Modeling Customer Satisfaction from Yelp Data

Li Liu*

6/10/2019

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 shows a generalized machine learning framework is better for businesses to predict the satisfaction of all customers.

JEL classification: M31, C53, L83

Keywords: Online Ratings, Customer Satisfaction, Quantitative Marketing

*M.A. Student in Computational Social Science, The University of Chicago
Email: liu431@uchicago.edu
I would like to thank Richard Evans and my classmates for their suggestions. All remaining errors are mine.
Code and replication files are available at Github.

1 Introduction

Identifying customers' satisfaction is crucial for firms' marketing research. Surveys, interviews, focus groups are examples of traditional methods to measure satisfaction. In the digital age, customers interact directly with firms and other customers through channels such as e-commerce websites and review forums. This research project focuses on understanding the factors that affect customers' satisfaction of the dining experience. More specifically, we use the large-scale Yelp social media data and machine learning approach to find the optimal model for predicting customer satisfaction.

Topics and sentiments extracted from the reviews are good measures of the subjective and diverse customers experience. This area of research has been popular in the natural language processing and data mining communities. However, businesses couldn't adopt the findings for predicting consumer satisfaction as only a small fraction of the customers write reviews. The ground truth is that most of the customers would only leave reviews when they have strong opinion to share or the businesses incentive them to do so. As a result, this project constructs features from the Yelp data that are generally observable for all the customers, such as gender, date of the visit, with accompanies or not, etc. We also add features that are business attributes, such as aggregated review stars and counts, and food categories. Then we compared the multiple linear regression and eight machine learning algorithms for predicting the satisfaction scores.

2 Literature Review

Yelp provides a large data-set to academic researchers who are interested in mining and analyzing the undiscovered insights from the business data, user meta-data, and user-generated content (text reviews and uploaded photos) (Yelp, 2019) . Many research communities have explored this data to conduct empirical research and improve machine learning algorithms. We surveyed some literature that provide the foundation for our research.

2.1 Text as Data for Social Scientists

The advances in natural language processing, combined with traditional data and models, have many applications in the business and marketing research. By extracting information from documents to serve as new input variables, text are new sources of data for social scientists (Gentzkow, Kelly, & Taddy, 2019).

Text reviews, as one common form of the user-generated content, is valuable for capturing marketing insights and creating values for both businesses and customers (Balducci & Marinova, 2018). User-generated content analyzed with machine learning methods is a better alternative to identify customer needs, compared with interviews and focus groups (Timoshenko & Hauser, 2019). Their results show promising potential for us to identify the factors that influences customers' dining experience from the machine learning approach.

2.2 Yelp Data in Economics

Researchers have used Yelp data to measure economic and business activities. Michael Luca at Harvard Business School estimated the effect of online reviews on restaurant demand by regression discontinuity method. Besides from the Yelp review data, he aggregated it with the restaurants revenue data from the Washington State Department of Revenue (Luca, 2011). Thus, he operationalized reputation by rating stars and consumer demand by reported revenue. This suggests adding external data source would be useful. Another example is the government survey (e.x. County Business Patterns (CBP) offered by Census Bureau) could be greatly complemented by the new economic activities metric predicted from Yelp data (Glaeser, Kim, & Luca, 2017). For our project, we thought the data on customers' tip amount would be interesting argument to quantify customers' satisfaction. However, we didn't find such business proprietary data that could match with the Yelp data. In addition, the micro-level data by the Current Population Survey offers a wide range of information on neighborhoods' demographics. This might help us infer consumers' possible demand for food and willingness to pay. For instance, a mediocre Asian restaurant might receive good ratings in the suburban areas as people don't have many options for exotic food. By contrast, diners in the downtown area of Chicago have higher expectation of food quality. We plan to use the CPS data when we extend the project beyond one region.

Given the unobservable heterogeneity of reviewers, aggregating the rating scores by taking the arithmetic average has potential issues for missing the underlying changes in product quality and prior knowledge (ex. learning from previous reviews and scores). Thus, an adjusted weighted average algorithm is necessary for aggregating the individual scores by accounting for the reviewers' characteristics and review histories (Dai, Jin, Lee, & Luca, 2012)

3 Data

Yelp data enables us to study customer satisfaction through the online reviewers (Yelp, 2019) . It is available online (https://www.yelp.com/dataset) for academic learning and research purposes. The dataset contains 7 files in JSON format, which are business, users, review, business hours, business attributes, tip, and check-in. 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.

3.1 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 are labeled as "open". The assumption is that tourists in Las Vegas are unfamiliar with the qualities of the 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 attributes for restaurants, the most informative ones are location, average 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. Each restaurant has several tags to classify their services. Figure 2 is a word cloud visualization of the large varieties of restaurant services in Las Vegas.

