ECON 7930 Assignment 2

Yelp Ratings Predictions

Huiyan Zhang 21436576

1. INTRODUCTION

Yelp is a well-known online review platform in the United States where users can rate and review businesses. This report predicted customer ratings with Sentiment Analysis.

2. DATA PROCESSING

2.1 Data Preprocessing

In data processing, we cleaned up the data and ran some EDA analysis. The train set has 2 columns and 650,000 rows, but the test set has only 50,000 rows. We first performed some basic text manipulations by removing unnecessary symbols and punctuations. Then, we connected 1 and 2 to Negative, and 4 and 5 to Positive to create Sentiment variable. In the third phase, we randomly sampled 10,000 corpus data and tokenized them.

2.2 EDA Analysis

The word clouds revealed that the top 3 frequencies in both sets are "food", "place", "excellent". Furthermore, the cumulative frequency graphs demonstrated that cumulative frequencies rose quickly before 3,000 then subsequently slowed.

Figure 1: Word Clouds of Test Set

3. MODEL SELECTIONS

3.1 Baseline Model: Naïve Bayes

Naïve Bayes is straightforward and fast to run. It estimates the probability of the Sentiment given a series of predictors under conditional independence criteria.

3.2 Ridge Regression

Ridge Regression is another better classifier in sentiment analysis since it often exceled other complicated models and reduce overfitting and complexity by adding L2 penalty.

3.3 Lasso Regression

Lasso Regression is computationally efficient and performs variable selections and regression simultaneously. It can reduce overfitting and complexity by adding L1 penalty.

3.4 Support Vector Machine

Support Vector Machine is highly accurate and can handle many features in high-dimensional data. It can train the models by assigning new examples to one category or the other.

4. MODEL EVALUATIONS

4.1 ROC Curve and AUC Value

Figure 2 revealed that Ridge Regression achieved maximum AUC Value of 0.95, followed very closely by Support Vector Machine with a value of 0.92 and Naïve Bayes with a value of 0.89. Lasso was the lowest with a score of 0.49.

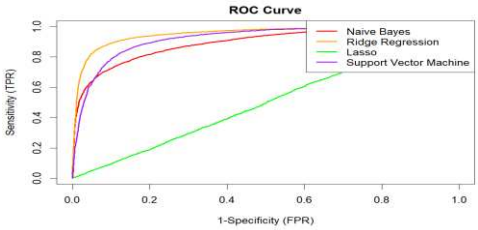

Figure 2: ROC Curve

4.2 Precision, Recall and F1 Score

Precision relates to the percentages of which are correctly identified as positive, whereas recall usually refers to the percentage of which are actually positive. Although Ridge Regression had a higher recall rate of 0.91, its precision rate 0.85 was still slightly behind the Support Vector Machine with a recall rate of 0.86.

F1 Score stands for the trade-off between precision and recall rate. Ridge Regression received highest F1 Score of 0.88, followed very closely by Support Vector Machine (0.85) and Naïve Bayes (0.75). Lasso was the lowest with 0.5.

Model	Precision	Recall	F1
Naive Bayes	0.66	0.85	0.75
Ridge Regression	0.85	0.91	0.88
Lasso Regression	0.51	0.5	0.5
Support Vector Ma	0.86	0.85	0.85

Figure 3: Classification Metrics

4.3 Distinguishable Terms

Ridge Regression was the most distinguishable since both the positive and negative terms sounded similar to the actual emotional expressions. Top 3 positive one is "pleasantly", "amazing" and "awesome" while negative is "worst", "mediocre" and "horrible". Support Vector Machine seemed to be indistinguishable with some verbs.

Model	Positive	Negative
Naive Bayes	great/good /place	buyer/agency /insulting
Ridge Regression	pleasantly /amazing /awesome	worst /mediocre/horrible
Lasso Regression	delicious/amazing/pleasantly	worst/horrible/rude
Support Vector Machine	last /small /paid	agree /larger/putting

Figure 4: Top 3 Distinguishable Terms

5. CONCLUSIONS

In this report, we trained several models then evaluated each supervised model's performance. We found that Ridge Regression was the best solution.

6. References

[1] Monroe, B. l. (2022, April 16). Text as Data Tutorial - Introduction to Text Classification (in R). https://burtmonroe.github.io/TextAsDataCourse/Tutorials/TADA-ClassificationV2.nb.html