Application of text analysis and a graph network for predicting ratings of grocery items on Amazon

Aaron Osher, Amanda Short, & Michael Wade
DS-5640 Big Data Scaling
Data Science Institute, Vanderbilt University
7 May 2021

Introduction

Amazon.com, Inc. currently represents the largest e-commerce business in the United States, and is quickly growing to accomplish that title globally[1]. Amazon has been offering grocery items on its platform since 2006, but a recent rise in competition from other e-retailers in the food and grocery industry has motivated Amazon to rapidly upscale and diversify this particular layer of its platform[2]. A large component of Amazon's sales platform, including in grocery items, is user ratings. Customers who purchase an item from the website are invited to provide a written review and 1-5 star rating of the product. These ratings are significant contributors to forecasting sales of specific items, both within and outside the Amazon platform[3]. As such, analysis of these ratings is a valuable exercise in understanding major economic trends, especially as they relate to e-commerce.

Our approach to analyzing reviews of food items on Amazon is twofold. First, we will attempt to predict the numeric star rating of products based on a corresponding written review. Second, we will construct a graph network representation of the grocery products offered by Amazon's online platform, which will provide insight into which products may be viewed or purchased in sequence. Each of these provides a valuable component of an overall greater understanding of grocery and food sales, as well as user behavior in providing reviews of online products and navigating a robust online sales platform such as Amazon. This information would be valuable for many applications, particularly in product marketing, where trends in customer behavior are critical to understand.

Methods

Our data[4] were provided upon request by Julian McAuley, Jianmo Ni, and the Computer Science Department at the University of California San Diego. More information about this dataset can be found here: http://jmcauley.ucsd.edu/data/amazon/index_2014.html. Their repository includes data for all categories of products sold by Amazon.com. For our project scope, we required use

[1] Garner, BA. (2018). Amazon in the global market. *Journal of Marketing and Management 9(2)*:63-73.
[2] Devi R. (2008). Amazon's foray into E-Grocery market: Successful venture. *ICFAI Journal of Consumer Behavior 2008.*
[3] Etumnu CE, Foster K, Widmar NO, Lusk JL, & Ortega DL. (2020). Does the distribution of ratings affect online grocery sales? Evidence from Amazon. *Agribusiness 36*:501-521.
[4] Ni J, Li J, & McAuley J. (2019). Justifying recommendations using distantly-labeled reviews and fined-grained aspects. *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

of reviews only within the “Grocery and Gourmet Food” category as well as their corresponding metadata. The reviews data include identifiers for each product and reviewer, the star rating, and the entire written review, among others. Metadata include descriptors of each product itself (such as brand, sub-category, and price), as well as a list of IDs for products that were also viewed or purchased after the selected product. Of the full ~50gb dataset of 233 million reviews, the selected grocery data include five million reviews of 200,000 products.

To approach our first problem, predicting rating given review text, we created a pipeline to transform the text using NLP such that it could be passed to a model. The complete pipeline included many common preprocessing features in Spark to extract and transform features. First, we took in the raw review text and used a tokenizer to split out individual words. Next, we looked at removing common stop words, however, this seemed to have a negative impact on the overall performance of the model. This may be due to the fact that some words signifying particularly positive or negative sentiment were removed thus creating some difficulty in accurately assigning a label. Similarly, we tested using n-grams (specifically two) in the pipeline, but again saw a decrease in performance. The final step of the pipeline was a count vectorizer to obtain term frequency. Once the data was transformed, we were able to pass it to the model. For the model, we used both a logistic regression model as well as random forest to perform multiclass classification. The results and conclusions are further detailed below.

One challenge that came up was the high class imbalance present in the data. Over three million of the five million reviews had five stars, causing our first run of the model to predict almost everything as five stars to maximize accuracy. However, this led to very underrepresented predictions for all other rating categories. To solve this issue, we decided to sample an equal number of reviews from each of the groups. Along with this sampling technique, we also selected a small subset of the data for training, due to the immense amount of time that it took to run because of the size of the dataset.

Using a multiclass classification approach led to poor results, especially with regard to the middle range of ratings. To deal with this challenge, we decided to bin the results such that ratings of one to three stars were assigned a “low” score, and ratings of four and five were assigned a “high score.” Similarly, we also looked at just predicting the sentiment of reviews (positive, neutral, or negative) using TextBlob. These two methods allowed us to understand how the user felt about a product which is ultimately the end goal. Essentially, we are more concerned with the overall sentiment of the customer than the precise score they gave so that we can more accurately predict future sales.

