Predicting Star Ratings from Text Reviews in Yelp

I. Intro

The purpose of this project is to investigate the relationship between the text in a Yelp review and the corresponding number of stars. Specifically, this work addresses the following**:**

Primary Question: Can the number of stars in a Yelp review be predicted from the text alone?

To begin investigating this problem, **exploratory data analysis** was done to see if there were visually apparent differences between the text of 1-star and 5-star ratings. Word clouds were constructed for 1- and 5-star reviews using the stemmed words. In building the clouds, the top 15% most frequent words were excluded since these are very common English words. The results are shown below.

1-Star Ratings

5-Star Ratings

From observing this comparison, it is apparent that differences in word frequencies can distinguish high ratings from low ratings. This suggests that text information can be used as an effective predictor of ratings.

II. Methods

In order to predict ratings from text, this problem was modeled as a classification problem. The reasons for this choice of modeling were twofold: 1) it alleviates the problem of predicting values out of the valid range (e.g. 8 stars or negative number of stars) which many regression methods are susceptible to, and 2) it allows effective ensembling which can increase prediction accuracy.

A) Response Variable and Features

The response variable was the number of stars (treated as five distinct factor levels). To obtain the features, the text from the reviews was first stemmed. The stemmed text was then represented in a document-term matrix using term frequency/inverse document frequency (TF/IDF) weighting. The features were the TF/IDF score of the review text for each stemmed word.

B) Prediction Algorithm

A model stack was used for the prediction algorithm. This stack employed three component classifiers: 1) stochastic gradient boosting, 2) quadratic discriminant analysis, and 3) random forest. A random forest classifier was used as the stacking classifier.

C) Training and Testing Data

This project was completed using a sample of 25,000 Yelp reviews. Specifically, it contained 5,000 randomly-chosen reviews from each star level in order to provide a balanced set of data. The reason for sampling is because text mining is very computationally intensive and it was not feasible to use the full set of reviews with the author's computing resources. This dataset was broken into 60% used for training data and 40% used for testing.

In order to obtain this data, a short Python program was written to first convert the JSON into CSV format for easier consumption by R.

D) Code

All code for this project can be found in the following GitHub repo:

https://github.com/chrisgarcia001/Explorations/tree/master/yelp

III. Results

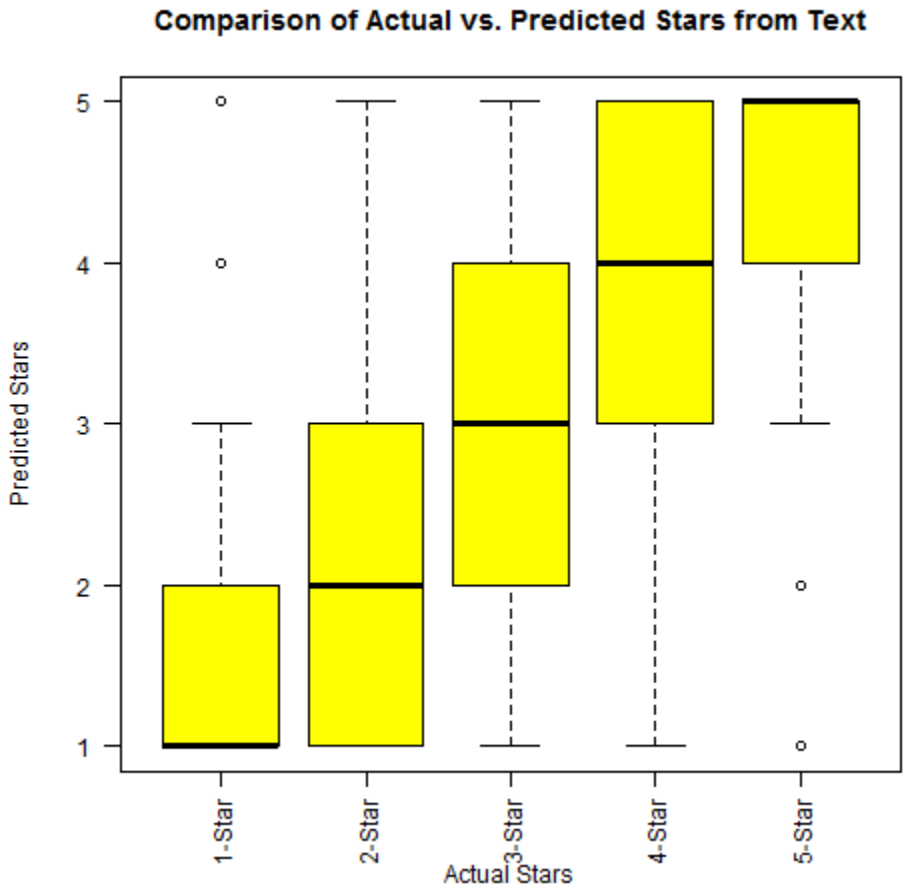
The results shown and discussed in this section are based on the test data. The raw performance results of each individual component classifier are summarized below:

Component Classifier	Accuracy	P-Value	Kappa
Quadratic Discriminant Analysis	0.4405	< 2.2e-16	0.3006
Stochastic Gradient Boosting	0.4393	< 2.2e-16	0.2991
Random Forest	0.4398	< 2.2e-16	0.2997

At first glance it appears that there is relatively low accuracy from each of the component models. However, because this is a regression problem implemented as a classification problem the specific types of misclassifications are important. Specifically, misclassifying a 5-star rating

as a 1-star is far more erroneous than classifying it as a 4-star. Additionally, ensemble methods (such as model stacking employed here) can utilize multiple weaker classifiers to create a stronger one. Next, the overall model stack performance is examined.

The **results of the primary model** (i.e. random forest stacking model) are summarized in the boxplot below:



In the above plot it is seen that for each (actual) star level, the majority of predictions are correct. Additionally, the majority of errors fall within a range of +/- 1 star. This shows that even when a prediction is incorrect it is unlikely to be far from the true value. It is additionally worth noting that the model had the highest accuracy for 1- and 5-star reviews.

IV. Discussion

The results of the primary model show that the predictions are correct in the majority of the time, and that prediction errors which deviate more than +/- 1 star are infrequent. While the predictions are not perfect, **it can be concluded** that Yelp ratings can be predicted from the text alone with reasonable accuracy. Accuracy may increase by using larger datasets than those used here.