rate, number of factors, lambda value, and training steps. The entire process is not efficient. By increasing or decreasing one or more hyperparameters, it is unpredictable whether the MSE would increase or decrease. The only consistent change is that increasing the training step would always bring MSE down. However, reaching a significantly low MSE would require a significantly large training iteration, which could take very long to finish running. We also tried to implement early stopping inside the training step function but the objective does not consistently decrease as interactions increase, which nullified this approach.

Overall, even though a latent-factor model might yield a good result after extensive training steps and careful hyperparameter tuning, it is not the most efficient one in this situation.

Advantage: Has the potential to outperform other models by optimizing hyperparameters.

Disadvantage: The tuning process is extremely burdensome and time-consuming.

*4.4. **Linear Regression***

Linear Regression was chosen as an initial model for this predictive task due to its simplicity and efficiency. Being a fundamental technique that is also covered in class, it helps in understanding the relationship between the various features and the overall rating.

Our optimization method for this model would be through feature selection and cross-validation. The model's performance will be evaluated on different combinations of features to find the most impactful set. We started off with only the ABV (Alcohol by Volume) feature which yields a relatively high MSE of 0.504, which in this case is even worse than our baseline, which is 0.41094117629368526. The possible reason is that linear regression assumes features have linear relationships while it's not the case in our dataset.

After discovering the fact that ABV is not highly correlated with the overall rating, we moved on to text analysis. The review text feature which contains the content of the reviews is transformed using TF-IDF Vectorization, a technique that underscores the significance of words within the text and converts them into a numeral format. After using this technique our model improved by a significant amount.

But we shouldn't stop there as there are numerous other features. After implementing the four categorical ratings: aroma, appearance, taste, and palate, our model was improved by another level. To push for perfection, we transform the time feature into DateTime objects, allowing us to analyze trends. These explorations lead to the conclusion that incorporating temporal trends would enhance the predictive capacity of our model.

Advantages: Efficiency and speed in training and prediction

Disadvantages: Assumes that there is a linear relationship between the features and overall rating which limits our ability to capture complex or non-linear relationships.

4.5. Gradient Boosting Regressor

Gradient Boosting Regressor was chosen for the prediction of overall rating because of its capability to find the potential non-linear relationships in the dataset. Also, its combination of multiple decision trees effectively reduces the risks of overfitting data when the depth gets too large.

Using our dataset, we include one-hot encoded styles, normalized ratings of ABV, appearance, aroma, palate, and taste, and tfidf of review text. To optimize the performance of the model, we then tried different combinations of features and attempted to perform fine-tuning of parameters focusing on the max_depth and n_estimators. During feature selection, we found that 'beer/ABV' has the lowest correlation with the overall rating while that of the other numerical features are all greater than or equal to 0.5. Therefore, we attempt two sets of features, one excludes ABV and the other includes all features mentioned above.

After running the model on different features and hyperparameters, we get the best features including all features, 500 estimators and max_depth of 3. This model will give us the MSE or 0.1479. Therefore, though the correlation between ABV and overall rating is not very high, adding this feature still slightly improved the performance as removing this feature will increase a little of the MSE for the same hyperparameters.

Advantage: Using multiple decision trees effectively prevents overfitting and able to find non-linear relationships just like random forests. Its runtime is better than Random Forest.

Disadvantage: The upper limit of this algorithm is relatively low. The hyperparameter tuning usually did not significantly improve the model performance. Compared with other machine learning algorithms, its runtime is slow still.

4.6. Model Evaluation

We will use Mean Square Error (MSE) to evaluate the models' performance, which quantifies the average squared difference between the estimated values and the actual value. The MSE is calculated as the following formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{1}$$

5. Results

After we attempted different combinations of features and carefully tuned hyperparameters for all models, we found that the Random Forests Regressor gives us the best model, whose lowest MSE is 0.124. Additionally, Gradient Boosting Regressors and Ridge Regression (regularized linear regression) are in second place, and their most optimal MSE is 0.148 and 0.152 respectively.

In contrast, the latent Factor model yields an MSE of 0.390. A comprehensive overview of the performance evaluations is presented in the following tables. Notably, the optimal MSE for each model is associated with distinct feature combinations. The complete table for different models and hyperparameters is in the Appendix.

	Combo 1[1]	Combo 2[2]	Combo 3[3]
Random Forest	0.1299	**0.1236**	0.1238
Gradient Boosting Regressor	0.1502	0.1493	**0.1479**
Linear Regression	**0.1529**	0.1531	0.1578

Table 1: Best Performance Models Comparisons

Combo1: Tfidf + ABV + style + appearance + aroma + palate + taste + time

Combo2:Tfidf + style + aroma + palate + taste + time

Combo3:Tfidf + ABV + style + appearance + aroma + palate + taste

6. Discussion

In this section, we discussed the results obtained from our experiments with different models and feature combinations.

