Huiying Ba | He Chen | Zhaoying Chen | Cheng-Ting Lee | Yiying Wang | Qifan Yang

Github: https://github.com/zychen96BU/BA888Spring.git

1. Problem Statement

For our Capstone Project, we analyzed Olist eCommerce platform; it is an eBay like online shopping platform in Brazil that connects all small businesses across Brazil. Olist was founded in 2015 and has been fastly booming ever since. We specifically targeted the Brazil eCommerce market, as Brazil is one of the most promising online retail industries in the world, and the largest economy in Latin America, generating more than BRL 50 billion in annual eCommerce sales. Brazil is also the 4th largest internet market in the globe; it has 150 million internet users out of a total population of more than 209 million. Brazilian online stores are booming and becoming a captivating market opportunity for merchants and online retailers in Latin America and the world.

According to André Pereira, Head of Marketing for Brazil and LA-South at SAP, Brazilian consumers are increasingly looking for convenience and comfort when they shop. This is also proven by the survey conducted by Ipsos, 80% of the Brazilians claimed that they prefer to seek out new brands and purchase products online rather than brick-and-mortar stores. By 2019, the amount of Brazilians who shop online is forecasted to grow by 18%. Due to this massive growth in the eCommerce industry, this project studies what is influencing this change, how different parties are affected by it, and how it can be improved to increase the overall efficiency of sellers and the experience of customers.

Therefore, we combined the two public datasets found on Kaggle, Brazilian eCommerce Public Dataset by Olist and Marketing Funnel Dataset, to analyze both the customers' and sellers' behavior and to provide Olist with more effective insight to advance its business.

And our project goals are as following:

1. Use text analysis to study the customer's sentiment and attitude towards product purchases.
2. Map out customer and seller distribution to find the heavily trafficked area in Brazil.
3. Learn each seller's sales performance, customer purchase behavior, most popular payment method, customer satisfaction, and order distribution.

2. Dataset

https://www.kaggle.com/olistbr/marketing-funnel-olist
https://www.kaggle.com/olistbr/brazilian-ecommerce/home

Our Capstone Project investigates Brazil's eCommerce sector by combining two datasets, Brazilian eCommerce Public Dataset by Olist and Marketing Funnel Dataset. We linked the two datasets together by using the primary key, seller_id.

Brazilian eCommerce Public Dataset consists of over 100,000 online transactions between 2016 and 2018. The data was collected by Brazil's largest department store, Olist, it connects with numerous small businesses over Brazil who sell their products through Olist stores and ship them directly to customers using Olist logistics partners. The dataset contains 30 variables on the information of the transaction, product, seller, and customer from 27 different states in Brazil.

Data in the Marketing Funnel Dataset was collected from sellers that filled-in requests of contact to sell their products on the Olist eCommerce store. The dataset consists of 8,000 leads that requested to join the Olist eCommerce platform between 2017 and 2018. And they were randomly selected from the total leads. Marketing funnel dataset is featured to allow views of a sales process from multiple dimensions, such as business type, average stock, lead category, catalog size, behavior profile and etc.

3. Methods

We aim to help Olist company to develop its market share globally by attracting more sellers and buyers to the platform. We plan to make recommendations from three angles: sellers, customers, and Olist website platform.

First, our team used Tableau to do the Exploratory Data Analysis to visually summarize and understand the main characteristics of the eCommerce dataset and market funnel dataset. We viewed orders from multiple dimensions: order status, price, payment, and product attributes to customer and seller location, customer reviews, and ultimately marketing channel. Next, we did supervised and unsupervised machine learning in R. For unsupervised machine learning, we used text analysis to gain valuable service insights from comments written by the customer. The top topics returned from the results reflect the problems that the sellers have. For supervised machine learning, we first went over all of the variables and made hypotheses on the factors that would affect review scores. Then we used forward and backward methods to prove their significance. Finally, we built a model with significant variables by applying random forest and XGboosting. The model would help the sellers predict their review scores.

