Olist Business Analysis

Li PinZhao, Liu YanYu, Dong GengShang, Vincent William

May 2024

Abstract

This project conducts a comprehensive analysis of the Olist dataset, a Brazilian e-commerce platform, to gain insights into various aspects of its operations. The dataset encompasses nearly 100,000 orders from 2016 to 2018, detailing order status, prices, payment and freight performance, customer locations, product attributes, and customer reviews.

The analysis begins with preliminary data examination, identifying trends, patterns, and anomalies. This includes:Monthly trend on the total revenue and total number of orders monthly, Geographical distribution of buyers and sellers, Merging with Brazil Population Data.

A detailed time series analysis predicts future freight values, fitting the closest ARIMA models and transformations to ensure stationarity and address autocorrelation (or residuals) issues.

Product analysis explores the relationship between price and freight value, sales volumes, and customer reviews.

Sentiment analysis of reviews uses both supervised and unsupervised methods, with tools like AFINN and TextBlob.

Association rule mining and item-based collaborative filtering are applied to understand purchasing relationships and to generate product recommendations, respectively. These techniques help optimize inventory management and enhance marketing strategies.

Key findings indicate significant monthly revenue growth, a positive correlation between product price and freight costs, and valuable insights from sentiment analysis of customer reviews. The project also suggests logistic optimizations and market strategy adjustments based on geographical and demographic insights.

This analysis provides a robust framework for e-commerce platforms to understand consumer behavior, optimize logistics, and enhance product offerings, ultimately contributing to better business decisions and customer satisfaction.

1 Background and Problem Setup

Olist is a Brazilian startup that operates in the e-commerce segment , mainly through the marketplace. It is well spread within the country. This project is a detailed analysis on the comprehensive Olist data. The original Olist 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. The connection chain of the datasets is as below:

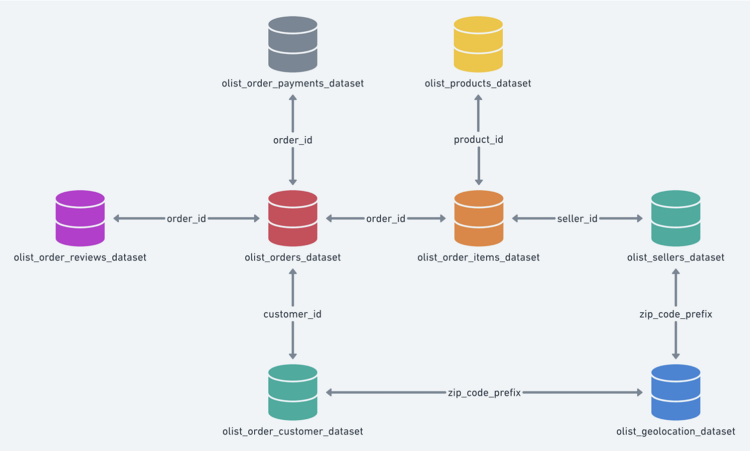


Figure 1: Data Structure

Multiple analysis tasks are carried out on the dataset, ranging from descriptive analysis to forecasting and predictive analysis.

Preliminary data analysis:

It visualizes and summarizes the original and the combined datasets, to find trends, patterns or faults. This analysis gives a holistic view of the dataset.

Freight value prediction

The freight value is the shipping value associated with each order.This section is created in order to conduct a meaningful time series analysis.

Product analysis

This section is devoted to an examination of the product's properties with the objective of providing guidance to the user in the selection of the product.

Association rule mining

It explores the application of association rule mining to analyze transaction information from the dataset. Association rule mining is a popular data mining technique used to discover interesting relationships, such as frequent patterns, associations, or correlations among large sets of data items. This analysis helps e-commerce platform understand the purchasing relationships between different products, thereby optimizing inventory management and promotional strategies.

Reviews sentiment analysis

Sentiment analysis is carried out on the reviews offered by customers. The section contains Supervised and Unsupervised methods for sentiment analysis, Reviews Sentiment Analysis. The supervised technique uses the rating provided as label for sentiment analysis, whereas the unsupervised technique lexicons- AFINN, TextBlob to perform the analysis.

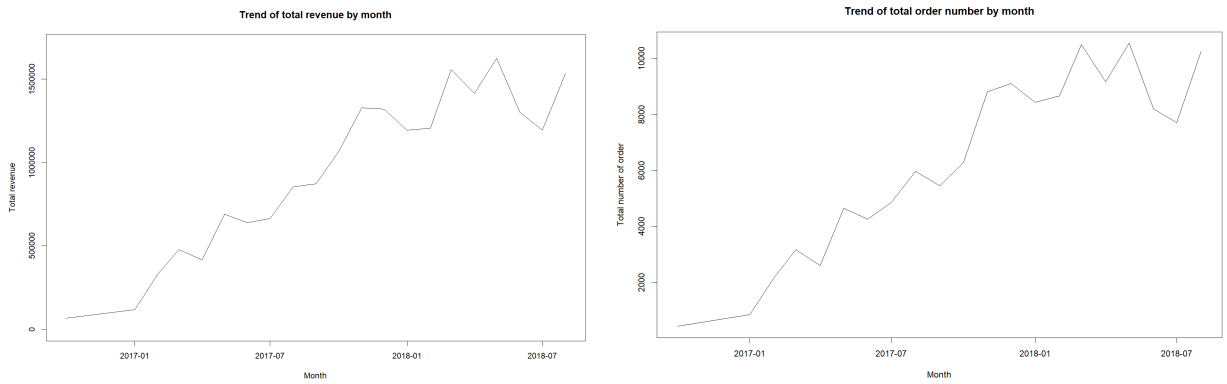
2 Preliminary Data Analysis

2.1 Monthly trend on the total revenue and total number of order monthly

Before commencing the monthly trend analysis, the rows with order status showing canceled and unavailable must be erased because:

- The company did not generate any revenue from those orders.
- The sales quantity is not counted.
- The freight value is ignored because it was not transported.

Then, we calculate the quantity of sales, revenue, and freight value based on every month. All monthly data where majority of its days are gone will not be considered in the monthly trend analysis.



The left chart shows that the total payment value increased as time goes on overall. It started from a minimum value of 67,845.25 in October, 2016. Then, the value reached its peak in May, 2018, with a total payment value of 1,625,161.

Not only does the right chart shows the same overall pattern, but it also strictly shares the same votality pattern. In fact, the lowest total number of order was also recorded at the same months as the The lowest total payment value recorded, with a total number of order of 434. Furthermore, its peak also happened at the same month time which is May, 2018, with a total number of order of 10553.

2.2 Geographical distribution of buyers and sellers

Take the orders in August 2018 as example, the connection between buyers and sellers is based on the geographical distribution of buyers and sellers.

