CSE151A-Final_Report

Jian-peng Li, Yang Han, Leonard Shi, Zhaoyu Dou, Wen Hsin Chang

July 2024

1 Introduction

Beer is the most consumed alcoholic drink in the United States. It accounts for at least 40% of overall alcoholic drink consumption. Therefore, it would be very useful and profitable to predict certain features of specific types of beers that would offer better quality beers to the people who prefer them, which promotes sales. Our model focuses on predicting the overall score and classifying the beer style based on the other features. Such models can help the breweries and marketers gain insights into what aspects of different kinds of beer (appearance, aroma, palate, taste) are most influential in determining consumer satisfaction. This can help in tailoring products to meet consumer preferences. Breweries can also use the model to ensure consistent quality by monitoring the predicted overall scores based on various review aspects. This project aims to create a model that satisfies the stated demands. The past beer review database is being used as a training and test data set to develop the model. The model uses liner regression, Neural Network, and XGBoost methods. Since the dataset is large, cross validation is used to test the performance of this model to ensure consistency of training and testing data during developing this model. The model could later be used on other real world datasets to test its performance on different datasets.

2 Methods

2.1 Data Exploration

The dataset employed for our rating prediction task is sourced from the RateBeer dataset. It encompasses comprehensive information such as beer name, Alcohol By Volume (ABV), style, ratings covering various aspects (including overall rating), and the corresponding review text (refer to Figure 1). For the specific task at hand, we will focus exclusively on the 'review/overall' and other feature ratings. There are a total of 2785525 rows where each row corresponds to the feedback of one user toward a specific beer. The mean of the overall rating is 13.23 and the overall rating distribution is shown in the following figure 3. In this figure, we can see that the majority of the beers has the Overall rating between 10-17.5, which is a left-skewed distribution.

2.1.1 Descriptive Statistics

The table below summarizes the key statistics for each numerical feature in the dataset:

	beer/ABV	review/appearance	review/aroma	review/palate	review/taste	review/overall
count	5.0	5.0	5.0	5.0	5.0	5.0
mean	5.72	4.1	3.9	3.9	4.28	4.2
std	1.05	0.26	0.74	0.33	0.43	0.48
min	4.2	3.8	3.0	3.5	3.7	3.6
25%	5.0	4.0	3.5	3.8	4.1	4.0
50%	5.8	4.0	3.8	3.9	4.2	4.1
75%	6.5	4.2	4.5	4.0	4.6	4.4
max	7.1	4.5	4.7	4.3	4.8	4.9

Table 1: Descriptive Statistics

Analysis of Descriptive Statistics The dataset consists of 5 beer samples. Here are some interesting observations:

- The average Alcohol By Volume (ABV) is 5.72%, with a standard deviation of 1.05%, indicating a moderate variation in ABV among the samples.
- The review scores for appearance, aroma, palate, and taste have means around 4, with taste having the highest mean of 4.28.
- The standard deviations for the review scores are relatively low, suggesting consistent ratings across the samples.

2.1.2 Histograms

Histograms of the numerical features provide a visual representation of the distribution of each variable. This helps identify the central tendency, spread, and presence of outliers.

