Predictions of Yelp Ratings using Regressions

University of California, San Diego

0. INTRODUCTION

As more restaurants and businesses open up with better qualities, it is very difficult to choose your perfect meal. Within this fast-paced society, it is also impossible to visit every place yourself. In order to address this problem, Yelp has been providing a platform for people to rate their experiences in different businesses for others to utilize. The overall rating, recommended reviews, and photos of the business can provide insight into how the experience of the specific businesses may be. However, with many reviews to read and features to consider, Yelp may not be able to reduce your problem. Can there be a prediction of the rating for that business that incorporates different information about the user and the review? In this paper, we will attempt to predict the rating of one's experience in a business, considering multiple factors.

1. DATASET

To address the problem above, we have chosen the Yelp open dataset, which stores approximately seven million reviews from numerous businesses and users spread throughout the United States. The advantage of the Yelp Dataset is that, along with the review dataset, there are datasets that list additional features of businesses and users themselves registered on Yelp. To condense down the dataset, we have randomly selected 100,000 reviews of businesses located in the city of Philadelphia, which consists of 14573 unique businesses and 59785 unique users. This would provide a good amount of data that can be used for this study.

The three datasets provide diverse information, from *review_id* to *friends* of the specific user, totaling 43 features that can be studied. However, with repeated information amongst three datasets and certain multi-layered features, working with all features may create too much data to work with. To decide which features to specifically work with, we conducted data analysis on these features to explore if there were interesting correlations between the feature and the ratings.

1.1 Explanatory Analysis

To understand the data we are working with, we start by understanding the distribution of the ratings and later see how it relates to other features. First, looking at Figure 1, the distribution of the review ratings is a left-skewed distribution with a higher median of 4.0 than the mean of 3.78868. The ratings of 4 and 5 are significant compared to others. Overall, users seem to enjoy their experience with the businesses registered on Yelp. In addition, we can see that there are more 1 star reviewers than the 2 star reviewers. The reason may be that 1 star reviewers feel more strongly about their experience, writing more reviews, or that if the restaurant is bad, people are more likely to review a 1 star rather than a 2 star. Either way, we will take into account that more 1 ratings will lower the mean.

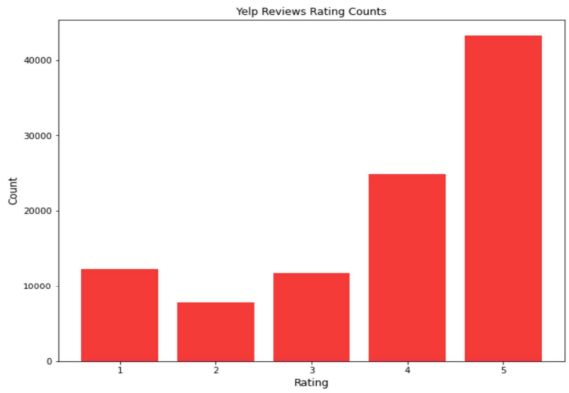

Figure 1. Rating Distribution

Since we are studying the city of Philadelphia, we can also examine the proportion of review ratings spread throughout the city. The business dataset provides the longitude and latitude of the businesses. Using this, we can graph all the review ratings as a scatterplot, resulting in Figure 2. While different review ratings are distributed throughout the city, we can see that the northwest and center of Philadelphia consist of lots of businesses with positive ratings, but southwest Philadelphia is filled with more red through yellow points, signifying lower ratings in general. This shows that the location of the business may be related to the overall ratings.

Since there exists a few outliers among the businesses in terms of geolocation, we may remove these outliers that lie outside the boundaries of the city of Philadelphia. Some data cleaning may help if using geolocation features in the study because most of the latitude and longitude in the review dataset have a very slight difference, meaning many outliers can affect the predictions. However, only a total of 19 samples lie outside of the main part of Philadelphia, which may not have a significant impact on the overall predictions.