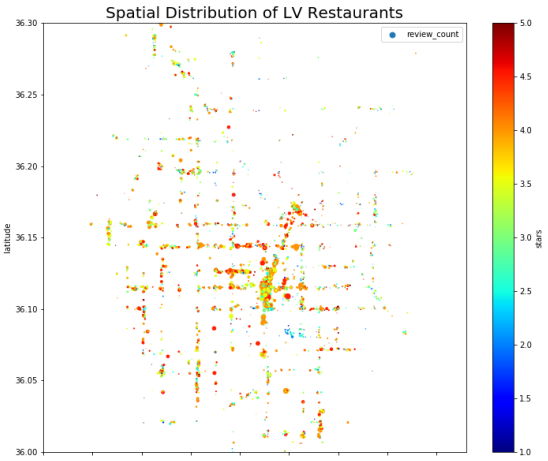

Figure 1: Dot size represents the review counts and color indicates the average star

3.2 User

There are 1326100 users in the original data. 22 columns provide information for researchers to classify and cluster the users, such as name, review counts, date of account creation, 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 online. 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 information about customers.

Figure 2: Grouping the restaurant from categories keywords

3.3 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.

3.4 Merging Tables

We merged the business (only open 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.

4 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 4, we find that roughly 67% of the rating stars are 4 or 5, while the remaining one-thirds are below 4. One problem exists for review ratings (also prevalent for survey research) is people have different standards for numerical scaling. The current ratings are positively correlated with the past ratings (See Figure 3). In order to control for such heterogeneity of the customer satisfaction, we reweighed the current ratings to 0-100 scale by accounting for customers' average ratings for past reviews.

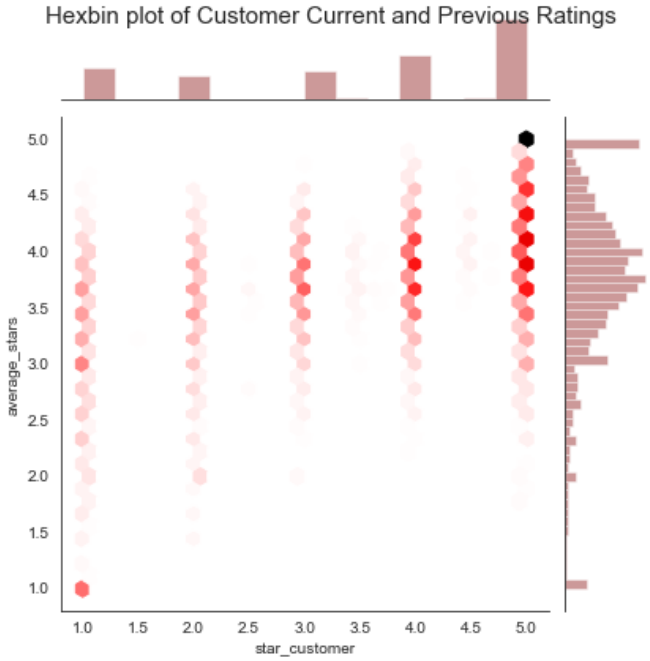

Figure 3: Rating Stars: 1 is the worst and 5 is the best

Suppose the current rating for current restaurant is $r1$ and the previous average rating is $r0$, then we estimated the re-weighted rating $\hat{r1}$ as $r1 + \frac{r1-r0}{r0}$. The minimum, average, and maximum for $\hat{r1}$ are 0.22, 3.43, 6.84, respectively. We constructed the customer satisfaction variable y by converting $\hat{r1}$ to 0 - 100 scale according to the formula $\frac{100-0}{max_{\hat{r1}} - min_{\hat{r1}}} * (\hat{r1} - min\hat{r1}) + 0$. For example, $r1 = 5$ and $r0 = 1.2$ corresponds to $y = 100$ (very satisfied). On the contrary, $r1 = 1$ and $r0 = 4.9$ corresponds to $y = 0$ (very unsatisfied). Figure 5) shows the histogram of new statisfaction socres.

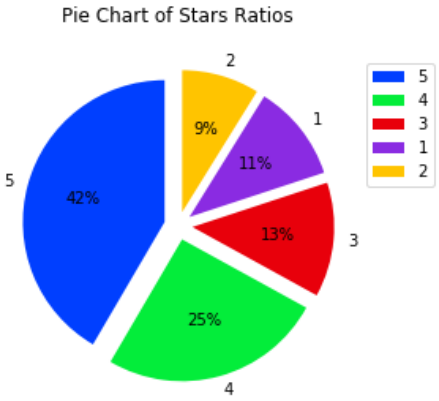

Figure 4: Rating Stars: 1 is the worst and 5 is the best

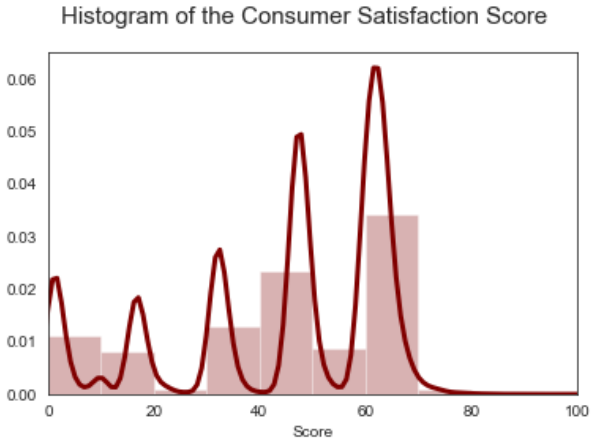

Figure 5: New customer satisfaction variable is left-skewed

5 Independent Variables

We are interested in finding the possible determinants of the customer satisfaction function. There are dozens of feature variables available in the data, but not all are relevant for our project. The principal criteria for selecting the variable is whether it is available information for customers, no matter they write Yelp reviews or not. As a result, we didn't use the review text directly as the input to the natural language processing models. Instead, we extracted relevant objective information for the data. There are two categories of the independent variables: information about the restaurants (reputation and categories) and information about the individual customers (weekend visit or not from the review-created date, groups visit or individual from the pronouns, gender from the name, etc).

