Generalized Customer Satisfaction Prediction

Li Liu

5/21/2019

Abstract

Predicting customer satisfaction is vital for businesses to understand the need and improve services. Besides from surveys and focus groups, the Yelp data provides a measure of satisfaction level by reviews and rating stars. However, most of the research and projects on Yelp data couldn't be applied into practice as only a small fraction of the customers write reviews. Using the restaurants in Las Vegas as an example, this paper would provide a generalized machine learning framework for businesses to predict the satisfaction of all customers.

1 Data

Yelp dataset (Yelp, 2019) enables us to study customer satisfaction through the online reviewers. It is available online (https://www.yelp.com/dataset) for academic learning purposes. The dataset contains 7 files in JSON format, which are business, users, review, business hours, business attributes, tip, and checkin. The size of the data is around 5 GB after we converted it to the spreadsheet format. In our analysis, we focus on exploring the first three files.

1.1 Business

The business data section contains 13 features of 174567 businesses from different regions, such as Las Vegas, Phoenix, Toronto, Charlotte, and Scottsdale. We selected the 3990 restaurants in Las Vegas which were still labeled as "open". The assumption is that tourists in Las Vegas are unfamiliar with the qualities of vast varieties of restaurants. Thus, they rely more on review forums (such as Yelp) to inform decisions. At the same time, the restaurants know very little about the demand of the new customers and their satisfaction levels. Among the feature columns, the most informative ones are location, review stars, review counts, and categories. Figure 1 shows the busiest restaurants are located on the major tourist attractions area around the South Las Vegas Blvd and Flamingo Rd. In the categories column, each restaurant has several tags to highlight their services. Figure 2 is a word cloud visualization of the large varieties of restaurant services in Las Vegas.

1.2 User

There are 1326100 users in the original dataset. 22 columns provide information for researchers to classify and cluster the users, such as name, review counts, date of Yelp account, number of friends, compliments received, tags (useful, funny, cool, elite, etc). Although this seems like a large sample of customers, only a very small fraction of the customers would leave a review after dining at the restaurants. The customers using Yelp could also potentially be a non-representative group of the total customers. In our analysis, we want our predictable models to be generalizable for predicting the satisfaction of customers not using Yelp. Thus, we will only keep the relatively objective customers information.

1.3 Review

Exploring the latent topics and sentiments has been the focus of most of the natural language processing and machine learning research projects using Yelp data. The review data has around 5 million review entries, with the date, stars, tags (useful, funny, etc), ids for review, user, and business.

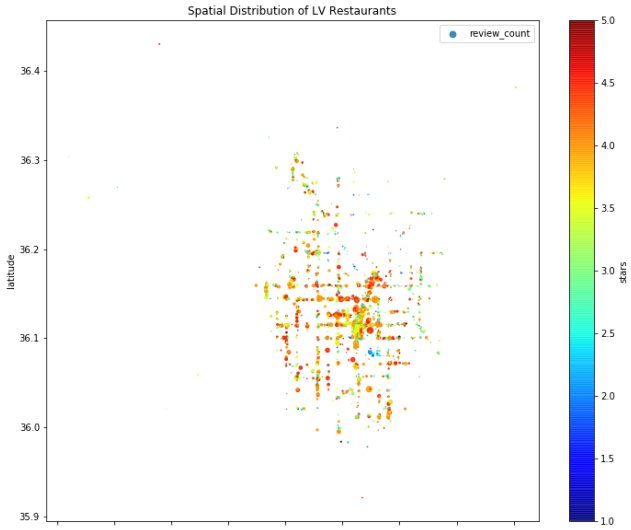

Figure 1: Dot size is the review counts and color is the stars

Figure 2: Grouping the restaurant from categories keywords

1.4 Merging Tables

We merged the business (only restaurants in LV) and review data by left joining on the business-id. Then we merged it with user data by left joining on user-id. The aggregated data has 795588 observations, where each line is a review summarizing the dining experience at a particular restaurant by a certain customer.

