**Executive summary**

**Introduction-** OList, a Brazilian platform integrator for e-commerce, Tiago Dalvi established Olist in February 2015. On the one hand, the OList focuses on merchants that wish to promote their products on marketplaces such as Mercado Livre, B2W, Via Varejo, and Amazon. On the other side, it consolidates all vendors' items into a single shop visible to the final buyer. After obtaining $186 million in Series E funding led by Wellington Management, the company reported a $1.5 billion value. They are market integrators focusing on developing an API interface between their system and numerous marketplaces. In this way, we can vouch for the firm's data, and the data we acquired from Kaggle is open source and published by the company, and it covers a variety of areas ranging from customer purchasing to logistics.

**Problem statement- We already have from our previous paper**

**Research methodology**

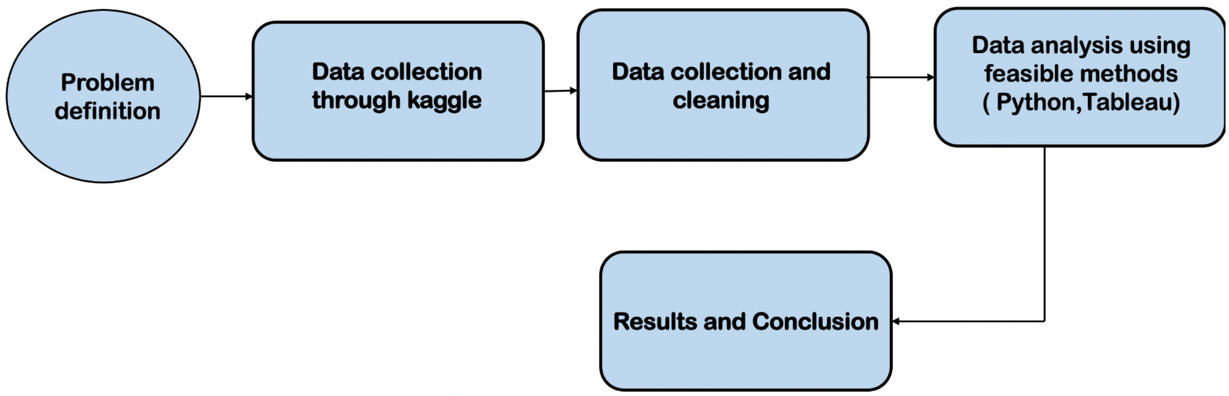
Our primary research question consisted in attempting to estimate the Net Promoter Score, or NPS, for OList. In order to understand NPS, we wanted to create a classification model that was able to predict which users would end up becoming *Promoters,* and which users would end up as *Detractors.* The relevance of investigating Net Promoter Score for the platform is high since the problem we identified for OList has to do with customer satisfaction and loyalty, and how this would translate into a strategy for them to grow their market share as an online selling platform. What Net Promoter Score allows us to do is to help the company identify what factors contribute most to the key driver of growth— the customers who not only return again to buy and sell, but would recommend the platform to their friends, or fellow vendors[1]. Through our initial data exploration we found that the review score served as a useful proxy to estimate OList's NPS. We binned users according to two broad categories— users who rated 4 and above out of 5 were coined "Promoters", and users who rated 3 and below were named "Detractors". We made the assumption that generally users who tend to rate 4 and above on platforms could be counted on to recommend to their friends since their experiences were more likely to be good, and those who rated lower were likely to report experiences where they either would not be willing to provide a recommendation, or go so far as to recommend against the platform. Given that both a lack of recommendations and negative recommendations hurt the brand in almost equal ways, if the goal is to generate a user base with high loyalty, we could treat these two categories as homogenous, as far as OList were concerned. After creating a model that would successfully identify promoters from detractors, our plan was to estimate a NPS for a given sample (test/train split) based upon the model's predicted true positives for each category. Once accomplished, we could use the NPS formula to calculate an estimate:

NPS = Promoters% of Sample - Detractors% of Sample

As for investigating what components may contribute in classifying promoters vs detractors we decided early in the process to use, in addition to logistic regression, some form of a tree-based classifier, a supervised tree model where data is split on according to a certain parameter. Not only would this produce a model with a high degree of accuracy relative to a logistic regression model, but it would also help us better identify which features of the model best contribute to our prediction success—thereby helping OList identify what areas of their platform needed improving. In addition, a tree-based model would also be suitable for any non-linearity in the data unlike the logistic regressor.

We produced the following flow chart documenting the development steps in our data analytics approach:

[1] Frederick F. Reicheld, Harvard Business Review (2003).




**Sample size:**

The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to latitude and longitude coordinates.

**Gnatt chart- yet to be made**

**SMART FRAMEWORK**

**SMART Analysis**

| | |
|---|---|
| Specific | We intend to enhance the promoter score while decreasing the detractors by improving logistics and customer satisfaction. |
| Measurable | We conducted exploratory analysis and discovered that consumers prefer to pay using credit cards rather than vouchers or boletta. As a result, we want to enhance cash-backs, discounts, and VIP access to promoters while the sale is ongoing. |
| Attainable | We need more genuine sellers on the platform to help us boost the review score, which allows the promoter to make a large purchase on the platform. We want to enhance rebates, make it easier to return items, and provide a smoother delivery experience. |
| Relevant | To achieve the above statements, we would like to collaborate with more logistic firms and distributors for services. We also intend to collaborate with banks on the rebates program plan to improve marketing for both ends.<br>This October, we plan to boost customer satisfaction by 20%. We also intend to leverage SEO insights to improve user and navigation experience. |
| Timely | We have received over 1 lakh 38 thousand orders in the last three years. We intend to increase the number by 30% in the following year, and we aim to employ the techniques above to serve as a better marketplace intermediary between buyers and sellers. |