5.1 Reputation

When customers search for the restaurants, the current restaurant average stars and counts provide customers with a measure of reputation and have a lot of variations (See Figure 6). However, there is no strong correlation between the two variables (See Figure 7). The restaurants with most reviews are not necessarily the best ones, and vice versa.

	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 6: Statistics Table for Review Counts and Business Stars

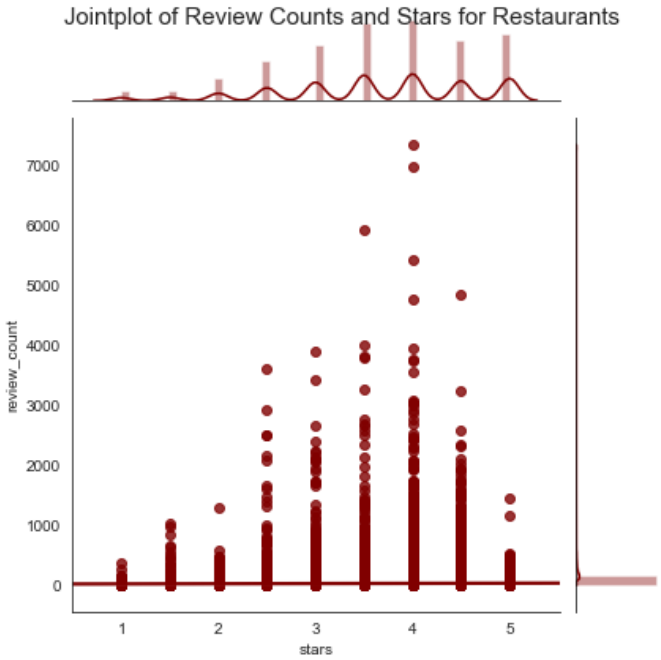

Figure 7: The correlation of review count and stars is -0.13

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 8). 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. We detrended the ratings by accounting for the time factor.

5.2 Categories

Categories (ex. cuisine type) influence people's expectation for the quality and service. Each restaurant have multiple self-chosen tags to categorize their business function. However, multiple tags are not useful information for the predictive model. So we compiled keyword lists (See Appendix B for the word lists and clouds) to cluster restaurants into six cuisine types: American, European, South American, Asian, Middle East, and Drinks. Then we we created dummy variables for indicating whether the restaurant belongs to certain group or not. Figure 9 shows that American restaurants have the highest market share (41%) while only 2% restaurants serve Middle Eastern food. Figure shows the Gaussian kernel densities of the customer satisfaction for each category of the restaurants.

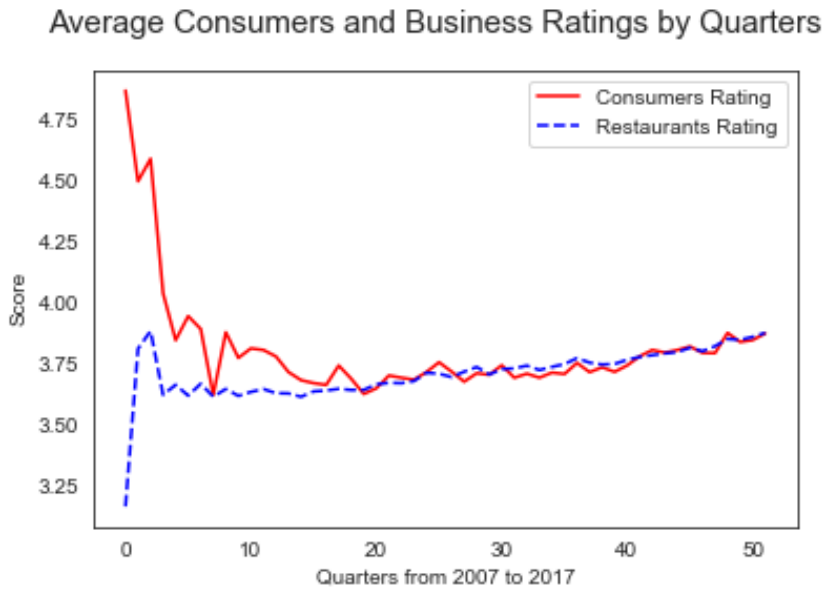

Figure 8: The Business and Consumers Ratings are increasing since 2010

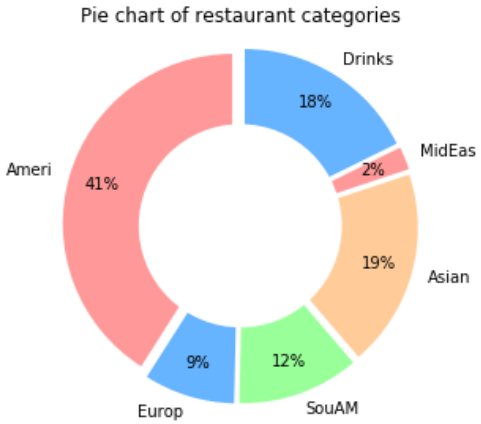

Figure 9: Grouping the restaurant from categories keywords

5.3 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 11 shows that the female group has very similar estimated satisfaction density with the male group.