4. Exploratory Data Analysis

Our dataset includes data from mid-2016 to mid-2018. When analyzing the trend of orders in different months, we chose to select 2017 the whole year. As a fast booming business that was founded in 2015, the overall sales volume showed an upward trend. In addition, from *Figure 1*, we can see that the second half of the year performed much better due to more national holidays from October to December. In December, the number of orders reached a peak.

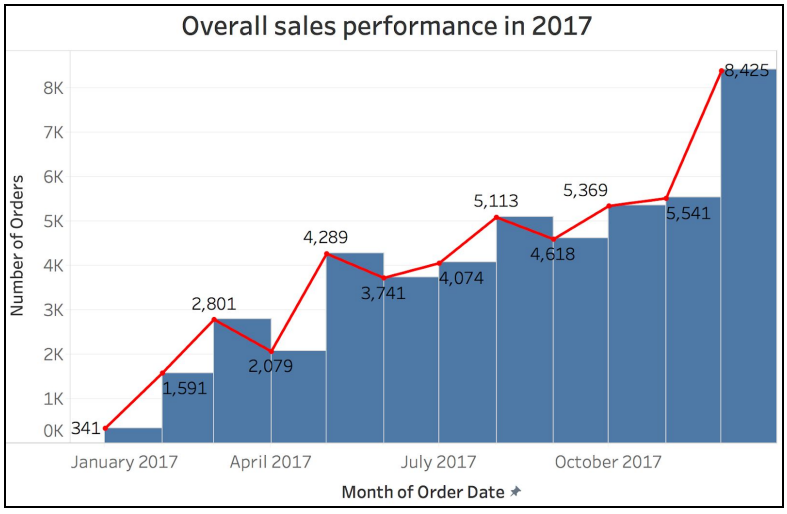

Figure 1. 2017 Sales Performance

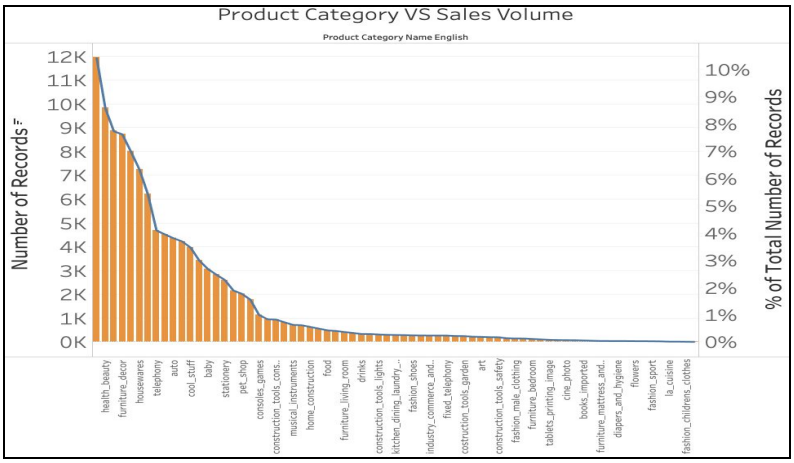

Figure 2. Product Category Needs

We divide the products into different categories and rank the categories from the highest payment value to the lowest. It is obvious that the bar chart is right-skewed. In other words, 20% of the products brought nearly 80% of total revenue, which means that the sales volume is seriously polarized. The revenue from some top several categories is over $10k, while most other categories of products are even lower than $1k. It shows that a limited selection of product types may lead to insufficient products to meet customer needs.

We then combined the number of orders, sellers, customers with geographic factors, and get graphs below in *Figure 3*. We listed the top 5 states with the highest number of orders based on each unique seller and we found that the number of orders in Sao Paulo is much higher than in other states. Then we pin-pointed the cities with sellers, and the result shows that most sellers are in the wealthy southeastern parts of Brazil, including SP. Based on the number of orders placed by customers, the state with the highest number of orders is still Sao Paulo. However, the difference between the top 1 state and the other states is not as big as the sellers. The map shows that Olist is more popular on the east coast of Brazil than in inland areas.

