horizontal line

Project Summary

|  |  |
| --- | --- |
| Batch details | DSEFT MUM JUL’20 |
| Team members | Asad Karel  Atul Gaikwad  Harshil Roy  Praveen Kamat  Sugandh Srivastava |
| Domain of Project | E-Commerce |
| Proposed project title | Prediction of Customer Churn (‘1’ for Customer Churned and ‘0’ for Customer not churned) of the purchase from the Brazilian e-commerce site Olist |
| Group Number | Group 3 |
| Team Leader | Asad Karel |
| Mentor Name | Mr. Ankush Bansal |

Date: 23rd April, 2021

Signature of the Mentor Signature of the Team Leader

# Table of Contents

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 4 |
| 4 | Critical assessment of topic survey | 5 |
| 5 | Methodology to be followed | 6 |
| 6 | References | 24 |

# 

# OVERVIEW:

Olist is operator of an online e-commerce site aggregation platform designed to facilitate direct sales on e-commerce sites. The company's platform connects entrepreneurs with major online retailers and allows shopkeepers to advertise and sell in the marketplaces without complication, enabling retail companies to reach out to the international marketplaces, improve the shopping experience and modify their purchasing behavior.

The company connects to the structuring of a business and focuses on the excellence of services provided to tenants and end consumers. It invests in a sustainable and fair business model for all involved. Olist was founded in 2015 and is based in Curitiba, Brazil.

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

# Business problem statement (GOALS):

1. Business Problem Understanding

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 can sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners. As per the dataset, after a customer purchases the product from Olist, very few customers repeats their purchase. The frequency of customers is very low.

1. Business Objective

The Machine Learning model to be developed to determine if the customer falls in the category of churned or Non-Churned by analyzing the details of the order history and its corresponding attributes. This shall help the business to predict which customer can be churned over the time and thus, concentrating more on those chunk of probable ‘Churned’ customers and converting them from probable churned customer to loyal customers by using various retention strategy.

Predict Customer Churn (‘1’ for Customer Churned and ‘0’ for Customer not churned) of the purchase from the Brazilian e-commerce site Olist.

*We are assuming that those customers who has not purchased for last six months are considered as churned customer*

1. Scope:

As per Kaggle, the given dataset is subset of actual data. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil.

1. Limitation:

The whole analysis can be performed on the basis of the given dataset. As per the Kaggle, this dataset is only subset of actual data. Hence, model can behave differently with respect to entire/actual data. Also, the dataset has multilingual data. Multilingual data is translated into English language using Google Translator. The accuracy of translation depends upon Google Translator.

# TOPIC SURVEY IN BRIEF:

1. Problem understanding:

Olist is the largest department store in Brazilian marketplaces. We are predicting the customer churn for this store. Those who did not purchase any product in the last six months are considered Churned customer otherwise they are considered non-churned. Basically predicting customer retention.

1. Current solution to the problem:

Since according to the data, around 60% of the customers are churned, the focus should be to build strategies to reduce the percentage of churned customers, thereby increasing the percentage of non-churned customers.

By observing the plots from the data provided, we see there are various reasons for a high churn percentage. In some places the delivery time has been really long even when the approval time was the least. So a large proportion of the orders where not delivered on time. Many times it is seen that the product availability is also an issue. There is a huge demand for household products and lot of customers use credit card payment modes to order products from the store. When it comes to Electronic items or Cosmetic items, the product prices are quite high and the sale volume is comparatively low.

Because of all these issues, apparently it seems that some good offers can be made on Electronic and Cosmetic items. And a lot of cashbacks or discount coupons and Gift Vouchers can be given to customers to keep the churn rate low.

1. Proposed solution to the problem:

Out of all the Machine learning algorithms applied to this problem where we are trying to predict the possible churn and non-churn for any customer, Random Forest classifier algorithm and the Ensemble techniques like Adaptive Boosting has performed quite well on both training and test data set, and is very likely to perform well in production, so that we can predict the possible churned customers. Though a bit of over fitting has also been a problem here, but most of the target points have been predicted accurately, given the domain we are working in. Once we know who can be the possible churned customer, the business can simply target those customers in some way and device some strategy to avoid losing them in the long run.

# CRITICAL ASSESSMENT OF TOPIC SURVEY:

1. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business:

Customer attrition, also termed as customer churn or loss of customers is critical as it affects the business growth. It is bad but inevitable, so it is important to track and try to keep it as low as possible. It can calculated on monthly, quarterly and yearly basis based on the business requirement. In this project we have analyzed why churn occurs, how to spot it and target those “at-risk”/unhappy group of customers based on their behaving similarity.