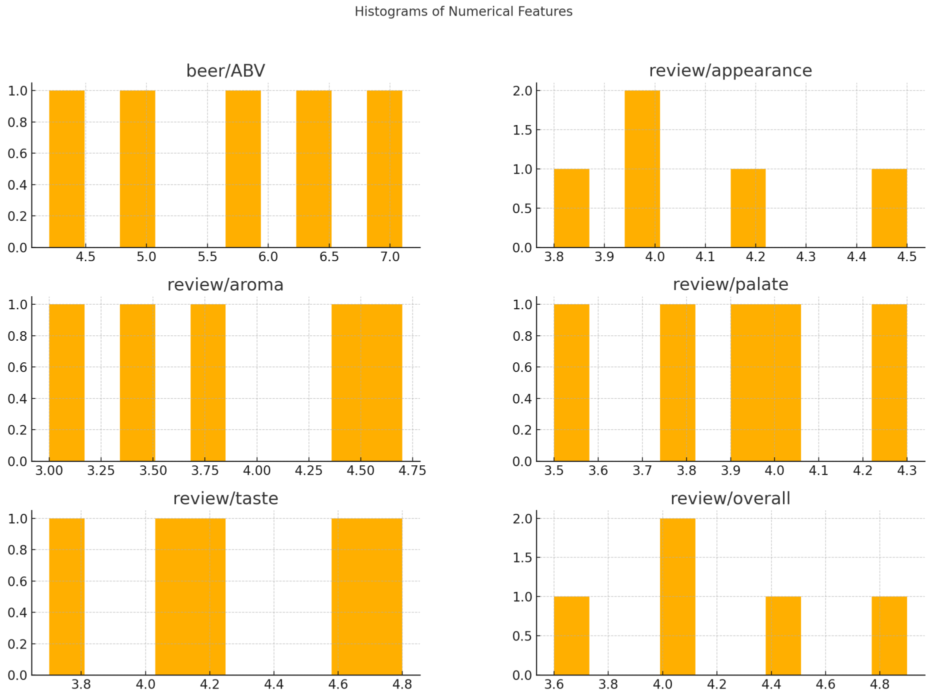


Figure 1: Histograms of Numerical Features

Analysis of Histograms The histograms reveal the following:

- The distribution of ABV is slightly skewed to the right, with most values concentrated around 5% to 6%.
- The review scores for appearance, aroma, palate, and taste are fairly normally distributed, with no significant outliers.

2.1.3 Correlation Matrix

The correlation matrix shows the relationships between pairs of numerical features. It helps identify which features are positively or negatively correlated.

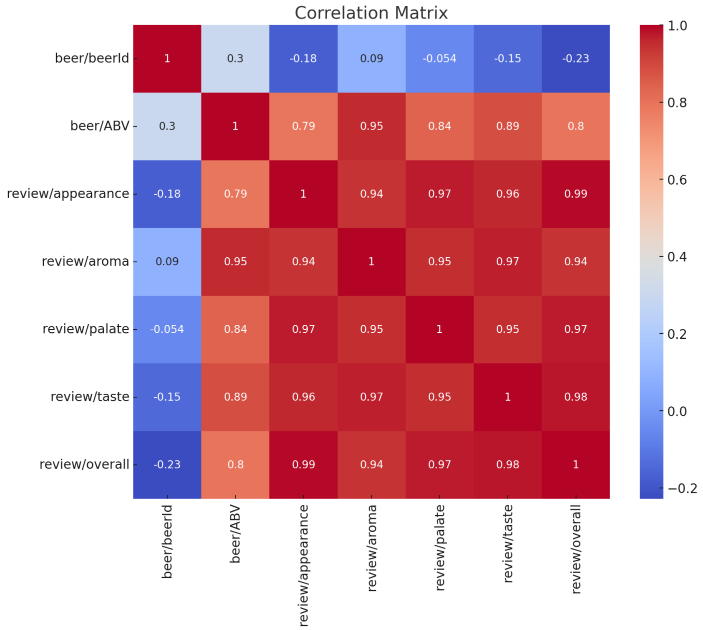


Figure 2: Correlation Matrix

Analysis of Correlation Matrix The correlation matrix reveals:

- There is a strong positive correlation between the overall review score and the review scores for taste, appearance, and aroma.
- The highest correlation is between the overall review score and the taste score, indicating that taste is a crucial factor in determining the overall rating of a beer.

2.1.4 Scatter Plots

Scatter plots of pairs of features reveal potential relationships and patterns between variables.

