DATACO SUPPLY CHAIN - LATE DELIVERIES FORECASTING

KARTHIK SAI TWARAKAVI

Domain Knowledge:

The supply chain Industry mainly stands on the pillars of Planning, Sourcing, Production, Distribution, and Returns.

As the DataCo is facing problems on Late deliveries, we would be focusing on Distribution.

Distribution: Once products are manufactured, they need to be distributed to retailers or customers. It involves in multiple steps.

- **Warehouse Management**: After production, goods are stored in warehouses until they are needed. This involves organizing and maintaining inventory in a way that facilitates easy access and efficient dispatch.
- **Order Processing**: When a retailer or customer places an order, this triggers the order processing system. This includes order confirmation, picking, packing, and preparing for shipment.
- **Transportation Management**: This step involves selecting the most efficient mode of transport (like truck, train, ship, or air) based on cost, distance, and product type. It also includes route planning to minimize transit time and cost.
- **Logistics Coordination:** Coordination among various stakeholders (like suppliers, carriers, warehouses, and customers) is crucial. This includes scheduling pickups and deliveries, tracking shipments, and managing logistics service providers.
- **Delivery and Last-Mile Logistics**: The final delivery of products to the customer's doorstep or the retailer. In last-mile logistics, the focus is on delivering the product most quickly and cost-effectively.

- **Reverse Logistics**: Handling returns, exchanges, or recycling of products, if necessary, is also part of distribution.

What are the Benefits to the Company by concentrating on Distribution?

• **Customer Satisfaction**: Effective distribution translates to timely delivery, and accordingly increases customer satisfaction as well as loyalty compared with the traditional one.

• **Cost Reduction**: Optimum distribution will lead to reduction of transportation and storage costs.

• **Market Expansion**: Modern distribution networks enable organizations to spread their market base.

• **Supply Chain Visibility**: New distribution practices that are more advanced create improved situations regarding tracking as well as transparency, assisting in better decision-making.

• **Inventory Management**: Optimal levels of inventory holds can be sustained through carefully developed distribution strategies and this reduces holding costs.

Risks to the Company:

Transportation Risks: Crucial links of the supply chain can be disrupted due to delays, accidents or damages during transit.

Warehousing Risks: Inefficient warehouse management can result to its highest costs fulfillment delays, therefore increasing the cost of order processing.

Compliance and Regulatory Risks: Not adhering to regulations for transportation and trade also results in legal consequences leading to fines.

Technology Risks: Technology dependency for logistics and order processing could be a threat if systems go down or are hacked.

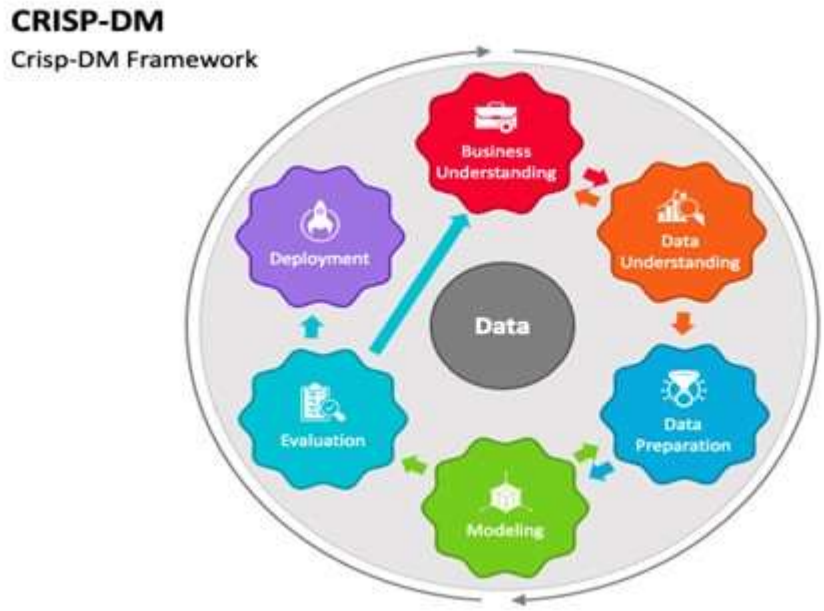
Market Demand Fluctuations: Poor matching of the strategies for distribution with demand can result in instances of over-ordering or stockouts.