1. What key gaps are you trying to solve?

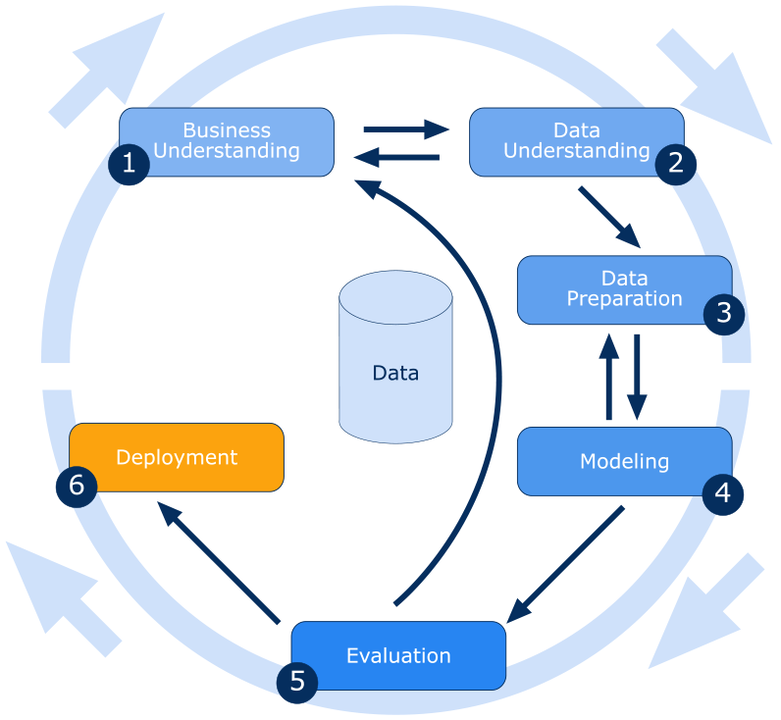
In this case study we find that customer attrition, is pretty high. This affects the growth of business and in turn year on year profits. To generate profit in this business model, customers need to be retained. To retain customers is easy rather than acquiring new customers as it adds cost. To mitigate this risk the customer loyalty and reward programs needs to be strengthen. It can be done by incentives, discounts and bundled offerings. At the same time priority should be given to optimize the supply chain systems with digitization and resolving customer complaints within a stipulated time.

# METHODOLOGY:

We will be using CRISP-DM methodology to implement our data science project.

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model with six phases that naturally describes the [data science life cycle](https://www.datascience-pm.com/domino-data-science-lifecycle/).It’s like a set of guardrails to help you plan, organize, and implement your data science (or machine learning) project.

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modelling?
4. Modelling – What modelling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