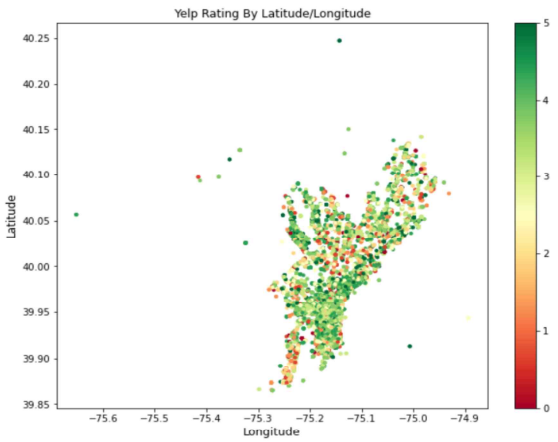

Figure 2. Ratings By Location

In addition to the geographical features, we can also see how the categories of the business may factor into the ratings. The businesses in the dataset consist of multiple labels, as they can be known for more than one specific aspect. In total, 1026 unique labels were used to categorize the Yelp businesses, with the top five labels being *Restaurants, Food, Shopping, Nightlife, Beauty&Spas.* Figure 3. shows the businesses with specific categories and the average ratings of the reviews with the respective category labeled business. To see the difference amongst multiple categories, the red bars represent the top 10 categories that label the businesses (at least 929 businesses labeled), and the black bars represent the lesser selected categories (at most 54 businesses labeled). Looking at the bar graph, it can be seen that well known categories have an average rating that is closer to the mean of 3.79, but the lesser chosen categories have largely varying average ratings. This phenomenom may happen due to lesser chosen categories having a lower number of reviews. Another reason may be that these businesses have a specific audience. For example, the physical therapy business will usually be visited by people who will absolutely need some therapy. After getting the therapy, they may feel better and leave a higher rating. As such, the special categories should be specially considered in the study.

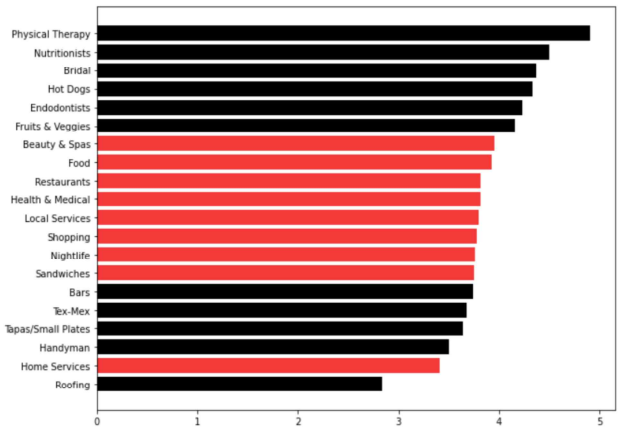

Figure 3. Ratings of Reviews with Specific Categories

Lastly, reviews that are given with the ratings should be considered. Specific words like *awesome* or *dirty* tend to have a high correlation with the experiences of the users, and can be a great feature for our study. The word cloud in Figure 4. shows the words with higher frequencies bigger compared to less frequent words. Words that are prominent, like *good*, *delicious*, and *great,* represent positivity within the reviews, while only the smaller words, like *little*, *need,* and *tried,* represent negativity. Studying the word cloud, we can acknowledge that there exists a good correlation between the words and the mean of 3.789 since most words seem to have positive connotation and the mean is on the higher side of the distribution. As such, the words in the review can be a great use to our study.

With the studies of different features that can be used in the dataset, we see that certain features may benefit the predictive tasks while others may disrupt the models. To make an accurate model, we will choose features that will likely have a good corealation with the ratings, such as the business average rating or number of reviews of the business.

Figure 4. Wordcloud of Reviews (In Yelp Logo)

2. PREDICTIVE TASK

With the conducted data analysis, we learned that different features, such as the geolocation, category of the business, and the words within the review, are related to the overall review ratings. Taking advantage of this knowledge, we can attempt to predict future users' overall ratings of a certain type of business using these features in the dataset. This predictive task can be helpful for someone who may want to visit a certain business, as it will provide them with an estimate of the ratings that they may experience.

From the dataset, we have chosen 14 different features to represent the business and the user in order to predict an accurate rating of the user's experience. These features include numerical values that do not need preprocessing but also include specially extracted features. The first 5 features are related to the business: *longitude*, *latitude*, *rating of business*, *review count of business*, and one hot encoding of *categories*. Then, the next 6 features include user-related features: *average ratings given*, *review count of user*, *"useful" count*, *number of friends*, *number of fans*, and *number of "elite" years*. The review-related features include *"useful" count*, *length of review*, and one hot encoding of *top 500 words in review*. In total, the feature vector has 613 dimensions, including the offset value.

2.1 Feature Selection

As studied in the data analysis portion, *longitude* and *latitude* were chosen as a feature since the scatterplot has shown a relationship between the location and the ratings. Additionally, the *rating of business* and the *review count of business* were chosen due to their good numerical representation of the business' overall rating and its popularity amongst the population. Lastly, the one hot encoding of *categories* may be a key factor in describing the business; as shown in Figure 3, businesses seemed to have a higher or lower rating depending on their categories. To know which category the business was in, we have incorporated the top 100 significant categories as one hot encoding for our feature.

