

# THE HATUCHELEWI MODEL

- Presented and developed by

**Otieno Calvin**

- Date

**7/9/2024**

# Overview

The **Hatuchelewi** model is a late delivery prediction model aiming to help curb customer dissatisfaction in the booming e-commerce industry

What well be looking at :

1. Problem Statement
2. Data description
3. Model selection
4. Methodology
5. Results
6. Insights and recommendations

# 1. Problem Statement

- In Puerto Rico's e-commerce sector, timely deliveries are crucial for customer satisfaction and maintaining a good reputation. Late shipments increase costs and upset customers. This dataset highlights key factors that influence delivery times.
- By predicting potential delays, businesses can take proactive steps to manage resources and improve customer experience.

## 2. Data description

The dataset provides insights into various aspects of customer orders, product categories, and shipping.

Here's a brief summary of what each column contains:

**Type:** Describes the transaction type, with the most frequent being "DEBIT," followed by "TRANSFER" and "PAYMENT."

**Days\_for\_shipping\_(real):** Most orders take 2, 4, 3, or 6 days to ship, with some being shipped immediately (0 days).

**Days\_for\_shipment\_(scheduled):** The majority of shipments are scheduled for 4 days, followed by 2 and 1 day.

**Benefit\_per\_order:** Varies widely, with small benefits being common, but there are also some negative and high positive values.

**Sales\_per\_customer:** Displays a range of sales per customer, with values clustering around common price points.

**Delivery\_Status:** Most deliveries are marked as "Late delivery," while fewer are marked as "Advance shipping" or "Shipping on time."

# Data description

**Department\_Id and Department\_Name:** Products are categorized into departments such as "Fan Shop" and "Apparel."

**Market:** Indicates geographic market regions like "LATAM," "Europe," and "Pacific Asia."

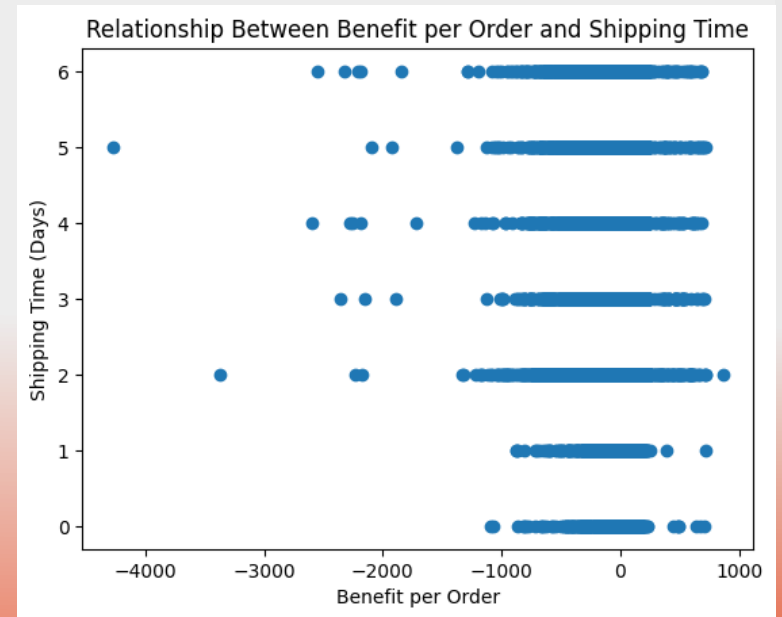
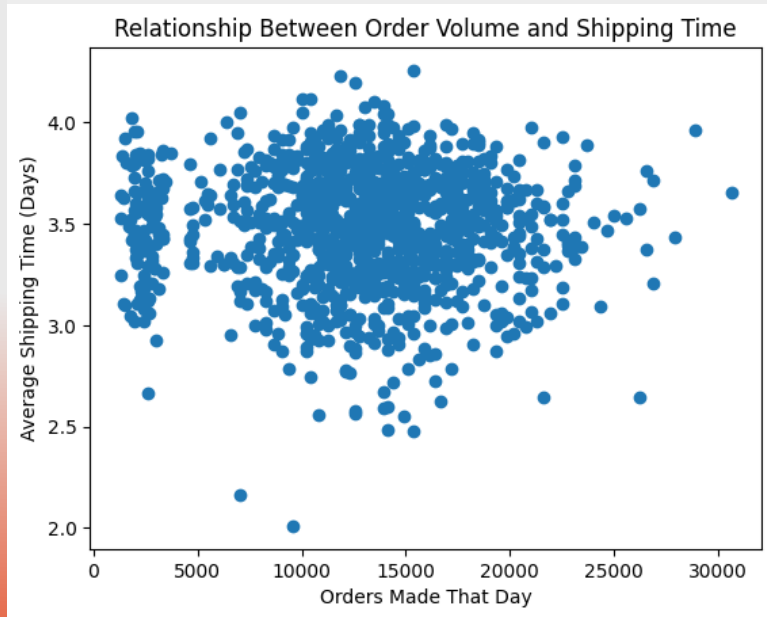
**order\_date\_(DateOrders):** Shows the distribution of orders over various dates.

