

**CAPSTONE PROJECT** - **INTERIM REPORT**

**OLIST-ECOMMERCE**

**Submitted By:**

* **Sivakanth Reddy**
* **Md. Sohail Abrar**
* **Zehna Erum Khan**
* **Prudhvi Raj**

**MENTOR: VIBHA SANTHANAM**

**LITERATURE SURVEY**

* The Olist store is an e-commerce business headquartered in Sao Paulo, Brazil. This firm acts as a single point of contact between various small businesses and the customers who wish to buy their products.
* This is a Brazilian ecommerce public dataset of orders made at Olist Store. 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 have a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
* This project tries to predict customer satisfaction on the item they had ordered based on the given data. This case study was converted into a binary classification task. Various machine learning were used to accomplish this task. The performance of these models was assessed using performance metrics such as confusion matrices and the F1 score. Check the link given below to find the blog which explains this project in detail.

**PROBLEM STATEMENT**

What we purchase on e-commerce websites is affected by the reviews which we read about the product posted on that website. This firm can certainly leverage these reviews to remove those products which consistently receive negative reviews. It could also advertise those items which are popular amongst the customers.

**PROJECT OUTCOME**

Our observations should serve as basis for Olist’s managers to actually make decision to increase

the business value.

The average review score being quite high already, Olist should

work on customer retention to ensure product quality, for example, by having a charter of

integrity signed by the sellers, penalizing them and offering customer benefits in the event

of a problem that is the responsibility of the seller. and to have issues with late deliveries. One might be interested to

further our analysis to try to spot geographical locations of late deliveries. Hence, if this

analysis come conclusive, Olist managers could develop relay points to improve the fast

delivery service.

**INDUSTRY REVIEW**

In today’s world, E-commerce is growing rapidly around the globe.

Ecommerce is the buying and selling of goods and services over the Internet. It is conducted over computers, tablets, smartphones, and other smart devices. Almost anything can be purchased through ecommerce today. It can be a substitute for brick-and-mortar stores, though some businesses choose to maintain both

The Olist store is an e-commerce business headquartered in Sao Paulo, Brazil. This firm acts as a single point of contact between various small businesses and the customers who wish to buy their products. Recently

They uploaded a dataset on Kaggle that contains information about 100k orders made at multiple marketplaces between 2016 to 2018. What we purchase on e-commerce websites is affected by the reviews which we read about the product posted on that website. This firm can certainly leverage these reviews to remove those products which consistently receive negative reviews. It could also advertise those items which are popular amongst the custom

Olist, a Brazilian e-commerce marketplace integrator, confirmed it is now valued at $1.5 billion after securing $186 million in Series E funding, led by Wellington Management.

**DATA SET AND DOMAIN**

A dataset is a collection of data, and it can be structured or unstructured.

* This is a Brazilian ecommerce public dataset of orders made at Olist Store. 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 have a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
* We are presented with a 5-star rating system that summarizes the overall satisfaction of the customer with the product which he or she had just purchased. We can convert this aspect into a multi classification problem by treating the 4 and 5-star ratings as the positive class ,3 as moderate and the rest as the negative class.

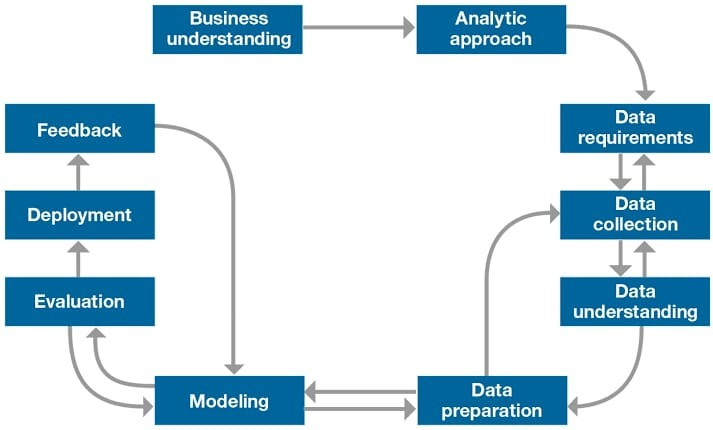
**DATA DESCRIPTION:**

This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.

The dataset has 117329 rows with 39 features. Refer to the below-detailed structure of the dataset.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | |
| Customer ID | key to the orders dataset. Each order has a unique customer\_id. | |
| customer\_unique\_ID | unique identifier of a customer. | |
| customer\_zip\_code\_pref ix | first five digits of customer zip code | |
| customer\_city | Customer city name | |
| customer\_state | Customer-state | |
| geolocation\_zip\_code\_p refix | first 5 digits of zip code | |
| geolocation\_lat | Latitude of the customer | |
| geolocation\_lng | longitude | |
| geolocation\_city | City name of customers | |
| geolocation\_state | state | |
| order\_id | order unique identifier | |
| order\_item\_id | | sequential number identifying number of items included in the same order. | |
| product\_id | | product unique identifier | |
| seller\_id | | seller unique identifier | |
| shipping\_limit\_date | | Shows the seller shipping limit date for handling the order over to the logistic partner. | |
| price | | item price | |
| freight\_value | | item freight value item (if an order has more than one item the freight value is splitted between items) | |
| payment\_sequential | | a customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments. | |
| payment\_type | | method of payment chosen by the customer. | |
| payment\_installments | | number of installments chosen by the customer. | |
| payment\_value | | transaction value. | |
| review\_score | | Note ranging from 1 to 5 given by the customer on a satisfaction survey. | |
| review\_comment\_title | | Comment title from the review left by the customer, in Portuguese. | |
| review\_creation\_date | | Shows the date in which the satisfaction survey was sent to the customer. | |
| review\_answer\_timesta mp | | Shows satisfaction survey answer timestamp. | |
| order\_status | | Reference to the order status (delivered, shipped, etc). | |
| order\_purchase\_timesta mp | | Shows the purchase timestamp. | |

**PRE-PROCESSING DATA ANALYSIS:**



**Data Preparation**

Data preprocessing is the process of transforming raw data into an understandable format.The quality of the data should be checked before applying machine learning or data mining algorithms.To make the process easier, data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation

Data preprocessing is an important step to prepare the data to form a QSPR model.Data cleaning and transformation are methods used to remove outliers and standardize the data so that they take a form that can be easily used to create a model..

Acquire the dataset:

<https://www.kaggle.com/olistbr/brazilian-ecommerce>

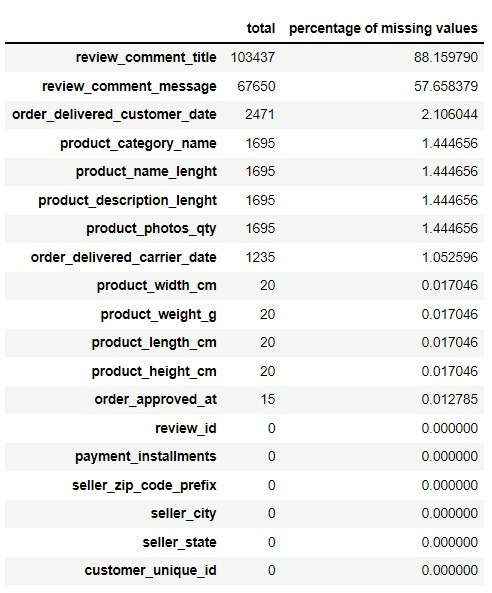
**Missing/Null Values**

Impute or drop features with missing values based on the percentage of missing values and relevance for model building.

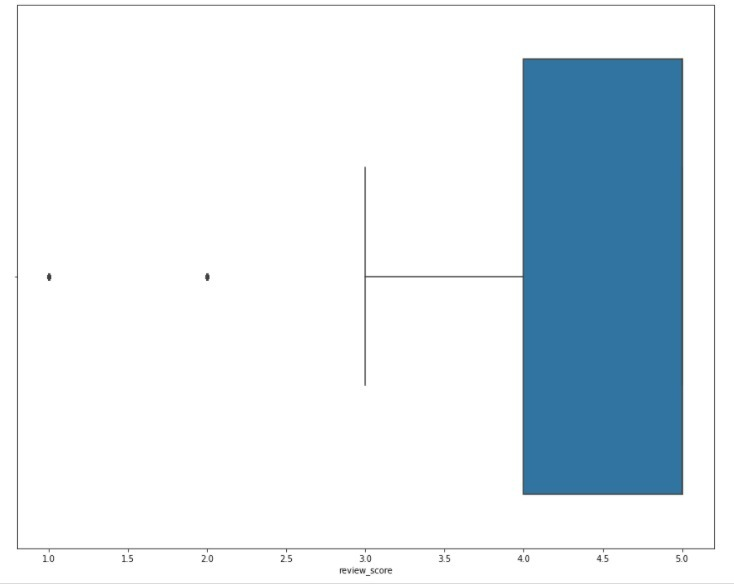
In the dataset order approved date, order delivery date,product\_category\_name,order\_delivered\_customer\_date have null values less than 10%. So, we can drop the rows.