**Conclusions and recommendations-ppt**

Based upon the results of our model we developed the following insights and recommendations for OList to improve their customer loyalty and satisfaction:

**Number 1: Improve the Delivery Logistic Management**

The common highest-importance feature that appears in both test and train set models is the delivery time difference measured in hours, or delivery delta in hours. What we can conclude from this result is that a major decider in classifying a promoter from a detractor was the time it took for OList to hand off the product from their warehouse to the delivery partner, and the final delivery date at the customer's address. Generally, we might say that longer wait times on these deliveries may contribute to a negative perception of the platform and may mark someone out as a detractor, and vice versa for Promoters. Wang et al note in their review of the Post-Purchase influences on customer loyalty for E-commerce businesses that shorter fulfillment times (regardless of whether they are advertised or unadvertised) contributes to the perceived risk of the E-commerce platform[2]. A higher perceived risk amongst customers reduces

2

customer loyalty. Our model corroborates their analysis given the feature importance plot for both data sets.

Our recommendation to OList is to develop a strategic partnership with their logistic partners to increase delivery efficiency, or reduce fulfillment time for each order, which in turn will result in more brand Promoters. We also recommend further studies using numeric regression between review score and delivery time differentials across different product groups to investigate the magnitude of the impact that this delivery time delta plays in increasing review scores, and ultimately in increasing Net Promoters.

In this regard, OList is already set to act upon the insight uncovered from analysis of our model's results with their recent announcement of acquiring the logistics company PAX, which may allow OList the opportunity to vertically integrate their fulfillment operations and improve brand loyalty further.

**2. Leverage Customer Comments for Platform Improvement**

The second insight the model's feature importance plot generated for us is that the review time delta, or the difference between when a review is posted and when it is responded to also highly contributes to the model's predictive capability. What we are able to find is that for those customers who did leave reviews and comments, it is likely that a shorter review time may contribute to a better review score, and higher likelihood of being a promoter. The model does not currently provide the ability to investigate the magnitude of the impact, and that would be an area of recommendation that we have for OList—further exploration of the relationship between review score and the contents of the reviews may yield interesting results. Likewise, we also recommend OList attempt further analysis of the comments using natural language processing, as the review time differentials may be connected with what kinds of content or keywords customers may be leaving for the vendor, platform, and service overall. Better understanding of these comments may in turn help identify what areas of the platform to highlight in promotional and marketing campaigns, as well as identify what platform features need fixing.

**3. Improve Understanding of Geographic Features on Customer Loyalty**

From our Tableau visualizations we were able to identify the highest value regions with regards to contribution to OList's Net Promoter Score. However, we currently lack the ability within this model to identify and understand how geography actually influences Review Score or helps classify promoters from detractors. Our Tableau visualization identifies the São Paulo region as the number one contributor to Review Scores, and therefore NPS. Our recommendation to OList is to further investigate the relationship between consumer's geographic distribution and their likelihood to be a promoter, or review score.

**References-**


Reichheld, Frederick F. 2015. "The One Number You Need to Grow." *Harvard Business Review*. Harvard Business School. July 16. https://hbr.org/2003/12/the-one-number-you-need-to-grow.

**Wang, Siwei, M. Meral Anitsal and Ismet Anitsal. "A Review of E-Commerce: The Influence of Post-Purchase Factors on Relationships between Customer Loyalty and Perceived Risk." (2016).**

**Hall, Christine. 2021. "Brazil's Olist Gets Its Horn with New $186M Funding Round." *TechCrunch*. TechCrunch. December 15. https://techcrunch.com/2021/12/15/brazils-olist-gets-its-horn-with-new-186m-funding-round/.**