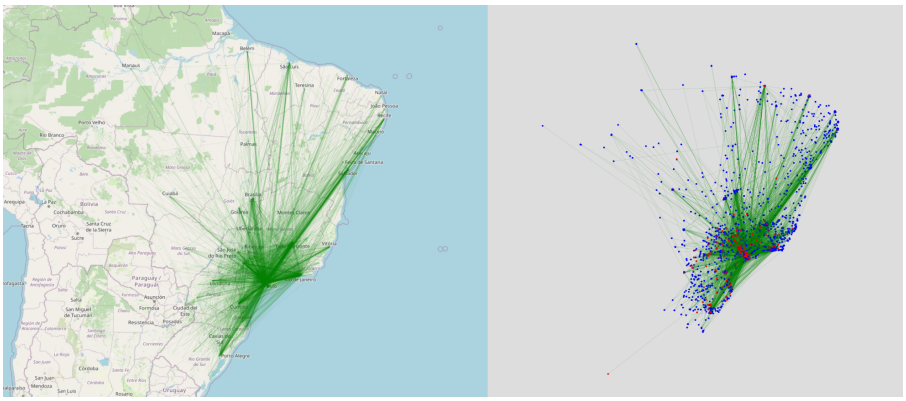


Figure 2: The location distribution of selles and consumers

2.2.1 Concentration of Flow

- The flow of orders is predominantly radiating out from São Paulo, highlighting its role as Brazil's largest commercial and industrial hub.
- The order flow extensively covers major Brazilian cities, especially densely populated and economically active eastern coastal cities like Rio de Janeiro, Salvador, and Recife.

2.2.2 Geographic Coverage

- The radiation lines cover a broad area from north to south and east to west, showing the wide reach and high accessibility of e-commerce services.
- The inclusion of cities in neighboring countries such as Buenos Aires and Asunción indicates cross-border operational capabilities.

2.2.3 Regional Differences

- Northern regions near the Amazon, such as Belém and Manaus, show fewer orders, likely due to geographical remoteness and infrastructure challenges.
- Southern and southeastern areas like Florianópolis and Belo Horizonte exhibit stronger order flows, correlating with their economic development and population density.

2.2.4 Logistics Network Optimization Suggestions

- Enhance logistics support in remote areas by forming partnerships or establishing regional distribution centers.
- Consider setting up additional warehouses or distribution centers in high-volume areas to reduce delivery times and costs.

2.2.5 Market Strategy Adjustments

- Adjust marketing strategies based on regional order volumes, promoting more activities in high-demand areas and adjusting strategies in low-demand areas.
- For international orders, tailor marketing and promotional strategies to accommodate cultural differences and consumption habits.

2.3 Merging with Brazil Population Data

Brazil Cities Population Data: The current population of Brazil is 215,653,799 as of Friday, July 22, 2022, based on Worldometer elaboration of the latest United Nations data.

`population_brazil_cities.csv`

A csv file containing population data of major cities in Brazil.
By merging and comparing to the number of sellers and customers in cities, we have obtained the following figures.

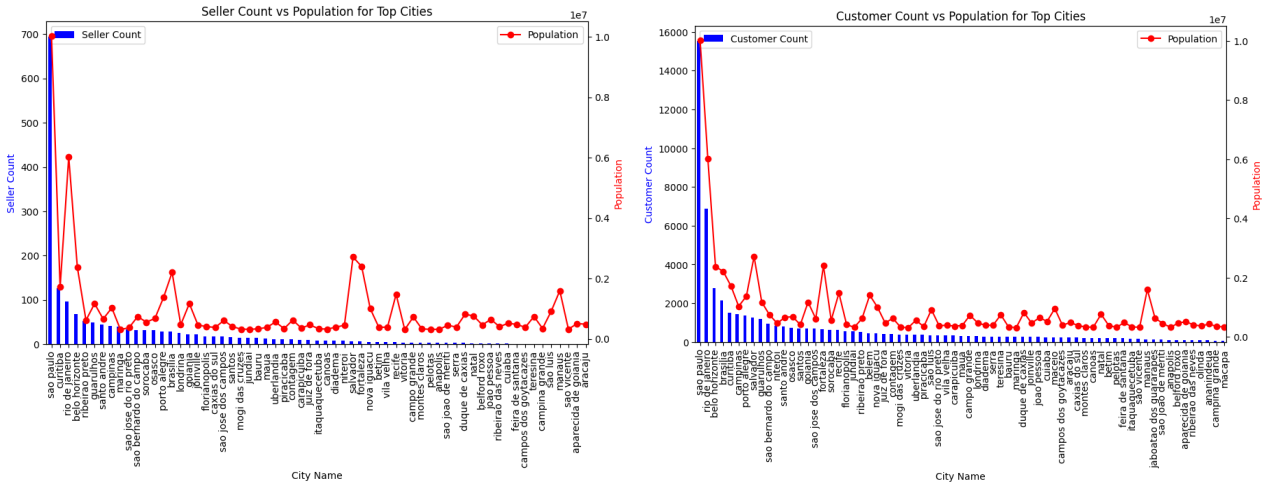


Figure 3: Seller and Customer VS Population

- Both charts highlight **Sao Paulo** as the city with the highest number of sellers and customers, indicating a strong market presence.
- The difference between the seller and customer counts in **Rio de Janeiro** suggests that while the city has a high number of potential customers, it might have fewer sellers relative to the customer base.

- Smaller cities with high penetration rates, such as **Campinas** and **Porto Alegre**, show significant engagement in both seller and customer metrics despite their smaller populations.
- There is a noticeable disparity in some cities where population size does not correlate directly with the number of sellers or customers, suggesting other factors at play such as economic activity or market reach.

So the concept of **Penetration Rate** was introduced: $\frac{\text{customers}}{\text{population}}$ and $\frac{\text{sellers}}{\text{population}}$.

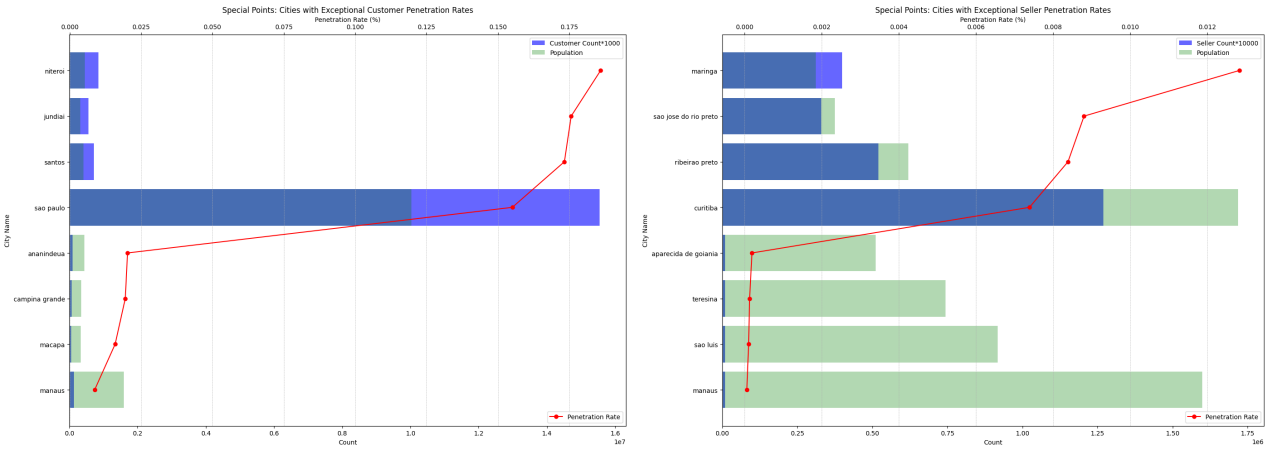


Figure 4: Special Points: Cities with Exceptional Customer and Seller Penetration Rates