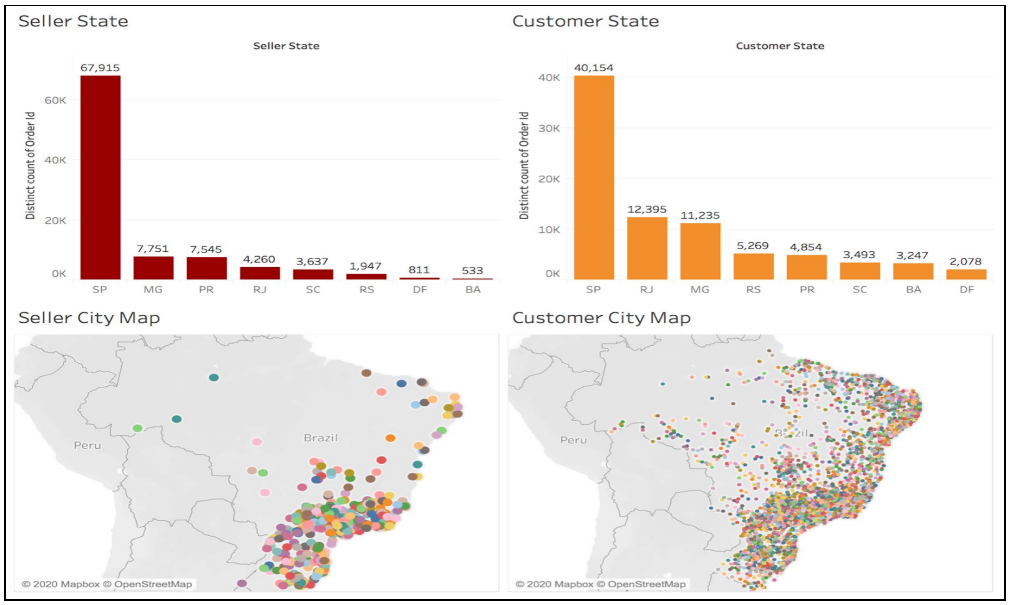

Figure 3. Seller and Customer Location

The second most relevant topic is the amount of payment. To figure out more about the sellers' performance and customers' performance, we did a deep exploration of the preferred payment method, hotcakes, and product ratings. Compared to other payment methods, consumption methods of credit cards and Boleto are more popular among Brazilians. Boleto is a payment method in Brazil and can be paid at ATMs, branch facilities, and internet banking of any Bank. In terms of monthly payment value, Boleto and credit cards grow more rapidly than others. Since people can pay online or cash offline by Boleto, many Brazilians find this method more convenient. Additionally, the payment types vary among 5 popular product categories, as watches are often imported from other countries, payment of credit cards is much higher than Boleto. If an international business tries to do business on Olist, we would suggest they provide Boleto as one of the payment methods to better assist Brazilian customers.

