RateBeer Interaction and Recommendation

CSE 158 Fall 2023 Assignment 2

Introduction

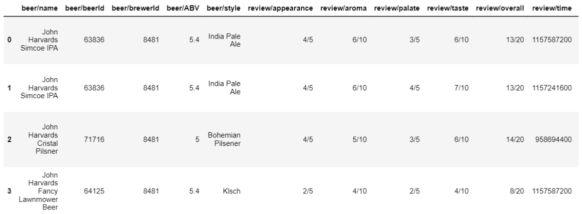
Predictive analysis plays a pivotal role in providing users with tailored recommendations. The ability to anticipate user preferences and deliver personalized suggestions is particularly valuable in the context of beer consumption. For our analysis, we turn to the RateBeer dataset, which is composed of reviews, aspect-specific ratings, as well as Alcohol By Volume (ABV) information. Understanding the dynamics of user preferences and the inherent characteristics of the items they consume is crucial for crafting an effective model. This report delineates our approach in building a robust predictive model tailored to the unique features of the RateBeer dataset, offering insights into the realm of beer recommendation systems.

Dataset

In this assignment, we chose to analyze a dataset of beer reviews from *RateBeer*, which can also be found on the *Recommender Systems Datasets* website provided by professor Julian McAuley. This dataset has 2,924,163 unique entries.

Before we generated any predictive models for the beer data, our initial endeavor was to unravel and comprehend the diverse fields within the dataset, and employ exploratory analyses to glean a foundational understanding. For this assignment, we elected to split the data into a training set and a validation set with 70% of the data being used to train.
The beer data is comprised of 13 fields, with each contributing to the holistic understanding of the review .Key attributes include the name field, housing the moniker of the beer, while the beer ID and brewer ID serve as unique identifiers for both the beer and its brewer.

Going through the dataset, we encounter features such as ABV, beer style, and scores for appearance, aroma, palate, taste, and overall impression. These numerical evaluations, ranging from 5 to 20, paint a picture of the sensory experience associated with each beer. The time of the review is included, as well as the profile name and review text.



	beer/name	beer/beerId	beer/brewerId	beer/ABV	beer/style	review/appearance	review/aroma	review/palate	review/taste	review/overall	review/time
0	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale	4/5	6/10	3/5	6/10	13/20	1157587200
1	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale	4/5	6/10	4/5	7/10	13/20	1157241600
2	John Harvards Cristal Pilsner	71716	8481	5	Bohemian Pilsener	4/5	5/10	3/5	6/10	14/20	958694400
3	John Harvards Fancy Lawnmower Beer	64125	8481	5.4	Klsch	2/5	4/10	2/5	4/10	8/20	1157587200

Fg 1. Overview of the Dataset

We first created a pandas dataframe with our data, and then cleaned it. We ensured the data was appropriately filtered, all the null values were changed to values we could parse, and any added characters in the dataframe were removed. After performing an exploratory data analysis, we found some interesting insights. There are 110299 unique beers, 610 unique ABVs, 29,265 unique users, and 2,844,800 entries with review text.

Predictive Task

The predictive task undertaken in this analysis revolves around forecasting beer reviews based on relevant features. This is pivotal in offering users personalized recommendations aligned with their taste preferences. The goal is to determine whether there exists a discernible relationship between the ABV of a beer and its overall review rating. To evaluate the performance of our predictive model, we will employ standard regression metrics, such as Mean Squared Error (MSE) and R-squared, which provide a quantitative measure of the model's accuracy in predicting the overall review rating. We will also qualitatively inspect scatterplots to visually compare the predicted ratings against the actual ratings. To establish a baseline for comparison, we will employ a simple linear regression model, treating ABV as the independent variable and the overall review rating as the dependent variable. This straightforward approach allows us to establish a baseline for predictive accuracy.

In the pursuit of a comprehensive analysis, we meticulously processed the data to extract relevant features. The first bar chart illustrates the distribution of the top 100 Alcohol By Volume (ABV) values in the dataset. ABV is a critical feature, as it encapsulates the alcohol content of a beer, which often influences taste preferences. The chart provides a nuanced breakdown, enabling us to discern the prevalence of different ABV values within the top 100. This information is essential for understanding the landscape of beer preferences and contributes to feature selection.

