What urban factors affect the success of the restaurant?

Applied Data Science

Aparna Bhutani, Di Xin, Guilherme Louzada, Yutong Zhu, Zheyuan Zhang

Dec 17 2019

Github
https://github.com/JasonZhangzy1757/Applied-Data-Science-Capstone

Abstract

New York is a city with a huge diversity of people from all over the world and a wide variety of restaurants. This motivated us to think of what exactly makes a restaurant successful. In our project we are trying to explore these factors which influence the restaurants' success using yelp data and urban data. In our approach, we first do clustering to find the clusters with similar types of restaurants. We further explore these restaurants along with NLP for customer reviews to analyze the customer satisfaction and we eventually find the factors which drives the success of restaurants in New York.

Introduction

Restaurants are an important piece of the American economy. With more than 1 million restaurants across the United States (Wickford, 2018), the industry is massive, but opening a new restaurant and maintaining it is a difficult task: most restaurants close during their first year of operation (Wickford, 2018).

At the same time, due to the Internet blossoming, restaurant reviews are everywhere. Yelp is the new standard for urban people making decisions (Dellarocas, 2003). Many researchers have analyzed online review data to impact consumers' behavior and decision making. (Zhu & Zhang, 2010, Mudambi & Schuff, 2010, Duan & Whinston, 2008) Although reviews and rates also have a significant impact on sales and business revenues (Chen et al., 2003), less research put their vision on the business's side.

The aim of our work was to understand **what urban factors affect the success of a restaurant**. In order to achieve that goal, we use both yelp data and urban data like income and population.

Literature review

Word of mouth (WOM) has been a way to receive information about any recommendation since ancient times. However, the WOM depends on the nature of the sender-receiver relationship, the richness and strength of the message and its delivery. It includes numerous personal and situational factors while people receive reviews from others. (Sweeney, Soutar & Mazzarol, 2008）

While we enter the digital era, review websites and apps transferred the old- fashion word of mouth. (Tucker, 2011& Luca, 2016) Electronic word of mouth is defined as "any positive or negative statements made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau, 2004).

Digital information provides a multi-dimensional way for visitors to view and select their next-stop. Yelp creates an environment where consumers can communicate about their experience at a particular business in the daily and practical language (Funk, 2009). Hence, these online reviews weaken Sweeney's research about the nature of the sender-receiver's relationship. In contrast, these reviews influence the restaurant by playing a significant role in the restaurant's revenues. (Duan, W., Gu, B., & Whinston, A. B. 2008).

Amazon's study and yelp study all come to two similar findings that the more the review, the higher the revenue. Secondly, consumers react more to review-content, not the statistical summary.(Chen, Wu & Yoon, 2003 & Chevalier & Mayzlin,2006 & Luca, 2016) Amazon's study also indicates that people react more to a one-star review compared with a five-star review. (Chevalier & Mayzlin, 2006). Yet, less viewed merchants are highly related to reviews compared with the popular merchant. (Chen, Wu & Yoon, 2003)

Yelp's study indicates consumer response more when a restaurant contains more information. In other words, consumers are more likely to pick and rate a restaurant with a relatively high amount of reviews and the "elite" reviewers amount. (Luca, 2016)

Data

All restaurant data is collected and processed from Yelp.com. For clustering and classification, we used information from the United States Census Information website to aggregate Phoenix income per zip code. The shapefiles of the zip codes from Phoenix (AZ) were downloaded from the Esri Website. All information was linked together by zip code for further analysis.

In order to simplify the access of the information from Yelp, we reduce the total amount of data to be processed. The main dataset to be used was the Review Information(5,261,668 lines and 3GB), since working with such kind of dataset would be burdensome, we decided to filter only the data related to Phoenix (AZ), reducing the dataset size in 90% (resulting in 570k records).

The business attributes file from Yelp.com contains all business information for each restaurant. To avoid missing data failure, we sum similar features and clean out 29 features that contain 0 entry. (Appendix 1)

Other data processing will be explained in each methodology later.

Methodology

Sentiment Analysis

In our project, we have used Natural Language Processing for sentiment analysis of Yelp reviews for every restaurant in Phoenix. Each review gives us a polarity score of the review and we calculate the overall polarity score for each restaurant. We follow the following process for sentiment analysis:

Data pre-processing:

Data cleaning consists of the part where we get our reviews data ready for analysis. For data preprocessing, we do some steps by ourselves and some by using the inbuilt NLTK library. (Loper & Bird, 2002)

Word Tokenization: In our first step, we divided each work into tokens, which are separate words.

