Analysis of Amazon Fine Food Reviews

Ezgi Siir Kibris[1]

***Abstract*— This paper is about the analysis of the Amazon fine food reviews data set. I made NLP, Latent Dirichlet Allocation (LDA), and Sentiment analysis to the reviews in this paper. Then I tried linear, ridge, and lasso regressions to predict review scores and reported the Root Mean Square Error (RMSE) of the models.**

I. INTRODUCTION

This paper is about the analysis of the Amazon fine food reviews data set using Text and Sentiment Analysis. The next part of this paper is about the descriptive statistics and the text analysis of the reviews. In the model part, I made sentiment analysis. Then I tried Linear, Lasso, and Ridge regressions to estimate the scores of the reviews from the data set. I tried to reduce RMSE using normalization.

II. DATA

Amazon fine food reviews data set consists of 568.454 reviews of fine foods from Amazon. The data span more than ten years, including reviews up to October 2012. The reviews include product and user information, ratings, and a plain text review.

A. DESCRIPTIVE ANALYSIS

This dataset consists of 568.454 reviews of fine foods from Amazon. The data set has more than ten years, including reviews from October 2012. However, after deleting missing variables, the data set consists of 35.173 observations and 11 variables.

I calculated the Helpfulness ratio: Helpfulness Numerator/Helpfulness Denominator

Figure 1 shows the descriptive statistics of the review length, title length, score, and helpfulness ratio. What I find interesting from this table is that the mean score is 4.18 out of 5, which looks high to me.

	TitleLength	ReviewLength	Score	HelpfulnessRatio
count	568411.00	568411.00	568411.00	298372.00
mean	23.45	436.24	4.18	0.78
std	14.03	445.35	1.31	0.35
min	1.00	12.00	1.00	0.00
25%	13.00	179.00	4.00	0.60
50%	20.00	302.00	5.00	1.00
75%	30.00	527.00	5.00	1.00
max	128.00	21409.00	5.00	3.00

Fig. 1. Descriptive Statistics of the review length, title length, score, and helpfulness ratio

Figure 2-5 are line plots that show review length, title length, score, and helpfulness ratio. I observe from the plots that there has been a stabilization recently. There is less volatility in all four graphs. It seems like less variance over the years in scores and review length. Since there have been more reviews recently, scores, helpfulness ratio, title length, and review length converge to the mean.

Furthermore, I am surprised about Figure 4, which shows that review scores have been higher recently. I think this pattern is a result of rising social media culture. Companies send products and give money to influencers to make them write a review. As a result of paid reviews, product scores may increase over the years. Another possibility of rising scores is that companies take care of past reviews and significantly improve product quality.