2 Customer Satisfaction

We are interested in predicting the customers satisfaction of their dining experience. Traditionally, companies would use survey and focus groups to gather such information. Yelp review rating is a new metric for the customer satisfaction using a 1-5 scale. In Figure 3, we find that roughly 67% of the rating stars are 4 or 5, while the remainings are below 4. One possible challenge is the heterogeneity of the customer satisfaction given the same products and services. As a result, a coarse grading would be more straightforward. We classified the satisfaction dummy variable as 1 ("good") if rating stars are 4 or 5, and 0 ("bad) if they are below 4 (See Figure 4).

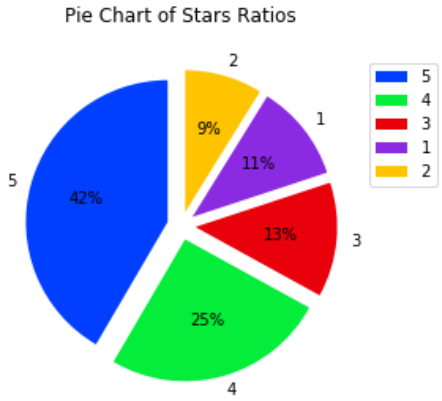

Figure 3: Rating Stars: 1 is the worst and 5 is the best

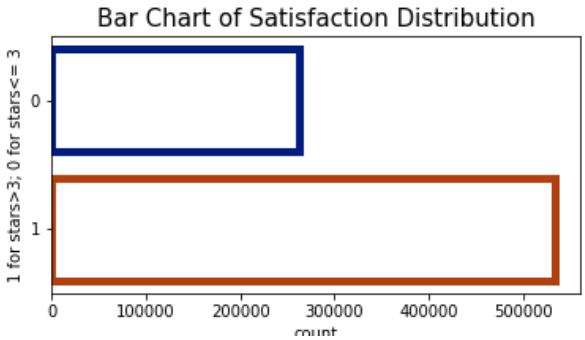

Figure 4: The ratio of group 1 and 0 is 2:1

3 Features Engineering

3.1 Restaurants stars and counts

When customers search for the restaurants, the current restaurant average stars and counts provide customers with prior belief of the service quality (See Figure 5). However, there is no strong correlation between the two variables (See Figure 6). The restaurants with most reviews are not necessarily the best ones.

	review_count	stars
count	174567.000000	174567.000000
mean	30.137059	3.632196
std	98.208174	1.003739
min	3.000000	1.000000
25%	4.000000	3.000000
50%	8.000000	3.500000
75%	23.000000	4.500000
max	7361.000000	5.000000

Figure 5: Statistics Table for Review Counts and Business Stars

The dates of the reviews range across a decade. There could be changes in platform and algorithms that lead to the inconsistency of the data. We plotted the average restaurant ratings and customers ratings by quarters (See Figure 7). Starting from 2005, we noticed there are discrepancies in the ratings at the early period. Then the two lines intertwined with each other and increased steadily starting from 2010.

