Amazon Product Ratings Analysis

Wenxiao Li **Phuong Truong** **Pearl Otti**

Abstract

The goal of this project is to determine how ratings influence the sale of products on Amazon, and to predict future sales based on these reviews and rating. This project provides a useful tool for Amazon and its vendors to understand buyer behaviour and ensure customer satisfaction, ultimately leading to increased sales.
Its also provides direction to future customers, by helping them find the most helpful reviews to guide their purchase.

1 Problem definition

In recent times, online shopping has gained popularity due to its speed, convinience and multifarious nature. However, due to the volume of information, it has become difficult to distinguish the quality of goods and services on the internet.
Luckily, reviews offer a way for potential buyers to get a clearer idea of their items of interest: people can determine the "value" of an item based on rate of different reviews. As such, reviews are of great importance to both business and customers alike.
In this project, we will carry out simple data processes on all ratings of goods on Amazon from 1998 to 2018 in order to determine the relationship between customer purchasing habits and past reveiews.
In addition, we will build a machine learning model to help in predicting the purchasing habits of customers in the future based on past and present reviews.

2 Methodology

2.1 Dataset

We will use the dataset from Amazon Product Data (https://nijianmo.github.io/amazon/index.html). The dataset's format is an large csv file.
We only use specific categories because part of the dataset is not large enough for our project,hence we can not get useful data result from them.
Therefore, for data processing part, we use "ratings only" version as our data set which has full number of total lines.
For all categories, the table looks like:

Asin	reviewer	rate	unix time
id of product	id of reviewer	rate of product	unix time of the review

For machine learning part, since the original dataset is very large and has multiple categories, we only used a small subset of Amazon review data for the video games category, which contains 231,780 reviews, to implement and experiment the machine learning model. The data's columns are:

Column name	Description
reviewerID	ID of the reviewer
asin	ID of the product
reviewerName	name of the reviewer
helpful	helpfulness rating of the review
reviewText	text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime	time of the review (raw)

2.2 Simple Data Process

Using the SFU cluster as out running enviroment, we work on annual change of sales and average of rates first. Afterwhich, we calculate annual change of items. We use hadoop for data storage and use spark+python for data processing.
We chose this combination because it provides a simplified interface to get a quick view on what the data looks like as well as a fast enough cluster for our program at this stage.

2.3 Machine Learning

The dataset's format is a json file compressed in gzip. We specified a schema to read the data into a Spark dataframe for the machine learning model, and used the provided code from the data source to read the data into a pandas dataframe for plot and analysis. For each review, we obtained the review text, overall rating of the product, and helpfulness rating of the review. We filtered out the reviews with fewer 10 helpful ratings because these reviews are less likely to be read by customers and will negatively affect our model's performance.

Our initial assumption was that longer reviews were more detailed and that the more detailed the review, the more helpful it would be to customers when purchasing products and to Amazon vendors in improving future sales.
Therefore, we decided to add more features by calculating the review's length, sentence's count, and word's count. We also considered the overall rating in our features because products with higher overall rating are more likely to receive more helpful reviews from customers.

3 Analysis and Problem

3.1 Simple Approach

We chose "Cellphones and accessories" as our first product category because it is large enough and has been popular in recent years. We started with getting annual sales and average annual rate for this category. To do that, we build the table with 5 columns: year, average rate, amount of sales, change of rates and change of sales, where change of rates is the amount of change of average rate and change of sales is the percentage of change(0.1 means 10 percent). We get rid of years that have less than ten thousand sales in order to remove the outliers because around that time most people buy phones in the store instead of online.