- **Sao Paulo** appears prominently in the customer penetration chart but not in the seller penetration chart, indicating a higher customer base compared to sellers.
- **Curitiba** shows significant presence in both charts, highlighting its importance in both customer and seller markets.
- Cities like **Manaus** and **Macapa** have high customer penetration rates but do not appear in the seller penetration chart, suggesting a customer-centric market.
- Conversely, cities like **Aparecida de Goiania** and **Teresina** have high seller penetration rates but are not prominent in the customer penetration chart, indicating a seller-centric market.

3 Freight Value Prediction

3.1 Data Filtering

Three months period which are September 2016, December 2016 and September 2016 are discarded since majority of those days in each month is missing. Consequently, they have a neglectible sum freight values which are obviously outliers. The x-axis of the forecast graphs shows the timestep. Each time step represent the order of month. Our team decided to predict the freight value for the next 10 years (or 120 months).

3.2 Time Series Background Information

In order to conduct a meaningful time series analysis, it is imperative for us to know whether the model we are using is stationary and has no lag autocorrelation. A stcohastic process $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ is stationary if the following conditions hold:

- The expected value of time series is constant for all time t. Equivalently, the mean value of time series does not depend on t.

$$E(X_t) = \mu \quad \forall\, t$$

- The covariance of X_t and X_{t+h} is independent of t.

$$Cov(X_{t+h}, X_t) = Cov(X_h, X_0) \quad \forall\, t, h$$

Augmented Dickey Fuller (ADF) and Ljung-Box Tests are used for testing stationary and the existence of autocorrelation respectively. The hypothesis test for ADF test is

$$H_0 : \text{The model is NOT stationary} \quad \text{v.s.}$$
$$H_1 : \text{The model is stationary}$$

While the hypothesis test of Ljung-Box test is

$$H_0 : \rho(t, s) = 0 \quad \text{v.s.}$$
$$H_1 : \rho(t, s) \neq 0$$

where $\rho(t,s)$ denotes The autocorrelation function (ACF) of X_t and X_t.

$$\rho(t,s) = \frac{Cov(X_t, X_s)}{\sqrt{Var(X_t)}\sqrt{Var(X_s)}}$$

and its test statistics is given by

$$Q = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_h^2}{n-h}\right)$$

where $\hat{r}_i$ denotes the residual ACF of lag i for $i = 1, 2, \cdots, h$. Under H_0, $Q \sim \chi^2_{h-1}$.

So, rejecting H_0 in ADF test and failing to reject H_0 in Box-Ljung test are the desired test results. Then, we can forecast the model that we want to observe.

3.3 Time Series Tests and Forecast on the Monthly Freight Value

The reported p-values of ADF test and Ljung-Box test are 0.8206 and 8.516×10^{-5}. With p-value more and less than $\alpha = 0.05$ for ADF and Ljung-Box tests respectively, the model is not stationary and has autocorrelation lag.

and forecast is given below

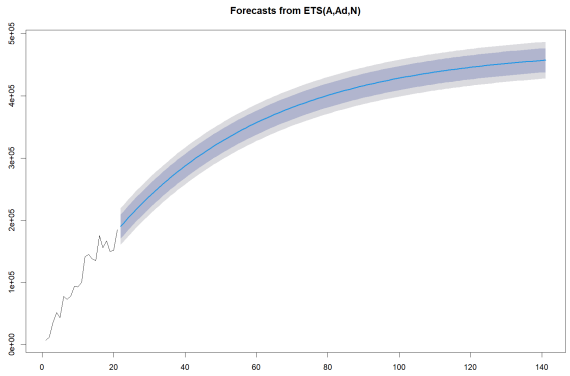


Figure 5: Forecast for monthly freight value

The forecast on the monthly freight value shows that the monthly freight value is expected to increase logarithmly and converges at one value.

To help understanding the reasons behind the non-stationary and autocorrelation existence properties of the model, we can fit the model into the closest ARIMA model. With auto.arima() function, the model is fit to ARIMA(0,1,1) with drift. The definition of ARIMA(0,1,1) , or known as Integrated MA model, with drift mathematically is

$$X_t - X_{t-1} = \delta + W_t + \theta W_{t-1}, \quad W_t \sim WN(0, \sigma^2)$$

To understand why IMA(1,1) with drift is not stationary, assume $t = 1$ is the initial time value and $X_1 = \delta + W_1$. Then, the model can be rewritten as

$$\begin{aligned}
X_t &= X_t - X_{t-1} + X_{t-1} - X_{t-2} + X_{t-2} - \cdots - X_1 + X_1 \\
&= X_1 + \sum_{j=2}^{t}(X_j - X_{j-1}) \\
&= \delta + W_1 + \sum_{j=2}^{t}(\delta + W_j + \theta W_{j-1}) \\
&= \delta + W_1 + (t-1)\delta + W_t + \theta W_{j-1} + W_{j-1} + \cdots + \theta W_1 \\
&= t\delta + W_t + (1+\theta)W_{t-1} + \cdots + (1+\theta)W_{-m} + (1+\theta)W_{m-1}
\end{aligned}$$

So that the mean value is

$$E(X_t) = t\delta$$

which is t-dependent.

3.4 Time Series Tests and Forecast on Differenced Monthly Freight Value

Since there is a linear non-stationary component in the mean value which is t model, differencing can fix the non-stationary issue as it will erase the non-stationary component in the expected value mathematically. Let

$$Y_t = \nabla X_t = X_t - X_{t-1}$$

Then,

$$E(Y_t) = E(X_t - X_{t-1}) = E(X_t) - E(X_{t-1}) = t\delta - (t-1)\delta = \delta$$

After we transformed the model through differencing, the reported p-value of ADF test and Ljung-Box test are 0.05266 and 0.0833. Although the differencing improves its stationary characteristic drastically, it still does not follow stationary time series if $\alpha = 0.05$. However, the autocorrelation issue has been solved. The forecast of differenced model is given below.