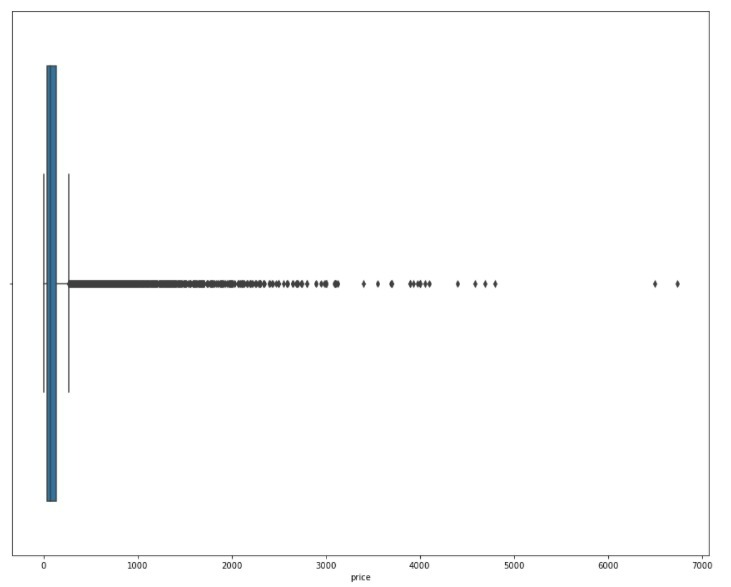
The null values of review\_comment messege are imputed by product name. Since the brand feature has null values of around 57% but these are important for classification.

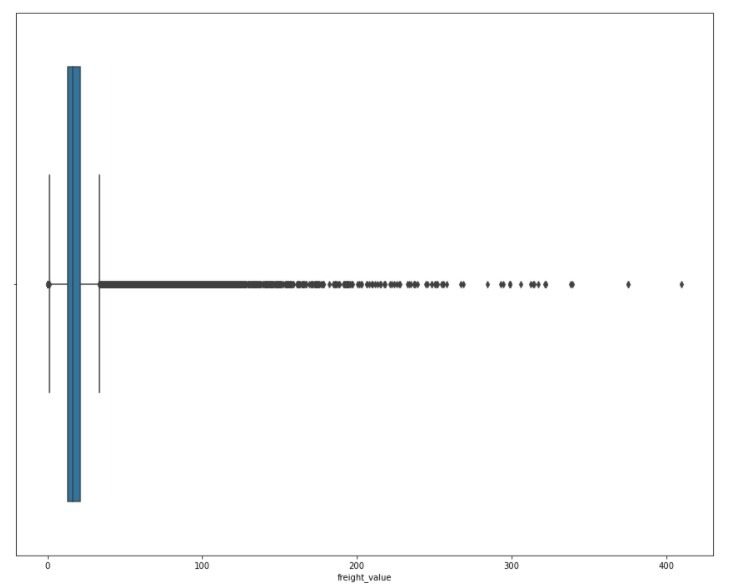
The review\_title have null values of more than 90%. So, we can drop variable.

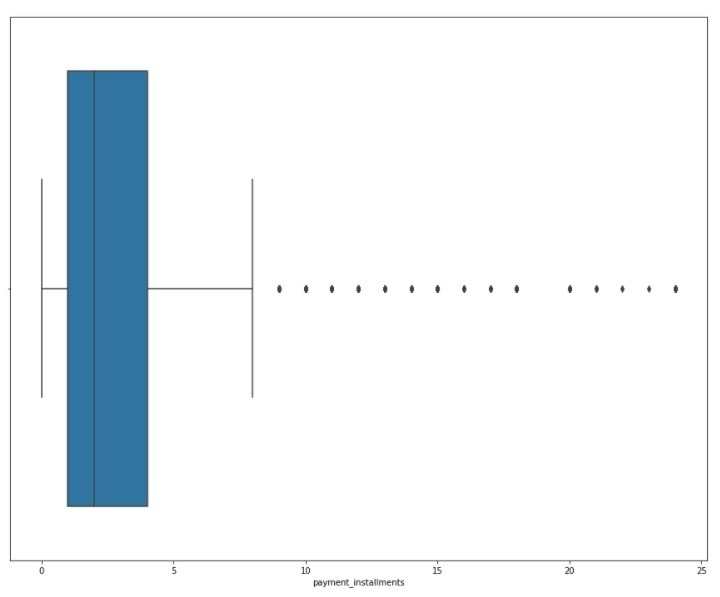


**OUTLIERS**









The outliers are present in the retail price and discounted price column. As the prices are significantly unique to each product category, which ranges from very low value to very high value, we are not excluding the outliers instead we would implement a transformation technique considering the outliers thereby reducing the skewness.

**REDUNDANT COLUMNS**

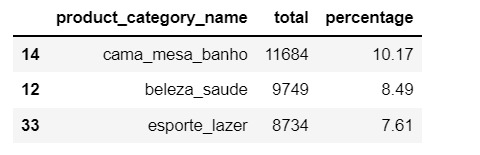
The below are the redundant features that have all the unique values and are dropped from the dataset.

