**Text Mining Assignment 2 – Yelp Data**

Task A) Ignore the text (reviews) and run a classification model with the numeric data (you can use standard methods like logistic regression, k-nearest neighbors or anything else). What is the best accuracy of your model?

We used logistic regression and KNN to predict the high or low ratings based on the numeric data. Figure 1 and Figure 2 show us the accuracies obtained by each model. The best accuracy was from the Logistic Regression model at 68.4%.

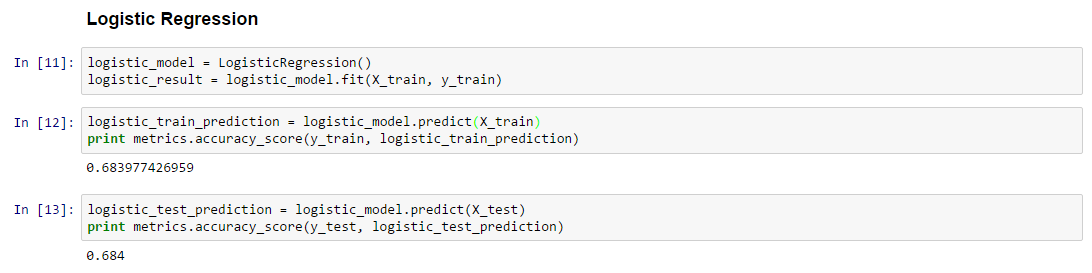


Figure : Logistic Regression Accuracy

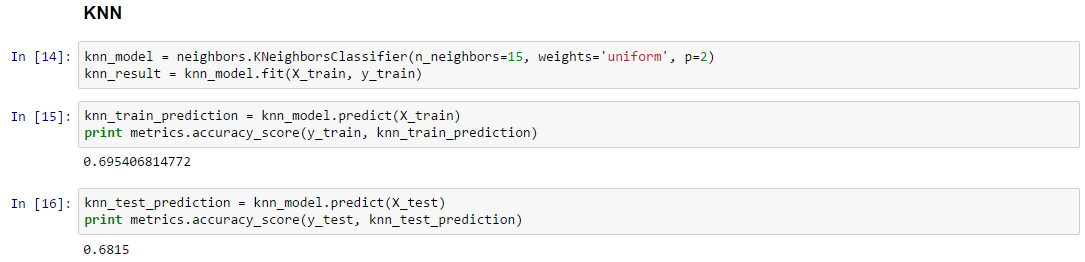


Figure : KNN Accuracy

Task B) Perform a supervised classification on a subset of the corpus using the reviews only. You can write your code in Python or R. What accuracy do you get from this text mining exercise?

We did this classification in two different ways, the first we used a randomly sampled training set [Figure 3] and the second we split the training set so we had approximately 50/50 of each class [Figure 4]. For the randomly sampled training set we achieved an accuracy of 69.6% and with the 50/50 training set we achieved an accuracy of 81.3%. In the randomly sampled training set we noticed that our model was heavily predicting ones which led us to test out the 50/50 training set.

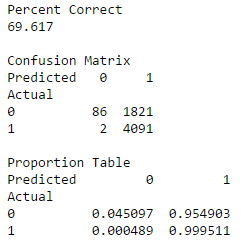


Figure : Random Sampling Classification

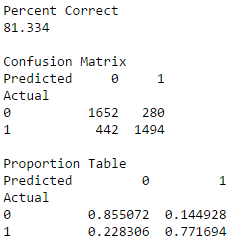


Figure : 50/50 Split (Undersampling) Classification

Task C) Combine the numeric data and the text classification model (in task B) to create a “hybrid” model. It is your task to figure out how to do this. Now run this hybrid classification model and compare the results with those in A and B.

Using the combined numerical data and text corpus we obtained in earlier parts gave us an accuracy of 79.4% using Multinomial Naïve Bayes [Figure 5]. This was roughly equivalent to using just the text data alone (max in part B was 81.3%). This method did outperform part A (max of ~68%).