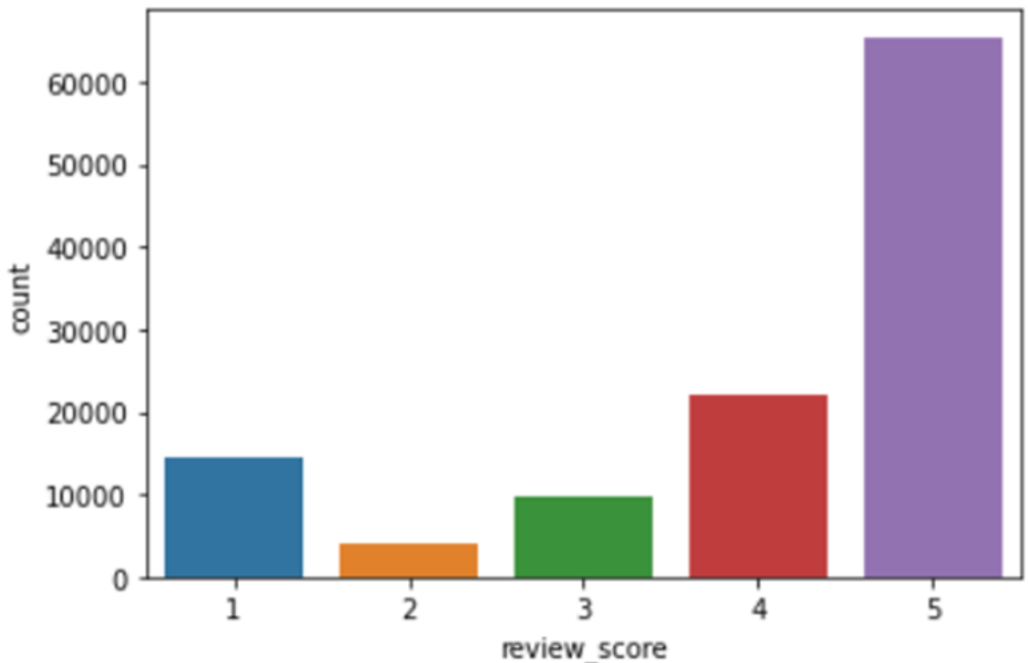
**Data Pre-processing**

After merging all the tables available we started the data cleaning and data wrangling process by handling the null values. Our dataset contained 117,329 observations with a null value rate under 3%. The columns containing null values were non-numerical columns, such as review titles, review comments, and DateTime data which can not be imputed. We dropped these columns along with some irrelevant columns such as product length and product weight which do not provide meaningful information on calculating the net promoter score of the customers.

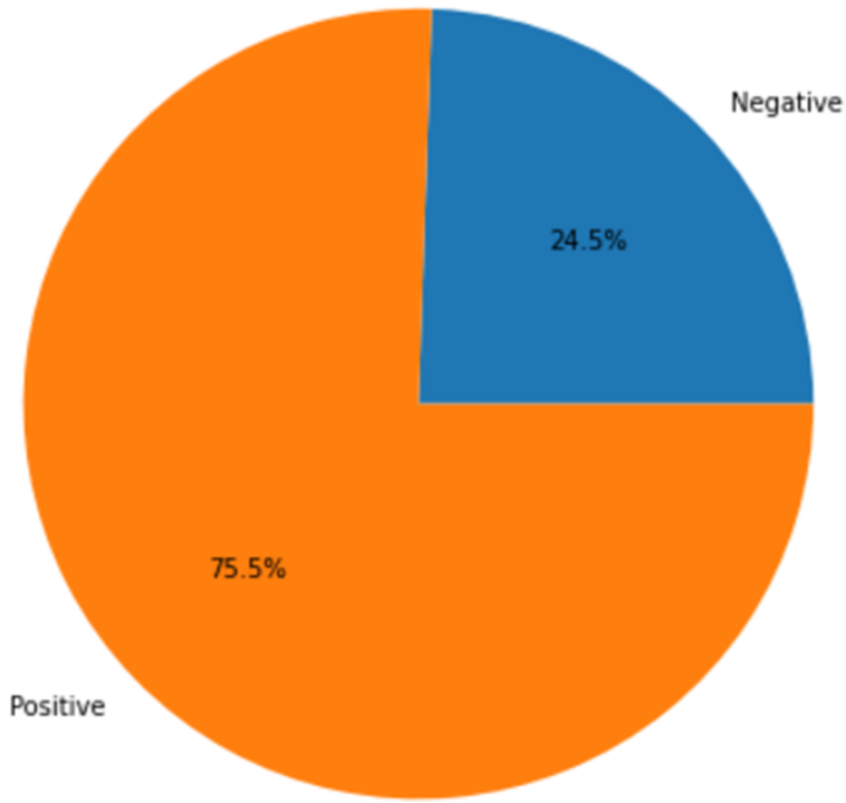
To acquire more information from the dataset, we extracted the date information and created extra columns that contain time delta information. The time difference between purchase and delivery gives us the number of hours and days of delivery delay for each order. The payment approval time difference gives us the number of hours and days of payment approval delay for each payment.

**Basic Data Exploration:**

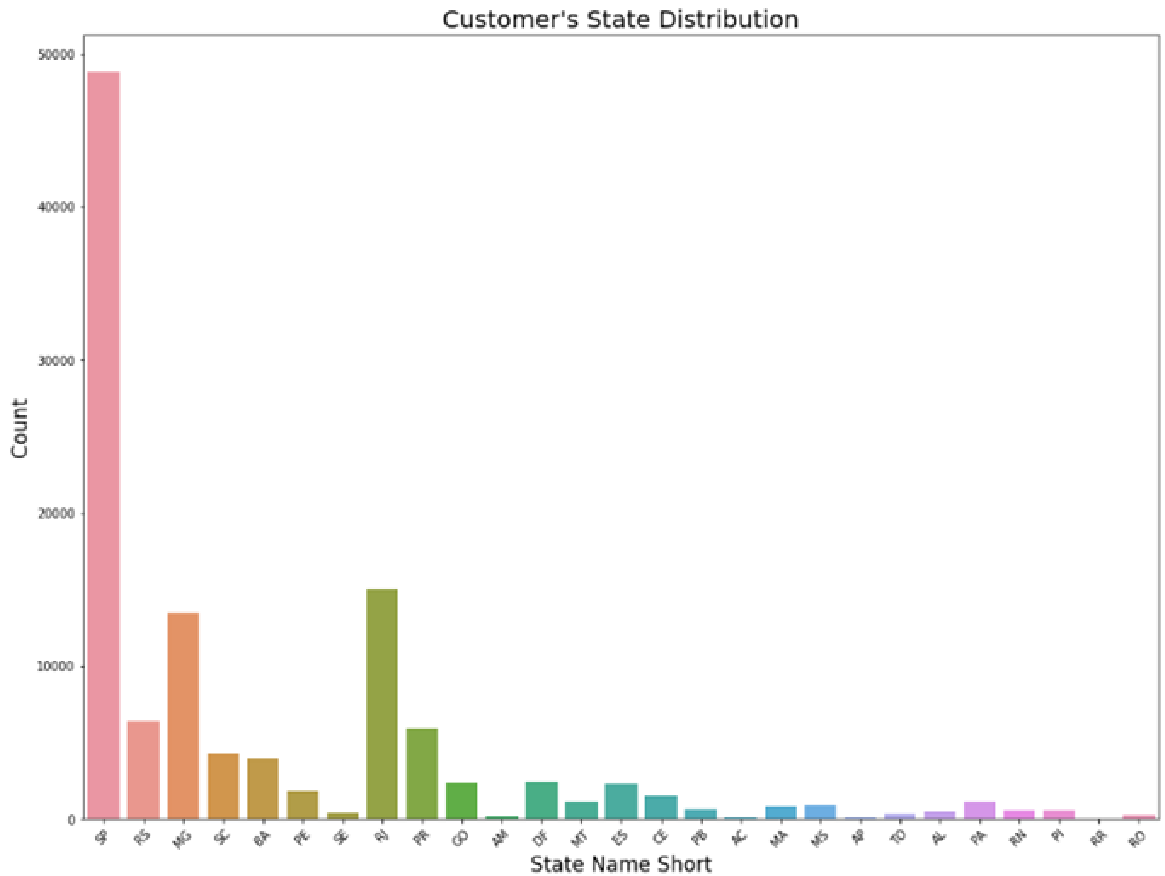
After data preprocessing, we have moved on to exploratory analysis. Below is an overview of the distribution of the review score. The range of the review score is from one to five, with most of the review scores concentrated on five.