**Late\_delivery\_risk:** A binary column where "1" indicates a late delivery and "0" indicates no risk.

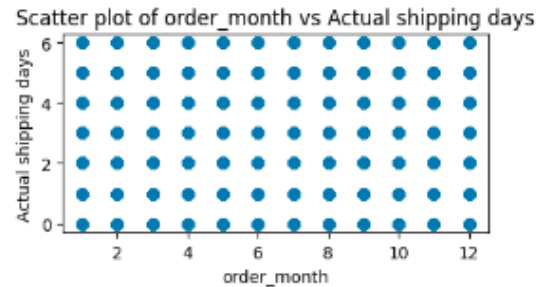
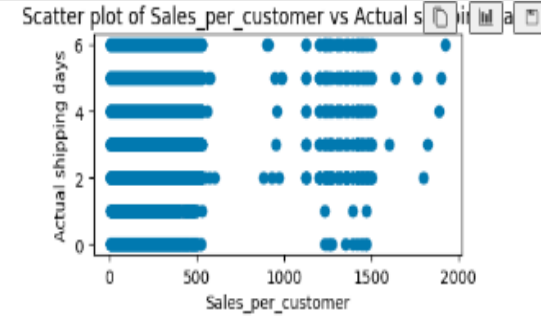
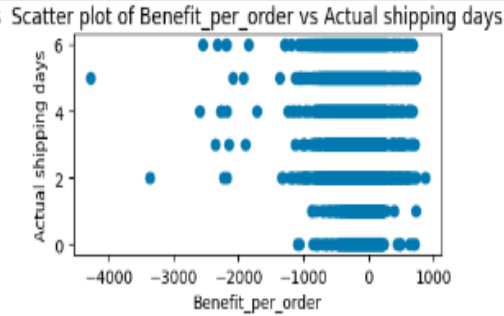
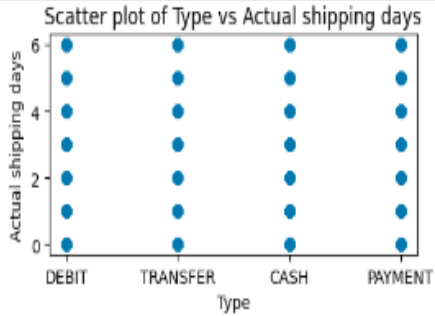
**Category\_Id and Category\_Name:** Each category ID maps to specific product types, with "Cleats" and "Men's Footwear" being among the most common.

# Data description

From basic EDA we were able to find that there was no correlation between most features and the target. With an optimized linear model yielding an  $R^2$  equivalent to 42% variance explained by model.



# Exploratory Data Analysis(EDA)



There was no direct linear relationship and so the Linear Model would be more difficult to engineer.

*Refer to appendix*

### 3. Model selection

I chose to use the **XGBoost classifier** model .( *XGBClassifier* )

**XGBoost (Extreme Gradient Boosting)** is a fast, efficient machine learning algorithm based on decision trees, designed for classification and regression tasks.

Given the dataset is imbalanced XGBoost, combined with Random Oversampling, offers a powerful and efficient way to tackle imbalanced datasets for late delivery prediction.

The model ran exemplary well on default parameters without the need of hyper-parameter tuning

Running on default parameters helps you create a baseline performance. This baseline gives you an understanding of how well the model performs without any optimization, providing a reference point to compare against later.



# Model selection

## Why Xgboost??

1. Combines weak learners (decision trees) using Gradient Boosting to create a strong predictive model.
2. Built-in regularization helps prevent overfitting, especially useful for complex data.
3. Handles missing data automatically, common in real-world datasets.
4. Optimized for speed and memory efficiency, ideal for large datasets.

## 4. Methodology

Given we now know what model we used . Lets now go through how , why and what we did prior to getting the results .

**- Training Process:** The model was trained with an 80/20 data split (80% for training, 20% for testing). Random oversampling balanced the late delivery classes, ensuring equal representation. 5-fold cross-validation was used for consistent validation across data splits.