Environmental Risks: Environmental externalities associated with the distribution activities and particularly transport may trigger regulatory alteration which will impose an economic burden on these firms.

Global Risks: Risk is risks such as geopolitical issues and currency fluctuations among others to global supply chains.

This Problem Statement follows the analogy of CRISP-DM, Which Involves

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



BUSINESS UNDERSTANDING

DataCo is a global e-commerce company committed to ensuring timely deliveries of goods to customers worldwide. Our primary goal is to uncover the reasons behind late deliveries and take proactive measures to reduce the risk of delays, ultimately preserving our valued customer base. Currently, 54.8% of our deliveries experience delays, which concerns our business.

To address this issue, we are leveraging data analysis of our delivery operations. We aim to identify the root causes of these delays and implement effective strategies. These strategies may include improving logistics in specific regions or enhancing communication with customers who might face potential disruptions. By pinpointing regions with the highest percentage of late deliveries, we can tailor our interventions and allocate resources strategically for maximum impact.

Our analysis identified that product categories such as Cleats, Men's footwear, and Women's apparel are among the top contributors to late deliveries. This information guides our focus on improving delivery timelines for these specific categories.

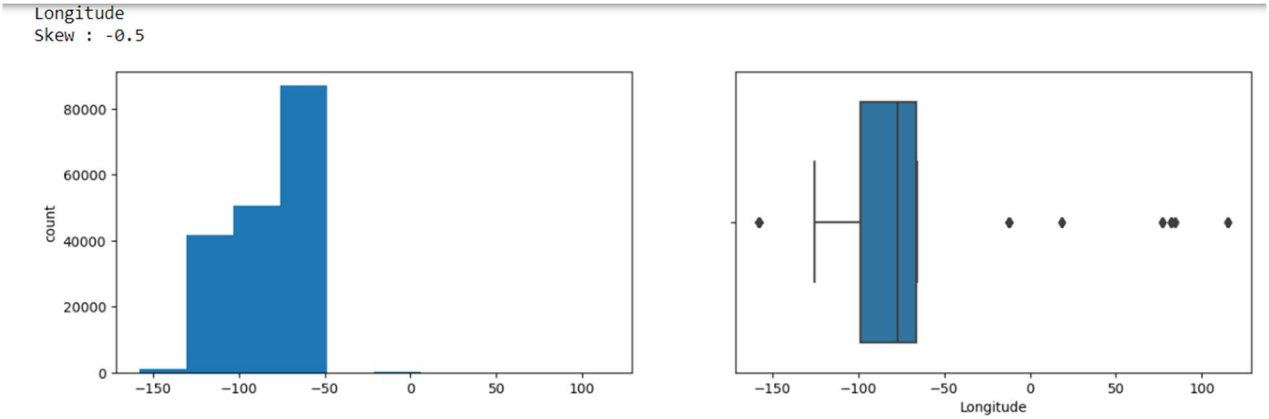
DATA UNDERSTANDING

Data was collected from the Kaggle. It has 53 columns and 1,80,519 rows. Of 53 Variables Late_delivery_risk was identified to be the Dependent variable and others are independent.

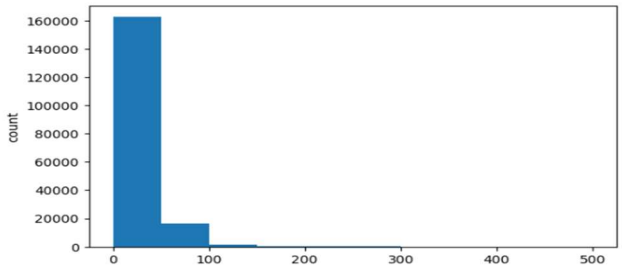
The dataset has included different types of Data types of which 21 consist of categorical variables, 27 numerical variables, and 2 date types. There are a few variables Customer Email, Customer Password must be dropped from analysis as these are confidential and would not be helpful for our analysis.