Convert to lowercase: This step convert text data to lowercase so that similar words which are uppercase/lowercase can be identified to be similar.

Remove punctuation, numbers : In this step, we removed the punctuation and numbers such as : , " ' 7 8 so that the similar words can be represented in the same way.

Remove stopwords: In this step, we remove the commonly occurring words which gives no meaning to the sentence such as over, under, once, here, there and so on. Removing stopwords gets rid of the unimportant words and provides us significant words which will be used for further analysis

Stemming: We converted words that share the same original form into its original. Words such as eating, eat, eats are all the same. In this step, we convert them to the same format. For example we converted eat, eating, eats all into eat.

Lemmatization: After getting the root words, we reduced our data by assigning similar words to their dictionary or canonical form. For example: is and are are assigned to “be”.

Sentiment Analysis:

After data pre-processing, we use our cleaned data for sentiment analysis. For this step, we use the Text Blob library. The Text Blob library automatically performs the functions of when a text blob object is created. The steps we followed for Sentiment Analysis were as follows:

Polarity calculation: We used the TextBlob library for calculating the polarity score of each review for every restaurant. This polarity value lies between [-1,1] where -1 means most negative, and 1 means most positive reviews. 0 means a neutral score. (Vijayarani & Janani, 2016)

Normalizing polarity score using VADER Algorithm: After calculating the polarity of every review, we normalize and find the compound polarity score of each restaurant. This process is done based on the concept proposed by VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm, which finds the compound or aggregated score by normalizing it. (Hutto & Gilbert, 2014, May)

After doing these steps, we got a polarity score for each restaurant. We used this score as a factor for our further analysis.

Clustering

Data Preprocessing

The dataset is created by merging income population data with category data. After setting the index as the postal code, we have the data prepared. Then the features are scaled and applied to PCA for features selection. We pick 22 components, which contains 80% of the original information (Figure 4-1). Finally, after the features are selected, the data has been preprocessed.

We tried K-Means, Gaussian Mixture, and DBSCAN and using silhouette score as a guideline to pick the number of clusters. Finally, we decided to use DBSCAN, for it gave the best output. So the silhouette score is not necessary but still could work as a validation. We tuned epsilon to get the optimal clusters.

We downloaded Arizona shapefile from ESRI, using geopandas to make a simple visualization (Figure 4-2). The clusters seem to be reasonable because the southeast part is where most commercial centers gathered in Phoenix City, according to google maps. Thus, it is safe to say we the clustering succeed in giving correct insights.

Furthermore, we used the folium library to create an interactive map, which makes a better visualization (Figure 4-3). Adding another layer for labels could give a clearer understanding of the graph (Figure 4-4).

We outputted the labels and zip code to a CSV file for further analysis in classification.

Classification

Based on the two clusters, the blue area stands for the commercial area, and the red area stands for the other places in Phoenix. We used the classification methods in order to see the differences among impacting factors, that can lead to the success of the restaurants in both areas. The implication is that we can provide restaurants with a series of suggestions.

In the data processing part, we take the rating stars as the Y variable and all the 23 columns such as the polarity of the comments, opening hours (weekday, weekend), category, reviews counts and some environmental factors like if dogs are allowed or if there is a parking lot and so on . as X variables. We scale the review counts, opening hours (average, weekdays, weekend) in training datasets, and testing datasets. For the missing values, we realized ‘neighborhood’ and ‘special food offered’ contain the missing values. As for the neighborhood, it contains more than 60% of the missing values. Thus we drop neighborhood. We use mode to tackle the missing values in the ‘special foods’ column. We also labeled the ‘Stars’, which is our predictive variable and divided by Stars=3.5, get the results of ‘good’ and ’bad’ by looking at the distributions of the values.

For the modeling part, we use SVM’ kernel function, which is SVR at here, but the out-of-sample f1-score is only around 0.53. Thus we tried Gradient Boosting, Xgboost, Light Xgboost. For the reason that Xgboost is a regularized model formalization to control over-fitting, but it also has a higher computing rate. The algorithms of gradient boosting fits the new model to new residuals of the previous prediction and minimizes the loss when adding the latest prediction.