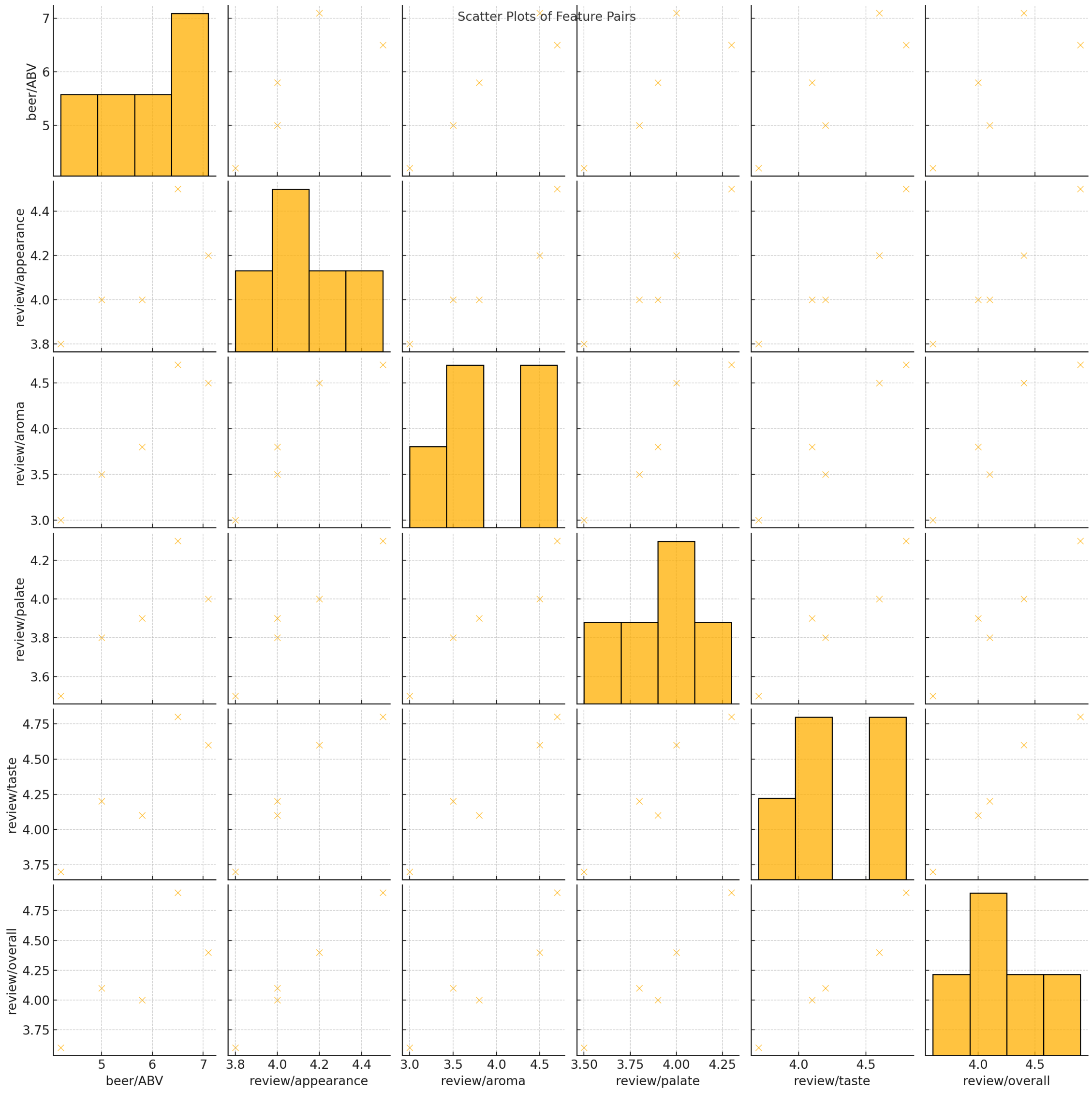


Figure 3: Scatter Plots of Feature Pairs

Analysis of Scatter Plots The scatter plots show:

- A linear relationship between overall review score and taste score, supporting the high correlation observed.
- Similar linear relationships between overall review score and appearance and aroma scores.

