Amazon Fine Food Customer Rating Analysis and Prediction

Yuan Wang
University of Rochester
yuan.wang1@simon.rochester.edu

Shijing Li
University of Rochester
sli111@ur.rochester.edu

Abstract

Review scores for e-commerce like Amazon is an important factor that will impact users' buying decisions. Generally speaking, there are relationships between the review score and the review text left by the buyers. It will be interesting to see how the sentiment within the review text will impact the scores and any potential correlations underlying the data.

1. Introduction

Sentiment analysis is an important natural language processing technique. This study applied various natural language processing tools such as **NLTK**, **Doc2Vec**, and **tf-idfs** on the amazon fine food review data collected from Amazon.com to analyze the review text to perform review score predictions. Three regression models were used to compare model performances and they are Linear, Ridge, and Lasso regression. The natural language processing tools appear to bring a significant increase in the model accuracy and the best model of performance was the linear regression for the review score prediction.

2. Methods

Various methods were used in this study. First, to get the text data preprocessed, **Lemmatization** and **Text Cleaning** were applied. Used both **Latent Dirichlet Allocation (LDA)** and **Non-negative Matrix Factorization (NMF)** for topic analysis. To perform sentiment analysis, we used **NLTK** and to improve the model accuracy, we applied **Doc2Vec** and **td-idfs**. Three predictive models that we tried are **Linear, Ridge and Lasso regression**. For feature selection, **Lasso for Feature Importance** and **RFECV for Feature Selection** were used.

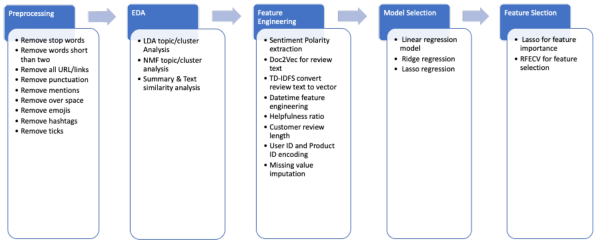

Preprocessing
• Remove stop words
• Remove words short than two
• Remove all URL/links
• Remove punctuation
• Remove mentions
• Remove over space
• Remove emojis
• Remove hashtags
• Remove ticks
EDA
• LDA topic/cluster Analysis
• NMF topic/cluster analysis
• Summary & Text similarity analysis
Feature Engineering
• Sentiment Polarity extraction
• Doc2Vec for review text
• TD-IDFS convert review text to vector
• Datetime feature engineering
• Helpfulness ratio
• Customer review length
• User ID and Product ID encoding
• Missing value imputation
Model Selection
• Linear regression model
• Ridge regression
• Lasso regression
Feature Slection
• Lasso for feature importance
• RFECV for feature selection

Figure 1. Methodology for Amazon fine food customer rating analysis and prediction

3. FINDINGS

3.0.1 Distribution of Numerical Variables over Time

Figure 2 shows the distribution of Review length, summary length, score, and helpfulness ratio over time. From all four-line graphs, we can see that from 2000 to 2003, the distributions of the length of reviews, length of the summary, score, and helpfulness ratio all showed to be pretty spread out. The range of the value is very large. From 2004 to 2006, they all reach a peak. After 2006, everything gets to its consistent state. The range becomes dramatically smaller and the frequency remains high. One reason behind this may be that from 1998-2004 is when amazon started to expand its services beyond books. Thus, the number of reviews increased drastically starting in 2004.

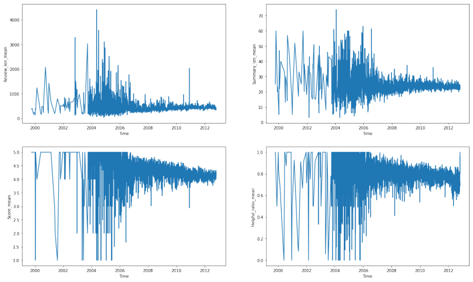
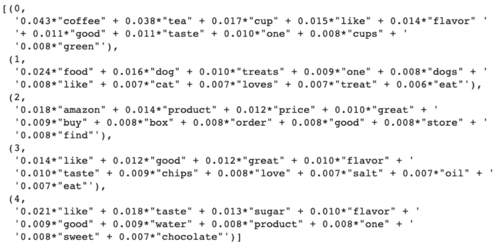