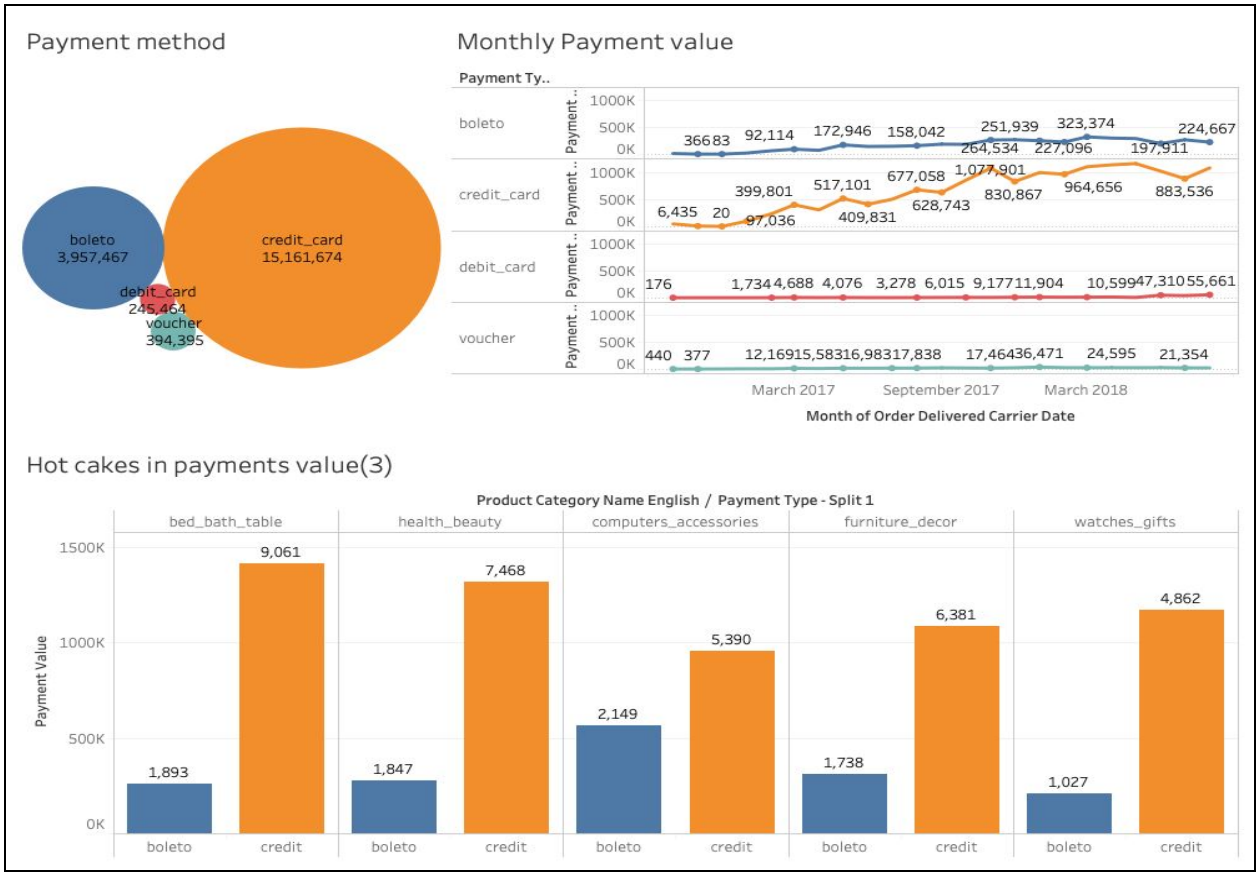

Figure 4. Payment Methods

The difference in hot cakes not only reflects on payment methods but also customers' satisfaction scores and monthly sales. With the gradual development of online platforms, the sales of bed_bath_table are more stable and the category of health/beauty is supported by steady growth. Since bed_bath_table is generally consumable which won't change a lot in the normal stage. Customers who have spare money will pursue a healthy life so health/beauty has the latent capacity to get more customers. Compared to the other popular products, the change in computer/accessories is most obvious. Since March is the new term of schools in Brazil, the sales are quite high in that period. Many students will purchase school supplies to support their studies. As mentioned before, SP, MG, and PR are the three states with the highest sales in all product categories and the average score is similar as well. Moreover, Brazilians prefer to shop on Olist during weekdays, especially Monday and Tuesday. Sellers can drive more traffic onto their site by placing more advertisements and to push more product recommendations during peak days.