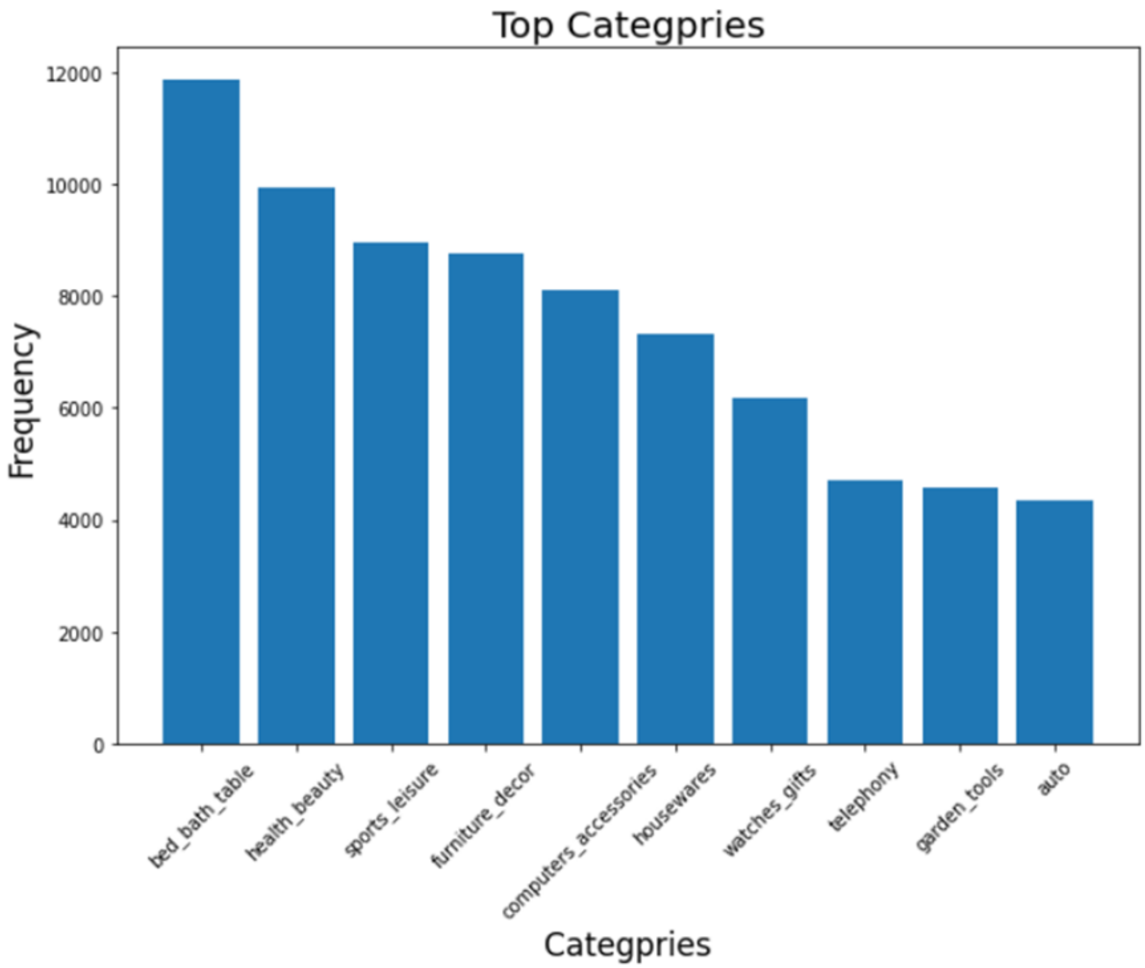

We converted this review score variable into a binary variable for the purposes of calculating an NPS with predicted values. We created a new column that contains only zeros and ones. Review scores in the range of one to three are labeled as zero (detractor), and those above three are labeled as one (refer as a promoter). On plotting the pie chart showing the percentage distribution of promoter and detractor, we see that the promoter accounts for 75.5% of the real data, whereas the detractor accounts for 24.5%. It indicates that the entire dataset is imbalanced, and we need to pay attention to the number of false positives for the modeling result.




