CKME-136: Analysis of Restaurant Features towards Higher Rating

**Introduction**

Yelp has always been a trusted source for finding high rated restaurants around the places. Their dataset provides a huge list of data containing each restaurant’s features and their ratings along with some user details, review details and check-in details. A lot of studies have been conducted to do the sentiment analysis on review text and predicting the star rating. But none of them are focused on if there is a relationship between high rating and any of the restaurant features. In this project I will try to analyze which features of a restaurant are more likely to contribute directly towards higher ratings; this result could become a great source for business owners or entrepreneurs to improve their ratings.

The theme for this project is Classification and Regression. The Yelp Dataset contains over 1.4 million business attributes like hours, parking, availability, and ambience for 188,593 businesses in 10 metropolitan areas1.

I plan to use Hadoop with pig/ hive to load the dataset, use R language to preprocess the data. Then build at least two/ three classifier models using different methods (Logistic Regression/ Naive Bayes/ Random Forest) to find out important features (i.e. hours/parking/availability/location etc.) of a business to get higher rating. At a later point, I plan to run some statistical test (i.e. ANOVA) to find the significant difference between classifier models built earlier.

**Literature Review**

Yelp dataset is one of the publicly available dataset which comprises a comprehensive set of restaurant features, their ratings and user comments. This dataset has been used in many researches to find valuable insight from within the restaurant industry in together with various data mining and machine learning techniques.

In [CFV14], the project involved investigating potential factors that might affect the business performance. The study first selected several features that came directly from the dataset like latitude, longitude, review count, price range, accepts credit cards, has take-out, has delivery etc. Then they generated additional features using clustering on latitude and longitude to determine the location of the restaurant (whether it is in downtown, or in shopping mall etc.). The features they have derived from k-means clustering are cluster size and cluster label. Then they have used Naïve Bayes Classifier for review sentiment analysis and derived 10 features (i.e. Horrible, Awful, Disappointed, Waited etc.). After feature selection they have used various methods of feature selections (Univariate Feature Selection, Recursive Feature Elimination and Tree Based Feature Selection). Then they used these features from each method separately to several predictive models (Support Vector Machine, Multinomial Logistic Regression, Multinomial Naïve Bayes, Gaussian Discriminant Analysis, Decision Trees and Random Forest Classifier). Then they presented top 5 features generated from each model as follows:

|  |  |
| --- | --- |
| **Method Used** | **Top 5 Features** |
| Recursive Feature Selection with Logistic Regression | Sentiment, Reservation, latitude, longitude, Review Count |
| Univariate Feature Selection with SVM | Sentiment, No. of reviews, Cluster Size, Reservations, Longitude |
| Tree based feature selection with SVM | Sentiment, Latitude, Longitude, No. of Reviews, Cluster Size |

In a study [A16] on Yelp Dataset Challenge a system or model has been proposed to predict the rating of a restaurant based on its review text. The study noted that when a user gives a rating to a restaurant, it is not possible for a user to consider all aspects and many times user assign an overall star rating solely based on a specific aspect. Therefore, it is important to have a rating prediction system which will show the rating based on user review text. In this study, they used four different feature extraction methods together with 4 supervised learning algorithms; creating a total of 16 different prediction models. They used 80% training data and 20% to validate results. Four feature extraction methods that they have used: Unigrams, Unigrams & Bigrams, Unigram Bigram & Trigram and Latent Semantic Indexing (LSI). Four supervised learning they have used: Logistic Regression, Naïve Bayes Classification, Perceptrons (a linear classifier that outputs class labels instead of probabilities) and Linear Support Vector Classification (SVC). Then they compared the result of all 16 models and found that Logistic Regression achieved the highest accuracy of 64% using the top 10000 Unigrams & Bigrams as features. The second best performing system was Linear SVC (63% accuracy) using top 10000 Unigrams & Bigrams. After that they test these two models using the test dataset, they found that Linear SVC achieved 56% accuracy and Logistic Regression achieved 54% accuracy, which are lower than the validation-fold scores indication possible overfitting. So, they proposed to add/ adjust regularization parameters as their future work.

