



INDIVIDUAL ASSIGNMENT

AICT009-4-2-IDA

Introduction to Data Analytics

UCDF2005ICT(DI)

NAME : Tin Eugene
TP NUMBER : TP061195
HAND OUT DATE : 15th November 2021
WEIGHTAGE : 50%
LECTURER : Ms. Hema Latha Krishna Nair

Table of Contents

1.0 Introduction.....	4
2.0 Business Goal and Objectives.....	8
2.1 Aim	8
2.2 Objectives	8
Scope.....	8
Deliverables	8
3.0 Data Analytics Life Cycle and Methodology	9
3.1 Crisp-DM Methodology.....	9
3.1.1 Business Understanding.....	9
3.1.2 Data Understanding	9
3.1.3 Data Preparation.....	10
3.1.4 Modelling.....	10
3.1.5 Evaluation	10
3.1.6 Deployment.....	10
4.0 Dataset Understanding	11
4.1 Dataset Schema.....	14
5.0 Dataset Preparation	15
5.1 Data Selection	15
5.2 Data Cleaning and Transformation.....	18
5.2.1 Removing Column.....	18
5.2.2 Filtering Column.....	19
5.2.3 Renaming Column	20
5.2.4 Replacing null or unidentified values	21
5.2.5 Data format and type.....	21
5.3 Data Merging	22
5.4 Data Grouping.....	23

6.0 Data visualization.....	24
6.1 OLAP Report	24
6.1.1 Visul 1	25
6.1.2 Visual 2	26
6.1.3 Visual 3	27
6.1.4 Visual 4	28
6.1.5 Visual 5	29
6.2 OLAP Data Cube Operation	30
6.2.1 Display details of each product categories.....	30
6.3 Outlier Analysis	31
7.0 Data Modelling	32
7.1 Regression analysis.....	32
7.2 Pearson Correlation.....	32
7.3 Train-test split	34
7.3.1 Configuration of Train-Test Split in Python.....	34
7.4 Logistic Regression.....	35
7.5 Confusion Matrix	38
7.6 ROC Curve.....	40
7.7 Model Coefficient by features	42
8.0 Data security and privacy.....	43
9.0 Conclusion and Recommendation	44
10.0 Personal Reflection Report	44
11.0 Workload Matrix.....	45
12.0 References.....	46

1.0 Introduction

The project will be focused on e-commerce activities in Brazil. Before diving into the e-commerce situation in Brazil, we need to understand what e-commerce is first. Although there are many different definitions of e-commerce online, e-commerce is essentially a paperless exchange of business information using electronic data interchange, electronic mail, electronic bulletin boards, electronic funds transfer, the Internet and other network-based technologies. There are also many different categories of e-commerce. Such categories include business-to-business (B2B), business-to-consumer (B2C), consumer-to-consumer (C2C), consumer-to-business (C2B) and many more. Each of these categories have different methods of implementation, pricing strategies and requirements. As e-commerce acts as a new medium for exchanging value, there are several benefits over the traditional method. One of them is the convenience e-commerce provides. E-commerce sites can operate 24 hours a day and 7 days a week and this allows customers to browse through the store's catalogue whenever and wherever they are. E-commerce also allows customers to shop without leaving the comfort of their home. Another benefit of e-commerce is more geared towards the sellers. The sellers now can sell their products to a larger market as e-commerce allows them to reach to international customers. This can increase their revenue as there are more customers to serve. However, e-commerce also has its drawbacks. Customers buying products from e-commerce sites will need to wait for their product to be delivered to them. This can take from a few days to a few weeks. This could decrease the customer's satisfaction as they must wait as compared to when shopping in physical stores where they get their products immediately. Customers may also need to pay the delivery fees when they purchase from e-commerce sites and this could increase the product's price several times, especially so when shopping from other countries. Although e-commerce has its advantages and disadvantages, when implemented right, it could offset a lot of those negatives.

Brazil has the ninth largest GDP (Gross Domestic Product) in the world and has a population of 204 million and half of this number is in the middle class of the economy. However, most of them are still underserved in retail sectors and this opens the window for retailers to jump in and seize this opportunity through the growing e-commerce sector in this country. As the middle-class incomes grow, so does the internet penetration in the country, and this will in turn endorse the growth of e-commerce. With growing orders and revenue

throughout the years and an expected Compound Annual Growth Rate (CAGR) of 10%, the e-commerce sector in this country cannot be ignored.

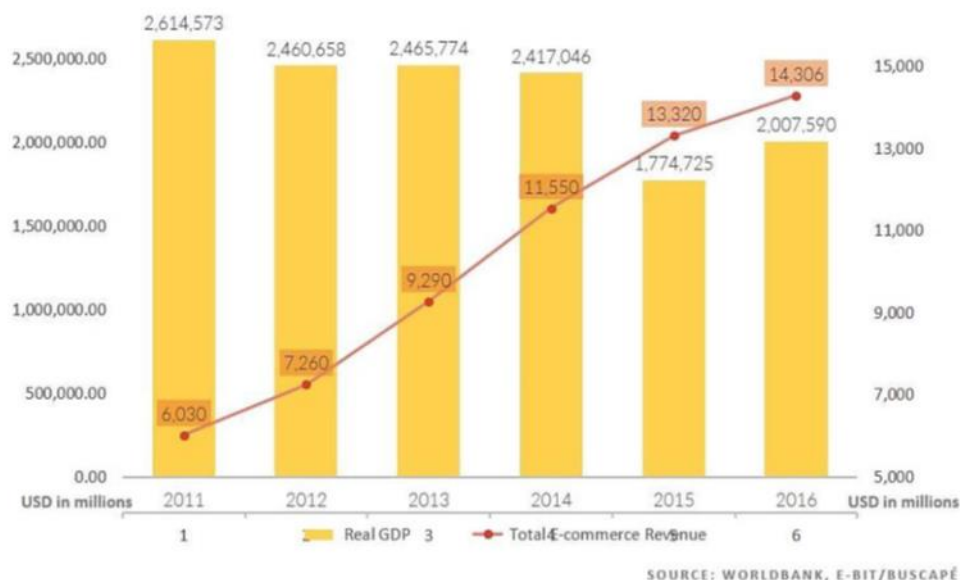


Figure 1 E-commerce growth in Brazil from 2011-2016

As shown in Figure 1, the e-commerce sector in Brazil has been growing steadily since 2011. However, even though this sector has been growing for the last decade, there are still some difficulties in this sector. One such difficulty is that there are high taxes, and the tax structure is very complicated. Businesses that wish to sell their products will need to understand this structure thoroughly to avoid any legal procedures in the future. Furthermore, the country's underdeveloped distribution infrastructure also proves to be a weak point for the e-commerce sector in the country. The picture below illustrates the comparison between the highway network infrastructure of Brazil in 2015 and United States of America.

This underdeveloped distribution network will impact the shipping of products by e-commerce as the unpaved and poor conditions of the road in Brazil pose a threat to the safety of the delivery trucks and drivers as well as lengthening the delivery times unnecessarily. Brazil is also infamous for its traffic congestion, with three of its cities in the top ten of most congested cities and Rio de Janeiro coming in fourth in the Traffic Index. These factors will affect the delivery schedules immensely. Although the e-commerce sector in Brazil was and is expected to flourish, the problems will still need to be addressed to further boost the growth of e-commerce and attract more foreign investment into this sector.

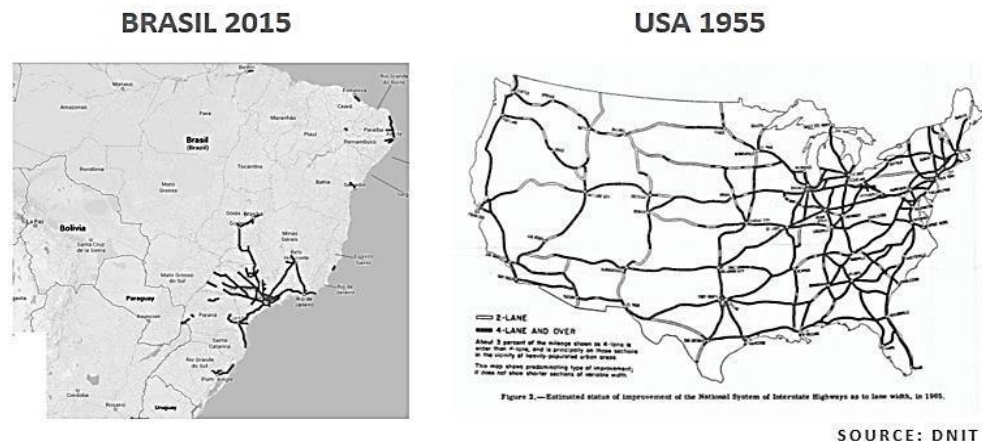


Figure 2 Highway network of Brazil and the USA

The analytics that will be done in this project will be used in hopes to solve some if not all the problems listed in the previous paragraph. The objectives and scopes of the project will be listed in one of the following sections. This project will analyse several aspects of the e-commerce activities in Brazil, such as the geolocations of the customers that bought items from e-commerce site in Brazil like Amazon from international sellers, B2W, Net shoes and so on. By analysing this information, courier companies or the sellers can plan their routes better to ensure that their customers can receive their item as soon as possible, providing the customers with the best shopping experience and retaining them. Besides that, this project will also focus on the type of products sold on the e-commerce platforms and the trending products in a certain time. With this information, recommendations can be given to the sellers to focus on certain product categories to boost their sales or to governments to provide incentives to certain product categories to grow their sales. This project will also use data analytics to produce a trend of the history of price movements in the e-commerce sector. This can help individuals to plan their expenses, like only purchase during holiday seasons or when the site is going to provide discounts.

This can also help the organizations or sellers to plan their investments into profiting product categories instead of focusing on loss making products. Furthermore, this project will also try to analyse the suppliers' information. This can help indicate how long it will take for the product to travel from the supplier's warehouse to the seller's warehouse. This can in turn

estimate how long it will take for an out-of-stock item to be restocked and solve supply and demand problems.

2.0 Business Goal and Objectives

2.1 Aim

To understand the trends and shifts of consumers behaviour and make data driven decisions that will assist in the growth of e-commerce.

2.2 Objectives

- ✓ To identify the bestselling product category based on the number of products sold, total sales generated by the respective category and predict the growth of the product categories in the future using the necessary OLAP and Predictive models.
- ✓ To analyse the product sold and its delivery time based on locations throughout certain city/states in Brazil.
- ✓ To find out if product price affects the purchase decision of the customers using Olist platform.
- ✓ To analyse and develop a predictive model to predict the future interest of consumers.

Scope

- ✓ To prepare an analytical dashboard that visualise bad reviews portrait from different perspective such as delivery time, product, and seller. Analytic will be conducted to study factors that causes poor product rating. The dataset used will be including Order Item Dataset, Order Review Dataset and Product Dataset provided by Olist.
- ✓ Create a predictive model that can understand factors that cause customer to leave bad review and predict bad customer reviews (scores 1 or 2) to minimize bad reviews and improve customer satisfaction.

Deliverables

- ✓ Explain the relative influence of each predictor on the model.
- ✓ Potential solution to reduce bad customer review
- ✓ Prediction model to classify bad review from customer

3.0 Data Analytics Life Cycle and Methodology

3.1 Crisp-DM Methodology

Crisp-DM as known as Cross Industry Standard Process for Data Mining, is implemented throughout the research paper as a blueprint. Crisp-DM consists of 6 phases that are used to analyse data. The phases include business understanding, data understanding, data preparation, modelling, evaluation and deployment.

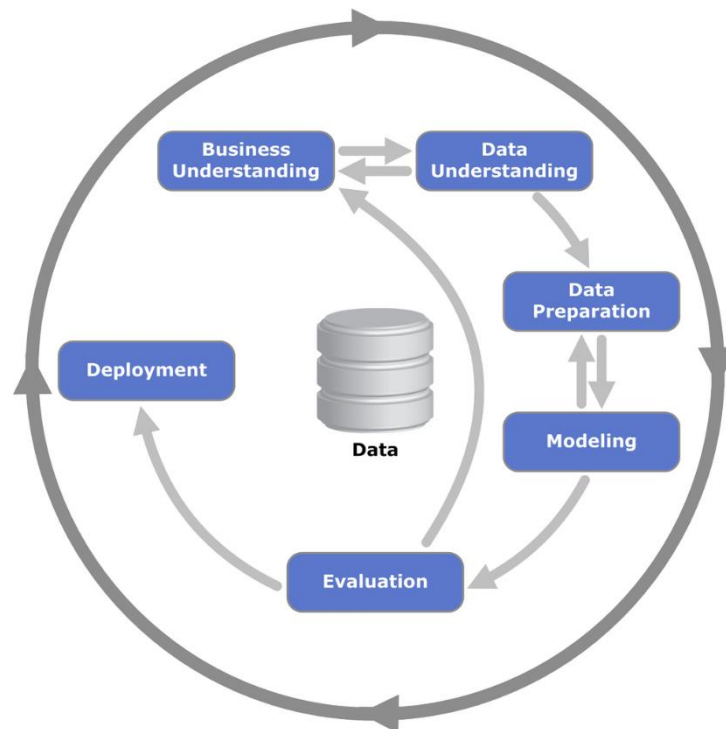


Figure 3 Crisp-DM Life Cycle (Wikipeddia, 2021)

3.1.1 Business Understanding

The first phase of Crisp-DM is that the data analyst shall thoroughly understand the business domain and draft a proposal including aim, objectives, and scopes for the analytics. After sorting out the proposal, individual should perform several assessments such as resources, equipment, knowledge, risk, and assumptions according to the goal of the research.