Furthermore, the light Xgboost can produce more rigorous trees by following leaf wise split approach rather than only the main factor in achieving higher accuracy normally.

However, we did not see a very high performance with these models. Then we tried random forest classifiers, and we achieved out of sample f1-score 0.76 for the blue area and 0.68 for the red areas by using the 6-fold cross-validation and GridSearchCV to do the hyperparameter tuning (Figure 5-3). However, with the limitation of computing power, we did not use the gen-stack to aggregate all the models.

Based on the feature importance graph of random forest classifications, we realized that for both areas, polarity of the reviews counts; average opening hours; review counts; weekday opening hours; weekend opening hours; category of the restaurants, dogs allowed or not; parking places have significant impacts on the stars rating on Yelp. However, the factors are slightly different between the red clusters and blue clusters(Figure 5-1), category places plays a significant role in red clusters and opening hours on weekdays is more critical for blue area (Figure 5-2), which may be because that the blue area has higher population density, the longer it opens, the higher income(stars) will it be. Comparably, the red cluster, which stands for the rural areas, the income (stars) will be highly influenced by local preference which is the category of the restaurants.

Conclusion

Given the fact that there is undeniable boosting in online reviews, our paper focused on different factors' impact on yelp rating.

Our findings provide strategic suggestions to business owners. Business owners in the commercial area should consider extending their weekday operation time to generate higher sales. Hence, more revenue. Future business owners in a residential area have to explore on which segment they are willing to enter.

Our finding indicates that the Yelp rating is more affected by the review's sentimental. In other words, the reviewer's language tone matters. It matches Chevalier and Mayzlin's study that the effectiveness of one-star review surpasses the effectiveness of five-star review. Moreover, the more reviews, the more likely a business has a higher rating. Average operating hours also have a high significance on businesses' ratings. It may suggest that businesses adjust their business hours to obtain a better rating.

Although we do not have business revenue data, we assume based on Luca's finding that the higher the rate, the more revenue. Nevertheless, it is still practical to merge with business revenue to reveal the philosophy between yelp rating and business growth.

Reference

C. Dellarocas. The digitization of word of mouth: promise and challenges of online feedback mech- anisms. Management Science, 49(10):1407–1424, 2003.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345-354.

Chen, P. Y., Wu, S. Y., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, 58.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision support systems*, *45*(4), 1007-1016.

Funk, T. (2009). Web 2.0 and beyond: Understanding the new online business models, trends, and technologies. Westport, CO: Praeger.

Hennig-Thurau, T., Gwinner, K., Walsh, G., & Gremler, D. (2004). Electronic word-of-mouth via consumer opinions platforms: What motivates consumers to articulate themselves on the Internet? Journal of Interactive Marketing, 18(1), 38-52.

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly*, *34*(1), 185-200.

Sweeney, J. C., Soutar, G. N., & Mazzarol, T. (2008). Factors influencing word of mouth effectiveness: receiver perspectives. *European journal of marketing*, *42*(3/4), 344-364.

Tucker, T. (2011). Online word of mouth: characteristics of Yelp. com reviews. *Elon Journal of Undergraduate Research in Communications*, *2*(1), 37-42.

Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenization tools-an analysis. Advanced Computational Intelligence: An International Journal (ACII), 3(1), 37-47.

Wickford, H. (2018, April 13). The Average LifeSpan of a Restaurant. Retrieved December 15, 2019, from https://yourbusiness.azcentral.com/average-life-span-restaurant-6024.html.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, *74*(2), 133-148.

Appendix 1:

Data description:

Yelp’s business attribute data set is a tremendous dataset contains comprehensive information about a restaurant. Indexing by unique business ids, each restaurant contains over 100 features. Indeed, a huge amount of missing data is a common side impact. Thus, We generate some new features to avoid a low efficient number in each feature.

PARKING is the number of parking options the restaurant could apply to the consumer. It includes garage, street parking, parking lot, parking validated and valet parking.

CUSTOMER is the service that could provide better customer satisfaction based on extra service a restaurant applies. Excludes 0 entries, we obtain good for kids, wheelchair accessible, bike parking, coat check, cater, good for dancing, outdoor seating, and takeout delivery services in our customer feature.

RECOMMENDATION shows consumer reviews and their recommendation to the restaurant. In the yelp app, you may find a restaurant is good for lunch. In Recommendation, we merged the Best nights for different days in a week and which scenario is the restaurant good for.