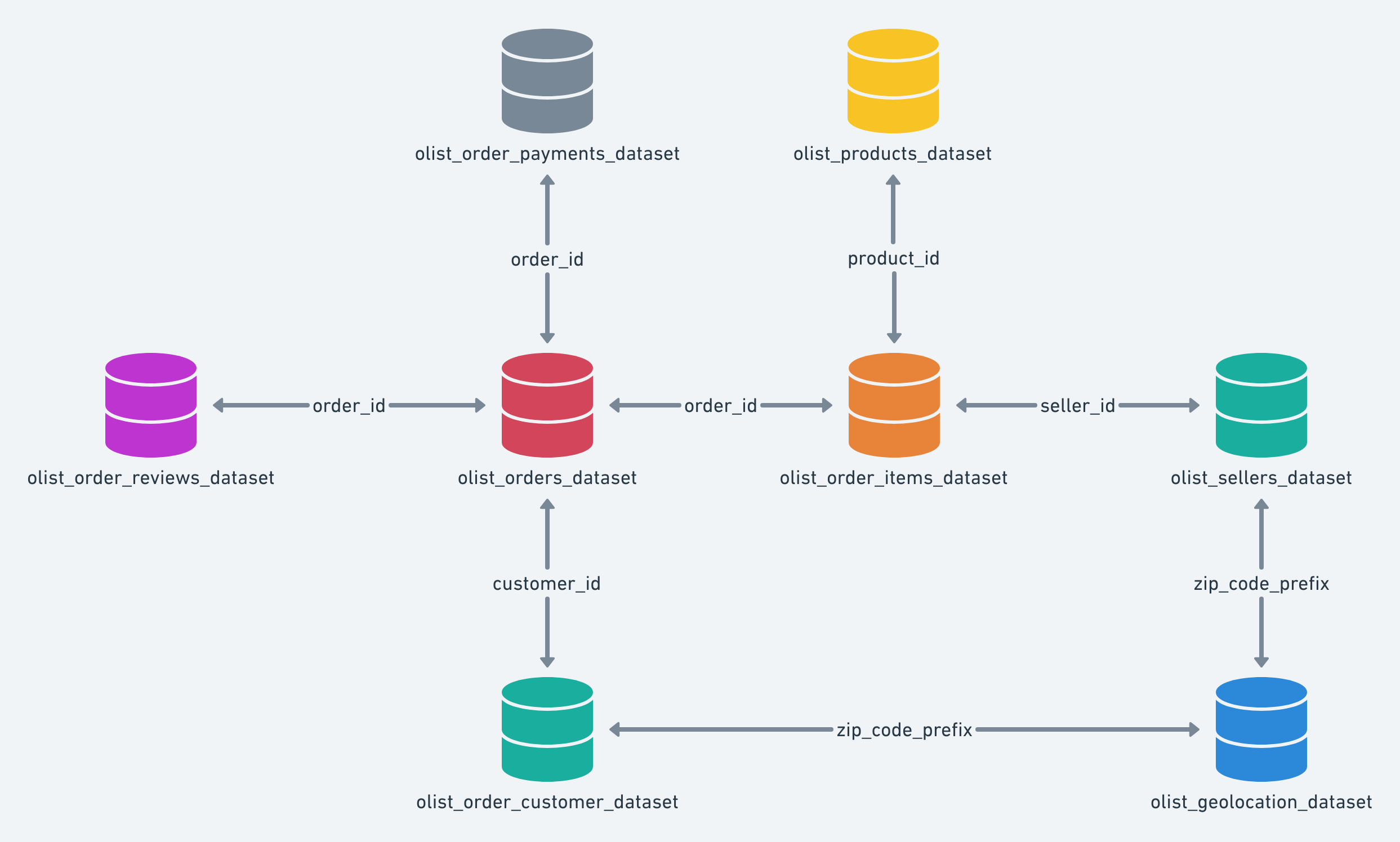


**1. Business Understanding:**

Olist is the largest departmental store in Brazil. 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. This is a Brazilian ecommerce public dataset of orders made at [Olist Store](http://www.olist.com/). 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

**2. Data Understanding:**

The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. There is data schema available for Olist.



Dataset comprises of 9 csv files

#### **Data Dictionary:**

###### 1. Customer Dataset (olist\_customers\_dataset.csv):

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.

* customer\_id: It is key to the orders dataset. Each order has a unique customer\_id.
* customer\_unique\_id: unique identifier of a customer.
* customer\_zip\_code\_prefix: First five digits of customer zip code.
* customer\_city: Customer city name.
* customer\_state: State of Customer.

###### 2. Geolocation Dataset (olist\_geolocation\_dataset.csv):

This dataset has information Brazilian zip codes and its lat/lng coordinates.

* geolocation\_zip\_code\_prefix: First 5 digits of zip code.
* geolocation\_lat: Latitude
* geolocation\_lng: Longitude
* geolocation\_city: The city name.
* geolocation\_state: The state name.

###### 3. Order Items Dataset (olist\_order\_items\_dataset.csv):

This dataset includes data about the items purchased within each order.

* 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)

##### 4*.* Payments Dataset (olist\_order\_payments\_dataset.csv):

This dataset includes data about the orders payment options.

* order\_id: unique identifier of an order.
* 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
* payment\_type: method of payment chosen by the customer.
* payment\_installments:number of installments chosen by the customer.
* payment\_value: transaction value.

##### 5. Order Reviews Dataset (olist\_order\_reviews\_dataset.csv):

This dataset includes data about the reviews made by the customers.

After a customer purchases the product from Olist Store a seller gets notified to fulfil that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

* review\_id: unique review identifier.
* order\_id: unique order identifier.
* 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\_comment\_message: Comment message 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\_timestamp: Shows satisfaction survey answer timestamp.

##### 6. Order Dataset (olist\_orders\_dataset.csv):

This dataset includes data about the items purchased within each order.

* order\_id: unique identifier of the order.
* customer\_id: key to the customer dataset. Each order has a unique customer\_id.
* order\_status: Reference to the order status (delivered, shipped, etc).
* order\_purchase\_timestamp: Shows the purchase timestamp.
* order\_approved\_at: Shows the payment approval timestamp.
* order\_delivered\_carrier\_date: Shows the order posting timestamp. When it was handled to the logistic partner.
* order\_delivered\_customer\_date: Shows the actual order delivery date to the customer.
* order\_estimated\_delivery\_date: Shows the estimated delivery date that was informed to customer at the purchase moment.

##### 7. Products Dataset (olist\_products\_dataset.csv):

This dataset includes data about the products sold by Olist

* product\_id: unique product identifier.
* product\_category\_name: root category of product, in Portuguese.
* product\_name\_lenght: number of characters extracted from the product name
* product\_description\_lenght: number of characters extracted from the product description.
* product\_photos\_qty: number of product published photos.
* product\_weight\_g: product weight measured in grams.
* product\_length\_cm: product length measured in centimeters.
* product\_height\_cm: product height measured in centimeters.
* product\_width\_cm: product width measured in centimeters.