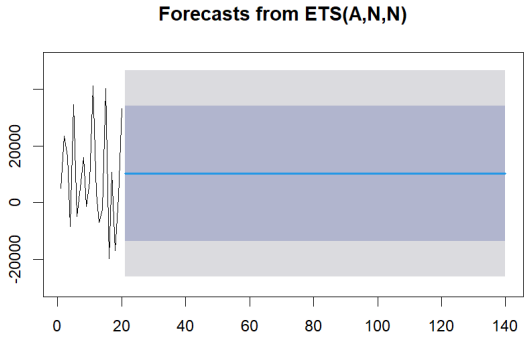


Figure 6: Forecast for differenced monthly freight value

The forecast shows that each month of the 10 years ahead of August 2018 shows that the fright value increases by a value of approximately 10,000.

3.5 Time Series Tests on the Log-transformed Freight Value

The log transformation is another well-known transformation technique as it can stabilize variance and even mean value. The reported p-value of ADF test and Ljung-Box test are 0.02425 and 0.0005701 respectively. The ADF test finally shows p-value less than $\alpha = 0.05$ which means that such transformation solve the stationary problem. However, the autocorrelation lag problem appears again. The forecast plot is not given as we are unable to extract meaningful information from it.

3.6 Time Series Tests and Forecast on the Differenced of Log-transformed Monthly Freight Value

Lastly, the differenced of log transformed monthly freight value is observed. Notice that we take logs first and then compute first differences - the order does matter. When a stochastic process has a relatively stable changes from one time period to the next, making such changes will make the model relatively stable and perhaps well-modeled by a stationary process. The reported p-value of ADF test and Ljung-Box test are less than 0.01 and 0.387 respectively. Finally, such model finally has the property of white noise which is stationary and no autocorrelation lag exists which is what we desire. Specifically, we have an extremly strong evidence to accept that the transformed model is stationary. The forecast of the differenced of log transformed monthly freight value is shown below.

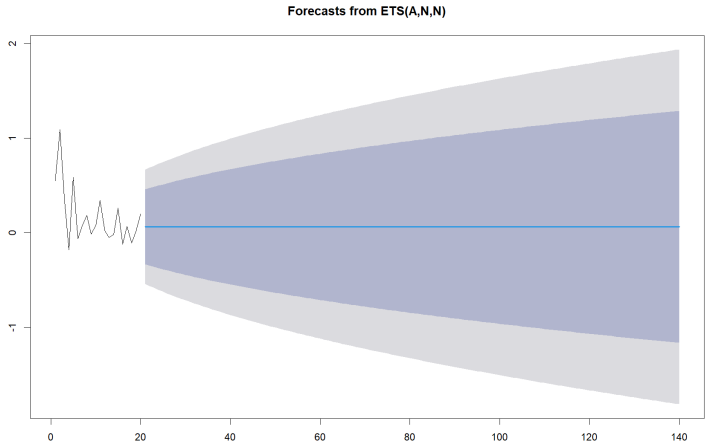


Figure 7: Forecast for the differenced log-transformed monthly freight value

The forecast shows that the differenced log-transformed monthly freight value has a constant mean of approximately 0.125. Since such transformation is popular for stochastic processes that tend to have relatively stable percentage changes from one time period to the next, it means that the forecast predicts that each month will yield an approximately 12.5% increase on the total monthly freight value.

4 Product Analysis

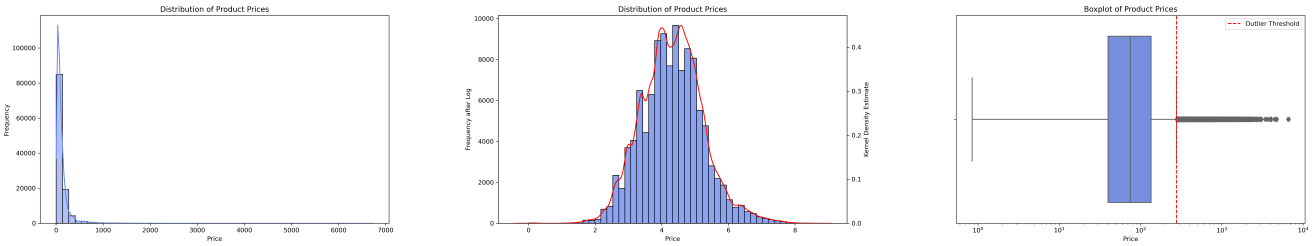
This section is dedicated to an investigation of the product's characteristics with the aim of identifying the user's preferences in the selection of the product.

4.1 Data Preprocessing

Due to the presence of a few missing values in the olist_products_dataset.csv file, it was deemed necessary to delete these rows directly. However, upon translating the product category using product_category_name_transaction.csv, it became evident that there were 13 missing items in the 'product_category_name' column. All of the missing values were replaced with the terms "computer gamer" and "kitchen stand and food prep".

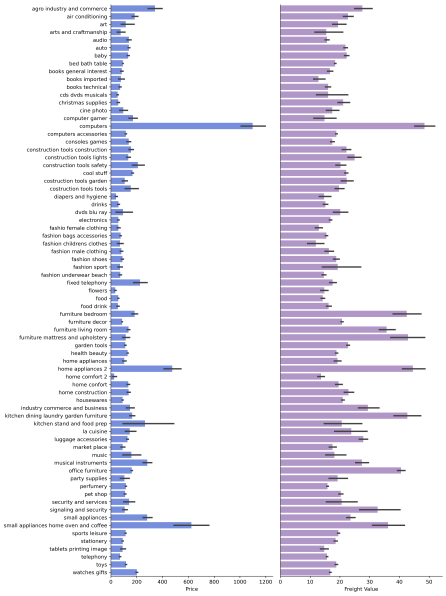
4.2 Price Distribution

The initial step is to generate a histogram that represents the distribution of price data for each product in question. The left-hand graph in the figure below represents the price distribution chart. Given that the distribution is skewed to the right, the logarithm of the price was calculated. Following the application of the logarithm, the price distribution chart is presented in the middle graph of the following figures.



25% Quartile: 39.9 Median: 74.99 75% Quartile: 134.9 Outlier Threshold: 277.4

4.3 Relationship between Price and Freight Value



The plot above illustrates a trend: there is a positive correlation between price and freight. This positive correlation indicates that as the price of a product increases, the associated freight cost also tends to increase.

Several factors may contribute to this observed trend.

- Firstly, higher-priced products often require more expensive transportation methods, such as expedited shipping or specialized handling, which can lead to higher freight costs. Additionally, the geographical distance between the point of origin and destination may affect freight costs, with longer distances typically resulting in higher shipping expenses.
- Moreover, the nature of the product itself can impact freight costs. For example, bulky or heavy items may incur higher shipping charges due to increased handling and transportation requirements.
- Understanding this positive correlation between price and freight is essential for businesses to accurately estimate total costs and set pricing strategies. By considering the relationship between product price and associated freight expenses, companies can optimize pricing decisions to maximize profitability while remaining competitive in the market.