In the study of [P17], a novel approach has been proposed to predict Yelp restaurant rating not only based on business features (i.e. images, descriptions) and user features (i.e. average previous ratings) but also considering the network features (i.e. which businesses a user has rated before). In their study, they demonstrated that a mixed approach combining node level features with network information can lead to a better Yelp star rating prediction. They split their data into training (80%), validation (10%) and test (10%) dataset and applied Linear Regression, Ridge Regression, Bayesian Regression, Neural Network and Random Forest. Although in training set Random Forest had the lowest RMSE, but in both validation and test set, Neural Network outperformed as the best supervised learning method with the lowest RMSE indicating that Random Forest seems to be overfitting the data.

Another study [FK14] was done to predict the Yelp rating solely based on its review text. In their study they have used three feature selection methods and four learning methods (Linear Regression, Support Vector Regression, Support Vector Regression with normalized features and Decision Tree Regression). The first feature selection technique was analyzing raw data and choosing top K-frequent words used in all reviews and counting their frequency. In the second feature selection method they analyzed Part-of-Speech (POS) per sentence. Based on the analysis, they selected top K frequent words amongst all. In the third feature selection technique they have extracted top K frequent adjectives. For each feature selection techniques, they have used K = 30,50,100,200,300,500,1000 and selected the one with least RMSE. In their study they found that no matter what feature selection they choose, Linear Regression always performs better than other learning model and within the feature selection techniques, Top Frequent Words from Raw Data with Linear Regression had the lowest RMSE.

In [LLJ+11], authors have argued that just doing a sentiment analysis on the review text is not enough because same text could carry different sentiments for different reviewers. Therefore, it is necessary to scale the sentiment based on reviewer and product specifics. In their study they have designed a predictor function to scale that. Then they compared their model against with Regression, Regression + PSP [PL05], SVM and SVM+PSP. They showed that their model outperformed rest of the model with the least RMSE.

In another study [K17] on Yelp dataset, the author has proposed a model which is a hybrid of neural network, decision tree and logistic regression to predict the review rating given the location (longitude, latitude) and business category. This model predicts if a new restaurant is opened in a certain location, whether it will belong to positive class (start rating greater or equal to 4) or negative class (less than 4). They have used a feature engineering technique called Extremely Randomized Trees [GEW06]. Extremely Randomized Trees is a tree induction algorithm that selects splits both attribute and cut-point totally or partially at random and in extreme cases it is capable of building totally randomized tree, whose structure is independent of output values of learning sample.. Then they compared their model with the results from Random Forest and Neural Network and showed that this model outperforms other models with a accuracy of 67.3% and highest Area Under Curve (AUC) of 73%.

Another study [SC00] was done on 63 Toronto restaurants to find out which attributes or features have statistically significant effect on customers’ ratings. The source of their data was from Zagat Survey and Toronto Life Magazine. Then they surveyed those restaurants’ managers to get data about average check, restaurant volume and restaurant size. All the restaurants used in the study were similar in size and volume. The study basically focused on nine attributes (i.e. Food, Décor, Service etc.) and finding out which attribute contributes most to the average check. They gathered the average check data from Zagat Foundation (which collects data from customers) and then they validated the information directly with the restaurants. They found it to be largely consistent. The correlation between the average check reported by the foundation and managers was 0.78, which indicates the strong level of agreement between two parties. The methodology they have used to find out important features that affect positively or negatively is Two-tailed correlation significance with α = .05 level. Some interesting findings from their study are as follows:

* 81% of their sample’s high rated restaurant had smoking section, 63% offered catering and 54% offered takeout.
* Three features that were not offered by all but still showed strong presence: dress code, parking and outside dining.
* Three uncommon attributes: internet presence, late-night menu and some sort of entertainment (i.e. music, dancing, comedy)

**Dataset**

For this project, I am using the Yelp Open Dataset [Y18]. The dataset contains information about 188,593 businesses in 10 metropolitan areas consisting of almost 6 million user reviews and 280,992 pictures.

My research question is to identify important business features that significantly contributes towards the higher ratings (4 or more start ratings). So, my dataset will be limited to business dataset. Yelp’s business dataset is in JSON format and contains the following:

### business.json

Contains business data including location data, attributes, and categories.

