CSE 158 Assignment 2

Group members: Kelly Huang, Zhoutianing Pan, Michelle Tong

Abstract

This report includes the exploratory data analysis which describes the Ratebeer's review dataset which serves for the predictive task: rating prediction. There are three models: baseline model, Jaccard Similarity Model and Unigram Bag-of-Word Model.

1. Exploratory Data Analysis

We are going to use Ratebeer reviews in our assignment. This dataset consists of beer reviews from ratebeer. The data span a period of more than 10 years (Apr 2000 - Nov 2011).

Dataset:

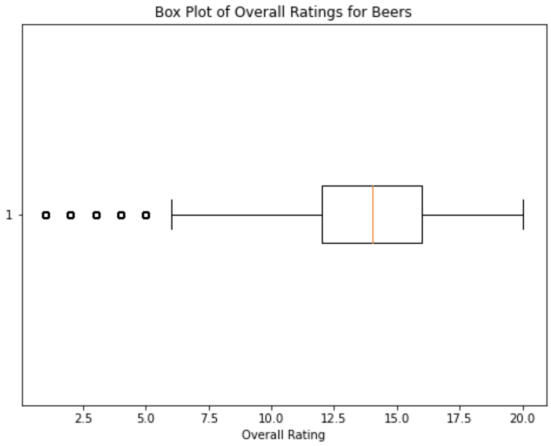
beer/name	Name of the reviewed beer
beer/beerID	Unique ID of the reviewed beer
beer/brewerID	Unique ID of the brewer of the reviewed beer
beer/ABV	Alcohol by volume of the reviewed beer
beer/style	Style of the reviewed beer
review/appearance	Rating of the appearance of the reviewed beer (x/10)
review/aroma	Rating of the aroma of the reviewed beer (x/10)
review/palate	Rating of the palate of the reviewed beer (x/10)
review/taste	Rating of the taste of the reviewed beer (x/10)
review/overall	The overall rating of the reviewed beer(x/20)
review/time	The time stamp when the review is made
review/profileName	Name of the reviewer
review/text	The review text content of the reviewed beer

The dataset has 2924127 number of reviews, in our analysis we will only use 80000 data entries which is fair enough.

There are 3273 different beerIDs and 7005 different reviewers, 20 different overall ratings in our 80000 data entries. The most popular beer in our dataset is Chimay Bleue with beerID: 53.

Distribution of the overall ratings

count	80000.000000
mean	13.661038
std	2.946579
min	1.000000
25%	12.000000
50%	14.000000
75%	16.000000
max	20.000000

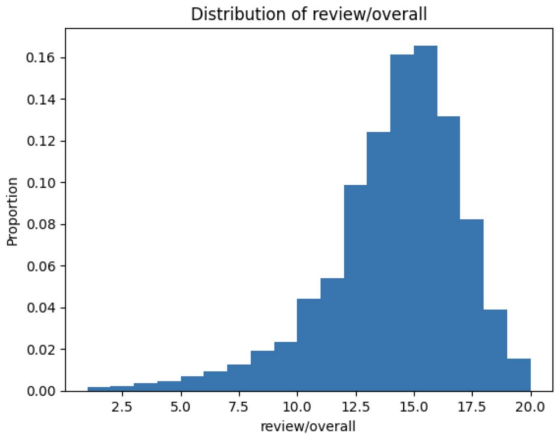


The statistical summary provides key insights into the distribution of beer review ratings. The dataset, with a total count of 80,000 ratings, exhibits a mean rating of approximately 13.66, indicating that, on average, the reviews tend to be slightly above the midpoint of the rating scale. The standard deviation of 2.95 suggests a moderate level of variability in the ratings. The ratings range from a minimum of 1 to a maximum of 20, showcasing the full spectrum of the rating scale.

