Restaurant Review Ratings Prediction

Mohit Matta

DSC680

Bellevue University

August 7,2021

Email: Mmatta@my365.bellevue.edu

Instructor: Prof. Catherine Williams

Abstract

In recent years, abundant online review data has become available for analysis through different websites connecting customers and businesses. Many websites allow users to post online reviews and share their experiences about the food restaurants. The wealth of information hidden behind the online review data can help the restaurant business and startups to set up defining goals to improve the revenue. A large amount of plain text data that customers have posted can help on sustainability and companies can come up with a sustainable marketing strategy to understand and meet diverse customer demands and maintain competitive advantages. This case study can help restaurants to determine how well they are performing as well as improve customer satisfaction.

Introduction

In recent years, eating out at restaurants/takeaways/deliveries has become increasingly popular in the United States. Except during the Covid pandemic year, sales of restaurants have increased for over 20 years. A leading business school conducted research to find out the relationship between review ratings and profit and found a direct proportional relationship between them.According to National Restaurant Association, table service restaurants make approximately $300 Billions in sales yearly.In the era of widespread internet access, people like to consult on to-go websites like Yelp and Google before deciding to go where to eat. Online reviews have a great influence on people to decide the places as well as how a restaurant business performs.

There are some common habits of customers like, to leave a positive review after they have a good experience(about 43%) and stay away from restaurants that don't maintain cleanliness(73%). Few years back , people used to rely on critic reviews but now they are preferring peer reviews(online). Timing is an important factor for customers as they try to read the reviews that are most recent.

Users provide feedback for their experience through the review comments. Although reviews are unstructured texts they provide very useful information that can reveal great trends and patterns. Users justify the ratings and express the sentiments through review comments. Different customers correlate with different attributes of restaurants such as ambiance, food quality, service, cleanliness, etc in review comments. Many of the restaurants fail within few years of their opening while some survive for a few more. The restaurants should also target to help in reducing the unemployment rate and grow the local economy along with increasing the revenue. A strong self-sustained business model can help businesses achieve that.

Users also try to convey hidden emotions and linguist styles through textual review comments that are often ignored since the performance is always measured in numerical factors like ratings. Sometimes the relationship between ratings and reviews is not obvious and users just look at the rating and decide to go ahead with the restaurants.

Review rating prediction has great importance because the users can decide whether to look at the textual reviews or ignore them. The review comment given by two different users can lead to different ratings because the reason for the online reviews for the users will be different. The online data related to reviews is increasing at a very high rate. Yelp recommendation website data has increased with an annual rate of 9% from 2009 to 2020 with a total of around 224 million reviews. This project will not only inform what factors matter most for the customers but also tells what features the new restaurant should focus on in order to successfully run it.

Methods

Data

The dataset comes from 2013 Yelp Recsys dataset challenge.It is a small subset of Yelp data containing data for Phoenix, Arizona location.The raw compressed dataset contains five files,: business, review, user, check-in and tip information representing each object type.Each file is in JSON format.The first dataset is business dataset that was scraped from yelp.com has 11537 entries with different business attributes like locations , ratings, types and categories etc.

```
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   business_id   11537 non-null  object
 1   full_address  11537 non-null  object
 2   open          11537 non-null  bool
 3   categories    11537 non-null  object
 4   city          11537 non-null  object
```

```
 5   review_count   11537 non-null  int64
 6   name          11537 non-null  object
 7   neighborhoods  11537 non-null  object
 8   longitude      11537 non-null  float64
 9   state         11537 non-null  object
 10  stars         11537 non-null  float64
 11  latitude       11537 non-null  float64
 12  type          11537 non-null  object
```

The second and the most important dataset is the review dataset, it has observations for more than 229K text reviews and contains below list of columns, it has the review text that were entered by users.

```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   votes        229907 non-null  object
 1   user_id      229907 non-null  object
 2   review_id    229907 non-null  object
 3   stars        229907 non-null  int64
 4   date         229907 non-null  datetime64[ns]
 5   text         229907 non-null  object
 6   type         229907 non-null  object
 7   business_id  229907 non-null  object
```