**- Performance Metrics:** Accuracy, precision, recall, and F1 score were used to assess the model, focusing on identifying actual shipping days (recall) and reducing false positives (precision).

## 5. Results

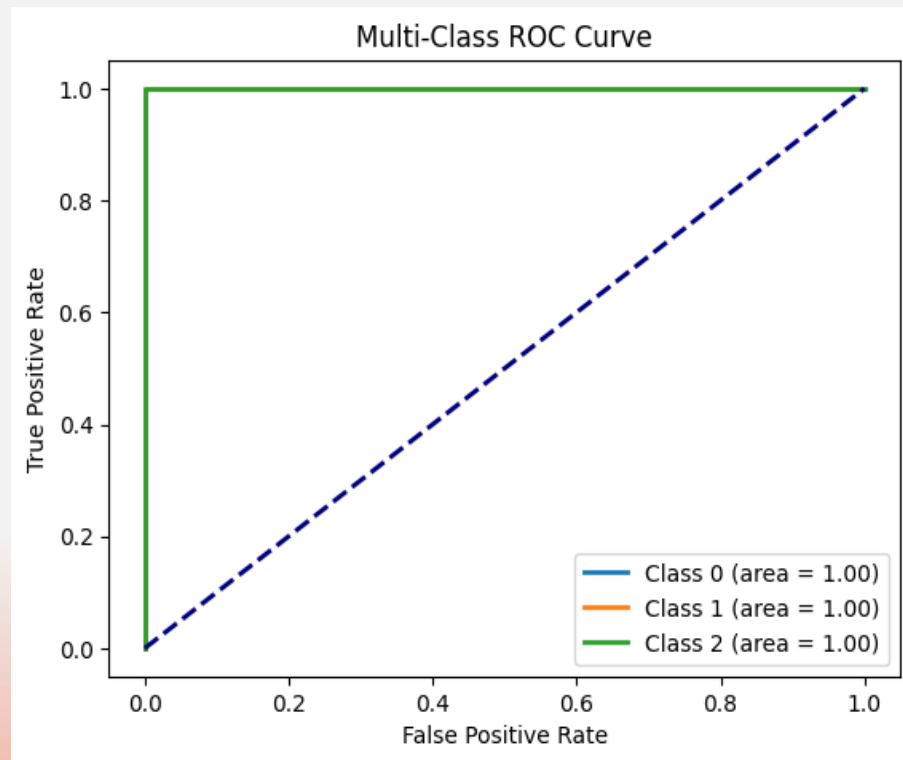
	precision	recall	f1-score	support
0	1.00	1.00	1.00	165
1	1.00	1.00	1.00	151
2	1.00	1.00	1.00	2276
3	1.00	1.00	1.00	1250
4	1.00	1.00	1.00	1233
5	1.00	1.00	1.00	1170
6	1.00	1.00	1.00	1239
accuracy			1.00	7484
macro avg	1.00	1.00	1.00	7484
weighted avg	1.00	1.00	1.00	7484

**Evaluation Metrics:** After applying random oversampling, the model achieved an accuracy of 100%, with precision at 100%, recall at 100%, and an F1 score of 100% for all days (classes) to be predicted. The improved recall indicates that the model is better at identifying actual shipping days without sacrificing any precision