Figure 2. Distribution of Review length, summary length, score and helpfulness ratio over time

3.0.2 Topic Analysis

Figure 3 is the five topics extracted by the Latent Dirichlet Allocation (LDA) analysis and figure 4 is the five topics extracted by Non-negative Matrix Factorization. We discovered some similarities and differences with the topics generated by these two methods. In the NMF model, there is a topic that covers amazon service and pet foods. However, in the NMF model, tea and coffee are clustered into two different clusters. This is different from the LDA model where coffee and tea are clustered into one cluster. In the NMF model, there is a topic that covers ingredients whereas the LDA model is clustered into snacks and desserts. It appears that the clusters in the LDA results are centered in major food categories such as snacks, desserts, drinks, and pet foods. Yet, the NMF model outputs clusters in a more specific level, such as tea, coffee, ingredients, and pet foods. These four categories are what American people focus on daily. Coffee, tea, and pet foods are necessities for the majority of people. Also, people nowadays pay great attention to the food ingredients. It is interesting to see that the NMF model can extract the differences between coffee and tea whereas the LDA is not able to.



```
[(0,
  '0.043*"coffee" + 0.038*"tea" + 0.017*"cup" + 0.015*"like" + 0.014*"flavor" '
  '+ 0.011*"good" + 0.011*"taste" + 0.010*"one" + 0.008*"cups" + '
  '0.008*"green"'),
 (1,
  '0.024*"food" + 0.016*"dog" + 0.010*"treats" + 0.009*"one" + 0.008*"dogs" + '
  '0.008*"like" + 0.007*"cat" + 0.007*"loves" + 0.007*"treat" + 0.006*"eat"'),
 (2,
  '0.018*"amazon" + 0.014*"product" + 0.012*"price" + 0.010*"great" + '
  '0.009*"buy" + 0.008*"box" + 0.008*"order" + 0.008*"good" + 0.008*"store" + '
  '0.008*"find"'),
 (3,
  '0.014*"like" + 0.012*"good" + 0.012*"great" + 0.010*"flavor" + '
  '0.010*"taste" + 0.009*"chips" + 0.008*"love" + 0.007*"salt" + 0.007*"oil" + '
  '0.007*"eat"'),
 (4,
  '0.021*"like" + 0.018*"taste" + 0.013*"sugar" + 0.010*"flavor" + '
  '0.009*"good" + 0.009*"water" + 0.008*"product" + 0.008*"one" + '
  '0.008*"sweet" + 0.007*"chocolate"')]
```

Figure 3. Five topics extracted by the Latent Dirichlet Allocation (LDA) analysis

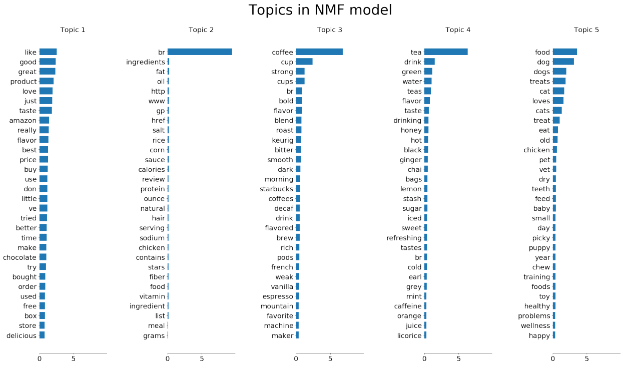


Figure 4. Five topics extracted by Non-negative Matrix Factorization (NMF)

3.0.3 Cosine Similarity

We can see the similarity score density plot in Figure 5 that the similarity value between text and summary is not very high. The majority of the score stays below 0.6. The density is at its highest when cosine similarity is 0. The density hit 9 when the cosine similarity ranges between -0.05 to 0.03. When cosine similarity is between 0.04 to 0.6, the density stays below 2. As the conclusion, the summary does not predictive text, only very few summaries describe what the text says. For the majority of the summaries, they do not tell a good story of the texts.