##### 8. Sellers Dataset (olist\_sellers\_dataset.csv):

This dataset includes data about the sellers that fulfilled orders made at Olist. Use it to find the seller location and to identify which seller fulfilled each product.

* seller\_id: seller unique identifier
* seller\_zip\_code\_prefix: first 5 digits of seller zip code
* seller\_city: seller city name
* seller\_state: seller state

##### 9. Category Name Translation (product\_category\_name\_translation.csv):

Translates the productcategoryname to English.

* product\_category\_name: category name in Portuguese
* product\_category\_name\_english: category name in English

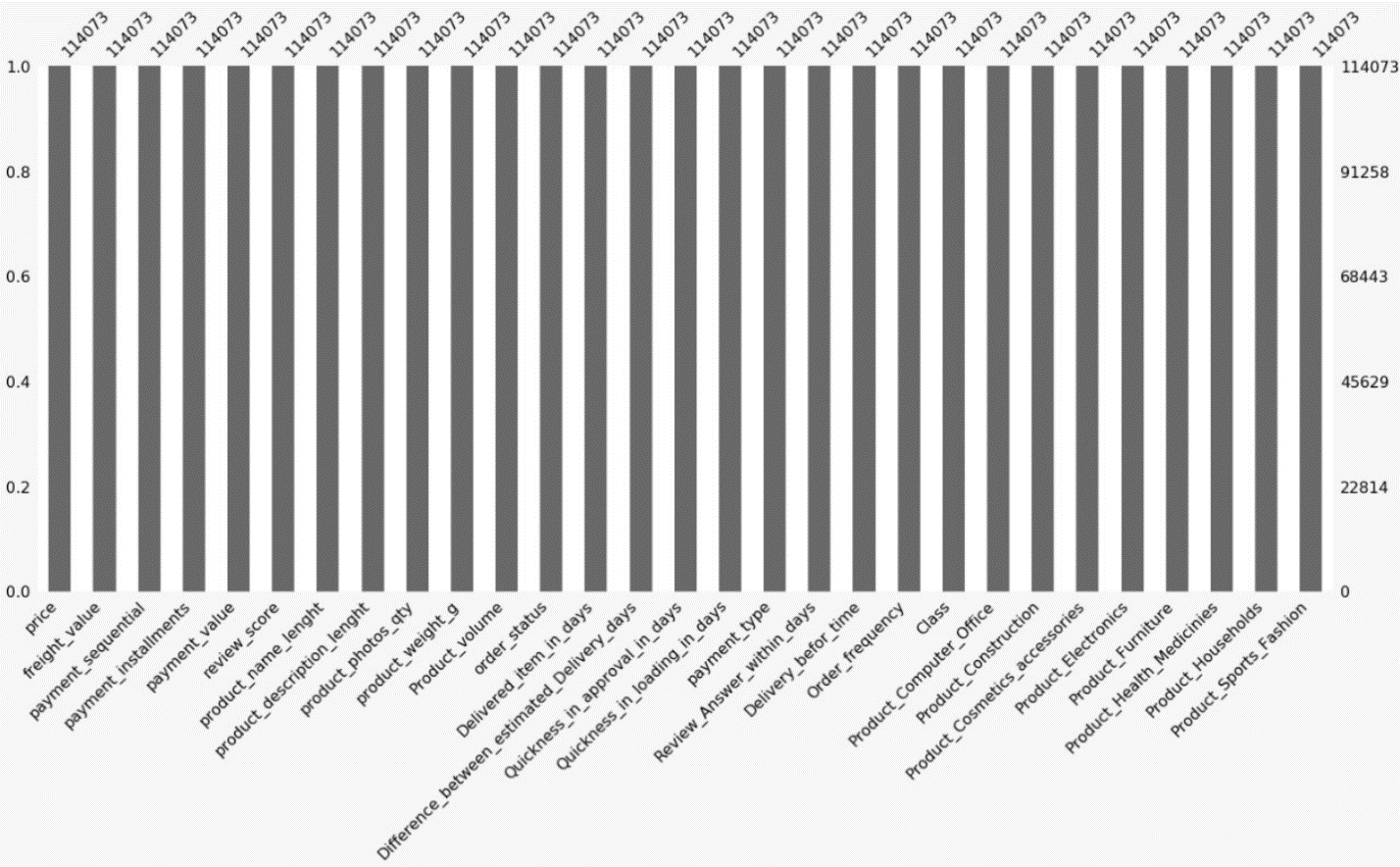
**3. Data Preparation:**

We have merged 9 csvs together and build common dataset for further analysis.

We have followed following steps for the preparation of dataset for model building.