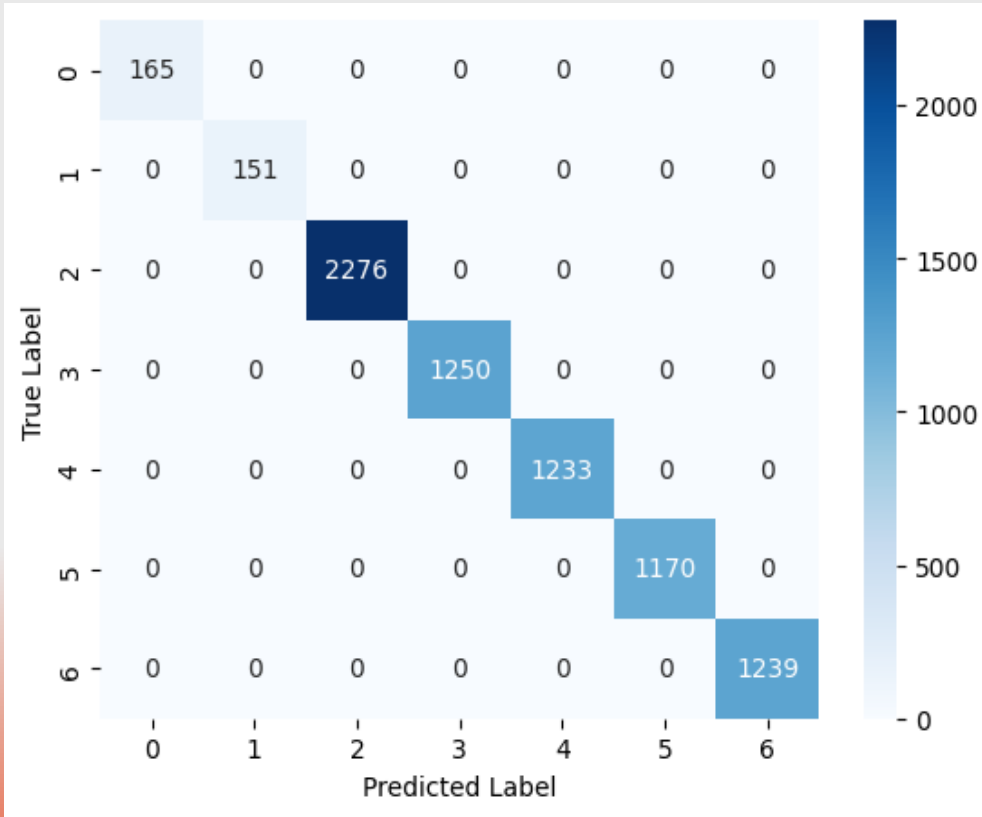
Cross-validation scores: [1. 1. 1. 0.99986636 1. ]

# Results

- **Confusion Matrix / ROC Curve:** A confusion matrix was generated to display the number of true positives, false positives, true negatives, and false negatives. The ROC curve showed an AUC of 1.00, signifying strong model performance.