In figure 1, we can see that from 2008 to 2017, the sales and ratings generally increase at the same time, and after that both are decreasing. It seems like the sales are affected by ratings, but the problem is that development of cell phones is very fast hence people in 2018 hardly buy phones that was made in 2008.
To fix this problem, we choose another category "Books", where the decreasing of values of books caused by time is relatively small. Figure 2 shows some difference with cell phones. The sales generally keep increasing till 2017 and decrease after, but the rating is up and down around 4.2 from 2000 to 2011, then increases till the end. The reason we get this result is that the average rating is relatively high and some of the low-rate book has more effect on the average than high-rate book, but the decreasing sales of low-rate book is relatively small comparing to the increasing sales of high-rate book. Because of that, it is hardly to get real information through year average. Thus, we start working on specific items.

```
{"year":2008,"avg_rate":3.596,"sales":13523,"rate_change":0.174,"sales_change":1.287}
{"year":2009,"avg_rate":3.627,"sales":22563,"rate_change":0.03,"sales_change":0.668}
{"year":2010,"avg_rate":3.63,"sales":36178,"rate_change":0.003,"sales_change":0.603}
{"year":2011,"avg_rate":3.611,"sales":52472,"rate_change":-0.019,"sales_change":0.45}
{"year":2012,"avg_rate":3.686,"sales":118519,"rate_change":0.075,"sales_change":1.259}
{"year":2013,"avg_rate":3.781,"sales":281490,"rate_change":0.095,"sales_change":1.375}
{"year":2014,"avg_rate":3.855,"sales":826050,"rate_change":0.074,"sales_change":1.935}
{"year":2015,"avg_rate":3.932,"sales":1500956,"rate_change":0.076,"sales_change":0.817}
{"year":2016,"avg_rate":3.99,"sales":2546904,"rate_change":0.058,"sales_change":0.697}
{"year":2017,"avg_rate":3.947,"sales":2602572,"rate_change":-0.044,"sales_change":0.022}
{"year":2018,"avg_rate":3.942,"sales":1461074,"rate_change":-0.005,"sales_change":-0.439}
{"year":2019,"avg_rate":3.922,"sales":590194,"rate_change":-0.02,"sales_change":-0.596}
```


Figure 1: Annual change of sales and rates for cellphones and accessories


```
{"year":1998,"avg_rate":4.446,"sales":13143,"rate_change":0.209,"sales_change":344.868}
{"year":1999,"avg_rate":4.398,"sales":62373,"rate_change":-0.047,"sales_change":3.746}
{"year":2000,"avg_rate":4.319,"sales":110486,"rate_change":-0.08,"sales_change":0.771}
{"year":2001,"avg_rate":4.284,"sales":348164,"rate_change":-0.034,"sales_change":2.151}
{"year":2002,"avg_rate":4.25,"sales":309243,"rate_change":-0.035,"sales_change":-0.112}
{"year":2003,"avg_rate":4.227,"sales":298189,"rate_change":-0.023,"sales_change":-0.036}
{"year":2004,"avg_rate":4.203,"sales":300236,"rate_change":-0.024,"sales_change":0.007}
{"year":2005,"avg_rate":4.153,"sales":343026,"rate_change":-0.049,"sales_change":0.143}
{"year":2006,"avg_rate":4.13,"sales":491534,"rate_change":-0.024,"sales_change":0.433}
{"year":2007,"avg_rate":4.179,"sales":552985,"rate_change":0.049,"sales_change":0.125}
{"year":2008,"avg_rate":4.252,"sales":724313,"rate_change":0.074,"sales_change":0.31}
{"year":2009,"avg_rate":4.231,"sales":785069,"rate_change":-0.022,"sales_change":0.084}
{"year":2010,"avg_rate":4.228,"sales":966891,"rate_change":-0.003,"sales_change":0.232}
{"year":2011,"avg_rate":4.223,"sales":1106990,"rate_change":-0.005,"sales_change":0.145}
{"year":2012,"avg_rate":4.227,"sales":1423791,"rate_change":0.005,"sales_change":0.286}
{"year":2013,"avg_rate":4.294,"sales":2388298,"rate_change":0.066,"sales_change":0.677}
{"year":2014,"avg_rate":4.38,"sales":6030546,"rate_change":0.087,"sales_change":1.525}
{"year":2015,"avg_rate":4.406,"sales":8122500,"rate_change":0.026,"sales_change":0.347}
{"year":2016,"avg_rate":4.428,"sales":8544603,"rate_change":0.022,"sales_change":0.052}
{"year":2017,"avg_rate":4.447,"sales":8091036,"rate_change":0.018,"sales_change":-0.053}
{"year":2018,"avg_rate":4.469,"sales":7383108,"rate_change":0.022,"sales_change":-0.087}
{"year":2019,"avg_rate":4.481,"sales":2915059,"rate_change":0.012,"sales_change":-0.605}
```