The third dataset is the users datasets that provide details of the users like number of reviews done by individual user ids , their names etc.

```
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   votes          43873 non-null  object
 1   user_id        43873 non-null  object
 2   name           43873 non-null  object
 3   average_stars  43873 non-null  float64
 4   review_count   43873 non-null  int64
 5   type           43873 non-null  object
dtypes: float64(1), int64(1), object(4)
```

The Yelp dataset contains different categories of business like auto, rentals, hotels, restaurants etc and the reviews. The reviews entered by customers may be different depending on categories so its important to filter the categories and apply sentimental analysis just on food/restaurant categories.The restaurants make around 60% of all businesses and around 80% of all the total reviews are related to restaurants.In this study , we will focus only on restaurant datasets for ratings prediction.

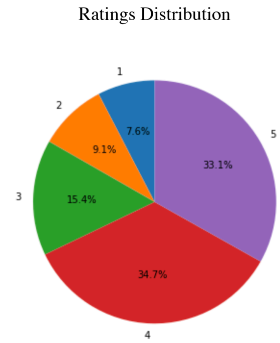


Figure 1

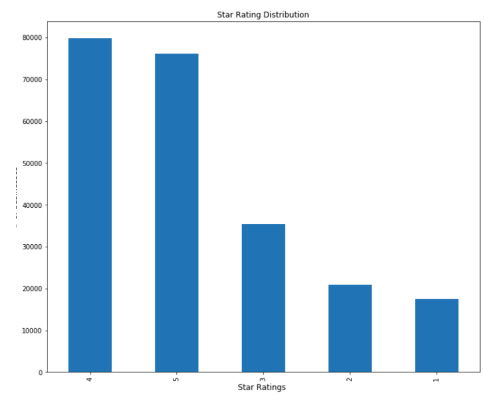


Figure 2: Ratings count

Preprocessing

Data was processed, cleaned and transformed using python scripts. After separating the restaurant businesses from rest, when data was explored it was found that reviews text field is in free form text format.Customers use different cases , grammatically incorrect words , punctuation marks