* Missing Value Analysis:

We have observed around 4% of values are missing. We have impute the null values using ‘group by mean’



* Data Type Conversion:

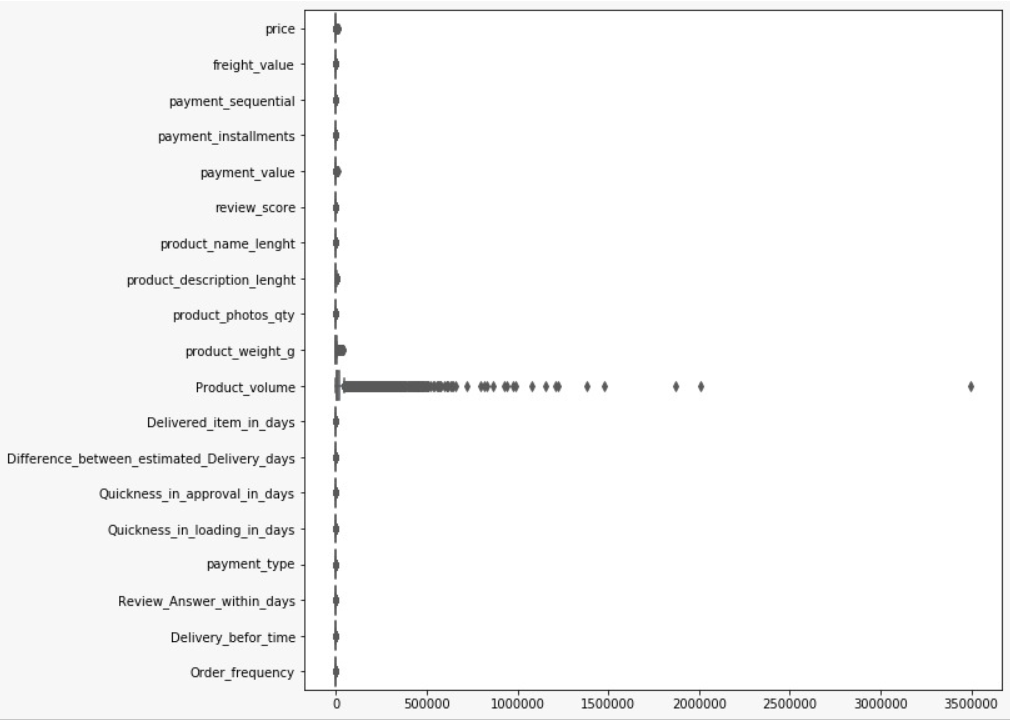
We have converted all the date related attributes features using ‘Date time’ conversion.

* Feature Engineering:

Order Frequency and Overall Price feature is added for each customer. Product Volume is added based on length, width and height attribute. Product Name is translated using Google translator. All the relevant ID columns has been dropped.

* Outlier Treatment:

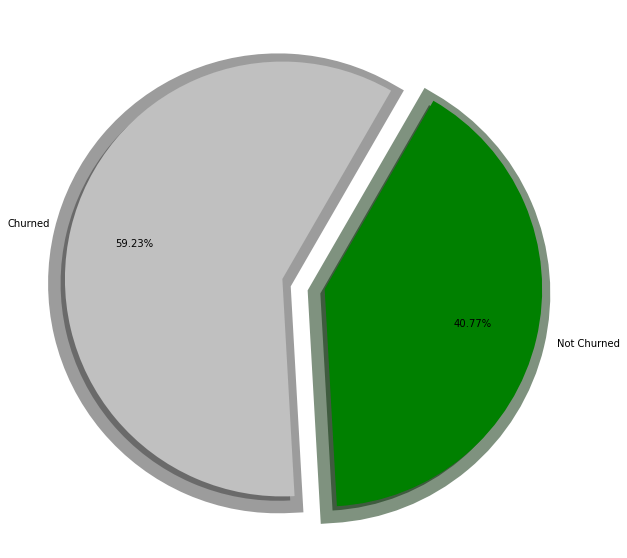
Almost all the columns has outliers. Hence we treat the outliers using ‘Log Transformation’



.

* Target Variable:

Target Variable Churned/Non-Churned has been added based on 6 months orders placed.



*As per our dataset, we have 59.23% are Churrned\_Customers and 40.77% are Non Churned Customers*

**Data Visualization:**

Chart, pie chart

Description automatically generated

* A large chunk of customers paid using Credit Cards.
* 16.86% customers paid using boleto
* 4.57% customers have used voucher for payment.
* 1.28% customers have paid using debit card.

Chart, pie chart

Description automatically generated

* Household items are essential items hence bought frequently.