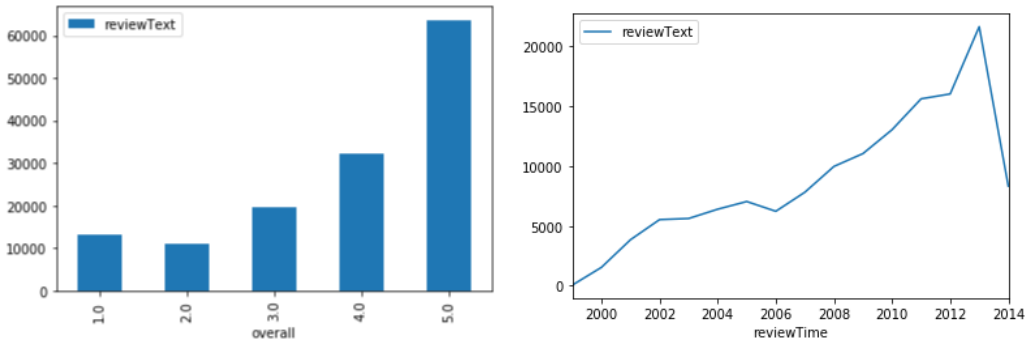

Figure 2: Annual change of sales and rates for Books

3.2 Working on items

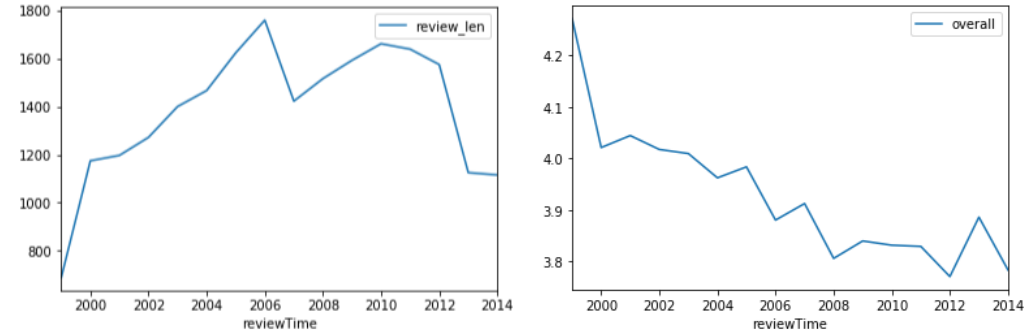
Similar to simple approach, we want to get the change of ratings and sales on different items between two consecutive years. However, due to the large amount of data, we only choose the data that has significant change in sales and ratings. We select items that has more than 0.1 annual change of rate and more than 10 percent annual change of sales. Figure 3 shows parts of our table. The result is slightly different from our simple approach, only 24.7% of items have both positive change of rates and sales, and 12.9% of items have both negative change of rates and sales. That means more than half of items has one positive and one negative. But in general, items with increasing sales have average rate of 4.35, and items with decreasing sales have average rate of 4.19.This is due to the fact that we only use ratings in our calculation, but in reality, review text is also important. Sometimes people find some small problems on goods but they may not give a low mark. However, these problems may be unacceptable and cause a customer not to purchaseb that item. Thisis a common real life situation and will affect the result of data analysis. To avoid this problem, we try to use machine learning on a more detailed data set that includes review text in it.


```
{"asin":"1447286200","this_year":2017,"last_year":2016,"rate":3.611,"sales":1009,"rate_change":0.109,"sales_change":-0.5}
{"asin":"0385537859","this_year":2015,"last_year":2014,"rate":4.029,"sales":5565,"rate_change":0.124,"sales_change":-0.523}
{"asin":"0385344430","this_year":2016,"last_year":2015,"rate":4.764,"sales":2920,"rate_change":0.106,"sales_change":-0.55}
{"asin":"0307932540","this_year":2015,"last_year":2014,"rate":4.416,"sales":1048,"rate_change":0.117,"sales_change":-0.556}
{"asin":"038568231X","this_year":2017,"last_year":2016,"rate":4.114,"sales":16350,"rate_change":0.108,"sales_change":-0.568}
