BIA | BOSTON INSTITUTE OF ANALYTICS ®

E-Commerce Product Delivery Prediction

PRESENTED BY:
Aryan Thube

Agenda

Introduction

Problem Statement

Methodology

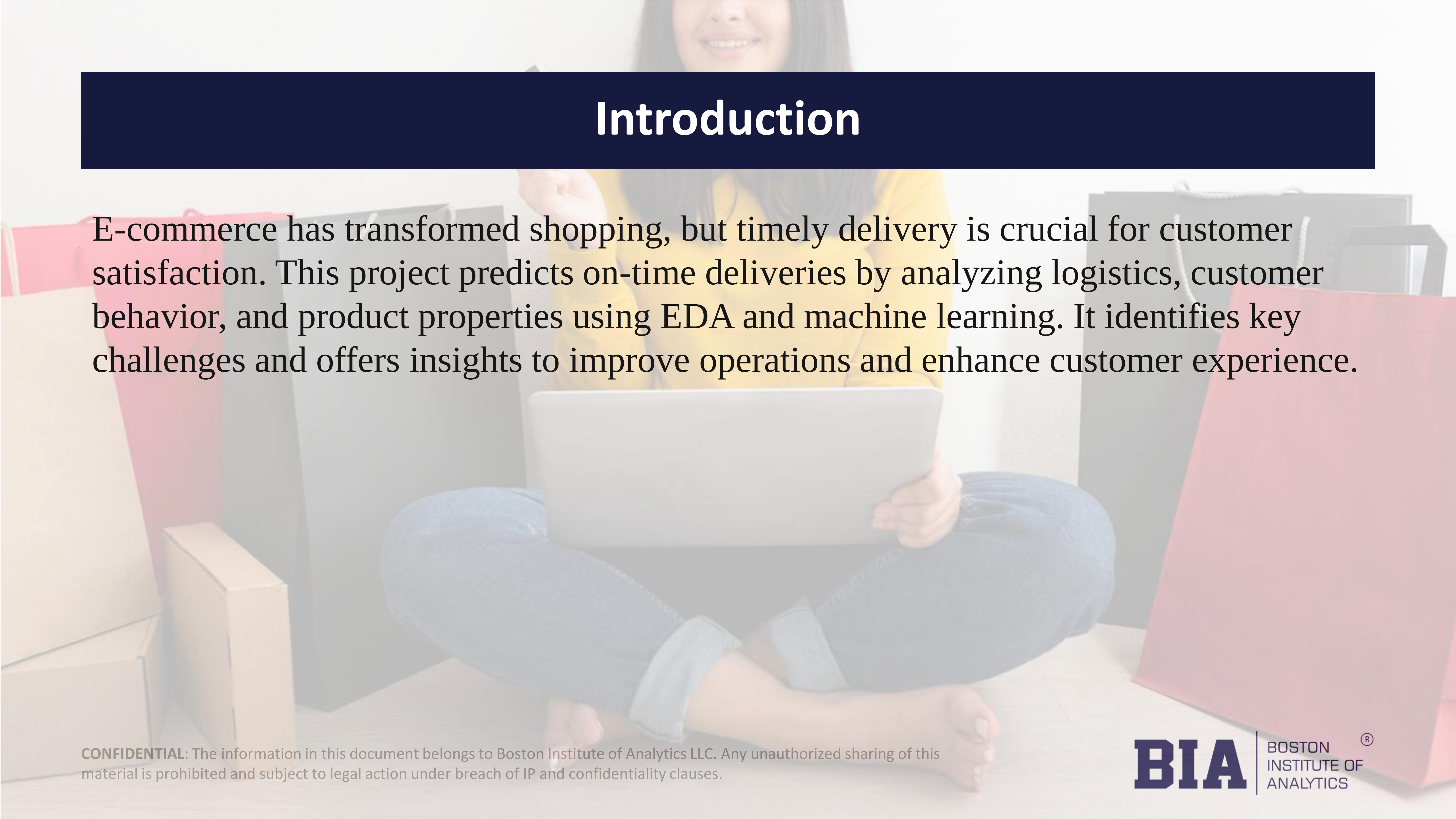
Data Preprocessing

Exploratory Data Analysis (EDA)

Model Building

Model Evaluation

Conclusion

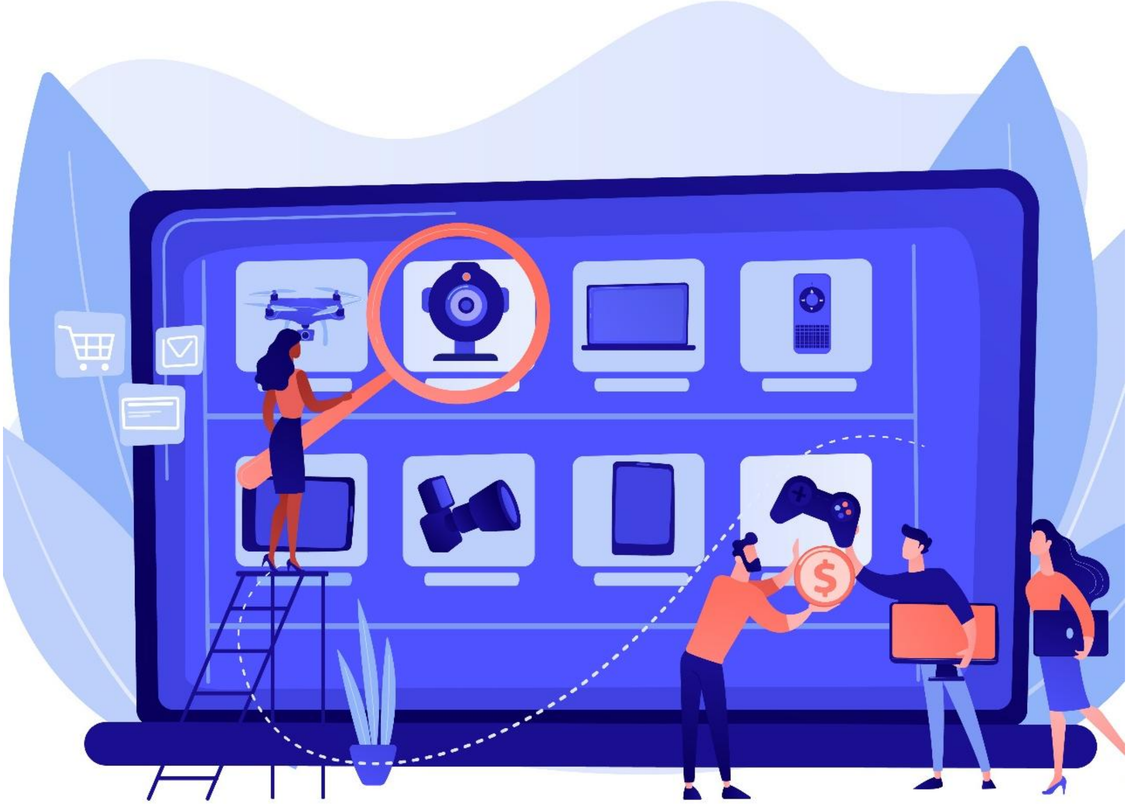


Introduction

E-commerce has transformed shopping, but timely delivery is crucial for customer satisfaction. This project predicts on-time deliveries by analyzing logistics, customer behavior, and product properties using EDA and machine learning. It identifies key challenges and offers insights to improve operations and enhance customer experience.

Problem Statement

E-commerce businesses face a critical challenge in ensuring timely delivery of products. Delayed shipments not only impact customer satisfaction but also business reputation. The aim of this project is to predict whether an e-commerce product will be delivered on time by analyzing factors like customer behavior, product properties, and logistics.



Methodology

Data Preprocessing: Cleaned and prepared data, handling missing values, duplicates, and irrelevant columns.

Exploratory Data Analysis (EDA): Investigated distribution of variables, customer behaviour, and logistics factors using visualizations.

Feature Engineering: Transformed categorical variables using label encoding.

Model Building: Deployed machine learning models like Random Forest, Decision Tree, Logistic Regression, and KNN to predict delivery outcomes.

Model Evaluation: Assessed models based on accuracy, confusion matrix, and classification reports.

