# BAX 452 Machine Learning: Final Project Report

**Topic:** Accelerating Brazilian E-commerce Deliveries: Insights and Predictive Models for Efficient Service

Group Members: Shayan Farshid, Qinyi Qiu, Srinivas Abhilash Chintaluru

Master of Science in Business Analytics

University of California, Davis

# Table of Contents

## Executive Summary

In the dynamic world of e-commerce, getting products to customers swiftly is the prime key to success. Olist, a Brazilian e-commerce platform that connects small and medium-sized businesses with customers across Brazil, understands this well enough. However, despite this awareness, Olist faces numerous challenges in keeping up with prime delivery companies' speedy service. Our goal is straightforward: to dig into the reasons behind these delays and use predictive analytics to help deliver faster and more efficiently.

To achieve this, we are taking a thorough approach of conducting exploratory data analysis, and intend to analyze our data to find specific order segments using techniques like K Means clustering. This will help us understand different types of orders and what makes some arrive faster than others. After which, we will use supervised learning methods like Linear regression, Random Forest, and XGBoost, draw contrasts that can predict how long delivery will take based on factors like location, payment method, and product details. By doing this, we hope to give Olist the tools that it needs to streamline operations, cut delivery times, and make customers happier in the long run.

## Background

Olist, a significant player in Brazilian e-commerce since its inception in 2015, operates as a marketplace connecting small and medium-sized businesses with customers across the country. Through Olist, merchants can list their products and services for online purchase by customers. The dataset under analysis, obtained from Kaggle, contains information on about 100,000 orders spanning 2016 to 2018 from various Brazilian marketplaces. This publicly accessible dataset provides insights into order status, pricing, payment, freight performance, customer location, product attributes, and customer reviews.

## Context

In the competitive e-commerce landscape, timely product delivery is critical for success. Swift delivery not only reduces operational costs but also significantly enhances customer satisfaction, positioning companies favorably amidst competition. However, Olist faces challenges as its delivery times often lag behind those of prime delivery companies.

## Competition

Olist operates in a competitive landscape where it contends with prime delivery companies renowned for their rapid and efficient service, such as Unbox Showcase, Avec Brasil, and local express delivery services. These competitors have established a reputation for delivering products within exceptionally short timeframes, setting a high standard in the industry. As Olist strives to optimize its delivery times and enhance customer satisfaction, it must closely monitor the strategies and performance of these competitors to identify areas for improvement and maintain a competitive edge.

## Domain Knowledge

Understanding e-commerce operations and customer expectations is crucial in addressing delivery time challenges effectively. Key factors influencing delivery times include location, payment type, product weight, and volume. Through advanced analytical techniques on the available data we aim to uncover unique order segments with characteristics conducive to faster delivery and develop predictive models to accurately estimate total delivery time.

## Traditional Strategies

Within the e-commerce industry, firms typically have targeted strategies to address delivery time challenges. One specific tactic involves optimizing distribution networks and partnering with regional carriers to expedite last-mile delivery. Additionally, firms may invest in advanced inventory management systems and predictive analytics to forecast demand and strategically position inventory closer to customers. Such approaches align seamlessly with the business model of e-commerce companies, as they prioritize swift delivery to enhance customer satisfaction and loyalty. By fine-tuning logistical operations and leveraging data-driven insights, firms attempt to minimize delivery times while maximizing operational efficiency.

## Objectives

1. Identify distinct order segments with unique attributes conducive to faster delivery.
2. Develop predictive models to estimate the total delivery time for orders accurately.

## Analysis

**Data:** We utilized the orders data, below is a brief description of the data fields.

| Column Name | Description | Data Type | Example Values |
|---|---|---|---|
| order_id | Unique identifier for order | lphanumeric | "e481f51cbdc54678b7cc49136f2d6af7" |
| order_status | Current status of order | Text | "delivered" |
| customer_state | State of the customer | Text | "SP" |
| product_category_name | Category of product | Text | "utilidades_domesticas" |