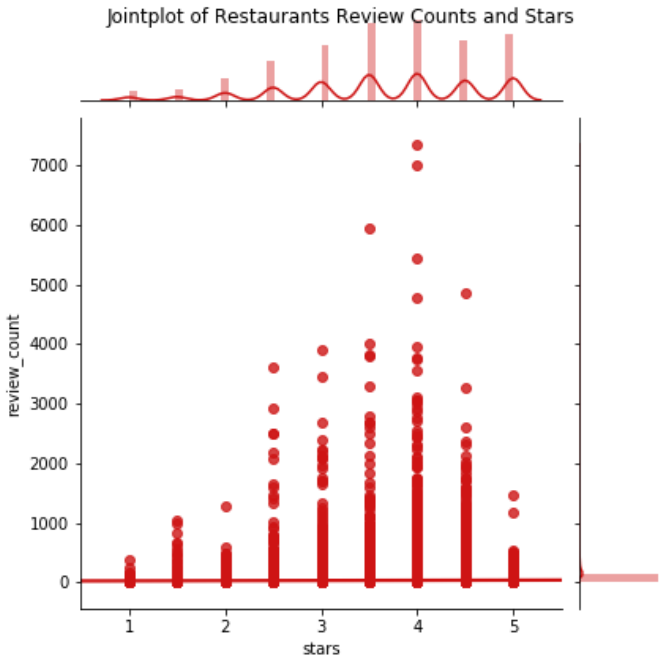

Figure 6: The correlation of review count and stars is -0.13

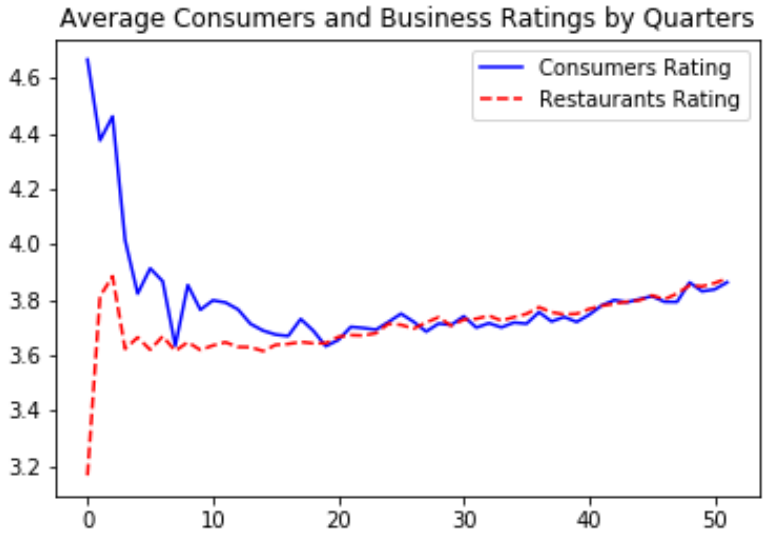

Figure 7: The Business and Consumers Ratings are increasing since 2010

3.2 Review Sentiment

We applied the VADER algorithm (Valence Aware Dictionary and sEntiment Reasone) to infer the sentiment level in the review, which is a a lexicon and rule-based sentiment analysis tool trained with social media data (Hutto & Gilbert, 2014). The final score is a metric for magnitude of the sentiment intensity normalized between -1 and 1. Table 1 shows how the score corresponds to the sample restaurant reviews.

Figure 8 is the histogram of the sentiment scores of all reviews. The distribution is left-skewed and the average is 0.63 with standard deviation 0.56. Figure 9 separates the scores by the satisfaction variable, where the left one are sentiment scores associated with satisfied customers. The unsatisfied customers have more negative scores than the satisfied ones.

Sentence	Sentiment
"amazing food"	0.59
"terrible service"	-0.48
"convenient and fast food"	0.00

Table 1: Sentiments Example Table.

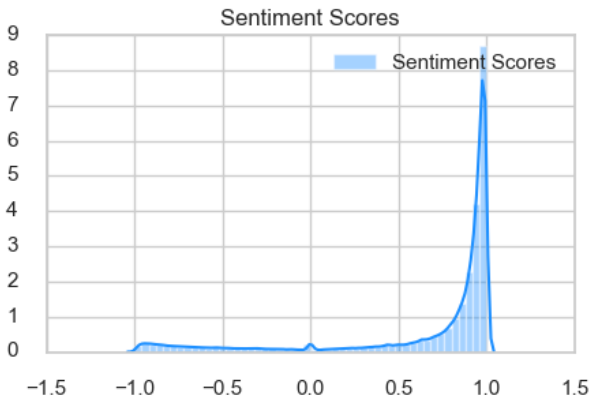

Figure 8:

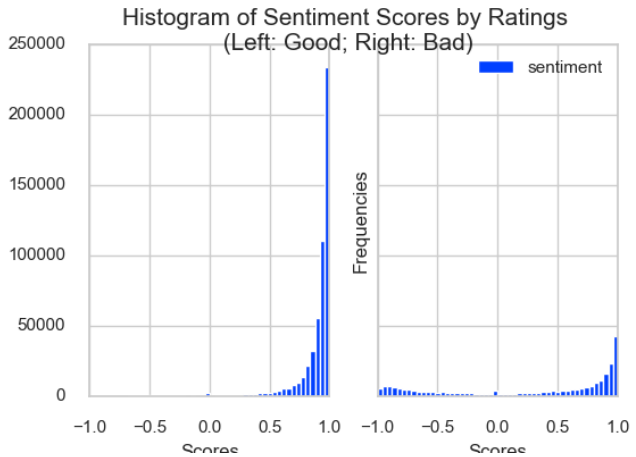

Figure 9:

3.3 Gender

Gender is an objective information that restaurants would usually ask the customers in the questionnaire. There is no variable on gender in the original data. We applied the gender-guesser package (PyPI, 2016) to guess gender from customers' first names (assuming they are the same as the user names). We labelled 43% of the customers as females. Figure 10 and Figure 11 shows that the female ratio in satisfied group is slightly higher than the one in unsatisfied group.

sat	femalects	malects	gender
0.0	109916.0	152602.0	0.42
1.0	234753.0	298317.0	0.44

Figure 10: Gender Ratios Table

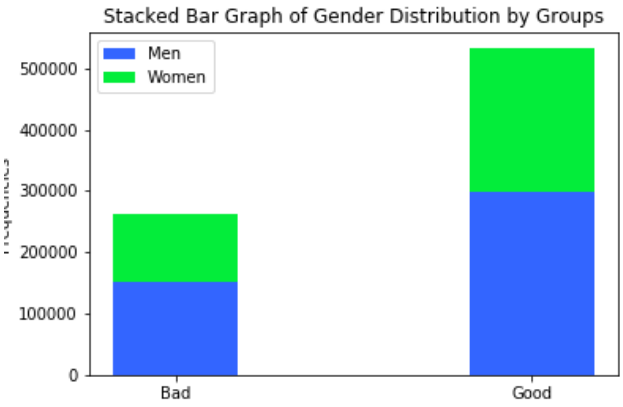

Figure 11: The gender ratios in two groups are close

3.4 Weekend

The visit date of the week could potentially influence customers' expectation of restaurants' services. For example, customers might know Friday, Saturday, and Sunday are busy periods and they would have a higher tolerance for waiting time. While on weekends, they might want the food served immediately and better service. Using the raw date of the review, we tagged each row with a dummy variable of whether a visit is on weekends (Friday to Sunday) or not (Monday to Thursday). Figure 12 illustrates that for both groups, 5 stars are the most frequent while 2 stars are the least frequent.

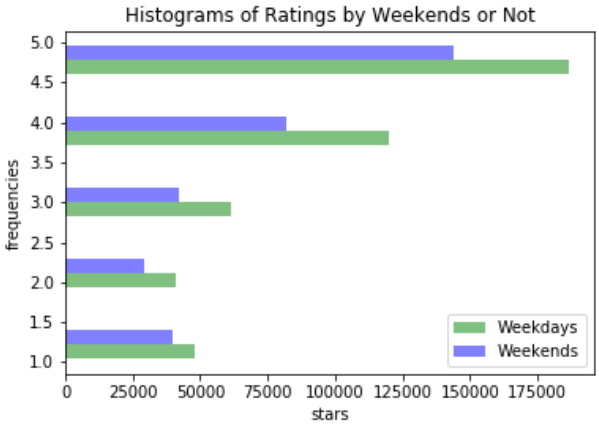

Figure 12: The gender ratios in two groups are close

3.5 Groups

Having accompanies or not might change customers' expectation. The hypothesis is that they are more likely to be satisfied if they are alone, compared with dining with other people. There are no information on the number of people per group visit in the data. This information is objective and could be observed directly by the restaurants. We leveraged the review text to infer whether the reviewer is with other people. Firstly, we defined keyword corpus such as "we", "our", "husband", "wife", "friends", "colleagues", etc. Secondly, we looped through each tokenized review text to check if it contains the keywords. Figure 13 shows that the observations tagged with individuals are about three times larger than the ones with groups. The average rating of individual segment is 3.58 (with standard deviation 1.45) is slightly greater than the one for the groups segment (mean 3.53 and standard deviation 1.46).