After exploring the distribution of the review score, we looked at where the orders came from to study where our target customers are located. The plot below shows the distribution of states from where the customers shopped.. Most of the customer's shopped from Sao Paulo, Rio de Janeiro, and Minas Gerais with Sao Paulo being the most targeted place. It might be because Olist’s headquarter is based in Sao Paulo, so most of the customers living in Sao Paulo are aware of the services of Olist.




We further explored the most frequently bought categories to study customer preferences. The most popular category contains products of beds, baths and tables. The second popular category contains products of health and beauty. The category ‘bed bath table’ is significantly more popular than most other categories. It is the only category whose sales have crossed 10000 units. The category in the 12th position ‘auto’ only had around 4356 units sold.



Top Categpries
Frequency
12000
10000
8000
6000
4000
2000
0
bed_bath_table
health_beauty
sports_leisure
furniture_decor
computers_accessories
housewares
watches_gifts
telephony
garden_tools
auto
Categpries

| | product_category_name_english | Frequency |
|---|---|---|
| **7** | bed_bath_table | 11847 |
| **43** | health_beauty | 9944 |
| **65** | sports_leisure | 8942 |
| **39** | furniture_decor | 8743 |
| **15** | computers_accessories | 8105 |
| **49** | housewares | 7331 |
| **70** | watches_gifts | 6161 |
| **68** | telephony | 4692 |
| **42** | garden_tools | 4558 |
| **5** | auto | 4356 |

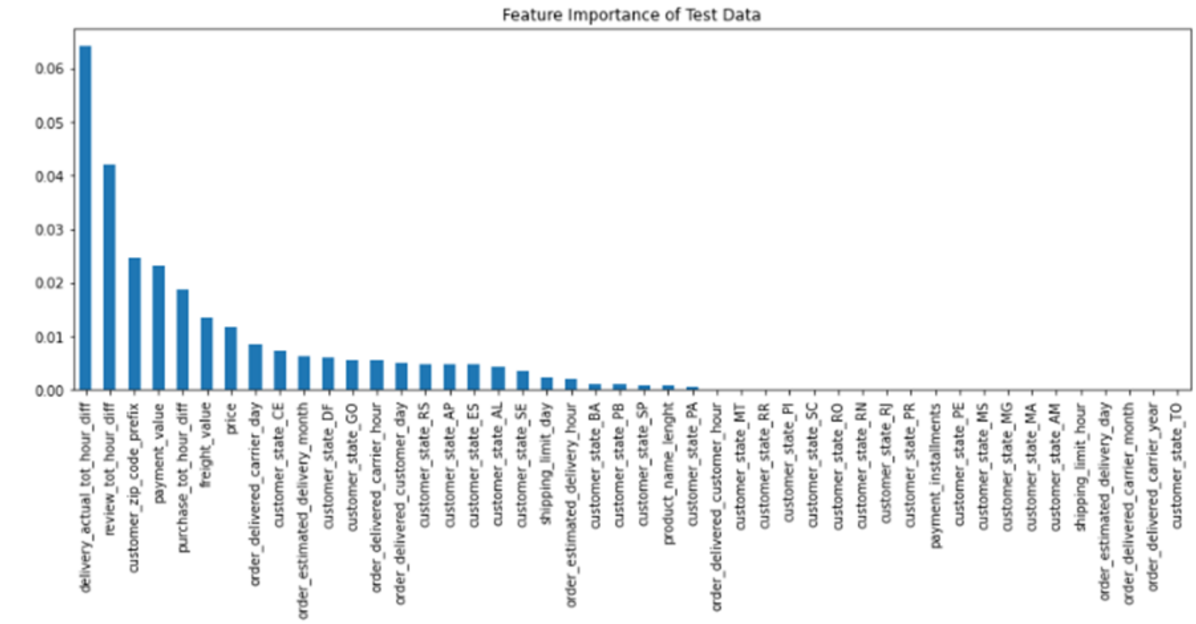
**Feature Engineering**

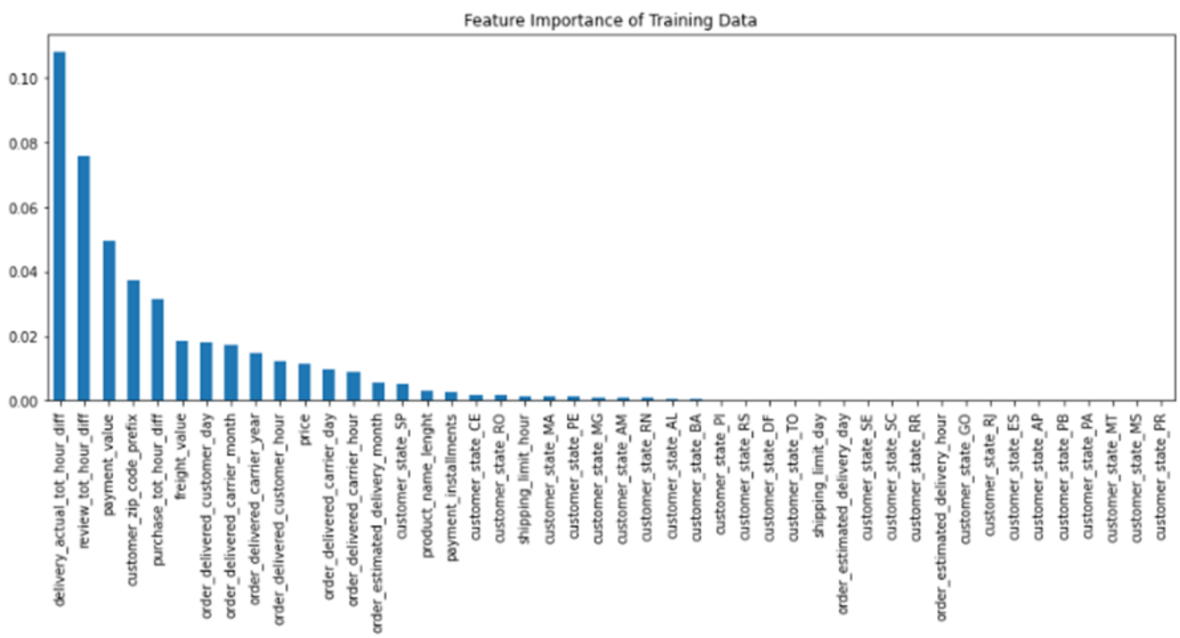
```
to_drop = [column for column in upper_tri.columns if any(upper_tri[column] > 0.85)]
print(to_drop)
```

```
['review_tot_mins_diff', 'purchase_tot_mins_diff', 'delivery_actual_tot_mins_dif
f', 'delivery_actual_tot_days_diff', 'review_score_dummy', 'order_delivered_carrie
r_week', 'order_delivered_customer_year', 'order_delivered_customer_month', 'order
_delivered_customer_week', 'order_estimated_delivery_year', 'order_estimated_deliv
ery_week', 'shipping_limit_year', 'shipping_limit_month', 'shipping_limit_week']
```

