Yelp Predictions and Success

Eric Brownrout, Philip Meyers IV, Roshun Patel
EECS 349, Northwestern University

Summary:

We want to see whether certain attributes of a restaurant impact whether it receives positive reviews on Yelp. This is useful for business owners trying to gain maximum positive exposure on the web, as a business's Yelp page is one of its first Google search results. More specifically, our task is to predict the classified category of a rating variable that is a function of average review (stars) and number of reviews, as average review indicates quality and number of reviews indicates popularity, and an ideal business is deemed of high quality and popularity. We believe this to be an important problem because understanding the factors behind the success and failure of a restaurant could prove very useful in the commercial market. We investigated this hypothesis by obtaining a publicly available Yelp dataset, refining the dataset to just consider restaurants and creating a proprietary means of evaluating the success of a restaurant given its average rating and number of reviews. Using Weka, we attempted to generate an accurate prediction algorithm for the success of restaurants. However, we were ultimately unable to generate an algorithm that yielded reliable and accurate results.

Report:

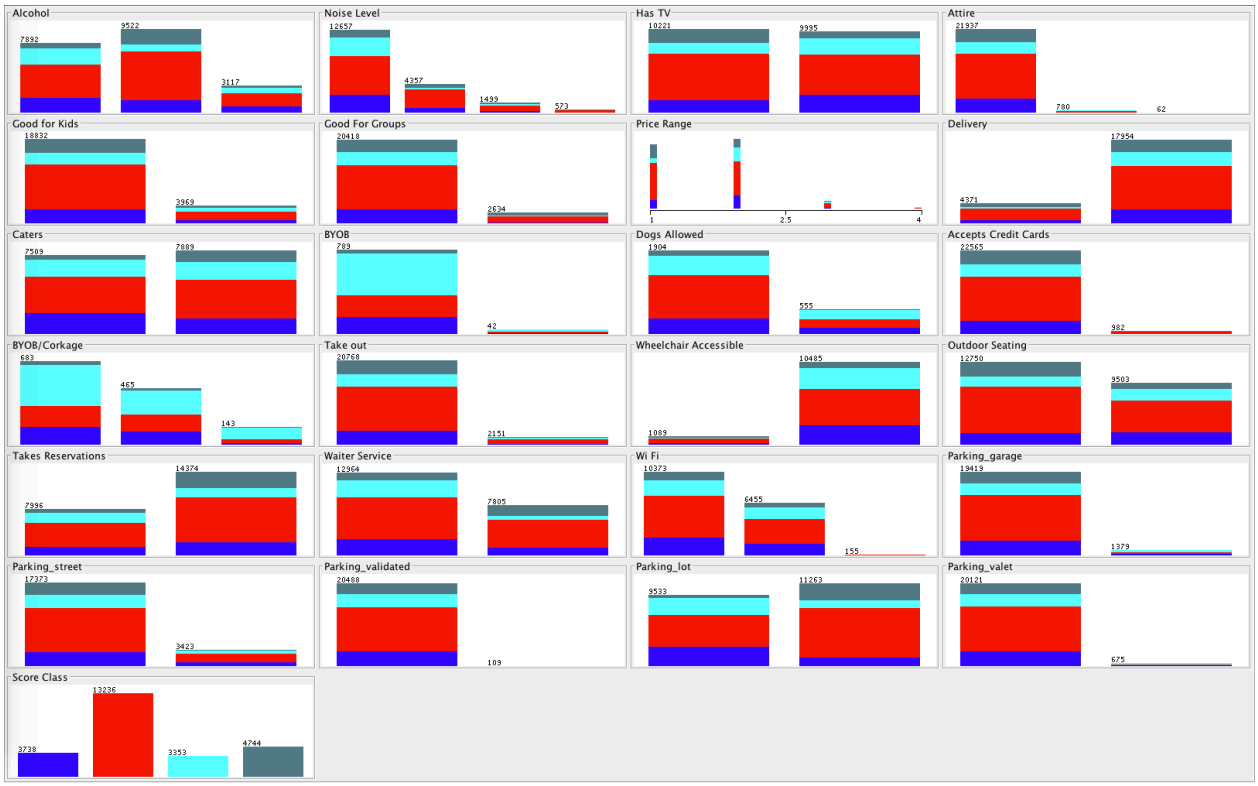
We find this task interesting for two reasons. First, given a new restaurant, we would be able to use our learner to predict its success and popularity. Second, our learner would prove useful in determining how to improve a restaurant's success by varying attributes of the restaurant. For example, if a pizzeria currently lacks dedicated customer parking but wants to know how much (if any) providing dedicated parking for their customers would boost their ratings. Our learner would provide a means of testing the impacts of potential improvements without actually forcing a restaurant to spend any money.

Our data, as indicated previously, comes from a large academic dataset that Yelp publishes as part of its yearly “Yelp Dataset Challenge.” The entire dataset contains 2.2 million reviews, 591,000 tips for 77,000 businesses, 566,000 business attributes, the Yelp social network, check-in times for all business, and 200,000 pictures from the businesses. The dataset was initially quite daunting, however, we elected to just focus on the businesses and their attributes in an attempt to investigate the relationship between attributes and the success of the business. We further restricted our dataset to only restaurants because we observed that restaurants had, on average, more reviews than other businesses. We believe that this higher review rate is because the Yelp user population has stronger opinions about food than services offered by other businesses. After refining the dataset, we were ultimately left with information on approximately 25,000 restaurants.