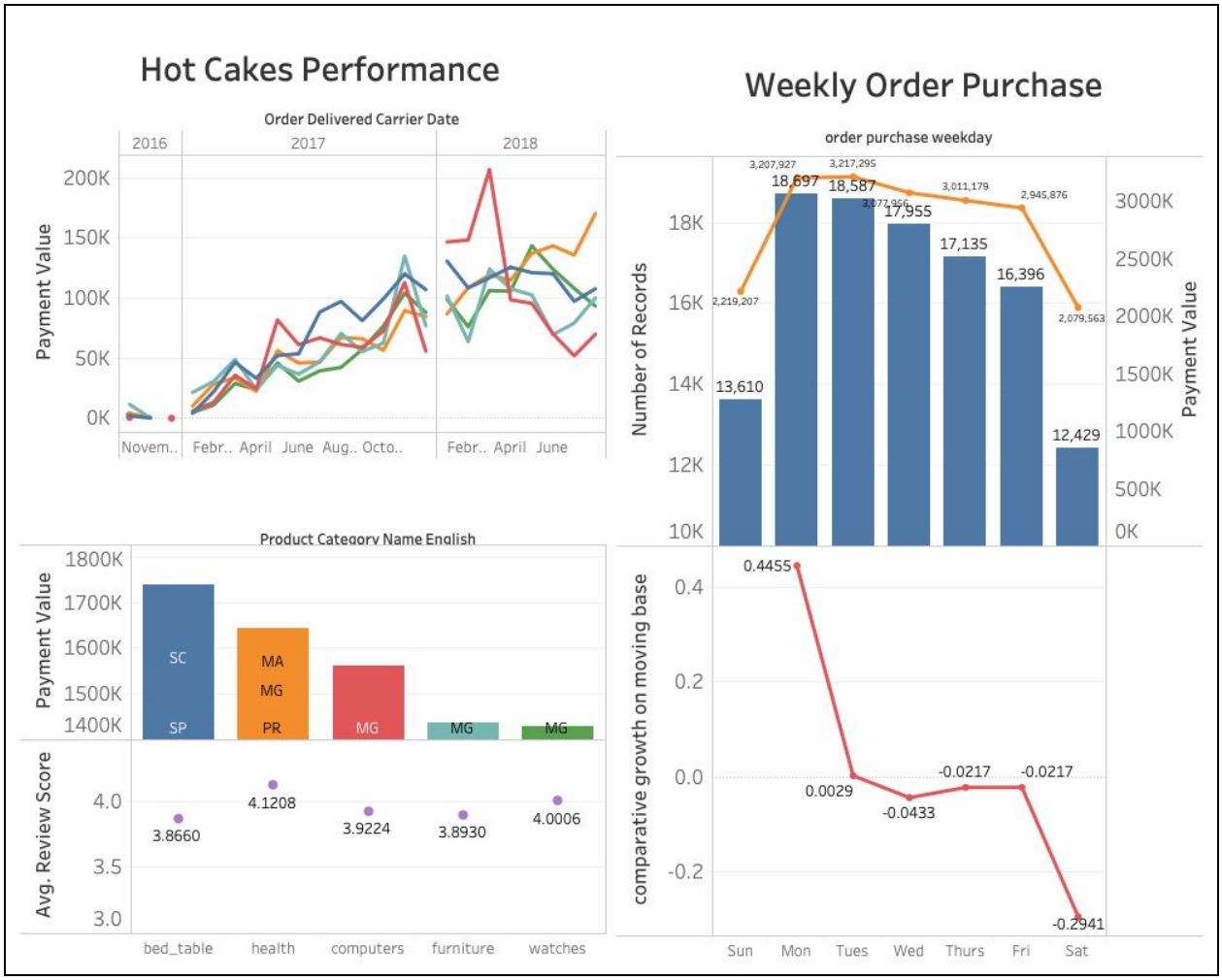

Figure 5.

5. Results

5.1 Regression Model

We built a Random Forest Model on the review score (Y). Following is the whole process that introduces how we eliminate meaningless variables, check MSE, and how we use the model to make recommendations to the Olist.

