Technical University of Denmark

Advanced Business Analytics

42578

Final Project
Executive summary: Olist

Authors:
Betlem Aguado Perez - s182245
Ignacio Agudo Rodrigo - s185227

06-12-2019

Contribution

Both students have equally developed this project. All parts have been solved, coded and written together.

1 Introduction

One of the main issues in e-commerce is the difficulty to understand the key of the success or the problems as a company. This kind of business requires a complex architecture that usually involves several kinds of customers, sellers, products or broad location coverage. In some cases, even with a good architecture, it is difficult to address the weakest points, as the number of variables can be immense.

Throughout this project, the Olist dataset is analyzed to understanding how Business Analytics can help on the comprehension of its situation, in order to create future strategies or develop hypotheses to reinforce their business.

Considering the information from the dataset, the project has used the following information:

- Prices
- Seller and customer location
- Type of payment
- Expectations (how good is described the product on the web)
- Dimensions of the product
- Category of the product
- Rating from the customer
- Dates

Having this information, two different research questions are addressed:

1. What affects the rating from the customer?
2. How can the network of customer-seller can be analyzed?

2 Customer rating score

One of the main issues of an e-commerce company is to address the dissatisfaction from the customers. It is one of the main attributes that evaluate some of the most important characteristics of the company, like their popularity, growth or customer trust. For this reason, it is so important to understand where are the weak points of the company in these terms and how they can solve these problems, or at least, to know which can be a potentially angry customer.

In order to understand the relevant attributes that model the satisfaction towards the company, it has been done one analysis of the dataset by using different classification models. This means, that would be possible to predict the review score of a customer using available data. The solution will try to allow the company to take actions for avoiding a situation where the reputation of the company can be damaged.

As a starting point, the studied variables were the first seven previously presented. The analysis has been divided into the following sub-analysis:

- Analysis of the problem as a classification problem: This means that the bins are considered independent between each other and will leave the computer understand if there is any relationship between them.

  1. Application of regular classifiers
  2. Application of advance classifiers that require more computational power.

- Analysis as a regression problem: This time, the bins are considered as integers where the relation between the number is not broken ("having a score review of 2, is closer to be 1 than 5")

- Feature engineering: Developed and study of the impact of a the variable that evaluate the timeliness of a deliver.

The conclusion of the first two implementation is that according with the results, there is no real reason to have a complex classifier. This is because the structure of the database. Almost all the ratings were really positive, between 4 and 5, which end up on a huge bias where the model tend to say that the service or was great, or was really bad (Rating "1").

During the "feature engineering", the study confirmed what was expected from the very beginning, which is the importance of the timeliness on the deliver. Customers based a lot their satisfaction on having their purchase on their hands on time, and the model proved it. As well, it is interesting to see, the lower importance of the location seller/customer. This would probably meant that the service is more or less the same for all the regions in overall. Moreover, right after the timeliness, are all the variables related to the price. This have a complete sense, because depending what a customer paid, for the shipment or the product itself , they would probably expect a better or worse service, affecting their expectations. On the following picture, is represented the importance of the most important variables for the most effective model:

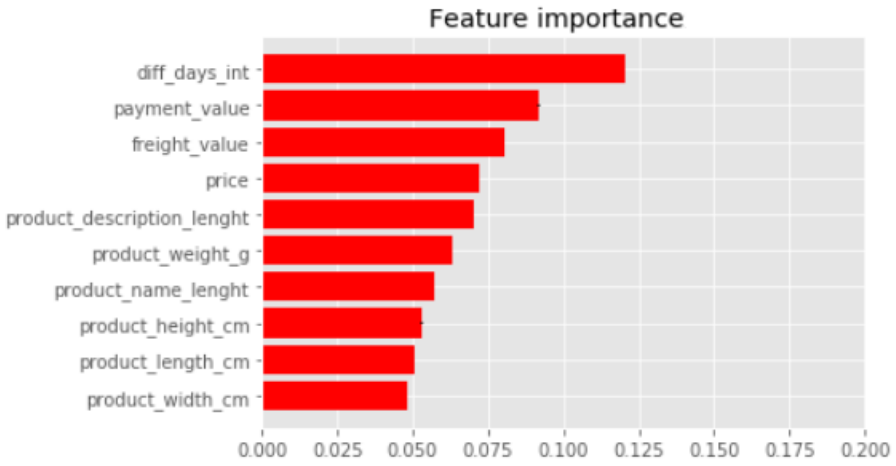

Figure 1: Importance ranking of the studied variables

3 Network customer-seller analysis

Another way of getting information from all the orders this dataset provides, is by focusing in the tendencies that make sellers popular between the customers. For taking a look of this, a network analysis was done to try to extract the top sellers, and after, check if there was any characteristics of the customers, products or something else that can give any information of

why this is.