2.2 Model 1 : Linear Regression Model to Predict Beer Overall Score

2.2.1 Data Splitting

We split the data into training and testing sets using an 80/20 split ratio:

```
X_train, X_test, y_train, y_test = train_test_split(df_num.drop(['Overall'], axis=1), df_num
    .Overall, test_size=0.2, random_state=10)
```

2.2.2 Model Training

We train a linear regression model on the training data:

```
model = LinearRegression()
linear_model = model.fit(X_train, y_train)
print(model.coef_)
```

2.2.3 Model Evaluation

We evaluate the mean squared error by the sklearn library:

```
yhat_train_lin = model.predict(X_train)
yhat_test_lin = model.predict(X_test)
print(f"MSE for train: {mean_squared_error(y_train, yhat_train_lin)}")
print(f"MSE for test: {mean_squared_error(y_test, yhat_test_lin)}")
```

2.3 Model 2 : XGBoost

We employed the XGBoost regression model to predict the overall rating based on other feature ratings. We chose to examine XGBoost because it is an efficient and scalable implementation of gradient boosting, which builds an ensemble of trees sequentially. During preprocessing, we chose to one-hot encode the attributes [’Appearance’, ’Palate’] as they are categorical data, and standardized the attributes [’Aroma’, ’Taste’] given their normal distribution, as demonstrated in the figure below.