* Customer\_id
* Order\_item\_id
* Review\_id
* Seller\_id
* Product\_id

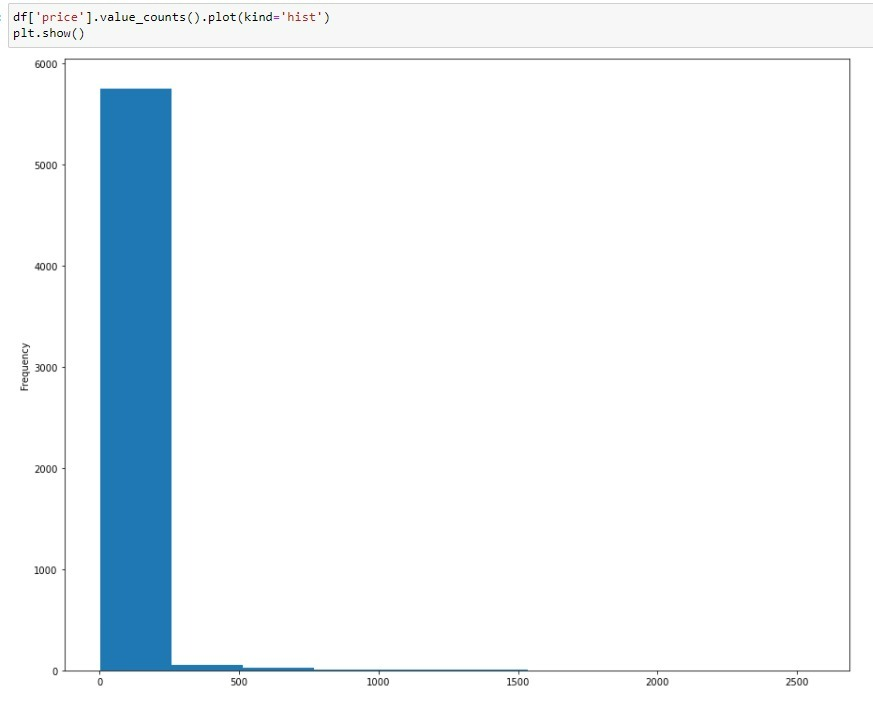
**EXPLORATORY DATA ANALYSIS & BUSSINESS INSIGHTS**

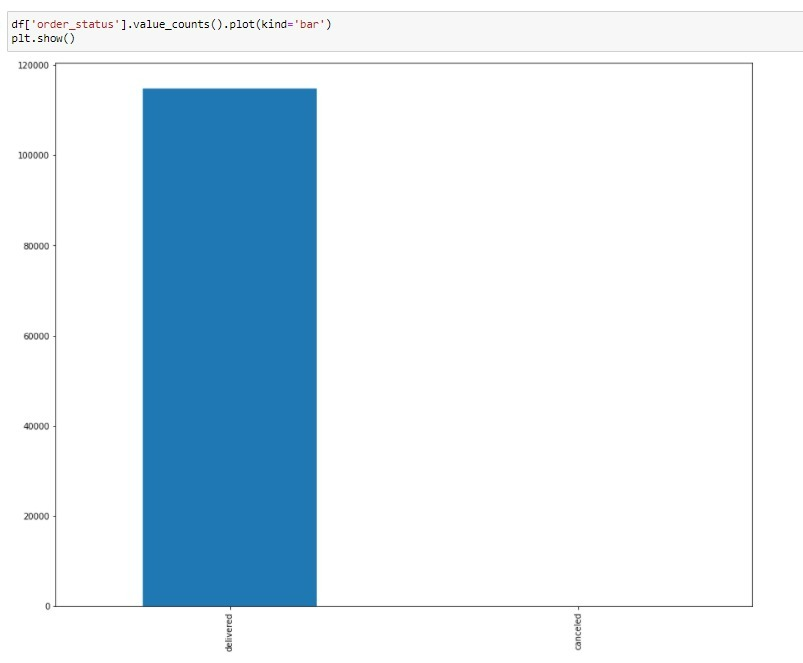
**Analysis on top performing products and product categories:**

The below are the high-performing products based on sales.

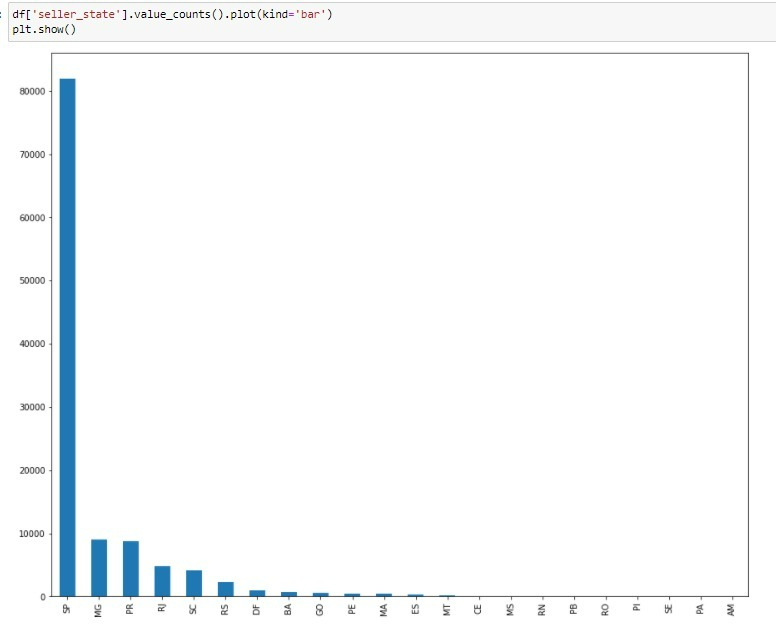


**UNIVARIANT ANALYSIS**

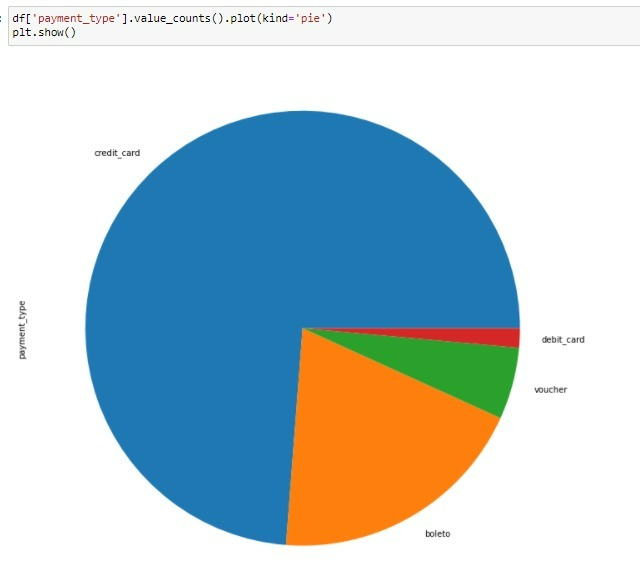




We can see that most of the order status are delivered from the above graph

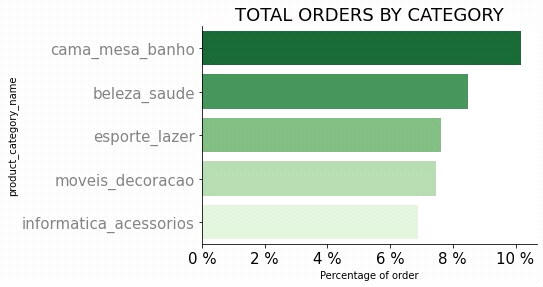


We can see that the state SP has more sellers compared to other states.



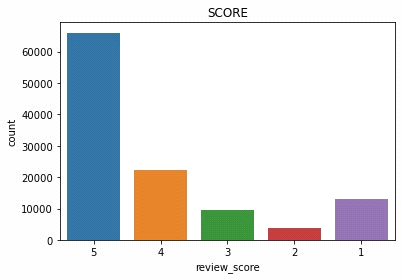
We can see that the most of the payments are done by the credit card and next is boleto from the above pie chat.

**Analysis of orders by category**



The bar plot infers that the cama\_mesa\_banho as saled more in the olist as shown in the above graph.

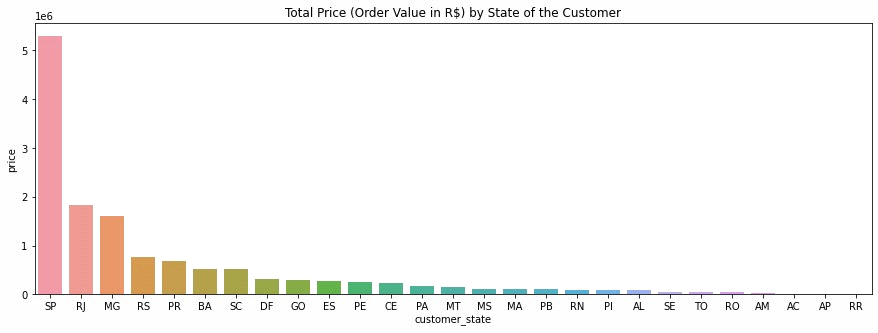
**Analysis of reviews**



The above graph shows that the review\_score contains more 5 and 4 star rating.

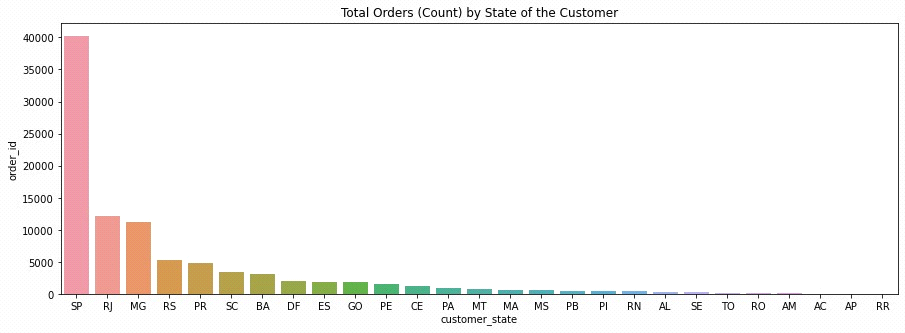
**BIVARIANT ANALYSIS**

**Analysis on Total price(Order Value in R$)by the state of customer**

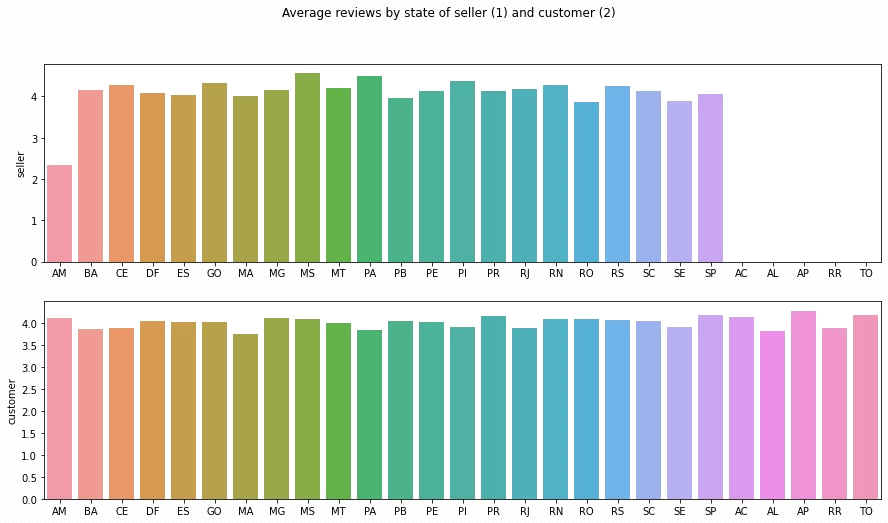


The above graph say that the sp have more value of order compaired to the others state.