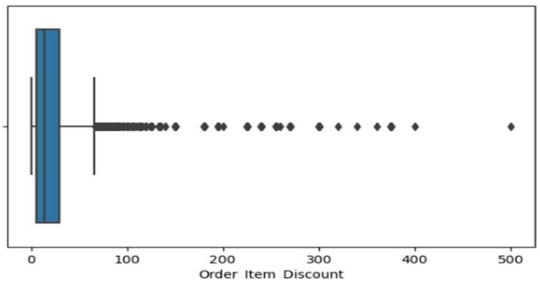
Exploratory Data Analysis

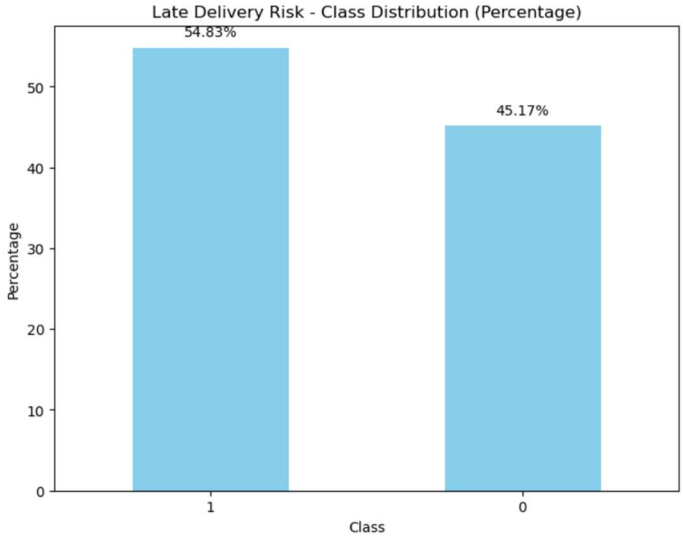
- It is observed that 100% of the Product Descriptions and 86.23% of the Order zip code have null values. 0.004% of Customer last name, 0.001% of customer zip code has null values.
- Distribution of Dependent variable Late delivery risk is imbalanced.



Order_Item_Discount
Skew : 3.04

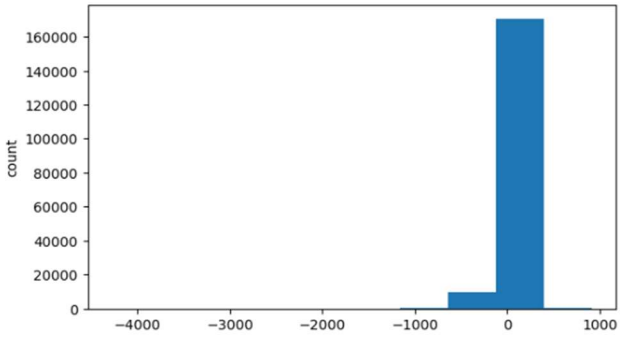


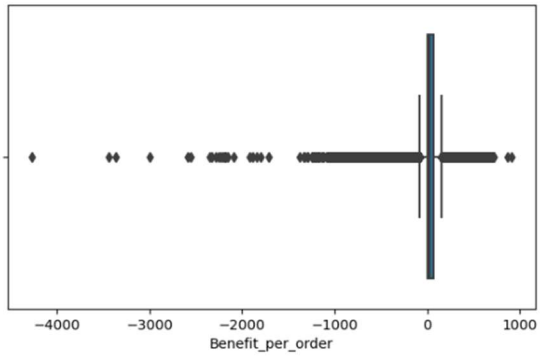




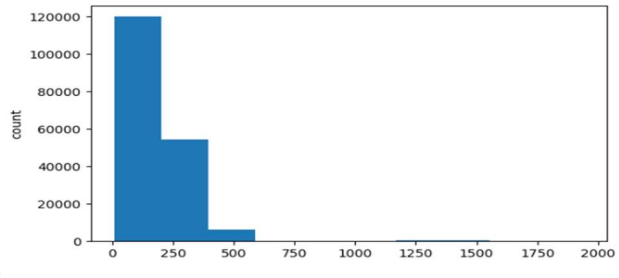
Univariate Analysis:

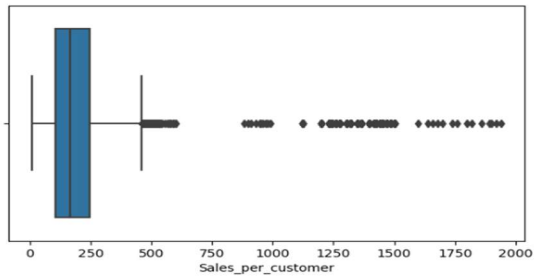
Benefit_per_order
Skew : -4.74



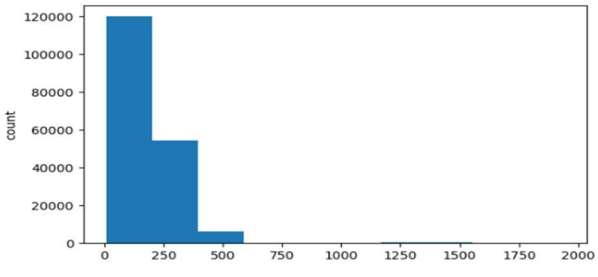


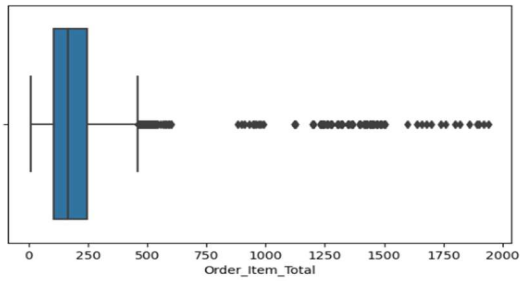
Sales_per_customer
Skew : 2.89

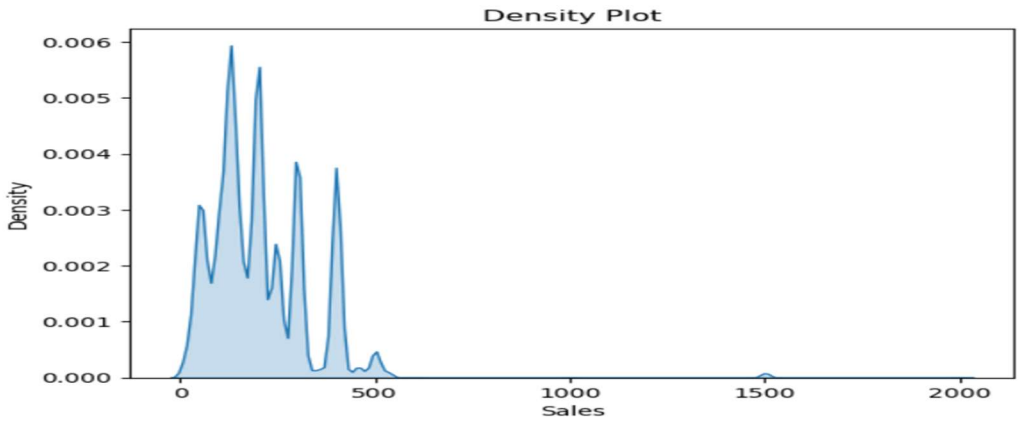




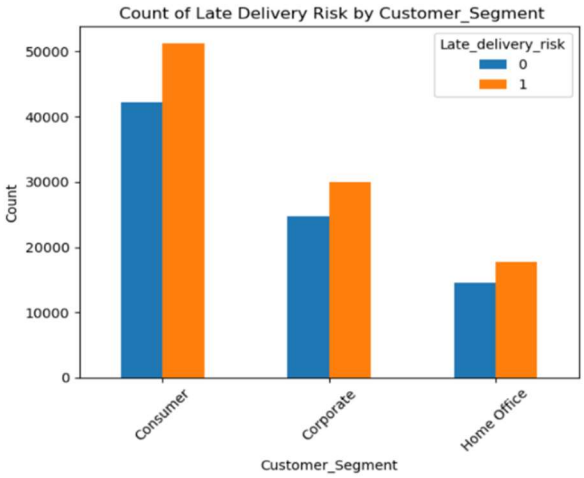
Order_Item_Total
Skew : 2.89