5.4 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 prefer the shorter waiting time and better service.

Using the raw date of the review, we tagged each row with a dummy variable of whether a visit is

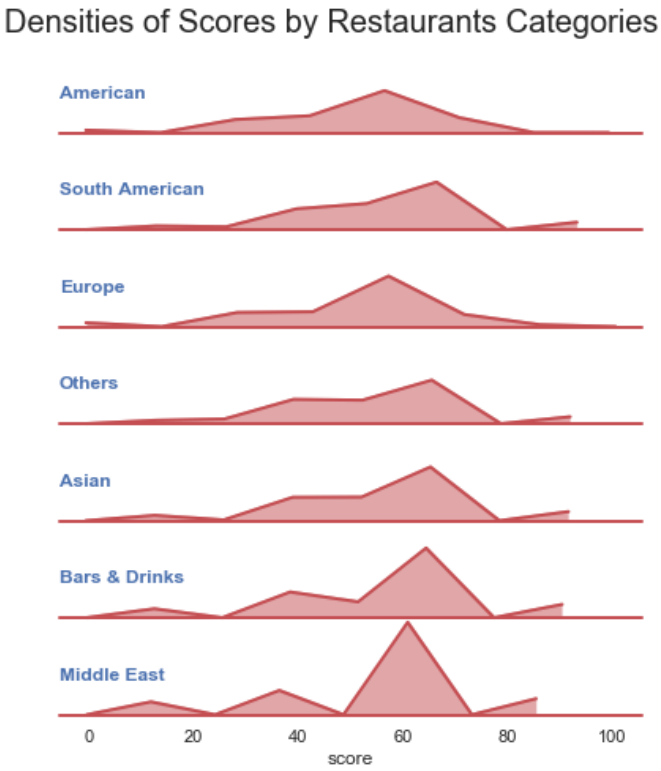

Figure 10: Satisfaction varies with the categories

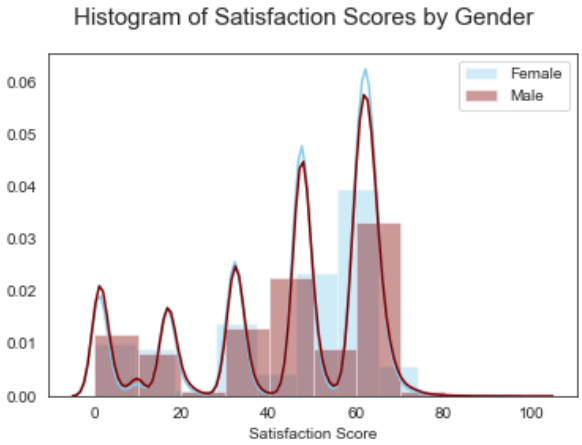

Figure 11: Female groups have slightly higher scores

on weekends (Friday to Sunday) or not (Monday to Thursday). 42% of the visits are identified as on weekends. Figure 12 illustrates that the estimated densities for both groups are identical.

5.5 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. However, 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

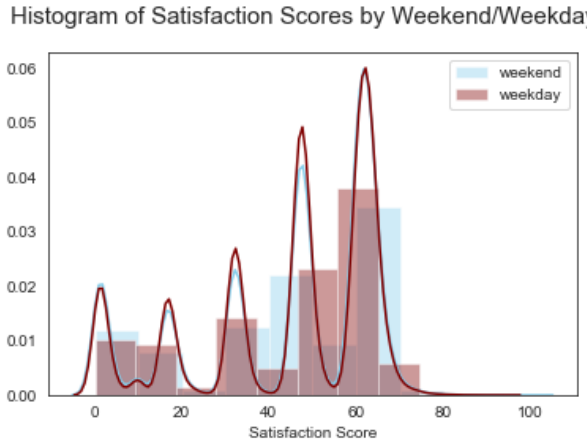

Figure 12: Mean for weekdays: 44.32; for weekends: 44.24

review text to check if it contains the keywords. 51% of the rows are labeled as group visit. Figure 13 shows that the observations tagged with individuals are about three times larger than the ones with groups. The average rating of the individual segment is 45.34 (with standard deviation 20.2) is slightly greater than the one for the groups segment (mean 43.28 and standard deviation 20.9).