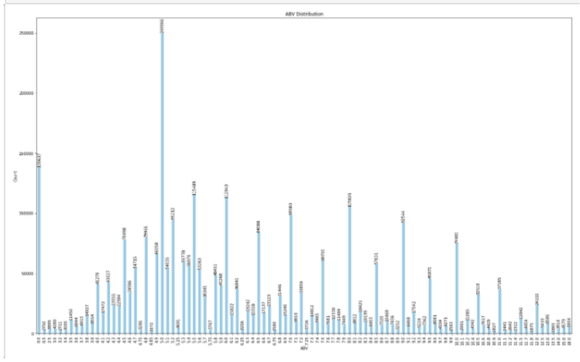
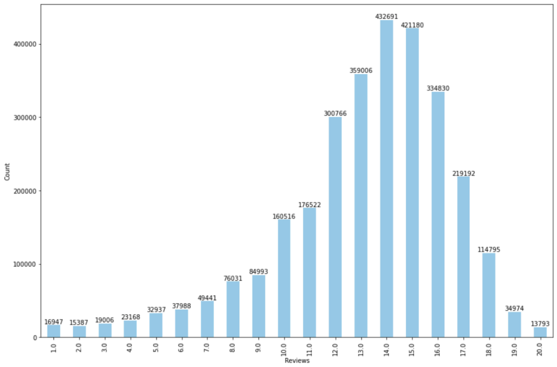


Fig 2. ABV Distribution

The second bar chart depicts the distribution of overall review ratings. This chart showcases the frequency of each overall rating, allowing us to identify the dominant trends in the dataset



.

Fig 3. Overall Review Distribution

Furthermore, the scatter plot depicting the relationship between ABV and overall review ratings visually conveys potential patterns and correlations. This exploratory visualization serves as a precursor to model development, offering insights into whether certain ABV ranges correlate with higher or lower overall review ratings. The plot lacked a clear linear trend, indicating that a simple linear regression model won't suffice for our predictions. This observation underscores the need for a more sophisticated model that can capture the relationship between ABV and overall reviews.

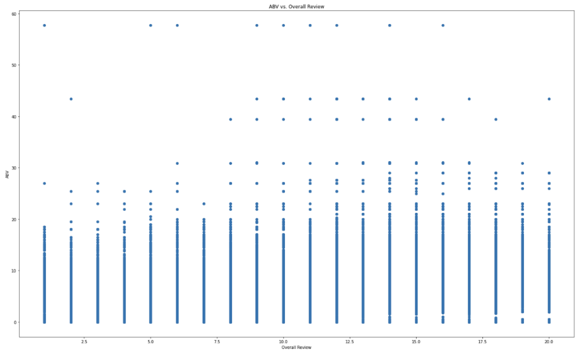


Fig 4. Scatter Plot: Reviews vs. ABV

The primary feature considered for this predictive task is the ABV of the beers. The ABV values were extracted from the 'beer/ABV' column of the dataset. As the independent variable, ABV serves as a proxy for the alcoholic strength of a beer, a factor that could potentially influence consumer perceptions and preferences. Preprocessing steps involve handling missing data, normalizing features, and potentially encoding categorical variables.

To ensure the validity of our model, we will perform cross-validation, splitting the dataset into training and testing sets. This approach allows us to train the model on one subset and validate its performance on another, giving us a more reliable estimate on how well the model generalizes to unseen data.

Model

For the linear regression model, the chosen features include the 'review/taste' aspect-specific rating. This feature was selected as it represents a critical dimension of consumer perception, capturing the taste quality of the beer. The 'review/overall' rating serves as the dependent variable, quantifying the overall satisfaction with the beer. The use of linear regression was decided under the assumption that the overall review rating can be linearly modeled based on the taste aspect, which may oversimplify the complex nature of beer reviews. The Mean Squared Error (MSE) is computed to assess the performance of the linear regression model, providing a quantitative measure of the prediction accuracy.

The logistic regression model addresses the classification aspect of the predictive task, aiming to distinguish between high and low overall review ratings.In this case, an overall rating of 15 or above is considered a "high" rating. We created a feature function that processes the taste values by converting them into a numerical scale and normalizing them on a scale of 0 to 10. The linear regression model is trained on these features to predict the overall review rating.

We decided to use logistic regression because of its interpretability and suitability for binary classification tasks. The model's performance is evaluated using the Balanced Error Rate (BER), a metric that considers both false positives and false negatives, offering a balanced assessment of classification accuracy. We received a resulting BER of 0.1948 as a baseline.

The linear regression model is optimized through minimizing the MSE, achieved by adjusting the coefficients during training. For logistic regression, the balanced class weight helps mitigate issues related to imbalanced datasets, ensuring that both high and low ratings contribute equally to the model training. The resulting MSE is regularized by a constant so that the overall results are easier to visualize. We received an MSE of 2.609 after this regularization.

Model Improvement Attempts

The initial attempt to improve the predicting accuracy of the model was done through the Bag of Words model, which extracted features from the review text to use as a model. The data is split with a 90-10 training-test data split. Then, an instance of the CountVectorizer class is created. This vectorizer is then fit to the training data (X_train) using the fit_transform method, which converts the text data into a bag-of-words representation. The same transformation is applied to the test data (X_test) using the transform method. An instance of the StandardScaler class is created with the parameter with_mean set to False. This means that the scaling is applied without centering the data (removing the mean). The scaler is then fitted to the bag-of-words transformed training data (X_train_scaled) and applied to scale it. The same scaling is applied to the test data (X_test_scaled). Since the dataset contained over 2 million unique entries, the model only contained a random subset of 10,000 entries to save time with the data processing. Breaking the entire dataset with over 2 million unique entries into training and testing subsets would yield even more accurate results. Our model yielded an overall accuracy of .2113 or a 21.13% chance of guessing the correct overall review based on the review rating text. Though low, this model provides some insight to a possibility in creating analyses purely based off human sentiment data.

