Predicting Yelp Business Ratings

Stanley Yip
A11319967
syip@ucsd.edu

ABSTRACT

The topic of rating predictions has become a classic data mining task with large public datasets such as the Netflix Challenge and Yelp Dataset Challenge becoming readily available. The core of these predictions focus on user rating prediction as the overall rating of a product is simply the average rating given by users. However, what if user ratings are not given and an overall rating of a product needs to be displayed? Of course, the review text left by a user is surely the most indicative of a product's overall rating, but there are still important factors related to the product itself that must be considered. In this paper, we look at the most recent Yelp Dataset Challenge and propose an approach to predict the overall business rating without using user ratings. Given the direct correlation between overall business rating and user ratings, it comes as no surprise that sentiment analysis on the review text is the most predictive. However, the categories and attributes of a business also contribute a substantial amount.

Keywords

Yelp, Rating Prediction, Features, Linear Regression.

1. INTRODUCTION

In the last few years, data science has grown to become the hot field within computer science. As data storage becomes cheaper and cheaper, companies often log and track user actions in order to gain insight into potential usage trends and act upon that insight. One obvious example would be recommender systems which are popular on product review sites such as Amazon, Yelp, and Netflix. By profiling user behavior on an online platform, advertisements can become more targeted and have higher success rates. At the heart of these recommender systems are rating predictions: what will a given user rate a given product and how will they review the product? In some cases, however, a summarized rating may not be sufficient or only the review text is given.

For this paper, we work with the most recent Yelp Dataset Challenge and look to predict the overall business ratings without using ratings given by reviewers. While the review text is expected to be the most predictive, we expect some predictive power through features inherent to a business such as its type.

The remaining paper is organized as follows: in Section 2, we perform exploratory analysis on the given dataset and highlight key phenomena. In Section 3, we identify our predictive task, identify a baseline for comparison, and select our features based on our exploratory analysis. In Section 4, we evaluate various models against our baseline and in Section 5, we discuss our results. Finally, in Section 6, we discuss related work and conclude in Section 7.

2. EXPLORATORY ANALYSIS

2.1 Overall Statistics

The Yelp Dataset consists of approximately 1.6M reviews for 61K businesses across multiple cities in the United States, United Kingdom, Germany, and Canada. For businesses, the following fields are given:

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not
business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

For reviews, the following fields are given:

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Since we aim to predict business ratings, we will not be considering user ratings as it would defeat the purpose of this predictor.

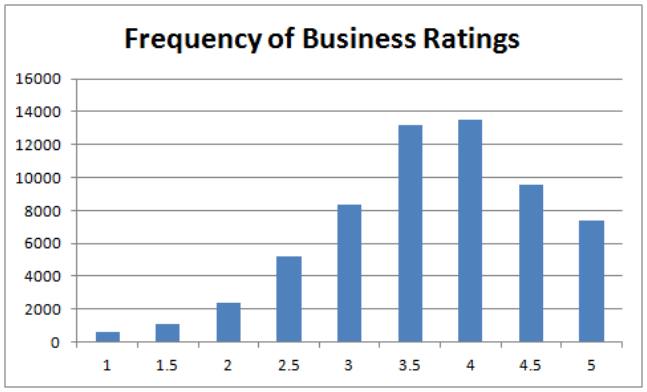

Figure 1a

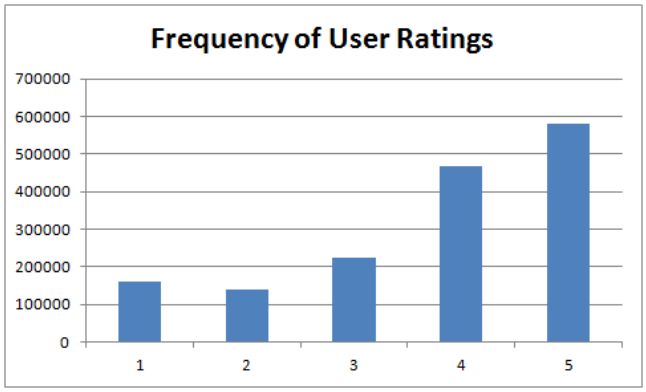

Figure 1b

First we will look at the rating frequencies in Figure 1. As expected, the general frequency of 1 to 5 star business ratings follow a Gaussian distribution with the mean centered around 3.67 (Figure 1a). Intuition explains that a business typically needs to be very poor or very good to be rated 1 or 5 whereas the majority of businesses will be "okay" and be rated around 3 stars.