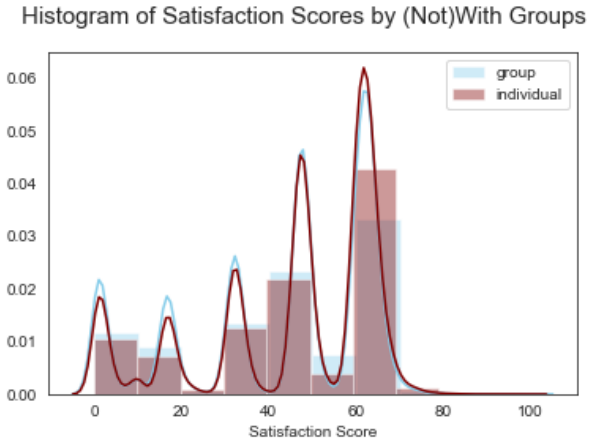

Figure 13: Two groups have similar distributions

6 Models

6.1 Building Models

Mathematically speaking, our output variable y is the customer satisfaction variable and input variables X are the features that we had constructed (business stars, review counts, weekend, gender, group). The goal is to find the optimal target function f which maps X to y. The other way to think of this problem is that satisfaction is a utility function determined by the features, but we pose no assumption on its functional form.

Before training the models, we applied the random forest algorithm to measure the relative importance of each predictor by counting the number of associated training samples in the 500 trees. After the iterations, we found the importance scores for food categories are close to 0. So we put them aside and proceeded with 5 features. Figure 14 and Figure 15 displays the features ranked by the predictive importance.

features	score
star_bus	0.71
cts_bus	0.17
weekend	0.03
gender	0.02
group	0.02
Ameri	0.01
Asian	0.01
Drinks	0.01
SouAM	0.01
Europ	0.01
MidEas	0.0

Figure 14: Tables of Features Importance by Random Forest

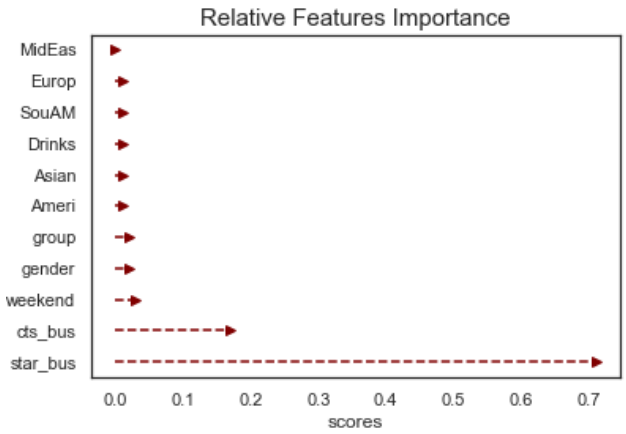

Figure 15: Business reputation is very predictive

We measured the prediction accuracy by the average squared difference between the true value (y) and predicted value ($\hat{y}$), which is known as the MSE (mean square error). The formula of MSE is

$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where n is the total number of observations (795588 rows) and i indicates the index of a given visit.

We randomly divided the data into training and test sets, with ratio 3:1. 25% of the data will be left aside while training the models, which enables us to measure the final model's performance on unseen data.We didn't see any orderings in the data, so the training and set sets should have identical underlying distributions.

6.2 Multiple Linear Regression

We started with the multiple linear regression, which is the classic linear model used in most of the social science fields.
The equation is:
$Y_i = \beta_0 + \sum_{j=1}^{5}\beta_j X_{ji} + \epsilon_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$

Calculating the coefficients by $\hat{\beta}_j = \frac{\sum(X_{ji} \breve{} \bar{X}_j)(Y_{ji} \breve{} \bar{Y}_j)}{\sum(X_{(}ji) \breve{} \bar{X}_j)^2}$ and intercept by $\hat{\beta}_0 = \bar{Y} \breve{} \sum_{j=1}^{5}\hat{\beta}_j \bar{X}_j$, the estimated model is (See full regression results in Figure 16):
$\hat{Y} = 6.13 + 0.76 * ReviewCounts + 56 * Ratings + 0.76 * Gender - 0.19 * Weekend - 2.51 * Group$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.157
Model:	OLS	Adj. R-squared:	0.157
Method:	Least Squares	F-statistic:	2.226e+04
Date:	Sun, 09 Jun 2019	Prob (F-statistic):	0.00
Time:	14:15:03	Log-Likelihood:	-2.6016e+06
No. Observations:	596691	AIC:	5.203e+06
Df Residuals:	596685	BIC:	5.203e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.1301	0.126	48.644	0.000	5.883	6.377
x1	0.7559	0.136	5.560	0.000	0.489	1.022
x2	56.1628	0.172	326.033	0.000	55.825	56.500
x3	0.7608	0.050	15.297	0.000	0.663	0.858
x4	-0.1895	0.050	-3.818	0.000	-0.287	-0.092
x5	-2.5083	0.049	-50.825	0.000	-2.605	-2.412

Omnibus:	43250.155	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53302.199
Skew:	-0.730	Prob(JB):	0.00
Kurtosis:	2.889	Cond. No.	12.8

Figure 16: Regression Output

All coefficients are statistically significant at 95% level. The coefficients of Weekend and Group are negative. This implies that all other things being equal, customers dining at weekend or with other people are predicted to have relatively lower satisfaction. However, female are on average more satisfied than males. The restaurant aggregated ratings have the largest coefficient. The estimated linear model accounts for 15.7% of the variation, which suggests the fit is not the best.

6.3 Machine Learning

6.3.1 Cross Validation and Hyper-parameters Tuning

We deployed eight supervised machine learning algorithms to find $\hat{f}$ which produces the best prediction values for the continuous $\hat{y}$.