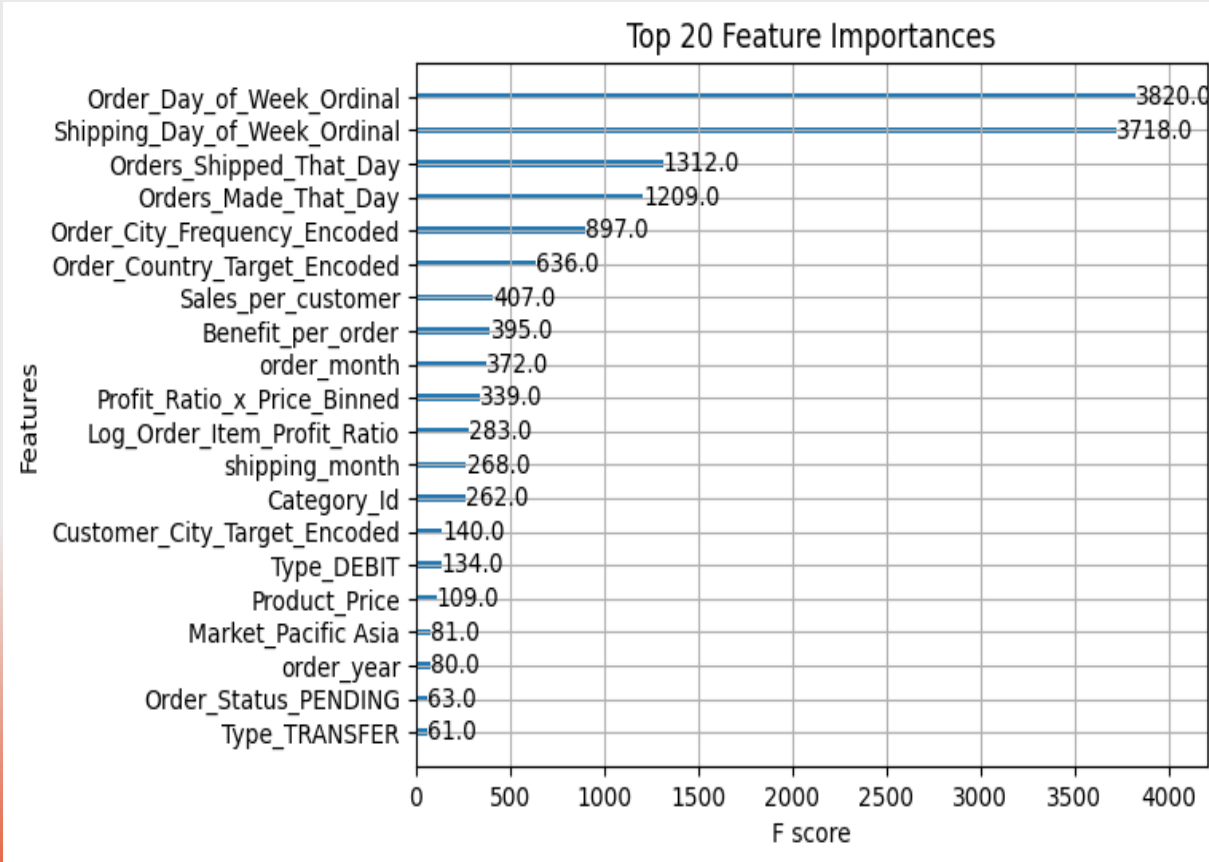


# Model selection



The confusion matrix shows that the model is performing well in predicting all classes although class 1 and 2 have lower values maybe due to them being the most premium shipping options.

# Model selection



**Feature Importance:** Feature importance analysis showed that “order\_day\_of\_week “ and “shipping\_day\_of\_week “were the most influential in predicting actual shipping days among others. These insights help identify key factors that affect delivery performance, allowing businesses to focus on optimizing these areas.

# 7. Model Insights

## Model Insights

Key Findings: The XGBoost classifier model highlighted several important patterns:

- **Order Day of Week** and **Shipping Day of Week** stood out as the biggest factors affecting delivery times.
- How many **Orders Shipped That Day** and **Orders Made That Day** also played a large role in predicting delays.
- Other key drivers include **Order City**, **Order Month**, **Market Region** where the order was placed among others.

## 8. Challenges and Limitation

**Model Limitations:** While the model performed well, it may still be prone to overfitting due to random oversampling. There is also potential bias if the dataset does not represent all regions equally, as some factors specific to Puerto Rico may not generalize to other markets.

**Challenges Faced:** Managing class imbalance and ensuring model consistency across different data subsets was a major challenge. We addressed this by using random oversampling and 5-fold cross-validation to maintain balanced training and reliable validation.



