IDENTIFYING INFLUENTIAL FACTORS FOR YELP BUSINESS RATINGS

Dhruv Sharma
A53101868
UC San Diego

ABSTRACT

In this paper, we investigate potential factors that may influence business performance on Yelp. We considered businesses' overall star ratings as a measure of their performance. In order to account for user sentiment and location dynamics we constructed additional features from business and user review data. We experimented with regression (Linear and Decision-Tree) as well as classification (Naive Bayes, Decision Tree and Random Forest) models and found that regression models achieved lower error that classification models. We found that across feature selection techniques, the important factors included sentiment of reviews, business location, neighbourhood, ambience of place, etc. However, user review sentiments tend to greatly influence star ratings in comparison to other factors.

1. INTRODUCTION

Yelp [5] was founded in 2004 to help people find great local businesses like dentists, hair stylists and mechanics. Today, the website and their mobile application publishes crowd-sourced reviews about local businesses as well as certain metadata about them that play a role in customer's decision making process. By the end of September, 2015, Yelp hosted more than 90 million reviews written by its reviewers. Yelp uses automated software to recommend the most helpful and reliable reviews for the Yelp community out such large and diverse dataset. The software looks at dozens of different signals, including various measures of quality, reliability, and activity on Yelp.

We on the other hand aim to focus on the problem from the perspective of a local business and aim to identify certain key attributes that tend to influence the overall performance and therefore, the success of a business. To chose a business' star rating as a quantitative measure of its performance and use various models to accurately predict the ratings and identify the features that influence such models.

Total Number of Businesses	13601
Total Number of Reviews	617358
Total Number of Users	173697
Average Number of Reviews per User	3.554224
Average Number of Reviews per Business	45.390633
Total Number of Business Categories	664
Average Review Length (chars)	714.568484
Average Review Length (words)	132.597243
Average Review Length (sentences)	10.885227
Average Star Rating	3.713651

Table 1: Yelp Dataset Statistics for Las Vegas.

To address this problem we used the business metadata and customer reviews from the dataset provided by the Yelp Dataset Challenge [6]. The dataset is a large collection of user reviews, business metadata, business check-ins, users' social network data, user tips for businesses across 10 cities spread across 4 countries.

In order to keep the problem at hand tractable we chose to focus on business data and user reviews for them for the city of Las Vegas, NV, USA.

2. DATASET ANALYSIS

The Yelp's dataset for businesses and user reviews comprises of 61K businesses and 1.6M reviews written by 366K reviewers collected over a period of 8 years starting from 2006. However, as we restrict ourselves to the businesses located in Las Vegas, we filtered the dataset to include businesses located in it and the reviews associated only with these businesses. Table 1 summarizes our basic statistical analysis of this filtered dataset.

Moreover, we studied the relationship between the number of review each business gets as well the number of reviews each reviewer writes. As depicted by Fig. 1 and Fig. 2, we observed that both relationships tend to follow a Power Law distribution where a large number of users and businesses have very few reviews associated with them, whereas in contrast, very few businesses and users have of received or written large number of reviews.

Therefore, in order to reliably predict the star ratings for businesses we discarded all reviews that were written by users with less than 5 reviews and all business that did not have atleast 10 reviews. Our decision was based on an

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

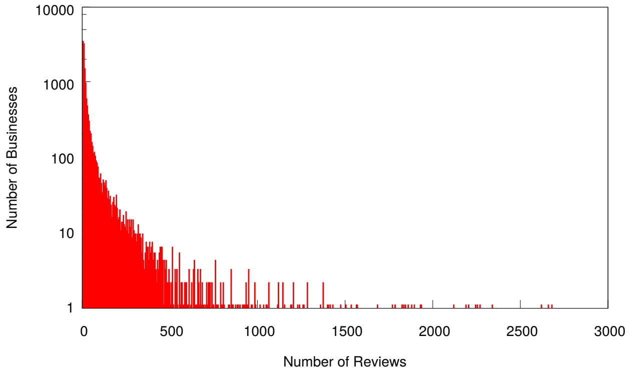

Figure 1: Distribution between Number of Businesses and Reviews.