To compare the performance of the models without using test set, we applied cross validation method to split the training set randomly into K folds. Then the first fold is evaluated by MSE

with the model fitted with the remaining K-1 folds. The procedure is repeated for K times to get K MSE values. The average CV MSE is helpful as we don't have to leave out the validation set (James, Witten, Hastie, & Tibshirani, 2013). We use 5 as the fold number in the project.

For each machine learning model, we further tuned the hyper-parameters by randomized search method. At each iteration, it will select a random combination of the hyper-parameters values and then evaluate the performance by average MSE of 5-fold cross validation (Geron, 2017).

6.3.2 Tree-based Methods

We started with the tree-based methods, as they are flexible and usually performs well for cases with relatively few features. This class of methods partition the feature space (which is 5 dimensional space for this project) into many rectangles and return the average y in the sub-region $f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$ (Hastie, Tibshirani, & Friedman, 2009). After we tuned the max depth, minimal sample splits, and minimal sample leaves, we generated a decision tree visualization in Figure 17.

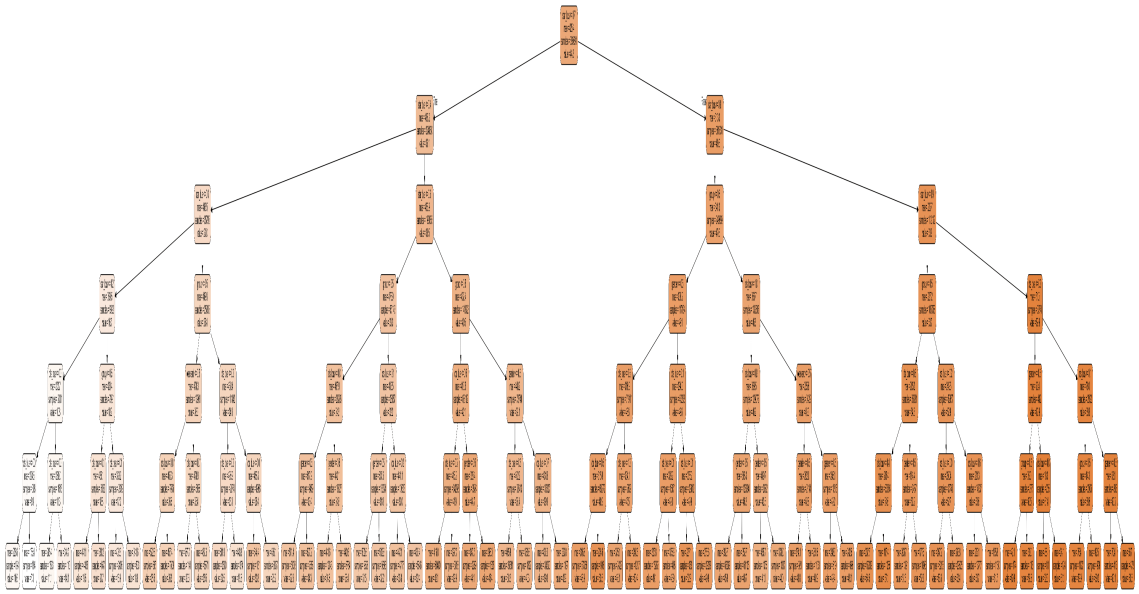

Figure 17: Decision Tree Visualization

The Random Forest algorithm searches for the best feature among a random subset of features, which brings in more randomness and diversity, as well as lower variance (Geron, 2017). If we grow B trees $(T(x;\Theta_b)_1^B)$ with this method, the predictor would be $f_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$ ((Hastie et al., 2009)).

Boosting is a popular ensemble method for improving tree-based methods. By training the predictors sequentially while correcting its predecessor, boosting combines weak learners together as a relatively better learner (Geron, 2017). The method we applied is gradient boosting, which fits the new predictor to previous predictor's residual errors (Geron, 2017).

6.3.3 Regularized Linear Methods

Multiple linear regression take all features as input, which leads to over-fitting issue. Instead, we applied the shrinkage method to regularize the coefficient estimates by the tuning parameter.

The ordinary least square procedure in multiple linear regression tries to minimize $RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$. In the Ridge regression, the new error function is $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$ where the second term is the shrinkage penalty. In the Lasso regression, the new error function is $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$ where the second term is the shrinkage penalty (James et al., 2013). In the Elastic Net regression, the new error function is $RSS + \lambda \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$ where the second term is the combined shrinkage penalty (Hastie et al., 2009).

A relatively new method that we included is called Bayesian Ridge Regression. The method estimates the regularization parameters by maximizing the log marginal likelihood ((?, ?))

6.3.4 Neural Network

Multi-layer perceptrons learn a non-linear function approximator by using back-propagation with no activation function in the output layer (Scikit-learn, 2019). The optimal model uses logistic function as activation, assigns 1.6 to α (regularization term), and has 18 hidden layers.

6.4 Model Evaluation

In Figure 18 and Figure 19, the eight models have very close performance in terms of CV MSE. One possible reason is the models have done their best with the given data after the hyper-parameter tuning. Another reason is that the feature number is too small. The mean MSE of 357 suggests that on average, the predicted value is off by around 19 with the true value.