To approach our second problem, understanding relationships between products to make suggestions, first a complete network representation of the products was generated. This was done by using the product IDs as the nodes, and connecting these nodes with edges using the ‘also bought’ and ‘also viewed’ fields. The network was prohibitively large to analyze, so a subgraph was generated based on one of the node attributes. We sought to explore products bought from Trader Joe’s, so the subgraph was generated by filtering on the ‘brand’ attribute. General network statistics were generated and the subgraph was plotted for interpretation. The processing of the graph data was done using PySpark, pandas and NetworkX. The visualization was handled by Cytoscape, an offline network drawing application. Cytoscape was utilized as

NetworkX struggled to handle drawing the larger network graphs. This was another project challenge that we faced due to the size of the dataset.

Results

The logistic regression model successfully classified 39.7% of reviews correctly without binning the data. The random forest performed surprisingly worse with only 17.6% classified correctly. When binning the data, the logistic regression again performed better and obtained a 74% accuracy. In the final iteration of model testing, we increased the size of the dataset used in the model which led our accuracy to increase even more to 84%. Again, we find that the model most accurately predicts extreme values, in this case one star and five stars, while neglecting to understand what exactly should lead to a two, three, or four star review (Figure 1).

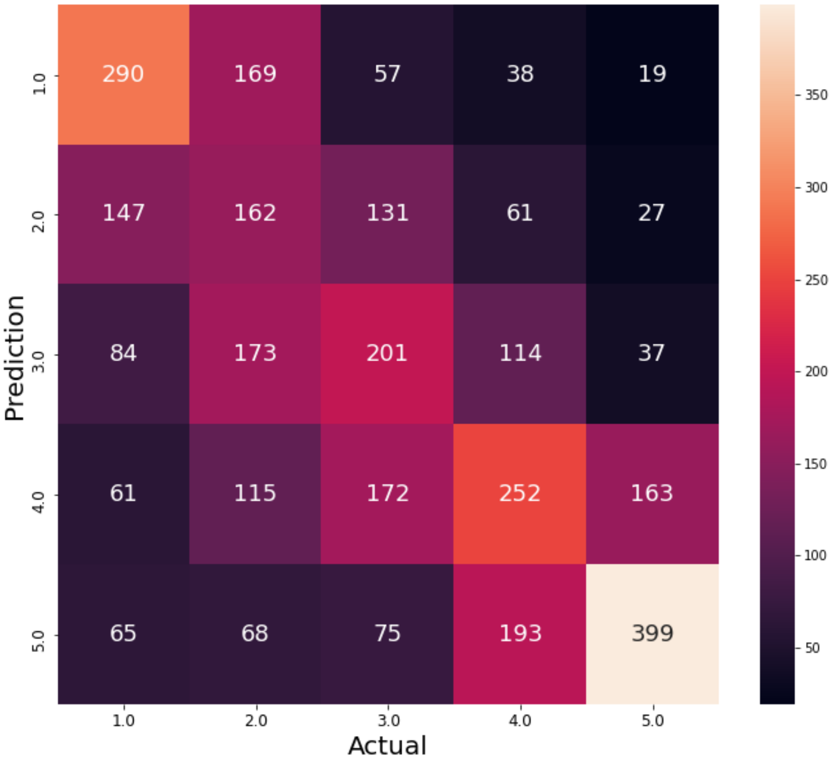
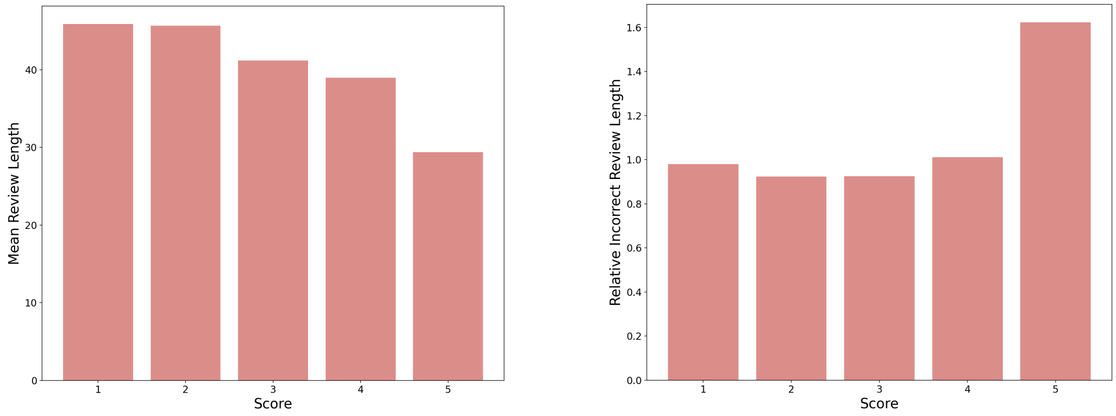