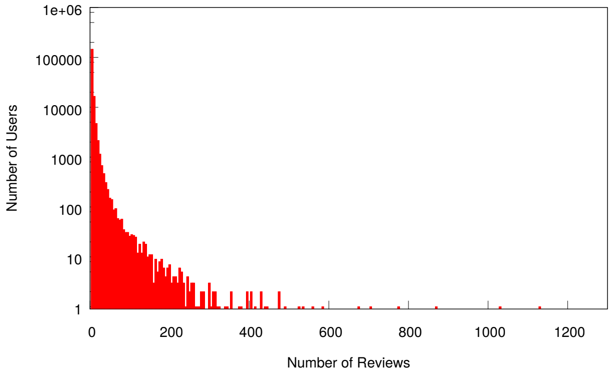

Figure 2: Distribution between Number of Users and Reviews

observation that casual reviewers and newly setup businesses do not tend to have very helpful reviews, in general.

In general, the structure and metadata for each record in the business and the reviews dataset is described by Fig. 3 and Fig. 4.

3. FEATURE POOL

As described in § 2, the business records structure Yelp dataset for businesses, tends to contain numerous features that we intended to use to generate meaningful predictions of the business’ star ratings. However, on closer observation of the dataset entries we observed that certain intuitively helpful features were available in very small number of entries and hence we decided to discard them for our analysis now and only consider features that are at the very least available in 75% of the dataset. Based on this scrutiny, we decided to consider the following set to our feature pool for analysis:-

- Price range
- Review count
- Business coordinates
- Accepts credit cards?
- Good For kids?
- Good for groups?
- Takes reservations?
- Has take-out option?
- Provides delivery?
- Wheelchair accessible?
- Meal times it’s good for?
- Type of ambience

```
{
    ’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),
        ...
    },
}
```

Figure 3: Business meta-data structure

```
{
    ’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)},
}
```

Figure 4: Review structure

In the above list, the interrogative features that either a positive or negative response such as ’Is it good for kids?’, ’Does it provide delivery?’, etc. were intuitive modeled to have a value of 0 or 1 based on whether the response was negative or positive respectively. The feature of price range even though available as an integer, was modeled as 4 different classes of prices (each class a decision variable). This was done as instead of know whether price range affects business ratings we are more interested in knowing which price class attracts positive responses from the reviewers. Similarly, meal times (breakfast, lunch, dinner, etc.) and ambience (classy, touristy, hippy, casual, etc.) were modeled as individual classes to investigate which categories people like more.

3.1 Feature Generation

In addition to the features discussed above, we generated four additional features out of the available dataset - three features to account for location of the business in the city and the third to account for user sentiments in the reviews. They are covered in the following discussion.

3.1.1 Location

While the business location captured by the latitude and longitude are already counted in the feature pool, we also wanted to capture the influence of the general neighbourhood of the city in which it is located. While this field is available in the general structure of the business data, we observed that the neighbourhood field was empty for a large fraction of the businesses in the dataset.

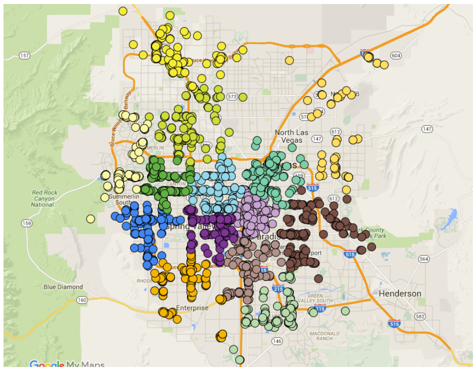

Figure 5: Business Clusters in Las Vegas

Therefore, we decided to cluster the businesses into groups to approximate the neighbourhood they belong to. This was based on a general observation that businesses tend to cluster around popular neighbourhoods rather than being uniformly distributed across the city.

For generating the clusters of businesses, we used the *K-means++* algorithm [8]. It uses an initial seeding heuristic for cluster centers where the next cluster center to be allocated is chosen based on the probability proportional to its square-distance from other clusters. This improves the quality of clusters for a given number of iterations. We used K=14 for the number of clusters, corresponding to the number of popular neighbourhoods in Las Vegas. Fig. 5 shows the approximate clusters we obtained from the algorithm. We observed that the clusters indeed roughly coincide with the neighbourhoods we considered.

The first feature was size of the cluster to which a particular business belongs. The intuition behind this choice was to identify if business ratings are indeed affected by the number of businesses in the surrounding area. Second, we used the businesses' membership across various clusters. This was constructed as a list of 14 decision variables each standing for a business' membership in that cluster. The idea of such a feature set was to identify if the presence of a business in certain selective neighbourhoods than others influences its ratings. For example, businesses in and around a popular tourist spot are expected to perform better than those in some less popular areas (such as residential neighbourhoods). Finally, we used the cluster center co-ordinates as the third cluster location feature to experiment if approximate co-ordinates of the business influence the ratings more than it's precise coordinates.