3.1.2 Data Understanding

In this phase, analysts are responsible to perform data gathering within the domain to perform data mining. Individual should prepare a data schema report that includes the attributes, quantity and description of dataset used to carry out in the report. Data description shall be well sorted in order to evaluate the integration of datasets to match the purpose of the business goal as compiled in the report section 4.0.

3.1.3 Data Preparation

The analysts are responsible to perform data cleaning and transformation to serve the purpose of business goal by removing inconsistencies and outliers from the data. The goal in this step is to increase data quality for the analysis requirement. Integration of data shall implement in this phase to create relationships between attributes in order to perform aggregation that are able to be used in the analysis.

3.1.4 Modelling

After integration between datasets, individual is required to assess a suitable modelling technique. The selection of model shall be based on the criteria of business goal and class attributes provided. The model will be selected to perform evaluation such as training, testing, finding out correlation between features. The model can be built based on evaluation of values and parameters. The model will be assessed with tools such as Power Bi and Python.

3.1.5 Evaluation

Evaluation phases are important for analyst to find out the solution and problems forecasted from the data mining model implemented. From the model evaluation, individual will have the ability to propose business solution. The accuracy of the model is important as it is directly impacted on the business decision that are made from the performance of the model.

3.1.6 Deployment

In this final phase of Crisp-DM, the evaluation result should be implemented by creating a business strategy and perform improvisation and maintenance to ensure the data mining result can be serve as a guidance efficiently in the deployed environment.

4.0 Dataset Understanding

The datasets that were used to analyse the customer ratings on e-commerce platform in Brazil are all sourced from the Kaggle website. The datasets provided are open-sourced.

Datasets Source: [Brazilian E-Commerce Public Dataset by Olist](#)

Dataset's reference including:

1. Olist_customers_dataset
2. Olist_order_items_dataset
3. Olist_order_reviews_dataset
4. Olist_orders_dataset
5. Olist_product_dataset
6. Olist_sellers_dataset
7. Product_category_name_translation

Olist_customers_dataset

Column name	Details	Values
Customer_id	key to the orders dataset	100% valid values
Customer_unique_id	unique identifier of a customer.	100% valid values
Customer_zip_code_prefix	first five digits of customer zip code	100% valid values
Customer_city	customer city name	100% valid values
Customer_state	customer state	100% valid values

Olist_order_items_dataset

Column name	Details	Values
order_id	order unique identifier	100% valid values
order_item_id	sequential number identifying number of items included in the same order.	100% valid values
product_id	product unique identifier	100% valid values

seller_id	seller unique identifier	100% valid values
shipping_limit_date	Shows the seller shipping limit date for handling the order over to the logistic partner.	100% valid values
price	item price	100% valid values
freight_value	item freight value item	100% valid values

Olist_order_reviews_dataset

Column name	Details	Values
review_id	unique review identifier	100% valid values
order_id	unique order identifier	100% valid values
review_score	Note ranging from 1 to 5 given by the customer on a satisfaction survey.	100% valid values
review_comment_title	Comment title from the review left by the customer, in Portuguese.	12% valid values 88% missing values
review_comment_message	Comment message from the review left by the customer, in Portuguese.	41% valid values 59% missing values
review_creation_date	Shows the date in which the satisfaction survey was sent to the customer.	100% valid values
review_answer_timestamp	Shows satisfaction survey answer timestamp.	100% valid values

Olist_orders_dataset

Column name	Details	Values
order_id	unique identifier of the order.	100% valid values
customer_id	key to the customer dataset. Each order has a unique customer_id.	100% valid values

order_status	Reference to the order status	100% valid values
order_purchase_timestamp	Shows the purchase timestamp.	100% valid values
order_approved_at	Shows the payment approval timestamp.	100% valid values
order_delivered_carrier_date	Shows the order posting timestamp.	98% valid values <2% missing values
order_delivered_customer_date	Shows the actual order delivery date to the customer.	97% valid values <3% missing values
order_estimated_delivery_date	Shows the estimated delivery date that was informed to customer at the purchase moment.	100% valid values

Olist_orders_dataset

Column name	Details	Values
product_id	unique product identifier	100% valid values
product_category_name	root category of product, in Portuguese.	98% valid values <2% missing values
product_name_length	number of characters extracted from the product name.	98% valid values <2% missing values
product_description_length	number of characters extracted from the product description.	98% valid values <2% missing values
product_photos_qty	number of product published photos	98% valid values <2% missing values
product_weight_g	product weight measured in grams.	100% valid values
product_length_cm	product length measured in centimetres.	100% valid values
product_height_cm	product height measured in centimetres.	100% valid values
product_width_cm	product width measured in centimetres.	100% valid values

Olist_orders_dataset

Column name	Details	Values
seller_id	seller unique identifier	100% valid values
seller_zip_code_prefix	first 5 digits of seller zip code	100% valid values
seller_city	seller city name	100% valid values
seller state	seller state	100% valid values

Olist_orders_dataset

Column name	Details	Values
product_category_name	category name in Portuguese	100% valid values
product_category_name_english	category name in English	100% valid values

4.1 Dataset Schema

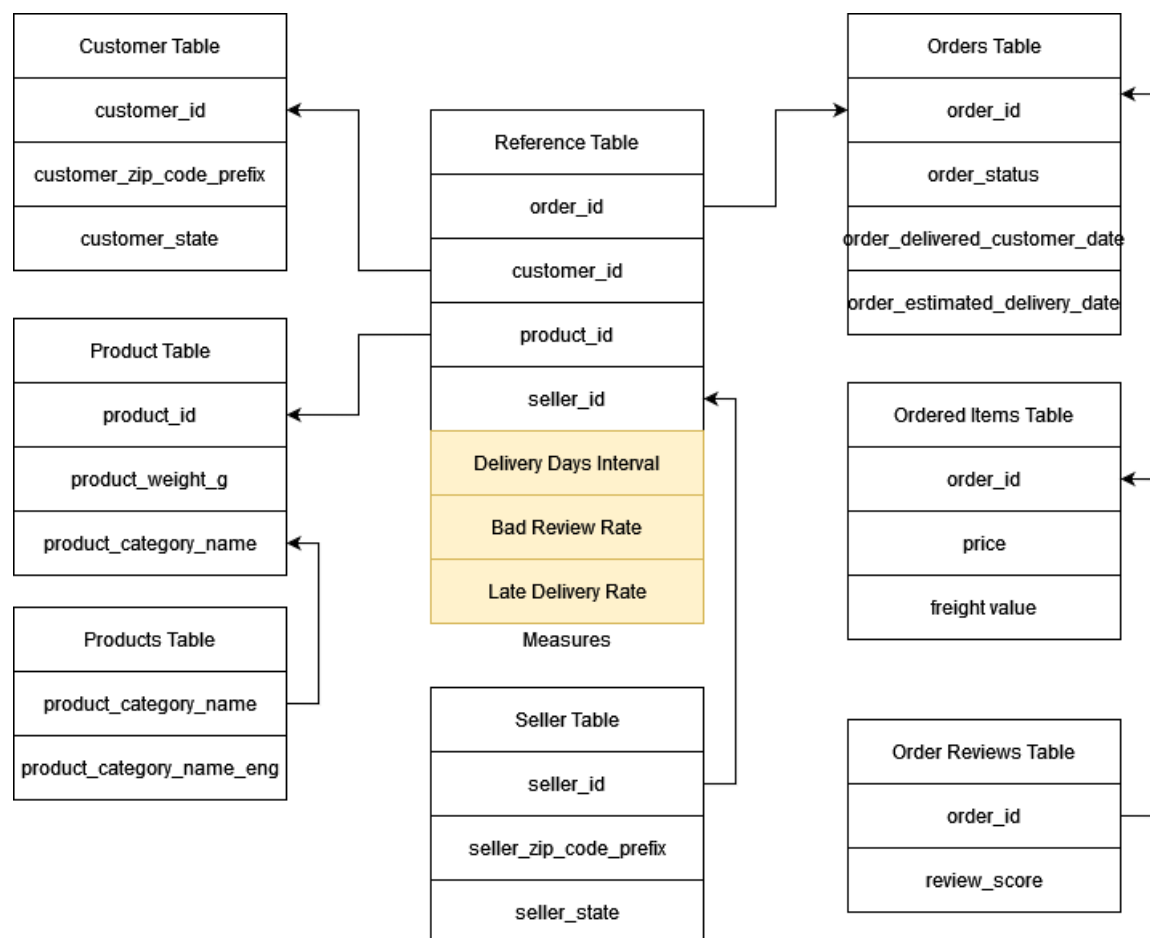
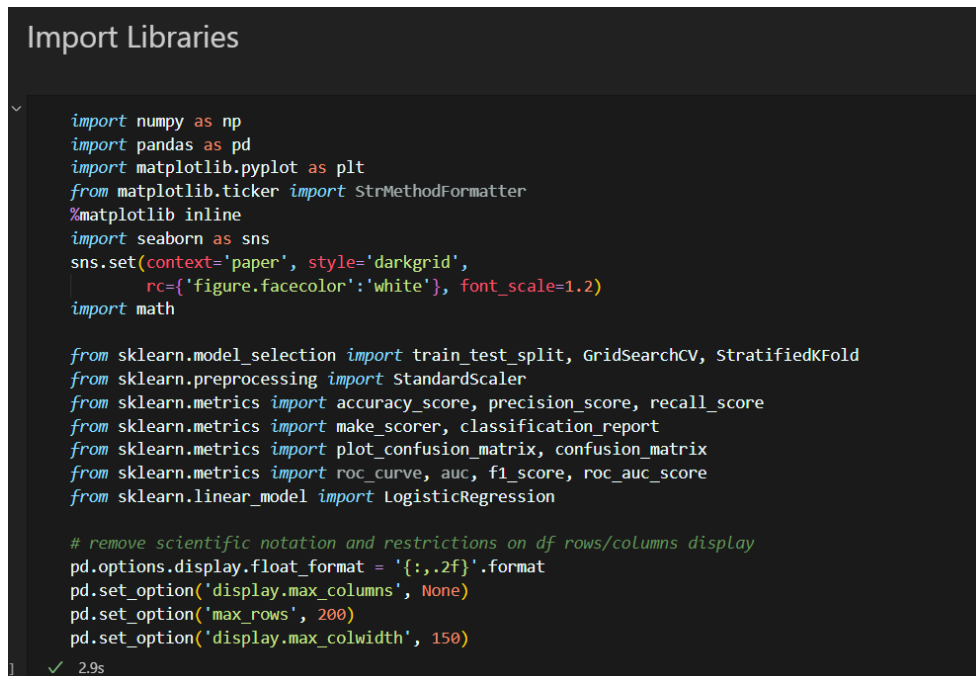


Figure 4 Star Schema

5.0 Dataset Preparation

5.1 Data Selection

To build an analytical dashboard for the datasets selected in section 4.0, the datasets must integrate into a single file with tools such as Python, Microsoft, Power Bi, Microsoft Excel, and Tableau. To prepare a visualisation and predictive model, we must filter and merge the data that are necessary and clean to acquire a precise and effective interpretation of data analysis. Microsoft Power Bi and Python are utilized simultaneously to create a functional visual report regarding bad review affecting product sales and proposed solution for the concurrent issue.



```
Import Libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.ticker import StrMethodFormatter
%matplotlib inline
import seaborn as sns
sns.set(context='paper', style='darkgrid',
        rc={'figure.facecolor':'white'}, font_scale=1.2)
import math

from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score
from sklearn.metrics import make_scorer, classification_report
from sklearn.metrics import plot_confusion_matrix, confusion_matrix
from sklearn.metrics import roc_curve, auc, f1_score, roc_auc_score
from sklearn.linear_model import LogisticRegression

# remove scientific notation and restrictions on df rows/columns display
pd.options.display.float_format = '{:,.2f}'.format
pd.set_option('display.max_columns', None)
pd.set_option('max_rows', 200)
pd.set_option('display.max_colwidth', 150)
```

Figure 5 import library in Python

Before importing data into a Jupyter Notebook, several Python libraries are required to prepare the data analysis as shown in Figure 5 above. After importing every necessary component into the workspace, the repetitive process of importing datasets and performing data check will proceed.