We first made a hypothesis that “price”, “freight _value”, “delivered_difftime”, “delivered_days”, “payment_value”, “product_category_name”, “product_photos_qty”, “product_description_length”, “product_name_length” are the main variables that affect review scores. We applied a linear regression to

see whether the coefficient of each variable is significant. We found the results in *Figure 6* match our hypothesis.

```
Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.891e+00  2.870e-01  17.043  < 2e-16 ***
declared_monthly_revenue           NA         NA      NA       NA
price                       1.068e-03  1.591e-04   6.714 2.19e-11 ***
freight_value               3.369e-03  1.684e-03   2.000  0.04559 *
product_name_lenght         1.633e-03  2.026e-03   0.806  0.42024
product_description_lenght  5.265e-05  3.281e-05   1.605  0.10857
product_photos_qty          2.350e-02  1.282e-02   1.833  0.06689 .
product_weight_g           -3.459e-05  1.019e-05  -3.396  0.00069 ***
product_length_cm          -3.209e-03  2.068e-03  -1.551  0.12090
product_height_cm           1.391e-03  2.212e-03   0.629  0.52936
product_width_cm            6.463e-03  3.009e-03   2.148  0.03177 *
payment_sequential         -1.104e-01  4.621e-02  -2.390  0.01688 *
payment_installments       -1.779e-02  7.972e-03  -2.232  0.02567 *
payment_value              -1.131e-03  1.216e-04  -9.305  < 2e-16 ***
mean_lat                    8.836e-03  4.865e-03   1.816  0.06940 .
mean_long                   7.087e-04  6.163e-03   0.115  0.90846
major_state                -7.047e-02  5.969e-02  -1.181  0.23784
deliverd_difftime          -3.721e-03  2.304e-03  -1.615  0.10639
delivered_days             -5.364e-02  3.222e-03 -16.646  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Coefficient Significance

Second, we applied all variables into a forward selection (*Figure 7*) and backward selection (*Figure 8*) model to filter out the significant variables. Because we have over 20 variables in our dataset, then eliminating unrelated variables is a very important step to increase the working efficiency and accuracy of the model. We found that after adding delivered_days, payment_value, price, product_weight_g, the MSE decreased significantly. Combining these results with linear regression results, we decided to eliminate other meaningless variables from the model. The variables that we finally used in the model are: price, product_weight_g, payment_value, delivered_days, freight_value, product_width_cm, payment_sequential, and payment_installments. The graph below shows how the MSE changes when each factor is added to the model.

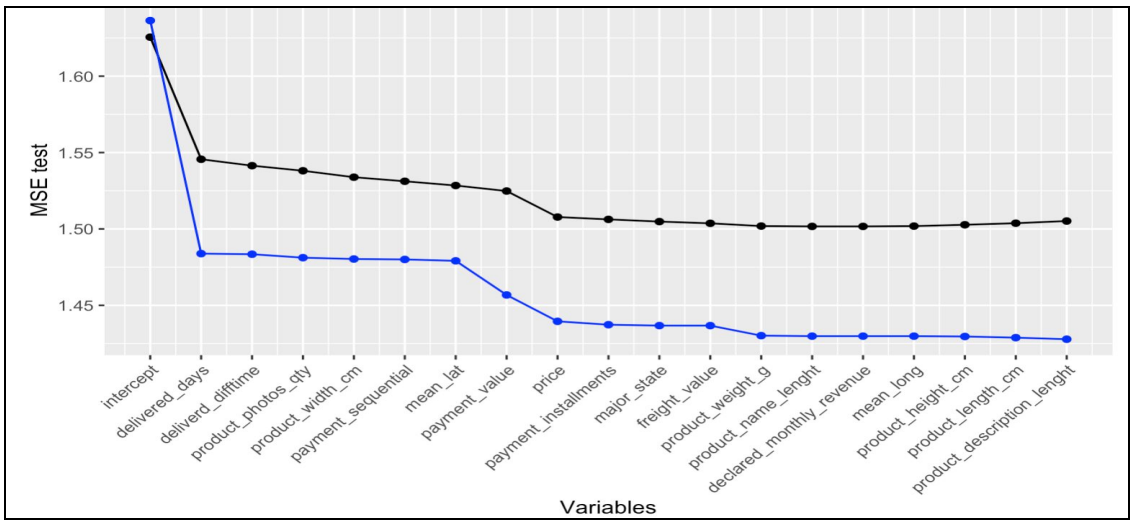

Figure 7. Forward Selection

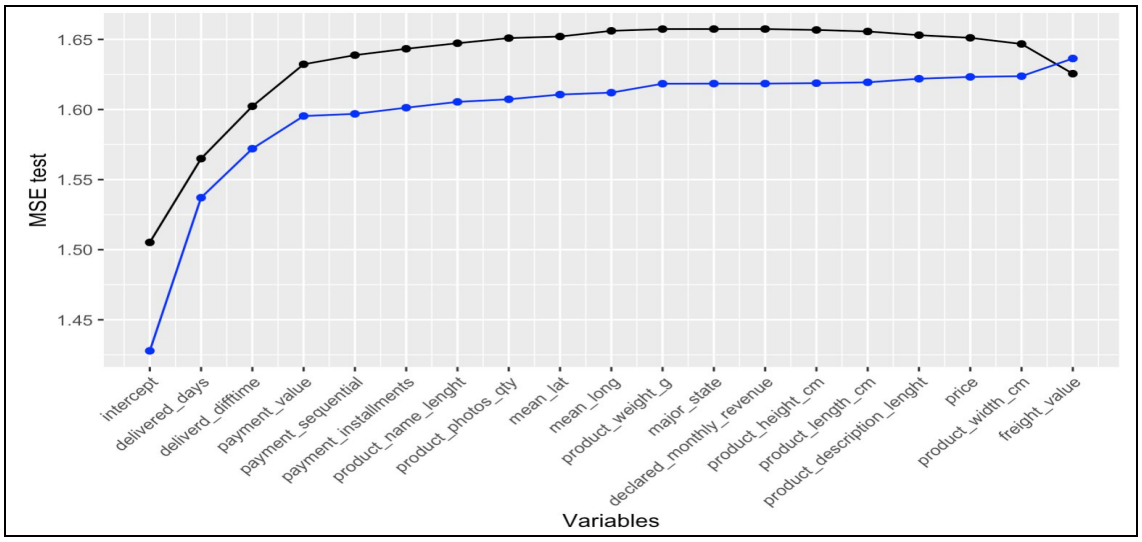

Figure 8. Backward Selection

Finally, we built the review score model with the chosen significant variables. We tried different models, such as Random Forest and XG Boosting. Given the results in *Figure 9*, Random Forest has the lowest MSE (TrainMSE: 0.21, TestMSE: 1.12), so we recommend Olist sellers to use Random Forest to do the prediction. In addition, the Random Forest can tell which variables weighted more in the model by its important function. The IncNodePurity (*Figure 10*) shows the importance of each variable in the model. We recommend the sellers pay more attention to the top four variables − delivery days, payment value, freight value, and price since any improvement on them can effectively increase the review score.

