Annika Amlie, Rahul Shukla

CS 396

18 March2020

Final Project Report

Introduction and motivation

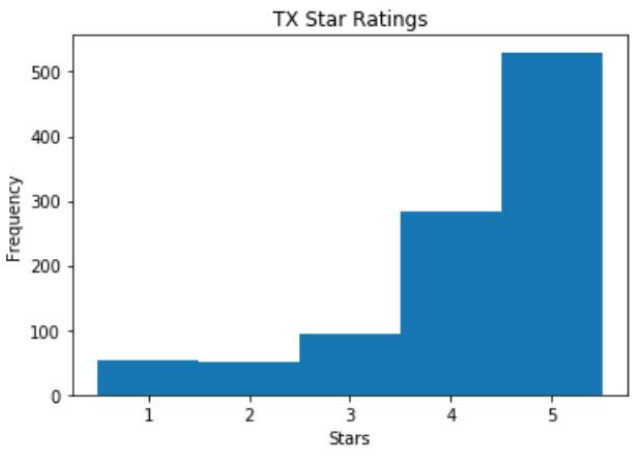
For our project, we would like to understand language patterns in reviews that give 5-star ratings to restaurants. The goal of our model is to take a given review text and predict whether it gives a 5-star rating. Through this model and through our EDA, we can better understand specific words and concepts that users care about when eating at restaurants. This knowledge can help restaurants focus on certain aspects of their businesses, and it can give users and critics a more quantitative value of 5-star reviews.

Data Cleaning

In order to obtain clean and comprehensive data, we first decided to narrow down the businesses to get only restaurants. We did this by searching for the word "Restaurants" in the 'categories' column of business.json. Once we accomplished this, we realized that our dataset was still very large. To narrow down our dataset more, we decided to only focus on restaurants in the state of Texas. After narrowing down the data, we then merged the businesses by the 'business_id' column and concatenated the 'review text' and 'stars' columns from review.json with the 'categories' and 'state' columns from business.json. Because the 'state' values for Texas were just 'TX', we did not need to do any cleaning for this. Similarly, the 'stars' and 'business_id' columns also did not need to be cleaned. The main cleaning we did was for 'review_text'. We used nltk.stopwords in order to clear out words such as "the", "and", etc. This helped us keep only words that would have a significant effect on the sentiment of the review, and would therefore be telling of what a user would value in a restaurant.

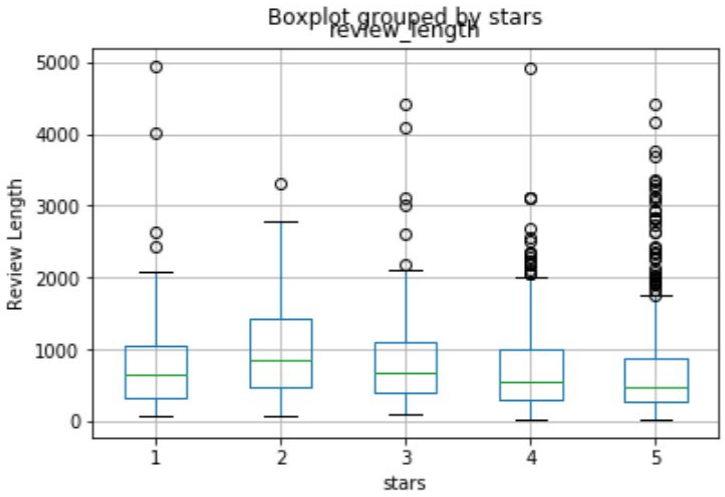
EDA

Our first step in the EDA process was to understand the distribution of 5-star ratings of restaurants in Texas. To do this, we used a histogram with a range of 1-5 on the x-axis (stars) and a range of 0-500 on the y-axis (frequency of star rating).



From this histogram, we can see that there is a left skew, meaning there is a much higher distribution of 5-star ratings in Texas than there are 1-4 stars. This is helpful to us because it allows us to visualize the amount of 5-star reviews we are analyzing, and it also allows us to get a bigger picture of the Texas users' tendencies.

In our second EDA strategy, we wanted to see if there was a large difference in review text length depending on the amount of stars the reviewer gives the restaurant. This could help us understand the weight of the words being used in the 5-star review texts, and it would just give us a better understanding of the data we are working with. If some reviews were only a few characters long, they might not be helpful to have in our dataset. We used a boxplot to better understand this distribution.



From this boxplot, we can see that the average review length is around 500 characters for all star ratings, but there is more variation for 4-5 star ratings, with many more long reviews. The lower star ratings seem to have a higher average than the 4-5 stars, but we know that there are fewer of these ratings in general, and there don't seem to be too many long reviews.

For the third step of our EDA, we wanted to see which words were most common in the 5-star review texts in Texas. This would first help us understand how effective the TFIDF method is, and it would also help us understand the words that Texas reviewers are using most frequently. To accomplish this step, we vectorized all of the review texts of the 5-star rating reviews in Texas. After doing this, we used TFIDF to assign a numeric value to the most common words in these ratings. TFIDF is a preferable metric because it combines the frequency of a word with the power of that word. Since stopwords were already taken out of the text, this enabled us to find words that were both powerful and frequent.

Here is a snapshot of some of the most common words we found using TFIDF:

```
'unbelievable': 0.703,
'fago': 0.664,
'word': 0.652,
'fun': 0.641,
'med': 0.632,
'affair': 0.628,
'hands': 0.624,
'five': 0.618,
'war': 0.611,
'pricy': 0.609,
'cheap': 0.609,
'kristen': 0.609,
'whenever': 0.605,
'terrifying': 0.593,
```

ML/Text processing/Social Network Analysis

Using TFIDF, we vectorized the review_text.