# 9. Recommendations.

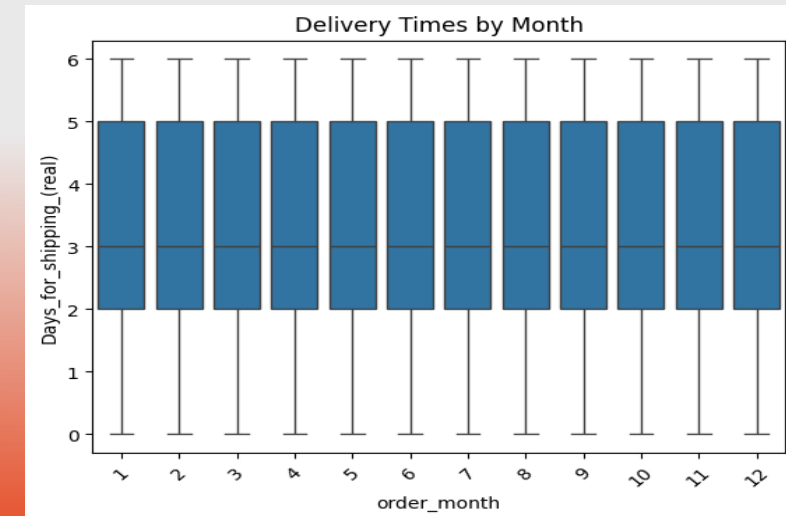
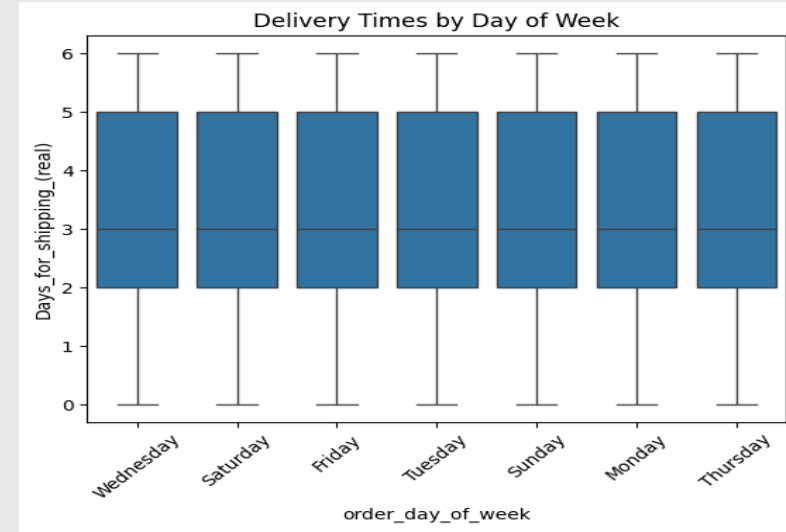
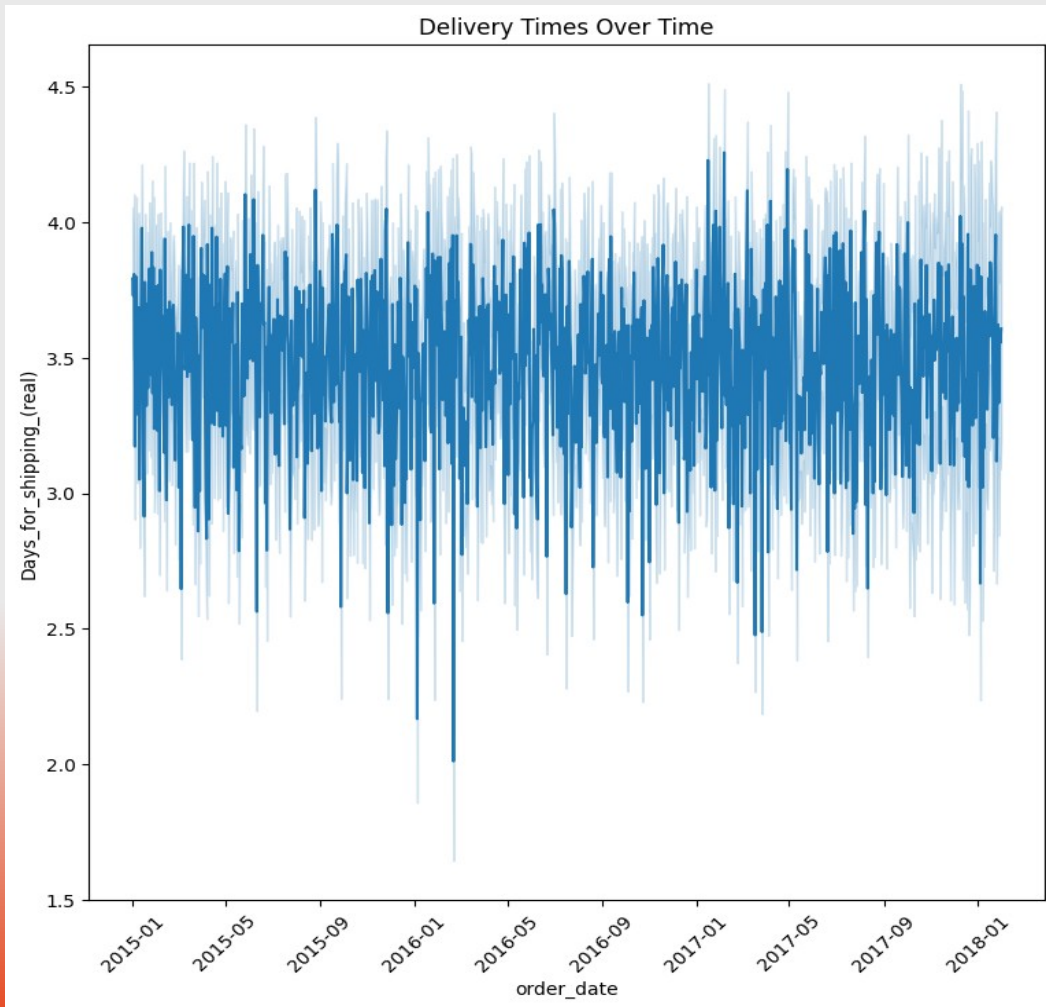
## Actionable Steps

1. **Adjust Shipping Schedules:** Since order and shipping days play a big role in delivery times, focus on making sure shipments go out on days with faster delivery patterns. Tuesday orders typically arrive faster prioritize those orders and allocate more resources on that day.
2. **Target Key Markets:** Some regions have a bigger impact on delivery times e.g. LATAM , Africa and Pacific Asia. By concentrating logistics efforts in these areas, like boosting staff or using different shipping methods, businesses can cut down delays in those key markets.
3. **Improve Order Management:** Use the model's insights to manage orders in real-time. When certain days have higher order volumes, plan ahead by scaling up staffing and transportation to meet demand.