4.4 Sales volume

4.4.1 Single order item quantity

As illustrated in the accompanying figure, a negative correlation exists between the price of a product and the volume of sales generated by a single order. This negative correlation indicates that as the price of a product increases, the volume of sales generated by a single order tends to decrease.

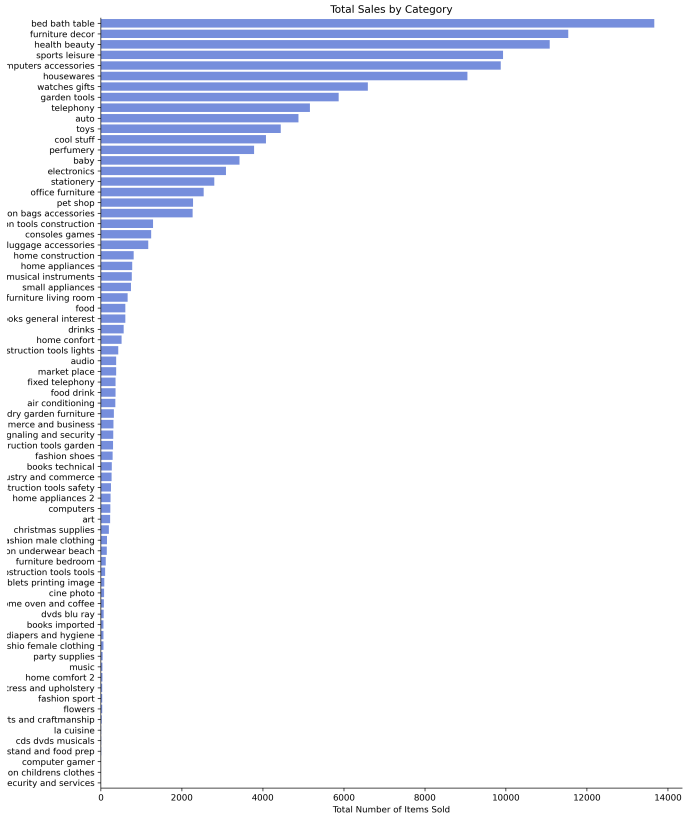
This relationship is typical in many markets and can be attributed to various factors. For example, higher-priced products may appeal to a narrower segment of consumers, leading to fewer sales overall. Additionally, high prices can deter impulse purchases and may require more consideration from consumers before making a purchase decision. On the other hand, lower-priced products may have broader appeal and may be more accessible to a larger number of consumers, resulting in higher sales volume per order.

Understanding this negative correlation between price and sales volume is crucial for pricing strategies and revenue optimization. It allows businesses to balance profit margins with sales volume, taking into account factors such as consumer behavior, market competition, and product positioning.

4.4.2 Category

From the plot below, we can find out that:

- The product category ”bed bath table” has the highest sales, indicating that its products have the highest demand compared to other categories during the observed period. This could be due to the high demand for items such as bedding, bath products, and tables in people’s daily lives.
- The product categories ”furniture decor”, ”health beauty”, ”sports leisure”, ”computers accessories”, and ”housewares” follow with relatively high sales. Although not as high as ”bed bath table”, these categories still occupy prominent positions among all product categories, suggesting significant consumer demand for home decor, health and beauty products, sports and leisure items, computer accessories, and household goods.
- There may be underlying market trends and consumer preferences reflected in the high sales of certain product categories. For example, the popularity of ”bed bath table” products may indicate a strong consumer focus on comfort and home living quality, while high sales in the ”health beauty” category may reflect growing concerns about health and personal care.
- These findings have implications for marketing strategies and inventory management. Understanding which product categories have high sales can help businesses develop more effective marketing strategies and inventory management plans. For high-selling categories, increasing inventory and enhancing marketing efforts can meet consumer demand and boost revenue. Meanwhile, for lower-selling categories, adjustments in marketing strategies or product features may be considered to improve competitiveness.



4.5 Product Score

The review_score is weighted based on all reviews for each product, as well as the number of reviews, and is combined to arrive at a final score for each product.

The five products with the highest score is below:

product id	category
aca2eb7d00ea1a7b8ebd4e68314663af	furniture decor
422879e10f46682990de24d770e7f83d	garden tools
99a4788cb24856965c36a24e339b6058	bed bath table
368c6c730842d78016ad823897a372db	garden tools
389d119b48cf3043d311335e499d9c6b	garden tools

4.6 Shipping Limit Date

To compare, we then analysis the top 50 with the highest scores and bottom 50 products.

- For the top 50 products, the time elapsed between the merchant's dispatch of the product and the customer's receipt of it is within the stipulated timeframe. The average delivery days are 12.98.
- Among the top 50 products, six merchant sent the product after the stipulated time limit, and five customers received the product after the limit time. The average delivery days are 14.36.

5 Association Rule Mining

5.1 Introduction

In this section, we explore the application of association rule mining and collaborative filtering to analyze transaction information from the dataset. Association rule mining is a popular data mining technique used to discover interesting relationships, such as frequent patterns, associations, or correlations among large sets of data items. This analysis helps e-commerce platform understand the purchasing relationships between different products, thereby optimizing inventory management and promotional strategies.

5.2 Data Preprocessing

Before mining association rules, it's essential to prepare the data adequately:

- **Customer and Geolocation Data Integration**:
  - **Files**: `olist_customers_dataset.csv`, `olist_geolocation_dataset.csv`
  - Merge customer profiles with geolocation data to enrich the customer dataset with geographic coordinates, stored as `customer_data.csv`.
- **Order Data Integration**:
  - **Files**: `olist_orders_dataset.csv`, `olist_order_items_dataset.csv`
  - Combine order data with item details to create a comprehensive order dataset that includes prices and quantities.
- **Product Category Translation**:
  - **Files**: `olist_products_dataset.csv`, `product_category_name_translation.csv`
  - Align product categories in English by merging product data with translation data, facilitating analysis across different regions.
- **Final Customer Order Dataset Preparation**:
  - **Output File**: `customer_order.csv`
  - Integrate enriched customer data with detailed order and product information, including English product categories, preparing the data for detailed analysis.
- **Data Transformation for Analysis**:
  - **Output File**: `transaction_data.csv`
  - Convert the comprehensive dataset into a format suitable for mining association rules, where each transaction is represented as a set of items.