Customers

```
# Load customers dataframe
file_custs = '../datasets/olist_customers_dataset.csv'
customers = pd.read_csv(file_custs, dtype={'customer_zip_code_prefix': str})
customers.head()
```

✓ 0.3s

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	09790	sao bernardo do campo	SP

Figure 6 Data import

In figure 6, the data are able to import with 'pandas' library that are available in Python. The data are retrieved from Excel Spreadsheet and certain column values are parsed into Python readable format. Several rows of data are queried to validate any error. Several datasets mentioned in section 4 are also imported using the similar line of codes as shown in Figure 7.

```
# Load orders dataframe
file_orders = 'data/olist_orders_dataset.csv'
orders = pd.read_csv(file_orders, parse_dates=['order_purchase_timestamp',
'order_approved_at',
'order_delivered_carrier_date',
'order_delivered_customer_date',
'order_estimated_delivery_date'])

# Load order items dataframe
file_order_items = 'data/olist_order_items_dataset.csv'
order_items = pd.read_csv(file_order_items, parse_dates=['shipping_limit_date'])

# Load products dataframe
file_products = 'data/olist_products_dataset.csv'
products = pd.read_csv(file_products)

# Load product categories dataframe
file_product_cat = 'data/product_category_name_translation.csv'
product_cat = pd.read_csv(file_product_cat)

# Load sellers dataframe
file_sellers = 'data/olist_sellers_dataset.csv'

# Load reviews dataframe
file_reviews = 'data/olist_order_reviews_dataset_delim.csv'
reviews = pd.read_csv(file_reviews, parse_dates=['review_creation_date', 'review_answer_timestamp'])
```

Figure 7 Remaining Data import

After sorting out the code, the remaining data cleaning process are decided to switch to Microsoft Power Bi for simplicity. To inject the data into Microsoft Power Bi, a new file with extension 'pbix' are required to launch Power Bi. Inside Power Bi, on the ribbon tab section, click on transform data as shown below.

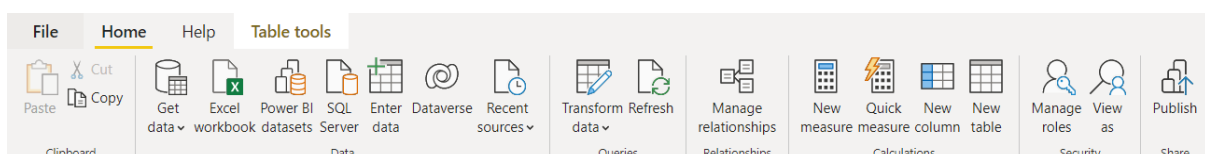


Figure 8 Power Bi Ribbon Tab

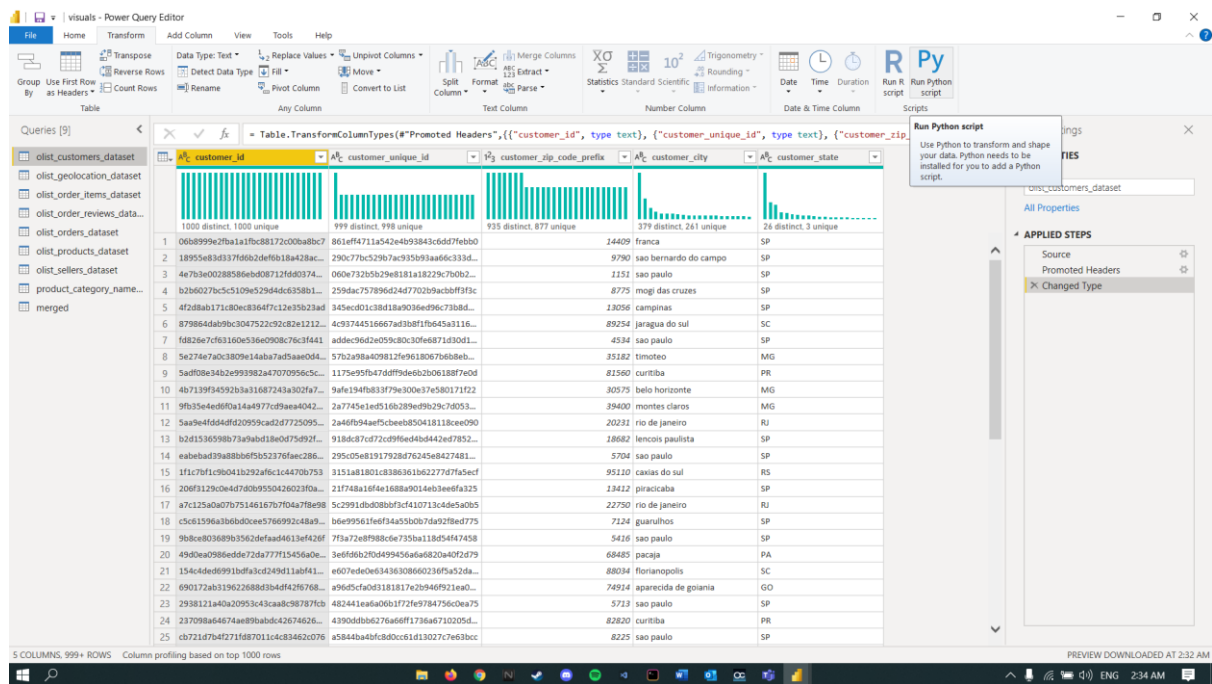


Figure 9 Power Bi Power Query Editor

On the Transform Tab, hover towards the 'Run Python Script' to inject the python source code into the Power Bi as shown in Figure 10 below.

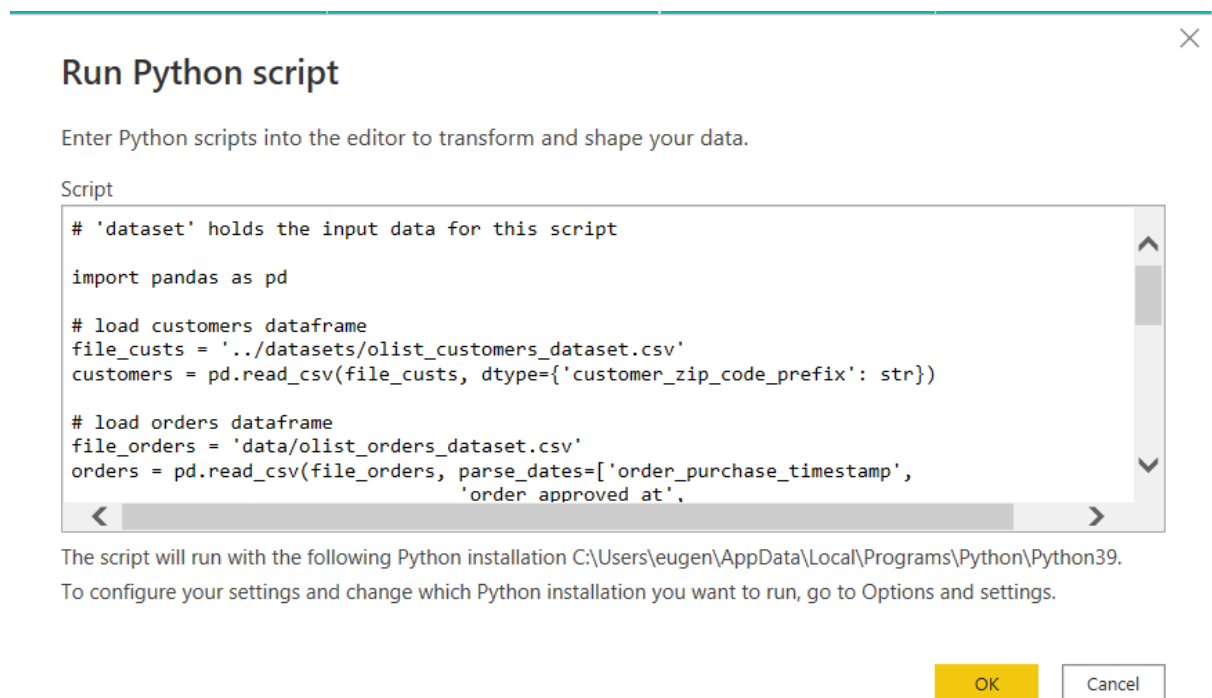


Figure 10 Power Query Editor 'Run Python Script'

Upon completion of data injection, data cleaning process are able to begin.

5.2 Data Cleaning and Transformation

Data filtering process happens here to remove null values and noisy data that might affect the output of the result at the end of the report. As mentioned earlier, several tables have a certain missing values or outliers are needed to be filled or replaced. However, certain columns that has faulty may not needed in the data analysis. In this case, analysis will be surrounding the order reviews and deliveries, columns that are unrelated to these objectives are suggested to be dropped. The columns include:

5.2.1 Removing Column

Column	Source Table
customer_unique_id	olist_customer_dataset
shipping_limit_date	olist_order_items_dataset
order_approved_at	olist_orders_dataset
order_delivered_carrier_date	olist_orders_dataset
product_name_lenght	olist_products_dataset
product_description_lenght	olist_products_dataset
product_photos_qty	olist_products_dataset
review_comment_title	olist_order_reviews_dataset
review_comment_message	olist_order_reviews_dataset

Although review comments title and messages correlate with the objectives and scope of the data analysis. It has been removed because it has more than 59% null values which nullifies the accuracy of the data during the text analysis process. Text translation is also a difficulty considering its commented language is in Portuguese.

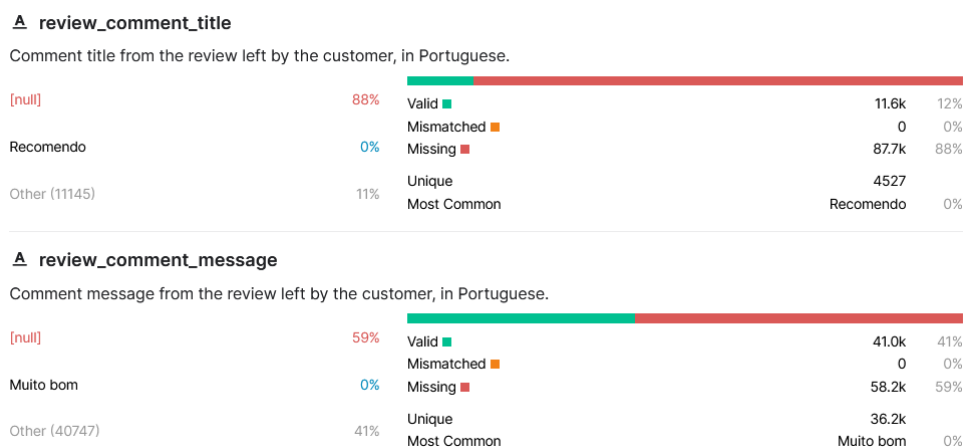


Figure 11 order reviews column information

5.2.2 Filtering Column

A order_status

Reference to the order status (delivered, shipped, etc).

delivered	97%	Valid	99.4k	100%
		Mismatched	0	0%
shipped	1%	Missing	0	0%
Other (1856)	2%	Unique	8	
		Most Common	delivered	97%

Figure 12 order status column details

There are several options when it comes to filtering data columns. The column that remains must be validated and fulfil the needs of the analytic process. In the case of resolving bad reviews issues, data that are being testified must be a delivered order. It is because ratings that are received from the customer must be based on the delivered orders and does not create any inaccuracy and bias of the data. It is reasonable to remove other type of status as it only consists 3% of the total order counts, which has less to nonsignificant impact on the analysis. Hence, the most important column, which is `order_status` within the `olist_orders_dataset` must be filtered to only remains order status which is delivered as shown in the figure below.