3.1.2 *Review Sentiments*

Apart from location, reviewers experience with the business is expected to have a strong predictive power in influencing the ratings it receives in general. To capture this effect we analyzed the Yelp reviews dataset for the businesses in consideration.

Bag-of-Words Analysis. We performed a bag-of-words analysis on the reviews of the businesses, where we considered K most popular n-grams approach to filter the tokens that influence the star ratings of the business. We used Ridge regression [3] model provided by the Sci-Kit Learn library [4] to perform a regularized linear regression over the popular n-gram feature set to predict the business rating.

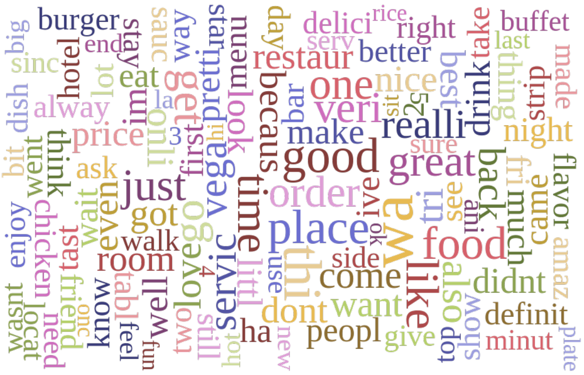

Figure 6: Word Cloud of 300 most frequent words

Token Count (K)	Training Error	Test Error
100	0.506	0.590
200	0.473	0.562
300	0.453	0.541
400	0.442	0.543
500	0.428	0.549
1000	0.380	0.566

Table 2: Variation in RMSE wrt Token Count for 1-gram analysis.

Filter Technique	Training Error	Test Error
Punctuation	0.457	0.552
Punctuation+Stemming	0.453	0.541
Punctuation+Stemming+ Stop-Words	0.461	0.555

Table 3: Variation in RMSE wrt Token Count for 1-gram analysis.

Initially, we used individual words (1-gram) for analysis and focused on optimizing the number of words (K) that optimizes the root-mean square error of the star rating prediction. Our analysis showed that 300 most-popular words tends to give lower error in comparison to the full set of words. This indicates that there is comparatively smaller set of words that occur more frequently in reviews but tend to have significant influence over the ratings. Table 2 presents the results from this analysis, whereas the word cloud shown in Fig. 6 shows the relative frequency distribution of these top performing words.

For this analysis we performed n-gram frequency computation with combinations of filtering mechanisms such as punctuation removal, word stemming, and stop words elimination. We found that punctuation removal coupled with word stemming gives comparatively lower error than punctuation removal alone. Also, we observed that elimination of stop words from the reviews actually causes the error to marginally increase. This seems counter-intuitive that stop words do not contribute to the expression of text sentiment and replace more, the analysis indicates that stop words indeed have some positive correlation with user sentiment and perhaps with the businesses' star ratings. Table 3 presents the result of this observation for 300 most popular words.

Finally, we use the predictions using our optimized feature set and Ridge regressor to generate an approximate star ratings for the businesses based on the our sentiment analysis. However, instead of considering these as final predictions, we use them to add another feature to the feature pool that stands for reviewers sentiment quantifying their experiences with a particular business. A lower value of this feature indicates bad customer experiences whereas higher value indicates the business is liked by customers in general.

4. EVALUATION

In this section, we present the evaluation of our feature set using various models available in the Sci-Kit Learn library and discuss our reasons behind choosing to evaluate using the particular predictor.

However, prior to that we required to pre-process the dataset in order to account for the missing values of some features in the dataset. Such missing values usually existed in decision features such as accepts credit cards, wheelchair accessible, etc. For replacing the missing values, we used the Imputer in the SciKit Learn's preprocessing module which provides various replacement strategies such replace with mean, median, most frequent occurrence (mode) or fixed values (such as 0). Since, most missing values existed in decision or membership variables that had 0/1 values, we chose to use most frequent occurrence replacement strategy as mean or median strategy would also give similar results based on the range of data values under consideration. As fixed values (such as 0) have chance of creating outliers we refrained from using this approach.