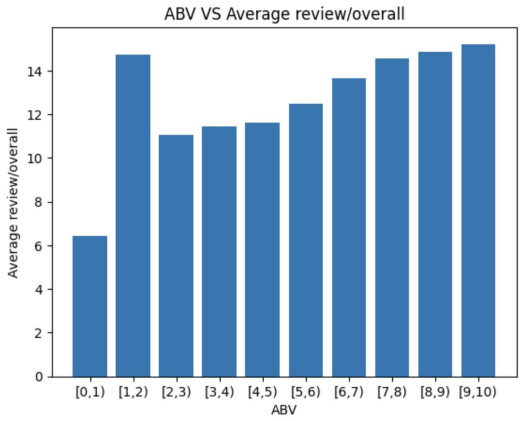
The interquartile range (IQR) provides a measure of the dispersion within the middle 50% of the data. The 25th percentile (Q1) is 12, the median (50th percentile) is 14, and the 75th percentile (Q3) is 16. This distribution indicates that a majority of the ratings fall within the range of 12 to 16.

The presence of outliers on the left side, specifically with ratings of 1, 2, 3, 4, and 5, is noteworthy. These outliers suggest that there are a few instances where the reviews deviate significantly from the norm, reflecting exceptionally low ratings.

The most common overall rating is 16, which occupies 16% of all the ratings.



Here is the barplot of ABV versus average overall ratings. Beers with ABV between 1 and 2 (not inclusive) have the highest overall rating - 14.7. On the other hand, beers with ABV between 0 and 1 (not inclusive) have the lowest overall rating - 6.43.



The interesting finding from our EDA is that the beer Chimay Bleue was reviewed 3056 times in our dataset (the beer with most reviews) and one reviewer called ‘fonefan’ has posted 416 reviews (the reviewer who posted the most reviews).

2. Predictive Task

In the predictive task, we are going to predict the overall rating would be given by specific user item pairs. Our total dataset has 80000 entries, we divide them into training (first 60000 entries), validating (60001 to 70000 entries) and

testing sets (the last 10000 entries). We choose MSE (mean squared error) as our loss function to assess the validity of our prediction:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(R_p - R)^2$$

n = total number of reviews

R_p = predicted overall rating

P = the label overall rating

Note: the overall rating we used will be the numerator of the overall rating for convenience. For example if the overall rating is 15/20, we will use 15 as the overall rating.

Baseline Model:

We chose average overall rating as the baseline model prediction for all user item pairs. We calculate the MSE of our prediction based on the validating and testing set to assess the accuracy.

Jaccard Similarity Model:

In the Jaccard Similarity model, we need features: beer/beerID, review/profileName, and review/overall in the training dataset. We would predict the overall rating a specific reviewer would give to a specific beer based on the historical overall ratings the user gave to similar beers.

To process the data, we will create six default dictionaries: ‘reviews per user’, ‘items per user’, ‘users per item’, ‘rating dict’, ‘user averages’, and ‘item averages’. We traverse the training dataset for each entry we extract the review/profileName as the key for ‘reviews per user’ and ‘items per user’ and add the whole entry as value for ‘reviews per user’ and beer/beerId as value for ‘items per user’. For ‘users per item’, through the same process, we use beer/beerId as the key and review/profileName as the value. For ‘rating dict’, we use the tuple (review/profileName, beer/beerId) as the key and review/overall as the value. For ‘user averages’, we use review/profileName as the key and the average value of all the ratings the reviewer gave. Similarly, ‘item averages’ stores the average value of all the ratings a beer receives for all beers.

Jaccard Similarity Model With Temporal Dynamics:

To improve the Jaccard Similarity Model, we added in temporal dynamics. In specific, ratings are weighted in terms of the percentile of their review length in relation to the maximum review length.

We get the new feature ‘max review length’ by traversing through all the review lengths in the whole dataset. We also formed a new dictionary ‘review percentile per item’ where beer/beerId is the key and the percentile of the review length is the value.