Data Description

ID: ID Number of Customers

Warehouse_block: The company has a large warehouse divided into blocks A, B, C, D, E

Mode_of_Shipment: The company ships products via Ship, Flight, and Road

Customer_care_calls: Number of calls made for shipment inquiries

Customer_rating: Ratings provided by customers, ranging from 1 (Worst) to 5 (Best)

Cost_of_the_Product: Cost of the product in US Dollars

Prior_purchases: Number of prior purchases

Product_importance: Product categorized as low, medium, or high importance

Gender: Male or Female

Discount_offered: Discount offered on the specific product

Weight_in_gms: Weight of the product in grams

Reached.on.Time_Y.N: Target variable: 1 indicates the product was delivered on time, 0 indicates it was NOT delivered on time

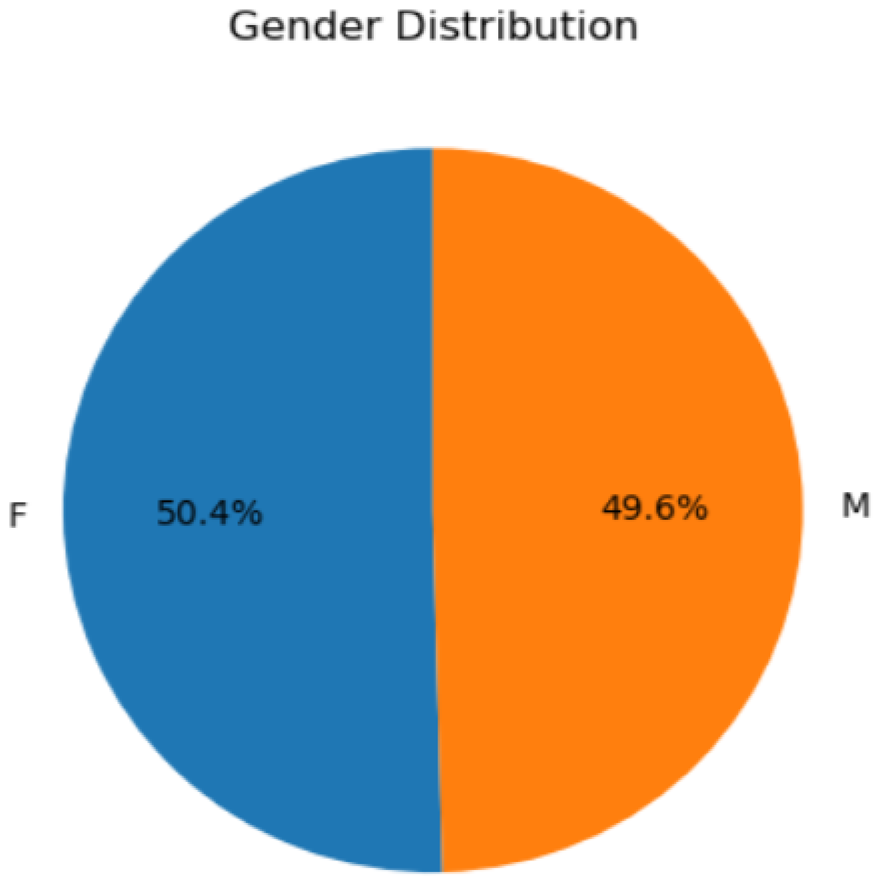
Data Preprocessing

1. **Data Cleaning:** Verified there were no missing values, checked for duplicate records as none were found.
2. **Column Transformation:** Dropped the irrelevant ID Column
3. **Encoding Categorical Data into Numerical data:** 1.Warehouse_block
   2. Mode_of_Shipment
   3.Product_importance
   4.Gender

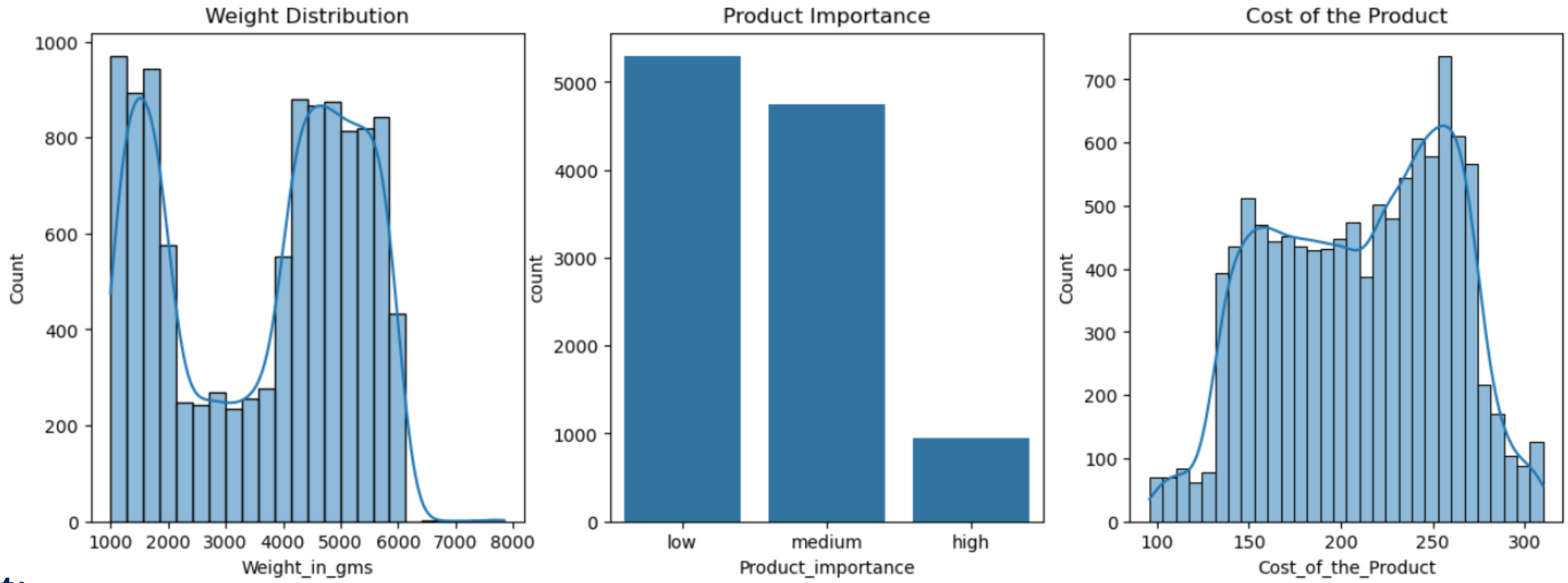
Exploratory Data Analysis

Customer Gender Prediction

Insight: The dataset has an equal number of both male and female customers, with percentages of 49.6% and 50.4%, respectively.