Features: Review Text (vectorized)

Output: 1 (5-star restaurant) or 0 (not a 5 star restaurant)

We used a train-test split of 80:20 and stratified the split to ensure approximately the same percentage of samples of each target class.

We implemented three models.

(1) Gradient boosting classifier which had a roc_auc score of 0.72 and an accuracy of 0.65.
(2) Random Forst Classifier (n_estimators = 800) which had a roc_auc score of 0.74 and an accuracy of 0.68.
(3) Logistic Regression which had a roc_auc score of 0.78 and an accuracy of 0.72

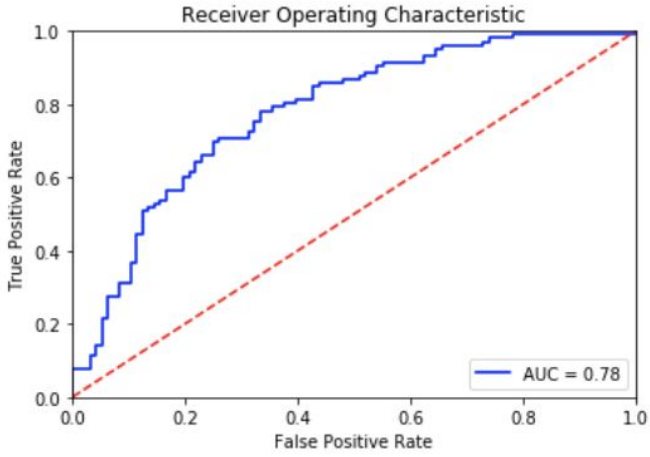
We decided to go with the Logistic Regression model because it had a better score for both metrics. However, we did value the roc_auc score the most because it can be more flexible to predict the probabilities of an observation belonging to each class in a classification problem.

Summary of findings

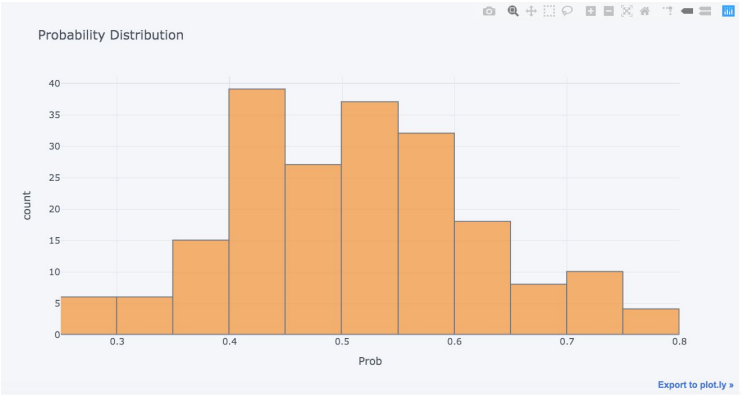
To analyze our model, we looked at the features that had the most impact on the prediction, the ROC curve of the logistic regression classifier, and a histogram of the prediction probabilities for the ‘1’ class.

-1.4169	salty	1.9754	amazing
-1.1793	the	1.4480	every
-1.0298	good	1.3680	best
-1.0277	overall	1.2695	place
-1.0266	salt	1.0841	also
-0.9309	like	1.0470	favorite
-0.9160	manager	1.0449	love
-0.8837	us	1.0302	make
-0.7864	better	1.0060	perfect
-0.7777	coupon	0.9841	ever
-0.7635	location	0.9512	delicious
-0.7150	money	0.8926	fantastic
-0.7093	not	0.8580	sirloin
-0.6882	however	0.8339	definitely
-0.6784	50	0.8071	steakhouse
-0.6761	pretty	0.7850	wonderful
-0.6732	bit	0.7786	beef
-0.6619	party	0.7411	vegas
-0.6605	nothing	0.7334	exceptional
-0.6602	but	0.7334	awesome

The visualization is interesting because we are able to find the words that have the highest impact on the prediction (highest or lower coefficient depending on class). Words like ‘amazing’, ‘best’, ‘love’ all seem to be impactful in predicting the ‘1’ class while words like ‘salty’ and ‘manager’ seem to be impactful in predicting the ‘0’ class.



The ROC-AUC plots False Positive Rate vs True Positive Rate at various threshold settings. As the blue line is always above the red line, the plot shows that our logistic regression model performs better than flipping a coin to determine whether a review is about a 5-star restaurant or not. The AUC score is 0.78 which means that there is a 78% chance that the model can distinguish between a 5-star restaurant and not a 5-star restaurant.



Please use this link to view the interactive visualization: https://plot.ly/~rss8119/1/#plot

If this link does not work, please try the Html document I've attached to the submission (should be interactive as well).

The histogram depicts the distribution of predicted probabilities outputted by the model of the positive class (the restaurant is 5-stars. The predicted probabilities refer to the confidence that the model has in its prediction. The histogram's mean is around 0.51, indicating that the model is, on average, 51% sure when predicting the positive class (the restaurant is 5-stars).

Potential implications and improvements

Potential Implications:
It would be interesting to use this model to provide a quantitative source for critics, reviewers, and users to test if a restaurant is five stars or not. Also, another extension of this project could be to generate word recommendations for when users fill out a five-star review.

Improvements:
In the future, we would want to take out more stop-words to better isolate key-words that contribute to whether a restaurant is 5-stars or not. On the machine learning step, we would want to parameter tune the logistic regression model to improve our results. We may also want to experiment with deep learning models like CNN to see if we can beat supervised learning models. More broadly, we would like to generalize this analysis to the entire country instead of just Texas.