6.1. Feature Importance

An intriguing observation from our experiments is the impact of different feature combinations on model performance. The tables present clear evidence that the optimal set of features varies across models. This underscores the importance of feature engineering and the need to tailor feature selections to the characteristics of each model. Also, selecting features based on correlations with our predicted targets provides insights about feature engineer.

6.2. Limitations and Future Work

It's important to acknowledge the limitations of our study. The choice of models and features may be influenced by the specific characteristics of the data set. Future work could explore additional models, conduct more extensive hyperparameter searches, or consider alternative feature engineering strategies.

6.3. Conclusion

In conclusion, our experiments have provided valuable insights into the performance of various regression models on the given data set. The Random Forest Regressor, with its low MSE, stands out as the preferred choice. Our findings contribute to the broader understanding of model selection, feature engineering, and hyperparameter tuning in regression tasks.

References

[1] Learning attitudes and attributes from multi-aspect reviews Julian McAuley, Jure Leskovec, Dan Jurafsky International Conference on Data Mining (ICDM), 2012

[2] From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews Julian McAuley, Jure Leskovec WWW, 2013

[3] Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," Proceedings of the 23rd international conference on Machine learning - ICML '06, 2006. doi:10.1145/1143844.1143865

Appendix A. Appendix Complete Hyperparameters, features, and Models Performance

Table A.2: Complete Model and Parameter Results

Model	Features	Normalized data?	MSE
Linear Regression(Default Parameter)			
	ABV	No	0.504
	Tfidf	No	0.3451
	Tfidf + ABV	No	0.3357
	Tfidf + ABV + style	No	0.3333
	Tfidf + ABV + style + time	No	0.3331
	Tfidf + ABV + style + time + appearance + aroma	Yes	0.1579
	Tfidf + style + appearance + palate + taste	Yes	0.1550
	Tfidf + style + aroma + palate + taste	Yes	0.1541
	Tfidf + style + appearance + aroma + palate + taste	Yes	0.1532
	Tfidf + ABV + style + appearance + aroma + palate + taste + time	Yes	0.1529
Ridge Regression			
α = 10	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1522
α = 15	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1521
α = 20	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1520
α = 21 (optimal)	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1520
Random Forests Regressor (n_estimator = 100, max_depth = 10)			
	ABV	No	0.3065
	Tfidf	No	0.3020
	Tfidf + ABV	No	0.3107
	Tfidf + ABV + style	No	0.3151
	Tfidf + ABV + style + time	No	0.3126
	Tfidf + ABV + style + time + appearance + aroma	Yes	0.1996
	Tfidf + style + time + appearance + palate + taste	Yes	0.1284
	Tfidf + ABV + style + time + appearance + palate + taste	Yes	0.1300
	Tfidf + style + time + aroma + palate + taste	Yes	0.1236
	Tfidf + ABV + style + time + aroma + palate + taste	Yes	0.1241
	Tfidf + style + appearance + time + aroma + palate + taste	Yes	0.1245
	Tfidf + ABV + style + time + appearance + aroma + palate + taste + time	Yes	0.1239
Latent Factor Model			
K = 5, Lambda = 0.00001, Training Step = 100	Interaction= profile name + beer id + overall rating)	No	0.39048
K = 5, Lambda = 0.00001, Training Step = 200	Interaction= profile name + beer id + overall rating)	No	0.38929
K = 5, Lambda = 0.0001, Training Step = 200	Interaction= profile name + beer id + overall rating)	No	0.42449
K = 5, Lambda = 0.00001, Training Step = 300	Interaction= profile name + beer id + overall rating)	No	0.39032
K = 5, Lambda = 0.00005, Training Step = 100	Interaction= profile name + beer id + overall rating)	No	0.41130
K = 7, Lambda = 0.00001, Training Step = 100	Interaction= profile name + beer id + overall rating)	No	0.39088
K = 7, Lambda = 0.00001, Training Step = 200	Interaction= profile name + beer id + overall rating)	No	0.39055
K = 7, Lambda = 0.00005, Training Step = 100	Interaction= profile name + beer id + overall rating)	No	0.41073
K = 10, Lambda = 0.00001, Training Step = 100	Interaction= profile name + beer id + overall rating)	No	0.39065
Gradient Boosting Regressor			
N_estimator = 50, Max_depth = 3	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1549
N_estimator = 100, Max_depth = 3	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1502
N_estimator = 500, Max_depth = 3	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1479
N_estimator = 100, Max_depth = 5	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.1493
N_estimator = 100, Max_depth = 10	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.153
N_estimator = 100, Max_depth = 20	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.164
N_estimator = 500, Max_depth = 10	Tfidf + ABV + style + appearance + aroma + palate + taste	Yes	0.155
N_estimator = 100, Max_depth = 5	Tfidf + style + appearance + aroma + palate + taste	Yes	0.15315
N_estimator = 500, Max_depth = 3	Tfidf + style + appearance + aroma + palate + taste	Yes	0.1486