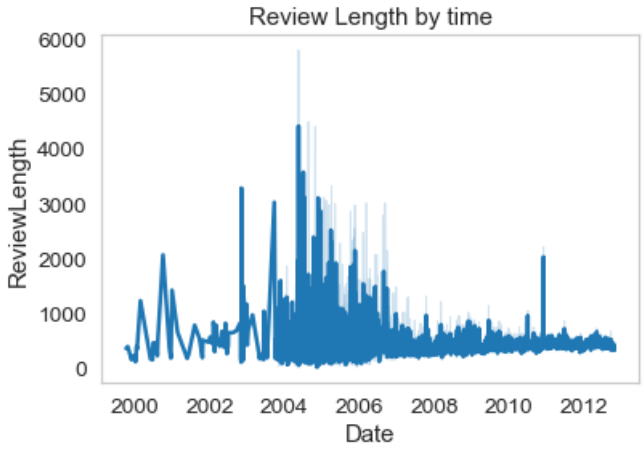

Fig. 2. Review Length over time

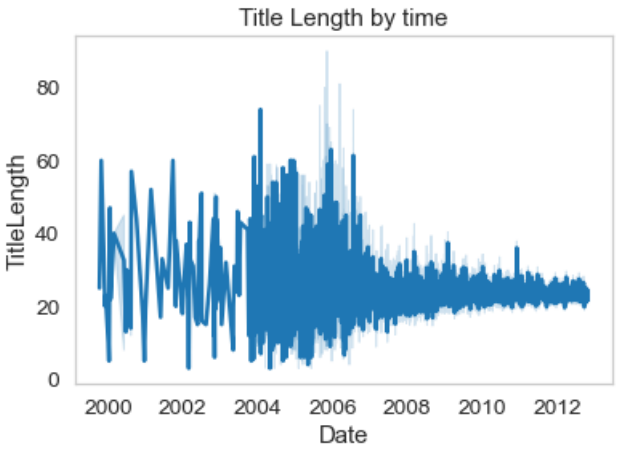

Fig. 3. Title Length over time

[1] MS Student at University of Rochester, Data Science Department

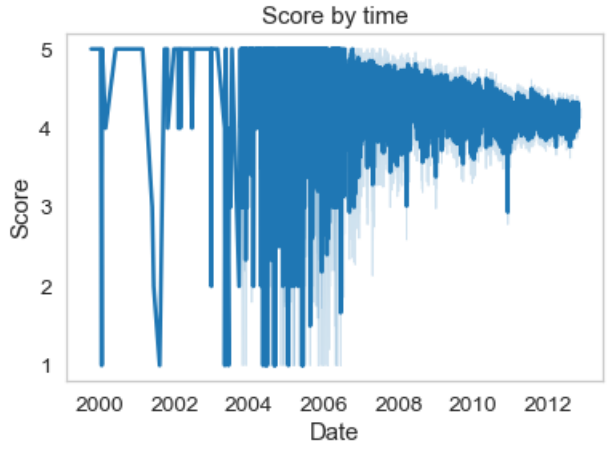

Fig. 4. Review Scores over time

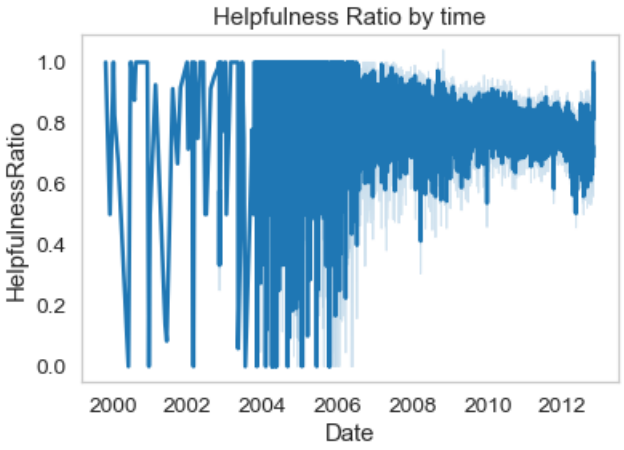

Fig. 5. Helpfulness Ratio over time

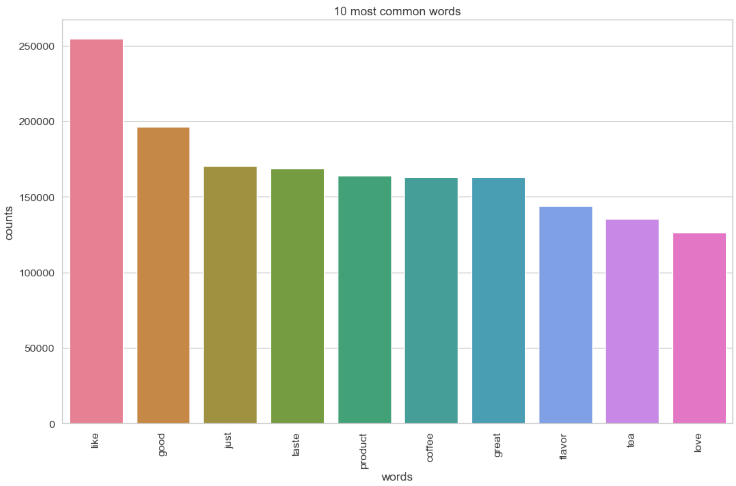

Fig. 6. Top 10 Common Words in Amazon fine food reviews

```
Topics found via LDA:

Topic #0:
food dog cat cats like eat product chicken dogs just

Topic #1:
coffee like flavor taste good chocolate just great cup love

Topic #2:
tea amazon product price great good order buy store love

Topic #3:
like tea water taste drink just flavor product sugar good

Topic #4:
like just great good eat love treats product dog little
```

Fig. 7. Topics with LDA Analysis

B. TEXT ANALYSIS

In this section, I talk about the LDA Analysis and Non-negative Matrix Factorization (NMF) methods on the reviews.