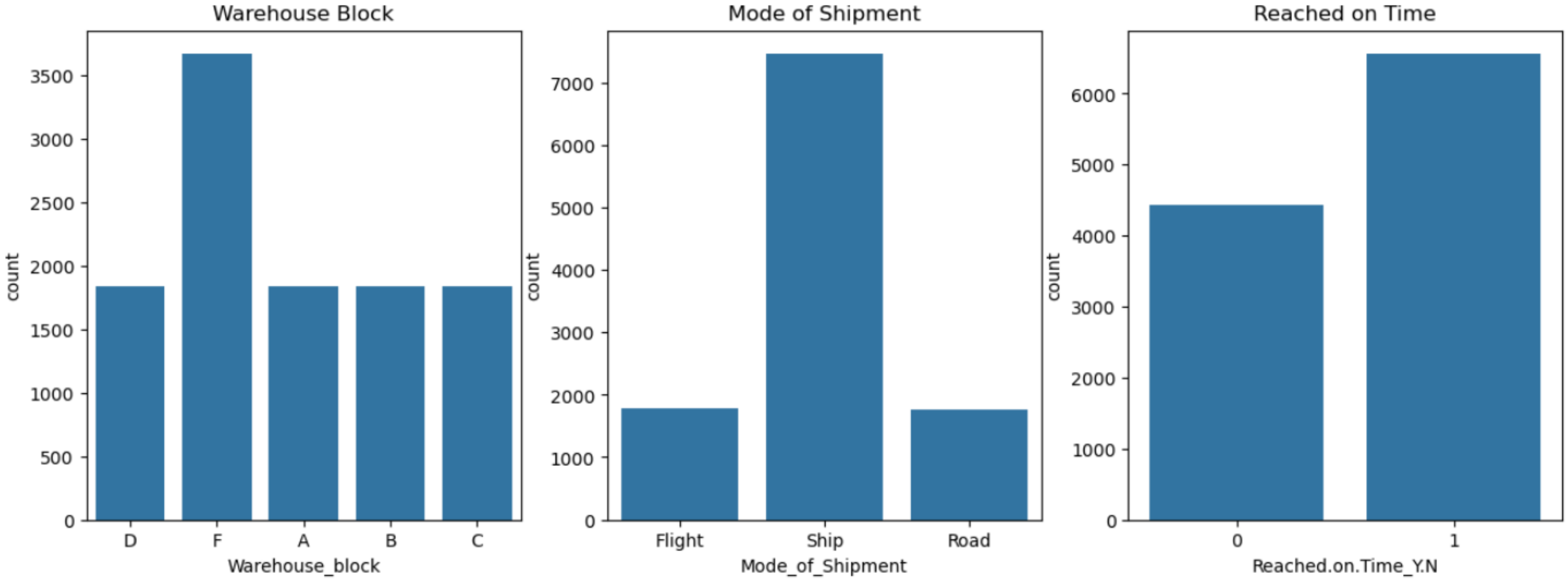
Product Properties



Insight:

1.Most products weigh between 1000-2000 grams and 4000-6000 grams.
2.Products are mostly categorized as having low or medium importance.
3. The majority of products are priced between 150-275 dollars.
This shows the company primarily sells lightweight, moderately important, and mid-priced products.

Logistics



Insight:

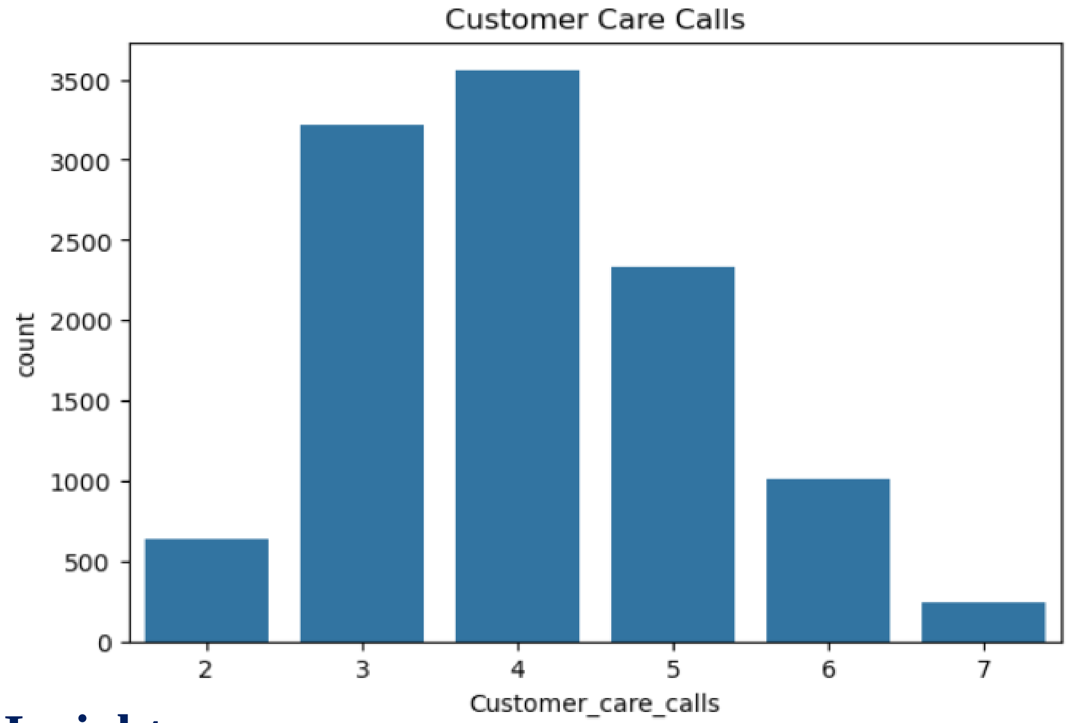
1.Warehouse F handles the most products (around 3500), while other warehouses manage fewer.

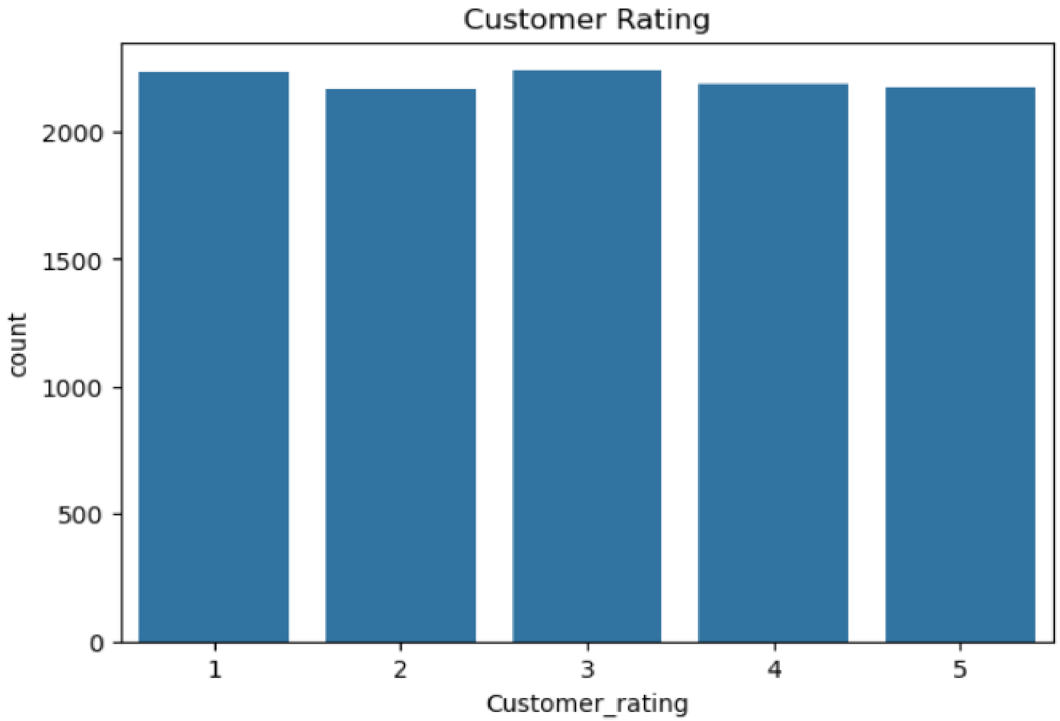
2.Most products are shipped by sea, with around 2000 transported by flight and road.

3.More products are delivered on time than late.

Warehouse F is likely located near a seaport, given its high product volume and reliance on shipping