Moving onto the user-related features, all six features did not require much pre-processing of the dataset. Features *average ratings given* and *review count of user* describe how often and the average rating that the user may experience. The remaining features, *"useful" count*, *number of friends*, *number of fans*, and *number of "elite" years*, then describe how others may relate to

this user's rating, making the ratings that this user gives more or less strength.

Lastly, we have previously seen that certain words are more prominent within the reviews and also have a mean rating of 3.79. To incorporate this text feature, we have made an one hot encoding of whether or not the top 500 prominent words existed in the review. Since specific words that are included in the review can justify how the user felt about the business, it would be a good indicator of the overall rating. Then, *"useful" count* and *length of review* features were included to also incorporate the strength of the review itself.

2.2 Model Validation

With this selection of features, we can create strong predictors with various models. However, before diving into modeling, we will set the standards for evaluation and the baselines that our models can be compared to.

First, we will divide our dataset into three subsets: training set, validation set, and test set of sizes 80,000, 10,000, and 10,000, respectively. By doing so, we can train our models with the training set and tune our hyperparameters with the validation set. Then, after determining the hyperparameters, the model can finally be tested with the test set for the best possible model. This type of resampling, called cross validation, allows us to check the performance of the model on data that is not seen by the model. An additional benefit from this approach is that the model will avoid overfitting the data since we can observe its performance on the validation set and the test set, which again is not used when training the model.

In terms of the accuracy of the model, since we are predicting numerical outputs with different features, we will use the mean squared error (MSE) when determining the strength of the models. We chose MSE as our method to evaluate since MSE can ensure that bigger errors will be heavily punished while smaller errors will be punished less. This would benefit our study because we do not want ratings that might heavily differ from what the user may rate, but smaller differences in the ratings will not matter as much.

Then, to evaluate the strength of our models against other models, we will have two different baselines in our study. Baseline 1 will always predict the mean rating of the training dataset, and baseline 2 will always predict the median rating of the training dataset. The reason for having two different baselines is that the distribution of the ratings is neither uniform nor normal, where a big proportion of ratings are above a 4 star. The resulting MSE for the two baselines is given in Table 1. Both the validation set and test set MSE are above 1.9, but baseline 1, which predicts the mean, seems to be a stronger baseline compared to baseline 2. We will use these baselines to check our models' strength.

	Valiation Set MSE	Test Set MSE
Baseline 1	1.9389	1.9085
Baseline 2	1.9925	1.9481

Table 1. Mean Squared Error of Baseline Models

3. MODEL

A good approach for prediciting numerical outputs using multiple features would be implementing regression models. In our study, we will implement three different regression models: linear, ridge, and random forest since each model have its advantages.

First, we test a linear regression model since it is a good place to start most types of predictive tasks with easy implementation and fast training time. Although it may be a difficult model to optimize with its limitations, since we expect a linear relationship between most features and ratings, this model would be a fair choice.

Second, we implement the ridge regression because the independent features in the feature vector are highly correlated with each other. For example, more *"useful" count* of the user would likely signify that the user's *review count* would also be high. Ridge regression model can account for these types of data and also can protect itself from overfitting, which would be helpful.

Lastly, we implement random forest regression that can easily deal with high dimensional features. Since our features total up to 613 dimensions, an implementation of the random forest model can accurately handle all the features while producing a reasonable prediction.