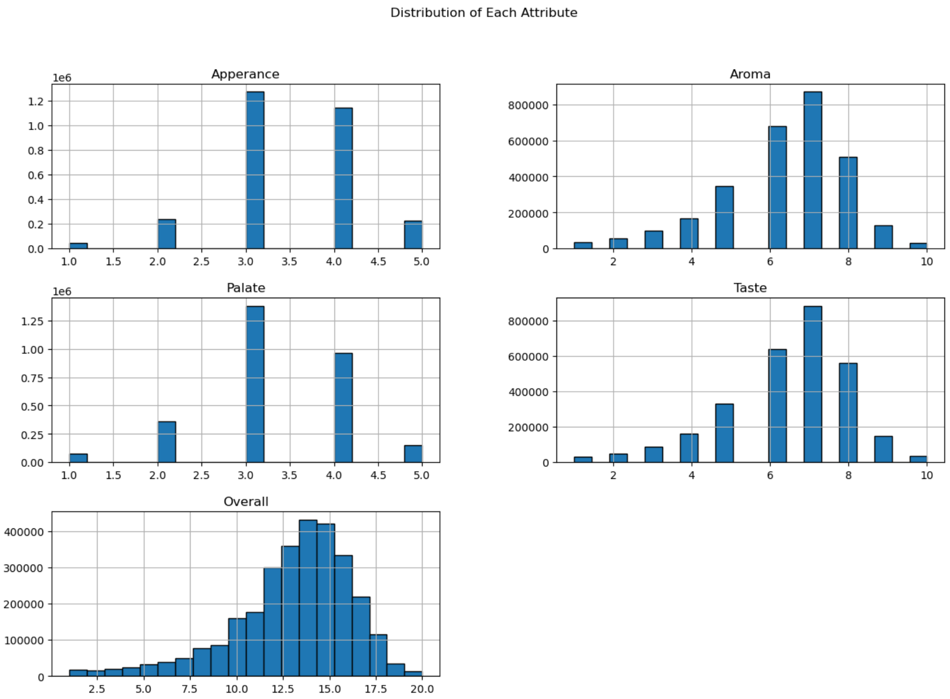


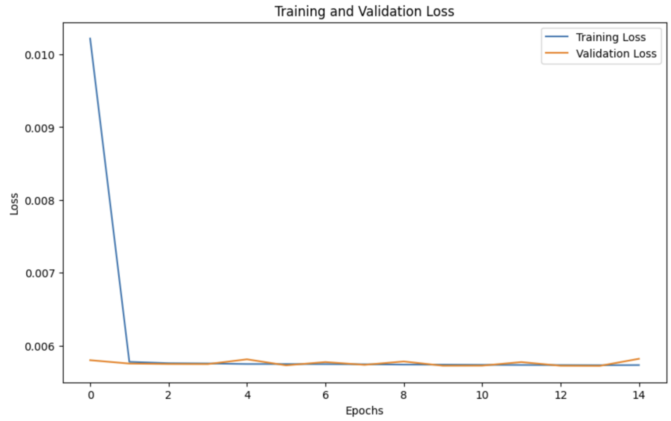
Figure 4: Feature distributions

After the data processing, we started with vanilla training without tuning the hyperparameters, which already yielded a reasonable output with MSE=2.09. This provided an optimal result, and we proceeded to examine grid search to find the best hyperparameters. With the findings *Best parameters:* {*`’colsample_bytree’: 1.0, ’learning_rate’: 0.1, ’max_depth’: 7, ’n_estimators’: 100, ’subsample’: 1.0`*}, And the mse drop from 2.09 to 2.06.

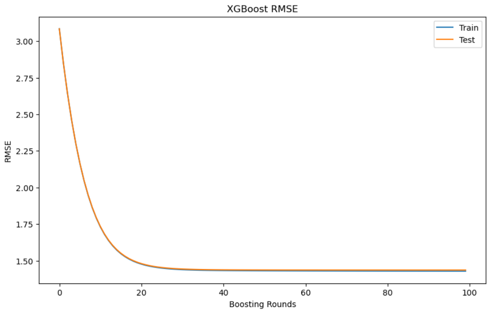
3 Results

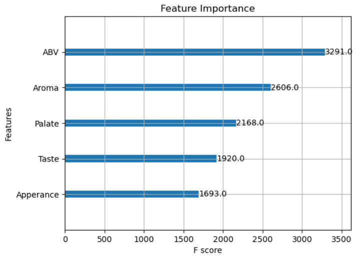
3.1 Results 1 : Linear Regression

In the model 1, Linear Regression Model, the MSE for the Training data is 0.006022, and the MSE for the Testing data is 0.006051 for the prediction of the overall score of the Beer.



3.2 Results 2 : XGBoost





For the baseline linear model, which is not tuned, we got the following results:

- Mean Squared Error: 2.0958
- R-squared: 0.8136

After tuning, we found that the best parameters are as follows:

- {'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'subsample': 1.0}

With these parameters, the Mean Squared Error drops to 2.0632. However, the data was not scaled correctly initially. Standardization was used on the non-normally distributed data in this case, which is the bad example we learn in classes. This results in significantly higher MSE. Figure below is the loss vs. boosting rounds.

4 Discussion

4.1 Disscussion 1 : Linear Regression

We chose linear regression for our first model since it is one of the simplest models. It would offer us a minimum expectation on how much time it would take to build a model on our dataset. Moreover, linear regression is less prone to overfitting, which is great for initial models since we have limited understanding of the structure of data. The low testing MSE justified the great capability of linear regression toward underexplored datasets. In addition, at the beginning of creating the linear regression model, we accidentally forgot to normalize our data, which resulted in an MSE around 2.1. After normalization, MSE decreased to around 6e-3, suggesting the necessity of normalization during preprocessing.

4.2 Discussion 2 : XGBoost

The XGBoost model demonstrated strong predictive capabilities for beer ratings, effectively capturing the relationships between the input features and the overall rating. The high R-squared value indicated that the model explained a significant portion of the variance in the overall ratings.