Interestingly enough, the frequency of individual user ratings are skewed towards 4 or 5 stars as seen in Figure 1b. This could be explained by the fact that reviewers will tend to avoid poorly rated businesses and favor highly rated businesses. Thus the raw number of high user ratings will be greater than the raw number of lower user ratings.

2.2 Business Features

Diving into the business features, we explore the categories. Figure 2 shows the top 10 categories by frequency as well as their average rating. Unsurprisingly, restaurants comprise the majority of reviewed businesses followed by shopping. No clear correlation can also be drawn between the frequency of a category and its average rating. Of course, businesses typically do not fall under a single category but rather a multitude. For example a nightclub would most likely be under the categories of "Nightlife" and "Bars". We then look at how does the number of categories a business belongs to compared to its rating.

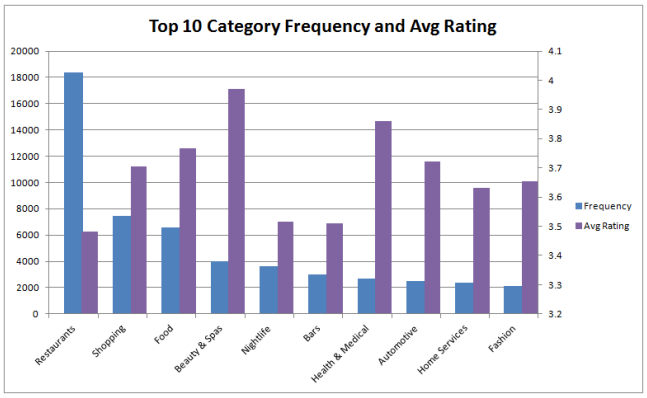

Figure 2

Figure 3a separates the distribution of the number of categories by their rating and we see that most businesses are in 2 categories and are rated 3 or 3.5 stars.

This distribution is very similar to the original distribution of just business ratings and we can see the gradient between number of categories and rating in terms of frequency. We then postulated that the average rating is correlated in some manner with the number of categories. Figure 3b shows this function and there is a generally positive trend.

Similarly, we look at the attributes of a business. Typical attributes include what kind of parking does the business have, does it deliver, does it have take-out, etc. We perform the same analysis on the attributes and find two distinct frequency distributions with the majority of businesses have 0-5 or 18-20 attributes (Figure 3c). We also see in Figure 3d a high average rating for businesses with 0-8 attributes followed by a sudden drop in rating that then steadily increases with the number of attributes. Possible explanations could be that businesses with minimal attributes have excellent products and do not need to "fluff" their business with extra attributes. In the case of restaurants, a restaurant with excellent food will continue to attract customers regardless if it delivers or not. The steady increase of average rating with the number of attributes could be explained by them actually incentivizing customers to rate higher.

An obvious feature to look at would be the price rating of a business. However, price rating was part of the attributes and fewer than 50,000 businesses listed a price range. Thus, we decided against incorporating it as a feature.

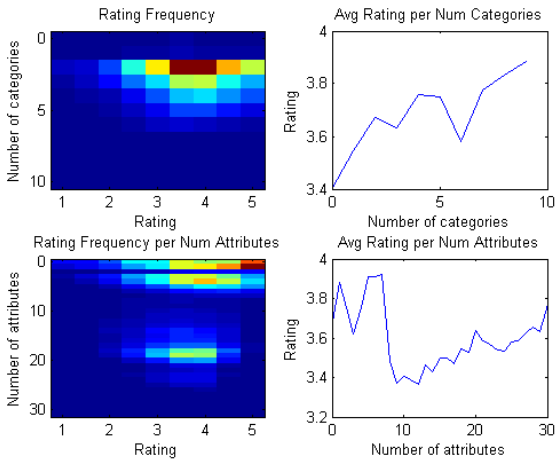

Figure 3a-d

We also heavily looked into the location of a business. Since the dataset is distributed across countries, the raw latitude and longitude of a business would be ineffective. Instead, we looked to see if there were any distinct differences between city averages. As shown in Table I, among the top 10 cities with the most reviewed businesses, there are some slight differences.