SPECIAL FOOD indicates if the restaurant serves specific food with dietary restrictions such as gluten-free or soy-free. Special food is generated from dairy-free, gluten-free, vegan, kosher, halal, soy-free and vegetarian.

WEEKDAY indicates how many hours the restaurant is open during the weekdays.

WEEKEND indicates how many hours the restaurant is open during the weekends.

Other features come from the original data set remain the same.

Appendix 2.0 (Tables and graph)

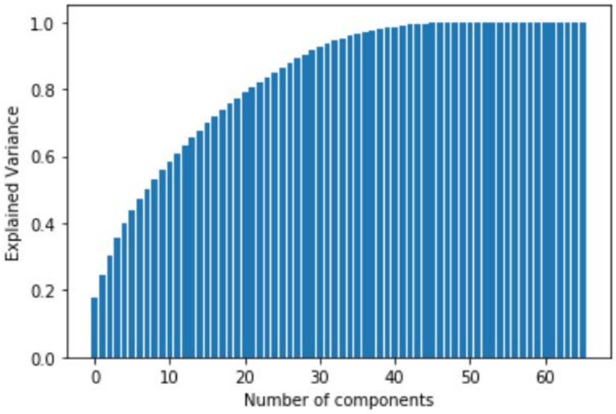


Figure 4-1

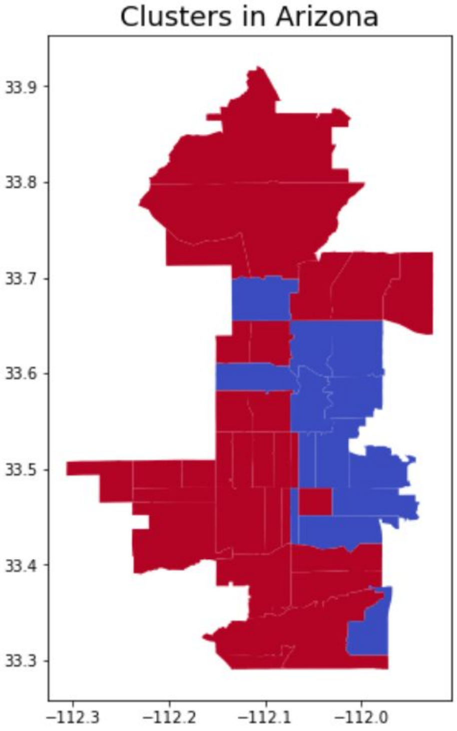


Figure 4-2

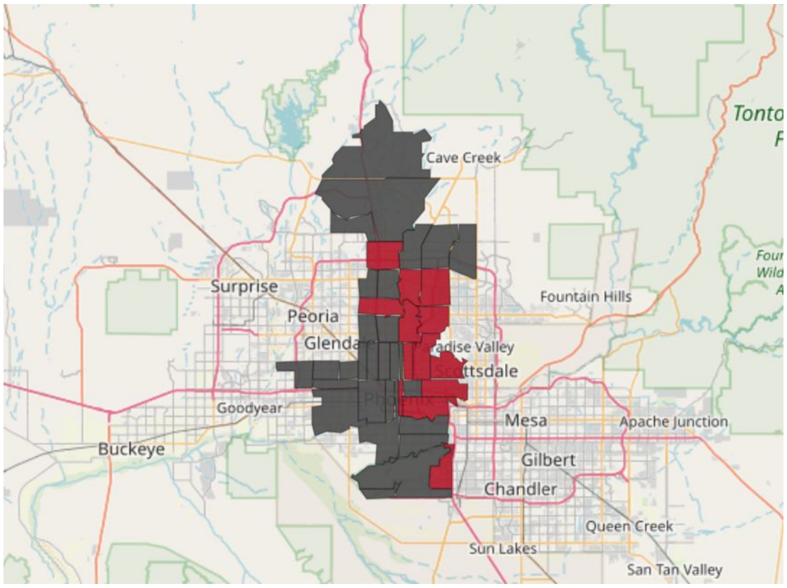
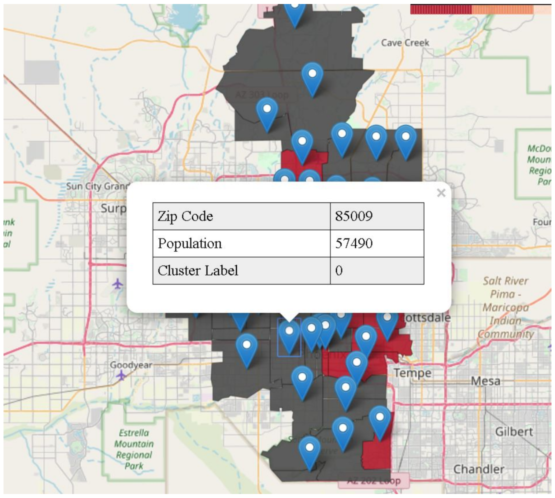


