EPFL

CS-401
Applied Data Analysis

Project

How is the trend for eco-friendly products evolving within Amazon ?

Author:
Kevin Pelletier
Maxence Petitpierre
Eliott Joulot

Supervisor:
Robert West

December 16, 2018

***Abstract*—Today more than ever, ecology is one of the key issues facing our society. This trend has gradually emerged, to become more and more essential, especially in our everyday products. By focusing on Amazon products in this study, we will first try to observe this trend, then understand it by analyzing the characteristics of these products and their consumers, and finally try to predict its evolution and to learn from it insights about our society.**

I. Introduction

This paper presents a study about the evolution of people interest concerning climate change, renewable energies, pollution and bio food using Amazon's dataset. There are many categories of products to work on in it: [Books, Kindle Store, TV-Movies, Home and Kitchen, Health - Personal Care, Tools and Home Improvement, Grocery and Gourmet Food]. Some of them need deep consideration analysis for relevant insights. We will then do some analysis and visualizations to show the evolution about the concerned products with the information provided by the huge amount of reviews and product descriptions.

The results, analysis or graph come from the main code referenced by $project\_general.pynb$, where more details can be found on our analysis, insight (code, graphs) for each section.

II. Data Collection - Dataset

The data set used is from Amazon reviews which contains products reviews and meta-data, including 142.8 million reviews spanning from May 1996 to July 2014. This year's project theme is "data science for social good", so it can be asked how data science could improve society through data analysis. The dataset contains many categories and the project will focus on few of them: [Books, Home and Kitchen, Health - Personal Care, Tools and Home Improvement, Grocery and Gourmet Food]. They have been selected by relevance and potential link to eco-products. Those categories gather over 3.200.000 products and for each at least 5 reviews. The metadata's products are separated from the reviews by categories. The total amount of data represents 30Gb.

III. Statistics - Algorithms

A. Tools and Libraries

The first step of our work is cleaning raw data. For that we use $Apache\ pyspark$. Once the data was filtered and thus more compact, we moved to $panda$ library for deeper analysis. In addition, $matplotlib$ and $seaborn$ were required for data visualization, $Scikit-learn$ for machine learning and finally $nltk$ for natural language analysis.

B. Cleaning phase

The cleaning phase consists of filtering out unnecessary information from metadatas (such as URLs images of products) and corrupted data. Once done, we save it in parquet to reuse the data very easily.
The second phase is extracting eco-friendly items from metadatas. A product is classified as eco-friendly if there is a match between its Title, Description or Brand with a word in an eco-friendly words list. This list was made by sentimental analysis plus an overview of the data. The dataset of products obtained was next joined to their respective reviews and save as parquet file. Finally, the dataset containing reviews has been converted to parquet file. At the end of the cleaning phase, we have for each category three datasets : metadatas, reviews and metadatas with reviews for eco-friendly product. This is summed up in the part 1 of the notebook.

C. Distribution of products over the years

The first part of our project was to analyze the number of eco-friendly products released every year. The release date of each product was not available, we made the assumption that the year of the first review of a product was its publication year. We then worked on both eco-friendly and general products. To do so, we grouped each review by its product ID, and aggregated taking only the minimal date. It was interesting to compare the number of products between the eco-friendly part and the general one, but we also computed the proportion to to observe the evolution within each part. We then computed these results using the histograms methods from Pandas for the first one, and from Numpy for the second one using the 'density' parameter. However, the data collection stopped on July 2014, so the results of this year could lead to some misinterpretation. That is why we implemented a linear regression model with degree 5 for each category, to define the mathematical evolution. We trained using data up to 2013, to then predict the most probable number of published products in 2014.

D. Reviews Analysis : Distribution and helpfulness

Now we are going to interest in the reviews distribution and the helpfulness associated. For the distribution of

reviews per products, we will reduce by productID all the reviews and implement a counter of reviews for all products and produce a histogram. On the other hand, regarding the helpfulness of the reviews, the ratings are in the form [3,5] where the first element represent the number of positive vote and the second is the total number of votes. We will then compute a percentage of helpfulness for each (3/5). We took care of filtering the reviews with less than 3 votes which could produce some bias. Indeed, there is a lot of reviews with no votes, or only 1 positive vote which is a bit sensitive and could produce wrong estimations. In order to have meaningful comparison plot, we transformed the rating votes histogram into percentages proportions. Finally, to be able to compare the categories we computed the mean number of reviews of the different categories, and the average helpfulness rating.