| | | | |
|---|---|---|---|
| order_item_id | lentifier for the item within the order | Integer | "1" |
| seller_state | State of seller | Text | "SP" |
| payment_type | Type of payment used | Text | "credit_card, voucher, voucher" |
| seller_dispatch_time | Descriptive category for seller dispatch time | Text | "Fast" |
| customer_delivery_time | Descriptive category for customer delivery time | Text | "Fast" |
| order_purchase_time_slot | Time slot of the day when order was purchased | Text | "9_12" |
| price | Price of the product | Decimal | "29.99" |
| freight_value | Shipping cost charged for the order | Decimal | "8.72" |
| product_weight_g | Weight of the product in grams | Integer | "500" |
| volumetric_weight | Volumetric weight used for shipping calculation | Decimal | "395.2" |
| payment_installments | Number of installments | Integer | "1" |
| total_payment_value | Total amount paid for the order | Decimal | "38.71" |
| seller_to_carrier_deliverytime | Time taken for the order to be delivered from seller to carrier (in days) | Decimal | "2.366493" |
| arrier_to_customer_delivery_time | Time taken for the order to be elivered from carrier to customer (in days) | Decimal | "6.062650" |

| | | | |
|---|---|---|---|
| total_delivery_time | Total delivery time from purchase to delivery to the customer (in days) | Decimal | "8.429144" |

**Data Preparation and Feature Engineering:**

We checked for null values, and fortunately enough, our data is quite clean and there is no need for data imputation.

Order purchase time slot: Bucketing the order purchase time into hour slots

Total delivery time: Adding the seller_to_carrier_delivery_time and carrier_to_customer_delivery_time

For categorical variables with a higher number of unique values, we reduced the number of categories by combining all the categories with lower frequencies to “Other”

Payment type: Created features that hold the count of each payment type for every order. Eg: For the value in payment_type = ‘credit card, credit card, voucher, voucher’ we created features payment_type_credit_card = 2, and payment_type_voucher = 2 which store the frequency of each payment type

**Exploratory Data Analysis:**

We examined various factors like customer locations, payment methods, and product categories to gain insights into delivery time, and the order payment value. We didn’t limit ourselves to the target variable, instead, we tried to analyze all the possible variables that might provide any possible insights to Olist.

We found the following insights for different attributes:

**Customer State:**

1. São Paulo(SP) has a significantly higher number of orders followed by Rio de Janeiro(RJ), and Minas Gerais(MG) states.
2. Roraima(RR), Amapá(AP), and Amazonas(AM) states have a very high average delivery time of over 25 days

3. Paraíba(PB), Acre(AC), and Amapá(AP) states have a higher average payment value as compared to the other states

**Seller State:**

1. Sellers from São Paulo(SP) state have a significantly higher number of orders processed as compared to any other state

2. The orders processed by sellers in the state of Amazonas(AM) seem to have a significantly higher average. delivery time as compared to the rest

3. The orders processed by sellers in the states Paraíba(PB), Bahia(BA), Amazonas(AM), and Rondônia(RO) seem to have a higher average. order payment value as compared to the rest

**Product categories:**

1. Bed, table & bathroom(cama_mesa_banho), beauty & health(beleza_saude), sport & leisure(esporte_lazer) product categories seem to have a higher proportion of orders as compared to the others

2. Office furniture(moveis_escritorio) product category seems to have a significantly higher delivery time as compared to the other product categories

3. Office furniture(moveis_escritorio), and Watches & Gifts(relogios_presentes) product categories seem to have a higher avg. payment value as compared to the other product categories

**Time slot:**

1. A higher proportion of orders are placed during the time slots 15_18(3 PM to 6 PM), 12_15(12noon to 3 PM), while time slots 0_6(12 AM to 6 AM) and 6_9(6 AM to 9 AM) have a significantly lower frequency of orders placed

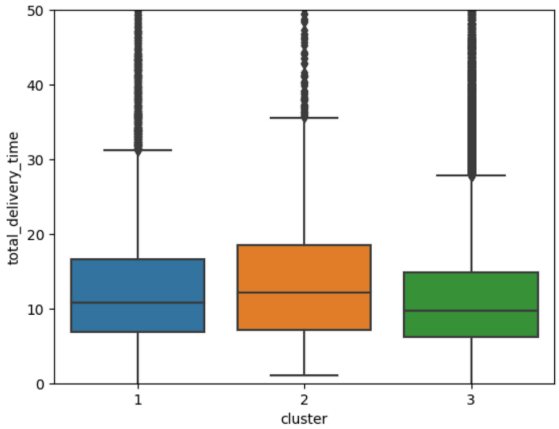
2. The orders placed in 0_6(12noon - 6 AM) seem to have a slightly higher delivery time as compared to the other slots

Next, we calculated the correlation between all the numerical columns in the data and found a very high correlation of 0.92 between variables (price, total_payment_value), and a value of 0.8 between product weight and product volume.