Since there are a lot of orders with a lot of different customers and sellers, the data was filtered in a way that only customers that have done more than 3 orders from 3 different sellers at least. The connections between these customers and sellers are presented in figure 2, the most connected sellers with different customers are represented with different colors depending on its degree of connections.

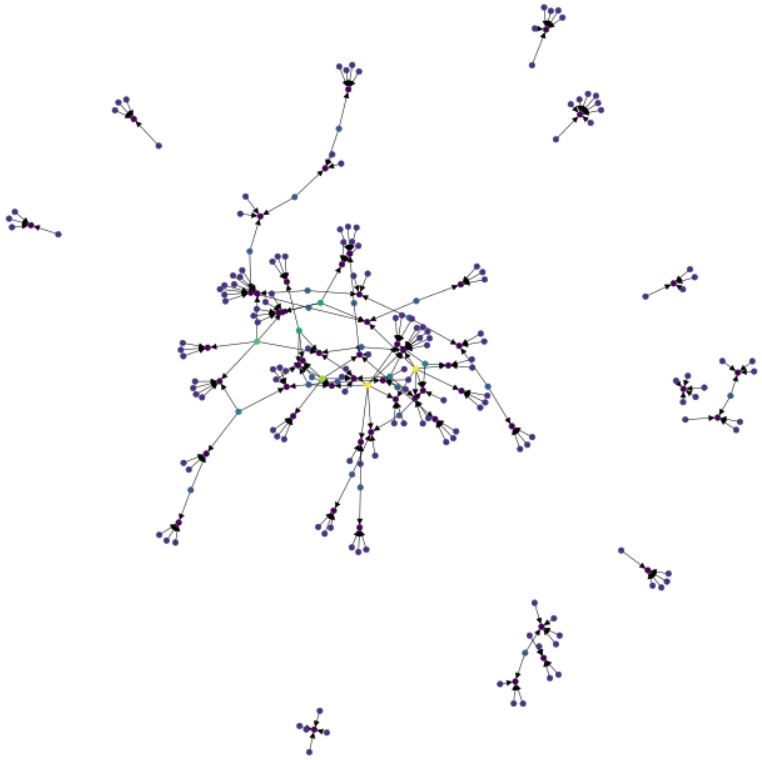

Figure 2: Network representation

This top connected sellers can be seen in the middle of the network, because they are related with a lot of different customers that have bought from some of them.

It was intended then, to try to find any set of characteristics of the customers than can be used for grouping them and set different customer segments. Unfortunately, it was not possible to determine it with the tools applied.

Trying to go further with this analysis, the list of top sellers was used to filter our dataset in order to find this set of characteristics not only with the customers but using all the information provided of each order. The segmentation was set to 3. After the convenient analysis, three characteristics raised to be the most decisive to group different orders. They are the price, the weight of the product and the length of the description.

It is possible to say that for heavy products, what usually implies a higher value, the length of the description is normally not very long, around 250 words. For cheaper prices, and small and not very weighted products a longer product description is used by most connected sellers.

4 Conclusions

Stronger models is not always the best solution and it is what this study has proved. Customers on e-commerce are really conditioned by their expectations, normally pushed from what they see on the internet, so how good is announced on the website, but as well of how much they paid. In any case, receiving with a delay a purchase can ruin absolutely everything. For this reason, the main suggestion to the company from this study is to focus and reinforce their transport network, as it will be crutial for their success.

As it has been stated previously, the dataset has a clear bias on the high-grades by having a lot of records with this rating, something that is really negative for applying predictive models. Considered this and the obtained results, to improve and expand the study would be interesting for the company to separate the datasets in ratings and study the key for each of the scores that they received separately, so maybe they can get a deeper understanding.