The tree-based models perform better than the regularized linear models, although the differences are not significant. We expect the gap in performance will be more obvious when we add more features into the modeling.

CV MSE	CV Std	Name	Class
356.73	1.09	Gradient Boosting	Boosting
357.74	1.04	Random Forest	Tree
357.93	1.02	Regression Tree	Tree
358.47	1.06	Multi-layer Perceptrons	Neural Nets
358.55	1.06	Lasso Regression	Linear
358.55	1.06	Elastic Net	Linear
358.55	1.06	Bayesian Ridge	Bayeisan
358.55	1.06	Ridge Regression	Linear

Figure 18: CV MSE and Std Table

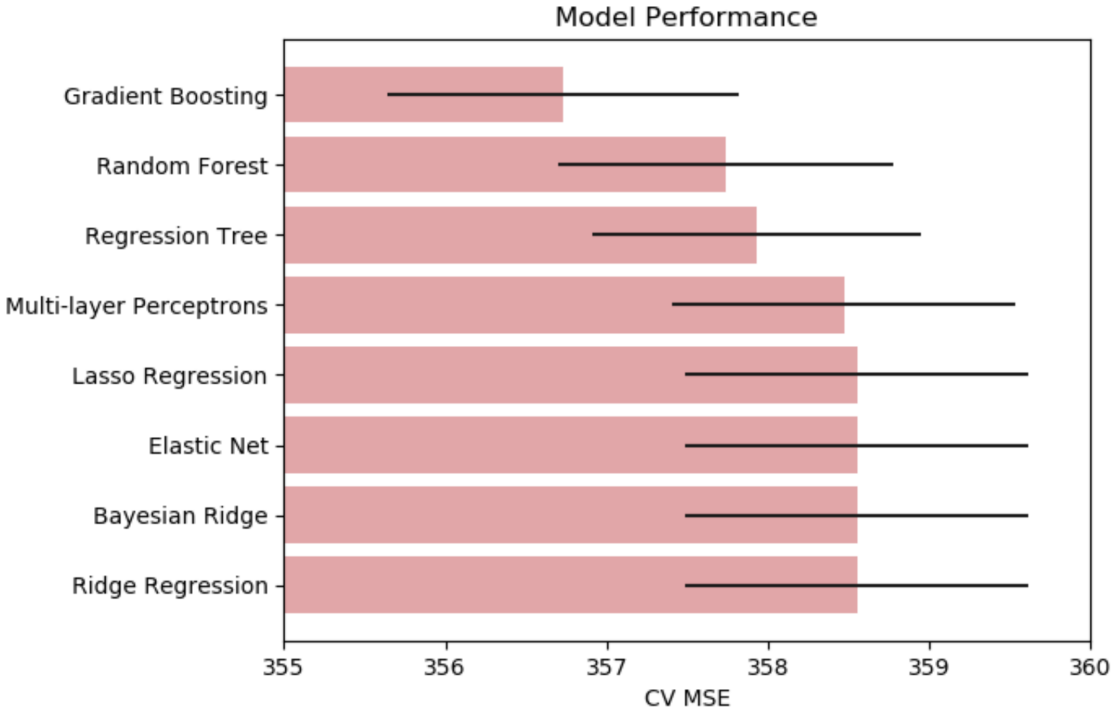

Figure 19: Model Comparison

7 Limitation and Future Work

Our project only focuses on the restaurants in Las Vegas. The original data covers more businesses categories and areas. So this limits our scope to only tourists seeking fine dining experience. It will be interesting to compare the predictions results with the ones where most of the customers are local residents. We expect to see stronger customer loyalty and network effect in that case.

We hypothesized that the restaurants could easily collect customers' information, such as the gender, with group or not. However, this requires restaurants time and investment to build such customers information database.

The models treat business stars and view counts as two predictive features. We further assume customers take restaurants reputation when they make the rational decisions. This information is vital for forming customers' prior belief of the restaurants' quality. Customers then update their belief after their dining experience. Our focus of the future work would model customers' learning process by the Bayesian framework.

8 Conclusion

This paper provides a generalized machine learning framework for businesses to predict the satisfaction of all customers. The data of restaurants in Las Vegas provides a lens to study customers who rely information from third-party source (such as Yelp app) to choose among restaurants and have no strong loyalty to certain brand.

The major finding is that restaurants reputation is most important for modeling customer satisfaction. The aggregated review stars and review counts are likely to be the best proxy to quantify customers' dining experience, such as services, atmosphere, prices, food, etc. We also didn't find significant differences in satisfaction caused by customers' heterogeneity. Although more customer-level information would be desirable, it is a challenge as people prefer to be anonymous online, especially when they leave a negative review. Also, restaurants don't have established routine to collect such information, except for asking customers to join the membership programs.

By comparing the eight machine learning models with multiple linear regression, we found the advantage of applying tree-based model in predicting the satisfaction score. The predictive advantage would be more significant if the inputs contain text and images, and this would be our next step in the research agenda. Meanwhile, Bayesian framework would possible provide a more concise and reasonable approach to model the customers' dynamic learning process.

