Capstone Report

Matthew Sheffer

Predicting Five-star and One-star Yelp Restaurant Ratings from English Yelp Review Text

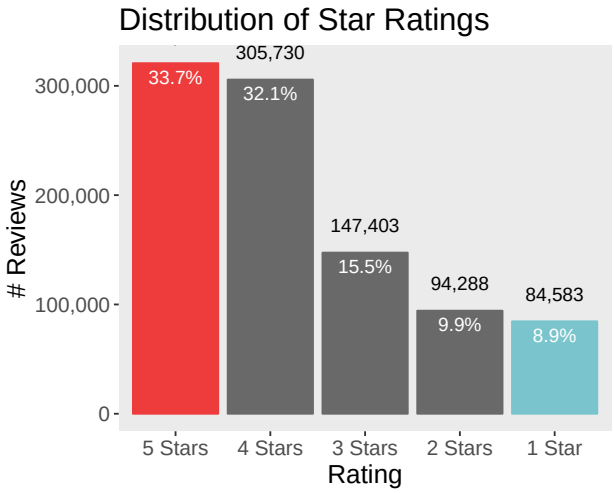
Introduction

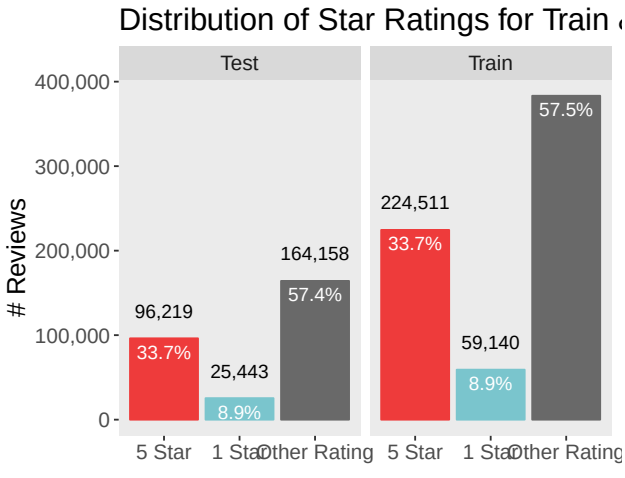
I was inspired by the example challenge ideas from the Yelp Dataset Challenge web page. I wanted to see if could accurately predict high (five-star) and low (one-star) ratings based on the review text alone, using restaurant data from English-speaking cities provided in the dataset. I believe that Yelp, as well as restaurants using the Yelp service, would be interested in knowing if there are specific review words that tend to predict very high or very low ratings as these words could point to specific customer service terms that would suggest categorization for Yelp or service improvement areas for the restaurants themselves. I ran 6 different predictive algorithms to see how well I could predict if the rating was a 5-star or 1-star rating. I also identified the words that best differentiate 5- and 1-star ratings from other types of ratings.

Methods & Data

Data for this project come from the Yelp Dataset Challenge. It contains approximately 1.6 million reviews by 366,000 users for over 61,000 businesses in 11 cities around the world. For this project, the data is limited to only restaurant reviews in the cities where English is the predominant language spoken (Edinburgh, UK, Waterloo, Canada, and Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison in the USA). The specific datasets used for this project include the business file (to identify restaurants vs. other types of businesses) and the review file (that contained the text of the specific reviews to be analyzed). I utilized a Python script provided by Yelp to assist in transforming the datasets (originally in JSON format) into a more amenable format for R and then conducted my analyses within R. Once the data was restricted to just restaurant reviews in the English-speaking cities and combined, it contained approximately 1,569,264 reviews for 18,953 businesses in the 8 cities of interest. Due to the size of the datasets and the execution time required to perform even basic analyses, all of the outputs shown in this report were saved previously and recalled for use in this document (all files provided in my Capstone github repository.