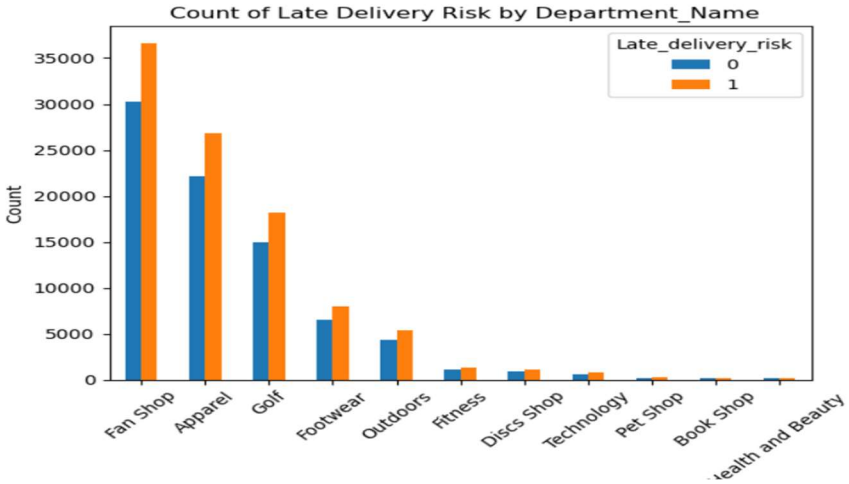




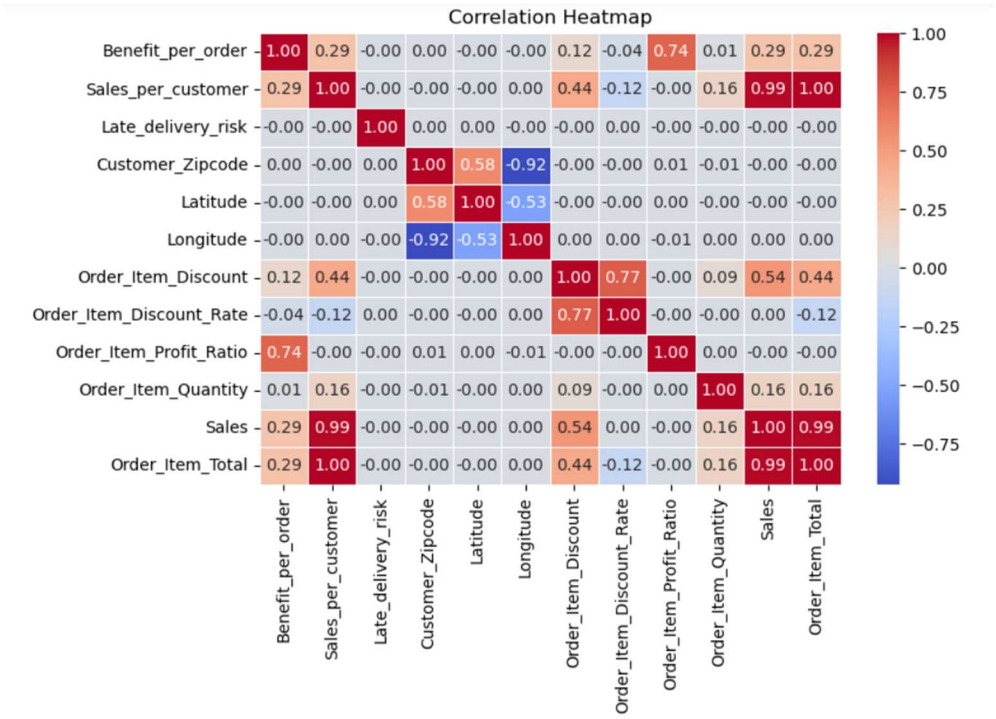


Bi-Variate Analysis:





Multi-Variate Analysis



From the above figures we can identify that many variables are both positively and negatively skewed and there is an imbalance in the dependent variable.