Chart, pie chart

Description automatically generated

* A large chunk of reviews have remained positive ranging from 4-5 ratings.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Shape, rectangle

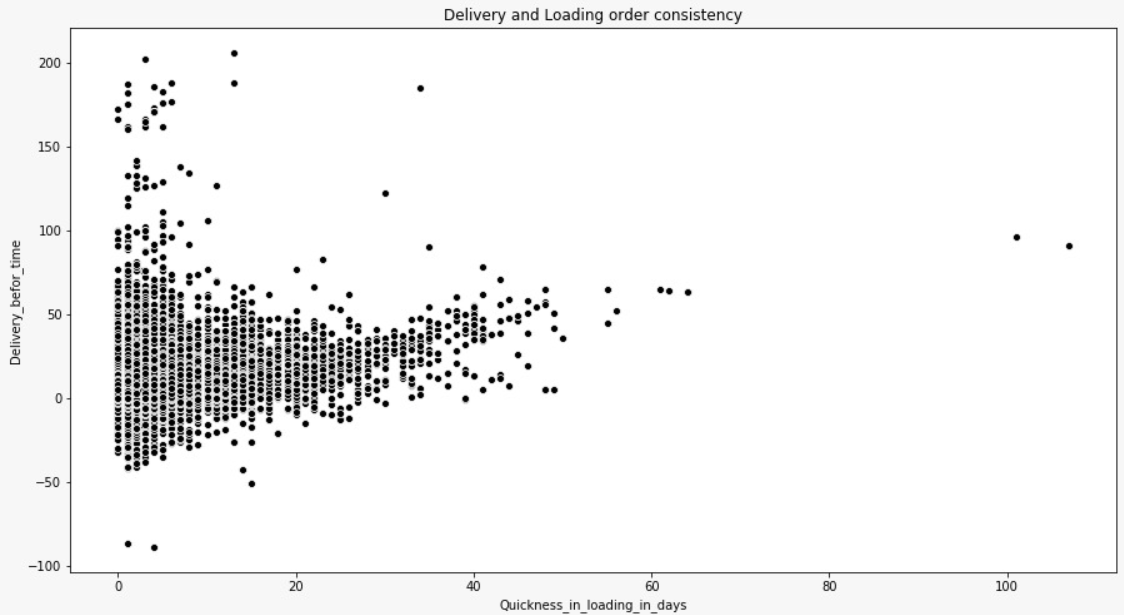
Description automatically generated

* The products of low price are exist in the market in a huge amount, which means people are buying these products only. Let us discuss this further for unique product and its frequency with respect to the customers.

Chart, scatter chart

Description automatically generated

* For many orders, we can see that, service approved the order at the same day, but delivery took place after several days.
* We can see there are some orders, which were approved so late, but delivered so quickly.
* For many orders, the order was quite soon approved, but delivery got very much late.



* We can see that there not more orders delivered before time, the proportion of the delivery late is more than that of earlier. Service Company should think of it very curiously, this can affect the customer churn.

Chart

Description automatically generated

* We found from the data is, no matter how many days it takes to approve customers' orders, the review is quite late
* There are many customers, who got their approval before 15 days, and company answered their review maximum by 8 days, in comparison to the approval time, review response is quite faster.

A screenshot of a computer

Description automatically generated with low confidence

* We sincerely observed that order frequency by 1 is very high from the service. Which means that the customers are mostly ordering only one product.

Chart, bar chart

Description automatically generated

* Most of the people churned the service who were the daily customers. This could be affecting the business, which will cause in future. Cause may be orders in time. If the order takes huge time they could be uncomfortable with it, remaining regular customer will leave service does not make any sense.
* From Stack bar plots we have got a clear analysis that, the rate of churn all the frequencies is quite larger than that of rate of non-churn.

Chart, bar chart

Description automatically generated

* Household product have been sold more any other products.
* Electronic items are bought less frequently.

Chart, bar chart

Description automatically generated

Additional Insights for Order Frequency are as below:

* Company should offer churned customer a good offer so that, they could purchase more than one product at a time, by which the order frequency will increase.
* The customers buying occasionally with low frequency, company should offer them some household products with reasonable price, if their count doesn't improve, then no benefit to entertain them. If they leave company does not suffer any huge difference.
* By messaging or call, attract those customers who are regular in ordering, get the knowledge of their leaving of the service. And increase the staff to place the order on the time. This could remain them connected.

**Statistical Tests:**

Statistical tests has been performed to determine relation between variables.

Statistical Tests: Hypotheses

1. Relationship between Approval times to the loading order time (One-way Anova test):

* Ho: Means are similar. There is no significant change difference between approval time to the loading time.
* H1: Means are dissimilar. There is a significant change difference between approval time to the loading time.

2. Relationship between order status and quickness in deliveries (Anova with olsmodel):

* Ho: Means are similar. There is indeed no specific relation in Order delivery consistency and status of the order.
* H1: Means are dissimilar. There is indeed a specific relation in Order delivery consistency and status of the order.