And finally, we get the dataframe with customers and the product bought in terms of category.

customer unique id	agro industry and commerce	air conditioning	art	arts and craftmanship	audio	auto	baby	bed bath table
0000366f3b9a7992bf8c76cfdf3221e2	0	0	0	0	0	0	0	1
0000b849f77a49e4a4ce2b2a4ca5be3f	0	0	0	0	0	0	0	0
0000f46a3911fa3c0805444483337064	0	0	0	0	0	0	0	0
0000f6ccb0745a6a4b88665a16c9f078	0	0	0	0	0	0	0	0
0004aac84e0df4da2b147fca70cf8255	0	0	0	0	0	0	0	0

Figure 8: Customer-items

5.3 Mining Association Rules

We employ algorithm FP-growth to mining the assiciation rules between groceries.

- **Setting Support Threshold**: Select a 0.00002 support threshold to filter out infrequent itemsets.
- **Setting Confidence Threshold**: Establish the level of confidence as 0.1 to ensure the reliability of the rules.
- **Rule Extraction and Evaluation**: Extract high-confidence rules from the frequent itemsets and assess their usefulness and significance.

The frequent itemset analysis is conducted using the FP-Growth algorithm. The analysis was aimed at identifying commonly purchased item categories among transactions within our dataset. Below is a detailed summary of each itemset along with its respective support value.

Bed Bath Table: Exhibits a support of 0.095839, indicating that approximately 9.58% of transactions include items from the "Bed, Bath, and Table" category. This category is one of the most popular among our customers.

Health Beauty: With a support of 0.090945, about 9.09% of transactions feature "Health and Beauty" products, highlighting significant consumer interest in personal care items.

Stationery: Holds a support of 0.024062, showing that stationery items appear in roughly 2.41% of transactions. This lower support suggests a niche market within our customer base.

Telephony: This category has a support of 0.043513, indicating that approximately 4.35% of transactions include telephony-related items, reflecting a moderate demand in telecommunications products.

Table 1: Frequent Itemsets and Their Supports

Support	Itemsets
9145.0	bed bath table
8678.0	health beauty
7515.0	sports leisure
6557.0	computers accessories
6317.0	furniture decor
……	……
2.0	furniture living room, toys
2.0	computers accessories, construction tools construction
2.0	furniture living room, baby
2.0	furniture living room, garden tools
2.0	cool stuff, consoles games

These generated rules show which products are likely to be bought (consequents) if the customer has already purchased a set of products (antecedents), with more than 10% probability. These rules can be utilized to understand relationship between products and demand within customers. It can also be used to recommend a product to a customer based on what the customer is currently buying. For example, If a man has bought bed bath table, construction tools lights, the probability that he wants to buy furniture decor is 66.7%, which is 10.070181 times of the probability that this category is bought randomly.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(bed bath table, construction tools lights)	(furniture decor)	0.000031	0.066202	0.000021	0.666667	10.070181
(home construction, housewares)	(furniture decor)	0.000042	0.066202	0.000021	0.500000	7.552636
(cool stuff, auto)	(bed bath table)	0.000073	0.095839	0.000021	0.285714	2.981176
(fashion childrens clothes)	(fashion bags accessories)	0.000084	0.018843	0.000021	0.250000	13.267519
(furniture decor, home confort)	(bed bath table)	0.000094	0.095839	0.000021	0.222222	2.318693
(perfumery, housewares)	(health beauty)	0.000094	0.090945	0.000021	0.222222	2.443471
(bed bath table, garden tools)	(housewares)	0.000168	0.061004	0.000031	0.187500	3.073570
(bed bath table, fashion bags accessories)	(housewares)	0.000115	0.061004	0.000021	0.181818	2.980431
(fashion bags accessories, housewares)	(bed bath table)	0.000115	0.095839	0.000021	0.181818	1.897112
(home confort)	(bed bath table)	0.004150	0.095839	0.000566	0.136364	1.422834
(bed bath table, auto)	(cool stuff)	0.000157	0.037885	0.000021	0.133333	3.519410
(health beauty, housewares)	(perfumery)	0.000157	0.032739	0.000021	0.133333	4.072557
(garden tools, housewares)	(bed bath table)	0.000252	0.095839	0.000031	0.125000	1.304265
(baby, housewares)	(bed bath table)	0.000178	0.095839	0.000021	0.117647	1.227543

Figure 9: The association rules

5.4 Item-Based Collaborative Filtering

Item-based collaborative filtering operates on the principle that users who agree on some items are likely to agree on others. This methodology focuses on the relationships between items themselves rather than user-user interactions. The theoretical foundation of this approach can be outlined through the following key concepts:

Vector Space Model for Items Each item in the dataset is represented as a vector in a multi-dimensional space, where each dimension corresponds to a user. The entries in this vector are the ratings or interaction metrics that users have assigned to the item. For binary interactions (such as purchase or no purchase), these vectors are binary.

Similarity Metrics The core of item-based collaborative filtering is the computation of similarity between item vectors. Cosine similarity is the most commonly used metric due to its effectiveness in high-dimensional spaces. Mathematically, the cosine similarity between two items i and j is defined as:

$$\text{similarity}(i, j) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \|\vec{j}\|}$$

where $\vec{i}$ and $\vec{j}$ are the item vectors, and the dot product of the vectors is divided by the product of their norms. This ratio measures the cosine of the angle between the two vectors, providing a similarity score between -1 and 1, where 1 indicates perfect similarity.

Prediction Computation To predict whether a user u will like an item i, item-based CF calculates a weighted sum of the user's ratings on similar items, weighted by the similarity scores:

$$\text{prediction}(u, i) = \frac{\sum_{j \in I_u} \text{similarity}(i, j) \times r_{u,j}}{\sum_{j \in I_u} |\text{similarity}(i, j)|}$$

where I_u is the set of items that user u has interacted with, and $r_{u,j}$ is the rating or interaction value of user u on item j.

5.4.1 Implementation Example

After establishing the theoretical foundations and computation methodologies, we illustrate the practical application of our item-based collaborative filtering system. Consider a scenario where a user has purchased items from the categories 'bed bath table' and 'toys'. We are interested in recommending other products that this user might like based on their previous purchase behavior.

Given the user's purchase history, the system's recommendations are computed as follows:

```
User purchased: ['bed bath table', 'toys']
Recommended products:  Furniture Decor: 0.296092  Home Confort: 0.272231
                       Housewares: 0.161075  Baby: 0.121167  Cool Stuff: 0.078814
```

5.4.2 Interpretation of Results

These recommendations are generated using item-based collaborative filtering, where each recommendation's value represents a similarity score, calculated based on other users' interactions with these categories. The scores indicate the likelihood that the user will be interested in these categories based on their similarity to items the user has already purchased.

The recommendation engine suggests that the user who purchased 'bed bath table' and 'toys' might also be interested in 'furniture decor' and 'home confort', with these categories showing the highest similarity scores. This implies a strong associative relationship between the purchased and recommended items, which could be utilized for targeted marketing campaigns or personalized product placements on e-commerce platforms.