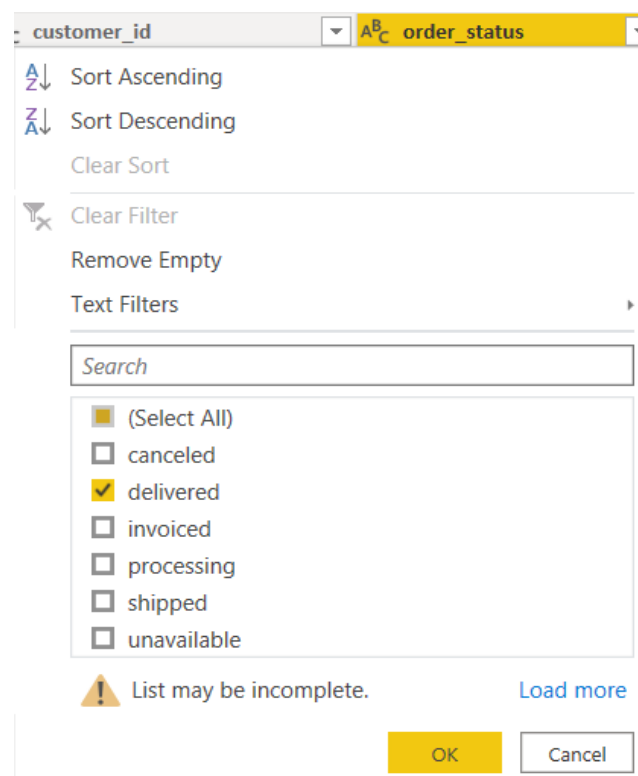


Figure 13 Filtering order status columns

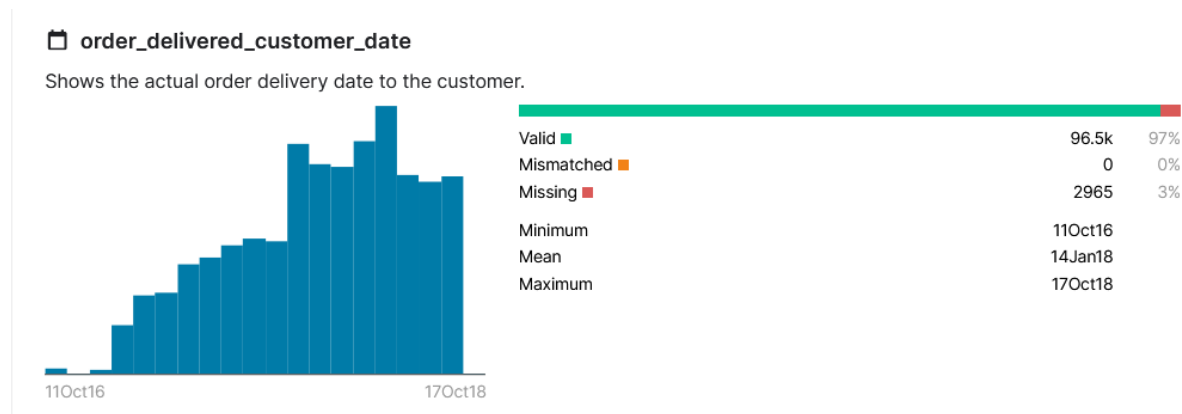


Figure 14 order delivered customer column details

Other than that, it is also important to include the range of the date that are going to be applied into the test set. After reviewing several timestamp columns that are included in the dataset, orders that are placed during the year 2016 will be filtered out due to its insignificant amount of order counts. Notice that there is small quantity of null values appearing in the rows because the column consists of data that are not yet been delivered. However, the dataset has been filtered to only include orders that has been delivered. Hence, no null rows that are necessary to be removed from the dataset.

5.2.3 Renaming Column

After filtering several datasets, renaming of columns should be considered as it simplifies the meaning of every column and prevent any misrelation between data. Several columns that have been rename including:-

Previous column name	Renamed column	Source
order_delivered_customer_date	act_delivery_date	olist_orders_dataset
order_estimated_customer_date	est_delivery_date	olist_orders_dataset
product_category_name	product category	olist_products_dataset

5.2.4 Replacing null or unidentified values

A product_category_name

root category of product, in Portuguese.



Figure 15 product category column information

Several rows that have null values shall be identified and eliminated as it will significantly affecting the visualization of report. One significant example from the dataset acquired are product category from the `olist_products_dataset` as shown in the figure above. It consists of 73 unique values and 610 units of missing values. There is several options to handle the 610 rows of orders, which includes removing and replacing it. In this report, latter decision was made. Instead of removing it, the empty values are able to convert into a new category named 'others'

5.2.5 Data format and type

Last but not least, the data format is also an important step to be assessed. It is because different data types responsible to different analysis methods. For instance, if the column that has values of price that are formatted into string, user will face issue when trying to perform measures on price column to calculate the total sum of price or average price. Hence, it should be taken consideration when preparing a set of clean data.

In the case of the dataset, several columns require a validation to prevent further complication to happen. Columns that should be formatted includes:-

Column	Format
Date	dd/mm/yy
Currency	R\$
Zip Code	99999-999

*Integer and float conversion should be careful as it may change the values on calculations.

5.3 Data Merging

Data merging are easy to execute within Power Bi, user are required to navigate into the transform data button inside the ribbon tab located on the upper section of the Power Bi interface. Upon clicking the transform data button, a new interface will prompt, which is the power query editor. Again, on top of the ribbon tab section, user shall select the merge queries button and bring up another popup that enable user to merge different datasets that are already imported into the workspace. On this interface is where user is allowed to perform data merging as shown in the figure below.

Merge

Select a table and matching columns to create a merged table.

olist_products_dataset

product_id	product_category_name_translation	product_weight	product_volume
1e9e8ef04dbcf4541ed26657ea517e5	perfumery	225.0	2240.0
6a2fb4dd53d2cdb88e0432f1284a004c	perfumery	400.0	2700.0
3aa071139cb16b67ca9e5dea641aaa2f	art	1000.0	10800.0
e3e020af31d4d89d2602272b315c3f6e	health_beauty	75.0	1911.0

olist_order_items_dataset

order_id	product_id	seller_id	shipping
00010242fe8c5a6d1ba2dd792cb16214	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	
00018f77f2f0320c557190d7a144bdd3	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	
000229ec398224ef6ca0657da4fc703e	c777355d18b72b67abbef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	
00024acbcd0a6daa1e931b038114c75	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	

Join Kind

Left Outer (all from first, matching from second)

☐ Use fuzzy matching to perform the merge

> Fuzzy matching options

Estimating matches based on data previews

OK Cancel

Figure 16 Data Merging Popup

Within the popup, user is required to select both primary keys that are related to each other to create a relationship between both datasets. After selecting the identical primary keys, user should proceed to decide on the join methods which includes left outer, right outer, inner join and so on. Primary keys and join methods are identified in the star schema plotted in the section 4.1 above.

5.4 Data Grouping

Groups [X]

Name: Field:

Group type:

Ungrouped values

- agro_industry_and_commerce
- air_conditioning
- art
- arts_and_craftmanship
- audio
- auto
- baby
- bed_bath_table
- books_general_interest
- books_imported

Groups and members

[Group] [Ungroup] ☐ Include Other group ⓘ

[OK] [Cancel]

Figure 17 Data Grouping

After transformation of data, there is one column of merged data that should be given attention. The values within the product category are similar such as furniture décor and furniture bedroom. These two values are similar items that are furniture related. Hence, both categories can be further group into a parent column called furniture. Other product categories are converted into subcategories by grouping them into larger and generalized categories as shown in the figure below:

- ✓ (Select all)
- ✓ Apparel & Accessories
- ✓ Automobile
- ✓ Children & Baby
- ✓ Construction Tools
- ✓ Cool Stuff
- ✓ Electronics
- ✓ Furnitures
- ✓ Groceries
- ✓ Health Beauty
- ✓ Holiday & Gifts
- ✓ Home
- ✓ Home Appliances
- ✓ Misc
- ✓ Sport & Hobbies

Figure 5 Product Category (Parent)

6.0 Data visualization

Data visualization can be implemented upon completion of data preparation. To create an analytical dashboard to visualize how different features may affecting customer rating, attribute such as bad review rate are often used to compare against order count, late delivery count, categories, and states of Brazil.

6.1 OLAP Report

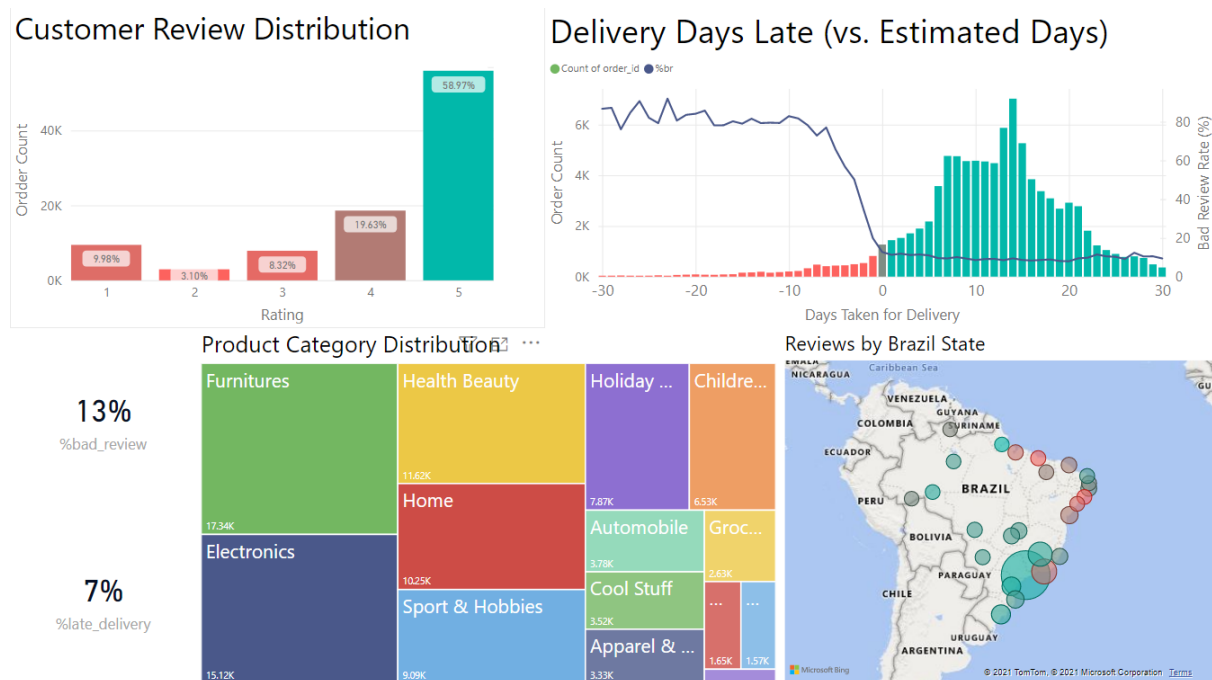


Figure 18 Bad reviews visualization

Source: [Visual Report of Bad Review Rate in Power Bi](#)

Description

Figure 15 above displays several attributes that are affecting the customer by submitting bad reviews on products, attributes that might directly or indirectly causes bad review including days taken to delivery, product quality, seller services, and delivery distance between customer and seller in between the year 2017 and 2018. Year 2016 has been excluded from the analysis because it does not contain significant count of reviews to affect the result of the visualization. The OLAP visuals displayed are able to hover around and perform drill down in the Power Bi application that would be explained at latter part.

6.1.1 Visual 1

Customer Review Distribution

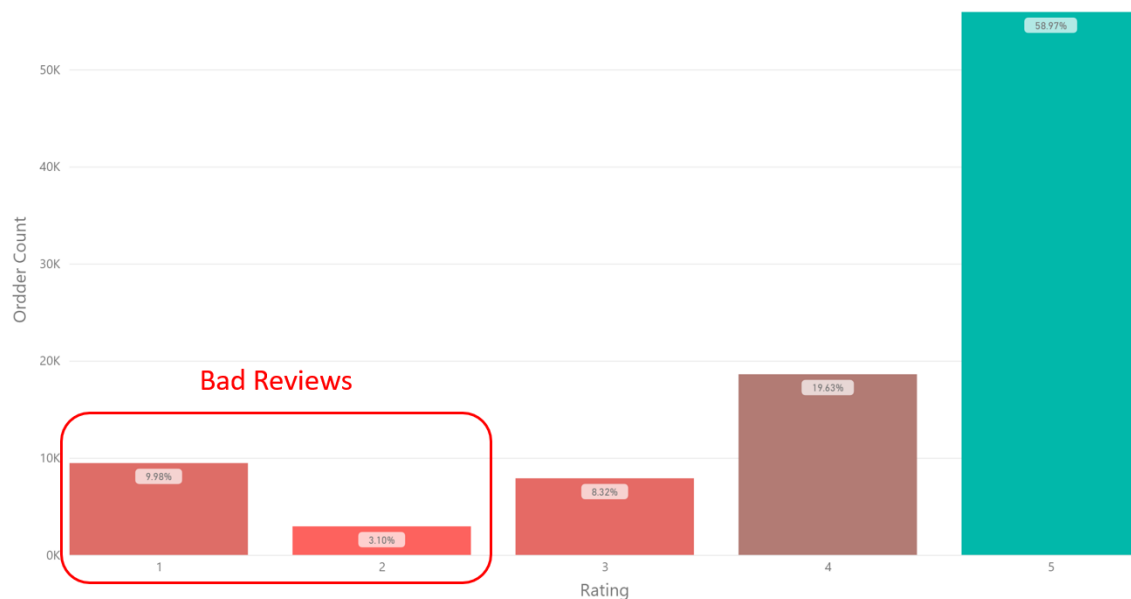


Figure 19 Customer Review Distribution

The visual of customer review score distribution are created using line and stacked column chart that are displaying the total order counts against rating score ranging from 1 (unsatisfied) – 5 (satisfied). The review score must be a binary target variable either 1 or 2 to be consider as bad reviews. As shown in figure above, the rating with value 1 are approximately 10% and approximately 3% on value 2 respectively which accounts to roughly 13% of the total sample of delivered orders. To investigate factors that causes such high negative rating in the e-commerce system, a machine learning classification model that can predict bad customer reviews are required.