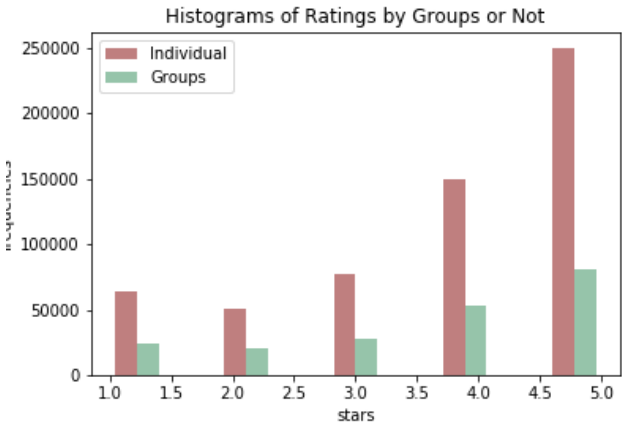

Figure 13: The gender ratios in two groups are close

3.6 Food Categories

We compiled a corpus to group the restaurants categories into cuisine types, such as American, European, South American, Asian, Middle East, and Drinks Bars. Figure 14 shows 41% of the restaurants are American food, such as steakhouse, pizza, and burgers. The Asian restaurants and Drinks bars account for 19 % and 18% of the total businesses respectively. We tagged the observations with whether they belong to these six categories or not.

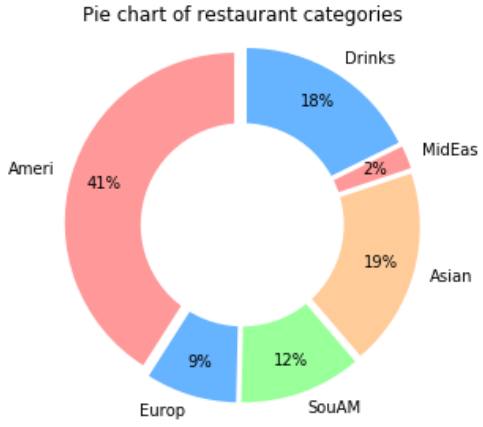

Figure 14: Grouping the restaurant from categories keywords

4 Machine Learning

Random forest measures the relative importance of each predictor by counting the number of associated training samples in the 100 trees. After first iteration, we found the importance scores for food categories are close to 0. So we put them aside and proceeded with 6 features. Figure 15 and Figure 16 displays the features ranked by the predictive importance. It seems that sentiment is very predictive of the satisfaction variable.

4.1 Model Training

We separated the data into training (75 %)and testing (25%) sets. We deployed six machine learning algorithms for classification problem: logistic regression, decision tree, random forest, support vector machine, K-nearest neighbors, and multi-layer perceptrons. For each model fitting, we further tuned the hyperparameters by randomized search method, using the "roc auc" (area under the receiver operating characteristic curve) as the scoring metric.

features	score
sentiment	0.74
stars_x	0.11
review_count_x	0.11
weekend	0.01
gender	0.01
group	0.01

Figure 15: Tables of Features Importance by Random Forest

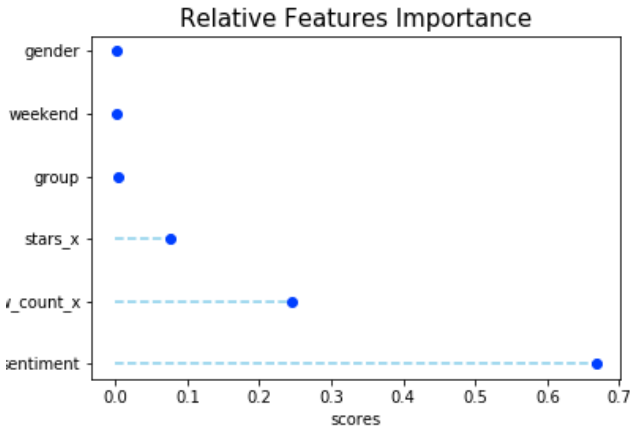

Figure 16: Sentiment is the most predictive variable

4.2 Model Evaluation

For each optimized model, we evaluated its performance by 5-fold cross-validation on the training set. In Figure 17, these six models have very similar performance in CV scores. The precision measures how many times the predictions are correct when the model predicts a customer's satisfaction variable. The decision tree and KNN models have the highest precision of 79%. The recall measures how many customers who have good experience are correctly identified. MLP model identifies 95% of the customers who are satisfied. It also has only half of the false negatives compared with the others. Decision tree achieves the highest AUC score. Random forest has the largest F1 score.