{"asin":"1444779761","this_year":2018,"last_year":2017,"rate":4.038,"sales":1123,"rate_change":0.125,"sales_change":-0.577}
{"asin":"0316055433","this_year":2017,"last_year":2016,"rate":3.927,"sales":1582,"rate_change":0.156,"sales_change":-0.633}
{"asin":"0316206873","this_year":2016,"last_year":2015,"rate":4.209,"sales":1017,"rate_change":0.105,"sales_change":-0.642}
{"asin":"0593073827","this_year":2016,"last_year":2015,"rate":3.972,"sales":1984,"rate_change":0.174,"sales_change":-0.656}
{"asin":"0425266060","this_year":2015,"last_year":2014,"rate":4.532,"sales":2139,"rate_change":0.539,"sales_change":-0.712}
{"asin":"0143142372","this_year":2017,"last_year":2016,"rate":4.246,"sales":1329,"rate_change":0.242,"sales_change":-0.728}
{"asin":"B000X1MX7E","this_year":2017,"last_year":2016,"rate":4.247,"sales":2652,"rate_change":0.243,"sales_change":-0.729}
{"asin":"0385537859","this_year":2016,"last_year":2015,"rate":4.155,"sales":1490,"rate_change":0.126,"sales_change":-0.732}
{"asin":"1447287967","this_year":2017,"last_year":2016,"rate":4.492,"sales":1837,"rate_change":0.116,"sales_change":-0.736}
{"asin":"0849946158","this_year":2016,"last_year":2015,"rate":4.48,"sales":1397,"rate_change":-0.145,"sales_change":-0.796}
{"asin":"0751565350","this_year":2018,"last_year":2017,"rate":3.85,"sales":1865,"rate_change":0.229,"sales_change":-0.807}
{"asin":"0099740915","this_year":2019,"last_year":2018,"rate":3.81,"sales":1187,"rate_change":-0.154,"sales_change":-0.834}
{"asin":"0345543254","this_year":2017,"last_year":2016,"rate":3.84,"sales":1459,"rate_change":0.112,"sales_change":-0.837}
{"asin":"1101946342","this_year":2017,"last_year":2016,"rate":4.327,"sales":1830,"rate_change":0.129,"sales_change":-0.843}
{"asin":"0297859382","this_year":2017,"last_year":2016,"rate":4.103,"sales":1375,"rate_change":0.173,"sales_change":-0.856}
```


Figure 3: Annual change of sales and rates for items in Books category



(a) Left: Number of reviews by overall ratings, Right: Number of reviews from 1996 to 2014



(b) Left: Average length of reviews from 1996 to 2014, Right: Average overall rating of products from 1996 to 2014