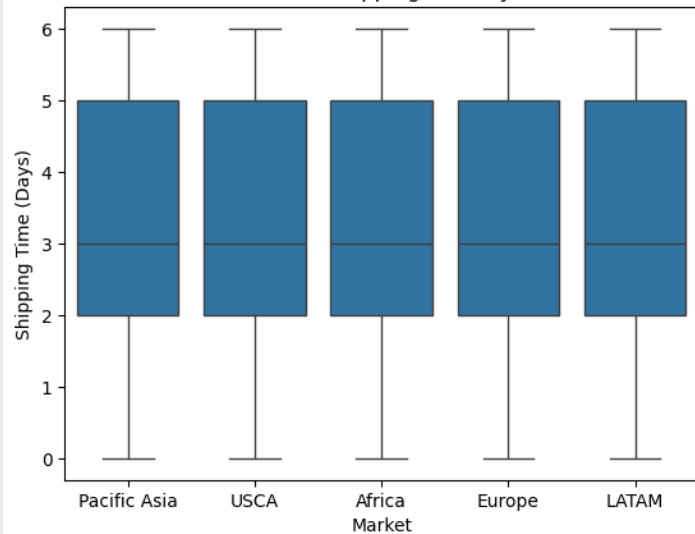
## Future Work

**Factor in External Conditions:** Going forward, adding factors like weather, holidays, or traffic might help fine-tune the model and make delivery time predictions even more accurate.