Customer Experience

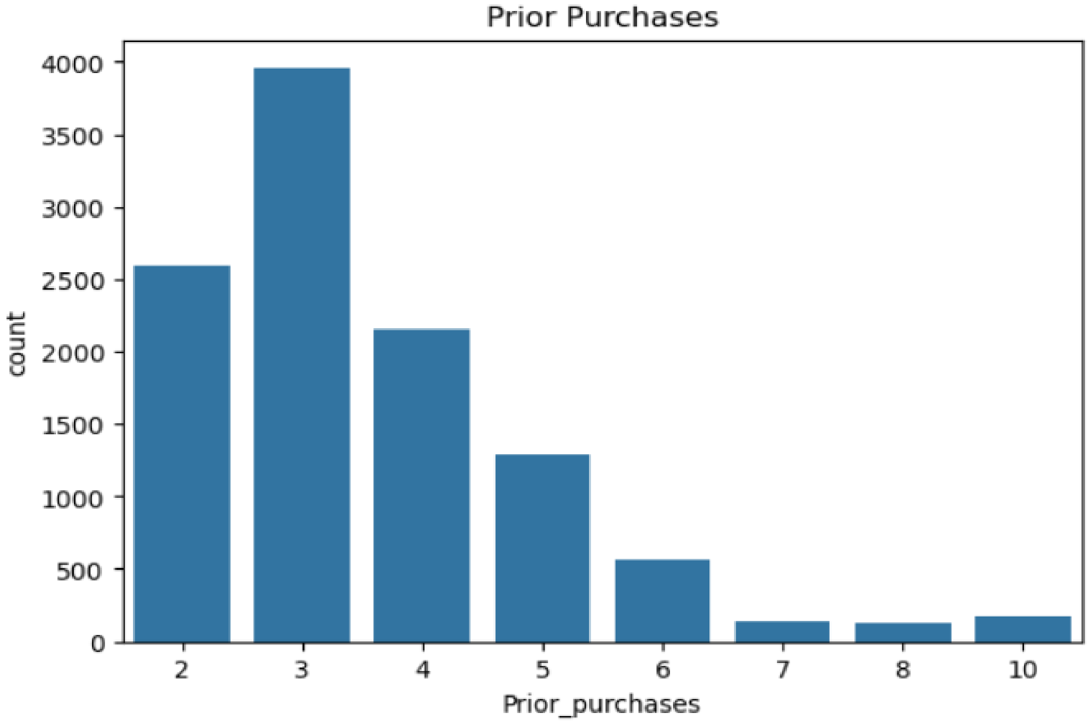


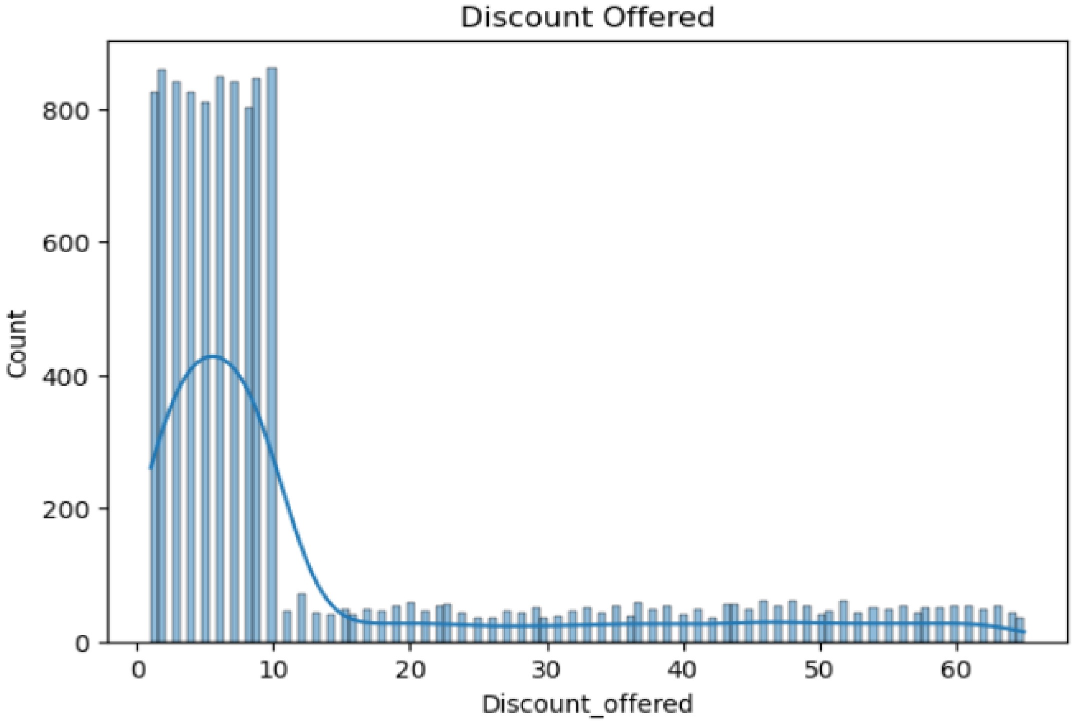


Insight:

1. Most customers make 3-4 customer care calls, possibly due to delivery issues.
2. Customer ratings are evenly distributed, but there are slightly more 1-star ratings, indicating some dissatisfaction.

Customer Experience

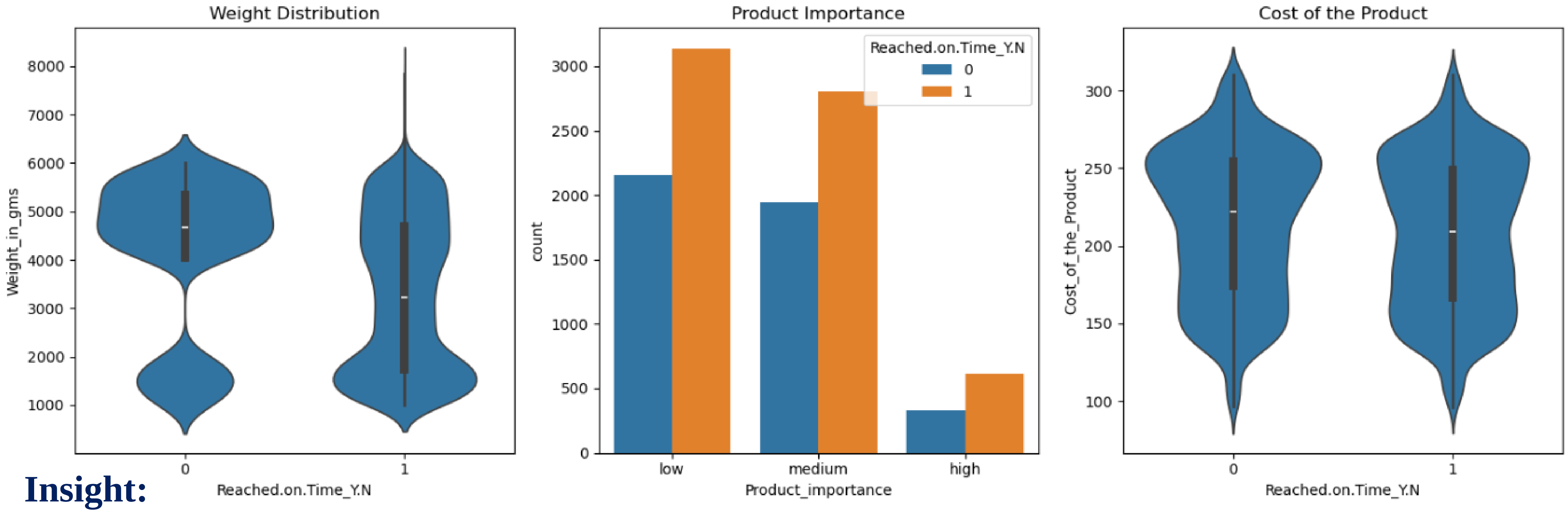




Insight:

1. Majority of customers have made 2-3 prior purchases, suggesting repeat customers are satisfied.
2. Most products have a 0-10% discount, showing limited discount offering.