Figure 4: Different figure for data analysis

4 Machine Learning

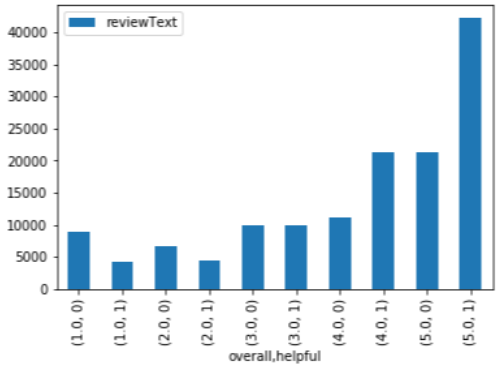
4.1 Data analysis

Before building the model, some measures were taken to better understand our data and gather useful insights from it. First, we examined how our data is distributed and discovered that the average of overall rating is 4.09. From Figure 4a-left, we can observe that there are more reviews with high ratings (4-5.0) than reviews with low ratings (1-3.0). Figure 4a-right shows the number of reviews from 1996 to 2014.
We can also observe that the number of reviews significantly increases over time. Note that the data for 2014 is only upto July.
Figure 4b-left shows the average length of reviews from 1996 to 2014. We can observe that the average length of reviews generally increases, which means that the reviews are getting more detailed over time.
Figure 4b-right shows the average ratings of products from 1996 to 2014 during which the average ratings were pretty high over time, but trending downward. It is important to note that this downward trend does not necessarily mean that the quality of products decrease over time. The lower average ratings could be due to the greater numbers of reviews and products over the years.

4.2 Classification Model

Our goal was to build a classification model to predict whether the review was helpful to customers and sellers or not. A review was considered helpful if the majority of people rate it as helpful, or specifically, the ratio of helpful rating to total rating exceeds a threshold. We used Binarizer to classify the data into 1 and 0 based on threshold. In our model, we set the threshold as 60%. We split the data into a training set and a validation set at ratio 0.75/0.25. We also used the Gradient-Boosted Trees learning algorithm for classification to create a classifier as a pipeline and printed out the evaluation score.

helpful	review_len	sentence_count	word_count
0	1298.091178	15.221390	246.864662
1	2644.876127	27.891306	493.171213



(a) Left: Average length, sentence count, word count of helpful and unhelpful reviews, Right: Number of helpful and unhelpful reviews by overall ratings

Figure 5: Different figure for result

		Predicted	
		0	1
Actual	0	1691	864
	1	707	3607

Figure 6

4.3 Results Discussion

Before running our model, we wanted to verify that our initial assumption was true. The figure 5a-left below strongly indicates that in general, the helpful reviews are much more detailed than the unhelpful reviews. In addition, Figure 5a-right shows that for high overall ratings (4-5.0), there are more helpful reviews than unhelpful reviews, while it is opposite for low overall ratings (1-3.0). Our model's score was 0.749, with the confusion matrix(Figure 6) as below: According to the result, our model successfully achieved the goal of the project. The model's score, the area under ROC by default, is 0.749, which is considered acceptable in evaluating whether or not the review is helpful. The runtime of the model is also fast. However, our model seems biased toward predicting helpful reviews because our data is unbalanced. Therefore, balancing the data should be considered. To reduce the uncertainty and improve the reliability of the model, additional features and other classification algorithms could be used in future works.

5 Conclusion

Based on our analysis, customers seriously consider indeed ratings and reviews when making a choice to purchase an item. Through out this project, we learned that the getting useful information and insight from large amounts of data is a big part of Big Data Analysis.
We need to fine-tune or simplify the data we get to make the result more readable and meaningful. The implementation of the processing code should be focus on how to transfer scattered data to some of the core data.

6 Project Summary

Getting the data: All data files were available on the website in json format compressed in gzip. No acquiring/gathering was needed. Point: 0/20
ETL: data loading, preprocessing and cleaning tasks were performed, but not much work. Point: 1/20
Problem: Problem was well defined and the overall goal was successfully achieved. Point: 9/20
Algorithmic work: Data preprocessing, features selection, and exploratory data analysis were performed. GBT classification algorithm from Spark ML was used to build the Machine Learning Model. Point: 5/20
Bigness/parallelization: The whole dataset is very large. We used small subsets of dataset to run on the cluster, but the model is scalable when applied to the larger dataset. Point: 1/20.
UI: No UI. Point: 0/20
Visualization: There are plots, tables, and screenshots to visualize results, but they could be represented more nicely. Point: 3/20
Technologies: No new technologies were used. We used technologies learned from class and assignments. Other popular tools were also used such as pandas dataframe, sklearn, matplotlib for plotting. Point: 1/20