Ridge Regression Model Using Bag-Of-Words :

The overarching objective of implementing the pipeline model is to streamline and simplify the process of leveraging text processing techniques and ridge regression to predict overall ratings within the context of beer reviews. This approach encapsulates data preprocessing, feature extraction, and model training within a unified framework, with a specific emphasis on optimizing hyperparameters to achieve superior predictive performance.

The features utilized in this analysis are derived from two key components of the beer reviews: 'review/text' and 'review/overall.' The textual content in 'review/text' is employed for feature extraction using a bag-of-words approach, enabling the creation of a numerical representation of the text data. Simultaneously, the 'review/overall' rating serves as the target

variable, providing the basis for the regression predictions.

Pipeline Overview:
The pipeline model encapsulates the entire predictive process, providing a systematic and organized methodology for handling beer review data. The key components of the pipeline include:

1. **Data Preprocessing:**
   - **tolower:** Convert text to lowercase for uniformity.
   - **removePunct**: Remove punctuation for cleaner representation.
   - **removeStem:** Optional stemming to reduce words to their root form.

2. **Feature Extraction:**
   - Implementation of the bag-of-words approach to transform textual content into numerical features.
   - Selection of the 'review/text' and 'review/overall' as the primary features for modeling.

3. **Model Training:**
   - Utilization of ridge regression for its regularization properties.
   - Exploration of different hyperparameters, specifically focusing on the 'lambda' parameter, to optimize predictive accuracy.

4. **Hyperparameter Optimization:**
   - Systematic exploration of hyperparameters, such as **'tolower**,' **'removePunct**,' and **'removeStem**,' to understand their impact on model performance.
   - Identification of the best-performing configuration based on validation Mean Squared Error (MSE).

3. Model

3.1 Baseline Model

For the baseline model, we use the global average of overall rating in the training set as the prediction for all the entries in validating and testing sets.

Global average	13.66
MSE of validate set	8.614
MSE of test set	8.546

3.2 Jaccard Similarity Model

We assess the similarity of two items based on their user group similarity.

$$Jaccard(u_i, u_j) = \frac{\left|u_i \cap u_j\right|}{\left|u_i \cup u_j\right|}$$

We first predict the overall rating of a user would give to an item solely by Jaccard similarity and here is the similarity function:

$$r(u,i) = \overline{R_i} + \frac{\sum_{j \in I_u \setminus \{i\}} (R_{u,j} - \overline{R_j}) \cdot Sim(i,j)}{\sum_{j \in I_u \setminus \{i\}} Sim(i,j)}$$

Note: When computing similarities, we return the item's average rating if no similar items exist, or the global average rating if that item hasn't been seen before.

MSE of validate set	4.439
MSE of test set	4.521

3.3 Jaccard Similarity Model With Temporal Dynamics:

$$r(u,i) = \overline{R_i} + \frac{\sum_{j \in I_u \setminus \{i\}} (R_{u,j} - \overline{R_j}) \cdot Sim(i,j) \cdot f(j)}{\sum_{j \in I_u \setminus \{i\}} Sim(i,j) \cdot f(j)}$$

Here, f(j) is the weighting function and we want

it to give higher weight to longer reviews. We had several unsuccessful attempts, such as directly multiplying the percentile of review length. We realized that exponential function is the one that fits our thoughts. As the value of x increases, the value of y increases exponentially, giving more weight to the longer reviews. We ended up with the following function and results.

$$f(j) = 3.5^{review\ length\ of\ item\ j\ /\ max\ review\ length}$$

MSE of validate set	4.436
MSE of test set	4.520

3.3 Sentiment Analysis: Unigram Bag-of-Word Model

The objective of the pipeline model is to leverage text processing techniques and ridge regression to predict overall ratings in the context of beer reviews. The process involves careful data preprocessing, feature extraction, and model training, with a focus on optimizing hyperparameters for enhanced predictive performance.