Product Properties and Product Delivery

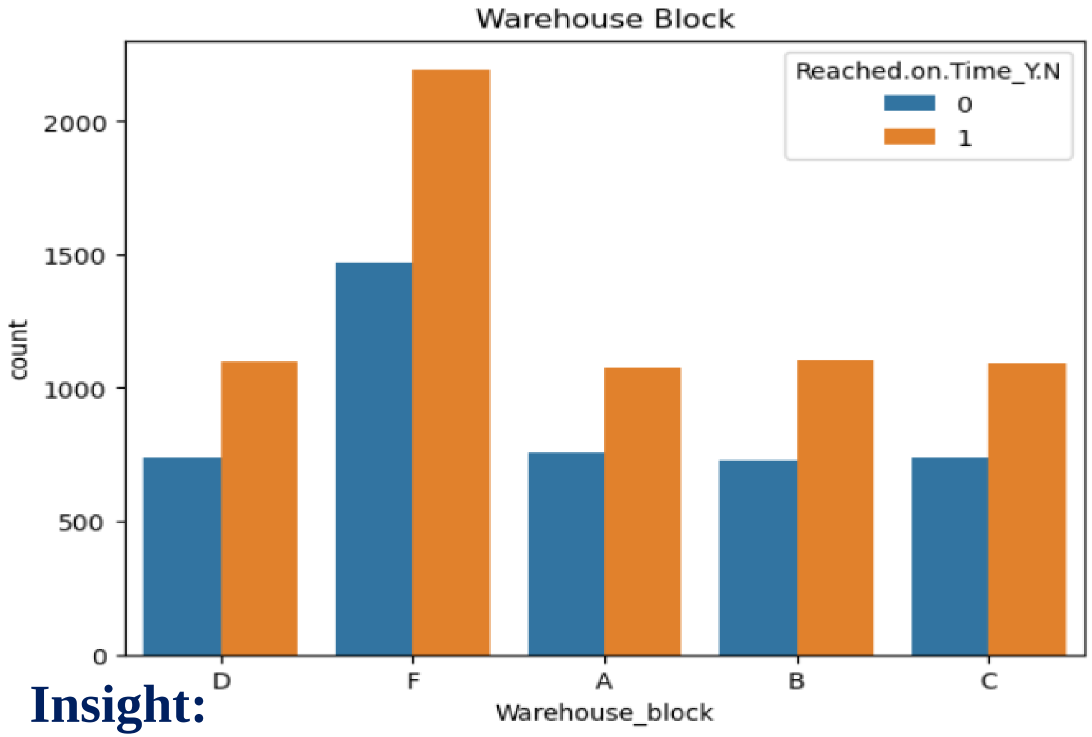


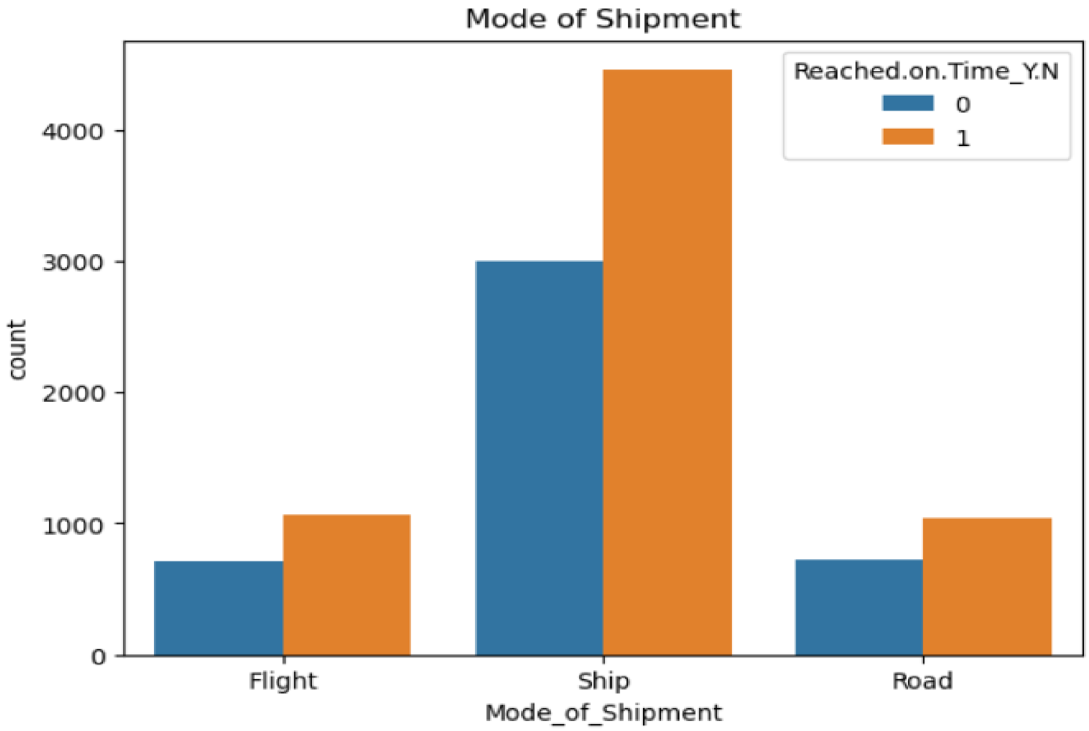
Insight:

1. Product weight affects delivery timeliness. Products over 4500 grams are often late, while those in the 2500-3500 gram range are delivered on time.
2. Product importance does not significantly impact delivery timeliness.
3. Products costing over $250 tend to be delivered late.

Product weight and cost are key factors influencing delivery timeliness.

Logistics and Product Delivery

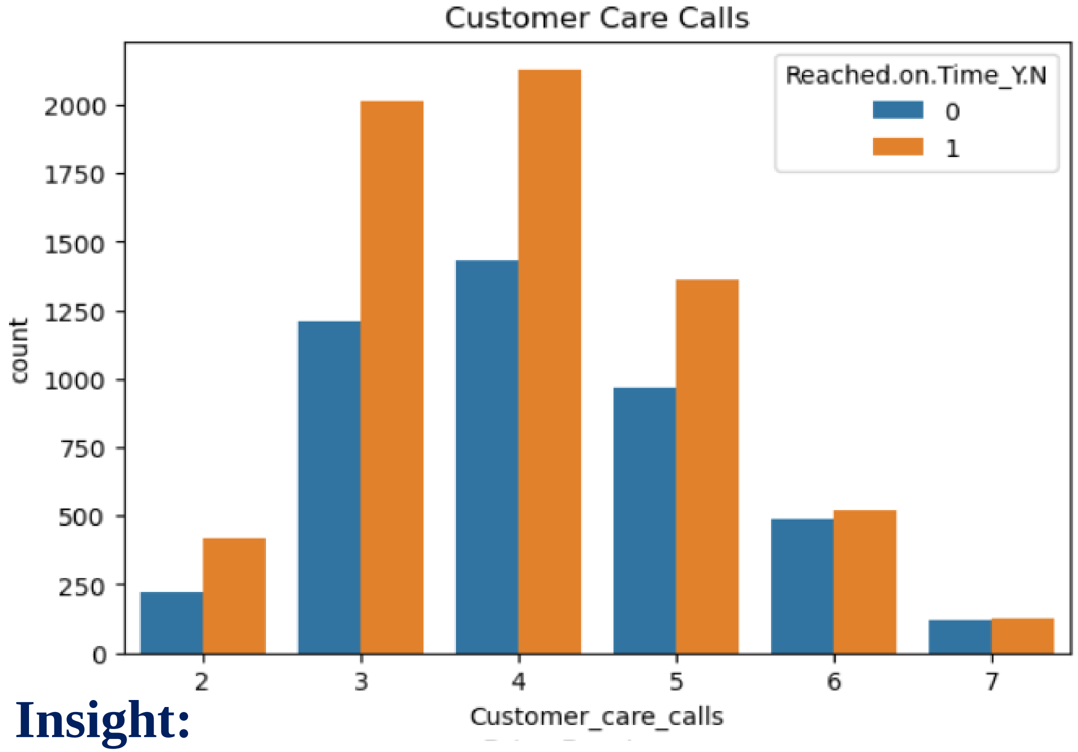


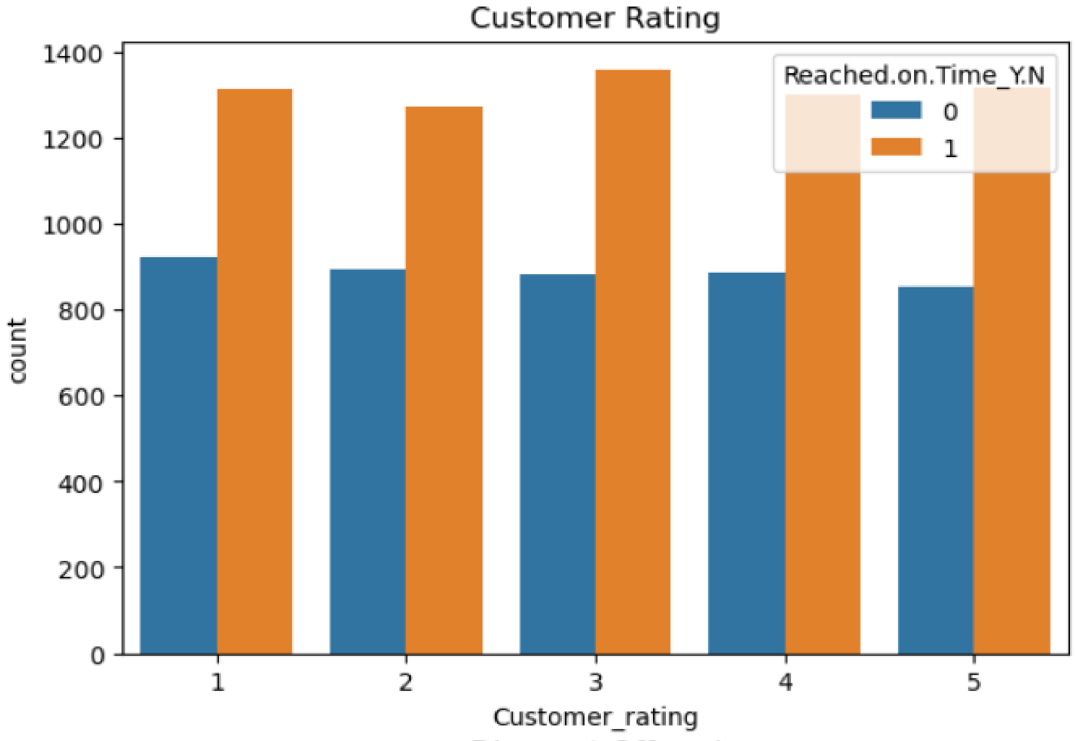


Insight:

1. Most products are shipped from warehouse F, likely near a seaport, and are primarily transported by ship.
2. Despite the high shipping volume, there is a consistent difference between on-time and late deliveries across all warehouses and shipping methods.
3. This consistency suggests that warehouse location and shipping method don't significantly impact delivery timeliness

Customer Experience and Product Delivery

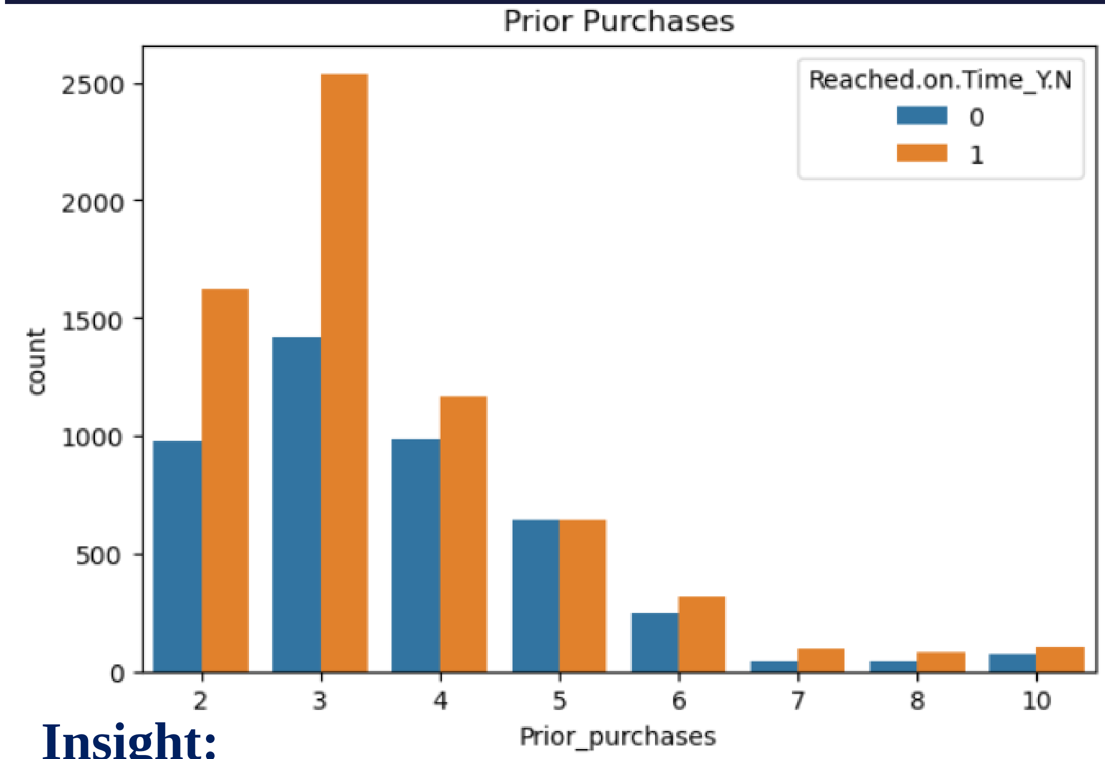

Customer Care Calls

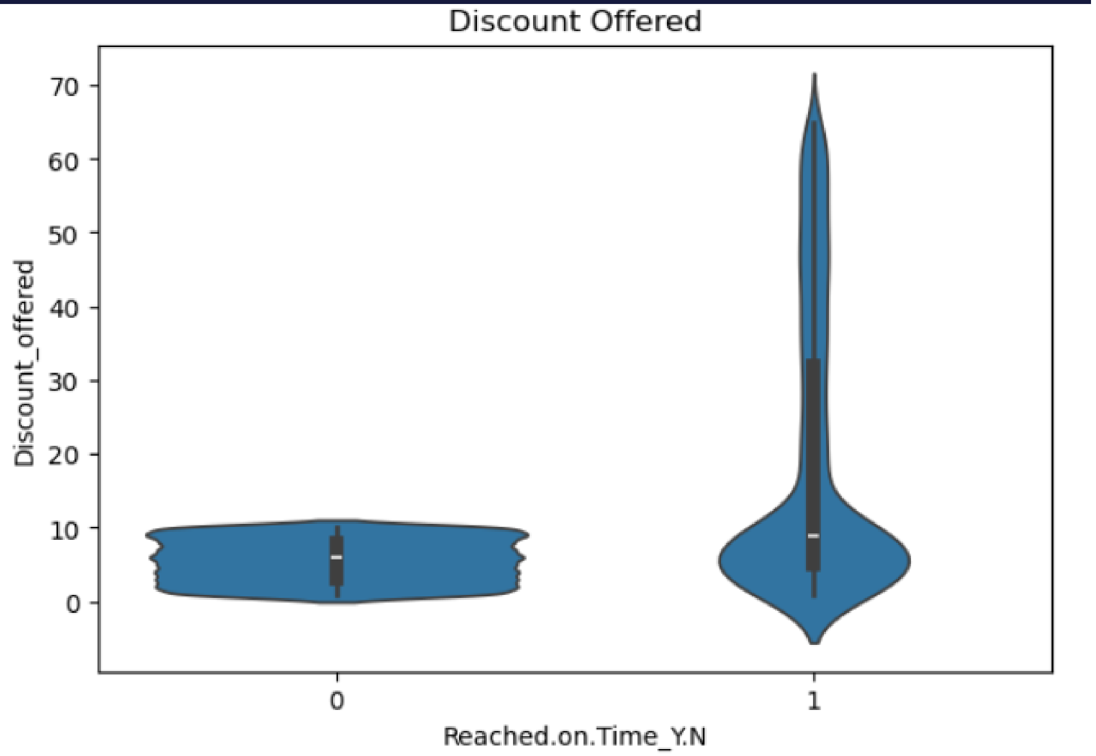

Customer Rating

Insight:

1. More customer care calls are made when deliveries are late.
2. Customers with higher ratings tend to receive their products on time.

Customer Experience and Product Delivery





Insight:

1. Repeat customers generally get their products on time.
2. Products with discounts under 10% are often delivered late, while those with higher discounts are delivered on time more often.