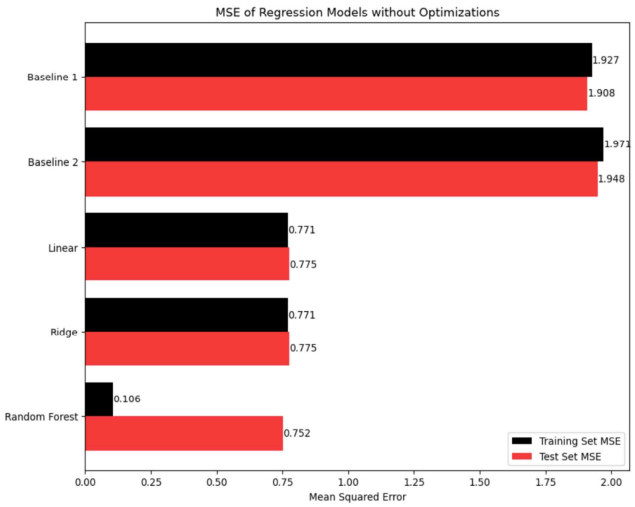

Figure 5. MSE of Regression Models

Even before optimizations, all three models have outperformed the baseline by a big margin in terms of MSE. Both linear and ridge regression models produced similar MSEs around 0.775, while random forest regression produced a MSE of 0.752. Although random forest has shown a slight advantage in terms of MSE, the training set MSE for random forest regression is very low at 0.106, meaning it may have overfitted the training data. Since we do not have any parameterization for this model, the random forest model currently does not have a limit for the maximum depth of the tree, continuously expanding until its leaves are all a size of 2 or less. To avoid overfitting, we need to find the right amount of depth. Along with this problem, we will also optimize the ridge regression model.

3.1 Model Optimization

For model optimizations, we first work with the word feature. There are a total of 97414 words even after removing punctuation, so more prominent words can be included in the feature. To accomodate more words, the one hot encoding feature of *top 500 words in review* was scaled up to 1000 to include more words that are prominent. However, if the feature vector is too large, the model training time would suffer while also causing the model to overfit. Thus, we will only include 1000 prominent words.

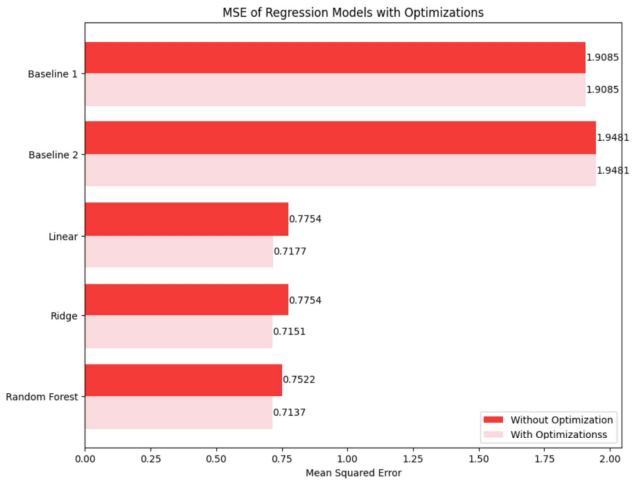

Figure 6. MSE with Optimizations

As the linear regression model is a simple one, it does not have any more optimizations. For the ridge regression model, we have tested multiple alpha values (1, 10, 100, 200, 300, 400, and 500), and the alpha value of 300 resulted in the lowest validation set MSE. We have implemented this alpha value as the final value for the model. Lastly, we tested various numbers of trees and numbers of depths for the random forest regression model to avoid overfitting while producing a good validation set MSE. The final hyperparameters for random forest regression were 150 trees with a maximum depth limit of 30. With these optimizations in place, we have achieved a lower MSE compared to the non-optimized version of the models, as shown on Figure 6. Once again, the random forest regression model had the lowest MSE of 0.7137. This is a very strong predictor compared to the original baselines and will likely predict the user's rating of the business with good accuracy.

3.2 Failed Optimization

As increasing the number of words in the one hot encoding helped in increasing the accuracy of the models, we have also attempted to incorporate word features other than the bag of words feature. The first substitute attempted was the N-grams word model. We attempted to use the number of n-grams as the one hot encoding vector since we assumed that the group of words would be a better predictor than a single word. However, the MSE of the models increased with this change. Not only the N-grams model, we also have attempted to use the TF-IDF words model but also have ended with bad results. Rather than lowering the MSE, it increased the MSE, harming the strengths of the models. Ultimately, the bag of words model was kept in as a feature in our feature vector.

4. LITERATURE

4.1 Description of Dataset

For the study, we have used the open dataset that Yelp has published online [1]. This dataset is a subset of what Yelp owns, which contains about seven million reviews up until the recent reviews of 2022. Additionally, not only does it also offer features about the reviews, it offers additional information on the users and businesses themselves, providing multiple useful features.

We have used this dataset to predict the ratings of the user and business pairs given specific features in the dataset, but this dataset can be used in diverse ways. The Yelp website mentions, "use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps." With diverse features and a huge number of samples, this dataset can be used for many machine learning or database related tasks.

4.2 Study on Similar Dataset

Google Local Reviews (2021), created at UCSD by our professor, McAuley along with other professors is an example of similar dataset [2]. This dataset also contains similar features such as average rating, review text, and geolocation. However, the study "UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining" focuses on text analysis and its uses in the evaluation of topics compared to our study of rating predicitons. It uses a method called contrastive learning, which differs from regression predictors. With this model, this study was able to conclude with a new model that is better able to do phrase representations.