When hovered, each bar of the rating will display a tooltip showing the amount of delivered orders and late delivery rate in that review score bar.

6.1.2 Visual 2

Delivery Days Late (vs. Estimated Days)

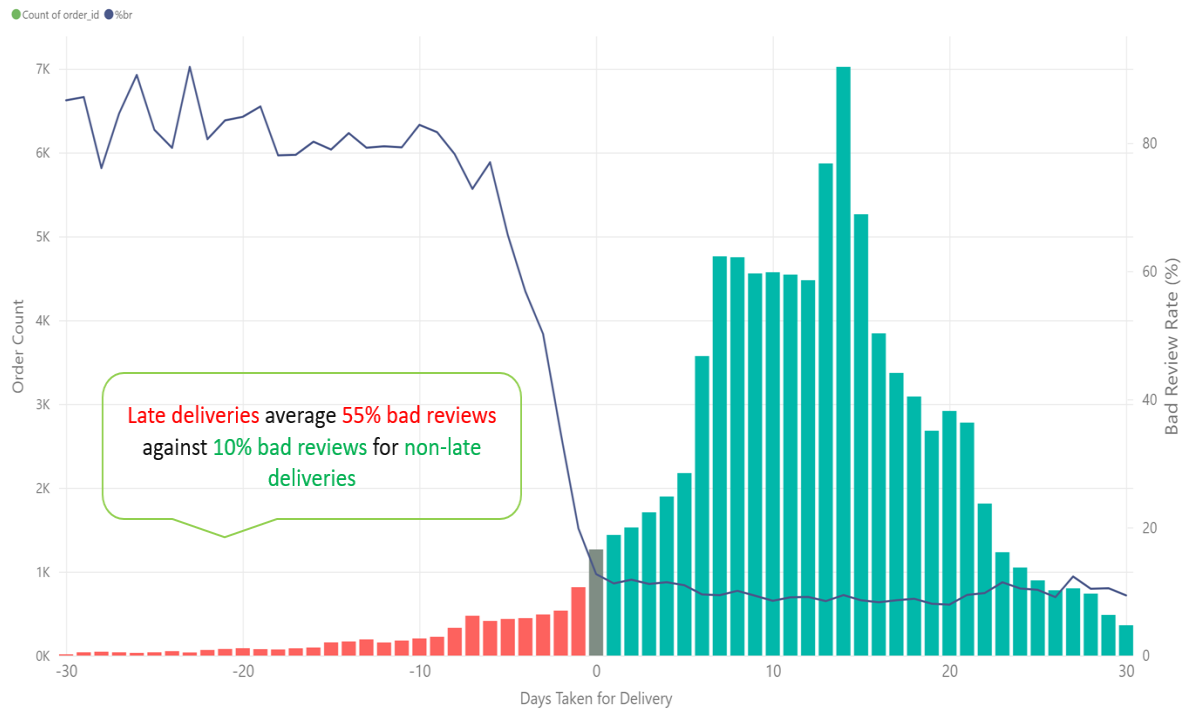


Figure 20 Delivery Days Late (vs. Estimated Days)

In the line and stacked column chart above displays the order count against the days taken to deliver the product. The e-commerce system provides an estimated delivery date to the customer, where it can be compared to the actual delivery date in a single order. By utilizing the estimated date and actual date attributes, we can construct a measure by calculating the days interval between two dates to visualize the relationship between late deliveries and bad reviews. By subtracting the actual delivery days from the estimated delivery days, we will get a value indicating the days taken to deliver an order. For orders that are past actual delivery date (negative value to estimated date), customer reviews show a significant increment by 55% compared to the 10% bad review submitted by consumer receiving their order on time. The line graph indicating the percentage of the bad review given by each day's interval, had proven that late deliveries will result in significantly higher percentage of bad rating compared to the order delivered on time, where bad review may result from product quality, seller attitude, and so on. The relationship concludes that the longer the day taken to deliver orders, the higher proportion of bad reviews will be given from customer.

6.1.3 Visual 3

Product Category Distribution

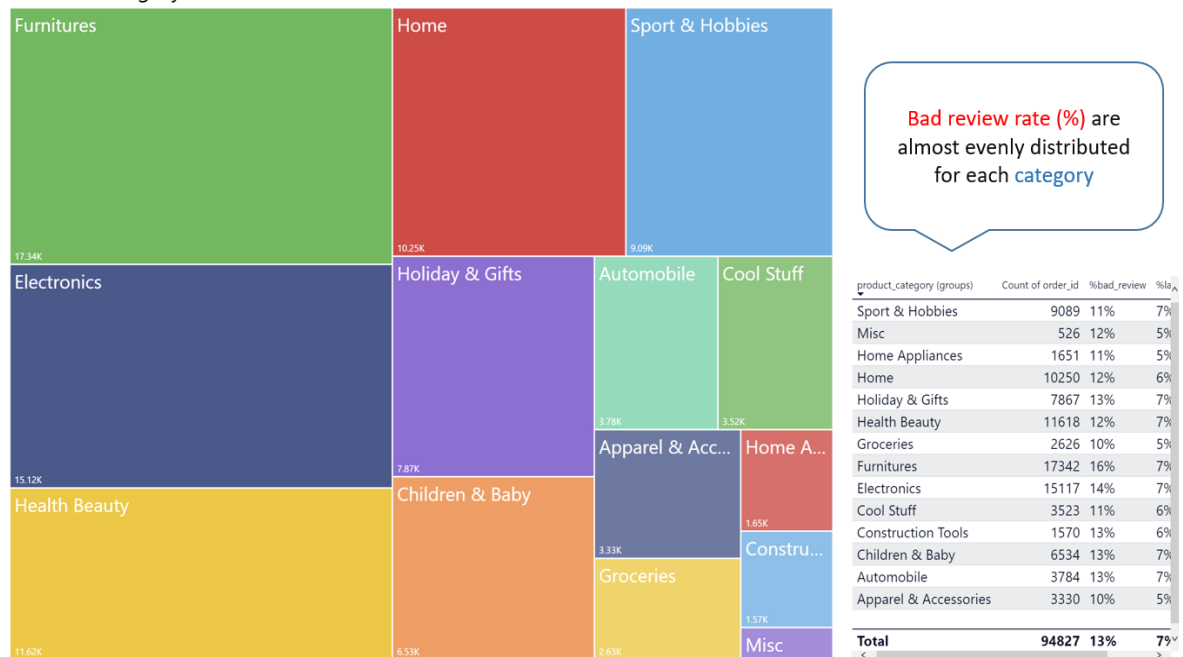
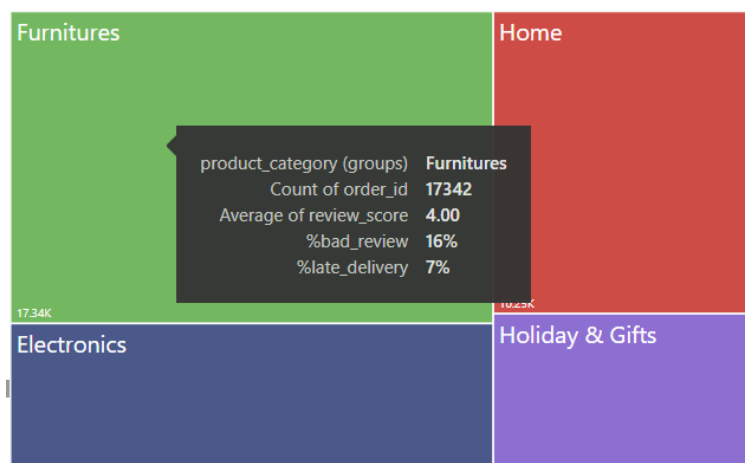


Figure 21 Product Category Distribution

The tree map above displays relationship between bad reviews rate by product based on categories. The intention behind this chart is to figure out whether product category affect bad review score more than other aspect, but results signify that every category has an evenly distributed bad review rate. On other hand, it concludes that product dimensions such as weight, volume, freight value and so on does not affect bad ratings. The tree map allow analyst to drill down into individual product by product id to find out which product has high bad review score where signify that product quality are clearly unmet.



Upon hover, analyst can bring up a tooltip displaying the order count of furniture in between year 2017-2018, with the average review score of 4. It also displays the quality of furniture selling on e-commerce with bad review rate and late delivery rate.

6.1.4 Visual 4

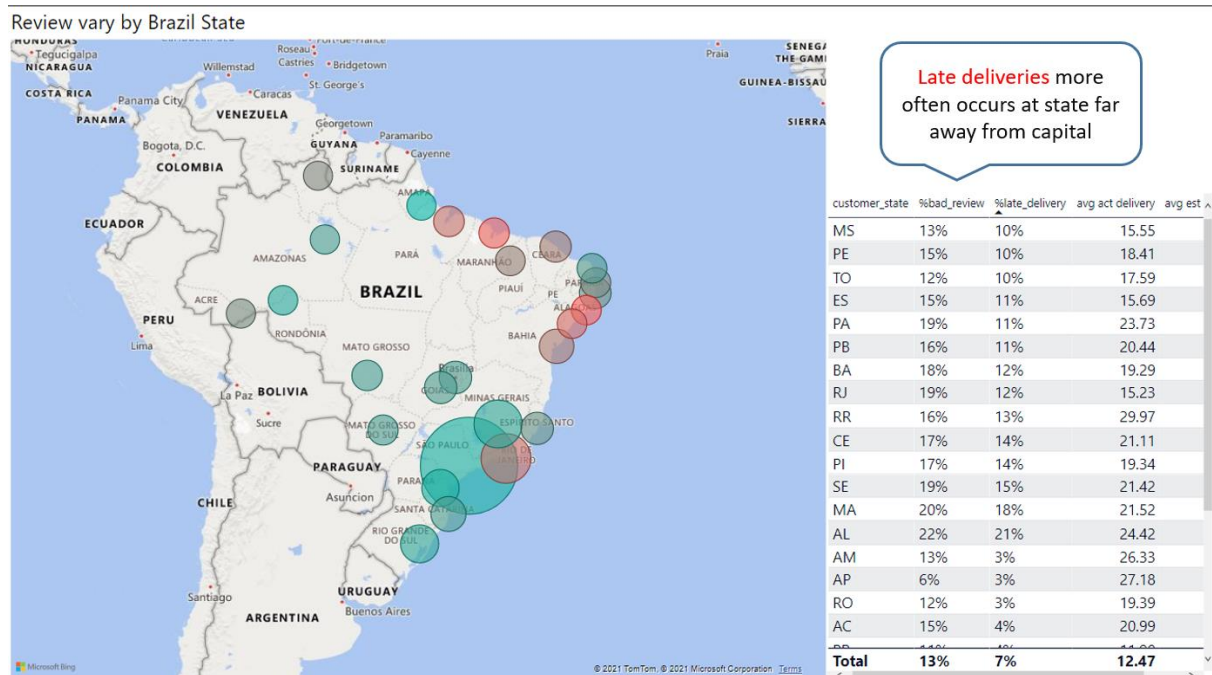
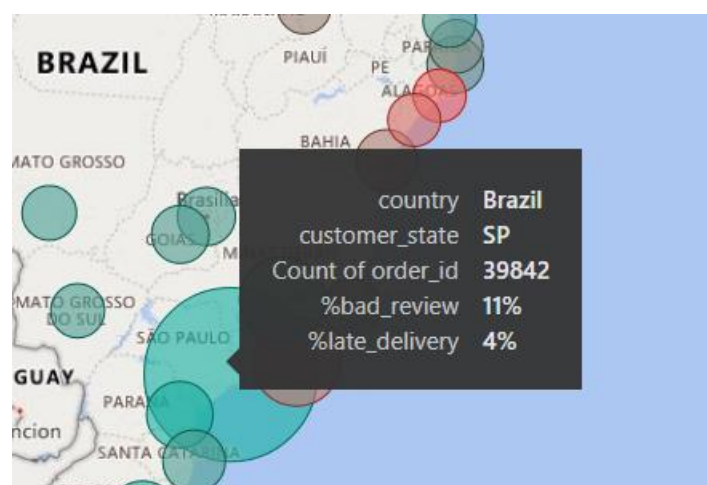


Figure 22 Review vary by Brazil State

As shown in the Geo Map designed from the Power Bi built in feature, it displays the relationship between bad reviews and states within Brazil, which includes 26 states and 1 federal district in Brazil. The references values to build the Geo Map is to display the bad reviews rate depending on the size of the order count per state. Therefore, state with red bubble indicates that the rating given by the customer are leaning towards negative impression. As observed from the bubble map above, it indirectly indicates the state that are far from the capital city tends to have a higher late delivery rate compared to other state.



Upon hover over a bubble, it will display a tooltip with the count of order within a state, in this case, which is 39,842 total order count in Sao Paulo, approximately 42% of the total order in all state. Accounting for 11% bad review rate and 4% late delivery rate.