Correlation Matrix Heatmap

Insight:

- **Variables with Strong Correlation:**

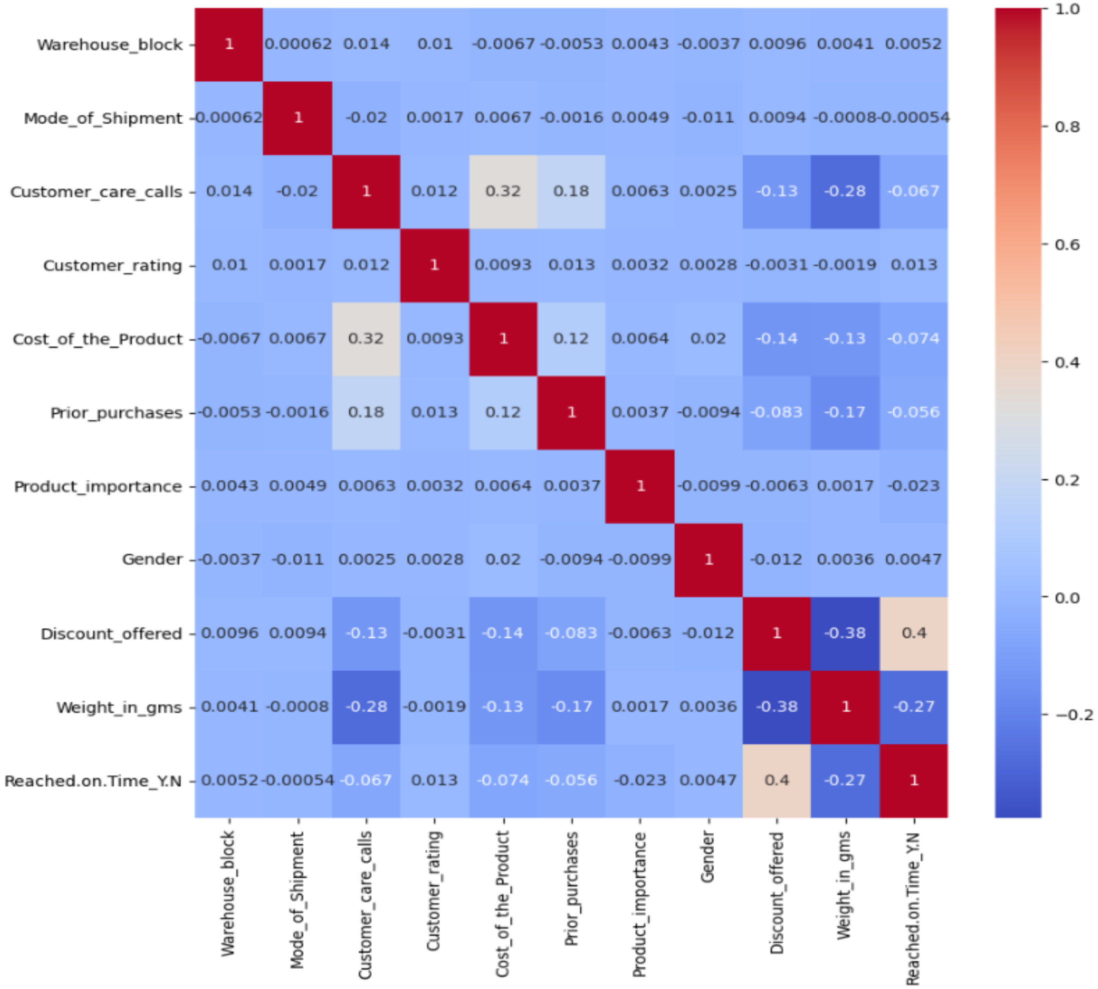
-Discount_offered & Reached.on.Time_Y.N (0.4): Higher discounts may impact delivery times due to logistical challenges.

-Cost_of_the_Product & Customer_care_calls (0.32): Expensive products lead to more customer inquiries.

- **Variables with Negative Correlation:**

-Discount_offered & Weight_in_gms (-0.38): Heavier products get lower discounts due to higher costs.

-Weight_in_gms & Reached.on.Time_Y.N (-0.27): Heavier products are less likely to arrive on time, likely due to delays in shipping.

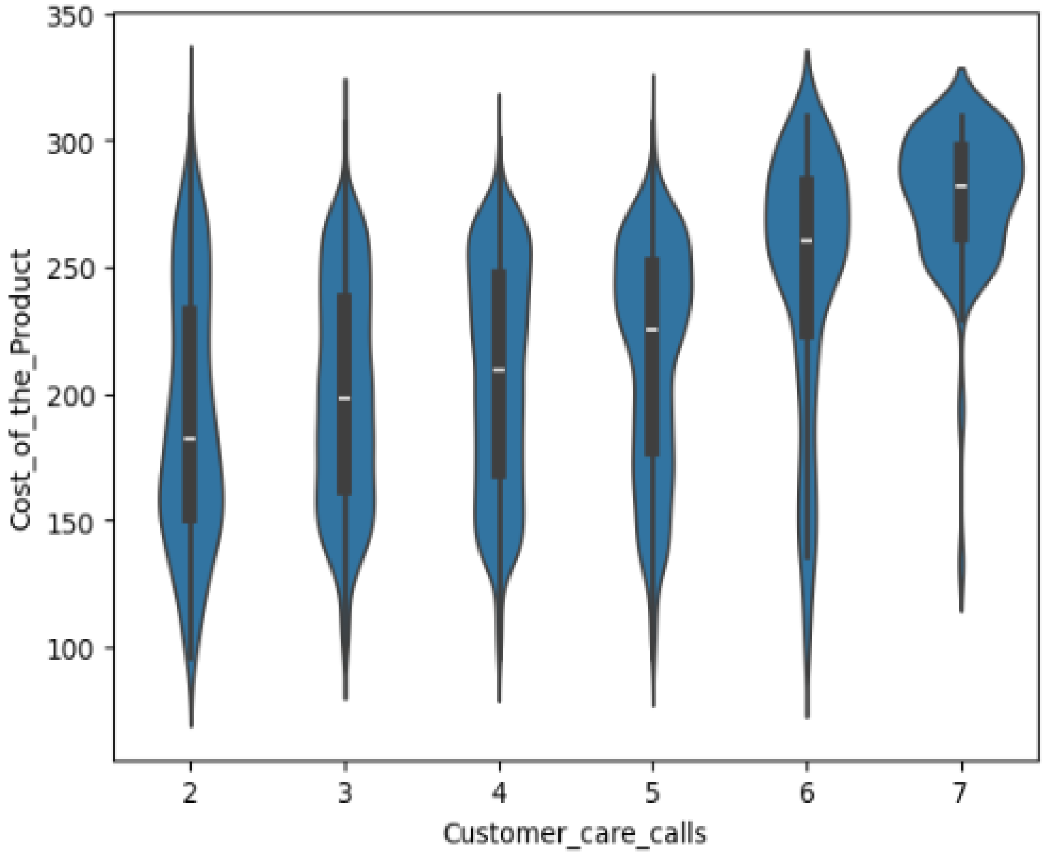


	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
Warehouse_block	1	0.00062	0.014	0.01	-0.0067	-0.0053	0.0043	-0.0037	0.0096	0.0041	0.0052
Mode_of_Shipment	0.00062	1	-0.02	0.0017	0.0067	-0.0016	0.0049	-0.011	0.0094	-0.0008	-0.00054
Customer_care_calls	0.014	-0.02	1	0.012	0.32	0.18	0.0063	0.0025	-0.13	-0.28	-0.067
Customer_rating	0.01	0.0017	0.012	1	0.0093	0.013	0.0032	0.0028	-0.0031	-0.0019	0.013
Cost_of_the_Product	-0.0067	0.0067	0.32	0.0093	1	0.12	0.0064	0.02	-0.14	-0.13	-0.074
Prior_purchases	-0.0053	-0.0016	0.18	0.013	0.12	1	0.0037	-0.0094	-0.083	-0.17	-0.056
Product_importance	0.0043	0.0049	0.0063	0.0032	0.0064	0.0037	1	-0.0099	-0.0063	0.0017	-0.023
Gender	-0.0037	-0.011	0.0025	0.0028	0.02	-0.0094	-0.0099	1	-0.012	0.0036	0.0047
Discount_offered	0.0096	0.0094	-0.13	-0.0031	-0.14	-0.083	-0.0063	-0.012	1	-0.38	0.4
Weight_in_gms	0.0041	-0.0008	-0.28	-0.0019	-0.13	-0.17	0.0017	0.0036	-0.38	1	-0.27
Reached.on.Time_Y.N	0.0052	-0.00054	-0.067	0.013	-0.074	-0.056	-0.023	0.0047	0.4	-0.27	1

Product Cost vs Customer Care Calls

Insight:

- Expensive products often lead to more customer care calls, as seen with 6-7 calls having higher median costs.
- Cheaper products have fewer customer care interactions and a broader cost range.



Preparing Data for Machine Learning

1. **Train-Test Split:**

- The data is split into two sets: **80% for training** and **20% for testing**.
- Setting a random state ensures consistent results, and using stratify=y keep the target variable distribution proportional in both sets.