**Unsupervised Learning:**

We performed clustering to identify the clusters that are similar to each other by providing all the data except the total_delivery_time as the input, and after we generated the clusters, we tried to evaluate if the clusters were significantly different from each other by looking at the target variable(total_delivery_time). Moreover, we ran an **ANOVA test**, for which we obtained the F-statistic of 176, and a p-value less than any significance level which, statistically proves that the difference between the clusters is significant.

The below visualization shows the box plot of total delivery time across the three clusters.




Next, we did a post-clustering analysis to observe the box plots for all the variables across these three clusters, and the results show that the orders with high **freight value**, **weight**, and **volume**, **payment installments**, **total payment value** take higher delivery time as compared to the rest of the orders.

**Supervised Learning:**

Our objective here was to predict the total delivery time using the available features, we tested the following models

1. Linear regression model: A higher F-statistic value and a lower p-value(less than any significance level) indicate that the model fits the training data well, while the out-of-sample R-squared value of 57% indicates that the variables explain 57% of the variation in the total_delivery_time.

2. Reduced Linear regression model: Used the FDR control technique with a q value of 0.2, and considered all the variables with $p < p^*(0.1514)$ in our reduced model. The F-statistic was higher than the full model and the p-value again was less than any significance level, indicating that the reduced model was

performing better than the full model, but, the out-of-sample R-squared lift was very minimal and didn't give us the confidence to go ahead and select this model

3. Random Forest model: The Random Forest model gave us an out-of-sample R-squared of 79% which was pretty high compared to the linear models we tested.

4. XGBoost model: The XGBoost model gave us an out-of-sample R-squared of 77% which was good enough and high as compared to the linear models we tested.

**Hyperparameter tuning:**

We tuned the hyperparameters for the Random Forest and XGBoost models, but since the tuning models were underperforming, we decided to choose the default models.

**Model Selection:**

We finalized the Random Forest model as that has the highest out-of-sample R-squared as compared to the other models, although XGBoost has a close value of 77%, we are not choosing it considering a few issues with XGBoost like interpretability, etc.

## Recommendations & Business Value:

From the exploratory analysis, we recommend the following:

1. As a significant proportion of customers are from the São Paulo(SP) state, Olist should expand its marketing campaigns to attract customers to increase the order frequencies from the rest of the states

2. Identify and acquire sellers, especially related to office furniture from Roraima(RR), Amapá(AP), and Amazonas(AM) and the nearby states to reduce the delivery times

From the post clustering analysis, we recommend the following:

As high value orders(i.e., orders with high total payment value) are expected to be delivered at the earliest by the customer, Olist should prioritize reducing the delivery time for the high value orders, and also considering strategize sellers location for high weight/volume orders to optimize the distance/delivery times

For predicting the total delivery time, Olist could leverage the Random Forest model and go ahead with the predictions for future orders. It would lead to the following business benefits,

1. **Customer satisfaction:** When customers know when to expect their orders, they are more likely to be satisfied, thus leading to **customer loyalty and retention**.
2. **Order conversion:** Providing accurate delivery estimates upfront would quickly help the customer place the order, instead of holding them in a cart, thus leading to **increased orders**.
3. **Inventory management:** Accurately forecasting the delivery times would help manage the inventory, thus leading to no overstocking or understocking, thus drastically reducing the **operational and infrastructure costs**.
4. **Enhanced communication:** As the order follows this process of seller -> carrier -> customer, there are a lot of stakeholders involved, and a delivery time would help them communicate better thus leading to better relationships with the seller, carrier, and customer, which would lead to **overall performance/satisfaction lift**

## Summary & Conclusions:

To Summarize, to solve our objective of identifying order segments as well as predicting the total delivery time, we started analysis to identify trends/patterns in the data and later tried to identify order attributes that would lead to higher/lower delivery times, after which we built predictive models to identify the delivery times.

To conclude, a properly estimated delivery time would be a critical data asset to the organization which would help it for further analysis and provide even higher business value going forward.