Yelp Project Report: What Makes a Successful Restaurant

Group 2
Katherine Wang
Jessica Ho
Alex Kan
Ishan Shah
Svetak Sundhar

Abstract

Restaurant owners face a great deal of fear when opening up new restaurants in a new city. To help them navigate through this competitive industry, we used Yelp's dataset to provide insight into what the biggest factors of a successful restaurant are. We employed three machine learning classification models – logistic regression, quadratic discriminant analysis (QDA), and Naive Bayes – to classify whether a restaurant will be unsuccessful (have a rating less than 3.5 stars), or successful (have a rating greater than or equal to 3.5 stars). We concluded that QDA was the better model because it was best at correctly identifying unsuccessful restaurants. Additionally, Naive Bayes categorized all of the observations as the majority from the sample, which was problematic. To remedy this, we suggest utilizing sentiment analysis to map keywords to certain categories. Random Forest also produced similar results to Naive Bayes, but our Decision Tree model did extract some insights from two important features. Finally, for future work, we suggest a different metric of measuring the "success" of a restaurant as the star rating metric clumps in fast food restaurants with 5 star and Michelin-star restaurants.

I. Introduction

Restaurant owners face a great deal of fear when opening up a new restaurant. Research shows that 59% of restaurants fail within their first year of opening, and 80% of restaurants fail within the first five years of opening [1]. Our team investigated the possible reasons behind this and the measures a restaurant can take in an area to maximize chances of success in this competitive industry.

II. Problem Definition

Using machine learning, we hoped to identify:

1. What factors contribute most to a restaurant's rating.
2. If a new restaurant will be successful or unsuccessful (precise definitions of successful and unsuccessful to be given in section IV).

III. Data Preprocessing

Our data was derived from two main sources. The first source was the Yelp Challenge Dataset. It consisted of 5.2 million reviews from 174,000 cities in 11 metropolitan areas across the U.S and Canada. It was spread across seven tables, some at the business level, some at the user level, and one at the review level. The three tables we decided to use were:

1. *yelp_business*, containing the name, location, stars, number of reviews, categories and attributes of each business;
2. *yelp_user*, which has the properties of each user, including their review count, the types of votes they have sent, and compliments received; and
3. *yelp_review*, which has a row with text of each review and the stars given.

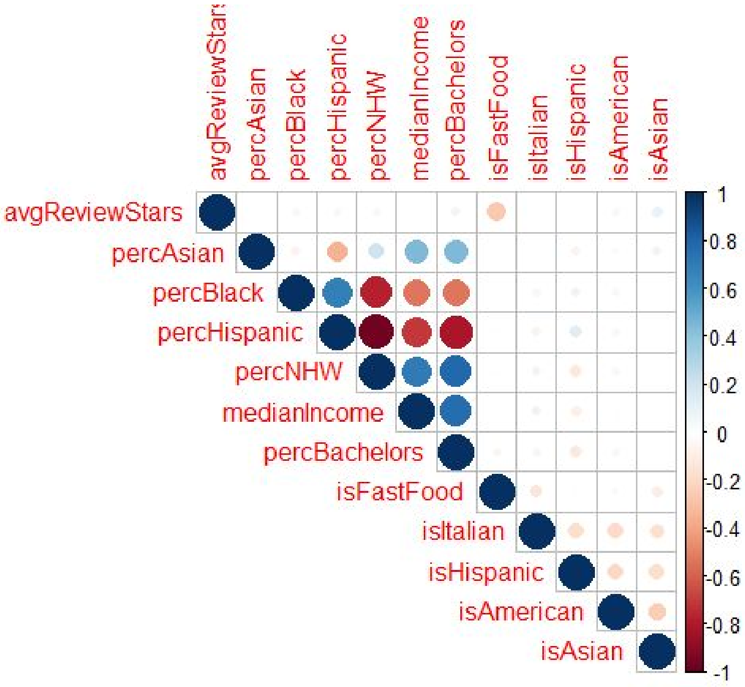
The second source was census demographic data pulled from *SimplyAnalytics*. Through the platform, we created a 30 mile radius around Phoenix, Arizona and retrieved data for all the ZIP codes in the data. We included the population, racial distribution, age distribution, median income, and percent with a Bachelor's degree for each ZIP code. We decided to isolate our data to only one city because every city varies in factors that would lead to better restaurant reviews. We chose Phoenix specifically because it was the largest city represented with a mainly local market for its restaurants. Las Vegas was the largest overall, but it had too many confounding factors, such as the large tourist population and geographic concentration.