4.1 Ridge Regression

We used the Ridge regression model with regularization provided by Sci-Kit Learn library to fit the full set of features in the feature pool as described in § 3. However, as this would generally result in too many redundant and/or less useful features to used in predictive task leading to unnecessary bloating of feature set and might result in less accurate predictions. In order to optimize the feature set to gives more accurate predictions of the star ratings, we used the recursive feature elimination technique provided by the Sci-Kit Learn's feature selection library. To be precise, used Recursive Feature Elimination with Cross Validation (RFECV) tool. The tool runs multiple iterations over which it trains the given estimator and cross-validates (we used 4-fold validation) the available features and ranks them based on the corresponding response of the prediction signal, and finally, eliminates one feature per iteration (can be varied).

The process of optimal feature selection result indicated that 7 is the optimal number of features to be used out of initial 48 features, and provided a ranking of the features in the order they performed in providing accurate predictions. The top 10 ranked features from this analysis were:-

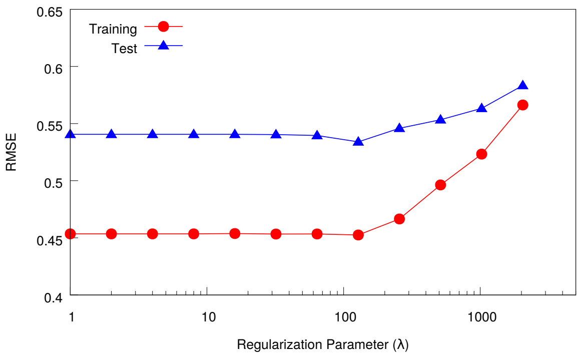

Figure 7: Variation of Training and Test RMSE with λ

1. Review sentiments
2. Presence in certain clusters
3. Business coordinates
4. Cluster coordinates
5. Accepts credit cards?
6. Good for groups?
7. Certain ambience types
8. Certain meal times?
9. Takes reservations?
10. Higher price ranges

While these results are quite interesting and are greatly influenced by the fact that we are considering businesses in Las Vegas, we defer the discussion to a later section. We used the features selected by RFECV tool to optimally train the linear regression as well as the following models to select the best model for predicting star ratings for businesses. The Fig. 7 shows the variation of error with the change in lambda.

4.2 Multinomial Naive Bayes

As the target star ratings are rounded to the nearest 0.5 stars, this gives us an opportunity to transform the task of predicting ratings into a multi-class based classification problem that predicts the most probable class i.e. star rating based on the optimized set of features found in § 4.1.

We use the Multinomial Naive Bayes algorithm available in Sci-Kit Learn library. As this requires the labels to integral values to form the correct set of classes, we scaled the training star ratings to a scale of 10 and reduced the predictions back to original star ratings without loss of precision.

4.3 Decision Tree & Random Forests

We build a decision tree classifier using similar label transformation as done in § 4.2 for classification. However, the Sci-Kit Learn's Decision Tree classifier requires positive values, we had to consider only the longitude values' magnitude. Also, instead of classifying the star ratings, we used Decision Tree regression to predict the numeric values for the ratings.

Random Forests on the other hand, is an ensemble prediction model that builds multiple decision trees using various combination of features and uses averaging to improve overall prediction and prevent the effects of over-fitting. We built our model with 10 decision trees estimators.

Model	Training Error	Test Error
Baseline (Mean)	0.657	0.799
Ridge Regression	0.452	0.534
Multinomial Naive Bayes	0.664	0.738
Decision Tree Classifier	0.122	0.658
Decision Tree Regression	0.103	0.663
Random Forests	0.186	0.638

Table 4: Training and Test RMSE for various models

5. DISCUSSION

We now discuss the significance of the features selected in § 4.1 and the results obtained by various models in § 4.

We know that Las Vegas is a city that is visited by large number of tourists round the globe and all round the year. We find that the most important feature identified by us is review sentiment which indicates the business ratings are most influenced by the positivity of the experience it's customers have with it. A better customer service therefore is likely to improve businesses' star ratings. Apart from that, we observed that business-coordinates, cluster-coordinates, membership in certain clusters and, to some extent, cluster size also influence star ratings. It indicates that businesses in some very popular neighbourhoods or hot spots tend to get higher ratings than others and hence opening business in such localities can provide boost to the ratings. Since, tourists usually visit Las Vegas in groups, the "good for groups" attribute has been given high importance. As a general observation people visiting Las Vegas tend to spend lavishly and therefore tend to like businesses that are pricier than others. Apart from that people tend to give higher ratings to businesses that have "touristy", "divey" and "classy" ambience. Also, people tend to care about facilities like whether the business accepts credit cards or takes reservations in advance or not. The ones that do provide seem to get higher ratings than the ones that do not provide these facilities.