Top 10 Cities with Most Reviewed Businesses	
City	**Avg Rating**
Las Vegas	3.6626351
Phoenix	3.6763971
Charlotte	3.5971827
Scottsdale	3.8529338
Edinburgh	3.7923127
Pittsburgh	3.6888766
Mesa	3.6001278
Tempe	3.6525686
Henderson	3.6967136
Montreal	3.7737967

Table I

Along the same vein, we wondered if the location of a business relative to similar businesses mattered. We defined similar business as any business that has a matching category and is within the same city. Since Las Vegas had the most businesses, we decided to focus solely on that city for prototyping. We computed all pairwise distances using the haversine formula. The haversine formula essentially projectes the coordinates onto a 3D cartesian coordinate system and calculates the distance there. We found that these distances were often negligible due to the cramped nature of property in Las Vegas; many stores had the same coordinates because of multi-level plazas. Extrapolating to another city, Phoenix, we found the same phenomena and decided against further exploration of relative location.

2.3 Textual Features

A basic property of the review text is the length. For exploratory purposes, we considered the length of a review text by just spitting by whitespaces. We plotted all review text length in a histogram, revealing a long-tailed distribution with the mean at about 125 words (Figure 4).

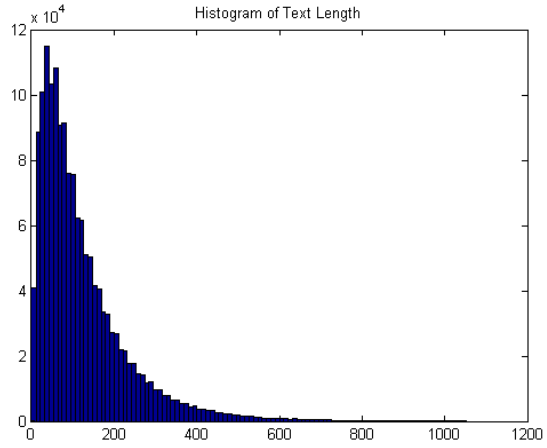

Figure 4

We then separated the text lengths by the associated rating the user gave to a particular business and plotted them as boxplots (Figure 5). There is a slight negative correlation between text length and rating, which can be seen through the whiskers of each boxplot as well as the average length. It appears that 1-star reviews tend to be longer, presumably because the reviewer had more to complain about, than 5-star reviews.

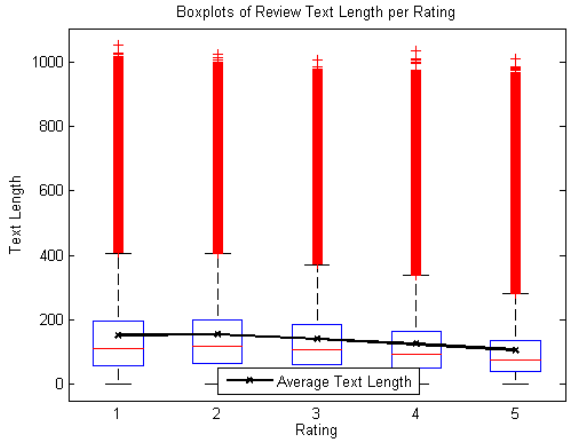

Figure 5

For sentiment analysis, we use TextBlob's polarity rating as an indicator of whether the review was positive or negative. We validate the usage of this library by looking at the boxplots of the review polarities separated by rating (Figure 6). As expected, there is a clear positive correlation between polarity and the rating in the average as well as median polarities. We deemed the library fit to use for our models.

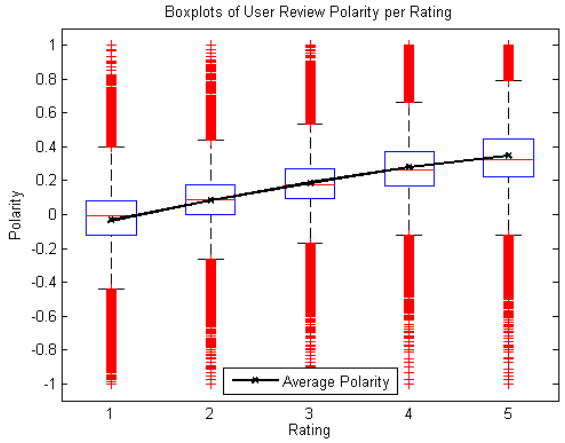

Figure 6

3. PREDICTIVE TASK

For this paper, we aim to predict the business rating. We will be evaluating our classifiers by splitting the dataset into approximately 60*K* training businesses, 5*K* validation businesses, and 5*K* test businesses. The split was done by randomly shuffling all 61*K* businesses given in the dataset and taking the last 5K of the shuffled dataset for the test set and the following 5K for the validation set. We will be using Mean Squared Error (MSE) to compare our models.