We then had to aggregate the reviews, user, business, and demographic data by business ID to form one coherent table. To do this, we first merged all the tables together by business ID, and demographics by ZIP code, creating a data frame with each row being a review with all the details of the location, business, and user. This provided us all the data to condense the table into having one row for each business. Examples of aggregated variables are:

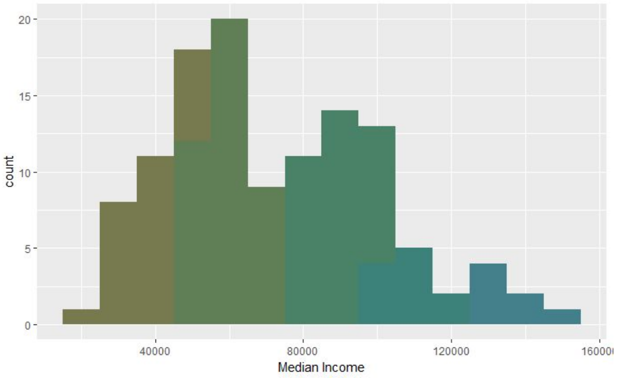
- *avgUserPastReview*: Average stars given by the user in past reviews
- *avgReviewStars:* Average stars for restaurant aggregated from review level (more precise than 0.5-rounded measure given)

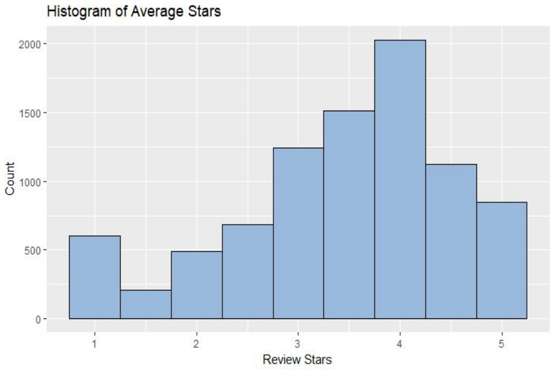
We further created binary variables based on string detection of keywords in the 'categories' field of the business. Examples include *isFastFood, isItalian, isHispanic,* etc. All these variables were value 1 in the range of 10% to 35% of the dataset. In the end, the data had 8742 rows for unique businesses, with 48 variables and data from 119 unique ZIP codes.

We created a correlogram with some of our demographic and food category variables, shown below.



Many of the demographic variables are heavily correlated with each other as expected, such as *medianIncome*, *percBachelors*, and the racial distribution variables. The binary cuisine variables are only loosely correlated with each other and demographics, implying they are viable predictors. We decided to use median income as a proxy for all demographic statistics to see the underlying distribution.





The left histogram displays the distribution of median income across ZIP codes, It is approximately normal, showing the wide range of demographics represented in the Phoenix area. On the right is the distribution of the response variable, average review stars. It is slightly skewed towards 3 and 4 stars, with the median being around 3.5 stars.

IV. Machine Learning

In order to measure the success of restaurants, we partitioned the star ratings into two categories: ≤ 3.5 stars, or > 3.5 stars. ≤ 3.5 stars are restaurants that we will classify as "unsuccessful", and > 3.5 stars are restaurants we will classify as "successful".

We used K-means clustering to test whether dividing the data into two categories was viable.