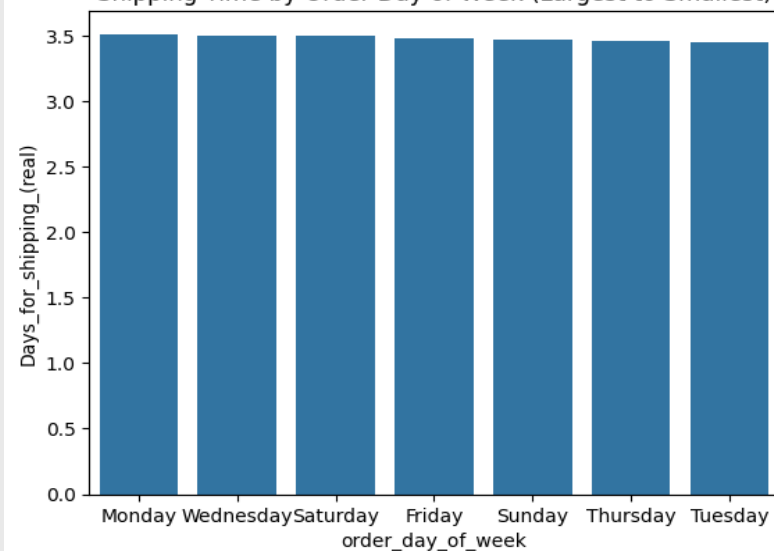
# Appendix



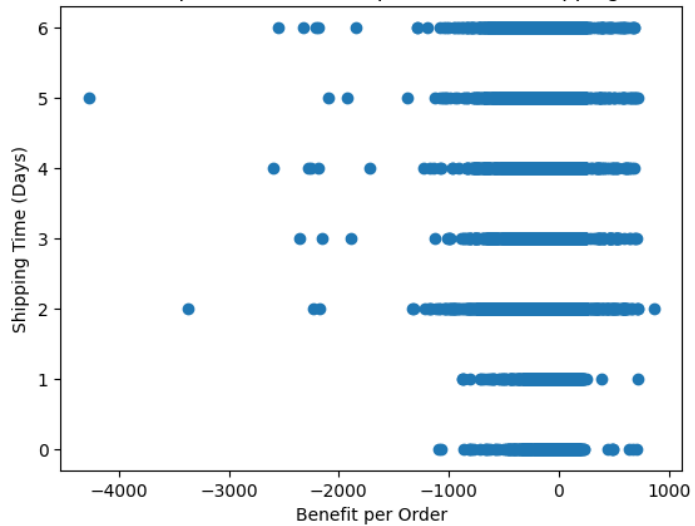
Distribution of Shipping Time by Market



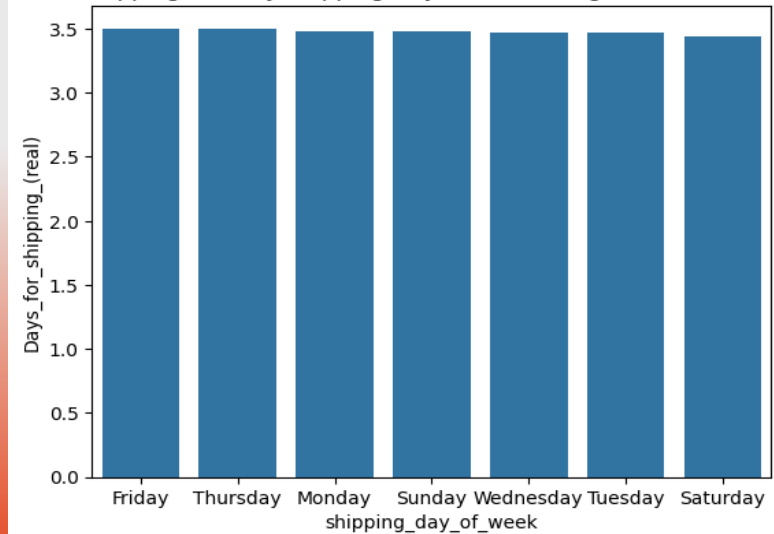
Shipping Time by Order Day of Week (Largest to Smallest)




Relationship Between Benefit per Order and Shipping Time



Shipping Time by Shipping Day of Week (Largest to Smallest)



# Data description tables

Column Name	Description	Assumed Impact on Deliveries
Type	Type of payment.	Certain types of payment may require longer and various verification methods.
Benefit_per_order	Profit margin per order.	Higher benefit orders might be prioritized in shipping to maximize revenue.
Sales_per_customer	Total sales value per customer.	High-value customers may receive better shipping as a loyalty strategy.
Category_Id	Identifier for product categories.	Some categories may be more prone to delays due to size, weight, or regulation.
Customer_City	City of the customer.	Geographic location can impact shipping time due to distance or logistics network.
Customer_Country 	Country of the customer.	International shipments may face delays due to customs or cross-border regulations.
Customer_State	State of the customer.	Regional differences in infrastructure can affect delivery times.
Department_Id	Identifier for the department handling the product.	Different departments may have varying efficiencies, affecting order processing times.
Market	Market segment for the product.	Certain markets may have faster shipping due to better infrastructure or priority.
Order_City	City where the order was placed.	Similar to Customer_City, affects logistics and delivery speed.
Order_Country	Country where the order was placed.	Affects delivery speed and logistics, especially for international orders.
Order_Item_Product_Price	Price of individual order items.	Higher-priced items may receive priority shipping to ensure customer satisfaction.
Order_Item_Profit_Ratio	Profit ratio of individual order items.	Items with higher profit margins might be prioritized in the shipping process.
Order_Item_Quantity	Quantity of items in the order.	Bulk orders may require different shipping methods, potentially causing delays.
Order_Item_Total	Total value of the order.	High-value orders may receive priority in shipping.

## Description table continuation

<b>Order_Region</b>	Region where the order was placed.	Regional logistics networks can impact delivery times.
<b>Order_State</b>	State where the order was placed.	Similar to <code>Customer_State</code> , affects delivery speed due to regional logistics.
<b>Order_Status</b>	Status of the order (e.g., shipped, delivered, pending).	Delays in status change could indicate potential delivery issues.
<b>Product_Price</b>	Price of the product.	Expensive products may be shipped faster to maintain customer satisfaction.
<b>order_date (DateOrders)</b>	Date the order was placed.	Orders placed on weekends or holidays may experience delays.
<b>order_date</b>	Same as <code>order_date (DateOrders)</code> , but formatted differently.	As above, impacts shipping time.
<b>shipping_date</b>	Date the order was shipped.	The difference between order and shipping date indicates processing time.
<b>order_day_of_week</b>	Day of the week the order was placed.	Orders placed on certain days may be processed slower (e.g., weekends).
<b>order_month</b>	Month the order was placed.	Peak seasons (e.g., holidays) may lead to longer processing times.
<b>shipping_day_of_week</b>	Day of the week the order was shipped.	Similar to <code>order_day_of_week</code> , affects delivery speed.
<b>shipping_month</b>	Month the order was shipped.	Shipping during peak seasons may result in delays.
<b>Orders_Made_That_Day</b>	Total number of orders made on that day.	High order volumes can strain logistics, leading to delays.
<b>Orders_Shipped_That_Day</b>	Total number of orders shipped on that day.	High shipping volumes could lead to delays if logistics capacity is exceeded.

## **Contact Information**

Otieno Calvin

Data Scientist, Techstride Equinova

Email: Techstrideequinova@gmail.com

Phone: +254 11426\*\*03

Github: [github.com/Otieno-Calvin](https://github.com/Otieno-Calvin)