**Analysis of Total Orders (count) by state of the customer:**

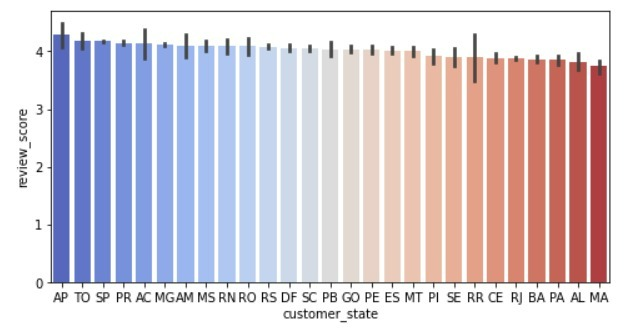


The bar plot infers that the more number of orders came from the state sp as shown in above graph.



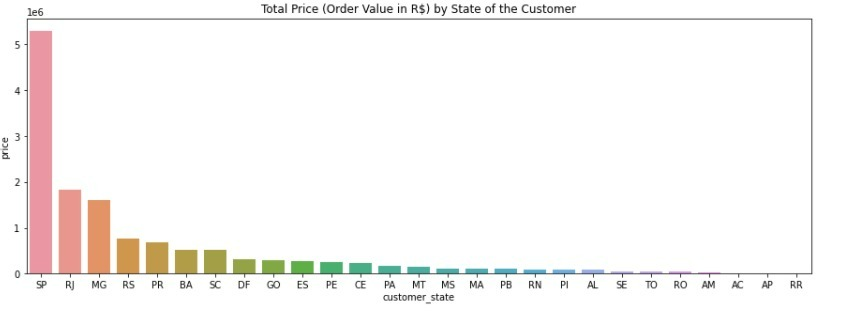
\*\* Status: Work in Progress \*\*

**ANALYSIS ON CUSTOMER\_STATE & REVIEW\_SCORE**



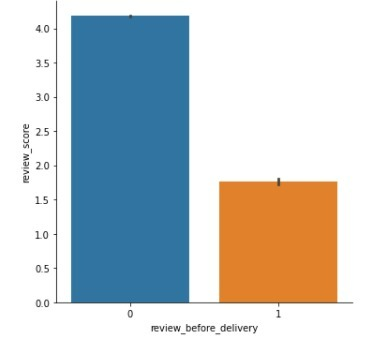
We can see that review\_score at customer\_state is high in AP followed by TO,SP,PR

**ANALYSING ON TOTAL PRICE(ORDER VALUE ) BY STATE OF THE CUSTOMER**



The above graph shows that the customer\_state SP is high in the terms of the order value.

**ANALYSIS ON REVIEW SCORE ON EARLY DELIVERY**



There is Huge impact of the lateness as you can see in the above graph.

We have to point out that the review score are somehow biased because the form is automatically received the day of the estimated delivery. Hence, if the parcel is late the customer will have to fulfill the review survey before having received its order. We notice that receiving and fulfilling the form before the delivery is impacting consistently the average review score.

However, receiving the survey before the delivery is not strictly speaking the real cause. The

first cause of this drop in scores is more about lateness, as it is shown in .But note

that the drop is even greater for people that filled the review form before having their order.

Indeed, imagine a customer whose order is delayed. It is no surprise that this customer will be

more rude in its review if he has to review the order before receiving it than if he had already

received it. Therefore, we recommend to Olist to change their survey mailing policy by sending.

We see that there is a strongly negative correlation between the average state

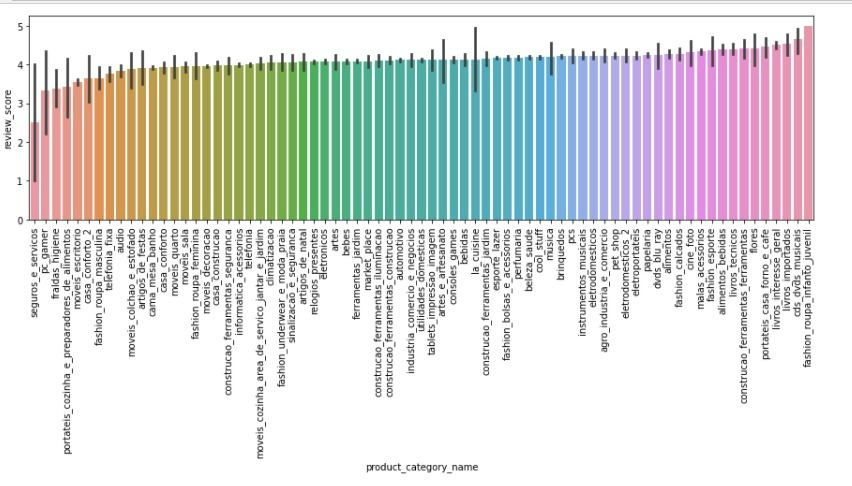
review score and the late delivery frequency. But what is really remarkable is the variation

of the lateness rate between state where in Alagoas (AL), more than 20% of the orders are

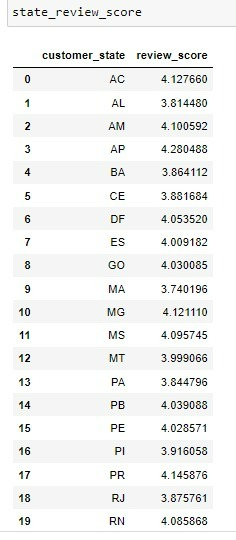
delayed. Olist should figure out why in some states the lateness frequency is high and find

ways to improve their shipping solutions. In parallel Olist should b less optimistic on their estimated delivery date.

**ANALYSIS ON PRODUCT CATEGORY NAME**

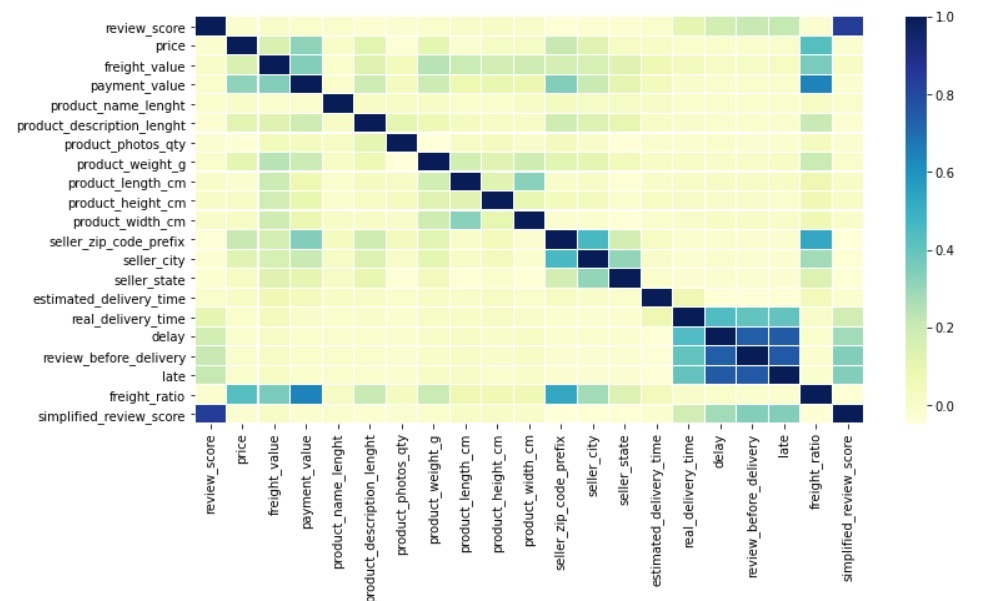


**ANALYSIS ON STATE REVIEW SCORE**



The above are the list of top states how had given the high review rating.