DATA PREPARATION

Dealing with Null Values: As 100% of the Product Descriptions and 86.23% of the Order zip code have null values they can be dropped from our Data. 0.004% of Customer last name, and 0.001% of customer zip codes have null values that can be dropped as they are negligible and will not affect the analysis.

Dealing with categorical variables:

For the Variables- Type, Market, Customer_Segment, Shipping_Mode, Customer_Country, and Order_Status as very few unique values and are not ordinal, we chose to perform **One Hot encoding.** Ensures that each category is represented as a binary vector with only one "hot" or "on" element (1) and all other elements set to "off" or "0"

For the Variables- Category_Name, Customer_State, Department_Name, Order_Region opted for **Binary encoding** as there are many Unique values for each variable, Binary encoding is ideal in such cases. It converts the variables into the binary format with combinations of 0 and 1.

For the Variables- Customer_City, Order_City, Order_Country, Order_State chose to do **Bayesian Mean Encoding** as these variables involve high cardinality, we encode categories concerning a binary target variable. It calculates the mean of the target variable for each category and then blends it with the overall mean of the target variable.

Frequency encoding for the variable Product_Name, where It variable with the frequency (or count) of that category in the dataset.

Dealing with Imbalance in Dependent Variable

As it is observed that there is a clear imbalance of the dependent variable with 54.8% of 1s and 45.17% of 0s, this can be taken care of by the Method SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic observations of the minority class.

MODELING

Feature importance was considered in the model, and identified top 5 features using a Random Forest Classifier and identified Latitude, longitude, order city, Shipping_Mode_Standard Class, and order state.

Model Selection and Tuning:

Chosen algorithms: Random Forest and XGBoost classifiers were selected due to their suitability for classification tasks and ability to handle complex relationships in the data.

Hyperparameter tuning: GridSearchCV was employed to systematically explore different hyperparameter combinations and identify the best-performing configurations for each model.

Parameter grids: The following parameter grids were used for tuning:

Random Forest: n_estimators, max_depth, min_samples_split

XGBoost: n_estimators, max_depth, learning_rate

Cross-validation: 5-fold cross-validation was used within GridSearchCV to assess model performance and prevent overfitting.

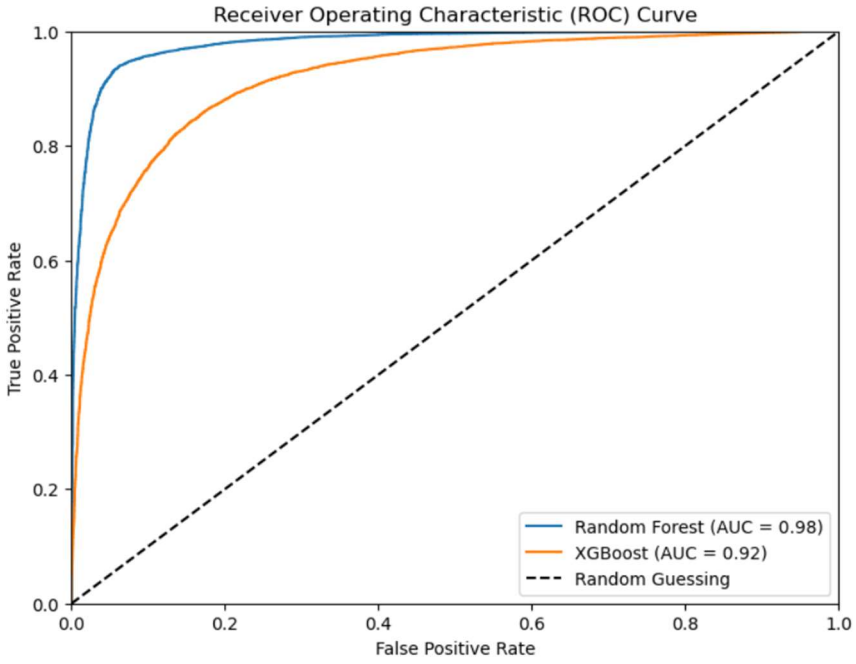
Model Training and Evaluation:

Trained models: Both models were trained using the optimized hyperparameters determined through GridSearchCV.