Figures 1 & 2: Star Rating Distributions



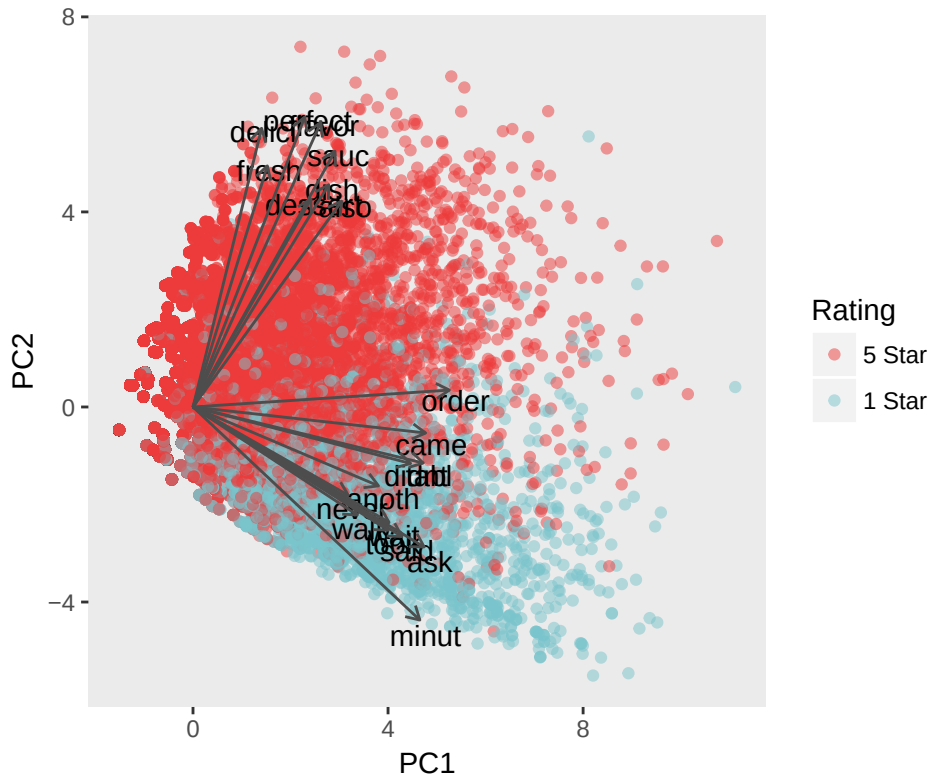


To construct a document-term matrix suitable for a predictive model, I first constructed a corpus by cleaning the the review text (removing non-ASCII characters, converted all words to lower case, stemmed the words, and removed common English stop-words and other rare words). Figure 1 shows the distribution of the ratings for the restaurants in the final, cleaned corpus. I am primarily interested in th 33.7% of the ratings that are 5-star ratings and the 8.9% of the ratings that are 1-star reviews. In order to predict the ratings, I run six different predictive algorithms to predict if the review is either a 5-star or 1-star review (so 12 models in total). The predictive algorithms I use include two classification trees (CART and C5.0), a Random Forest model, a Bagged Tree model, and Gradient Boosting Machine (GBM), and a Naive Bayes model. In order to assess the out-of- sample hit rate and avoid over-fitting the data, I break the dataset into two parts: about 70% of the cases are committed to a Training Data set and about 30% of the cases are committed to a Test Data set. Figure 2 shows the distribution of the different star ratings between the two sets. I will use the training data to fit the models and then the test data to assess goodeness-of-fit.

Exploratory Analysis

Before conducting the predictive models, I want to first explore the data to see if there appears to be any noticeable link between the types of words used in the reviews and the star rating. Figure 3 below shows a Principal Component Map of some of the most popular words uses in restaurant reviews in the Yelp database. For clarity, I've removed the points associated with other ratings and only plotted the 5-star and 1-star reviews.

Figure 3: Principal Components Map of Review Words by Star Rating



Although the data is multidimensional in nature, the two components that explain the most variation in rating shows a clear pattern toward high (5-star) reviews on the vertical dimension and low (1-star) reviews on the horizontal dimension. Word-stems such as perfect, delici, flavor, sauc, fresh, favorit, and dessert appear to be much more strongly associated with 5-star reviews than with 1-star revies. Word-stems such as minut, ask, took, didnt, anoth, came, never, wait, and bad appear to be much more stongly associated with 1-star reviews than with 5-star reviews. What's more, the separation is rather distinct between the high and low dimensions in the maps, suggesting that user reviews do contain distinctly different words depending on the rating.

Results

Given the exploratory results, I am confident that I can find a predictive algorithm to accurately predict if a review is high (5-star) or low (1-star). Tables 1 and 2 below show the results of the 6 different predictive algorithms used in this analysis. I report the overall accuracy, sensitivity (correctly predicting if 5 or 1 star), specificity (correctly predicting if some other rating), and the area-under-the-curve for both the training dataset and the test dataset.