6 Reviews Sentiment Analysis

Reviews and ratings are given to the products sold on the Olist website. These reviews can be used to understand the sentiment of the review towards a product. These sentiments can help recognize popular products on the Olist website.

There are two approaches to this problem. Supervised sentiment analysis can be performed using ratings given as the target. Unsupervised learning can be performed to estimate the sentiments of these reviews and hence the popularity of a product. For unsupervised, the reviews have to be translated to English.

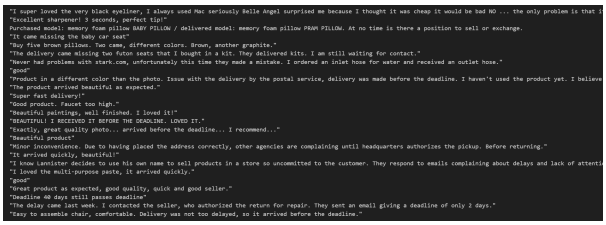
6.1 Implementing GPT-3,5-Turbo

In olist_order_reviews_dataset.csv , we got the review comments from customers in Portuguese.By using the pipeline of gpt-3.5-turbo, we translated the Portuguese to English for next analysis.

```python
def translate_to_english(text):
    if pd.notna(text) and text.lower() != 'nan':
        text = " ".join([word for word in word_tokenize(text) if word.lower() not in STOP_WORDS])
        prompt = f"Translate the following text from Portuguese to English: \"{text}\""

        try:
            response = openai.ChatCompletion.create(
                model="gpt-3.5-turbo",
                messages=[
                    {"role": "system", "content": "You are a helpful assistant."},
                    {"role": "user", "content": prompt}
                ],
                max_tokens=100,
                n=1,
                stop=None,
                temperature=0.5,
            )
            t_text = response.choices[0].message['content'].strip()
            print(t_text)
            return t_text
        except Exception as e:
            print(f"Translation error: {e}")
            return text
```

Openai Implement



Translation

Due to limitations of resources, only 10K reviews could be translated over a period of 2 days. Hence for the unsupervised method, only 10K reviews are utilized. For the supervised method, all the reviews can be used.

About 27500 data points were translated. Out of which 10K reivews are valid ones (not NaN). Few manipulations are performed to reduce the dataset to only the 10K valid reviews for further analysis.

For the unsupervised method, the predition of sentiment of a review is based on a-priori knowledgebases, lexicons and ontologies. Lexicon is a dictionary or a book of words, for this case specifically created for sentiment analysis.

Lexicons contains positive and negavtive words with scores. When applied on the data, based on Parts of speech, surrounding words, context, phrazes etc a score is calculated. Aggregation of the score then implies if the review is positive or negative.

There are two lexicons used - **AFINN** and **TextBlob**.

6.2 Implementing AFINN and Textblob

product_id	score	sentiment
37664a6100bf523c545b7040be69ad00	17.0	positive
f71954485428cf87945953ea0a0a6229	16.0	positive
f82a4b08cf7b2bf375fb77e519231f9a	16.0	positive
c696cd2db7870be08c37822f7c0ef831	15.0	positive
92857a2b590adeae8d5992511dc197e9	15.0	positive
1574ed2c73e4465c572f68dd77528203	14.0	positive
12e479f1c2c0417147a3e36302485e30	14.0	positive
4563095e06df1fa67de2eade86f4f01b	14.0	positive
0c96091f6f20b1d2c946a933370782a4	14.0	positive
e21af3d371523b2d9543645d3564a873	14.0	positive
7810b42c3450a044431d95c59c00b600	14.0	positive
dfb68ce52d4975e692909a1bf9d59f1b	12.0	positive
2b79d65ddb4051238ce31fbabc5763ce	12.0	positive
d37c7b44350fea9327f32f3e603f682b	12.0	positive
dfd2ea8377fb2838a47488c5b37558f1	12.0	positive
4c8dfecc09a4b15b7c71b219278d9026	12.0	positive
91421c5de3c0cb822ec9bf05f9220267	12.0	positive
9fa7e05ae5cddeb74ff2fe5046b742db	12.0	positive
546c7560260f60e3e086d0d106ced4e5	11.0	positive
323c53e2d6f8bace877b0df931ab58f9	11.0	positive

Top20 Products with scores

product_id	score	sentiment
77e8ff91e2cccad62b30a0a83bec431e	-6.0	negative
0cf3ab3383c2ae6f5750e5847c749e22	-6.0	negative
24c855c83d506d85c7386ff94d6c8db1	-6.0	negative
9e88896ff9b39f42a9652dc0730ecabc	-6.0	negative
1a0ef485e35802dd1d23ab5446cac458	-6.0	negative
093cd981b714bcdff182b427d87fc8fc	-6.0	negative
07797d4315ce5c1570e0a484472e04b5	-6.0	negative
ea768069ca084e95a752ee7d51a8ff70	-6.0	negative
d5abd3b28e2a7a0da8f0a4617f9188d6	-6.0	negative
18297d6ba9247aa8fb22c28df54bbcd4	-6.0	negative
f6691c7e7a0c12d28003c97608dfe605	-6.0	negative
04a3191f211edc97716020026d05d260	-6.0	negative
f8a9ed2d85e69c00f817d88ee3de7333	-6.5	[negative, neutral]
d9d3f97d19971e6e134c04428b7ef416	-7.0	negative
d68bd4dedccc5545b1ff6629de8fb021	-7.0	negative
b1443a3dadb0b2b1edf47ffe38366788	-7.0	negative
de75785dc5be712c4d4ed4ec5b422ce1	-7.0	negative
c3ae2001a36724730fbc90605b9278e2	-8.0	negative
51bb55b87cf953219cadc7deb9d282c2	-9.0	negative
d5a19cc8a92434231693043e185b4c5b	-12.0	negative

Least20 Products with scores

To understand why these reviews have such a score, a word cloud is created.

Positive WordCloud

Negative WordCloud

The wordcloud shows the most frequent words as bold. In the top 20 products with positive reviews, words such as - good, beautiful, super, loved, recommend, perfect, satisfied are used mostly. Hence the positive sentiment assigned.The wordcloud for least popular 20 products shows why these product's overall sentiment score is negative. The most frequent words are - bad, missing, upset, problem, defective etc.

Since these methods were unsupervised, there are no target values to check the predictions against. But there can be a comparision made between the two models, by assuming one model to predict actual values and the other model be tested agaisnt it. This way it can be said how many predictions match between the two. If most of the predictions between the two match, then that can be said as the true value for that review.