**HEATMAP**



For now, we keep only relevant features that may have an effect on review score, which are

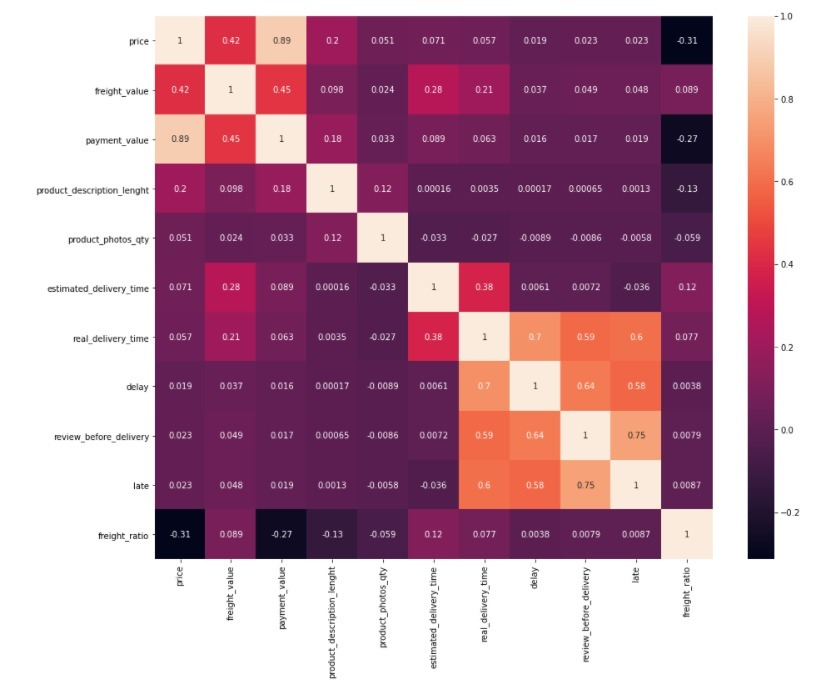
shown in the correlation matrix.We directly see that there is a correlation

between review the real delivery time, the delay and late deliveries. Note that these time

related features are correlated with each other, and we will mostly refer to late and delay for

the following statistics.

**CORRELATION**



**FEATURE IMPORTANCE**



**FEATURE ENGINEERING (STATISTICAL TESTING):**

**Two sample independent z-test:**

* The two-sample z-test is used to compare the equality of means of two populations for unpaired data.
* The hypothesis to test whether the population means are equal

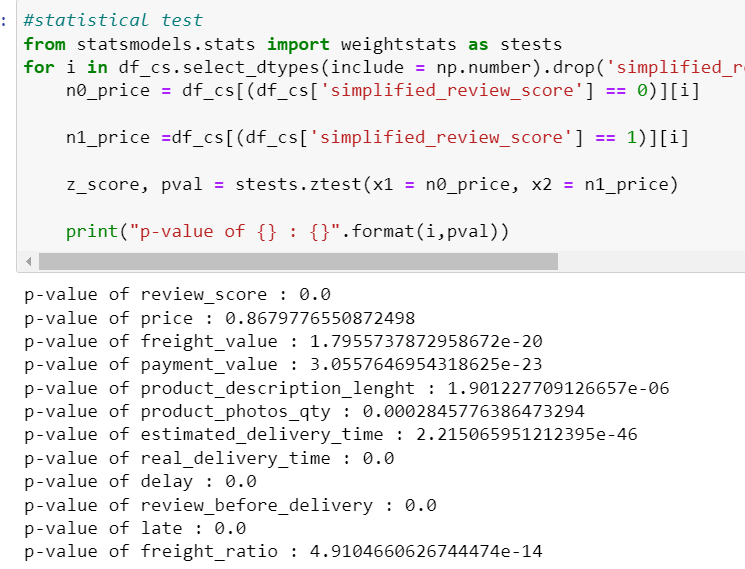
H0: µ1 = µ2 against Ha: µ1 ≠ µ1

It implies

H0: The two-population means are equal (i.e., µ1 = µ2) against

Ha: The two-population means are not equal µ0 (i.e., µ1 ≠ µ2)

**Two sample independent z\_test on every column and target (review\_score)features:**

****

**Interpretation**:

* It is observed that the p\_value is 0.0 or less than which is less than 0.05 significance level, hence Ho is rejected and Ha is selected.
* There is a significant effect of every column except price on the target variable, hence considering all features for analysis.we check the results of impact of that columns after building the model.

**Splitting the data into 70:30 training and testing proportions:**

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets.

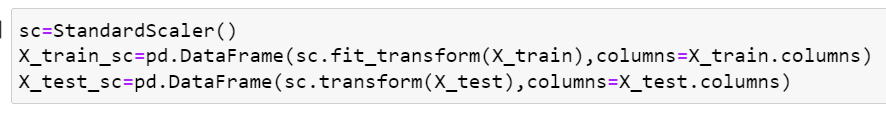
The simplest way to split the modelling dataset into training and testing sets is to assign 2/3 data points to the former and the remaining one-third to the latter. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we can evaluate the performance of our model

Separating data into training and testing sets is an important part of evaluating data mining models. ... Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

****

**Normalization of training and testing data:**

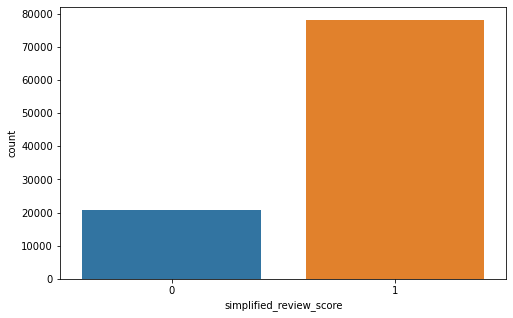
* Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization.



**SMOTE ANALYSIS:**

* SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.   
  It aims to balance class distribution by randomly increasing minority class examples by replicating them.

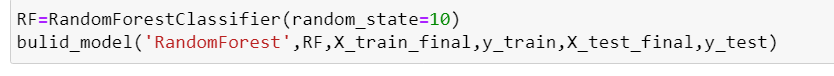
|  |
| --- |
| The data is an imbalance |

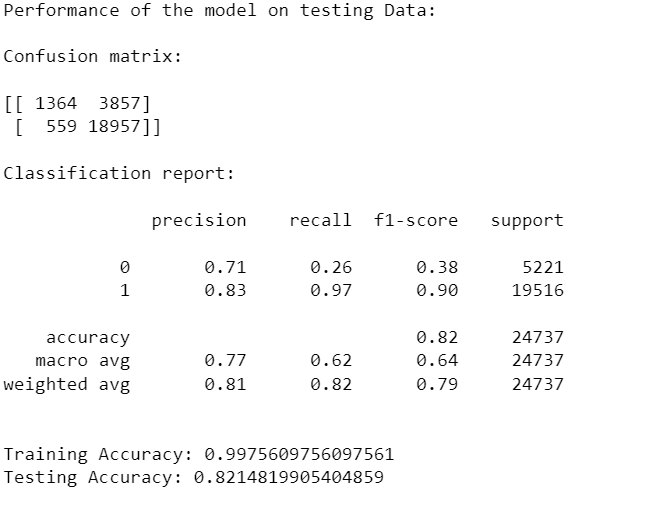


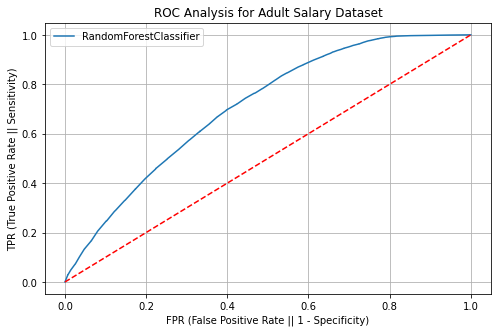
As data is imbalance so we need to smote but after doing over sampling false positives has been increased rapidly so we didn,t use smote