Figure 1: Confusion matrix of rating prediction

We looked into this matter further and found that the length of the text was correlated with the rating given by the user (Figure 2). Longer reviews tended to indicate worse reviews which the model may have picked up on, except in the case of five star reviews (Figure 3).



Figures 2 & 3: Length of review by predicted rating; Relative length of misclassified reviews by rating to average length of true rating

Sentiment analysis classification proved to overpredict positive reviews, despite the fact that the classes were balanced. Our expectation was that the positive and negative reviews would be about equally distributed with neutral reviews as the least represented group. However, the model seemed to incorrectly assign many review texts to a positive or neutral sentiment while assigning very few as negative (Figure 4).

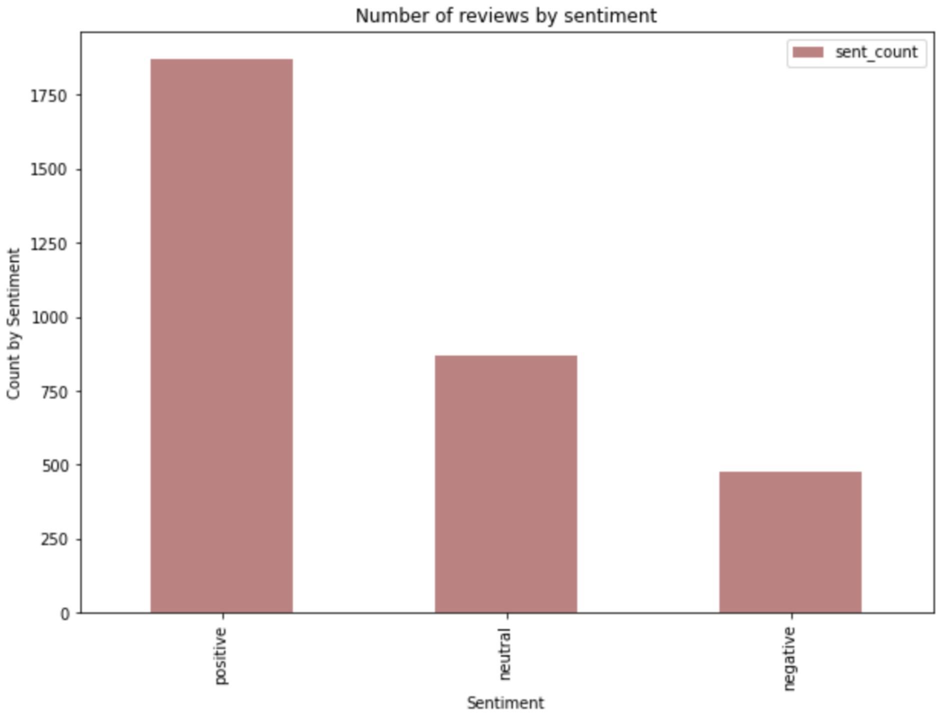


Figure 4: Predicted sentiment counts

The network analysis yielded a graph of the Trader Joe's branded products. This graph was far more diverse than expected and expanded beyond brands directly associated with Trader Joe's.

There was significant interconnectedness amongst adjacent products. Utilizing network information can yield a far more personalized online shopping experience. The results of the graph analysis suggest that making recommendations would not only be possible but likely beneficial to yielding future sales. With this in mind, expanding the dataset to include products outside of the grocery category is highly advisable in future work.

This project demonstrated that when a model is presented with longer review text, ratings are more inaccurately predicted. We also found that review length is inversely proportional to high rating. Finally, we conclude that the model performs better as a binary classifier than as a multiclass classifier. We believe that predicting rating in a binary way provides useful information in terms of future purchase predictions. The network graph can be used as a supplementary tool to create a personalized shopping experience and additionally yields more accurate recommendations. The hope is that by using these tools, sales will subsequently and consequently increase as a result.