The Table 4 summarizes the performance of various models we experimented with for predicting the star ratings. It turns out all of the above models indeed out-perform the baseline model that predicts the mean star ratings for every business. However, an optimized linear regression model out-performs other classification models. The Decision Tree classification and regression, as well as Random Forest models show very low training error but a poorer error when predicting on the test dataset. This can be attributed to the over-fitting problem that tree-based classifiers suffer from. Naive Bayes classifier tends to perform badly for both training and test datasets which indicates that the features used still have some inter-correlation with each other. This can be seen that multiple features that identify the businesses' location were selected leading to poor performance for Naive Bayes classifier.

6. RELATED WORK

The Yelp dataset challenge has run through multiple rounds over the last couple of years and has produced a number of research papers and innovative recommendation techniques.

We discuss a few those works that closely match to the prediction task we aim to solve in this paper.

Feng, et. al. [7] experiment with various predictive models to accurately predict the ratings a specific user is likely to give to a certain business, solely based on the Yelp reviews dataset. While they observe that collaborative filtering models tend to perform slightly better than multinomial logistic regression, decision trees classifiers and K-nearest neighbours algorithms, better accuracy is achieved by using ensemble models comprised of some combination of the these models.

Carbon, et. al. [1] follow an approach very similar to ours in evaluating various models and feature set construction, however, they tackle the problem as success or failure decision problem rather than multi class ratings prediction. Nevertheless, the features they identify to large extent conform with the optimal feature set that we found during our analysis of the problem.

Li, et. al. [2] on the other hand tackle the problem of predicting business ratings as a classification problem using the review dataset. While their evaluation lacks detailed discussion of the prediction models used, their support-vector regression model tends to provide marginal improvements in predictions using sentiment analysis process that closely matches our own.

McAuley, et. al. [9] introduce "Hidden Factors as Topics" model to exploit the review text to identify the relationship between latent factors of users and products/businesses across multiple dataset such as Amazon product reviews, Yelp dataset, Beeradvocate reviews, etc. to provide stronger correlation between ratings and the review text. They were successful in extracting meaningful topic identifiers that can be used to enrich the recommendation of useful reviews and products.

7. CONCLUSION

We present an analysis of the features that greatly influence a business' star ratings for the businesses located in Las Vegas. Such an analysis can prove to be quite useful for upcoming businesses to identify and according transform based on the features that their customers tend to like in general which can help them improve their ratings and consequently the revenue by attracting more customers. As review sentiments greatly matter is deciding the star ratings, businesses should focus more on the providing good customer service and more facilities that are more suited for tourists coming out of town.

Also, we analyze various models from linear regression to multi-class classification. We observed that, linear regression is better suited for predicting the star ratings than classification models. Moreover, decision tree-based classifiers and regressors suffer from the problem of over-fitting, whereas the naive-bayes classifier performs poorly in all cases due to the inherent correlation in the top performing features that we identified.

8. ACKNOWLEDGEMENTS

We thank our instructor Prof. Julian McAuley for his continued direction throughout the course, and to the teaching assistants for their guidance in the analysis process.

9. REFERENCES

[1] Applications of Machine Learning to Predict Yelp Ratings. http://cs229.stanford.edu/proj2014/Kyle%20Carbon,%20Kacyn%20Fujii,%20Prasanth%20Veerina,%20Applications%20Of%20Machine%20Learning%20To%20Predict%20Yelp%20Ratings.pdf.

[2] Prediction of Yelp Review Star Rating using Sentiment Analysis. http://cs229.stanford.edu/proj2014/Chen%20Li,%20Jin%20Zhang,%20Prediction%20of%20Yelp%20Review%20Star%20Rating%20using%20Sentiment%20Analysis.pdf.

[3] Ridge Regression. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html.

[4] Sci-Kit Learn. http://scikit-learn.org.

[5] Yelp. http://www.yelp.com.

[6] Yelp Dataset Challenge. http://www.yelp.com/dataset_challenge.

[7] Yelp User Rating Prediction. http://cs229.stanford.edu/proj2014/Yifei%20Feng,%20Zhengli%20Sun,%20Yelp%20User%20Rating%20Prediction.pdf.

[8] A. D. and V. S. k-means++: the advantages of careful seeding. In *ACM-SIAM'07*.

[9] M. J. and L. J. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys'13*.