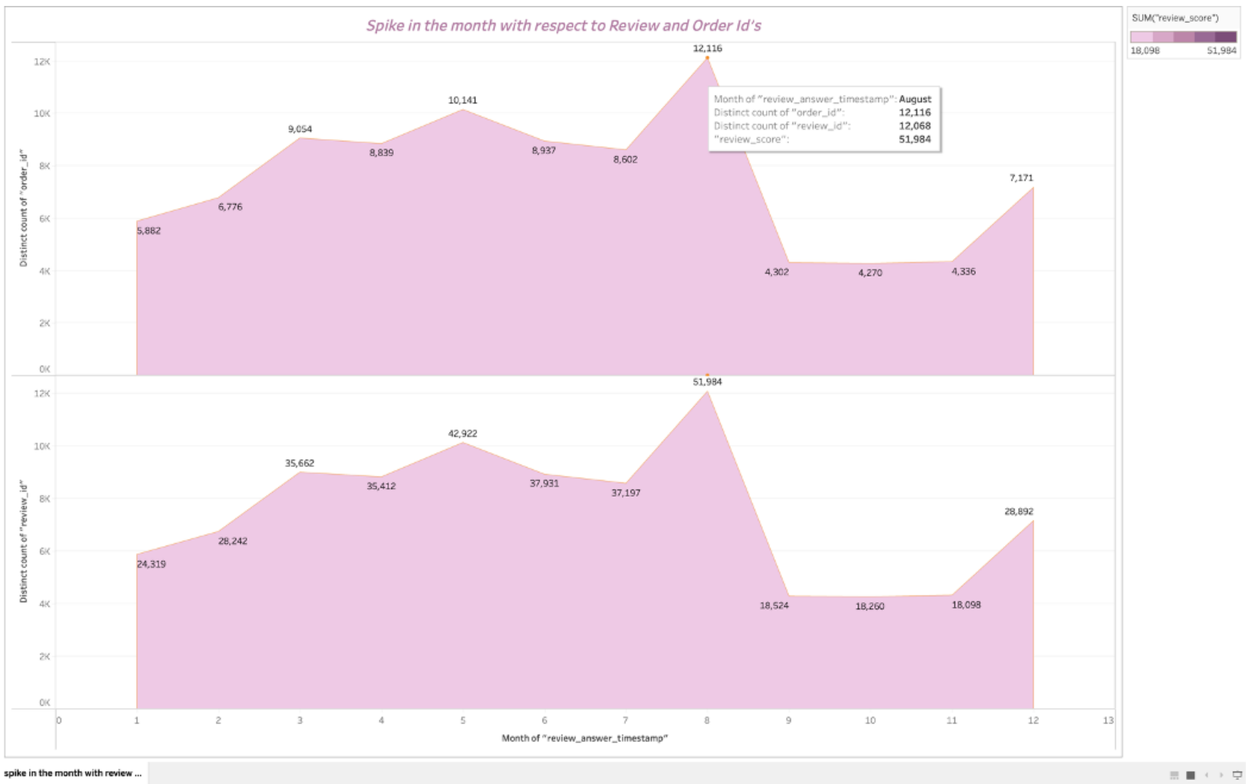
After applying the one-hot encoding to turn the only existing categorical feature, customer states, into numerical values, we started the feature engineering process by first checking and dropping the highly correlated columns to avoid multicollinearity.. We dropped the columns that are 85% correlated with each other.

Since the dataset contains multitudes of features, it would be inefficient to use visualization to study the hidden relationship between the target object and all other features. We utilized the technique called mutual information classifier from sklearn which provides a measurable value of the dependency between the target object and other features with a higher mutual information score indicating a closer connection between the feature and the target. The plots of feature importance of both test data and train data are presented below.