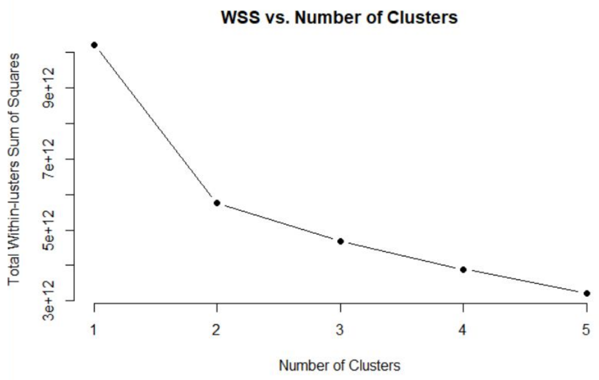
Next, we ran three classification models on our data:

1. Logistic Regression
2. Quadratic Discriminant Analysis
3. Naive Bayes

Lastly, we ran a decision tree and random forest model.

K-Means Clustering

We decided to visualize the effectiveness of our choice by plotting the "elbow curve" method discussed in class, graphing the total within sum of squares (WSS) error vs. number of clusters K from 1 to 5. The elbow value on this plot occurred at K = 2, confirming that the optimal number of clusters was indeed 2.

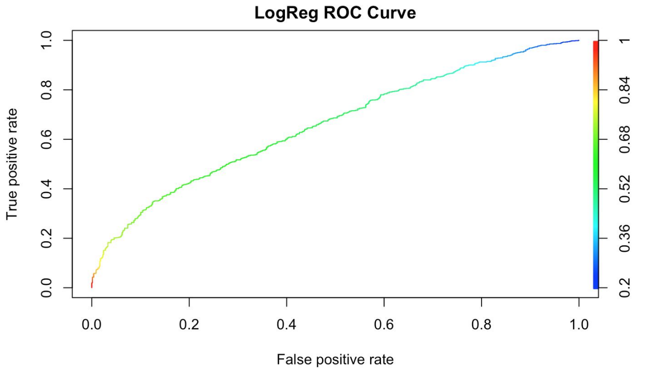


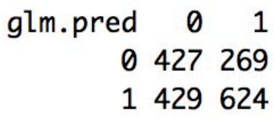
Splitting the data into 5 discrete categories would be difficult, because we would have to assign an even more arbitrary mapping to the rating values that occurred between whole numbers of stars (reviews in the dataset took on real values). Reviews at the lower end of the scale were likely to be more emotionally-charged responses, and we decided that a potential restaurant owner cared less about individual review semantics and more about a general distinction between "good" and "bad."

Logistic Regression

Logistic regression is a classification method for binary response variables that models the probability of an observation being in either class. For the purposes of this project, we decided to predict the probability of whether a restaurant will be a successful restaurant (i.e. average rating is greater than 3.5 stars) or a unsuccessful restaurant (i.e. average rating is less than 3.5 stars).

After training the model, we had a test error rate of 38.99%. While this seemed fair, upon inspecting the confusion matrix, we noticed that the model predicted most restaurants to be successful restaurants – meaning many poor restaurants were classified incorrectly. This may be problematic for new restaurant owners as they might develop misinformed expectations of high performance when, in reality, they are set up for failure.





```
glm.pred   0   1
       0 427 269
       1 429 624
```

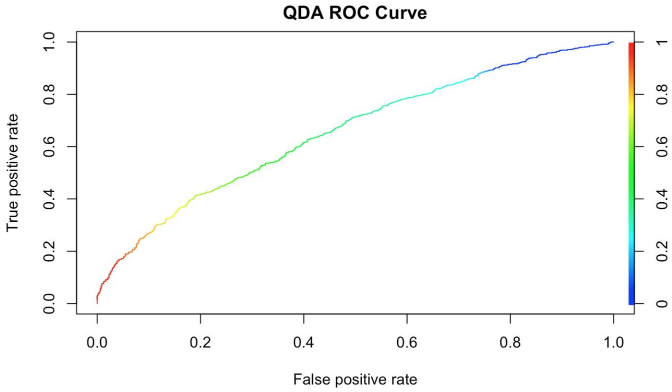
Quadratic Discriminant Analysis

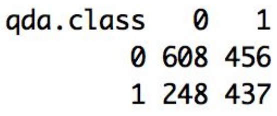
Quadratic discriminant analysis, or QDA, is an extension of linear discriminant analysis, which uses the Bayes theorem for classifying observations. We decided QDA was the better approach because it is more flexible and relaxes many assumptions. The assumptions we made were:

1. The observations were drawn from a Gaussian distribution,
2. Each class has its own covariance matrix, and
3. The classification boundary is nonlinear.

Additionally, the training set was large, meaning QDA would most likely yield more accurate results.

The test error rate was 40.60%, which was slightly higher than the error for a logistic regression. However, we decided QDA was the better model because it did a much better job of identifying restaurants that were unsuccessful. Hence, as a new restaurant owner, it would be more beneficial to think it would be unsuccessful and make improvements than think it was successful and suffer in the long run.



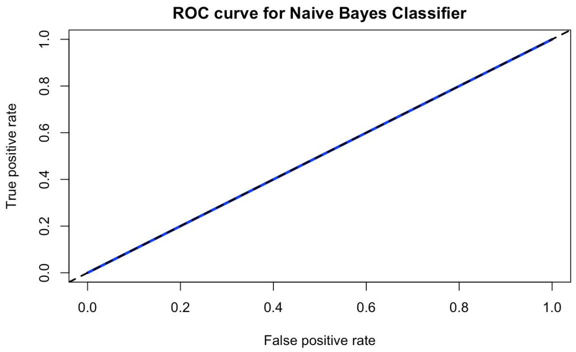


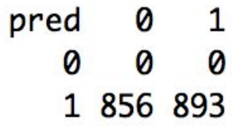
```
qda.class   0   1
        0 608 456
        1 248 437
```

Naive Bayes

Naive Bayes is a classification methods that we did not learn in STOR 565, but decided to implement it. Naive Bayes simply assumes that the predictors are all independent of one another, and uses Bayes theorem to calculate the probability that a certain observation will fall within a certain category.

We saw this as a good fit, given we were performing classification. When we ran our model, we noticed that all of our observations were being classified as a “successful” restaurant. We then tweaked our sample size and saw that all observations were being classified as “unsuccessful”. We concluded that this was simply because the model was not able to distinguish the attributes that classified a restaurant as successful or unsuccessful, and thus took the majority that the training data was classified by. In the discussion part of this work, we discuss the steps we can take to enable Naive Bayes to classify this data set more accurately.

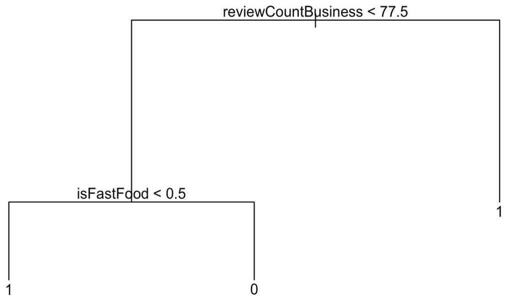




```
pred   0   1
   0   0   0
   1 856 893
```

Decision Trees

We decided to try and implement several tree-based methods because of their increased interpretability. Without any additional pruning or other modifications to our dataset, our decision tree model split the data into the two classes with 3 different terminal nodes. The two features that our model decided to split on were the number of reviews that a business had and whether or not it was a fast food business.



The selection of these two features can provide some insight into performance of restaurants based on their type and age/popularity, as evidenced by number of reviews. If a business has a large number of reviews, or a smaller number and is not fast food, then it will likely be rated well. If it has a low number of reviews and is fast food, then it will likely be rated poorly.