References

Balducci, B., & Marinova, D. (2018). Unstructured Data in Marketing. *Journal of the Academy of Marketing Science*, *46*, 557–590. doi: (2018)46:557âĂŞ590https://doi.org/10.1007/s11747-018-0581-x

Dai, W., Jin, G. Z., Lee, J., & Luca, M. (2012, November). Aggregation of Consumer Ratings: An Application to Yelp.com. *NBER Working Paper*, *No. 18567*.

Gentzkow, M., Kelly, B. T., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *forthcoming*.

Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Glaeser, E. L., Kim, H., & Luca. (2017). Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity. *Harvard Business School NOM Unit Working Paper*, *No. 18-022*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (1st ed.). Springer.

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School Working Paper*, *No. 12-016*.

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

Scikit-learn. (2019). *MLPRegressor*. Retrieved from https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Timoshenko, A., & Hauser, J. R. (2019). Identifying Customer Needs from user-Generated Content. *Marketing Science*, *38*(1), 1–20. doi: http://doi.org/10.1287/mksc.2018.1123

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

Appendix A Review Sentiment

This section uses the sentiment analysis to observe how it relates with the customers rating.
We applied the VADER algorithm (Valence Aware Dictionary and sEntiment Reasone) to infer the sentiment level in the review, which is 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.

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

Table 1: Sentiments Example Table.

Figure 20 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 21 separates the scores by the satisfaction variable, where the left one are sentiment scores associated with satisfied customers (original rating >3). The unsatisfied customers (<=3) have more negative scores than the satisfied ones.

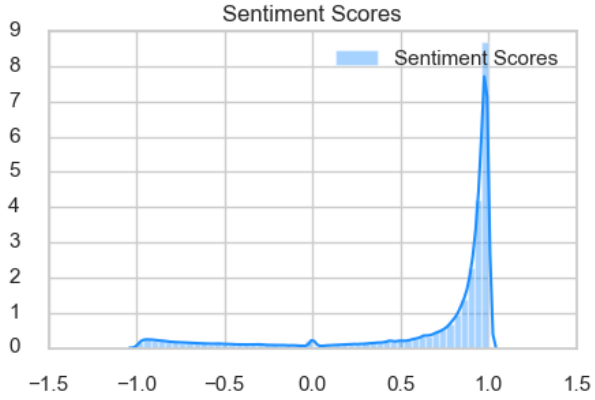

Figure 20: Histogram of Sentiment Scores

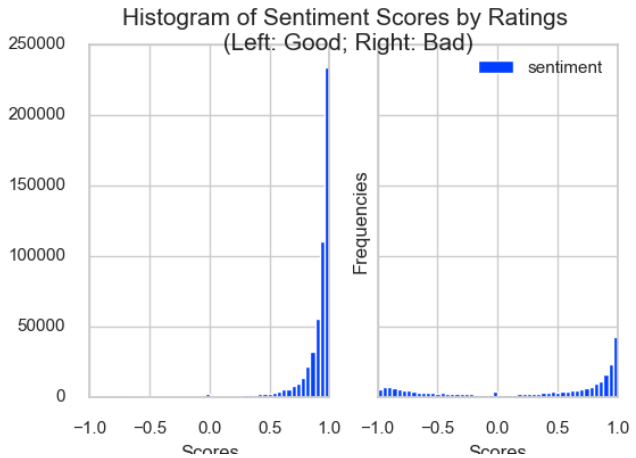

Figure 21: Histograms of Sentiment Scores by Rating Groups

Appendix B Restaurants Categories Word-clouds

This section shows the tags corpus for classifying the restaurants into categories. **American food**

"American (New)", "American (Traditional)", "Burgers", "Chicken Wings", "Fast Food", "Hot Dogs", "Salad", "Sandwiches", "Barbeque", "Smokehouse", "Pizza", "Cheesesteaks", "Cafeteria"

European food

"Italian", "French", "British", "Hungarian", "Lombian","Portuguese", "German", "Russian", "Spanish", "Ukrainian", "Fish Chips", "Greek", "Irish", "Modern European", "Polish", "Bulgarian", "Armenian"

South American food

"Mexican", "Argentine", "Cajun", "Caribbean", "Peruvian", "Honduran", "Salvadoran", "Venezuelan", "Puerto Rican", "Tex-Mex", "New Mexican Cuisine", "Nicaraguan", "Tacos", "Latin American"

Asian food

"Thai", "Chinese", "Japanese", "Fusion", "Taiwanese", "Korean", "Indian", "Cantonese", "Shanghainese", "Szechuan", "Dim Sum", "Filipino","Hawaiian", "Curry", "Asian", "Pan Asian", "Ramen", "Soba", "Vietnamese", "Singaporean", "Mongolian", "Guamanian", "Himalayan/Nepalese", "Hot Pot", "Malaysian", "Middle Eastern", "Sushi Bars", "Hong Kong Style Cafe", "Bubble Tea"

Mediterranean food

"Mediterranean","Kebab", "Pakistani", "Persian/Iranian", "Falafel", "Lebanese"

Drinks

"Beer", "Coffee", "Wine Bars", "Pubs", "Wineries", "Wine & Spirits", "Tea Rooms", "Irish Pub", "Juice Bars & Smoothies", "Bars"