Figure 4-3



Zip Code	85009
Population	57490
Cluster Label	0

Cave Creek
Sun City Grand
Surp
ottsdale
Salt River Pima - Maricopa Indian Community
Tempe
Mesa
Goodyear
Estrella Mountain Regional Park
Gilbert
Chandler

Figure 4-4

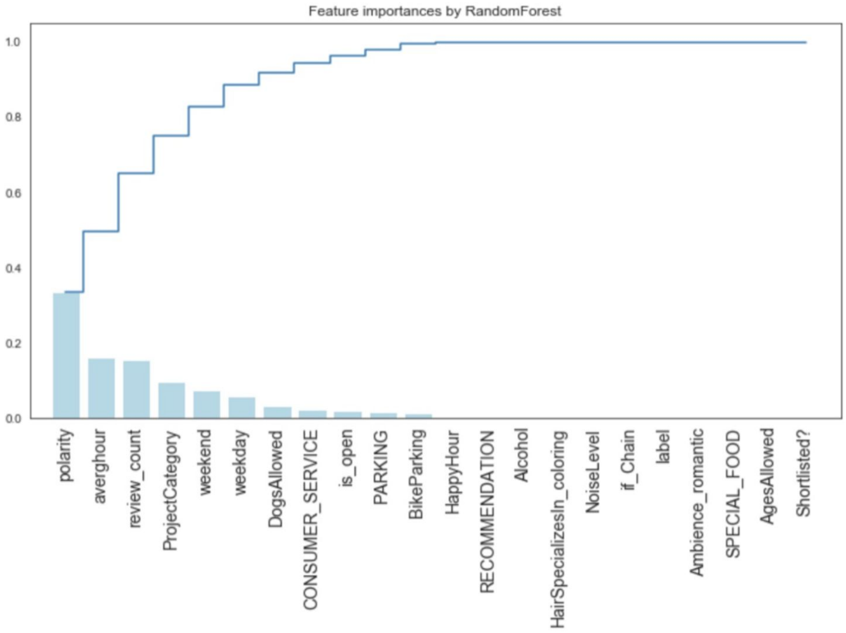

Feature importances by RandomForest
1.0
0.8
0.6
0.4
0.2
0.0
polarity
averghour
review_count
ProjectCategory
weekend
weekday
DogsAllowed
CONSUMER_SERVICE
is_open
PARKING
BikeParking
HappyHour
RECOMMENDATION
Alcohol
HairSpecializesIn_coloring
NoiseLevel
if_Chain
label
Ambience_romantic
SPECIAL_FOOD
AgesAllowed
Shortlisted?