These insights could be due to the fact that high amounts of competition within the restaurant industry ensure that only those with good service and food (and therefore a good rating) survive enough to have a large number of reviews. Because fast food restaurants often privilege quick service and low prices over quality, they are likely to be rated worse; this only gets worse if a small number of reviews exist and bad reviews are less likely to be balanced out by good ones.

Random Forest

Additionally, we applied a Random Forest model to classify reviews within our dataset. Unfortunately, similarly to Naive Bayes, this model performed clustering to the majority class of observations. The cause of this may be due to the selection of the $\sqrt{p}$ models that the model decided to predict upon. The test error for models that classify based upon the majority class can be very deceptive in terms of test error. Because our data had a more reasonable split closer to 50/50 for each class, the test error was close to 50%. If the split was more uneven, say 85/15, and we had not verified results by checking the confusion matrix, we could have come to a false conclusion that such models perform better than some of our earlier models.

V. Discussion

After running our models on the test set, we received test errors of 38.99% for Logistic Regression, 40.60% for Quadratic Discriminant Analysis, and 48.94% for Naive Bayes. For Phoenix, QDA performed better in predicting unsuccessful restaurants while logistic regression predicted successful restaurants better. Naive Bayes predicted to the majority class. From a business perspective, it could be beneficial to use the QDA model since it is better to see if you're more likely to be a poorly-rated restaurant that way you can take those results and improve your restaurant features and overall rating. Below is a chart of how our methods compare.

Technique	**Test Error**	**Comments**
Logistic Regression	0.3899371	● Predicted highly-rated restaurants better
QDA	0.4059463	● Predicted poorly-rated restaurants better
Naive Bayes	0.4894225	● Predicted all as majority classifier

We were able to find important predictors from the Logistic Regression model (predictors with p-value < 0.05), which are the *reviewCountBusiness, avgDaysSinceJoined, isFastFood, isAsian, isAmerican, isBreakfast* variables. The review count, average days since joined, Asian cuisine or not, breakfast or not, all contribute positively to the good rating, but if a restaurant serves fast food or American food, it negatively impacts a restaurant's rating.

VI. Future Work

We are able to conclude that these results hold in Phoenix, and cannot be extrapolated to other cities such as New York or Los Angeles. This is simply because customer's standards vary across city.

To remedy Naive Bayes, we could perform sentiment analysis on the text. Further we could map specific keywords such as "great" or "tasty" to a successful restaurant and words such as "poor" or "nasty" to an unsuccessful restaurant. The initial steps for this would be to first make a dictionary of "keywords" we would like to use. This would involve deleting common words such as "the" or "and" in the preprocessing stage. Then, the model could use the existing words in the dictionary to help classify and hopefully increase the correct classification rate.

We would like obtain more data on specific business attributes as it can define a restaurant's rating, such as cleanliness, service, chef, etc. The initial Yelp dataset provided a table of business attributes, however, majority of the table were NA values. Perhaps certain attributes are significant of the distinguishment between highly-rated and poorly-rated restaurants. Similarly, we would like demographic data at a more detailed level like census tracts. With only 119 different ZIP codes, we only had 119 different rows of demographic statistics, making it difficult for any of the demographic statistics to be significant when predicting rating. With more unique data points, we could see the business' true target market and consequently discover the true effect of demographics on how well certain types of restaurants are rated.

Additionally, we would like to create a new metric of success rather than just star rating, because star rating loops all restaurants under one category, while we know the restaurants vary from fast food to Michelin-starred restaurants. This will allow us to better gauge a business's success and provide a control for restaurants.

VII. Acknowledgments

We would like to thank Dr. Yufeng Liu for his guidance and support throughout this semester.

VIII. Citations

[1] King, Tiffany, Njite, David, Parsa, H.G., and Self, John T. (2005). Why Restaurants Fail Management, 46:304–322.

Gareth, J. (2017) *An Introduction to Statistical Learning: with Applications in R.* New York, NY: Springer.

Yelp, Inc. (2018). *Yelp Open Data* [Data file]. Retrieved from https://www.yelp.com/dataset.

VIV. Appendix