I used LDA analysis to extract the ten most common words in the fine food reviews. Figure 6 shows the top 10 most common words in the Amazon fine food reviews. The most common words include “like”, “good”, “just”, “great”, “product”, “taste”, “flavor” and “love”. However, it is interesting that “coffee” and “tea” are also among the most common words. That means people buy coffee and tea more than other fine foods on Amazon. Maybe it is because coffee and tea are more durable products, and they are not damaged as much during transportation. Maybe buying caviar, cheese, or wine as fine food is not possible on Amazon because of the transportation costs.

Figure 7 displays the topics found via LDA. Having “dog", “cat", and “chicken" in Topic 0 made me question whether the cat and dog foods can be considered fine food or not.

Figure 8 and Figure 9 show NMF for topic analysis. NMF analysis gives some more detail than LDA top 10 word analysis. Again, I see similar words such as “like”, “ taste” and “great”. In addition, “coffee” and “tea” are also in the common topics.

Moreover, some words that are used to describe coffee such as “bold”, “strong” and “roast” are among the most common words. It is also interesting to see the word “Keurig” appears on the list. Thus, people wrote reviews on Keurig brand coffee products. Similarly in Topic 3, it is possible to see the words that can be used to describe tea products such as “green” and “iced”. Like the LDA Analysis, Topic 4 made me think that dog and cat foods and treats are also counted as fine food. Topic 5 shows that people comment on shipping speed and order problems, giving more details in NMF Analysis than LDA Analysis.

Figure 10 is about the cosine similarity density plot between the review title and review itself. The density plot is skewed to the right and there are more observations when the cosine similarity is zero. Therefore, I believe it is not possible to say that the review title predicts what is written in the review. It seems like the review title and the review text are not usually relevant to each other.

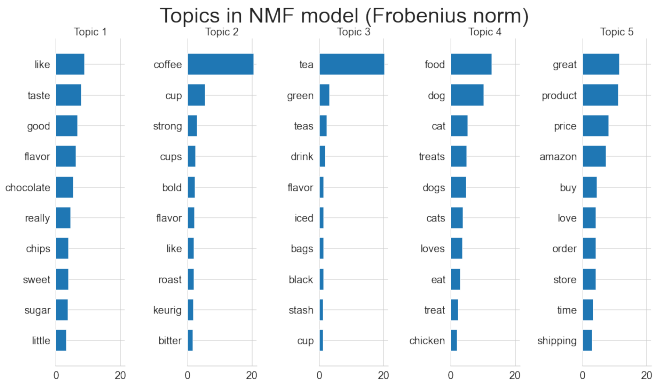

Fig. 8. Topics in NMF model (Frobenius norm)

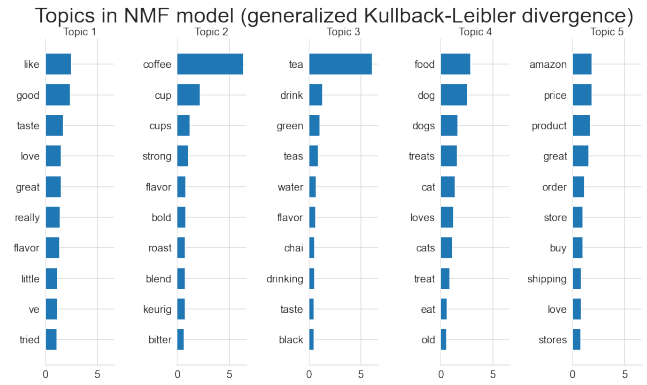

Fig. 9. Topics in NMF model (Kullback-Leibler divergence)

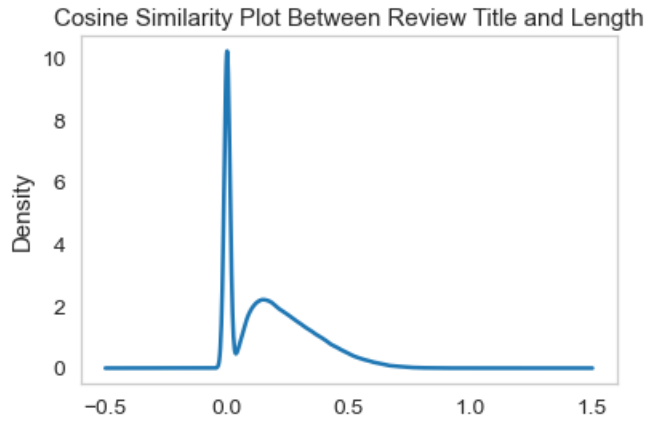

Fig. 10. Cosine Similarity

III. METHODS

I tried to predict the review scores from the available data in this part. To do so, I made Sentiment Analysis to review the title (*Summary*) and reviews (*Text*). I utilized nltk and VADER packages in Python for the Sentiment Analysis. Figure 11 shows the descriptive statistics of the sentiment analysis. I used the compound score in the models.