The Confusion Matrix

```
confusion_matrix(afinn_reviews.sentiment, textblob_reviews.sentiment,
labels= ['positive','neutral','negative']

[[5547    0    0]
 [   0 3212    0]
 [   0    0 1350]]
```

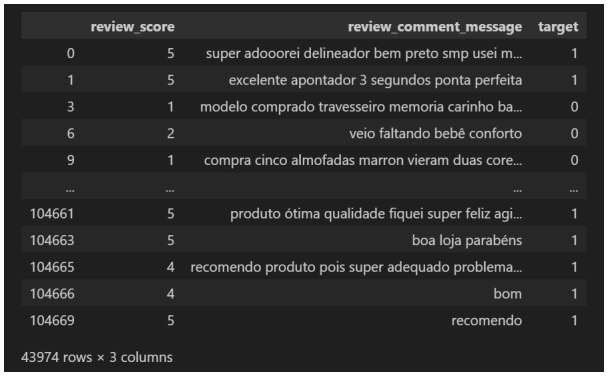
The confusion matrix shows that between the two models, the predictions match and hence are considered to be true labels for the reviews. Now that the sentiments with score for each review is knows, products with positive sentiments can be viewed to find what was it that customers liked. Similarly negative sentiment reviews can be analyzed to understand what customers disliked and how that can be improved. Based on wordcloud it was clear that defective, missing or delay are words mostly in negative sentiment reviews and are the causes for it.

6.3 Supervised Learning Model

The next part of the analysis, is to make a supervised learning model for the reviews. As compared to the previous method, unsupervised learning model is trained on the reviews and tested against a target variable. This ensures that the model is a perfect fit. Further, the model can be used to classify future reviews as they are posted.

For the model, the target has to be created. Each review has a rating assigned. Let us consider a benchmark - 2.5.Any rating below 2.5 is conidered as negative and above 2.5 as positive.

From the unsupervised method , the reviews are now cleaned and free of Stop words. Another important step in preprocessing is stemming the words. The reviews may have same words in different tense. By the model, these are picked up as different words for example - doing, does, done are all of the same root/stem do, but wont be treated as such.



	review_score	review_comment_message	target
0	5	super adooorei delineador bem preto smp usei m...	1
1	5	excelente apontador 3 segundos ponta perfeita	1
3	1	modelo comprado travesseiro memoria carinho ba...	0
6	2	veio faltando bebê conforto	0
9	1	compra cinco almofadas marron vieram duas core...	0
...	...	...	...
104661	5	produto ótima qualidade fiquei super feliz agi...	1
104663	5	boa loja parabéns	1
104665	4	recomendo produto pois super adequado problema...	1
104666	4	bom	1
104669	5	recomendo	1

43974 rows × 3 columns

Figure 10: For this model, the reviews have not been translated

A stemmer is applied to each word in the review.For sentiment analysis, it is important to understand how the reviews can be. A text can either be objective or subjective. Objective text usually inform a fact whereas subjective text show emotions. Emotions is what the sentiment analysis banks on. It is important to understand which words can show emotion and how crucial can the words be for the same. Say a word "product" in the reviews will be very common, and such words do not usually show emotions nor due to their frequency account for classification. A word which is rare in the reviews can really be crucial in identifying the sentiment of the review.

Tfidf Metric

Hence tfidf metric is chosen to give weightage to each word in the review. A very frequent word is given less weightage, whereas a rare word is given a heavy weightage. After applying the weighting, data is split into train and test.Before splitting we check the distribution of the target variable. This can give a good idea of the splitting ratio for train and test.

```
1    0.719562
0    0.280438
Name: target, dtype: float64
```

Logistic Regression

Now the train and test data are ready, the model is fit on the training data and tested on test data.

The model to try is the Logistic Regression. LogisticRegressionCV performs K-fold cross validation and grid search for C and l1-ratio values.

	precision	recall	f1-score	support
0	0.80	0.79	0.79	3436
1	0.92	0.92	0.92	8896
accuracy			0.89	12332
macro avg	0.86	0.86	0.86	12332
weighted avg	0.89	0.89	0.89	12332

The logistic Regression model, after cross validation gave an accuracy of 89%. The weighted average for precision, recall and f1-score are also at 89%. But it can be seen that precision and recall for negative class is way lower than positive class, this is because of the imbalance in the dataset.

RandomForest Classifier

The next model to be tested is the RandomForest Classifier. The RandomForest Classifier is an ensemble technique and powerful in removing biases and errors using weak learners.

	precision	recall	f1-score	support
0	0.81	0.80	0.80	3499
1	0.92	0.93	0.92	8814
accuracy			0.89	12313
macro avg	0.87	0.86	0.86	12313
weighted avg	0.89	0.89	0.89	12313

The RandomForest Classifier shows improved results in terms of precision and recall of negative class as compared to Logistic Regression, yet there is a gap between positive and negative class. The accuracy and weighted average of precision, recall and f1-score are all 89%.

SVC

The next model to be tested is the SupportVector Machine.

	precision	recall	f1-score	support
0	0.82	0.81	0.82	3476
1	0.93	0.93	0.93	8837
accuracy			0.90	12313
macro avg	0.87	0.87	0.87	12313
weighted avg	0.90	0.90	0.90	12313

The SupportVector Machine shows improved accuracy and weighted average of precision, recall and f1-score of 90%. The gap between positive and negative classes still exists.

Model selection

For model selection, ROC curve and AUC score is reported. The model with higher AUC is the model to be selected.

```
AUC LogisticRegression:  0.942435674180289
AUC RandomForest:        0.9450056715387773
AUC SupportVector:       0.9485774753873948
```

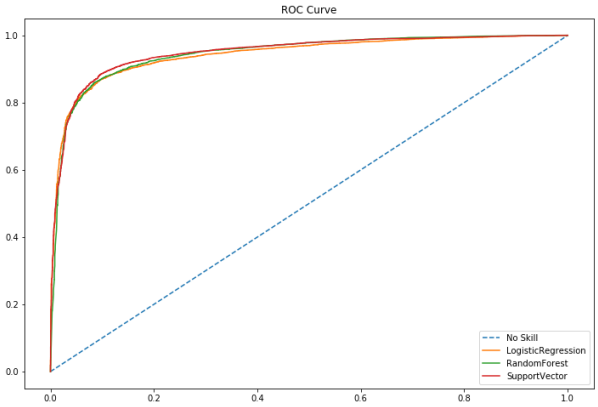


Figure 11: Models' ROC Curve

All the three models have very close AUC scores. Support Vector machine has the highest of them all. The SupportVector model has shown highest AUC and accuracy among all the models. Thus the SupportVector model is chosen as the final model. The trained model can further be used to classify new reviews posted as orders are placed.

Hence the final model is SupportVector Machine for the sentiment analysis.

References

[1] Olist, and André Sionek. *Brazilian E-Commerce Public Dataset by Olist*. 2018. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/195341

[2] Anand HU. *Population Brazil*. Data set available on Kaggle. https://www.kaggle.com/datasets/anandhuh/population-brazil

[3] Raj Tulluri. *Olist Business Analysis*. GitHub repository, 2022. https://github.com/rajtulluri/Olist-business-analysis