3.1 Baseline

For our baseline, we will simply use the global mean rating for businesses, meaning we will predict the global mean for every business. As shown in our exploratory analysis, most businesses tend to be near the global mean. The baseline actually performs decently well with MSE 0.7621633 on the validation set.

3.2 Feature Selection

For our first feature, we use the categories a business belongs to. Namely, we build a categorical binary feature where 0 indicates the business is not part of a particular category and 1 indicates that it is. We choose to use this binary feature rather than just the number of categories because our exploratory analysis of the top 10 most frequent categories reveals that they each have very different average ratings. It makes sense to include this binary representation to capture the actual categories a business is labeled. Since there are only 774 unique categories, we believe it will not add that much computational complexity to our model and the expected benefits outweigh the costs.

Our second feature uses the attributes in the same fashion. For true/false attributes, such as if the business does delivery, we instead use -1 if it is false and 1 if it is true. For all other types of attributes, we merely determine if it is present (1) or not (0). There are only 38 unique attributes so it again does not add much complexity to our model.

Our third feature is the average rating for the city a given business is located in. While not drastic, the differences revealed in our exploratory analysis prompted us to try the feature anyway. We computed this feature by building a dictionary mapping the city name to the average rating. The average rating was pre-computed on the training data by simply taking the mean of all the ratings per each city.

Our final feature is the average polarity of all of the business's reviews. We computed this feature by first pre-calculating the polarities for all 1*M* reviews and mapped the business_id to the average polarity for each business. If no reviews exist for a particular business, we default to a polarity of 0. Since the actual reviews will provide the bulk of the predictive power and we have shown in our exploratory analysis that TextBlob's polarity does a good job expressing sentiment, we believe this feature will have the greatest impact.

4. MODEL

In this section, we will discuss the different models we tried. A summary of the performances of our different features/model combinations can be found in Table II. We evaluate all classifiers using our validation set.

For our predictive task, we initially chose to go with a simple Linear Regression model, specifically Ridge regression. We elected to start with Linear Regression because it is the simplest model to train and validate for.

We decided to first evaluate our business features and textual features using separate classifiers at first just to see how they perform. As expected, the textual features performed the best at 0.4372922 MSE whereas the business features had 0.6091887 MSE. Still it is important to note that the business features alone improved the MSE by 0.1529746.

MSE Summary of Models on Validation Set	
Model	**Mean Squared Error**
Baseline	0.7621633
Category, Attributes, City Avg	0.6091887
Polarity	0.4372922
Bigrams	1.7538123
Trigrams	0.8015264
Quadgrams	0.7629564
Linear Regression	**0.3779972**
kNearestNeighbor (n = 5)	0.4837140
kNearestNeighbor (n = 10)	0.4609070
kNearestNeighbor (n = 20)	0.4577276
kNearestNeighbor (n = 40)	0.4750766
kNearestNeighbor (n = 80)	0.4937891

Table II

We also decided to explore n-grams instead of using polarity. We used the frequency of the top 2000 bigrams, trigrams, and quadgrams of the first 50000 reviews in separate classifiers and found little improvement from the baseline. Bigrams had an MSE of 1.7538123, trigrams had 0.8015264, and quadgrams had 0.7629564. We believed pursuing higher values n would result in overfitting to the training data and we were already feeling a massive increase in computation time. Exploring those top n-grams, we found many were food related such as "french onion soup." This makes sense since restaurants comprised of the majority of the dataset. However, as a predictor, they are unhelpful since "french onion soup" will probably not appear in reviews for a hair salon. It is extremely difficult to attempt to filter out all food-specific n-grams so we went back to polarity.

Combining our business and textual features, we achieve an MSE of 0.3779972 with a Linear Regressor.

We then tried a KNearestNeighbor Regressor. kNN Regression was an attractive option because presumably businesses with similar categories, attributes, etc. would be rated generally the same. If our training set covers the majority of the underlying variance in datapoints, then kNN should be a good predictor. As stated earlier, we randomly shuffled the dataset before splitting into training, validation, and test sets so we believe our training set should be sufficient. While training the model would be relatively straightforward (sklearn.neighbors) simply stores all the training data, actually predicting ratings will be the bulk of the computation time. Luckily, the library function allows parallelization, which greatly decreased the computation time. We trained our model using all of our features and performed a rough optimization of the number of neighbors by manually trying different values.