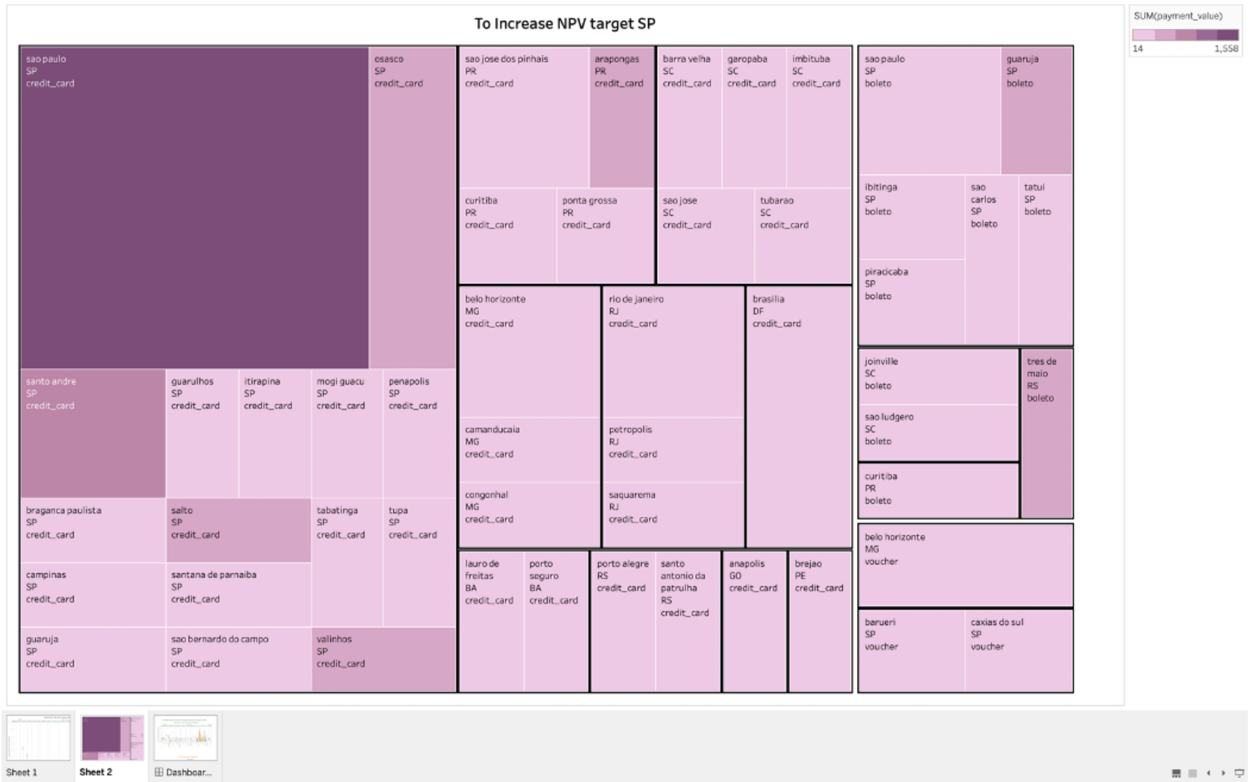



The result aligns with our original expectation that delivery delay plays an important factor in affecting the customer shopping experience and the review score. One thing to notice is that nearly half of the features do not provide helpful information for us to study the Net Promoter Score of the customers, so we decided to include only the top 20 most important elements for both the train and test datasets.




As noted in the preceding statement, the data we obtained was from the Brazilian company olist. Because the sample size of the information is enormous, with more than 1lakh columns, we utilized tableau prep to clean and filter it. We saw a significant increase in the products they were ordering. When we investigated this issue further, we discovered that Brazilian Independence Day was on September 7th, leading the statistics to reflect a significant increase.




As the problem description indicates, our primary aim is to improve operations, and logistics is one of the crucial areas where we have seen that SP. has the most significant number of customers. Assume we could narrow our emphasis to that area and target more buyers and sellers. We could see additional expansion in the future.

We also looked at how customers purchased; the majority used credit cards, while 30% used boleto, a legal currency partner for the Brazilian company; if you pay with this, there is no return policy because it is a voucher approved by the Brazilian government

**Predictive Model:**

As we have introduced earlier on this report, we have binned the customer ratings into two categories: Positive and Negative with the thought that customers who rated their orders 4 and above (positive) will be promoters and those who rated lower than 4 will be detractors. We intend to accurately predict the customer rating so that we can come up with a predicted NPS. In this endeavor the binary customer rating was our dependent variable.

In line with the previous section of this report, we have used the 20 most important features from the mutual information classifier as our independent variables in all our predictive models. We partitioned the data into training (70%) and test (30%) sets. Below table shows the accuracy scores of the train and test data sets for each of the models we have used.

| Models | Random Forest | Decision Tree Classifier | Logistic Regression | Bagged Tree | Boosted Tree |
|---|---|---|---|---|---|
| Mean Accuracy (Train Set) | 0.7998 | 0.8139 | 0.7847 | 0.9524 | 0.7981 |
| Mean Accuracy (Test Set) | **0.7904** | 0.7797 | 0.7310 | 0.7825 | **0.7973** |