Figure 5-1

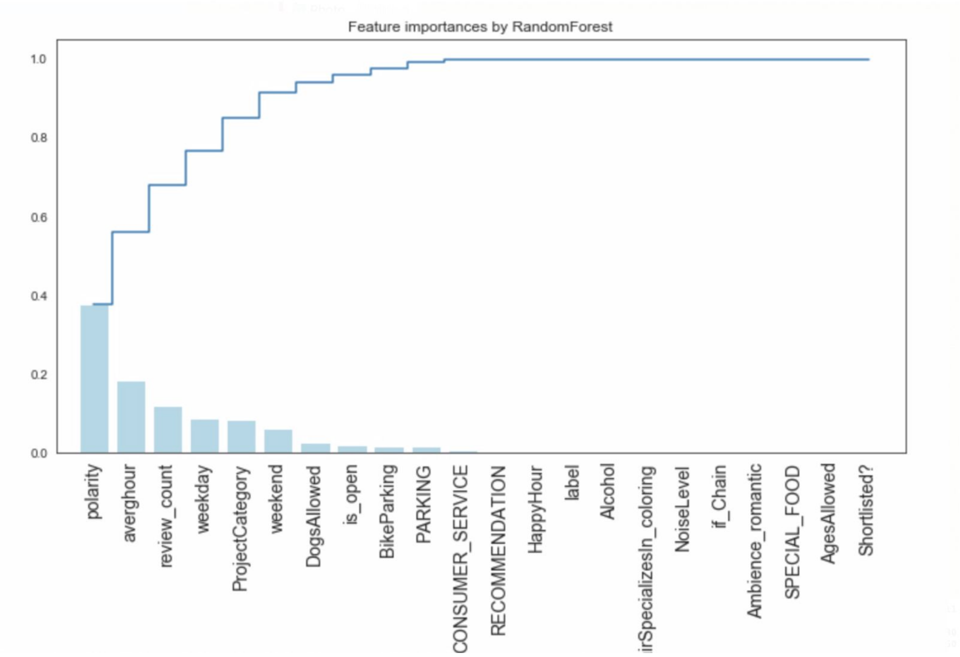
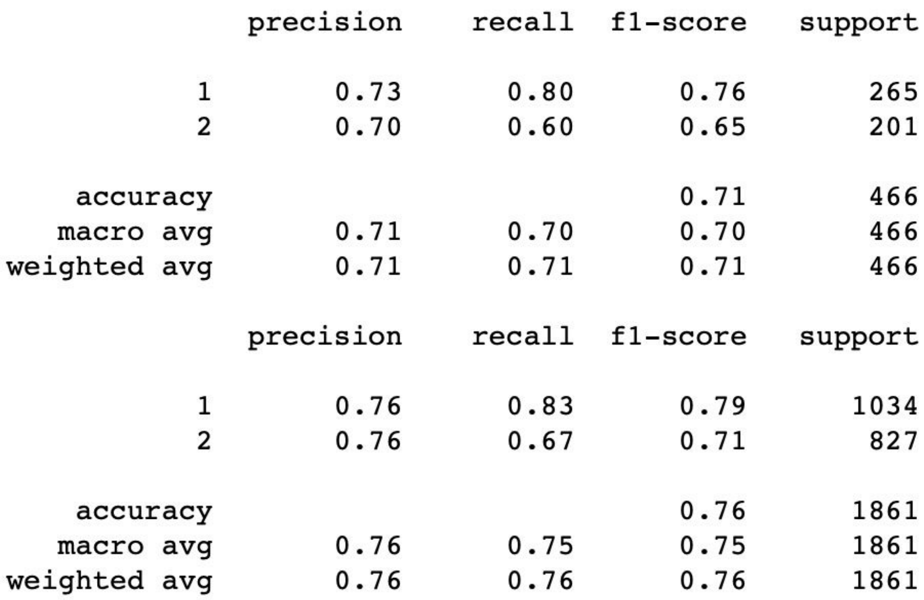

Feature importances by RandomForest
1.0
0.8
0.6
0.4
0.2
0.0
polarity
averghour
review_count
weekday
ProjectCategory
weekend
DogsAllowed
is_open
BikeParking
PARKING
CONSUMER_SERVICE
RECOMMENDATION
HappyHour
label
Alcohol
irSpecializesIn_coloring
NoiseLevel
if_Chain
Ambience_romantic
SPECIAL_FOOD
AgesAllowed
Shortlisted?

Figure 5-2



```
              precision    recall  f1-score   support

           1       0.73      0.80      0.76       265
           2       0.70      0.60      0.65       201

    accuracy                           0.71       466
   macro avg       0.71      0.70      0.70       466
weighted avg       0.71      0.71      0.71       466
              precision    recall  f1-score   support

           1       0.76      0.83      0.79      1034
           2       0.76      0.67      0.71       827

    accuracy                           0.76      1861
   macro avg       0.76      0.75      0.75      1861
weighted avg       0.76      0.76      0.76      1861
```

Figure 5-3

Contribution

Introduction: Guilherme Louzada & Yutong Zhu

Data Cleaning: Guilherme Louzada & Yutong Zhu

Literature review: Yutong Zhu

EDA:Aparna Bhutani

Cluster: Zheyuan Zhang

NLP: Aparna Bhutani, Di Xin

Classification: Di Xin, Zheyuan Zhang

Conclusion: Yutong Zhu

PPT: Guilherme Louzada

Github: All