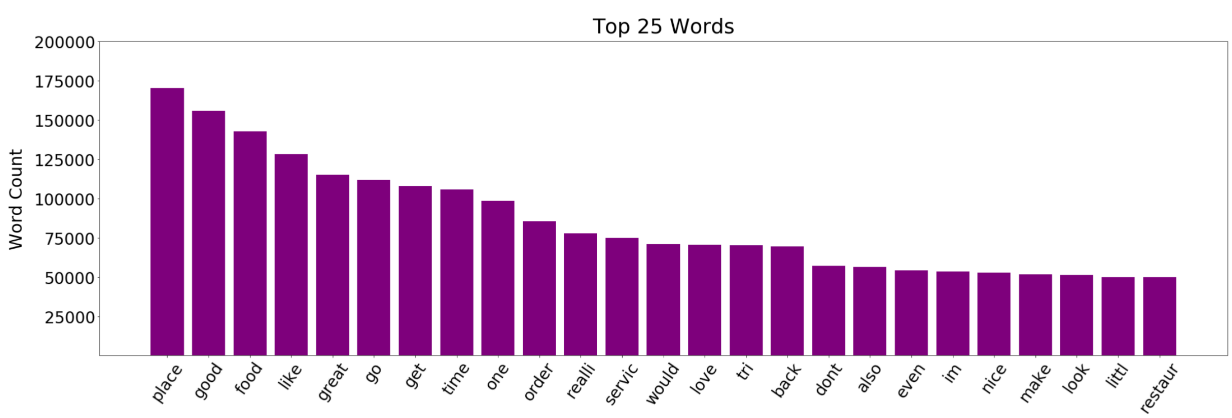


Figure 3: Top 25 frequent words

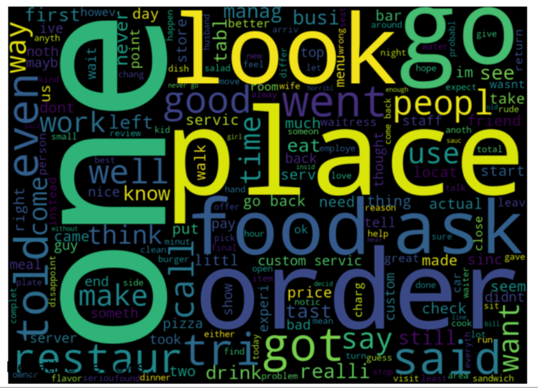


Figure 4 : Word Cloud positive reviews

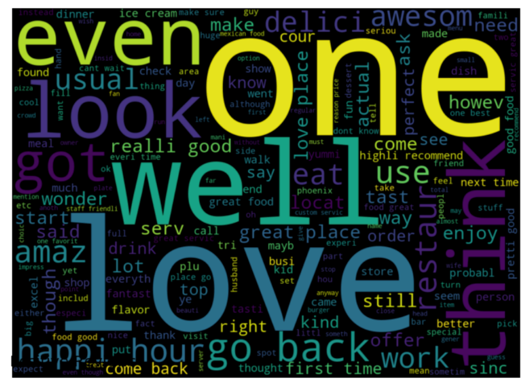


Figure 5 : Word Cloud negative reviews

and slang words in the review. They may use different words for expressing their likes or dislikes.Stop words are used very frequently by users which are not needed for machine learning models for analysis.So standard natural language processing techniques were used to convert all cases to lowercase , to remove punctuation, to remove stop words and apply porter stemmer to convert the sentences to stemmed sentences.

Feature Extraction

We used different feature extraction methods based on semantic analysis to extract useful features from the review corpus and built a feature vector for each review. Both unigram and bigram methods were used for the study. We used count vectorizer objects, word level TF-IDF(Term Frequency-Inverse Document Frequency) objects , n-gram level vectorizer objects and Character level vectorizer objects for vector representation of review text. Count vectorizer and TF-IDf vectorizer create a dictionary of words from review dataset and consider each unique word as a feature.Once the frequency of each word is obtained TFIDF weighing technique creates the final feature matrix.It adds high weights to words that are rarely occurring in the text and less weight to words that are frequently occurring. Bigram models gives better results compared to count vectorizers because it considers the relationship between two words. So when TFIDF weighing technique is applied on the text , more importance is given to combination of words like ‘delicious food’ and ‘coming back’ compared to single and more common occurrences.