On the other hand, studies such as "Yelp Dataset Analysis using Scalable Big Data" by Alam et al., or "An Empirical Study Identifying Bias in Yelp Dataset" by Choi use the Yelp Dataset to analyze features and find how the features affect what they are studying [3, 4]. Specifically looking at Choi's study, she also focuses on the text portion of the review, finding bias and discriminatory behaviors within user reviews [4]. However, she uses linear regression models to find correlations and concludes that users tend to be more bias against businesses with specific categories, like African or Western African.

It can be seen through these studies that similar or different models can be used for the same text analysis predictors. Both the regression models that we have learned in class can still make good predictors, while more complex models like the constrastive learning framework can also give useful understandings about the dataset. Overall, the existing works seem to have similar findings in the sense that all the studies have found meaningful relationships between the words in the review text and the feature that is being studied.

5. CONCLUSION

5.1 Results

In this paper, we have trained three different models to predict the true ratings of the reviews given the data of the user, business, and review. The results of all regression models have outperformed the baseline models of predicting the mean or the median, and out of the three, the random forest regression has produced the predictions with MSE of 0.7137, given in Figure 6. We believe that the random forest regression outperformed the other models due to its ability to deal with high-dimensional feature vectors, while not overfitting as much. Also, as this model can provide better stability in the prediction, it was likely to do better compared to other models. However, as all MSE were extremely close, all models successfully predicted the ratings of the review.

5.2 Understanding Features

Looking at how the features performed in Figure 7, it can be seen that having certain words like *not*, *great*, *delicious*, and *but* were good indicators for the regression model to predict the ratings. The best feature was, however, the user's average rating on their previous reviews. From this, we can assume that a single user will likely give similar ratings throughout their reviews on Yelp. Additionally, the business rating came in second when predicting the overall rating, meaning the overall experiences of various users' will likely be the experiences that others will have.

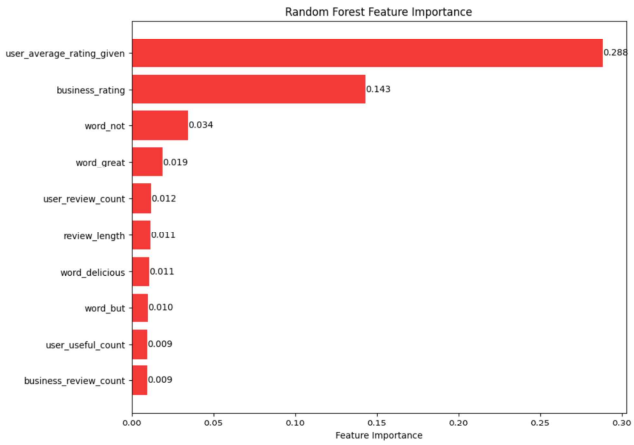

Figure 7. Feature Performance

On the other hand, the top ten least important features for the random forest regressor were all different types of categories for the businesses. While the data analysis showed that different categories had differing average ratings, the existence of certain categories did not seem to matter when predicting the overall rating. With these results, we can better understand what features were important when predicting the ratings of different businesses on Yelp. Random forest regression was able to take advantage of these features and predict the test set rating by MSE of 0.7137. In conclusion, given certain types of features in the reviews, we can predict how the user will rate the business in an accurate manner. In further studies, this study can be expanded upon with more complex models that may incorporate more features for a better overall prediction.

6. REFERENCES

[1] Yelp, "Yelp Dataset." (Feb. 16th, 2021). Distributed byYelp. https://www.yelp.com/dataset/ (accessed Dec. 12th, 2023).

[2] Li J, Shang J, Mcauley J. UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. Aclanthology. 2022;1:6159-6169. Accessed December 5, 2023. https://aclanthology.org/2022.acl-long.426.pdf

[3] Alam M, Cevallos B, Flores O, Lunetto R, Yayoshi K, Woo J. Yelp Dataset Analysis Using Scalable Big Data.; 2021. Accessed December 5, 2023. https://arxiv.org/ftp/arxiv/papers/2104/2104.08396.pdf

[4] Choi S. An Empirical Study Identifying Bias in Yelp Dataset. Published February 2021. Accessed December 4, 2023. https://dspacc.mit.cdu/bitstrcam/handlc/1721.1/130685/1251779073-MIT.pdf?sequence=1&isAllowed=y