I. **Data Preprocessing:**
The initial step involves processing the text data extracted from beer reviews. The training, validation, and testing sets were created by randomly shuffling the original dataset. Subsequently, the first 60,000 samples were assigned to the training set (**Xtrain** and **ytrain**), the following 10,000 samples were designated for the validation set (**Xvalid** and **yvalid**), and the final 10,000 samples were allocated to the testing set (**Xtest** and **ytest**). This random shuffling ensures that each set is representative of the overall dataset, capturing a diverse range of beer reviews and ratings for robust model training, validation, and evaluation.
The chosen text processing options include converting the text to lowercase (**tolower**), removing punctuation (**removePunct**), and stemming (**removeStem**). The model utilizes a defaultdict structure to efficiently count word occurrences.

II. **Feature Extraction:**
The pipeline then proceeds to extract features from the processed text data. A vocabulary of words is established based on word counts, and a subset is selected using a specified **dSize** parameter. The pipeline generates a sparse matrix X where each row corresponds to a beer review and each column corresponds to a unique word in the selected vocabulary.

III. **Model Training and Optimization:**
The **ridge regression model** is trained using different regularization parameters (ls). The pipeline systematically explores these parameters to find the optimal configuration. The training process involves splitting the data into training, validation, and test sets. The **Mean Squared Error (MSE)** is used as the evaluation metric.

tolower	remove Punct	remove Stem	Best Lambda	Valid MSE
True	True	True	100	5.683810
True	True	False	100	5.189601
True	False	True	1	5.800911
True	False	False	100	5.477513
False	True	True	100	6.071314
False	True	False	100	5.383918
False	False	True	10	6.173742
False	False	False	100	5.673993

From the comprehensive exploration of different combinations of text processing techniques and regularization parameters, the analysis reveals that the model configuration with tolower=True, removePunct=True, removeStem=False, and a regularization parameter (lambda) of 100 produces the most favorable validation MSE result, indicating superior predictive performance.

IV. **Testing Set Evaluation:**
To further validate the chosen model configuration, it was applied to the testing set. The MSE for the testing set, representing the model's performance on previously unseen data, is reported as **5.3396.**

4 Literature

We use the existing dataset from Stanford Large Network Dataset Collection. The data was used to build models which segmentize different features of the reviews and explain the rating of the beers based on users' preferences. Similar dataset such as BeerAdvocate reviews was also studied before in the same analysis.

4.1 Learning Attitudes and Attributes from Multi-Aspect Reviews

This paper discusses the methods to segmentize the reviews given by users into separate aspects and find out which words describe each aspect with associated sentiment.

The model determines which parts of a review correspond to each rated aspect, which sentences best summarize a review, and how to recover ratings that are missing from reviews. In this paper there were five dataset used and the dataset in our analysis is just one of them. The method of bag of words in this paper is similar to what covered in the lecture, but additionally the paper finds out which part of the bag of words corresponds to a certain aspect.

The model in this paper matches state-of-the-art approaches on existing small-scale datasets. Also the model is able to 'disentangle' content and sentiment words. The model is trained using gradient descent to maximize log likelihood with a regularizer, a common approach of learning nowadays.

The topic of the research paper also involved processing of reviews to predict rating, but they predict ratings of a specific product based on specific aspects which is different from our methodology.

4.2 From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

This paper discusses a recommendation system which involves the changes of taste of users over time. The research team models how tastes change due to the very act of consuming more products. They develop a latent factor recommendation system that explicitly accounts for each user's level of experience.

The researchers use the Ratebeer dataset only which is large and representative enough. They trained this dataset by minimizing mean square error and optimized parameters using L-BFGS, a quasi-Newton method for non-linear optimization, and dynamic programming.

They reached the conclusion that Users' tastes and preferences do evolve over time and the trend, the arrival of new products and so on may influence their rating behavior. The similarity of their research and ours is the same MSE loss function we used. Instead of considering the dynamic temporal effect, we include the length

of reviews as the strength of user's rating into consideration.