6.1.5 Visual 5

Sellers Bad Review % vs Late Delivery %

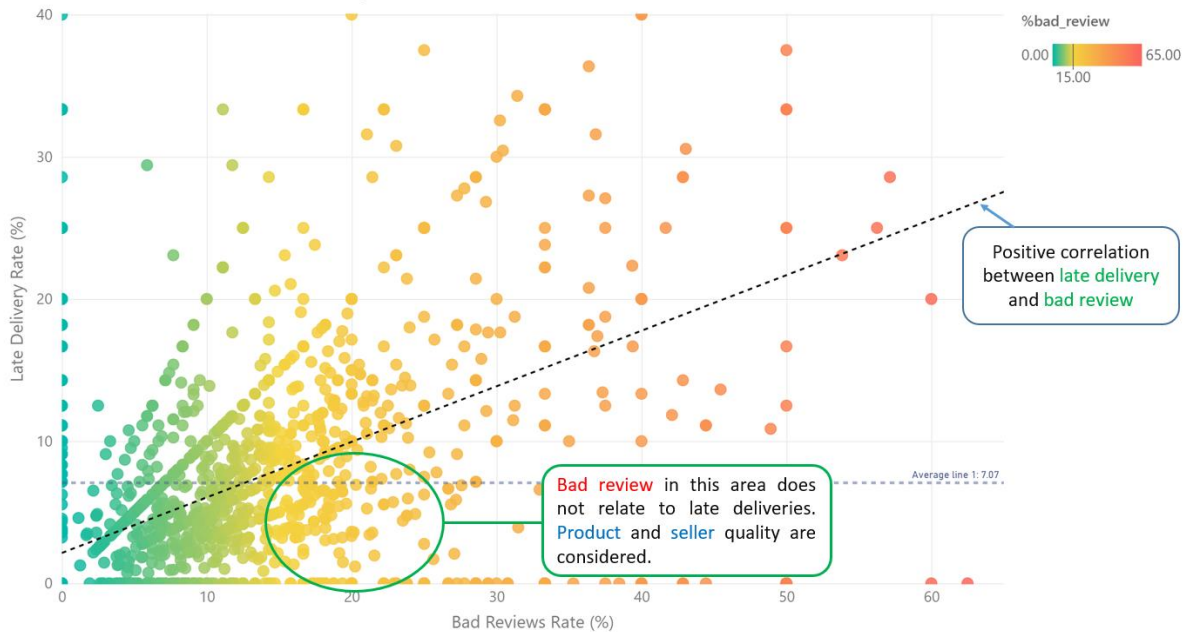


Figure 23 Scatterplot chart

The scatterplot chart signifies a positive linear trend between late deliveries and negative rating by seller id. From the observation of the chart, it depicts that most of the bad reviews are concentrated on the trend line where late delivery and seller/product quality affect reviews drastically. In the circled area, it does not explain that sellers with above expectations delivery performance are getting higher bad review rates, which concludes that other form of factor also impacts on the bad review rates.

6.2 OLAP Data Cube Operation

6.2.1 Display details of each product categories

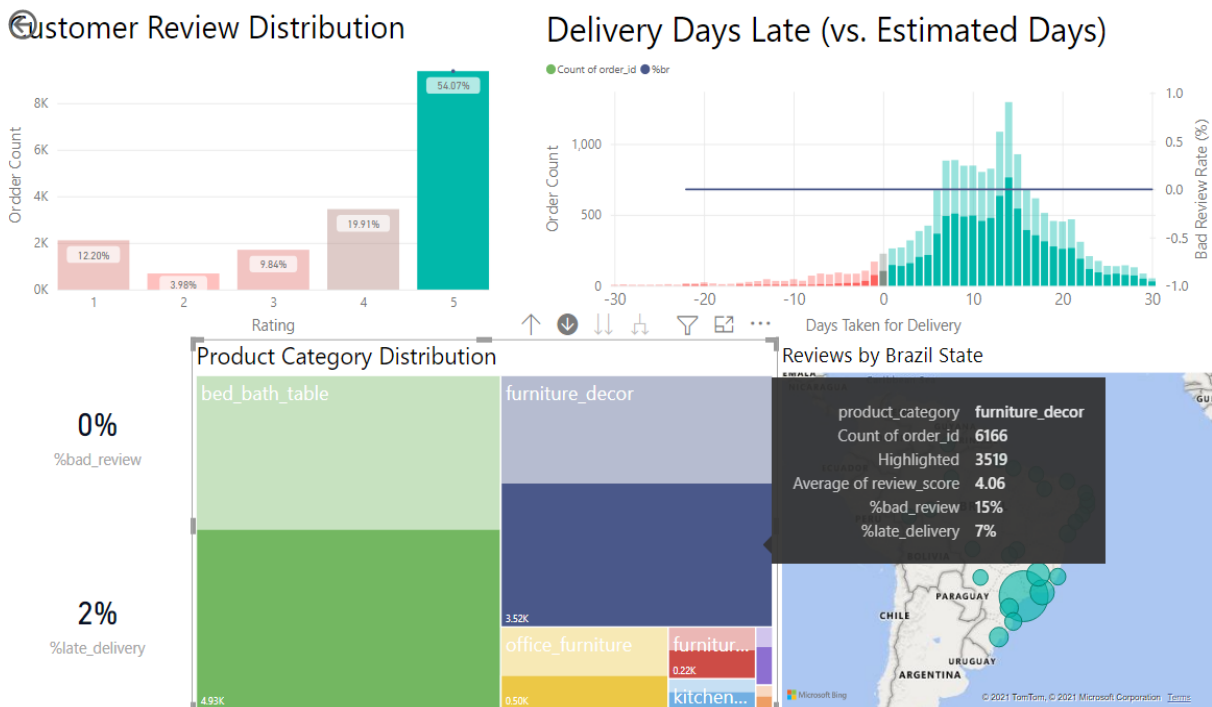


Figure 24 Drill down by product categories

From the scenario shown in Figure above, user is allowed to drill down into product category group to have a better insight on each individual category belongs to the same parent group. For instance, if the user attempt to view details such as order count of furniture décor within the year 2017 and 2018, instead of viewing overall details from the furniture parent group, user can drill down into the furniture group and view furniture décor details such as highlighted order count that has 5 points rating, average review score, bad review percentage and so on. In this case, furniture décor is having 3519 five-point rating in total order count of 6166. It depicts that furniture décor is a decent category item to purchase without high bad review rating. It also has a overall good delivery speed where late orders are only considered minority.

6.3 Outlier Analysis

```
def find_outliers(df, std_devs):
    """define a list to accumulate outliers
    """
    """calculates mean, standard deviation, lower/upper limits and outlier
    counts and % of total records in dataframe based on desired number
    of standard deviations.

    Parameters:
    df (pd.DataFrame): Name of Pandas dataframe
    std_devs (int): Number of standard deviations that define an outlier

    Returns: df_outliers (dict): Dictionary of outlier indices by column
    """
    df_outliers = {}
    print(f'outliers at greater than {std_devs} standard deviations')
    print('-----')
    for column in df:
        outliers = []
        df[column].astype(float)

        # Set upper and lower limit to 4 standard deviation
        series_std = np.std(df[column])
        series_mean = np.mean(df[column])
        outlier_cut_off = series_std * std_devs
        lower_limit = series_mean - outlier_cut_off
        upper_limit = series_mean + outlier_cut_off

        # Generate outliers
        for index, value in df[column].iteritems():
            if value > upper_limit or value < lower_limit:
                outliers.append({'index': index, 'value': value})

    return df_outliers
```

Figure 25 determine outlier source code

To develop a precise and accurate logistic regression model to depict the event occurrence, outliers must be eliminated as much as possible. This is because outliers increase the variability in the data, which the resulted model may be inaccurate. To find out outlier values from each predictor values, a Python function is written to calculate the mean, standard deviation, lower/upper boundary, and outlier counts of each predictor based on desired number of standard deviations, which is 6 in this analysis. The output is displayed below where every outlier's row shall be filtered before conducting the predictive model.

```
Outliers at greater than 6 standard deviations
-----
price | mean 136.9 | std 197.1 | ll=-1045.7 | ul=1319.4 | outliers=230 | 46648 outlier % 0.49%
freight_value | mean 22.8 | std 17.3 | ll=-80.7 | ul=126.3 | outliers=197 | 46648 outlier % 0.42%
product_weight | mean 2192.9 | std 3739.7 | ll=-20245.4 | ul=24631.2 | outliers=183 | 46648 outlier % 0.39%
product_volume | mean 15248.6 | std 22669.5 | ll=-120768.6 | ul=151265.9 | outliers=167 | 46648 outlier % 0.36%
review_score | mean 4.1 | std 1.3 | ll=-3.7 | ul=12.0 | outliers=0 | 46648 outlier % 0.00%
seller_bad_pct | mean 0.1 | std 0.1 | ll=-0.4 | ul=0.6 | outliers=82 | 46648 outlier % 0.18%
product_bad_pct | mean 0.1 | std 0.2 | ll=-0.9 | ul=1.1 | outliers=0 | 46648 outlier % 0.00%
product_order_cnt | mean 32.3 | std 69.1 | ll=-382.1 | ul=446.6 | outliers=0 | 46648 outlier % 0.00%
order_item_cnt | mean 1.0 | std 0.0 | ll=1.0 | ul=1.0 | outliers=0 | 46648 outlier % 0.00%
order_seller_cnt | mean 1.0 | std 0.0 | ll=1.0 | ul=1.0 | outliers=0 | 46648 outlier % 0.00%
order_value | mean 136.9 | std 197.1 | ll=-1045.7 | ul=1319.4 | outliers=230 | 46648 outlier % 0.49%
order_freight | mean 22.8 | std 17.3 | ll=-80.7 | ul=126.3 | outliers=197 | 46648 outlier % 0.42%
order_avg_price | mean 136.9 | std 197.1 | ll=-1045.7 | ul=1319.4 | outliers=230 | 46648 outlier % 0.49%
order_freight_ratio | mean 0.3 | std 0.3 | ll=-1.6 | ul=2.3 | outliers=117 | 46648 outlier % 0.25%
est_delivery_days | mean 26.5 | std 8.0 | ll=-21.6 | ul=74.5 | outliers=23 | 46648 outlier % 0.05%
act_delivery_days | mean 14.0 | std 9.0 | ll=-39.8 | ul=67.7 | outliers=47 | 46648 outlier % 0.10%
days_variance | mean 12.5 | std 9.9 | ll=-47.2 | ul=72.2 | outliers=22 | 46648 outlier % 0.05%
delivery_km | mean 723.4 | std 592.5 | ll=-2831.3 | ul=4278.2 | outliers=3 | 46648 outlier % 0.01%
```

Figure 26 Outlier Analysis

7.0 Data Modelling

7.1 Regression analysis

Regression analysis is a way of predicting future events between a dependant variable and one or more independent variables as known as predictor. The priority of regression analysis is forecasting, time series modelling and finding the cause-and-effect relationship between variables. (Appier, 2021) It is a effective way of determining the utmost important factor, least important factor and also correlation between factors. There are several linear regressions used for machine learning, including linear regression, logistic regression, ridge regression, lasso regression, and polynomial regression. In this report, logistic regression is preferred.

Linear Regression	Logistic Regression
continuous output	constant output
estimated using Ordinary Least Squares (OLS)	estimated using Maximum Likelihood Estimation (MLE)
OLS is a distance-minimizing approximation method	MLE is a "likelihood" maximization method

7.2 Pearson Correlation

Correlation coefficients are calculations that used to measure how strong a relationship is between two variables, commonly used in linear regressions. Correlation coefficient formulas always returns a value between -1 and 1 where 1 indicates a strong positive relationship whereas -1 indicates a strong negative relationship. Meanwhile 0 indicates no relationship at all. To find out correlations within predictive variables in the selected dataset, implementation of Pearson correlations against the dependant variable in the report, which is bad review by plotting the relations using Python libraries are necessary.

```
# create column correlation matrix
train_comb = pd.concat([X_train, y_train], axis=1)
corr = train_comb.corr()

plt.figure(figsize=(16, 8))

heatmap = sns.heatmap(corr, vmin=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Correlation between bad review and predictive dimensions', fontdict={'fontsize':16}, pad=12);
```

Figure 27 Source code for Pearson correlation

Parameters includes:-

- ✓ **X_train**: a list of predictive variables that are trained using train-test split evaluation.
- ✓ **y_train**: bad review, a target variable with binary values that are trained using train-test split evaluation.

By using the seaborn library and formula as shown above, the output of the correlation heatmap is created as shown below.