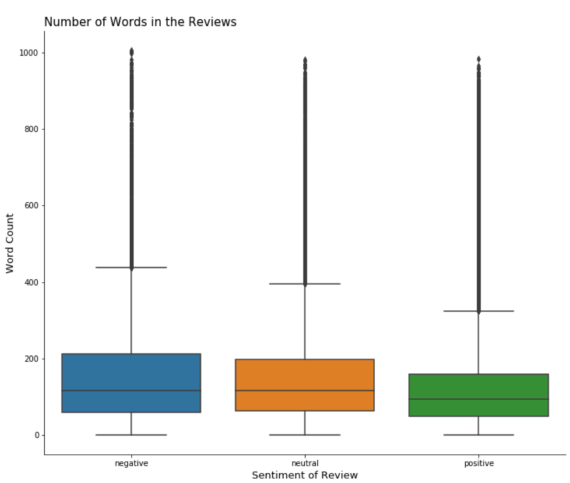


Figure 6: Number of words in positive , negative and neutral rating reviews

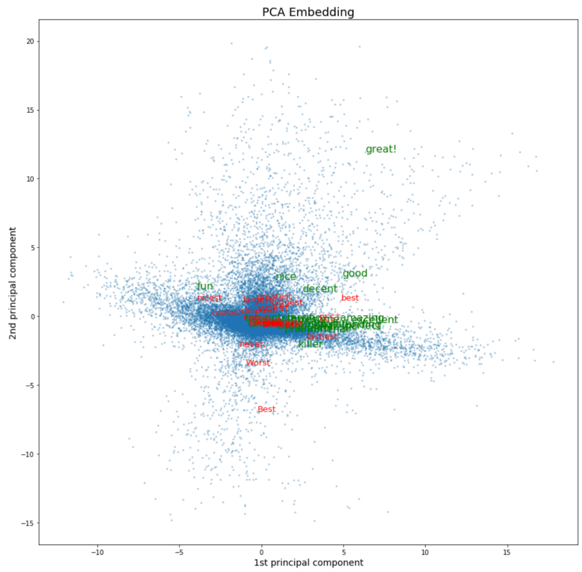


Figure 7: Cosine similarity of words in review

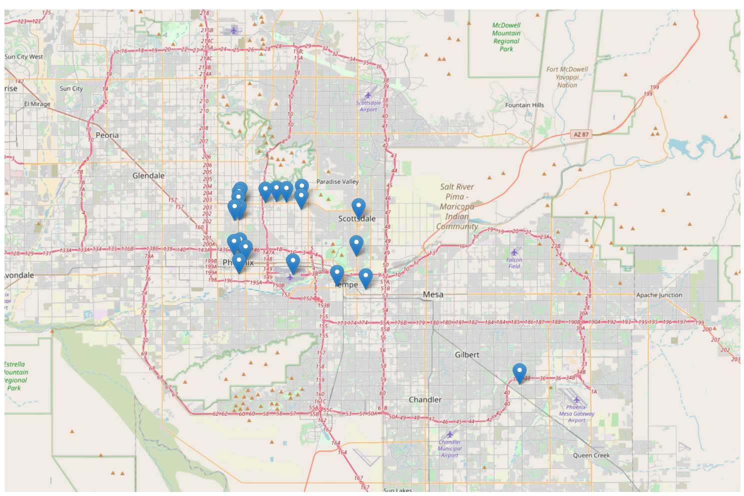


Figure 8: Live map for top 20 restaurants with top ratings

Supervised Learning

To train our prediction models, we use four supervised learning algorithms.

Logistic Regression

The logistic regression model predicts the conditional probability using the feature vectors and decides the class labels. The results that give the highest probability are mapped to the output as the final rating for the review.

Naive Bayes Classification

A Naive Bayes classification works on the basis of the Naive Bayes theorem of independent assumption between the features. It also works on conditional probability and constructs a classifier based on the probability model.

Linear Support Vector Classification (SVC)

In the SVM technique, we select the best hyperplane (decision boundary) from the data points and separates the labels. We are using linear SVM techniques where data is linearly separable. It is also known as a discriminative classifier. It also uses the Kernel trick for non-linear classification.

Gradient Boost Classification

Gradient Boost classification uses ensemble techniques which works on the principle that a collection of predictors(weak or strong) works better than individual predictors. In the Boost technique, the weak learners are converted into strong learners. The model tries to minimize the overall error of the strong learners.

Neural Network Classification

A neural network has the weight, score function, and loss function as the main components. It learns in a feedback loop. It adjusts the weight based on results from score function and loss function. The architecture of a neural network has an input layer, hidden layer, and output layers.

The neural network calculates, tests, calculates again, tests again, and repeats until the optimum and accurate solution is reached.

Word2Vec Gradient Boost Classification

Word2vec works on the idea of distributional semantics which means that we can understand the meaning of a word by understanding the context that a word keeps. It was developed to overcome the shortcomings of one-hot encoding. With the increase in word's vector dimensions, the relationship between words can be explained in detail.For example, if we have words like apple, mango, and element word2vec model will create features like is_fruit, is_eatable and is_animal. Fruits apple and mango will have the same context for is_fruit and is_animal but less contextual similarity for is_animal.

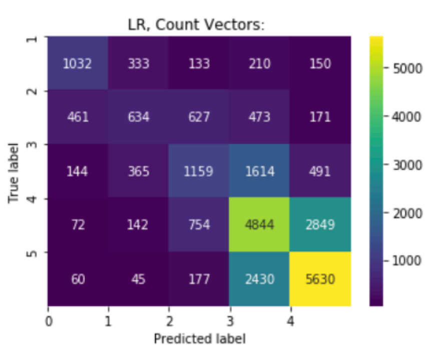
Accuracy, precision, and recall statistics were produced to determine the overall performance of each model.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

