Amazon Mini Project Report

Hao Ma

Abstract— **This project is mainly about predicting review scores using review text of fine foods from Amazon. Methods used include basic exploratory analysis, NLP (natural language processing), and linear regression. Topics of the reviews and relationship between scores and reviews are explored in this project.**

I. INTRODUCTION

The data set contains information on user, product, helpfulness, score, timestamp, summary, and actual text of 568,454 reviews of fine foods from Amazon. A simple predictive model is expected to predict scores from other given information.

Basic exploratory analysis is done to better understand the data, including creating table of summary statistics, time series line graphs, and leveraging NLP packages for text analysis. 5 topics of the reviews are identified through an unsupervised model. In addition, the exploratory analysis finds that more than half of the scores are greater than 3 (a score of 4 or 5). Text length, time, and topic may affect the pattern of scores.

The next step is building linear regression models and choosing the one with lowest root mean squared error. Several models are built using different predictors which are computed by transformation, grouping (based on exploratory analysis), and NLTK sentiment analysis package. RMSE values along with confusion matrices are reported in the notebook file.

II. METHODS

The main strategy of this project is choosing and creating predictors based on exploratory or descriptive analysis and building linear regression models. The following steps are taken to get to the final model.

Firstly, load the data and compute summary statistics for summary length, text length, score, and helpfulness ratio. New columns such as text length are added to the original data set.

Secondly, create line graphs of summary length, text length, score, and helpfulness ratio aggregated by day over time. Observe any pattern or relationship between the target variable and other features.

Thirdly, write a function to clean (removing stopwords, short words, links, emojis, and punctuation) and lemmatize the review text. Then apply NLP techniques (Latent Dirichlet Allocation) to extract 5 topics from review text and apply NLTK sentiment analysis package to compute numerical sentiment score for each data entry. In addition, cosine similarity scores between summary text and actual review text are computed. It shows that summary is not predictive of the actual review text.

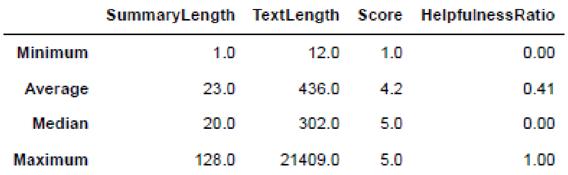
Next, take log-transformation for one of the numerical predictors to make its distribution more normal. Several models are built including ordinary least squares linear regression and Lasso & Ridge regularized regression. Models are evaluated based on root mean squared error and confusion matrix. According to the model performance, a weighted linear regression model is also built to better incorporate minority (scores other than 4).

Lastly, two new predictors are created to better predict lower scores (a score of 1, 2, or 3). Models used include ordinary least squares linear regression, Lasso, and weighted linear regression. The predictors used are topic group - indicating whether the topic of each review belongs to low-score group, product group - indicating whether the reviewed product belongs to low-score group, compound sentiment score, log-transformed text length, and time group - whether the posted date is earlier than the year of 2007. The best model is chosen to be weighted linear regression model using the above predictors.

III. RESULTS

Results obtained from exploratory analysis, NLP analysis, and modeling will be discussed in this section.

A. Summary Statistics



	SummaryLength	TextLength	Score	HelpfulnessRatio
Minimum	1.0	12.0	1.0	0.00
Average	23.0	436.0	4.2	0.41
Median	20.0	302.0	5.0	0.00
Maximum	128.0	21409.0	5.0	1.00

Fig. 1. Picture of the Summary Table

B. Time Series Line Graphs

Four line graphs with the following variables aggregated by day over time are created: summary length, text length, score, and helpfulness ratio. From the graphs below we can see that approximately after the year of 2007, text length became more stable, as well as the trend of score, so text length and whether the review time is prior to the year of 2007 would be good predictors. Based on the summary statistics above and Fig. 5, helpfulness ratio is not chosen as a predictor since there does not seem to be any relationship between score and ratio.