Our next attempt to improve the accuracy further was with a simple text classification pipeline that uses TF-IDF vectorization to convert text data into numerical features and then trains a logistic regression model on the transformed data. This process began with imports of the TfidfVectorizer class from scikit-learn, which is used for converting a collection of raw documents to a matrix of TF-IDF features. Similar to the Bar of words model, an instance of the TfidfVectorizer class is created. This vectorizer is then fit to the training data using the fit_transform method, which computes the TF-IDF weights for the words in the training data and transforms the training data into a TF-IDF matrix. The same transformation is applied to the test data using the transform method. Finally, an instance of the Logistic Regression class is created with the parameter fit_intercept set to False. The logistic regression model is then trained on the TF-IDF transformed training data with corresponding labels (y_train) using the fit method. This model slightly improved our accuracy to 0.2352 or 23.52%.

Literature

We see that there was a very important paper that was written based off of this dataset. This dataset comes from a website called RateBeer from the time period of Apr 2000 - Nov 2011. Consequently, this time period is known for the popularity of online forums that connect users based on personal experiences. This leads to an active forum with a vast sample size of users with different tastes and preferences. Similar datasets that have been used in projects like these are those from Tradesy and GameSwap. These datasets were studied in the same context of the RateBeer dataset. We see that they provide the same information about all the individual data points in the dataset: the have and want lists and the peer to peer interactions.

One talks about bartering systems. The goal of this paper is to discover that would happily swap a pair of items within their possession. We see that they use Matrix Factorization to come to the conclusions that they do and make predictions based on who will barter which beer. Another important contribution is that they want to factor in user behavior and previous purchases while recommending the beer. To achieve this, they added in the user's previous purchase history and recommended by taking that into account. They also compare their results with the state of the art item exchange method which I will elaborate on in detail later. The results of this experiment can not really be compared to ours as our model looks at the reviews that will be predicted while this looks at whether or not the given beer would be bartered off for another beer.

A similar paper using a similar dataset talks about how we can use implicit feedback to recommend beer using this dataset. The implicit feedback in question lies in the visual appearance of the beer. The paper talks about a similar paper written by professor McAuley in the past that uses CNN's to identify similarities between various pictures and then use this as feedback. This paper used the dataset that we are using for this assignment and various other datasets. This paper uses a dataset that comes from tradesy. Tradesy provides visual data as well as metadata. This information can be used along with the metadata to get feedback.

As previously mentioned, matrix Factorization is a state-of -the-art technique used to study these types of datasets. It basically has 2 vectors: one for the user and one for the items. Fow i is the embedding for user i in the user matrix and row j is the embedding for user j in the item matrix. A multiple of these matrices are approximations for the feedback matrix.

The item exchange method is another state of the art method in bartering systems. Bartering experiments are often performed on datasets such as this one. This method basically means that the users can publish lists that they want and another list of the items that they are willing to discard. This makes it easier to match users with each other.

Results and Conclusions

In conclusion, the RateBeer dataset consists of 40,213 users and 2,855,232 reviews, each consisting of entries on the beer's taste and a personalized review that allows for the creation of a model that can accurately predict the overall rating of a beer. In this report, we decided to first build simple linear and logistic regression models to better understand the relationship between the taste of the beer and the overall rating. We found that these two attributes of the RateBeer dataset correlate positively with each other, allowing us to obtain a small value MSE (2.609) for the linear model and a small BER (.1948) for the logistic model.

We then began to incorporate the review text attribute of the dataset to build a predictive model that would be solely based on the words in the review instead of taste. As mentioned previously, this was done through a Bag of Words that obtained an accuracy of 21.3%. This model did not perform as well as the regression models that incorporated the taste of the beer, but it was able to give us insights on how review text can be used to increase the performance of our previous logistic regression model.

We continued to increase the accuracy of the model by incorporating a simple text classification pipeline that uses TF-IDF vectorization to convert text data into numerical features. These numerical features allowed for our Bag of Words model to increase in performance, yielding 23.52% accuracy.

By comparing these results to our initial baseline models, we see that the sentiment analysis is not as accurate. The feature that relates taste to overall review is better at predicting, however, the Bag of Words gives a new approach to analyze the dataset. The RateBeer's nature of having an overall rating from 0 to 20 and taste from 0 to 10 allows for these base models to have a high accuracy despite their simplicity. Although our proposed model did not exceed the accuracy of the baseline, there is definitely room for improvement that incorporates sentiment analysis and the base regression models. By utilizing the text of these reviews, working to effectively grab the emotions of the RateBeer users, the model can increasingly become more accurate, opening the gateways to a more personalized predictive model.