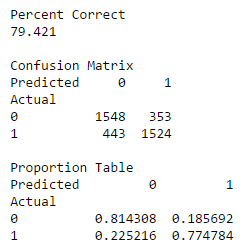


Figure : Classification with Corpus and Numerical Data

Task D) Use unsupervised sentiment analysis on the reviews (with SentiStrength or any other tool) and use the sentiment score to predict high/low rating. Compare and contrast the results of tasks B and D. What can you conclude from your analysis?

Utilizing only sentiment we achieved between 58.4% and 35.9% accuracy depending on how we classified neutral sentiment. Classifying neutral sentiment as a lower star rating resulted in 58.4% accuracy [Figure 6] whereas classifying neutral as a high star rating resulted in 35.9% accuracy [Figure 7]. This is in line with how we classified the stars; we considered only 4 and 5 star ratings as high whereas 1-3 were considered low. If you believe that a 3 star rating is supposed to be neutral then our classification accuracy, considering this, achieves a higher accuracy. Ultimately, using just sentiment was not as powerful of a classification tool as something like Multinomial Naïve Bayes with a bag of words model using the review text.

|  |  |
| --- | --- |
| Figure 6: Sentiment Classification (Neutral Classified Low) | Figure 7: Sentiment Classification (Neutral Classified as High) |

Task E) Use unsupervised clustering on the text. Does clustering achieve “good” separation between high and low rated restaurants? How can you explain the result?

Figure 8 shows the accuracy results for classifying our reviews using K-Means clustering. The reason we included two “percent correct” attributes is because depending on the randomness of the clustering, low reviews could have been clustered as either 1 or 0 and vice versa. Looking at the results manually we determined that cluster 1 was our high star cluster to which we achieved a 58.6% accuracy. The separation was better than random but not nearly as good as other models tested prior. We think this is due to significant overlap of words used in high and low reviews making it hard to systematically separate the two. (Part F will corroborate this)

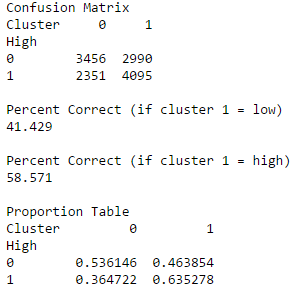


Figure : Clustering Classification Accuracy

Task F) What are the top 5 “attributes” of a restaurant that are associated with (i) high and (ii) low ratings?

Figure 9 shows that the top 30 nouns associated with a highly rated restaurant review and low rated restaurant. We generated the following five “attributes” based on these 30 nouns…

1. High: Food, Place/Atmosphere, Service, Menu, and Wait Time
2. Low: Food, Place/Atmosphere, Service, Order, and Cleanliness (based on table)

Finally, as extra, Figure 10 shows us the top ten nouns that were common for high reviews but less common for low reviews and vice versa.

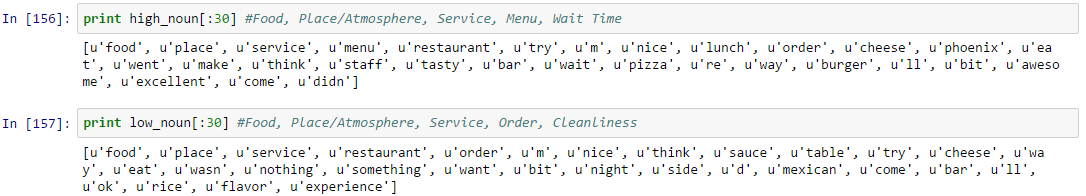


Figure : Top 5 Attributes of a High Rating and Low Rating

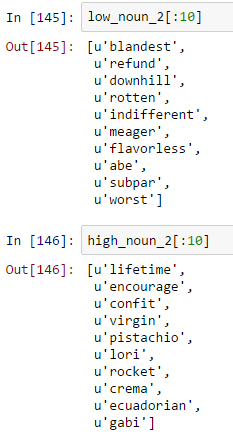


Figure : Top Ten Predictive Words For High/Low Rating