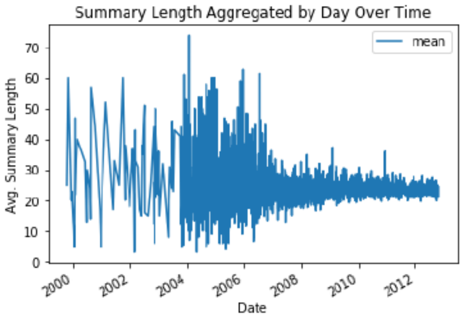


Fig. 2. Summary Length Aggregated by Day Over Time

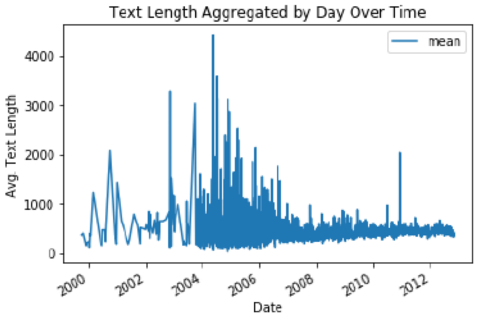


Fig. 3. Text Length Aggregated by Day Over Time

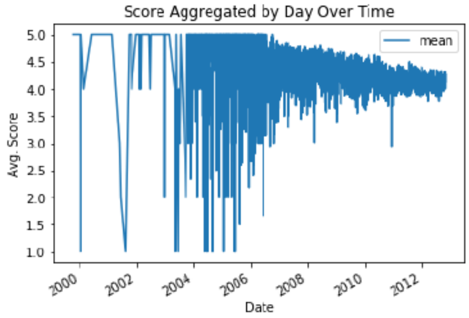


Fig. 4. Score Aggregated by Day Over Time

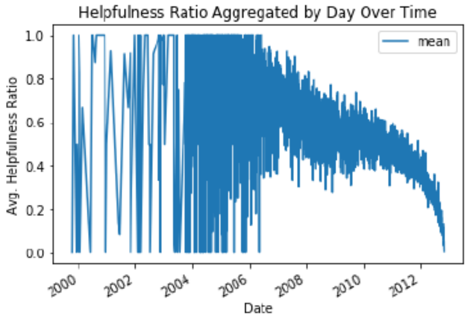


Fig. 5. Helpfulness Ratio Aggregated by Day Over Time

C. Topic Analysis - LDA (Latent Dirichlet Allocation)

Please note that this model is trained with cleaned and lemmatized review text and number of topics is set to 5. Please run the code from my notebook to see the interactive plot for intertopic distance map and top-30 most salient term for each topic.

Top 15 words from each topic:
good, like, product, great, taste, amazon, love, chocolate, price, flavor, box, store, bag, snack, bar;
food, dog, cat, treat, love, like, product, eat, good, time, year, old, day, great, chicken;
product, water, taste, like, oil, use, bottle, salt, sugar, good, drink, coconut, day, popcorn, time;
tea, flavor, like, taste, good, make, great, love, sauce, hot, use, add, really, green, ve;
coffee, cup, like, flavor, taste, good, drink, great, love, strong, tried, make, really, blend, roast;

The identified topics are: some good tasting sweet snack; some dog/cat food; some product and its ingredients; some tea drink; some coffee drink.

D. Topic Analysis - NMF (Non-negative Matrix Factorization)

Please note that this model is trained with raw data and number of topics is set to 5. Please see my notebook for visualization of the distribution of top words.

The coffee, tea, and dog food topics are similar to the previous result from the LDA model, but one of the topics seems like some mixture of topics from the previous result. Also, the other one can be discarded as it does not make sense. LDA model and its result are used for this project.

E. Cosine Similarity Analysis

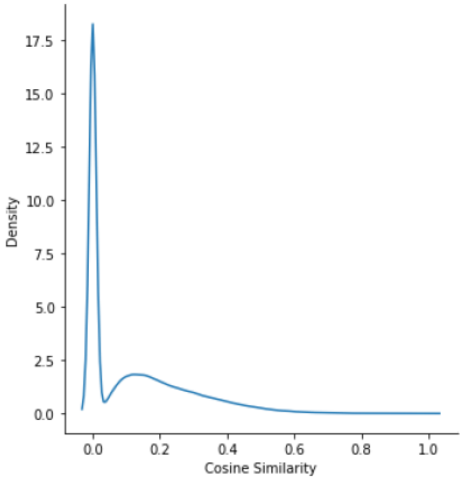
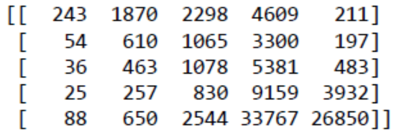


Fig. 6. Density Plot of Cosine Similarity Scores

Cosine similarity values between cleaned summary text and cleaned review text are calculated. We can see from the density plot that the cosine similarity scores are mainly distributed around 0. There are several very large values (greater than 0.9) where summary text is the same as review text or the entire summary is a part of the actual review text. Moreover, there are only 13108 scores that are greater than or equal to 0.5, so summary is not predictive of text.

F. Model Evaluation

The root mean squared error of the final model discussed in the previous section is 1.171098. A confusion matrix is also created:



```
[[  243  1870  2298  4609   211]
 [   54   610  1065  3300   197]
 [   36   463  1078  5381   483]
 [   25   257   830  9159  3932]
 [   88   650  2544 33767 26850]]
```

Fig. 7. Confusion Matrix

G. Discussion

During the model refinement process, there is one model that have a relatively high ranking on Kaggle but it is actually predicting everything to be 4, which is not doing any work. Note that the RMSE value of the final model on Kaggle (1.415) is higher than what I have in my notebook (1.171) and the reason is unknown.

Moreover, there are some potential improvements such as computing bigrams and trigrams before training the LDA model and performing grid search for number of topics and sample weights. Some classification methods or deep neiral networks may be used to get much better result, but this will be beyond the scope of this project.