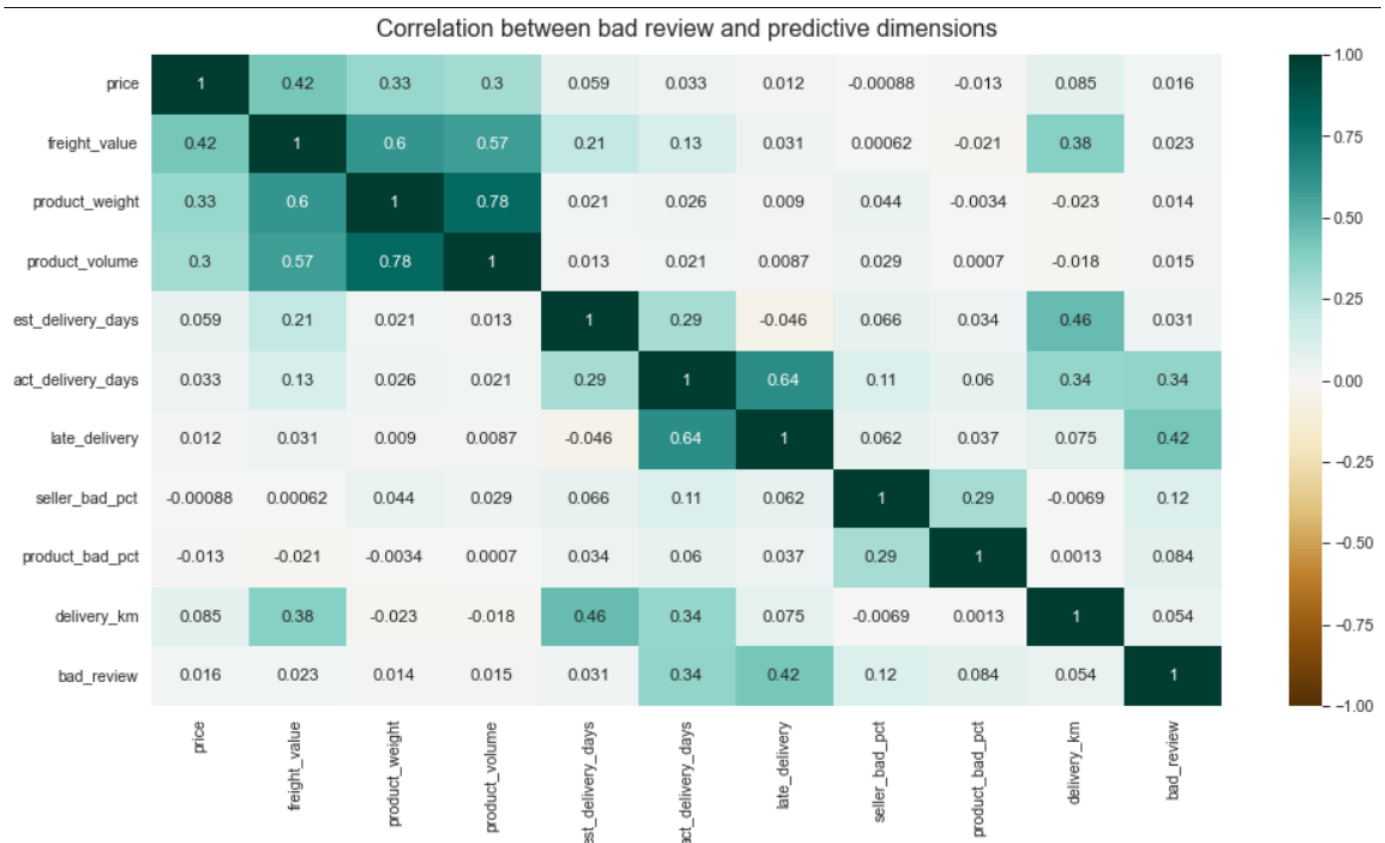


Figure 28 Pearson Correlation heatmap

As interpreted in the heatmap, the target variable, bad_review has the highest correlation with late deliveries and actual delivery days with the score of 0.42 and 0.34 respectively. There are several positive correlation pairs such as products and sellers overall bad review rates with the score of 0.084 and 0.12 respectively. Several minors' correlation includes freight value, order values, and delivery distance in km which barely affect the bad reviews. Hence, the independent variables with higher relationship on bad reviews will be filtered to carry out logistic regression modelling.

7.3 Train-test split

Train-test split is a technique for evaluating the performance of a machine learning. It is widely used in classification or regression models. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset. (Brownlee, 2020) The main goal of the technique is to measure the effectiveness of the machine learning model on new data: data not used to train the model.

7.3.1 Configuration of Train-Test Split in Python

Train Test Split

```
col1 = ['product_bad_pct', 'est_delivery_days', 'act_delivery_days', 'late_delivery', 'product_weight', 'product_volume', 'delivery_km',  
'seller_bad_pct', 'price']  
  
X_train, X_test, y_train, y_test = train_test_split(merged[col1], merged['bad_review'], test_size=0.20, random_state=0)
```

Figure 29 Train-test split source code

Parameters included:-

- ✓ **X_train**: a list of predictive variables that are trained using train-test split evaluation.
- ✓ **X_test**: a list of predictive variables that are tested using train-test split evaluation.
- ✓ **y_train**: target variable with binary values that are trained using train-test split evaluation.
- ✓ **y_test**: target variable with binary values that are tested using train-test split evaluation.
- ✓ **test_size**: 20%
- ✓ **train_size**: 80%

7.4 Logistic Regression

Unlike simple linear regression, logistic regression is primary method for identifying the binary classification issues, which includes values such as 1 and 0. Logistic regression can be used for various classification issues such as diabetes prediction, user click prediction and so on. Logistic regression describes and estimates the relationship between one dependant binary variable and independent variables. (Navlani, 2019)

Linear regression equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y: Dependant variable

β_0 : y-intercept

β_1 : slope

X: Independent variable

ϵ : Random error term

Sigmoid Function:

$$p = 1 / (1 + e^{-y})$$

Apply sigmoid function on linear regression:

$$p = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)})$$

Characteristics of logistic regression includes the dependant variable in the model follows Bernoulli distribution. Other than that, estimation of data is done through maximum likelihood. It also does not have R Square, model fitness that is calculated through Concordance.

Types of logistic regression

Type	Explanation
Binary Logistic Regression	The target variable has only two possible outcomes
Multinomial Logistic Regression	The target variable has three or more nominal categories

Ordinal Logistic Regression	the target variable has three or more ordinal categories
-----------------------------	--

To develop a Logistic Regression Model, several Python Libraries such as NumPy, Seaborn, and sklearn are required to be install and import to the Python file. Once all the required libraries are imported, datasets that are needed for the prediction model are required to be imported into the file using the `pandas.read.csv(file)` method. After importing the merged datasets from data visualization, several redundant and unused columns are suggested to filter using the `pandas.drop(col)` method for training performance purpose. Other than that, variables with outliers identified at section 6 earlier should be filtered before the implementation of data training, using the code `df.dropna(how='condition')`. Other than that, several additional predictors can be added by performing feature engineering process where new feature are introduced by adding measures into the merged dataset. For instance, the delivery distance between the seller and customer living state, using both party geolocation, it is possible to calculate the approximate distance by using the haversine function as shown in figure below.

```
# function to calculate haversine distance between lat-long coordinates
def haversine(coord1, coord2):
    R = 6372800 # Earth radius in meters
    lat1, lon1 = coord1
    lat2, lon2 = coord2

    phi1, phi2 = math.radians(lat1), math.radians(lat2)
    dphi      = math.radians(lat2 - lat1)
    dlambd    = math.radians(lon2 - lon1)

    a = math.sin(dphi/2)**2 + \
        math.cos(phi1)*math.cos(phi2)*math.sin(dlambd/2)**2

    return 2*R*math.atan2(math.sqrt(a), math.sqrt(1 - a))
```

Figure 30 Haversine function

```
# create Logistic Regression classifier object
model_lr = LogisticRegression()

X_test = X_test.fillna(0)
X_train = X_train.fillna(0)
y_train = y_train.fillna(0)
y_test = y_test.fillna(0)

model_lr.fit(X_train.values, y_train)
y_pred = model_lr.predict(X_test)
```

Figure 31 Creating logistic regression classifier from sklearn library

To create a logistic regression instance, independent variables must be identified beforehand, by creating an array of predictors as shown in figure above. From the Pearson Correlation Heatmap analysed on section 7.2, variables such as late delivery and actual delivery days must be included into the training model as it shows high correlation between itself and bad review rates. By using the `logisticRegression()` function, a logistic regression classifier object is created. Predictor used to train the model includes actual delivery days, estimated delivery days, price, late deliveries, product weight, product volume, delivery distance and so on. To fit the model with data, two parameters are required to insert into the method of classifier, which is bad review training data on x-axis, and predictors training data on y-axis.

After defining the grid parameters, using the train-test split model that are broken into 80% and 20% on model training and model testing respectively. Then, the model is then used to perform test set using `predict()` method.

7.5 Confusion Matrix

Confusion matrix is a set of summary data that is used to determine the effectiveness of a classification model. Classification accuracy is the ratio of correct predictions to total predictions made, which can be visualize by using the formula below.

$$\text{accuracy} = \text{correct prediction} / \text{total predictions}$$

By visualizing the confusion matrix, it shall be represented in a form of the array object using Python. The dimension of the matrix is 2 by 2 because logistic regression is a binary classification, classes including 0 and 1. There are 4 events in a confusion matrix as shown in the table below.

“True positive”	correctly predicted event values
“False positive”	incorrectly predicted event values
“True negative”	correctly predicted no-event values
“False negative”	incorrectly predicted no-event values

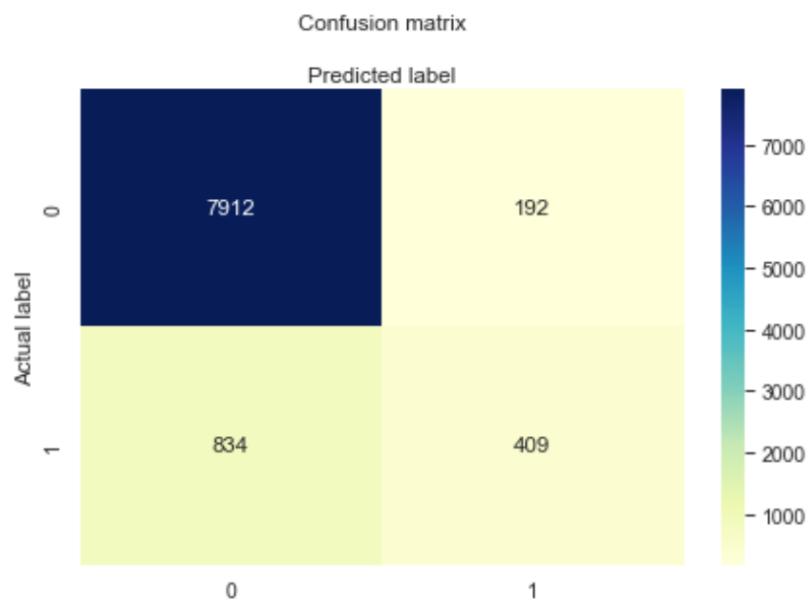


Figure 32 Confusion Matrix generated from Python

There are only two classes 0 and 1. Diagonal values represent accurate predictions, while non-diagonal elements are inaccurate predictions. In the output, 7912 and 409 are actual predictions, and 834 and 192 are incorrect predictions.

Other than that, confusion matrix also gives a summary regarding model evaluation metrics such as accuracy, precision, and recall. By using metrics method provided from sklearn Python library, it can perform calculation to acquire the evaluation from the model.



Figure 33 Confusion Matrix

By calling the methods that comes within the metrics function, including `accuracy_score`, `precision_score`, and `recall_score` using the predictor test values and prediction values, it will generate the output as shown below.

```

Accuracy: 0.8902321600513534
Precision: 0.6805324459234608
Recall: 0.32904263877715206

```

Figure 34 Model evaluation metrics

From the figure above, the classification rate is above 89%, which is considered accurate and eligible to perform coefficient test regarding the analysis. In the classification, it is measured that the precision is approximately 68%, which means there is 68% chance that if the delivery arrives later than the estimated date, the customer will submit a bad review on the order. Recall metric in the model does not looks accurate as it only has 32.9%, which means if the customer submits a bad review, Logistic Regression is able to identify it 32.9% times.