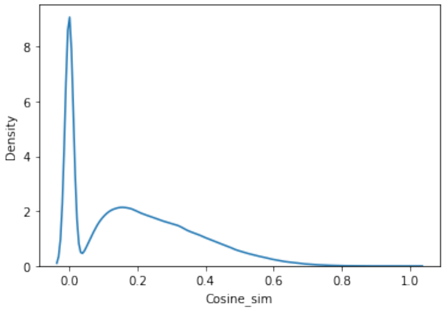


Figure 5. Similarity score density plot

4. Results

We applied three different prediction models to compare their performances and picked the best one for this study. The RMSE score we got for the ridge regression, linear regression, and lasso regression is 0.938, 0.938, and 1.304 accordingly. It appears that linear and ridge regression work better than Lasso. We decided to pick the linear regression model as our final predictive model. We have also performed feature reduction, however, the RMSE we got worsen than before. Thus, we kept all the features in the end. The final RMSE that we got for the linear model is 0.9164. Figure 6 is the confusion matrix for the test data set. Comparing the precision, recall nd F1-score for each class, it appears that score 5 has the highest F1-score of 0.84 which indicates that both the precision and recall are very high for score 5. However, the F1-score for the rest of the classes is only around 0.2.

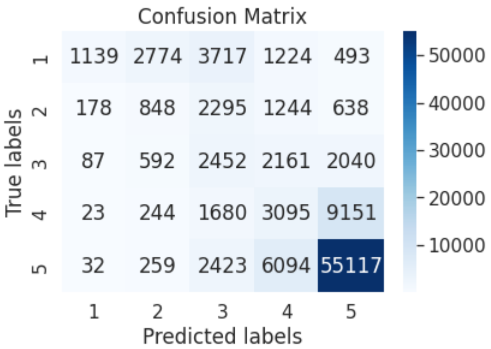


Figure 6. Confusion matrix for test data set

References

[1] Topic extraction with Non-negative Matrix Factorization and Latent Dirichlet Allocation https://scikit-learn.org/stable/auto-examples/applications/plot-topics-extraction-with-nmf-lda.html

[2] Topic Modeling in Python: Latent Dirichlet Allocation (LDA) https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