**MODEL BUILDING:**

**Random Forest Algorithm:**

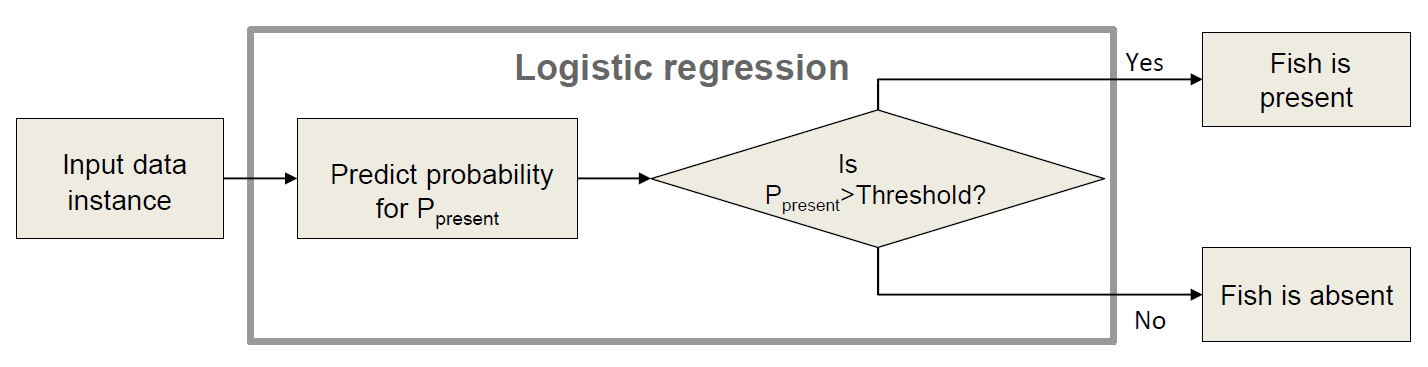
* Random Forest consists of several independent decision trees that operate as an ensemble.
* It is an ensemble learning algorithm based on bagging.
* 

****

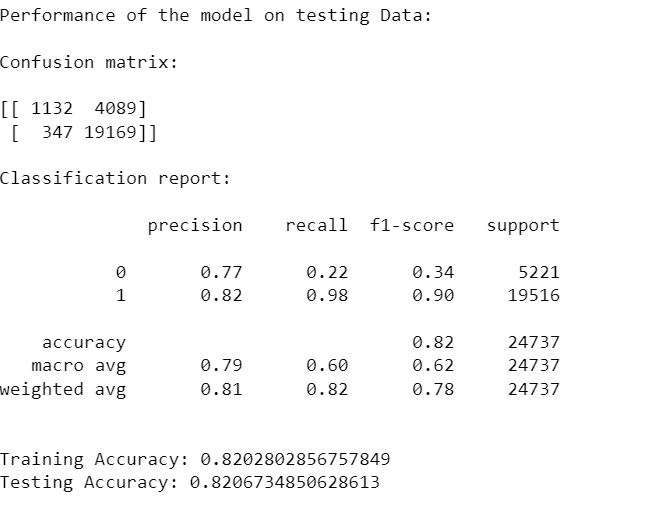
****

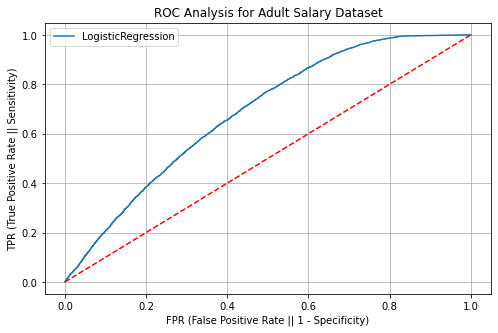
**Logistic regression:**

* Logistic Regression is a binary classification algorithm. It predicts the probability of occurrence of a label class.
* Consider that logistic regression is used to identify whether the product falls under the advantage category or not.



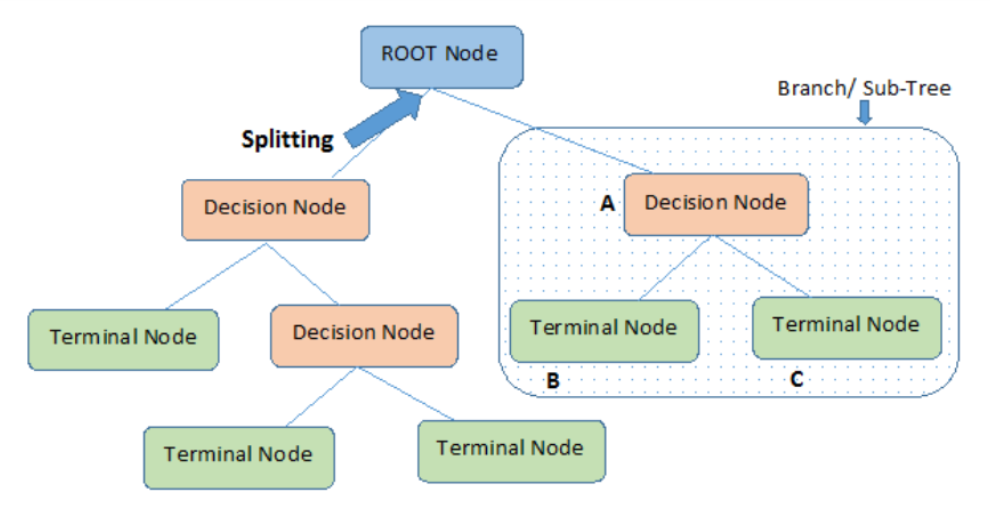




****

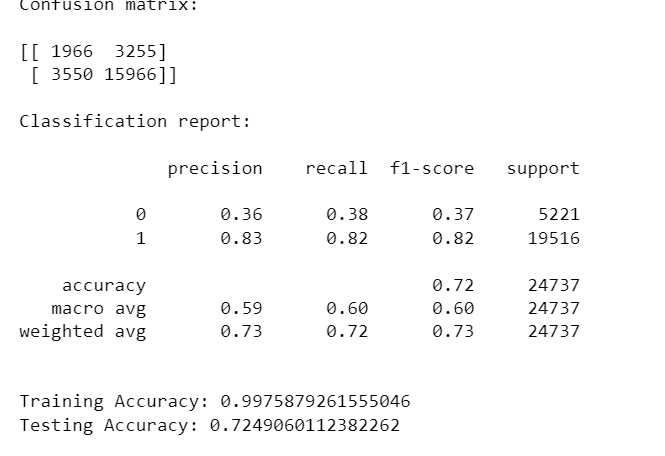
**Decision Tree Algorithm:**

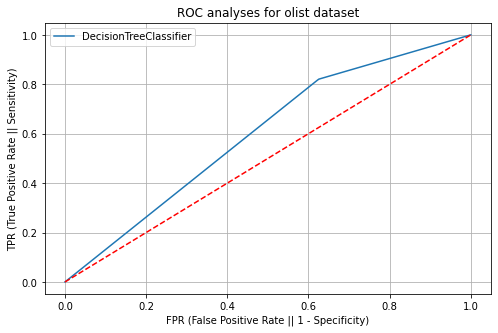
* Decision trees can be used for classification as well as regression problems.
* The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits.
* It starts with a root node and ends with a decision made by leaves.

****

**DT=DecisionTreeClassifier(random\_state=10)**

**bulid\_model('DecisionTreeClassifier',DT,X\_train\_final,y\_train,X\_test\_final,y\_test)**

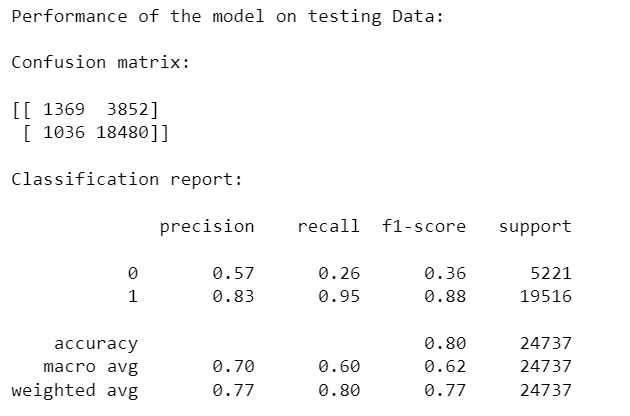
****

****

**KNN Algorithm:**

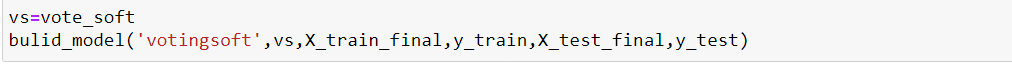
* The K -Nearest Neighbour algorithm classifies the data based on the similarity measure.
* K specifies the number of nearest neighbours to be considered.