{

// string, 22 character unique string business id

**"business\_id": "tnhfDv5Il8EaGSXZGiuQGg",**

// string, the business's name

**"name": "Garaje",**

// string, the neighborhood's name

**"neighborhood": "SoMa",**

// string, the full address of the business

**"address": "475 3rd St",**

// string, the city

**"city": "San Francisco",**

// string, 2 character state code, if applicable

**"state": "CA",**

// string, the postal code

**"postal code": "94107",**

// float, latitude

**"latitude": 37.7817529521,**

// float, longitude

**"longitude": -122.39612197,**

// float, star rating, rounded to half-stars

**"stars": 4.5,**

// interger, number of reviews

**"review\_count": 1198,**

// integer, 0 or 1 for closed or open, respectively

**"is\_open": 1,**

// object, business attributes to values. note: some attribute values might be objects

**"attributes": {**

**"RestaurantsTakeOut": true,**

**"BusinessParking": {**

**"garage": false,**

**"street": true,**

**"validated": false,**

**"lot": false,**

**"valet": false**

**},**

**},**

// an array of strings of business categories

**"categories": [**

**"Mexican",**

**"Burgers",**

**"Gastropubs"**

**],**

// an object of key day to value hours, hours are using a 24hr clock

**"hours": {**

**"Monday": "10:00-21:00",**

**"Tuesday": "10:00-21:00",**

**"Friday": "10:00-21:00",**

**"Wednesday": "10:00-21:00",**

**"Thursday": "10:00-21:00",**

**"Sunday": "11:00-18:00",**

**"Saturday": "10:00-21:00"**

**}**

}

From the above dataset, based on my research question, I am planning to use the following attributes:

// float, star rating, rounded to half-stars

**"stars": 4.5,**

// interger, number of reviews

**"review\_count": 1198,**

// integer, 0 or 1 for closed or open, respectively

**"is\_open": 1,**

// object, business attributes to values. note: some attribute values might be objects

**"attributes": {**

**"RestaurantsTakeOut": true,**

**"BusinessParking": {**

**"garage": false,**

**"street": true,**

**"validated": false,**

**"lot": false,**

**"valet": false**

**},**

**},**

// an array of strings of business categories

**"categories": [**

**"Mexican",**

**"Burgers",**

**"Gastropubs"**

**],**

// an object of key day to value hours, hours are using a 24hr clock

**"hours": {**

**"Monday": "10:00-21:00",**

**"Tuesday": "10:00-21:00",**

**"Friday": "10:00-21:00",**

**"Wednesday": "10:00-21:00",**

**"Thursday": "10:00-21:00",**

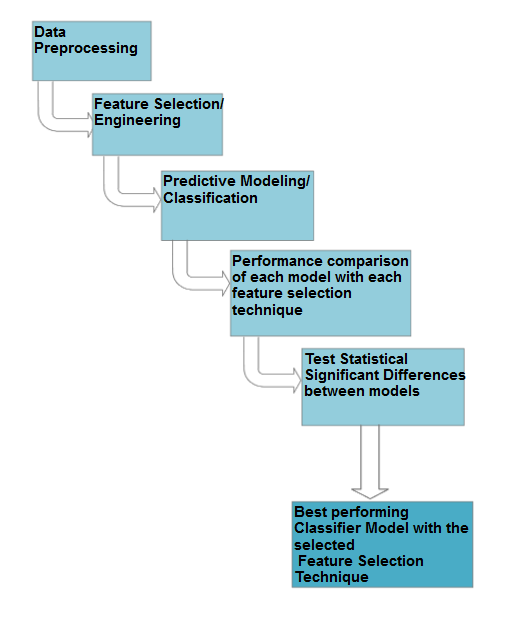
**"Sunday": "11:00-18:00",**

**"Saturday": "10:00-21:00"**

**}**

**Approach**

Following is the high level of approach taken to solve the research question. In the following sections, they have been further discussed in detail.