E. Price Analysis : Evolution and comparison

In this part we focused our attention to the average price of products per year. Indeed, the ability to compare the prices tendency between eco-friendly products and the averaged products, is interesting, as the products related to the bio area are usually known to be more expensive. Moreover, we will be able to draw some comparison between each category. For this task, we used the release year as a reference year for every product. We then grouped by year the dataset, aggregated by computing the mean of each group and finally outputted the histogram. Additionally, we took care of the outliers in the extracted eco-friendly products by implementing an error value for each bin. We calculated this error by taking the standard deviation of prices within the bin and dividing it by the number of products represented by this bin. Std(X) / len(X). Nevertheless, as the eco-friendly products tends to appear later in the years, we added some zeros values in the missing years to have a corresponding year index in both eco-friendly and general products. Finally, we plotted the bars representing the average price of each year of the eco-friendly products with the associated error, and the bars from the general products in the same figure. In order to have some resources to compare the different categories we computed the average price of each category by taking all the years, excluding the added zero values. Additionally, to have a more quantitative comparison, we delivered 2 dataframes outputting the differences of average prices between the eco-friendly part and the general ones of each category. The first one was meant to show the difference in percentages of the average price of the eco-friendly products compared to the general ones of the same category, whereas the second one is showing the differences in dollars.

F. Overall rating Analysis

In this part we implemented an histogram of the products ratings by counting the number of occurrences for the 5 distinct rating values. Then to combine the 5 categories we computed the mean rating of all the products of each category, separating additionally the eco-friendly part and the general one.

G. Sales Rank Correlation

This part will produce a correlation analysis between the features regarding the number of reviews, the overall rating, the sales rank position and publishing year of each products. Indeed, we could think that a popular product regarding this reviews number and overall rating should have a better position in the sales rank than another less popular. First, we grouped the reviews by product ID and then computed the average overall rating over these groups and extracted the associated sales rank, release year and number of reviews. Pearson coefficient :

$$p(X,Y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y}$$

However the result did not give any correlations. We then took the analysis a bit further by clustering the products by their number of reviews and their release year using the K-Means algorithms. We were able to see if the overall rating is correlated with the salesRank within products of the same popularity iterating over 80 clusters. Finally, after averaging the Pearson coefficient of each cluster, the result was not referencing any correlation as it was closely equal to 0. Indeed by analyzing the qualitative information, there is a lot of products with bad overall rating which have an excellent sales rank position.

H. Top Eco Brands

What are the criteria that can be use to define a brand as a top one : the number of product, reviews, grades ? and so on. Looking only at overall grades, a brand with only one product with a 5 note rating will have a higher rank than a brand with 70 reviews and a overall rating of 4.9. Actually, amazon's product grades are integers from 1 to 5 (stars). So we will round the averages overall to keep the same idea and then sort by descending order.

We wanted to look at two mains things, the average of score for all products for a brand plus the number of products. At first, group item by ids and count the number of reviews plus average score, and then we grouped by brand and compute the average of averages score and count the number of product per brand. Finally we sorted them by best score and number of reviews. Thus what we can see in Part 8 of the notebook, a left graph with top most reviewed brand and on the left the number of products available.

IV. Findings - Results

As we cannot include all graphics, you can go here to follow the process of every section and visualize the results.

A. Proportion of eco-friendly products

There is obviously a huge amount of products not concerned by the environmental aspects. The more represented category is the Garden. Books category has the smallest amount but that is not surprising. It is not very easy to find books that deal with ecological them, unless it is very specify to this domain and thus can be rare. We can ask if it is judicious to keep the books category, but 0.2% of 2.5 millions books represents 500k books which is sizable. Except books, We can observe on pie plots that the proportions are between 1.5% and 3.6% for each category.

B. Evolution of eco-friendly products by year

If we look at all histograms in log scale, we can see at first that they were not necessary present. Once they appeared on the market, although they can be in small proportion, they follow the same trend as non-eco product. First appeared in 1997 with books, then in 2000 with Home & Kitchen but we had to wait until 2004, ten years after Amazon's creation, to see eco-products belonging to Healthcare and Grocery.

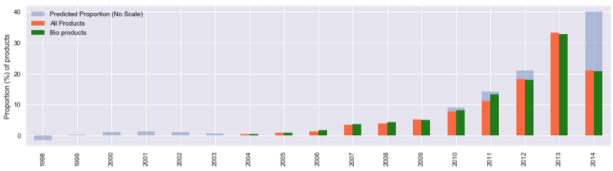

Figure 1: Product prediction per year for Healthcare

C. Distribution of the number of comments

The number of reviews whether for eco-product or not have the same curves. The most reviewed eco-products are in Home & Kitchen and Books with around 3000 comments whereas books category contains the most reviewed product so far with more than 20.000 votes.

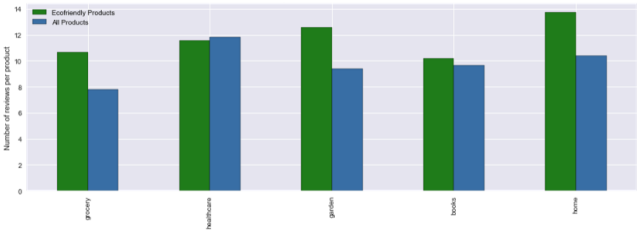

Figure 2: Average number of reviews per product

Now if we look at the figure above, we can see the average number of reviews per product per categories. Although it is eco-friendly product or not, the average helpfulness score is lower for books. This may be due to opinions that can be more personals/political/committed compared to other categories where opinions are less various and personals. Also there is a important difference of rating for grocery because people might want to leave a reviews/analysis concerning a consumable eco-product (taste, quality, additives,..) and public perception.