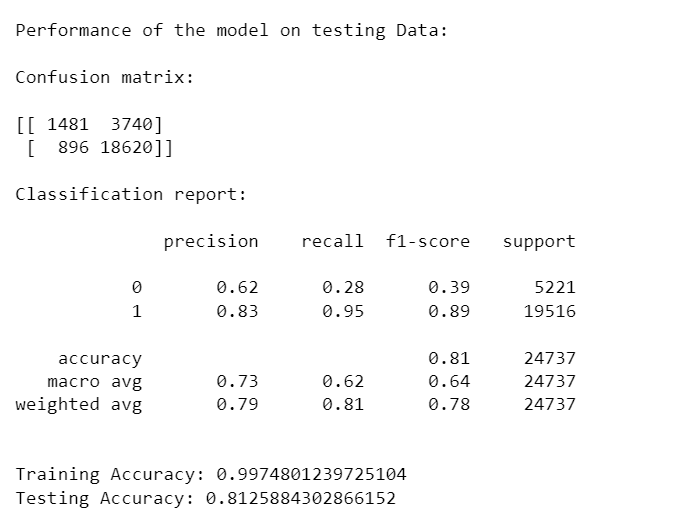




## Voting Classifier-Soft

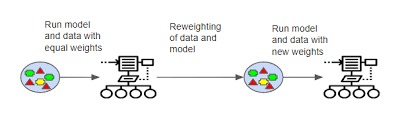
## A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output

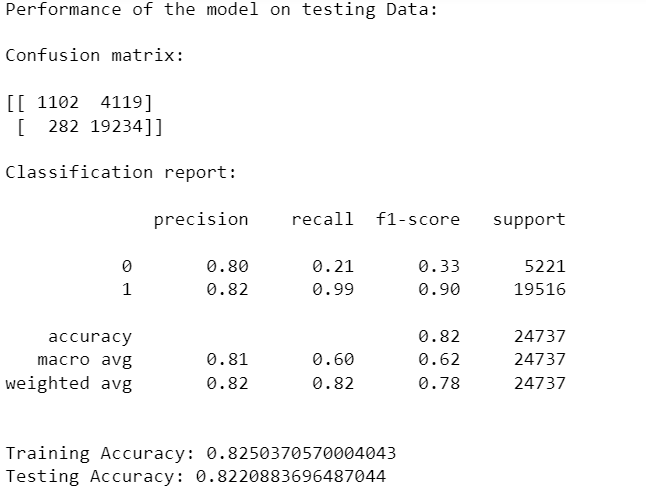
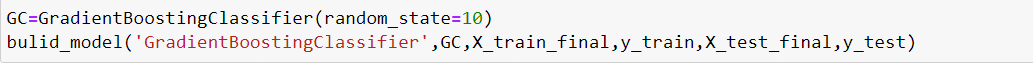


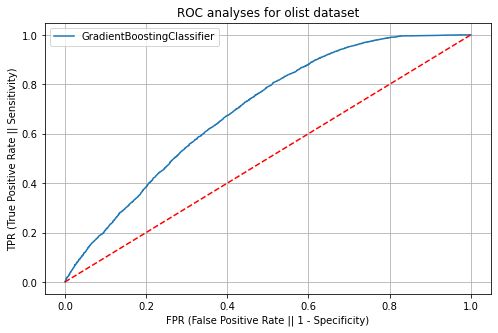


**5.GradientBoostingClassifier**

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.



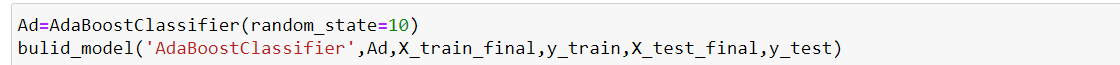
****

****

AdaBoostClassifier

Ada Boost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level



Performance of the model on testing Data:

Confusion matrix:

[[ 1079 4142]

[ 286 19230]]

Classification report:

precision recall f1-score support

0 0.79 0.21 0.33 5221

1 0.82 0.99 0.90 19516

accuracy 0.82 24737

macro avg 0.81 0.60 0.61 24737

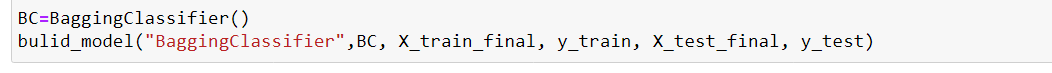
weighted avg 0.82 0.82 0.78 24737

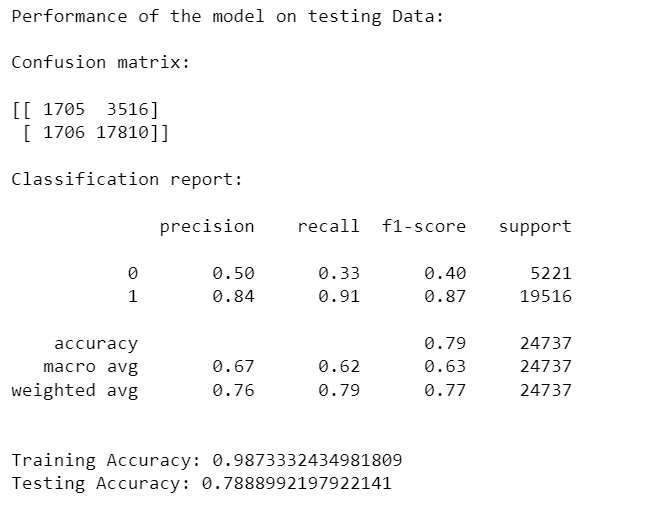
Training Accuracy: 0.8228001617032745

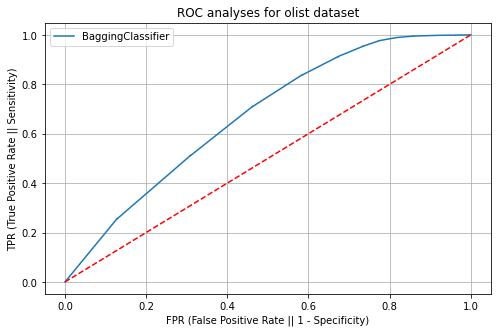
Testing Accuracy: 0.8209968872539112

Bagging Classifier

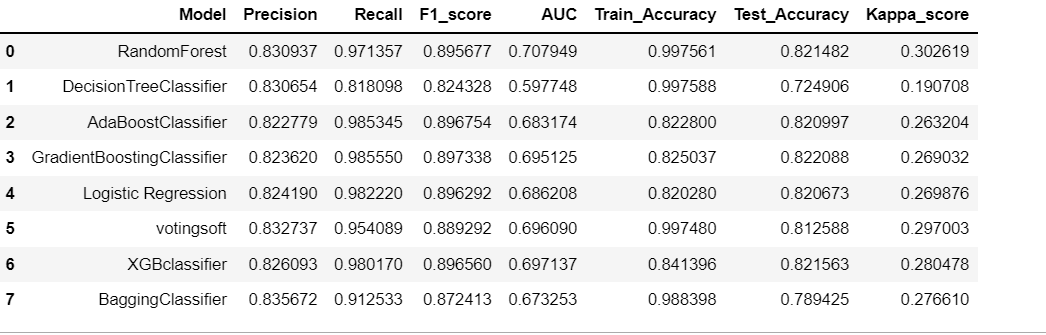
A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

****

****

****

From above base models we consider gradient boosting as base model as training and testing accuracy scores are equal(0.82)

****

We performed all above base models using automated function to check the best fit model we can see gradient boost has equal scores with better precision score

**PERFORMANCE METRICS:**

* **Confusion Matrix:**

It is the performance measure for the classification problem. It is a table used to compare predicted and actual values of the target variable.

* **ROC:**

ROC curve is the plot of TPR against the FPR values obtained at all possible threshold values.