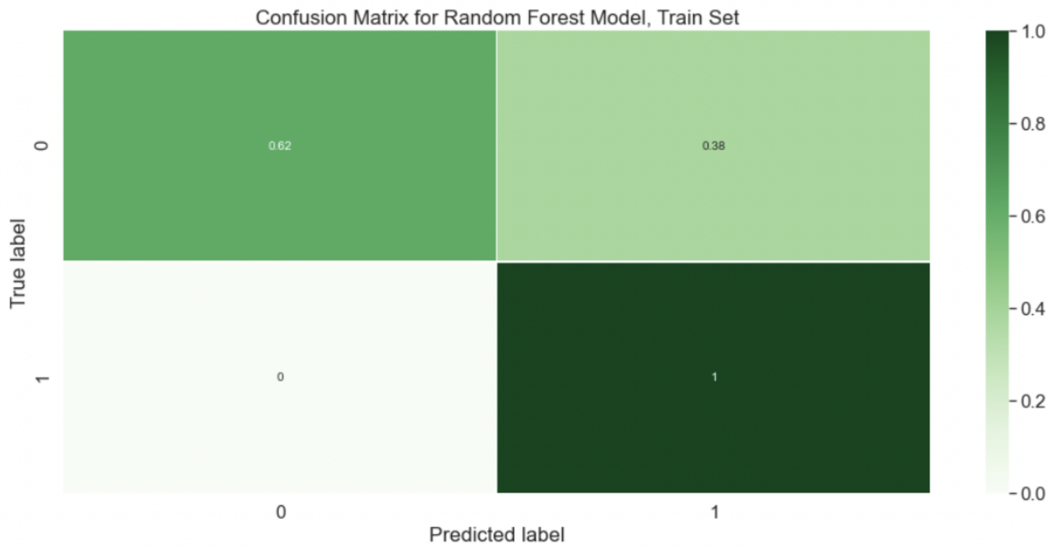
As seen from the table, the test accuracy scores of the Random forest and the boosted tree are the highest. We wanted to choose our final model based on the presented accuracy scores before conducting model optimization. Therefore, we needed to choose between the random forest and boosted tree algorithms.

Even though in our preliminary models boosted tree had a slightly higher test accuracy score than random forest, we have decided to go with random forest because of the differences in algorithms in both of thee models. Since boosted trees are sequentially built and they progress by correcting the error in each sequence, in the presence of noisy data, they tend to overfit. Given we have a large data set with roughly 117 thousand observations, we believe that the data is quite noisy. Therefore, we have decided to go with random forest as our final algorithm.

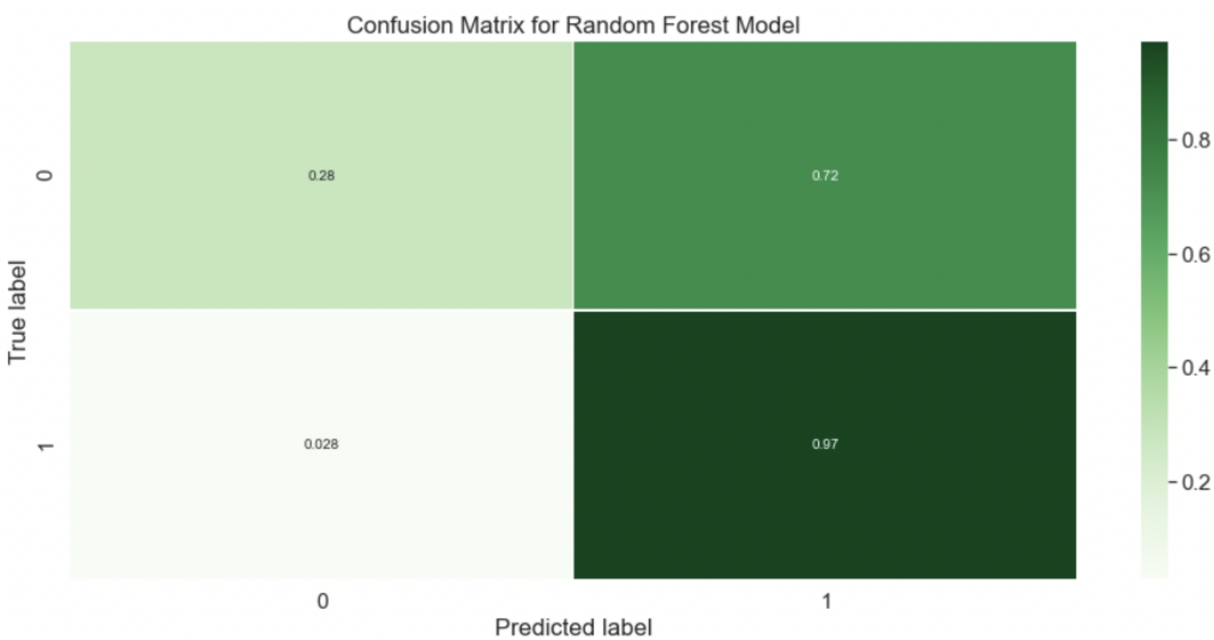
Once we have decided on the algorithm to use, we have proceeded with parameter tuning to optimize our model. To do this we ran a grid search where we aimed to find the optimum values for the following hyper parameters: Maximum tree depth, Maximum number of features in a given tree, and Number of trees. Our final model turned out to be a random forest model with Maximum depth of 20, Maximum number of features in a given tree of 10, and maximum of 150 trees. The accuracy scores from our final model is below

| | |
|---|---|
| Train Score | 0.91 |
| Test Score | 0.81 |

Compared to our preliminary model for random forest, we have managed to acquire an improved test score. Using our final model we have constructed confusion matrices to further understand the model performance




As seen from the confusion matrix above, the model perfectly classifies true positives. However, it is not as accurate when it comes to classifying true negatives (at 62%).



The model shows a similar performance with the test set as with the training set. We see that the model classifies true positives near-perfectly, but classifies true negatives correctly only 28% of the time.

Using the confusion matrix of the training set, we have calculated the NPS. Below table shows the predicted and true NPSs as well as the treu NPS of the entire data set.

| | |
|---|---|
| Predicted NPS of Train Set | 0.7125 = 71.25% |
| True NPS of Train Set | 0.535 = 53.5% |
| NPS of Entire Dataset | 0.5364 = 53.64% |

As seen above the training set is a true representative of the entire data set since the calculated NPS values are close to each other. However, since the model has a high rate of false positives, predicted NPS ends up being overestimated.