As can be seen in Table II, k = 20 performed by the best at 0.4577276 MSE with very similar MSE when k < 20 and k > 20. While still a significant increase in performance from our baseline, kNN Regression did not outperform our Linear Regression model. We believe this may be because there are very few 1-2 star rated businesses while many more 3-4 star rated businesses. This imbalance makes it difficult to correctly predict a

1 star business when it has similar features as the more frequent 3-4 star businesses.

We also considered Support Vector Machines, but given our large feature matrix and even larger dataset, it would be infeasible in terms of computation time. We also considered Logistic Regression, but that would necessitate converting our star ratings into classes and turning our predictive task into a multi-label classification problem. While the star ratings are discrete in some sense (values can only be 1-5 in increments of 0.5), the values should be comparable. A 5 star rated business should be "greater than" a 4 star rated business and so on. Thus, we decided to leave the star ratings as is and did not try Logistic Regression.

Since our Linear Regressor outperformed other comparable models, we selected that model as our final one.

5. RESULTS

As discussed in our model selection, our Linear Regression model outperformed our kNN Regression model on our validation set. On the test set, our baseline achieved 0.7876248 MSE. In comparison, our model achieved **0.3944618 MSE** for an improvement of 0.393163.

Top 10 Most Predictive Features	
Feature	**Theta**
Polarity	4.3572975
Pulmonologist	0.8692765
Ethnic Grocery	0.8611180
Trophy Shops	0.8137796
Personal Assistants	0.7916307
Propane	0.7872016
Immigration Law	0.7611711
Private Investigation	0.7546434
Real Estate Agents	0.7217493
Lice Services	0.7052526

Table III

The top 10 predictive features can be seen in Table III. Unsurprisingly, the polarity was the most predictive with theta value of 4.3572975. The remaining are various fairly specific categories with low frequency. We believe this is because the low frequency of reviews for these categories entails potentially less variance of ratings and thus easier to correlate.

6. RELATED WORK

In this section, we will discuss our findings in relation with work performed by others. Since we use the most recent Yelp dataset (6^{th} round) [1], there is a plethora of other papers written using the datasets from previous rounds. Many chose to focus solely on the review text [2] and propose methods on learning hidden dimensions within the text [3]. Since so much work has been done already on text analysis, we chose to focus primarily on other intrinsic features of a business.

In particular, we wanted to find any trends in the location of a business similar to Kothari, et. al. [4]. While their work focused on using proximity to compute average topic vectors and average rating, we wanted to see what minimum distance a business needed to be from a similar business in order to be successful. However, due to the number and broad definitions of categories, it was difficult to achieve a fine enough granularity for similarity determination. In this regard, we attempted a basic model using nearest similar business as discussed in Section 1, and found only a negligible improvement just as Kothari, et. al. did.

As for the textual features, many state-of-the-art sentiment analysis methods exist such as topic modeling (Latent Dirichlet Allocation). Understanding that this particular subtopic has been done through and through, we elected to use a simple library function [5] to cover our needs in this area for a stronger model. Our finding that the average polarity of a business is the strongest indicator was expected and consistent with other works.

Similar datasets that have been studied in the past include the Netflix Challenge as well as Amazon reviews; both contained many user reviews and typical work involved rating prediction of some sort.

7. CONCLUSION

In conclusion, we analyzed the Yelp dataset and sought to predict business ratings without using user ratings. Given that business ratings are simply the average review rating, we knew that the most predictive power would come from text analysis. Much prior work has already been done on the subject matter so we opted to dive into business features and attempt to find strong predictors. While our efforts were not as fruitful as we hoped, we did find decent predictors in the categories and attributes of a business.

In the future, further work could be done on building a stronger definition of similarity between businesses. Given that information, we could use the minimum distance from a similar business or the maximum number of similar businesses within a given radius as a predictor.

8. REFERENCES

[1] Yelp. Yelp Dataset Challenge. yelp.com/dataset_challenge, 2015

[2] Fan, M. and Khademi M. 2014. Predicting a Business' Star in Yelp from Its Reviews' Text Alone. CoRR abs/1401.0864

[3] McAuley, J. and Leskovec, J. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM conference on Recommender systems*, ACM, pp. 165-172.

[4] Kothari, M. and Wiraatmadja, S. 2015. Reviews and Neighbors Influence on Performance of Business.

[5] TextBlob. textblob.readthedocs.org, 2015