EVALUATION

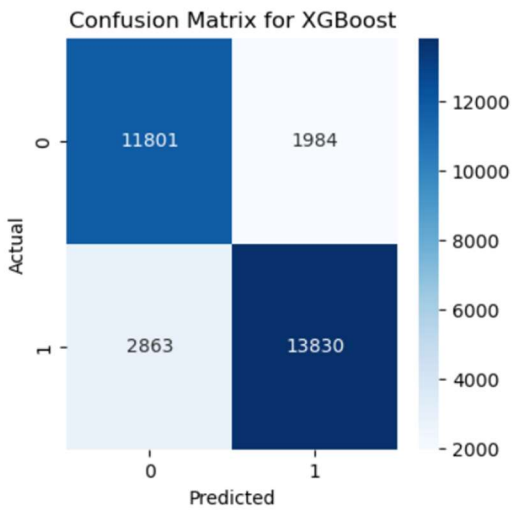
Evaluation metrics: Accuracy, classification reports, and confusion matrices were used to evaluate model performance on the test set.

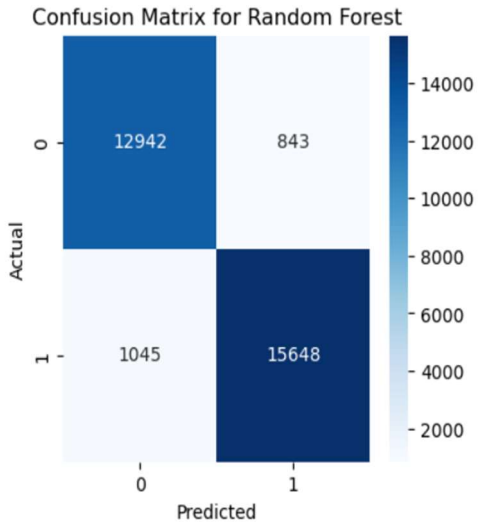
ROC Curve: The curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) at different thresholds.



From the ROC Curve, we can say that Random Forest did really well with AUC = 0.98 compared to XGBoost AUC =0.92.

Confusion Matrix: This is a table used to describe the performance of a classification model by displaying the actual versus predicted classifications.





Metrics:

Metric	Random Forest	XGBoost
Precision	94.89%	87.45%
Recall	93.74%	82.85%
F1 Score	94.31%	85.09%

Random Forest:

Precision (94.89%): High precision indicates that the model has a low false positive rate. When it predicts a class, it is correct most of the time.

Recall (93.74%): High recall shows that the model is good at identifying all relevant instances within the dataset.

F1 Score (94.31%): The high F1 score, which is the harmonic mean of precision and recall, suggests a well-balanced model in terms of both precision and recall.

XGBoost:

Precision (87.45%): This is lower compared to Random Forest, indicating a higher rate of false positives.

Recall (82.85%): This lower recall value suggests that XGBoost misses a higher proportion of relevant instances than Random Forest.

F1 Score (85.09%): The lower F1 score, relative to Random Forest, indicates that XGBoost is less balanced between precision and recall.

These results suggest that Random Forest is the more effective model in terms of both identifying relevant cases accurately and maintaining a balance between precision and recall.

ACTIONABLES

- Implement automated notifications for orders with a high risk of late delivery.
- Set up a warehouse in proximity to areas where late deliveries occur most frequently to optimize logistics.
- Establish effective communication channels between different departments to facilitate timely information sharing.
- Arrange for alternative shipment options if late deliveries are consistently caused by a specific carrier or transportation provider.
- Consider implementing dynamic pricing strategies, where customers for high-risk orders might pay slightly more to ensure on-time delivery, incentivizing prioritization.

- Refine demand forecasting methods to better predict customer demand and resource allocation for improved capacity planning.
- Maintain timely communication with customers regarding any potential or actual late deliveries to manage expectations and provide transparency.