7.6 ROC Curve

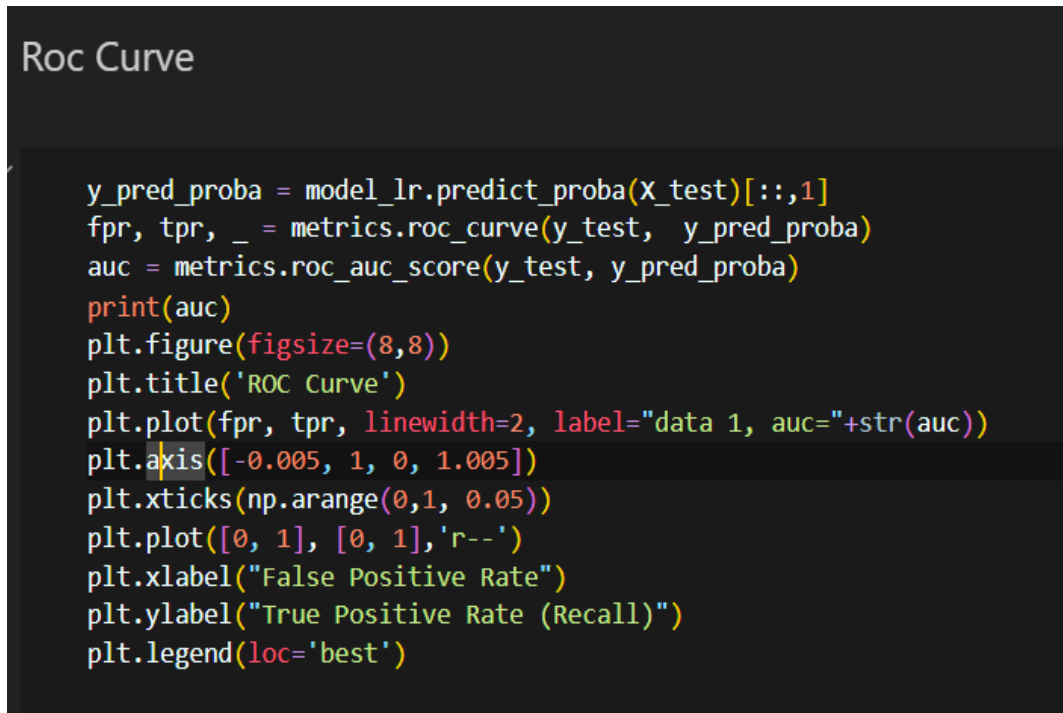


Figure 35 ROC Curve

ROC as known as Receiver Operating Characteristic curves; its main purpose is to visualize the comparison of classification models. The area under the ROC curve is a measure of the accuracy of the model. The benefit of using ROC curve is that it can show trade-off between the true positive rate and the false positive rate. For instance, true positive show instances that are correctly classified as positive, meanwhile false positive show instances that are incorrectly classified as negative.

By using the source code from python as shown above, it will produce an output that displays the AUC score, which used to identify the weightage of the classifier implemented. Other than that, it also plots a line graph which predicts the accuracy of the model.

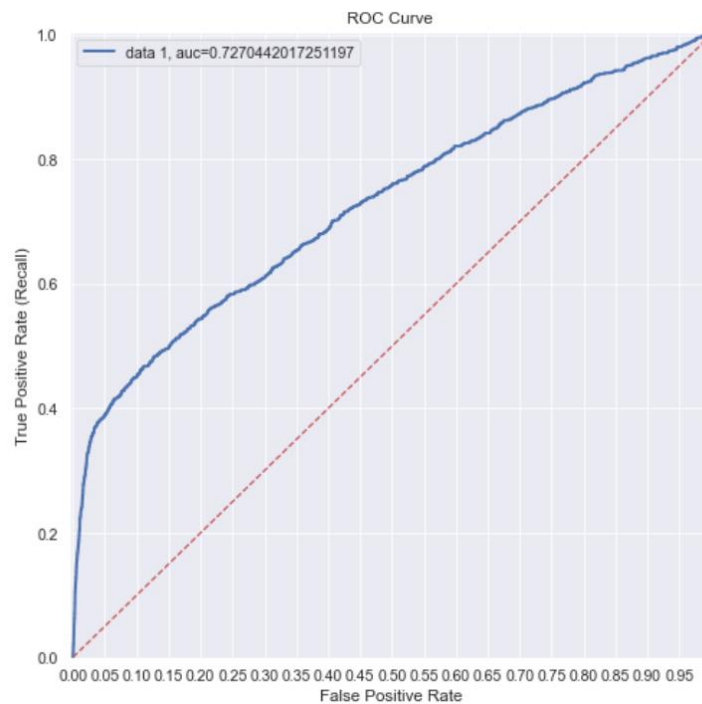


Figure 36 ROC Curve graph

AUC score for the case is 0.72. AUC score 1 represents perfect classifier, and 0.5 represents a worthless classifier, which is considered applicable in this situation. The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model. A model with perfect accuracy will have an area of 1.0, that means the value of TP rate should be one for all FP rates.

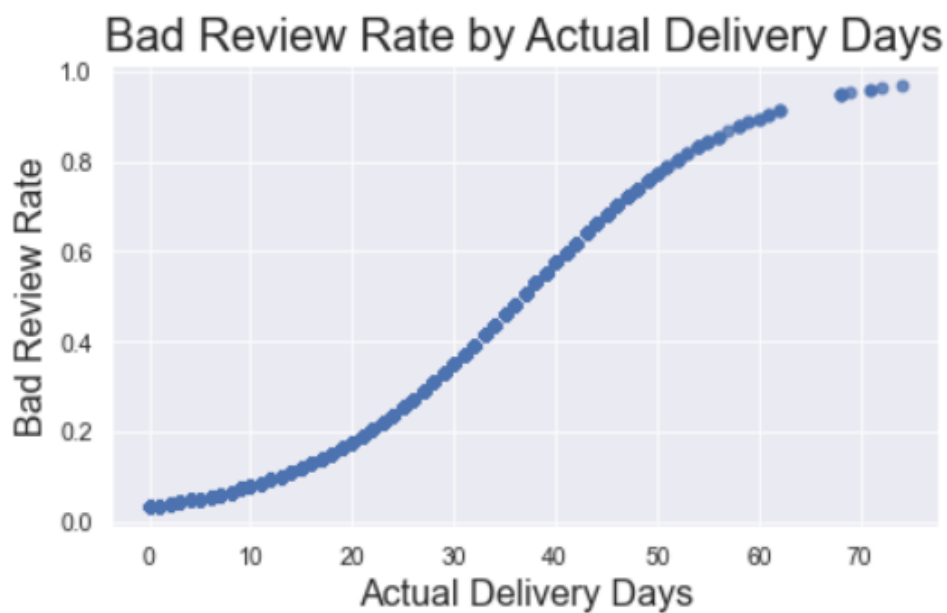


Figure 37 Logistic Regression on relationship between bad review rate and actual delivery days

7.7 Model Coefficient by features

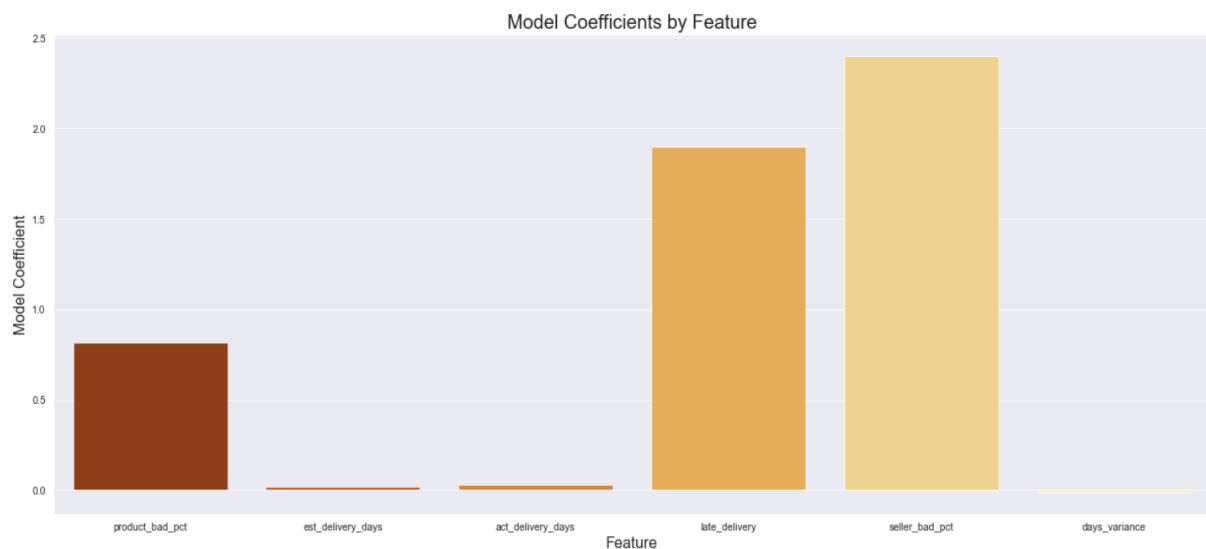


Figure 38 Model Coefficient by Features

In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one. (The Trustees of Princeton University, 2007)

The most important features in the Logistic Regression model included:-

1. Overall bad seller service rate
2. Late deliveries
3. Overall bad product quality rate

feature	coef
product_bad_pct	0.82
est_delivery_days	0.02
act_delivery_days	0.03
late_delivery	1.90
seller_bad_pct	2.40
days_variance	-0.01

Figure 39 Coefficient on every feature

In conclusion, bad reviews are most sensitive to overall seller service quality, late deliveries, and product qualities. Business owner should focus on improving the features that are stated in the analysis result, such as paying more attention and budget to fix the root causes for bad delivery performance. Another aspect of the bad reviews derived from seller's honesty and product quality. Business owner should come out with counteract measures such as penalties

to dishonest seller and provide full refund terms and condition to product that does not meet expectations.

8.0 Data security and privacy

During the pandemic, e-commerce platforms have a significant siege in user traffic. The E-commerce public dataset in Brazil has been chosen as the main topic to be conducted on analysis. Data security and privacy are the main concerns when it comes to platforms that require a user to fill in personal details. Hence, the article that is used as a reference is titled “Brazilian marketplace integrator Hari express exposed 1.75 billion records”. Within the article, it is mentioned that the company responsible for the data leak are refraining from taking steps to secure the data, exposing personal data and available online for the public for more than a month. Since the data was leaked to the public for some time, it is hard for the company to take action to retrieve the data and fix these issues. The consequence faced by the company is the loss of trust of the stakeholders (Waqas, 2021).

Regarding the scope analysed in the dataset, sensitive data such as customer billing information and address should be disclosed and generalized before being available online to the public. The responsible company should take preventive measurements of users’ personal information by including authentication and encryption service in data protection. Moreover, the company should also notify the customers on what data is gathered from them and what the data is used for. Failure to compromise on respecting users’ information will harm the company reputation, public fear causing distress and facing a lawsuit.

9.0 Conclusion and Recommendation

In a nutshell, Brazilian e-commerce has a decent quality when it comes to online purchase. With a promising 83% above average rating provided by each order that are delivered. Several improvements are inevitable as demands are drastically increasing over the years. However, there are also several aspects of the analysis that does not meet expectations. For instance, the sampled data which consists of approximately 100,000 orders that are associated with ratings are suitable for building a logistic regression but not representative of the population. Furthermore, sentimental analysis is not able to carry out since roughly 10% to 15% of customer fill up the review messages. Messages with comments are also not translated to English, which makes the test analysis even challenging and generalize against the full population of Brazil orders.

Although logistic regression model is able to provide data analytic with 89% accuracy and a high 32.9% recall rate, which means bad reviews can be predicted correctly 32.9% of the time. Due to knowledge restriction, several interesting classification model such as random forest model, AdaBoost model, SGD Classifier Model, and Gaussian Naïve Bayes model and so on. Certain classifier may suit this current analysis better with higher precision and recall rate. Furthermore, the datasets acquired may include more detailed information for more precise measurements such as customer services, customer returns rate, and demographics. Overall performance of the data analytic implementation is satisfied.

10.0 Personal Reflection Report

Personal experience with the implementation of the whole report gave me opportunity to have a deep immersing into the world of data analytics. Following the guidelines of Crisp-DM model, I can have a proper procedure in producing the analysis correctly. Other than that, it allows me to gain more knowledge in the utilization of Python programming language and Microsoft Power Bi ability to perform visualization of collected data neatly. Furthermore, it taught me to understand the steps to implement a predictive model, although the learning process is difficult but the achievement that I had after completing the logistic regression is immeasurable.

11.0 Workload Matrix

Name	Task
Bryan Hor Jin Hao	Scope: Sales and shipping fee Model: Predictive model (Linear Regression)
Chia Wen Xuen	Scope: Popularity of Product Sold Model: Statistical model (Time-Series Forecast)
Ngiam Jie-Hao	Scope: Payment Type Model: Predictive model (Decision Tree)
Tin Eugene	Scope: Bad reviews Model: Predictive model (Logistic Regression)

12.0 References

Appier, 2021. *5 Types of Regression Analysis And When To Use Them*. [Online]

Available at: <https://www.appier.com/blog/5-types-of-regression-analysis-and-when-to-use-them/>

[Accessed 9 9 2021].

Brownlee, J., 2020 . *Train-Test Split for Evaluating Machine Learning Algorithms*. [Online]

Available at: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>

[Accessed 10 10 2021].

Navlani, A., 2019. *Understanding Logistic Regression in Python*. [Online]

Available at: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

[Accessed 11 11 2021].

The Trustees of Princeton University, 2007. *Interpreting Regression Output*. [Online]

Available at: https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

[Accessed 10 10 2021].

Wikipeddia, 2021. *Process diagram showing the relationship between the different phases of CRISP-DM*. [Online]

Available at: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

[Accessed 10 August 2021].