Our first challenge of this project was simply handling the data. The business and review data sets were both provided as JSON files, however, Yelp apparently does not conform to JSON standards as the JSON files Yelp provided were improperly formatted. The JSON files themselves were too large to fix by hand (they were simply too large to even open in a text editor

like Sublime), so we developed our own script to read, evaluate, and fix the JSON files. Even the fixed JSON files were prohibitively large to open and process (~1.5 Gb, ~70 Mb), so we wrote another script to filter the data and generate a CSV file with the relevant attributes.

Yelp provided information pertaining to a staggering number of business attributes, but we chose to focus on a subset of 20 attributes that we predicted would be most relevant to the reviews and performance of the restaurant. These attributes are: Alcohol (nominal), Noise Level (nominal), Has TV (nominal), Attire (nominal), Good for Kids (nominal), Price Range (nominal), Delivery (nominal), Caters (nominal), BYOB (nominal), Dogs Allowed (nominal), Accepts Credit Cards (nominal), BYOB/Corkage (nominal), Take Out (nominal), Wheelchair Accessible (nominal), Outdoor Seating (nominal), Takes Reservations (nominal), Waiter Service (nominal), Wi Fi (nominal), Parking Garage (nominal), Street Parking (nominal), Parking Lot (nominal), and Valet Parking (nominal). The dataset also included the average rating (rounded to the nearest 0.5) and the number of reviews factored into the average. We used these two attributes to calculate the classification for each restaurant.



The distribution of attributes for each Score classification

The classification, called "Score" is a nominal class taking on four values: Low, Medium, High, Very High. These values ideally correspond to the performance of a restaurant, where a restaurant with a Low classification is unsuccessful and has the potential to shut down and a restaurant with a Very High classification is an extremely successful and respected business. We elected to use a discrete classification as opposed to a continuous classification (which would be the equivalent of trying to predict the restaurant's actual average score) for two reasons. First, our entire dataset is composed of nominal values and we expected it would be very difficult to

generate a continuous classification from only discrete values. Second, we decided that a continuous value was not desirable for our task. Our task is far more concerned with predicting the success of the restaurant than predicting whether a restaurant will receive an average rating of 3.4 versus 3.7.

In the process of attempting to generate appropriate Score classifications, we found that we need to weigh restaurants with more ratings as better than those of less reviews. That is, a restaurant with 4.5 star average across 3 reviews is worse in our model than a restaurant with 4.0 star average across 150 reviews because even though it has a higher star average, a proxy for quality, the number of reviews, a proxy for popularity, is significantly lower (remember, an ideal business has high quality and high popularity). In order to accomplish this, we used the function: score=p/2+5(1−e^(−q/Q)), where p is the average review, q is the number of reviews, and Q is a factor chosen to normalize the values. The choice of Q depends on what we deem "few", "moderate", "many" reviews. Consider a value M that we deem "moderate" and take Q=−M/ln(1/2)≈1.44M. So if we think 100 is a moderate value then take Q=144. In our case, we defined the moderate value as the median value for the number of ratings per restaurant in our dataset.

Our final dataset consisted of 24 nominal attributes and 1 nominal classification for 25,071 restaurant instances. We aggregated this data in a CSV file and used Weka to generate various approximate hypothesis for our prediction function. We began by running ZeroR on our data with 10-fold cross validation and established a baseline performance of 52.7% by classifying all instances as Medium (the most common classification by a significant margin). We continued to train and evaluate on our dataset with 10-fold cross validation on a wide range of classifiers, however, we were unsuccessful in generating a significantly more accurate prediction function.

Classifier	Classifier Type	Accuracy
ZeroR	Rules	52.7%
NaiveBayes	Bayes	49.9%
BayesNet	Bayes	52.9%
KStar	Lazy	49.8%
IBk	Lazy	49.3%
HyperPipes	Misc	15.0%
VFI	Misc	18.5%
ConjunctiveRule	Rules	52.3%
DecisionTable	Rules	57.1%
Ridor	Rules	50.2%
DecisionStump	Trees	52.3%
J48graft	Trees	54.7%
RandomTree	Trees	52.4%
MultilayerPerceptron	Functions	58.0%
RBFNetwork	Functions	55.5%
AdaBoost (DecisionTable)	Meta	58.4%

Our failure to generate a viable prediction algorithm was disappointing. We attempted to modify our dataset by omitting certain attributes or aggregating other attributes (like combining all

parking attributes into a single attribute representing the disjunction of the 4 types of parking). We experimented with changing our output space to be everything from a binary "positive" and "negative" reviews (above 3.5 stars + certain review count threshold = 1, else = 0) to having more than 4 outcomes. Ultimately, we were unable to come up with more meaningful results. Even boosting proved fruitless in improving upon our previous results.

Analyzing our results, or lack thereof, we conclude that there is no strong, reliable correlation between the set of attributes we considered and the performance of a restaurant. It is an unfortunate conclusion, however, it is also a logical conclusion. While it would be very interesting (and beneficial to our group project) for the attributes and amenities of a restaurant to significantly affect its Yelp rating (and transitively its success in real life), in reality the most important factor in a restaurant's success is the food itself. People go to restaurants for food and the ratings that they give to restaurants will be dominated by their opinion of the food. Our data is almost entirely independent of a restaurant's food (except perhaps the price range attribute) so it actually makes sense that our data is a poor predictor of the restaurant's success. Looking to the future, we believe that while our dataset failed to accurate predict success of restaurants on Yelp, the features we considered are still likely to be relevant to other types of businesses. If we had more time, we would re-examine the entire Yelp dataset in its entirety and choose a new type of business (or subset of business) to consider and attempt to generate a new success hypothesis.

The bulk of the work was completed together by all three members, however, Philip Meyers IV led the data processing and generation aspect, Roshun Patel led the training and experimentation aspect, and Eric Brownrout led the data aggregation and conclusion generation aspect.