	Train MSE	Test MSE
Linear Regression	1.42	1.51
RandomForest	0.21	1.12
XGBoosting	1.19	1.42

Figure 9. MSE comparison

	IncNodePurity	rownames(importance)
1	1365.29563	delivered_days
5	888.79392	payment_value
2	755.50874	freight_value
4	727.27485	price
3	638.52315	product_weight_g
6	469.62263	product_width_cm
7	314.89008	payment_installments
8	57.12157	payment_sequential

Figure 10. Variables importance

5.2 Text Analysis

In order to avoid any loss of information during translation, we did the text analysis in Portuguese from 36,000 customer reviews in order to prove our regression model. We first applied K-means and set k = 2; however, the results in *Figure 11* shows terms on both groups are too similar. Instead, we directly pointed out the main topics reflected in the word cloud plot, *Figure 12*. We found that product quality and delivery status are the two most important topics that customers care about. Both of them are the variables in our review score model, so the text analysis further supports our model's accuracy. The result also indicates that on-time delivery and product quality have a great impact on customer satisfaction and higher review scores. Thus, customers with a positive shopping experience on Olist would be more likely to recommend the platform and products to others.

	Topic 1	Topic 2
[1,]	"entrega"	"produto"
[2,]	"bom"	"antes"
[3,]	"prazo"	"prazo"
[4,]	"produto"	"recebi"
[5,]	"chegou"	"recomendo"
[6,]	"veio"	"chegou"
[7,]	"bem"	"entregue"
[8,]	"excelente"	"tudo"
[9,]	"qualidade"	"compra"
[10,]	"comprei"	"ótimo"

Figure 11: K-means

Figure 12: Word Cloud

6. Discussion around your results

After a thorough exploratory data analysis, the top five popular product categories are healthy and beauty, watches and gifts, housewares, pet shops, and sports leisure. The use of Boleto has become more popular in 2017. We suggest sellers who want to join Olist open the Boleto payment method. Moreover, most orders were placed on Monday and Tuesday, unexpectedly the purchase orders on the weekend were lower. We also learned that Olist customers have been continuously shopping except for sleeping time. Since our dataset only recorded transactions from mid-2016 to mid-2018, we mainly analyzed the year 2017 data since we are able to have a more complete view of the performance. One limitation is that there is no other year data to compare so that we can not find any trends between years or do time series.

When we did the text analysis, we found that there were two clusters and the topics of these two clusters are too closely related, both containing the keywords of delivery and quality. As we only 36,000 reviews out of the 100,000 transactions, low accuracy or bias may occur. This problem can be solved by carefully tokenizing the data and acquiring more text data.

If we can get more years of data, our model will be more accurate. For example, different events and holidays in one year have different demand and supply problems, we need another year's data as reference and comparison. We can not use only one year's data to train models because the result is inaccurate.

7. Criticism of the results and future work