Next, I made a correlation matrix to avoid multicollinearity when building models. Figure 11 shows the correlation heatmap between variables of the data. Generally speaking,

	SentimentSummary	SentimentText
count	568411.00	568411.00
mean	0.32	0.69
std	0.35	0.41
min	-0.95	-1.00
25%	0.00	0.62
50%	0.44	0.86
75%	0.62	0.94
max	0.99	1.00

Fig. 11. Descriptive Statistics of Sentiment Analysis-Compound Scores

a Pearson correlation coefficient value greater than 0.7 indicates the presence of multicollinearity. Therefore, it is not a good idea to put *Helpfulness Denominator* and *Helpfulness Numerator* to the same model because their correlation is 0.97 which is extremely high.

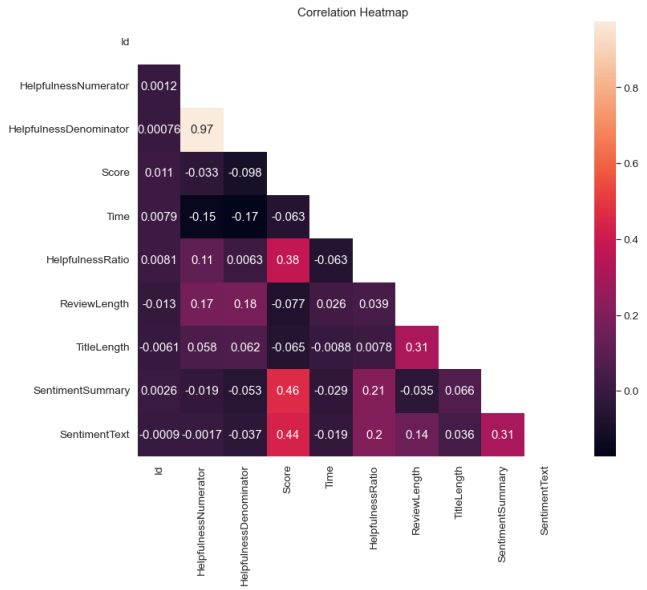

Fig. 12. Correlation Heatmap

Then, I tried Linear regression, Ridge regression and Lasso regression. I used 0-1 normalization. At the end, I have the ridge regression model with the least RMSE (0.26) with *SentimentSummary*, *HelpfulnessDenominator*, *Time*, *TitleLength*, *SentimentText*, *ReviewLength*.

The equation of my lowest RMSE model is a Ridge Regression:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 SentimentSummary + \hat{\beta}_2 HelpfulnessDenominator + \hat{\beta}_3 Time + \hat{\beta}_4 TitleLength + \hat{\beta}_5 SentimentText + \hat{\beta}_6 ReviewLength. + \hat{\epsilon}_i$$

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y \quad (1)$$

Figure 12 shows the confusion matrix of the scores of the predicted model and true scores using the Ridge regression model.

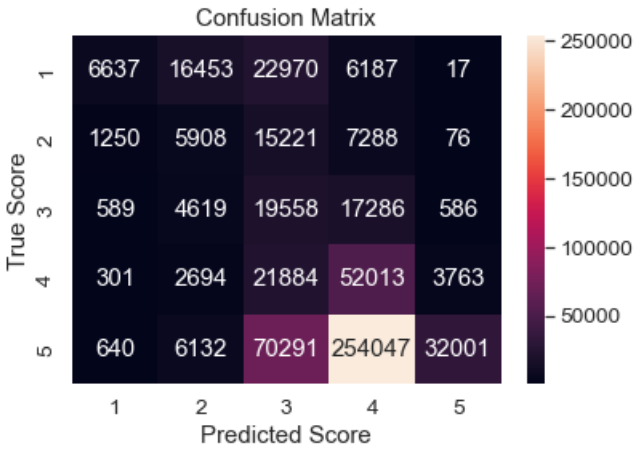

Fig. 13. Descriptive Statistics of Sentiment Analysis-Compound Scores

Finally, my analysis is missing dimension reduction methods.If I had more time, I would use PCA, t-SNE, and spectral embedding methods to have the more optimal model and reduce RMSE.