No	Model	CV Scores	True Negatives	False Positives	False Negatives	True Positives	Percision	Recall	F1-Score	AUC
1	Logistic	0.85	1817	1189	445	4049	0.77	0.9	0.83	0.75
2	Decision Tree	0.85	1965	1041	529	3965	0.79	0.88	0.83	0.77
3	Random Forest	0.85	1809	1197	406	4088	0.77	0.91	0.84	0.76
4	SVM	0.85	1786	1220	400	4094	0.77	0.91	0.83	0.75
5	KNN	0.84	1918	1088	503	3991	0.79	0.89	0.83	0.76
6	MLP	0.85	1327	1679	236	4258	0.72	0.95	0.82	0.69

Figure 17: Model Comparison

The accuracy of the fitted logistic regression on the test data is 0.79. Further analysis would try new models (ensemble and deep learning) and experiment with different features.

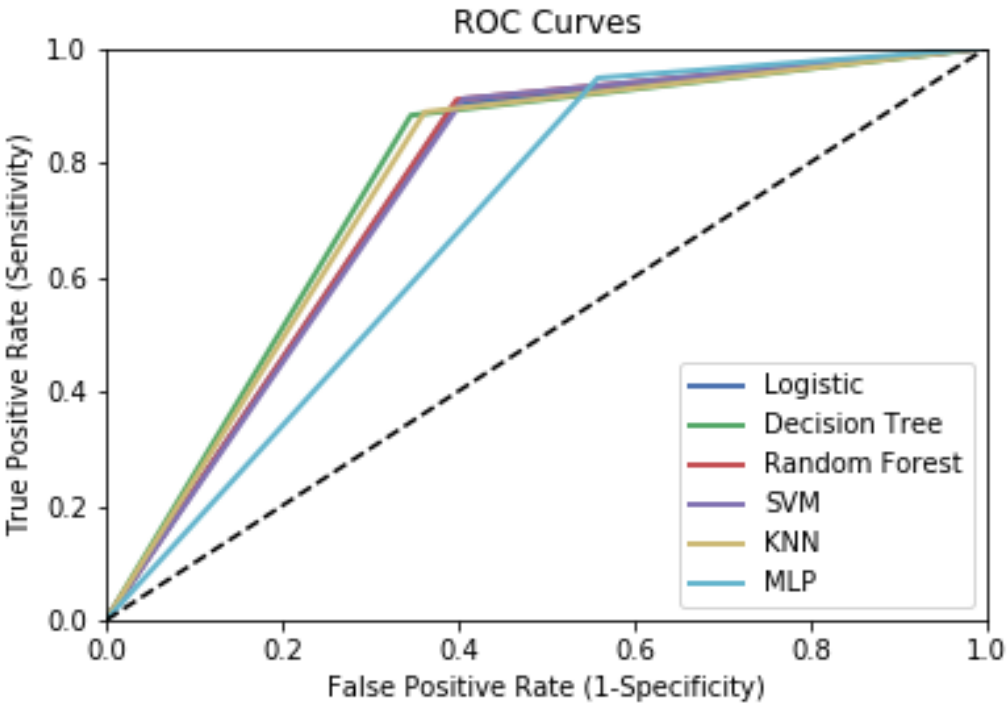

Figure 18: Performance Visualization

The following discussion would be on how to make the predictive model generalizable for estimating the satisfaction levels of customers who don't use Yelp, such as estimating sentiment scores from the amount of tips.

References

Hutto, C., & Gilbert, E. (2014, June). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Ann Arbor, MI.

PyPI. (2016). *gender-guesser 0.4.0.* Retrieved 2019-05-21, from https://pypi.org/project/gender-guesser/

Yelp. (2019). *Yelp Dataset.* Retrieved 2019-04-27TZ, from https://www.yelp.com/dataset/challenge