Given the current datasets acquired from Kaggle, our team has experienced several limitations when analyzing the data. First, the transaction data were only recorded from mid-2016 to mid-2018, which is not up to date. For future work, we will try our best to network with professionals from Olist to get support from them, including updated datasets to ensure the accuracy of our analysis. Secondly, in the Brazilian eCommerce dataset, we have over 100,000 transactions; however, there are a lot of NAs where we have to drop some rows and use the median to replace some of the NAs. This might result in lower precision and accuracy of the analysis.

Moreover, we do not have the revenue or sales variable to build supervised machine learning on revenue to predict the future fluctuation of the Olist business. This is definitely one piece of information that we will try to get from Olist in the future. Furthermore, our model on Review Score can also help Olist to further predict the review score of its products; however, we need more data to complete that step.

8. Conclusion and recommendations

To improve Olist Sellers service:

- Improve their current service quality through product service and logistics service in order to gain higher review scores, by setting up social media care and live chat to increase the efficiency of customer support and to help with more urgent cases.
- Sellers should provide survey and after-sales service to gain direct feedback from customers and to build a closer connection with them to increase customer retention rates.
- Sellers should also cultivate new products to enrich the product categories to meet customer needs.
- Sellers should consider to carry out more promotional activities over the weekend to boost weekend sales.
- Sellers should provide bundles options and buy one get one free strategy to attract more customers and increase sales per customer.

To increase Olist customer satisfaction:

- New coming sellers should provide Boleto as one of the payment methods, as many of the customers find this very convenient. More convenient method options can help to increase the speed of payment, simplify user's online shopping process, lower the user's online shopping threshold, and shorten the online shopping delivery time.
- Encourage customers to leave comments on the product that they purchased and to include a picture of the product, this will be very helpful when future customers are considering the same product. This will help sellers and Olist to identify their problems and be able to solve them promptly.

To increase Olist Platform overall quality:

- Though 2017 overall sales performance was on a rise, most of the orders and sales happened in cities with large populations. We recommend Olist to stabilize these areas and improve the sales of high product category marketing rates, but the platform should expand the market in small cities at the same time.
- Olist can recruit more diversified sellers on its platform to increase its diversity and to attract more potential buyers.
- Olist can partner with its sellers to release a reward system that can award customers for their purchase and give them points for the money spend. This may increase traffic to Olist website, increase sales, increase user stickiness, and retention rates.