5 Results

5.1 Jaccard Similarity Model

The Jaccard Similarity Model has a better performance as compared to the baseline model. The MSE on both valid set and test set decreased by **47% (**from 8.546 to 4.520).
This reflects that predicting the rating that a review would give a beer by the similarity of beers is a reasonable approach. In our model, the more reviewers overlap for two beers, the more similar they are. The implication is that these reviewers have a similar preference for beers, so their ratings have more referential significance in prediction and output more satisfactory results.

5.2 Jaccard Similarity Model With Temporal Dynamics

This model has a better performance than both the baseline model and the Jaccard Similarity model. The MSE on both valid set and test set decreased by 48% as compared to the baseline model. Yet, the decrease in MSEs as compared to the Jaccard Similarity model is mere, especially on the test set, only 0.001.
The significance of the result is that percentile of review length does have a positive impact on the prediction. Our attempt to give more weight to longer reviews is correct. Yet, the feature 'maximum review length' we used was the longest one in all data, which is 6450. In contrast, the maximum review length in valid set and test set are 3744 and 4820, respectively. Clearly, there is a huge difference between the range of review length in different sets. We verified the MSE on the training set is the lowest amongst all - **4.147**.
To improve, since the train, valid, and test sets are randomly chosen, we should explore the distribution of review lengths and choose a more appropriate value, such as 75 percentile of the whole datasets rather than the max value to avoid set bias.

5.3 Sentiment Analysis: Unigram Bag-of-Word Model

Results and Analysis:
The pipeline iterates through various combinations of text processing options (tolower, removePunct, removeStem, best lambda), recording the validation MSE for each configuration in a DataFrame (results_df). This enables a comprehensive comparison of the impact of different preprocessing steps on the model's performance.

tolower=True: Preserving the original casing of the text allows the model to capture subtle nuances associated with uppercase and lowercase expressions in beer reviews. This contributes positively to the model's ability to discern sentiment and context.

removePunct=True: Retaining punctuation aids in maintaining the structural integrity of sentences, ensuring that the model can leverage grammatical information for better comprehension.

removeStem=False: The decision to avoid stemming enhances the model's capability to capture detailed language patterns without simplifying words to their root form. This nuanced approach proves beneficial in the context of beer reviews where specific terms and expressions matter.

Regularization Parameter (lambda=100): The regularization parameter (lambda) of 100 signifies a balanced approach between fitting the model to the training data and preventing overfitting. This choice prevents the model from becoming too specialized to the training set, enabling it to generalize effectively to new, unseen data.

Analysis and Interpretation:
The observed configuration suggests that preserving the original casing of the text (**tolower=True**) and retaining punctuation (**removePunct=True**) contribute positively to the model's predictive accuracy. Interestingly, the decision to not perform stemming (**removeStem=False**) enhances the model's capability to capture nuanced language patterns within the beer reviews.

The identified regularization parameter (**lambda**) of 100 indicates a balance between fitting the model to the training data and preventing overfitting. This parameter choice likely aids in generalizing the model to new data, as evidenced by the favorable MSE results on both the validation and testing sets.

Comparison with Baseline Model:
To contextualize the model's performance, a baseline model predicting the global average was established, yielding an MSE of **8.546** on the testing set. The optimized model, incorporating text processing and regularization, significantly outperformed the baseline, with testing MSE of **5.3396,** demonstrating the efficacy of thoughtful parameter tuning and feature engineering.

The success of the proposed model compared to other configurations can be attributed to the careful balance struck between feature preservation, regularization, and model generalization. Other configurations may have failed to achieve optimal results due to a lack of consideration for the nuanced language patterns in beer reviews, overfitting to training data, or inadequate regularization.

5.4 Conclusion
In conclusion, this analysis underscores the importance of systematically exploring model configurations and fine-tuning parameters. The identified optimal configuration provides valuable insights into the interplay between **similarity scores, text processing choices and model performance**, showcasing the potential for accurate prediction of beer review ratings in this specific context. The substantial improvement over the baseline model highlights the success of the proposed approach in enhancing predictive accuracy.

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

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