A study on Amazon's distinctive features

Raphael Barman raphael.barman@epfl.ch
Hakim Invernizzi hakim.invernizzi@epfl.ch
Albane Descombes albane.descombes@epfl.ch

Abstract

Can we really trust the Amazon's review system? This document describes and discusses the results obtained in the context of the authors' project in ADA2017. The project aims at deepening the understanding of Amazon's whole system. Particular attention is put into detecting if the system seems to be used correctly or if there's a general bias. We found out that users tend to rate very high with low variation. That Amazon 'also bought' and 'bought together' features played a big role in the ranking of the products. Finally, we also found that the more a category has reviews, the less helpful they tend to be.

1 Introduction

The analysis is centered around three different aspects.
First, the review feature, with the goal to describe its usage from the userbase and answer the question of whether reviews are generally biased or not.
Second, the adverting feature which is composed of the "also bought", "also viewed" and "bought together" features. The intra-correlation among the sub-features as well as their impact on the whole system are the subject of study.
Lastly, the categorization feature and its impact on the user perception of a product.
The three features represent three different starting points for the analysis which should ideally converge in a single conclusion about bias.

2 Related work

3 Data collection

All the data comes from the Amazon product dataset created by James McAuley (McAuley et al., 2015). We wanted to use a quite big sample, however did want to have a significant number of reviews per user and per product, thus we settled on the 5-core dataset that guarantees reviews with a minimum of 5 reviews per reviewer and 5 reviews products

4 Review feature

The first step in trying to describe Amazon's reviews was trying to detect potential correlation between its features. This was done by defining two metrics: the "wordcount", which is the length of a review in number of words, and the "helpfulness", which is a measure of the helpfulness of a review in the range [0, 1]. By observing the distribution of these two metrics it was possible to remark that the majority of reviews are less than 1000-word long and that their average helpfulness is high [add figure]. The correlation between these two metrics and other descriptive features of the review was studied. The only relevant result was noticing the presence of a weak Spearman correlation (0.26) between helpfulness and the review's rating of the product. This hint at the presence of a monotonic component in the relationship, meaning that if a review gives a high rating to the review product then the review is slightly more likely to be judged as helpful. This is not the expected behaviour of the relationship, as the product rating of a review and its helpfulness should be independent variables.

The second step was trying to quantify the bias from the reviewers. A first analysis showed that the average number of reviews written per user is low [add figure?], therefore the sample will be composed of users who have written at least 5 reviews, a sample whose size is approximately 110'000 reviewers. The average rating for each reviewer was computed [add figure], then further averaged to obtain the mean average rating. This value is of 4.23 out of 5, with a mean standard deviation of 0.9. It can be argued that the mean average rating is high while the mean standard deviation is low. Furthermore by varying the sample based on the standard deviation, it was noted that the weaker the standard deviation, the higher the average for the reviewer [add figure?]. This is a peculiar result that could be argued as pointing towards a general bias in the review system, since the only situation that justify these results is that all products are excellent, an hypothesis which doesn't hold. A possible explanation to this behaviour could be that Amazon users tend to rate the product they're satisfied with and not those they're unsatisfied with.

The third step was studying the users behaviour with regard to the brand. The approach was to select reviewers with at least 5 reviews concerning the same brand and analyze the mean average rating and mean standard deviation. These values are 4.5 and 0.63 respectively on a sample of [size]. While these values are not directly comparable to the ones obtained in the second part because they refer to samples of different sizes, the higher mean average rating and the lower mean standard deviation let think of an even higher bias in the evaluation.

Furthermore, the correlation between the ranking of a product and its reviews' quantity, average helpfulness and average rating were studied without meaningful results.

5 Advertising features

The three advertising features subject of the study are the "also bought", "bought together" and "also viewed" features. As a first approach, correlation between these three features and the product ranking have been computed. While there were no relevant results concerning "also bought" and "also viewed", a Spearman correlation of 0.41 has been detected in the relationship between the ranking of two product that were "bought together". This result implies the presence of a monotonic component in the ranking of bought together items, which means the ranking of one product influences the ranking of the other.

Another goal of the research was to examine the intra-correlation between the three features, however due to the array-like nature of these attributes and the consequent exponential growth in the size of the dataframe, this was not feasible.

6 Categorization feature

The third part of this analysis aims at assessing the impact of the product categorisation on the whole Amazon's system. To do this, products are divided by category. Then, the number of reviews per category as well as the average helpfulness and average product rating per categories are computed.

As expected, categories which are conceptually broader ("Automotive") have many more reviews compared to niche category ("Pizza Kits"). The mean average helpfulness is 0.89, highlighting the fact that the usefulness of the reviews is judged in a positive way. Concerning the mean average rating, it is close to 3.98, a result which is lined with the result obtained in the analysis of reviews.

Furthermore, there are interesting results from the point of view of Spearman correlations.
The correlation between the average helpfulness of reviews and the average rating of reviews per category is 0.29, which means that there's a weak monotonic component in this relationship, similarly to what was observed in the first part of this analysis. This results means that if a review gives a high rating to the product, then it is slightly more likely to have a higher helpfulness.
The correlation between average helpfulness and number of reviews per category is -0.45. The interpretation of this result is that if a category has many reviews, then it is less likely that the reviews are considered helpful in average.

7 Conclusions

In the first step of the analysis, a weak correlation among the helpfulness and product rating of a review was detected, when theoretically they should be independent variables. Mean helpfulness is high (0.9). It was also noted that the mean average product rating was high (4.2 out of 5) and the mean standard deviation was low (0.9), facts that lead to think that users tend to favorise giving high ratings. This was enforced when noticing that the smaller the standard deviation, the higher

the mean average rating in the study sample.
This result is peculiar because it is hard to believe that this data really reflects the quality of the observed Amazons's products. The mean average product rating is too high: supposing that low quality products also exist, it should be more towards 3.0.
A possible explanation is that users tend to review the product they're satisfied with, and are less diligent in rating the product they're not satisfied with. This behaviour would bias the data and produce data similar to the ones we observed.

In the second part of the analysis, a relevant result was produced: a moderate correlation (0.41) between the ranks of products that were bought together. This speaks of a "boost" factor in the ranking of certain products, which gained positions by benefiting from the influence of a higher ranked product. This is not a form of user bias per se, but it's most likely a bias voluntarily introduced with the "bought together" feature, which suggests users to buy products that other users bought together with the observed one. Unfortunately, it was beyond our capabilities to examine the correlation between the "bought together" and "also bought" features, which would have allowed to add informations to the picture.

In the third part of the analysis, the correlation between average helpfulness and average rating was confirmed to hold even if these values are computed per category. The interesting result produced is the correlation between the number of reviews and the average helpfulness of reviews in a category (-0.45). The interpretation of this result is that if a category has many reviews, then it is less likely that the reviews are considered helpful in average. This might makes sense, because if many reviews are available, then the user can use a comparative approach in choosing which review is most helpful; however when there's few review, the lack of information brings the user to lower its standards in appreciation.

In light of these observations, it can be affirmed that bias plays a role in the Amazon's system: may it be user introduced bias as seen in the first, or "design bias" as seen in the second part, or even bias due to a lack of information as seen in the third part.

References

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.

Unknown Unknown. 2013. Spearman's rank-order correlation - a guide to when to use it, what it does and what the assumptions are. `https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php`, Retrieved on December 19, 2017.

Unknown Unknown. 2016. Interpret the key results for correlation. `http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results`, Retrieved on December 19, 2017.

References

Laerd Statistics. *https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php*. Last visited on 17.12.2017.

Minitab Express Support. *http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results/*. Last visited on 17.12.2017.