**PERFORMANCE EVALUATION METRICS:**

* **Accuracy:**

Accuracy is the fraction of predictions that our model got correct. Higher the accuracy of the model better is the model.

* **Precision:**

Precision is the proportion of positive cases that were correctly predicted.

* **Recall:**

A recall is the proportion of actual positive cases that were correctly predicted.

* **F1 score:**

F1score is the harmonic mean of precision and recall values for a classification model.

* **Cohen Kappa score.**

Kappa statistic is a measure of inter-rater reliability or degree of agreement.

**Business justification**

* The main motive is to improve the performance of the model. As per the business scenario, the model is predicting the positive or negative review if the product is wrongly classified under positive review In this scenario, the customer is betrayed and will lose the customer. This is a type 2 error.
* In the second scenario if the model is predicted the product is classified as negative review In this scenario, a customer might not choose the product which is also a loss to the company. This is a type 1 error.
* However, both the errors are costly but type 2 error is costlier than type 1. Here we need to reduce the false positives as minimum as possible. So, we choose precision as a performance metric.
* From the above model comparison table, the gradient boost model has good precision compared to other models. So, we are considering the gradient boost as a base model.

# **Hyper Parameters Tuning**

* On top of the base model, we have developed models , Random Forest, XGBoost, gradientboost
* In order to find the best parameters for each of these models we used Grid Search CV.

Hyper Parameter tuning for random forest: Max Depth, n Estimators, Min Samples Leaf, Min Samples Split.

* GridsearchCV best parameters obtained:

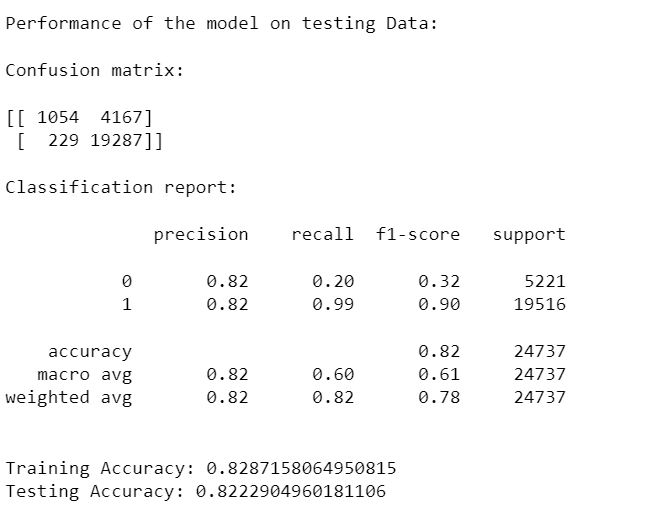
n\_estimators = 90,

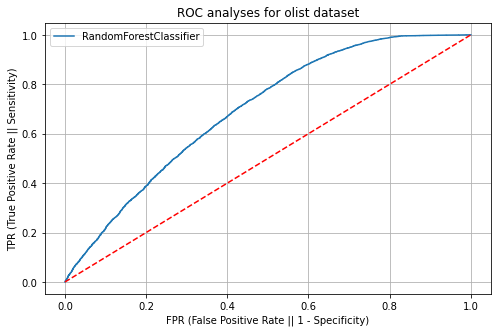
criterion = 'gini',

max\_depth = 10,

min\_samples\_split = 4,

min\_samples\_leaf = 1,

random\_state = 42, n\_jobs = -1) 

****

Hyper Parameter tuning for gradient boost: Max Depth, n Estimators,Min Samples Split.

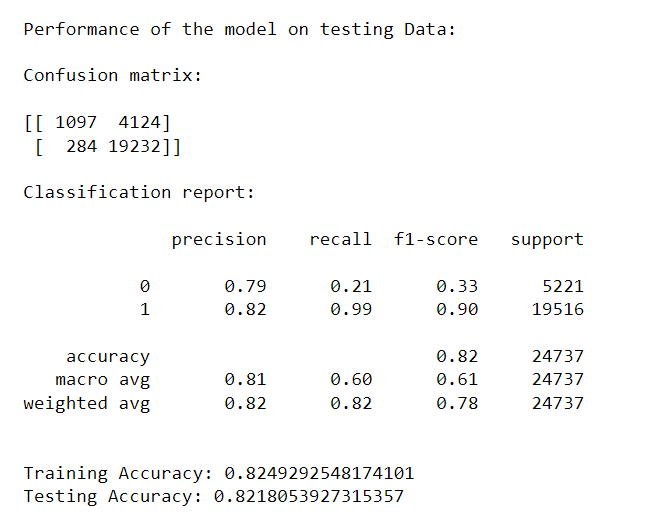
* GridsearchCV best parameters obtained:

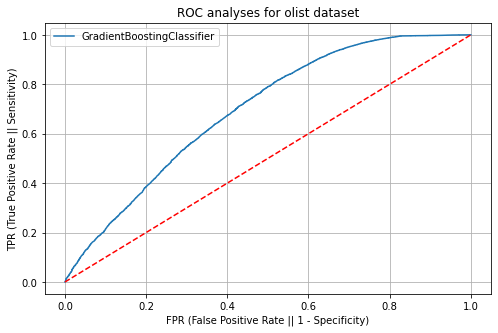
tuned\_gb = GradientBoostingClassifier(n\_estimators=100,

min\_samples\_split=10,

max\_depth=3,

random\_state=53)

****

****

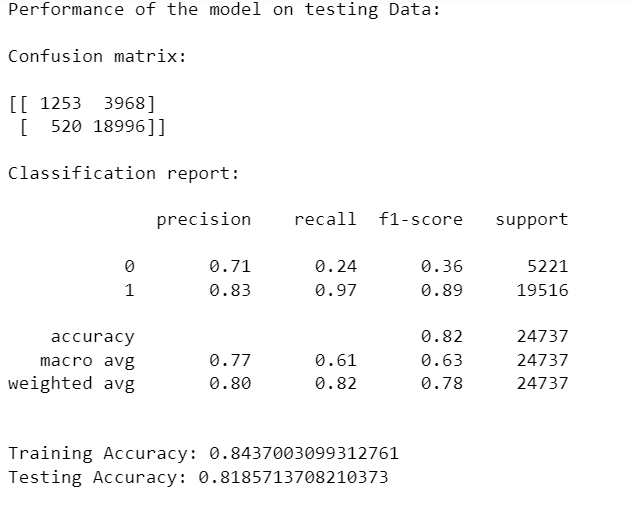
Hyper Parameter tuning for xg boost: Max Depth, learningrate,gamma,random\_state

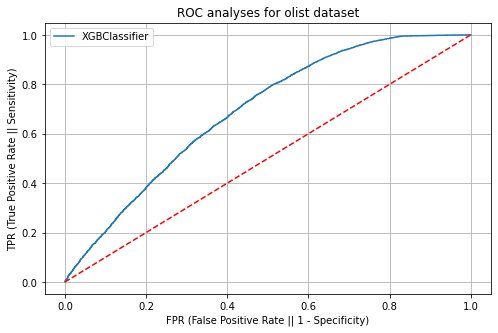
tuned\_xgb = XGBClassifier(max\_depth = 5,

learning\_rate = 0.6,

gamma = 2,

random\_state = 42)





**Inferences:**

Even after hypertuning with different models we can infer that model is

overfitted again if we compare with base model

Hyper tuning the parameters of three models the gradient descent modelgiving better results comparing with other two but the mosel has not given better results compared to the base model. Since the main requirement is to maximize the precision so consider the random forest base model as final model.

The gradientboost model provides better results (precision=0.824) with an accuracy of 0.82 without overfitting so we consider base model as best model

**CONCLUSION**

The goal of this project was to analyse Olist’s business using data analytic to predict the customer review rating and to find the high performing sectors to value and those that need to be improved.

**Business Strategies Insights**

Our observations should serve as basis for Olist’s managers to actually make decision to increase the business value.

**Customer Satisfaction**

The average review score being quite high already, Olist should

work on customer retention to ensure product quality, for example, by having a charter of

integrity signed by the sellers, penalizing them and offering customer benefits in the event

of a problem that is the responsibility of the seller. Furthermore, given the large size of

Brazil, it is not surprising to have issues with late deliveries. One might be interested to

further our analysis to try to spot geographical locations of late deliveries. Hence, if this

analysis come conclusive, Olist managers could develop relay points to improve the fast

delivery service.