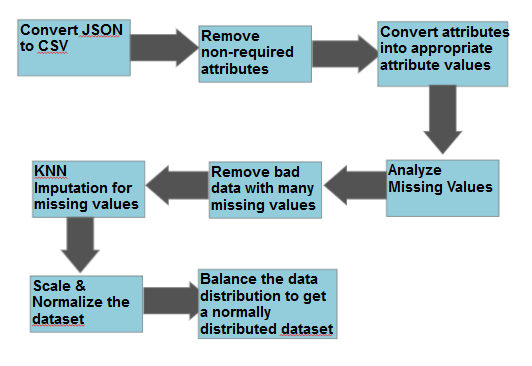
***Figure 1****:* ***High Level Overview of Approach***

**Step 1: Data Preprocessing**

* Convert the business.json file to csv format.
* Remove all non-required attributes.
* Convert the rating attribute to a class attribute with two values: YES (rating >=4), NO (rating <4)
* Analyze the review count attribute values and divide it into three ranges attribute and assign each business into one of these three values: LOW, MEDIUM, HIGH
* Analyze the values in categories and either convert each category into an attribute or create a new attribute called no\_of\_categories\_offered and assign the count of categories as value for this attribute.
* Then we need to convert the hours attribute into appropriate values. We plan to convert them into
  + offer\_breakfast\_hour,
  + offer\_lunch\_hour,
  + offer\_dinner\_hour,
  + offer\_latenight\_hour,
  + open\_seven\_days

attributes and assign YES/NO values for each row.

* Analyze missing value count for each row remove observations with multiple missing values if required.
* After removing bad data, we need to replace existing missing values in the dataset. Since our data is categorical, KNN data imputation could be a good option.
* For better analysis and classification, we need to scale and normalize the data.
* Balance the data distribution: we need to analyze the data and if it is not normally distributed, we need to resample the dataset for equal distribution.



***Figure 2****:* ***Data Preprocessing Steps***

**Step 2: Feature Selection**

Feature selection is the mechanism that will give us insight about which attribute has high correlation with the class attribute. We plan to use at least three feature selection techniques and later use the attributes found with different classifiers. For this step, we plan to use:

* Pearson’s correlation
* Information Gain Ratio to Class attribute
* Recursive Feature elimination

**Step 3: Predictive Modeling/ Classification**

In this step, we plan to use at least three classifiers. First we need to split the dataset in 80%-20% training and test set. Then we need to train each classifier with all the feature selection technique. Therefore we will have 3 x 3 = 9 resulting models. Three classifiers we plan to use are:

* Logistic Regression
* Naïve Bayes
* Random Forest

**Step 4: Performance Comparison of each model**

In this step, we will test all 9 models using our test data and record the Accuracy and Precision and put them into a matrix for a better comparison and visualization. From there we could get an understanding of which model and which feature selection technique is performing better and whether there is any data overfitting problem.

**Step 5: Test Statistical Significant Difference**

We also plan to test the model for any statistically significant differences. We plan to use a statistical method **ANOVA** to conduct this test.

**Step 6: Best Performing Model**

Finally, from all the testing conducted above we will try to draw a conclusion about which model is performing better and why and also which restaurant attributes seems to carry significant weight towards high rating.

**References:**

[A16] Asghar, N. (2016). Yelp Dataset Challenge: Review Rating Prediction. *arXiv preprint arXiv:1605.05362*.

[CFV14] Carbon, K., Fujii, K., & Veerina, P. (2014). Applications of Machine Learning to Predict Yelp Ratings.

[FK14] Fan, M., & Khademi, M. (2014). Predicting a business star in yelp from its reviews text alone. *arXiv preprint arXiv:1401.0864*.

[GEW06] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3-42.

[K17] Kaing, D. (2017, October). Yelp business rating classification using hybrid ensemble. In *Knowledge Engineering and Applications (ICKEA), 2017 2nd International Conference on* (pp. 30-33). IEEE.

[LLJ+11] ]Li, F., Liu, N., Jin, H., Zhao, K., Yang, Q., & Zhu, X. (2011, July). Incorporating reviewer and product information for review rating prediction. In *IJCAI* (Vol. 11, pp. 1820-1825).

[PL05] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115-124). Association for Computational Linguistics.

[P17] Perez, L. (2017). Predicting Yelp Star Reviews Based on Network Structure with Deep Learning. *arXiv preprint arXiv:1712.04350*.

[SC00] Susskind, A. M., & Chan, E. K. (2000). How Restaurant Features Affect Check Aberages: A Study of the Toronto Retaurant Market. *Cornell Hotel and Restaurant Administration Quarterly*, *41*(6), 56-63.

[Y18] Yelp Open Dataset. (2018). <https://www.yelp.com/dataset/>