olist

Executive Summary

Authors

Daniel Juhász Vigild - s161749

Lau Johansson - s164512

May 21, 2020

1 Project scope

This project analyses data from the public dataset of orders placed at the Brazilian e-commerce platform Olist Store. In the project we wish to analyse customer churn: when customers have bought their first item on Olist, what has an impact on whether or not they purchase something again in the future? We wish to answer this question in a manner that is as specific and actionable as possible, and maximize potential benefit for Olist. We structure the project around the following two research questions:

- **RQ1:** What variables affect customer churn the most?
- **RQ2:** How accurately can we predict customer churn?

1.1 Methodology

Our analysis is structured around the following steps:

- Text analytics of the review messages, including topic modelling.
- Logistic regression, using variables about the first order of the customer and the topics found during text analysis.
- Build feed-forward neural networks that can be used to predict customer churn based on variables about the first order of the customer and results from the text analysis.

Text analytics presents a powerful way to gain insights to the information-rich review messages given to orders on Olist, and doing topic modelling collects information about a predetermined number of topics within these messages that can describe the general latent themes. The logistic regression is meant to evaluate and compare the effect of different variables on churn - what impacts churn the most and how? Finally, the feed-forward neural networks predicting customer churn could be used by Olist for different purposes, for example giving customers with a high churn probability a discount voucher, or directing specific product adds towards them that generate higher revenue for Olist.

1.2 Key performance indicators (KPI)

This report will aim at providing business insights regarding customer churn. As Olist is an ecommerce company, the churn rate is a KPI that belongs within sales. The company can by classifying and potentially predicting churn customers, get a better understanding of how to decrease the churn rate. Decreasing the churn, more customers will return and buy more at Olist, which then increases the revenue. If the churn rate is defined as the number of unique Olist customers which only has one order divided by the total number of unique Olist customers, the churnrate is 87.5%. A customer with more than one order spends on average 96 Brazilian Reals (114 DKK) more than a customer with only one order. Having 84075 customers churning from Olist, making them return, could bring in around 8 million Brazilian Reals (9.5 million DKK).

1.3 Data

We have narrowed the dataset to only include the 30 sellers who have received the most orders. This allows us more flexibility in terms of computational resources, but also narrows the space in which data lies. This could have the advantage of making it easier for us to interpret the results of the analysis and gain actionable insights. We have focused on the 30 most selling sellers since we assume that this subset is also a relatively interesting one for Olist in terms of revenue generated.

Since we wish to analyse churn, we limit the dataset to only include the first order of customers. We enrich this data with whether or not the customer has returned for any number of purchases on Olist. Since it is only relevant to look at customers first orders, which have had a meaningful time frame to also purchase a second time, we do not count first orders made within the final year of the dataset. The assumption here is that it takes about a year to purchase something again on Olist. The table below shows all the included variables:

Variables	Comments
Order id	
Seller id	
Price	Price of product (numerical)
Freight value	Price of freight (numerical)
Review score	Score given in review
Product name length	(numerical)
Product description length	(numerical)
Product weight	in g (numerical)
Product length	in cm (numerical)
Product height	in cm (numerical)
Product width	in cm (numerical)
Volume	in cm^3(*numerical*)
Review comment message	
Product category name	

Table 1: Variables in dataset.

2 Results

2.1 Text analysis

We did topic modelling to find latent themes across the review messages. We found the following 5 themes:

Topic	Theme
Topic 0	Belated delivery
Topic 1	Satisfied with product, but delayed delivery
Topic 2	Great product that arrived earlier than expected
Topic 3	Fast delivery and a recommendation
Topic 4	Positive comment related to the packaging

Table 2: Topics found using Latent Dirichlet Allocation.

As the table shows, the themes are very related to the delivery experience, which makes sense since it is the service that Olist provides to the customers. These topics were used to train the logistic regression and deep learning models.

2.2 Logistic regression

We built a logistic regression model using the variables from Table 1 and topics presented above [1]. By doing logistic regression we obtain comparable estimates on the effect each variable has on the probability of churn/return. This enables us to

[1]Topic 4 was omitted to comply with the comply with the assumption regarding absence of multicollinearity - but this is a technical detail

say something about was is most important for churn and in which way. We present the most important results from a model with an accuracy of 62%:

- *Price:* The higher the price of a product, the lower the probability of buying something again. Based on this, Olist could consider focusing marketing efforts towards new customer around cheaper products as to heighten the probability of the customer returning for multiple purchases.

- *Product name length:* If the first order has a long name, it appears more likely that the customer will return. The explanation behind this could be that buyers who like long product names have a tendency to buy many items. So it is not the long product name itself, that impacts customers to return, but rather an unobserved characteristic that could result in an affection for long product names and returning.

- *Watches:* While analysing a group of outliers, we found that the product category *watches* is the largest category (in products ordered) where more customers churn than return. We advice Olist to look into this, and attempt to find ways in which churn in this category could be reduced - since it would have a large effect on the overall churn rate if a general solution for second purchase after a buying a watch was found.