Table 1: 5 Star Model Results

	Model	Train Acc.	Train Sens.	Train Spec.	Train AUC	Test Acc.	Test Sens.	Test Spec.	Test AUC
1	CART	69.2%	32.7%	87.7%	61.2%	69.1%	32.6%	87.7%	61.2%
2	C5.0	79.8%	63.4%	88.1%	85.8%	75.8%	57.3%	85.2%	81.4%
3	Random Forest	98.4%	96.7%	99.2%	98.8%	75.5%	46.8%	90.1%	80.6%
4	Bagged Tree	99.3%	98.5%	99.6%	99.1%	74.5%	52.2%	85.8%	78.5%
5	GBM	69%	13.8%	97%	71.5%	68.9%	13.8%	96.9%	71.4%
6	Naive Bayes	74.9%	82.5%	71%	87.5%	74.9%	82.5%	71.1%	87.6%

The training results for the 5-star models suggest that most of the models do a pretty good job of identifying if the review is a 5-star review or not. All have an accuracy of over 69% but accuracy can be misleading; because there are so many non-5-star reviews in the data, accuracy can be inflated by predicting mostly non-star reviews. For example, the GBM training model is 69% accuracy but the sensitivity is only about 14% compared to the specificity of 97% - the accuracy is inflated by predicting mostly non-5-star ratings. The test accuracy drops as one might expect but all are still rather accurate in that they correctly predict the review between 69% and 76% of the time. Only one model, though, achieve a high success rate y correctly identifying the 5-star reviews as evidenced by a high sensitivity rating - the naive bayes model. All of the others sacrifice sensitivity for specificity and since the goal here is to predict if a rating is 5-star, then the sensitivity measure is very important. The naive bayes model, therefore, does the best job of correctly identifying both 5-star and other-star reviews (83% and 71% of the time respectively), has a good overall accuracy rate (75%), and has the highest area-under-the-curve (88%).

Table 2: 1 Star Model Results

	Model	Train Acc.	Train Sens.	Train Spec.	Train AUC	Test Acc.	Test Sens.	Test Spec.	Test AUC
1	CART	91.1%	0%	100%	50%	91.1%	0%	100%	50%
2	C5.0	95.4%	53.2%	99.5%	96.3%	92.5%	33.7%	98.3%	89.4%
3	Random Forest	99.4%	93.8%	100%	98.6%	92.3%	20.7%	99.4%	88.1%
4	Bagged Tree	99.8%	97.8%	100%	98.9%	92.3%	28.3%	98.5%	85.2%
5	GBM	91.1%	0%	100%	77.2%	91.1%	0%	100%	77.4%
6	Naive Bayes	89.5%	88.7%	89.6%	95.2%	89.4%	88.5%	89.5%	95.1%

The training results for the 1-star models appear even better than the 5-star models based on accuracy, but here the small proportion of 1-star reviews is more evident. Most of the training models predict nearly 100% non-reviews as evidenced by the very high specificity scores. The test accuracy stays close the training accuracy but the same problem of over-predicting non-1-star ratings is also evident. Only the naive bayes model, does a good job of correctly identifying both 1-star and other-star reviews (86% and 90% of the time respectively), has a good overall accuracy rate (90%), and has the highest area-under-the-curve (95%).

Although the naive bayes models do a good job of predicting whether or not the review is high (5-star) or low (1-star), it is not possible to determine what particular words drive the prediction. This is a limitation of naive bayes predictive models in that it does not afford any type of predictor importance measure. In order to identify which words are most associated with high and low ratings, I instead rely on the C5.0 model, which performs well according to Tables 1 and 2 but not quite as well as the naive bayes models.