3. Relationship between order status and customer consistency (Chi2 Contingency test):

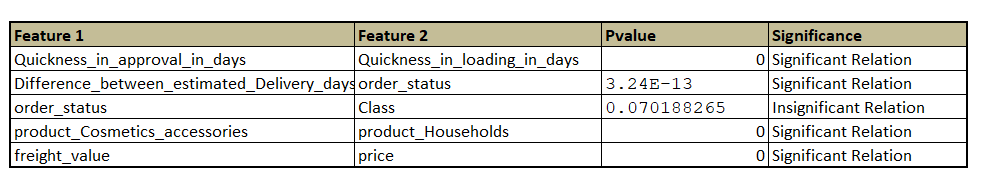
* Ho: Means are similar. There is no significant relation between order status and customer consistency.
* H1: Means are dissimilar. There is a significant relation between order status and customer consistency.

4. Relationship between order household products and cosmetic sales (Chi2 Contingency test):

* Ho: Means are similar. There is no significant relation between household products and cosmetic sales.
* H1: Means are dissimilar. There is a significant relation between household products and cosmetic sales.

5. Relationship between order product freight value and price of the products (One-way Anova test):

* Ho: Means are similar. There is no significant relation between freight value and price of respective product.
* H1: Means are dissimilar. There is a significant relation between freight value and price of respective product.



**Insights from Statistical Tests:**

* There is a significant change difference between approval times to the loading time.
* There is indeed no specific relation in Product and Price. Price is dependent on the product for some cases.
* There is indeed a specific relation in Customer consistency and Order frequency. The customers are ordering consistently they are either leaving the service or strongly stayed to the service.
* There is indeed a specific relation in Order delivery consistency and status of the order. For sure if the delivery gets earlier customers become happy with the services. They remain consistent with the service.
* There is no significant relation between order status and customer consistency. Which means if the order does not get approved by the customer, this does not lead to leave the service. But this scenario does not make sense for each circumstances.
* Customers purchasing households can attract towards the cosmetics accessories as well. Which means buyers are ladies for household stuffs. We can give them like cosmetic discount to increase the customers count. We have done so much Visualizations to prove this.
* Higher the price of products affect the frequency of orders.

People from the area are not keen in shopping. Company should bestow them greatest offers.

May be discount can be given to the customers who have a higher frequency in the order.

* There is an effect of description on reviews. Company should give the good description of the product to the customers. Elaborate the product very well.
* Price of the product depends upon the size for some cases like households or clothes and fashions. For some cases this is not related especially for the Electronic Devices. Hence, we can claim that most of the things in the service are available on household products. We also proved this in Visualizations.
* There is an effect of size and price on freight value. Some products are of lesser size with high expense can affect the freight value and vice versa.

**4. Model Building:**

Split: Data split is in by default manner 75:25 since the data is not imbalanced.

We have built following models to determine the accuracy.

Probabilistic Models:

1. Logistic Regression
2. Naïve Bayes

Deterministic Models:

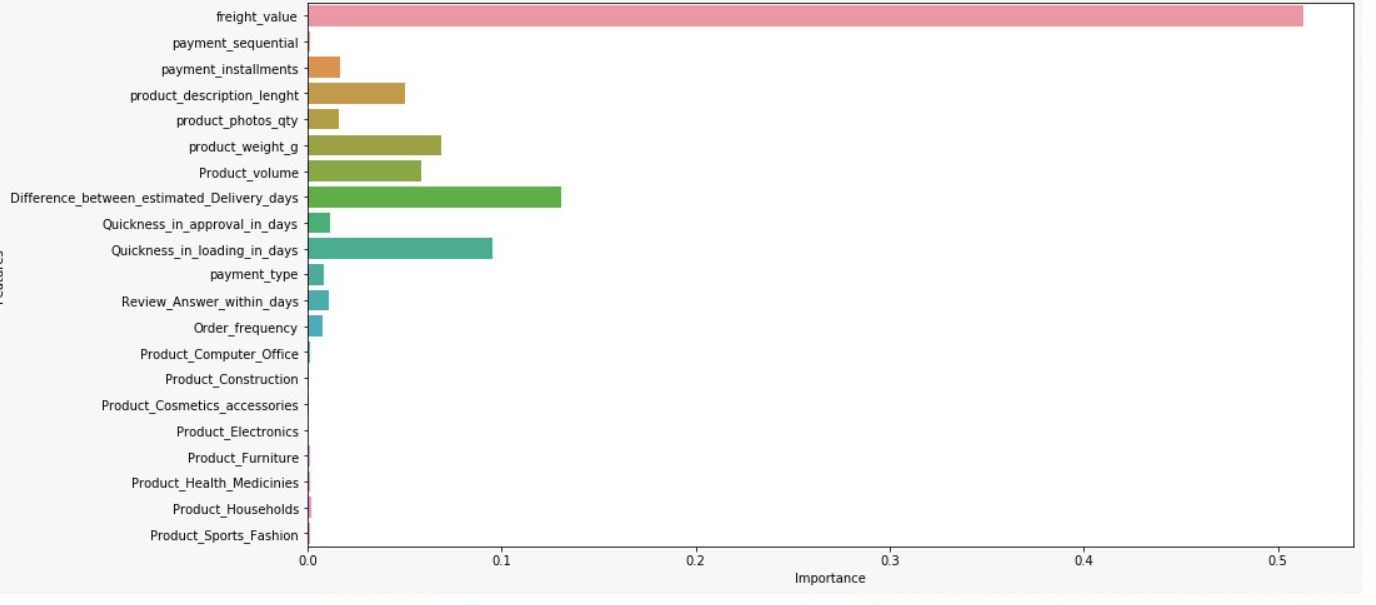
1. KNN
2. Decision Tree Classifier
3. Extra Tree Classifier