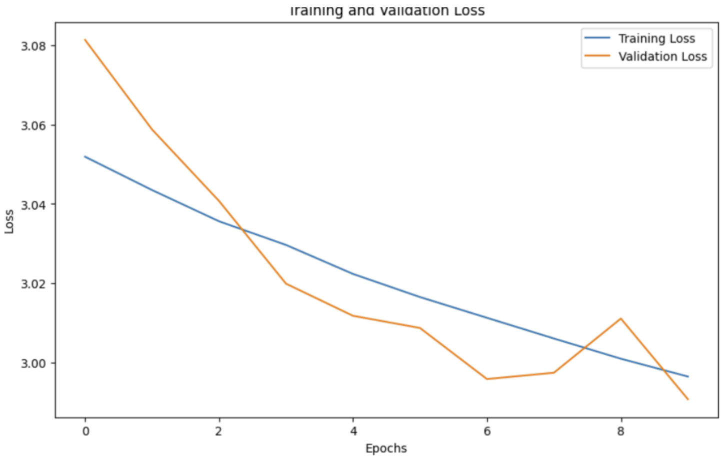
Feature Importance ABV emerged as a crucial factor influencing the overall rating. This aligns with the high correlation observed between ABV and the overall rating. However, the model could benefit from incorporating additional features, such as sentiment analysis of review texts, to capture more nuanced information.

Generalizability While the model performed well on the current dataset, its performance on other datasets should be evaluated to ensure generalizability. Different datasets may present unique challenges, and the model's accuracy could vary accordingly.

Model Complexity XGBoost is computationally intensive and can be sensitive to hyperparameter settings. Although our grid search improved performance, further hyperparameter optimization could enhance the model. Additionally, alternative machine learning algorithms or ensemble methods could be explored to compare performance and identify potential improvements.

4.3 Future direction

We also made a futuFor Neural Network prediction of the style of beer using other numeric features, despite the accuracy being low around 0.22, it is still significantly better than random, since probability of correctly predicting the style is 0.0124 (89 styles in total). Our neural network could be further improved by hyperparameter tuning, such as adjusting the number of units per layer and trying different activation functions, which should offer better accuracy of predicting the style. However, due to the time limitation, we were unable to conduct this process. This could be a future direction for our model development.



5 Conclusion

Our analysis and model development focused on predicting a beer's overall score using ABV, appearance, aroma, palate, and taste. Results showed that these characteristics are reliable predictors of overall scores, providing useful insights into factors that influence beer quality. However, the minor errors in our predictions highlight the complexity of assessing beer quality, which is influenced by many real-world factors.

For example, where the beer is sold and varying consumer preferences in different climates can significantly affect the overall score. By incorporating additional features such as marketing data, geographic sales information, and consumer demographics, we could achieve a more comprehensive understanding of the factors that affect beer quality.
However, adding more factors would increase the model's complexity, and it is impractical to account for all potential factors, as this could be an infinite proposition. Therefore, our model focuses on ABV, appearance, aroma, palate, and taste as the key attributes to predict a beer's overall score and achieve effective results. This approach allows us to provide valuable insights to brewers and marketers, helping them to improve product quality and better meet consumer preferences.

We observed that data normalization plays a crucial role in improving model performance. Initially, the data was not scaled, which led to suboptimal results. After normalization, the MSE significantly decreased, underscoring the importance of proper data preprocessing.
Despite these successes, there are several areas that could be improved. Firstly, incorporating additional features such as analysis of review texts could provide more nuanced insights and potentially enhance model performance. Secondly, exploring other machine learning algorithms or ensemble methods might offer further improvements. Thirdly, doing PCA would allow us to reduce the dimensionality of our data and thus discover new patterns in data, which offers better understanding of the structure of the dataset. Finally, evaluating the model's performance on different datasets would help ensure its generalizability and robustness.

In conclusion, while our models have shown promising results, there is always room for enhancement. Future work could focus on hyperparameter tuning for neural networks, exploring additional features, and testing the models on diverse datasets. These steps would improve the overall performance of our data on unseen real-world datasets.

6 Statement of Collaboration

Yang Han: Did data preprocessing and exploration, writing some part of report, improving codes, and proofreading.
Leonard Shi: Help to find the dataset and abstraction. Write the Read-Me and the written report.
Jian-Peng Li: Complete XGBoost training notebook, implement grid search. Complete DNN notebook, complete corresponding section on latex file.
Wen Hsin Chang: Train our first model, written report
Zhaoyu Dou: Train our first model and second model.