D. Price Analysis : Evolution and comparison

In people's minds, eco-friendly products tend to be more expensive than non-eco ones. It is not entirely false, and we came through the same result. Eco-products are on average at least 20% more expensive (see Part 4 notebook). But there is some differences within the different categories. Indeed, regarding the Healthcare category, eco-friendlies are not really more expensive than the average. In the contrary, the eco products from the garden category are way more expensive.

We wanted to verify our results and see if we can easily find some example which can illustrate it. Once on amazon.com, we searched for eco and non-eco cups. Those products can be assigned to the home category. What we found is that non-eco friendly cups cost $\$0.07/unit$ whereas eco-cups cost $\$0.15/unit$. Indeed this is more than twice the price of a non-eco product. Below the graph for Patio category where we can see relevant differences.

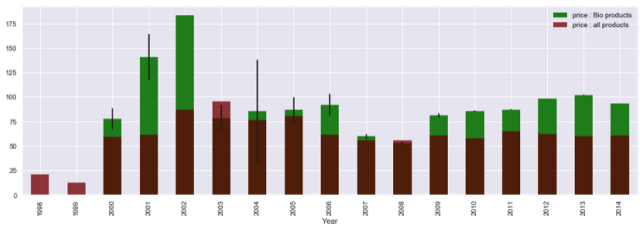

Figure 3: Garden products prices per year (in $)

E. Overall rating Analysis

By observing the mean overall rating by categories we can not establish a definitive conclusion.Indeed, the health-care seems to have highly better ratings for the eco-friendly products than the average and the grocery and the home categories ratings are slightly better, but the in the garden category the eco-friendly products seems to have a bad rating. After analyzing we noticed the correlation with the prices. Indeed within the health-care, the eco-friendly products are not really more expensive than the average products and those from the grocery and home categories are slightly more expensive. Whereas the eco-friendly products from the garden category are way more expensive. To summarize, the products overall ratings are linked to the prices, even if the products are labeled as bio.

F. Top eco-friendly product

We searched here for the best eco-friendly product in the category Garden, based on the sales rank. Te one found is a pack of 24 environmentally safe firestarters, with a sales rank of 21 for sale for 7.99$. We then searched for the equivalent firestarters available in the entire category and found 4 of them. Among them 3 were not including any environmental aspect with an equivalent or cheaper price. But their sales rank positions were far from the top product, which could be explained by the environmental safety it provides at a similar price. Indeed, the last one was environmentally friendly but way more expensive which were thus not competitive. Moreover, we wanted to compare the top product with the products of the same popularity using the K-means clusters. The products from the associated cluster were various and not really related between them, but they are all in the first 10% top sales rank and interestingly among them there are 3 articles from the same brand as the top product.

G. Top eco-friendly brands

For some categories such as Patio Lawn garden, the most reviewed brand Hydrofarm with 580 reviews is the one with the most products for sale (20). Whereas in Healthcare category, the top brand Diva Cup has only less than five products for sales (2) with more than 1000 reviews. Also brands in Home & Kitchen have a lot of reviews while the largest number of product for sale is only 6. If we go to amazon.com and search for the brand Zyliss, we only find 6 articles with a total number of reviews around 1000.

We can assume that the more a product has good reviews, the more people are willing to buy it and thus leave a good review after use (are they influenced by others' point of view ?). But that analysis does not only work for eco-friendly products but in general.

Finally we concede that we did not compute the weigthed means of the overall scores. It means that if an item with 100 reviews has a mean of 4.9, it will have the same "importance" or weight as a product with only 2 reviews and an overall mean of 1. This is what we also can observe for the brand Medline in Healthcare. They offer a lot or products but do not get a lot of reviews.

H. Mr/Mrs eco-friendly

This study was meant to find the customer who bought the largest number of eco-friendly products, combining all the categories. We worked on all the extracted environmental products reviews and grouped by the reviewer ID to get the number of products purchased by each client. We then added these results over all the categories. Finally, the top one Mr/Mrs eco-friendly was named... 'Love to Cook' and obviously most of his products are in the category Grocery food. After analyzing his reviews, the customer was indeed interested in the natural and environmental aspects of the products. Another client in the top 5 was called 'vegan compassion' and is willing to provide reviews to people taking care of the natural attribute and the packaging of the products. This slight analysis confirms the behavior of the customers from our extracted eco-friendly dataframes, who are setting an importance on the environmental aspects.

Beyond this analysis, those results could be reuse for targeted advertising,where online advertisers can target the most receptive audiences with certain traits, based on the product or person the advertiser is promoting.

V. Summary

Throughout our study, we first understood the green trend in more details by thoroughly analyzing the categories of amazon products, and the products that can be associated with this trend. Through many in-depth and well-focused studies, we were able to properly observe the evolution of this trend over time, as well as obtain a lot of information about its consumers and their behaviors. While we could expect to see general growth, we were able to detail this evolution precisely by looking at product rankings, their prices, and comparing the categories against each other.

REFERENCES

- Amazon Dataset
- Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering R. He, J. McAuley WWW, 2016 pdf