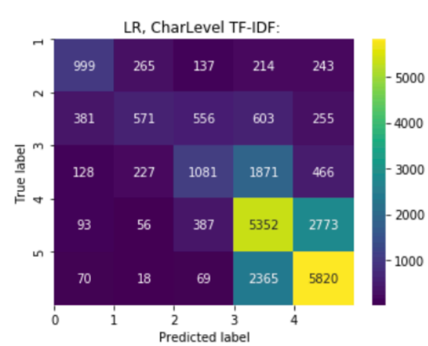
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

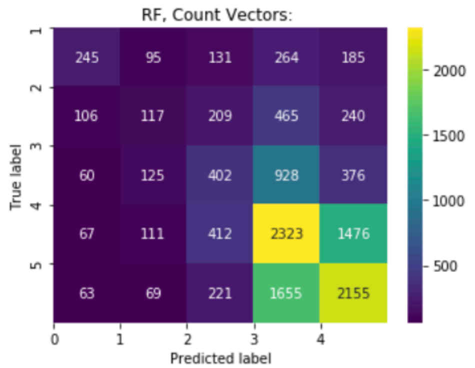
Where TP = True positive; FP = False positive; TN = True negative; FN = False negative.



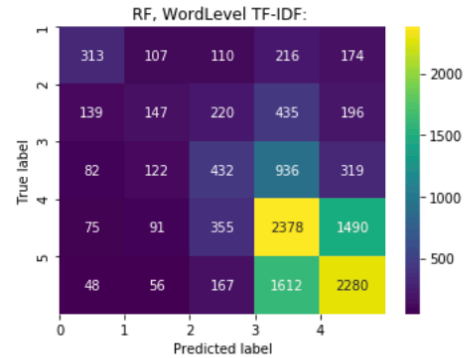
Accuracy of Liner Regression Count Vectors Model is: 53.20%



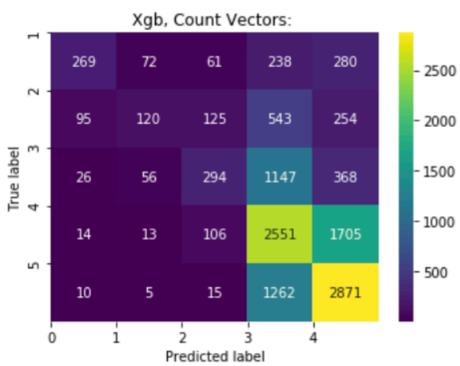
Accuracy of Liner Regression Char Level TFIDF Model is: 55.29%



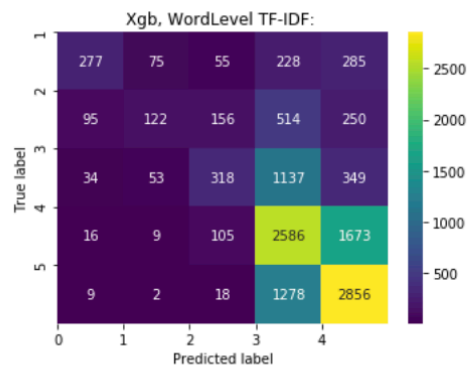
Accuracy of Random Forest Count Vector Model is: 41.94%



Accuracy of Random Forest word level Model is: 44.40%



Accuracy of Xgradient boost count vector Model is: 48.84%



Accuracy of Xgradient boost TFIDF vector Model is: 49.27%

Technology

The model was coded in Python, using Scikit-Learn's methods for the logistic regression, random forest, and extreme gradient boosting.We used Keras deep learning model to classify and predict the ratings of restaurants. The data cleaning , preprocessing and exploratory data analysis has been performed in python as well.We have used folium API to pull geo location map data using python. Matplotlib and seaborn were used to plot some of the graphs in python.

Results

As seen in ‘Model Accuracy’ table logistic regression model with word level TF-IDF classifier achieved the highest Accuracy (56.50%) and Precision (56%) ratings and performed well in Recall (55%) across Python. Char level Logistic regression classifier performed similarly with Accuracy (55.29%) and Precision (55%). Rest all models performed poorly with less than 50% accuracy. Ngrams neural network classifier performed poorly compared to all other models.Principal Component Analysis was not found to improve any model.