Ensemble Technique:

1. Stacking
2. Bootstrap Aggregation (Bagging)
3. Boosting
4. Adaptive Boosting
5. Gradient Boosting
6. XG Boost

**Machine Learning Models conclusion:**

We got following significant features from using Random Forest Classifier:



**Decision Tree Classifier:**

Classifier after tuning: Best parameters for the model are: {'criterion': 'gini', 'max\_depth':9}

Train, Test Accuracy: 0.806487, 0.791451

Precision Score: 0.7674058473736373

Recall Score: 0.9185201660735468

F1 Score: 0.8361906047516199

Auc\_Roc Score: 0.7622919756091611

**Stacking: Logistic, Naïve Bayes, Decision Tree Classifier**

Train, Test Accuracy: 0.788677, 0.777812

Precision Score: 0.7790448081566944

Recall Score: 0.8608392645314353

F1 Score: 0.8178648163150294

Auc\_Roc Score: 0.7578031936129235

**Bagging:**

Train, Test Accuracy: 0.822372, 0.805091

Precision Score: 0.7989551083591331

Recall Score: 0.9202995255041518

F1 Score: 0.8498105995973108

Auc\_Roc Score: 0.7938103324557251

**Random Forest Classifier:**

Classifier after tune: Best parameters for the model are: {'criterion':'gini', 'max\_depth': 9}

Train, Test Accuracy: 0.771590, 0.756266

Precision Score: 0.7346112632456244

Recall Score: 0.9175563463819691

F1 Score: 0.8150281363786824

Auc\_Roc Score: 0.7280667784957605

**Gradient Boosting:**

Train, Test Accuracy: 0.787504, 0.782373

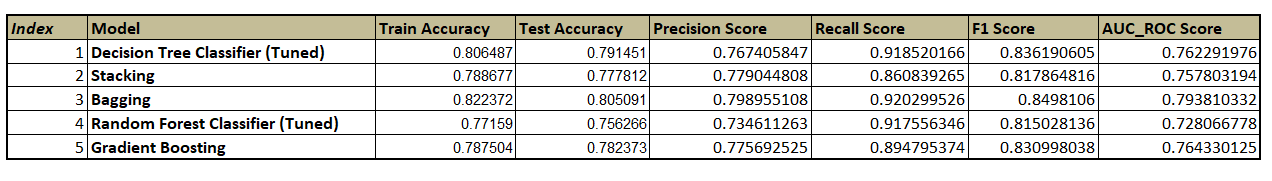
Precision Score: 0.775692525226557

Recall Score: 0.8947953736654805

F1 Score: 0.8309980376630978

Auc\_Roc Score: 0.7643301249359715

**Model Summary:**



**Best models for the entire project are:**

* Decision Tree after tuning
* Stacking
* Bagging
* Random Forest after tuning
* Gradient Boosting

**Additional Note:**

Multicollinearity Analysis for model improvement:

There was no improvement in the model accuracy and metric score after remove those features which were highly multicollinear. Indeed the accuracy went down.

Clustering:

Elbow point and Silhouette score suggested 2 optimal clusters for preprocessed data. Since this was not making a proper sense, hence K-mode clustering applied, though it deals with numerical as well as categorical data. Since the data was quite longer the code was not getting executed. We were planning to see the pattern of customers approaching to the market so that we can make a comfortable attraction towards them to increase our business.

**5. Deployment:**

Deployment is out of scope of our project.

# REFERENCES:

1. <https://www.kaggle.com/olistbr/brazilian-ecommerce>
2. <https://analyticsindiamag.com/7-types-classification-algorithms/>

.

**Notes For Project Team**

*Sample Reference for Datasets (to be filled by team and mentor)*

|  |  |
| --- | --- |
| Original owner of data | Kaggle.com |
| Data set information | 9 Csv Files,100000+ orders |
| Any past relevant articles using the dataset |  |
| Reference |  |
| Link to web page | <https://www.kaggle.com/olistbr/brazilian-ecommerce> |

### 

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*