5. Appendix

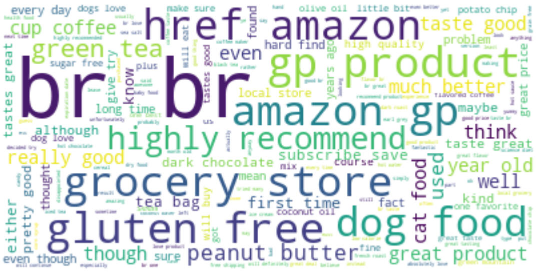


Figure 7. Word cloud of customer review

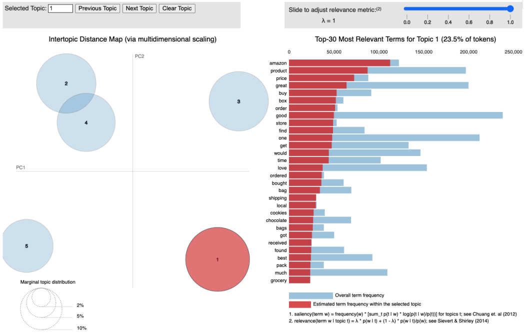


Figure 8. Topic-1 extracted by LDA clustering

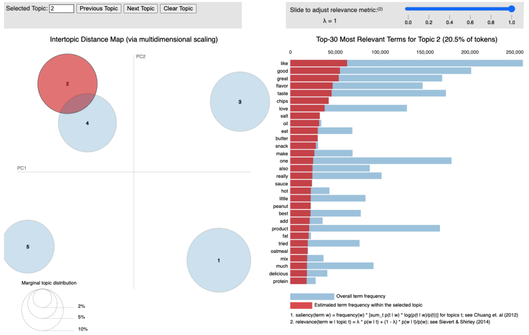


Figure 9. Topic-2 extracted by LDA clustering

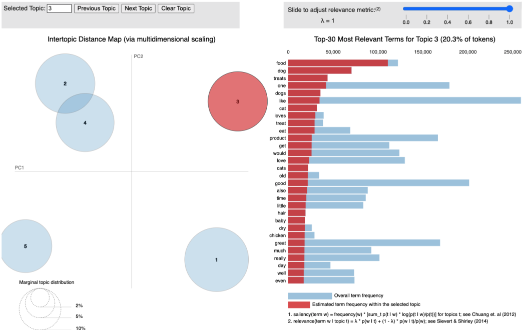


Figure 10. Topic-3 extracted by LDA clustering

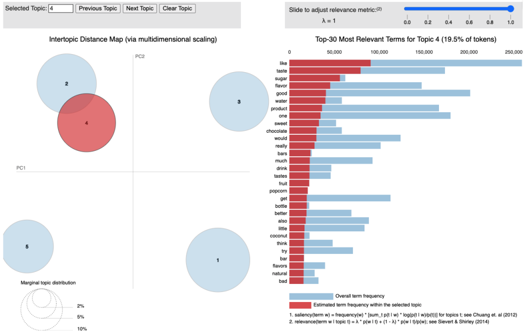


Figure 11. Topic-4 extracted by LDA clustering

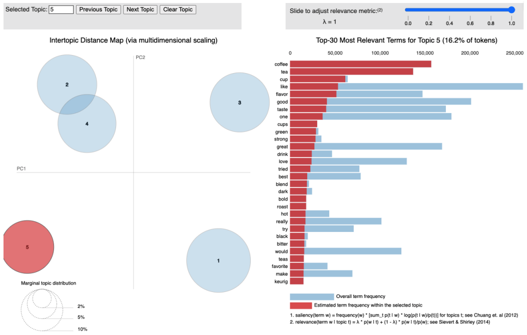


Figure 12. Topic-5 extracted by LDA clustering

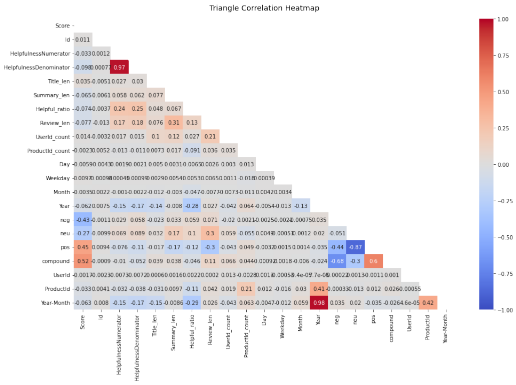


Figure 13. Feature triangle correlation heatmap

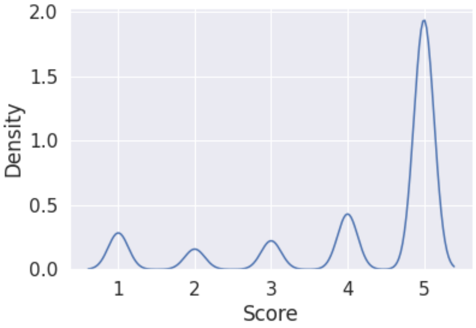


Figure 14. Test dataset customer review rating distribution

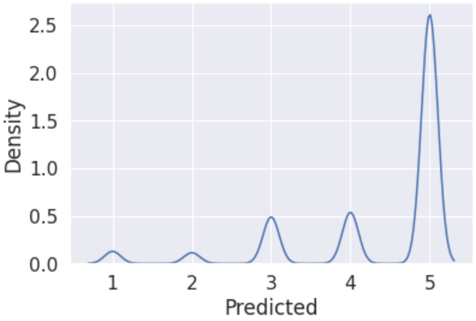


Figure 15. Test dataset Preidicted customer review rating distribution

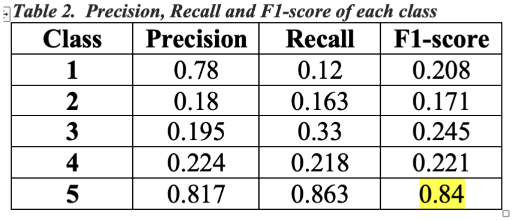


Table 2. Precision, Recall and F1-score of each class

Class	Precision	Recall	F1-score
1	0.78	0.12	0.208
2	0.18	0.163	0.171
3	0.195	0.33	0.245
4	0.224	0.218	0.221
5	0.817	0.863	0.84

Figure 16. Table 2. Precision, Recall and F1-score of each class