Figure 4: Predictor Importance by Model

Figure 5: Word Frequency of Most Important Predictors by Rating

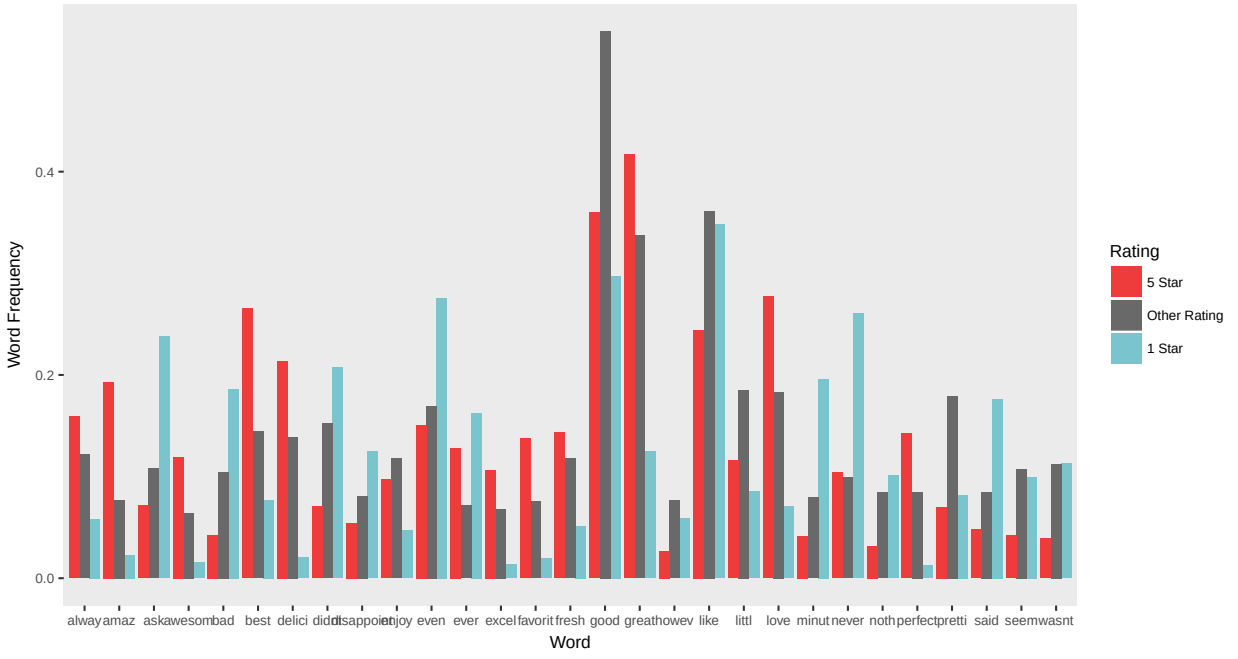


Figure 4 lists the top words based on an importance measure used with C5.0 models to predict 5-star or 1-star ratings. The importance score in this case is the percentage of samples that fall into all terminal nodes

after a split - the higher the value, the more "important" the predictor. As Figure 4 shows, similar words are considered important for predicting both 5-star reviews as well as 1-star reviews but there are some that are unique as well. Word stems like wasnt, noth, amaz, didnt, awesome, favorit, and like are important and unique to predicting 5-star ratings. Word stems like never, ever, good, enjoy, alway, even, disappoint, little, and fresh are important and unique to predicting 1-star ratings. There are some words, however, that are important to both sets of predictions.

Figure 5 suggests why this may be so; it shows the relative frequency of each of the top predictors among 5-star reviews, 1-star reviews, and the other star reviews. The word-stem amaz, for example, is one of the top 20 most important predictors and Figure 5 suggests that it is a highly frequent word-stem in 5-star reviews but has a very low word-frequency in 1-star reviews. Other star reviews fall somewhere in between. So this word is important for predicting both types of reviews. Other word-stems that are highly associated with 5-star reviews include delici, best, awesom, love, and great. Word-stems highly associated with 1-star reviews include bad, minut, didnt, ask, and said. Interestingly, many of these same words were identified in the exploratory PCA presented in Figure 3.

Discussion

I was able to identify a model that could accurately predict whether or not a Yelp restaurant review in an English-speaking city was high (5-star) or low (1-star). The out-of-sample accuracy for both models was high: 75% for predicting 5-star reviews and 90% for predicting 1-star reviews. What's more, both models did a very good job of separating the star reviews from other reviews; the 5-star model correctly identifies 5-star reviews in 83% of cases while the 1-star model correctly identifies 1-star reviews in 89% of cases. This suggests that it is possible to predict high or low reviews based solely on the content of the review itself.

Additionally, I was able to identify that there are distinct words associated with each type of review that are not related to the other types of reviews. High reviews tend to contain word that suggest a good experience, such as delicious (delici), best, awesome (awesome), and great. All of these words occur much more frequently in high reviews than any other type of review. Similarly, low reviews tend to contain words that suggest possible service problems, such as bad, minute (minut), didn't (didnt), and ask. All of these words suggest long waits or not getting the service that was expected.

For Yelp and the restaurants that use the site, these findings suggest that the review text could be a valuable source of information for service improvement and monitoring. It's possible to use the review text to identify the highest and lowest possible review types without requiring the respondent to input any actual star rating. Additionally, the contextual usage of the words suggests that it would be possible to use the presence of certain types of words in reviews as a warning system that there may be service-related issues at a particular restaurant, should someone choose to use this data in such a way.