Model Accuracy

Model Name	Accuracy
• Count Vector Naive Bayes classifier	52.26
• Word level-TFIDF Naive Bayes classifier	50.10
• N-grams Naive Bayes classifier	50.50
• Char level Naive Bayes Classifier	48.27
• Count Vector Logistic Regression classifier	53.20
• Word level-TFIDF Logistic Regression classifier	56.50
• N-grams Logistic Regression classifier	51.69
• Char level Logistic Regression Classifier	55.29
• N-grams SVM vector classifier	35.11
• Count vector Random forest Classifier	41.94
• Word level-TFIDF Random forest classifier	44.40
• Count vector Gradient Boost Classifier	48.84
• Word level-TFIDF Gradient Boost Classifier	49.27
• Char level Gradient Boost Classifier	49.25
• Count vector Neural Network Classifier	48.30
• Ngrams Neural Network Classifier	32.92
• Word2vec Gradient Boost Classifier	38.41

Discussion

The goal of developing this model was to see if it was possible to develop an application that can predict likelihood of a good or bad restaurant through review ratings. This initial model is not well promising and accuracy of models is in the range of 50-56%.The methods need to be improved to increase the accuracy of models. Considering the fact that people are spending more time to eat outside food and want to share their experience through review ratings, restaurants can use this information to improve their food quality and use the feedback to stay competitive in the market. This research work can act as a tool for restaurants to act swiftly to implement the process improvement steps for operational excellence based on the user feedback.

Acknowledgement

Thank you to the peers and professors in DSC680. This research was made possible with the resources provided by Bellevue University.

Compendium

1) What are the different aspects that determine restaurant ratings ? Is it all dining aspects (food/taste/value) or emotional aspect as well?
2) What are the factors most responsible for positive customer sentiments(Good taste/food quality/value for money/ambience)?
3) What are the factors most responsible for negative customer sentiments(Bad taste/poor quality/bad ambience/service quality)?
4) Does choosing a good high traffic location with surrounding well developed transportation infrastructure helps the restaurants attracting more customers?
5) Can restaurants filter large amount of online reviews and can they use it as a guideline to improve the service quality?
6) Can sustainable aspect of restaurants(that help conserve the environments) help in building the positive customer experience?
7) Can the machine learning algorithms catch fake reviews using the dataset ? Can they help in identifying the genuineness of reviews?
8) What is the variety that exists among different cuisines like Italian restaurants - Drinks, French restaurants - drinks and desserts?
9) Has the restaurant been following all the safety and health code protocols? Are they likely to be inspected by food and health safety departments?
10) Would the relationship of restaurants with good or bad ratings change if we choose different users?

References

[1] M. Anderson and J. Magruder. Learning from the crowd. 2011.
[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, (2):15–21, 2013.
[3] M. Fan and M. Khademi. Predicting a business star in yelp from its reviews text alone. arXiv:1401.0864, 2014.
[4] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 426–434. ACM, 2008.
[5] A. Mueller. A wordcloud in python. Andy's Computer Vision and Machine Learning Blog, June 2012.
[6] A. Ortony, G. L. Clore, and A. Collins. The cognitive structure of emotions. Cambridge university press, 1990.
[7] R. A. Stevenson, J. A. Mikels, and T. W. James. Characterization of the affective norms for english words by discrete emotional categories. Behavior Research Methods, 39(4):1020–1024, 2007.
[8] M. Woolf. The statistical difference between 1-star and 5-star reviews on yelp. http://minimaxir.com/2014/09/one-star-five-stars/, September 2014.
[9] Narendra Gupta, Giuseppe Di Fabbrizio, and Patrick Haffner. 2010. Capturing the stars: predicting ratings for service and product reviews. In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, pages 36–43. Association for Computational Linguistics.
[10]López Barbosa, R.R., Sánchez-Alonso, S. and Sicilia-Urban, M.A. (2015), "Evaluating hotels rating prediction based on sentiment analysis services", Aslib Journal of Information Management, Vol. 67 No. 4, pp. 392-407. https://doi.org/10.1108/AJIM-01-2015-0004