2.3 Deep learning models

To be able to help Olist with classifying whether an order is the first order of a customer who will comeback or not Feed Forward Neural Networks are implemented. The models takes the order variables and combines them in a non-linear complex way. The complex structure of neural networks can in some cases help making better predictions compared to models like logistic regression.

We build three different models, and their content and performance levels are summarized in the table below:

Model	**Variables**	**Performance**
FFNN1	Numeric order variables	69%
FFNN2	Numeric order variables + Topics from topic modelling	82%
FFNN3	Numeric order variables + single words	88%

Table 3: Results of FFNN models.

As can be seen in Table 3, the more information we use in the models, the higher the accuracy of the predictions. In fact we end with a prediction accuracy of almost 90% for the model that has a variable for each word used in the review messages. An accuracy of this size is quite acceptable for this type of problem.

In the following, so called "importance plots" will show which variables contributes the most to the model and thus have high predictive power.

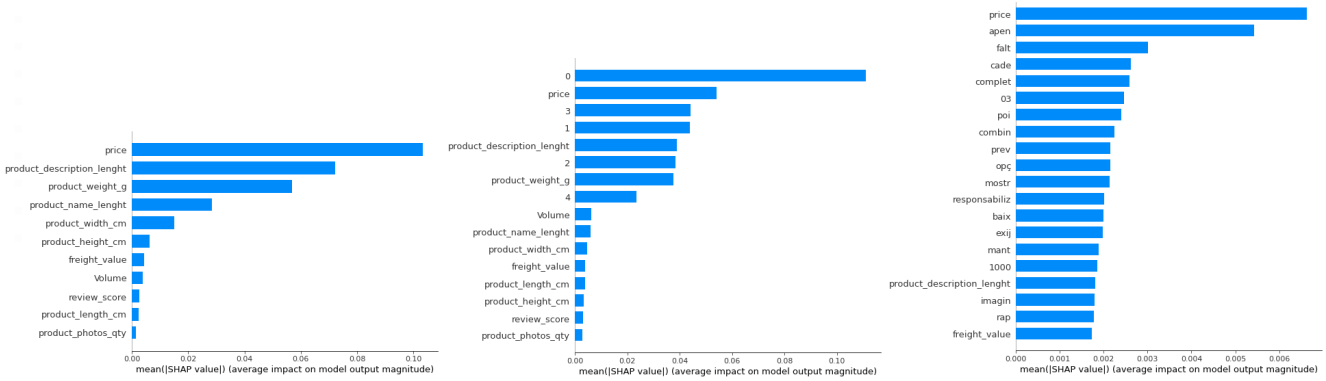

Figure 1: The importance of the variables in FFNN models. From the left: Model FFNN1, Model FFNN2 and Model FFNN3.

The most important variables from model FFNN1 are price, product description length and product weight. An in depth analysis of the variables showed, that comeback orders tend to have lower prices, smaller description lengths and heavier products. What is interesting to notice from model FFNN2 is that some of the most important variables are topics! The most important are topic 0: "Belated delivery". According to the results of this model, Olist should focus on this topic to influence the customers to return. In model FFNN3 the most important variable is the price. However, the following most important variables for customer to come back is specific words. Some of these are (translated to English): "only", "miss", "where is it", "full/complete", "combined", "accountable", "prevent", "optional", "imagine" and "demand". Looking into review comments with the portugese words "apen" (only) and "falt" (missing) some indications of unsatisfied customers pops up. Here are two of them:

"I paid for 6 chairs and delivered only 5"
"My order came incomplete. An item was missing"

The insights from looking into few of the important words for comeback customers - specifically "only" and "missing" - it seems that the customers on a general level complains about missing products. The most general tendency in the comments is that customers expecting X number of products, but receives fewer than X. Looking at the sizes of the importance values, topic 0 in model FFNN2 has a value of 0.1. The values of the words in FFNN3 is around 0.003-0.006. Olist should know that a topic can so to say "encapsulate" the importance of many words, and Olist can with more confidence base their text analytic decisions relation to reviews messages on model FFNN2.

3 Conclusion

RQ1 The results from the logistic regression showed that the price of a product lowered the probability of buying another item, that is heightened the risk of churn. Furthermore, it appeared that product name length is correlated with some highly influential, but unobserved, confounding variable, which has the effect that a longer product name of the first purchased item will result in a higher likelihood of returning. The topics found during topic modelling did not appear to have a meaningful impact when included in a logistic regression model. Based on the logistic regression it is recommended that Olist in the future targets new customer with marketing of lower priced products. The FFNN models supported the observation regarding price, but presented new information regarding the topics. In FFNN2, topics occupied three of the four most important variables in terms of prediction. This implies that the knowledge gained from topic modelling is easier accessed when using models with higher levels of complexity. Based on the insights of the neural network models, it is recommended that Olist should have focus on issues about packing and delivering of orders with more than one product associated with it. Perhaps Olist also should look into if there is a mismatch between what is registered in the system as sent and what is actually sent.

RQ2 If Olist in the future is interested in predicting comeback customers the most accurately, it is recommended to use model FFNN3 with a predicting correctness of 88%. If Olist wants to use the most computational efficient prediction model, they should choose FFNN2 (82%), at the expense of prediction correctness.