2. **Splitting Data into X and y:**

- **The dataset is divided into:** X: Independent variables (features used for predictions).

  y: Dependent variable (the target we want to predict).

Model Building

Using the following models to predict the product delivery:

--Random Forest Classifier
--Decision Tree Classifier
--Logistic Regression
--K Nearest Neighbors

Classification Report

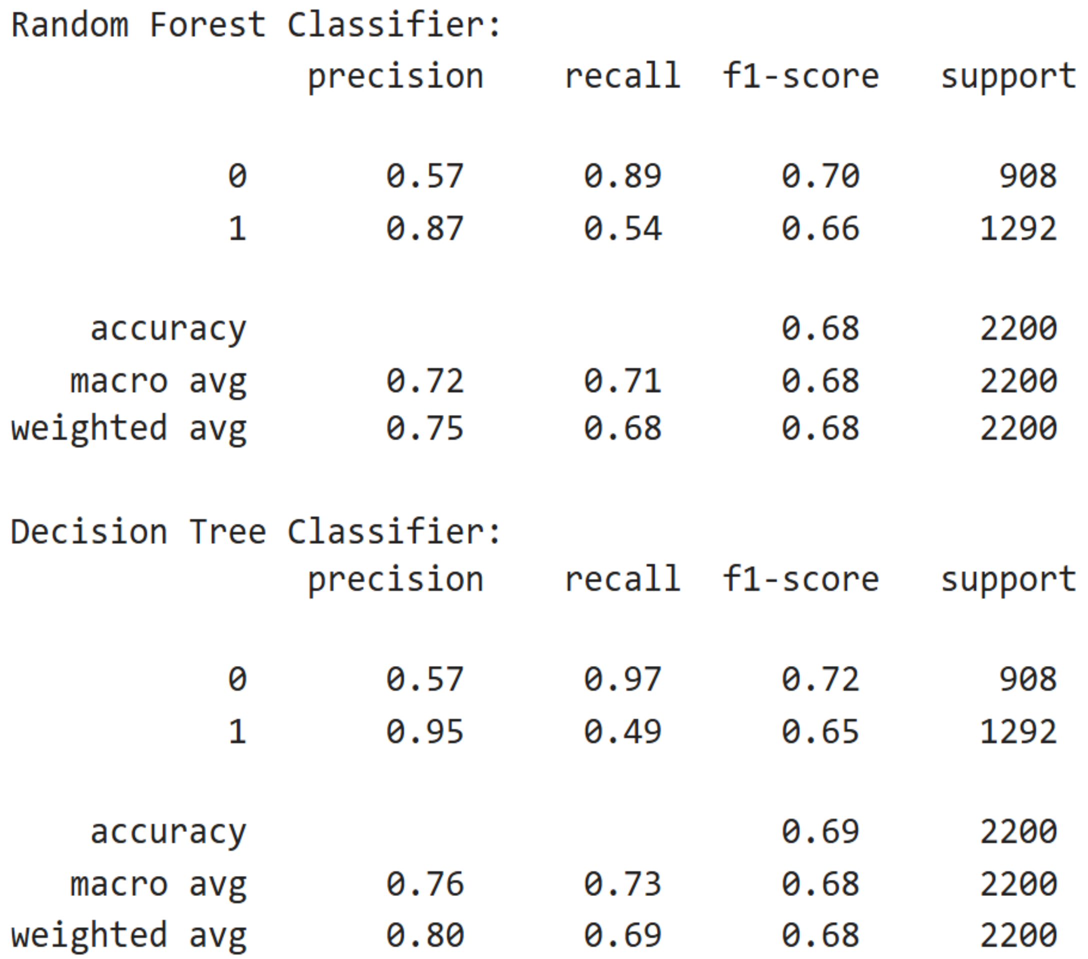
Insight:

1.Random Forest Classifier:

- High Recall (0.89) for class 0 indicates it identifies most negative samples correctly.
- Overall Accuracy (0.68) makes it the best-performing model in this comparison.

2.Decision Tree Classifier:

- Very High Recall (0.97) for class 0 but low recall for class 1 (0.49), meaning it struggles with positive samples.
- Accuracy (0.69) shows decent performance but less balanced compared to Random Forest.



```
Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.57      0.89      0.70       908
           1       0.87      0.54      0.66      1292

    accuracy                           0.68      2200
   macro avg       0.72      0.71      0.68      2200
weighted avg       0.75      0.68      0.68      2200

Decision Tree Classifier:
              precision    recall  f1-score   support

           0       0.57      0.97      0.72       908
           1       0.95      0.49      0.65      1292

    accuracy                           0.69      2200
   macro avg       0.76      0.73      0.68      2200
weighted avg       0.80      0.69      0.68      2200
```

Classification Report

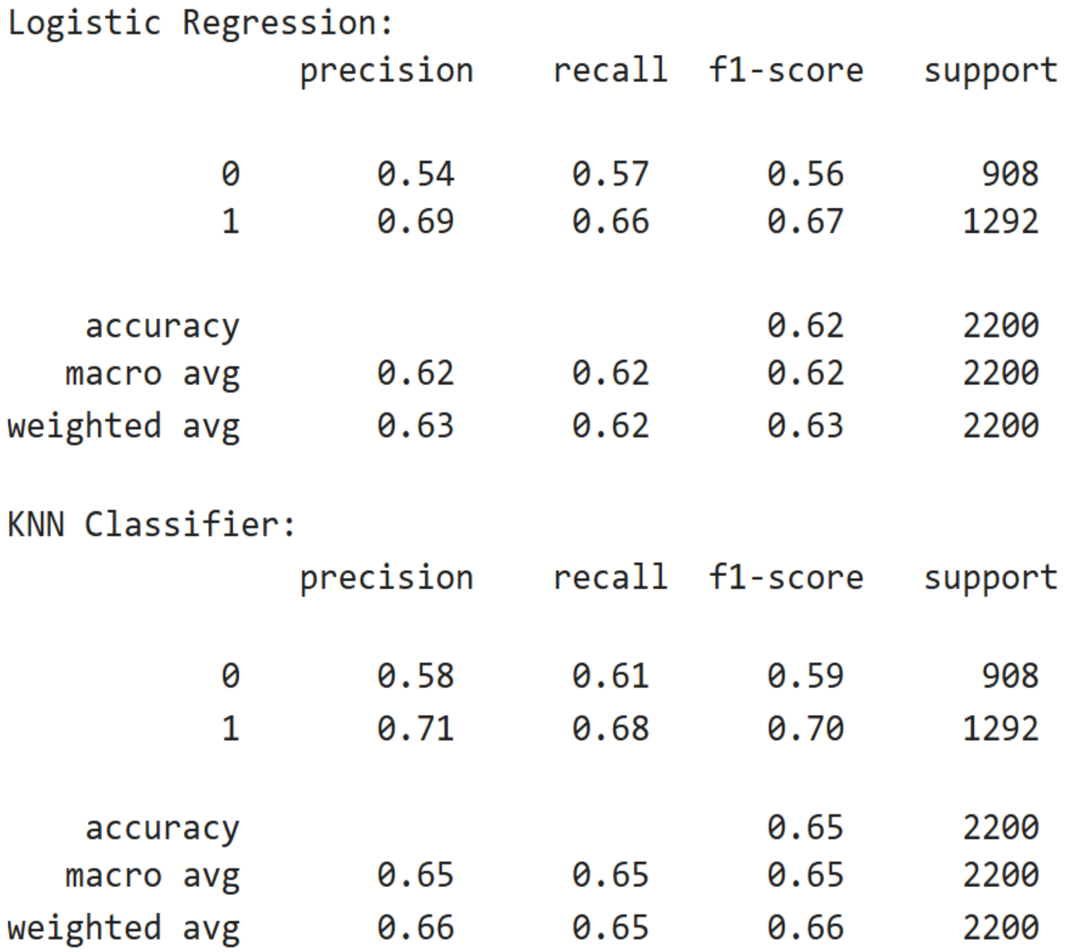
Insight:

3.Logistic Regression:

- Balanced scores across precision, recall, and F1 for both classes, but overall performance is lower (Accuracy: 0.62).
- This model is simpler and interpretable, making it suitable for basic tasks.

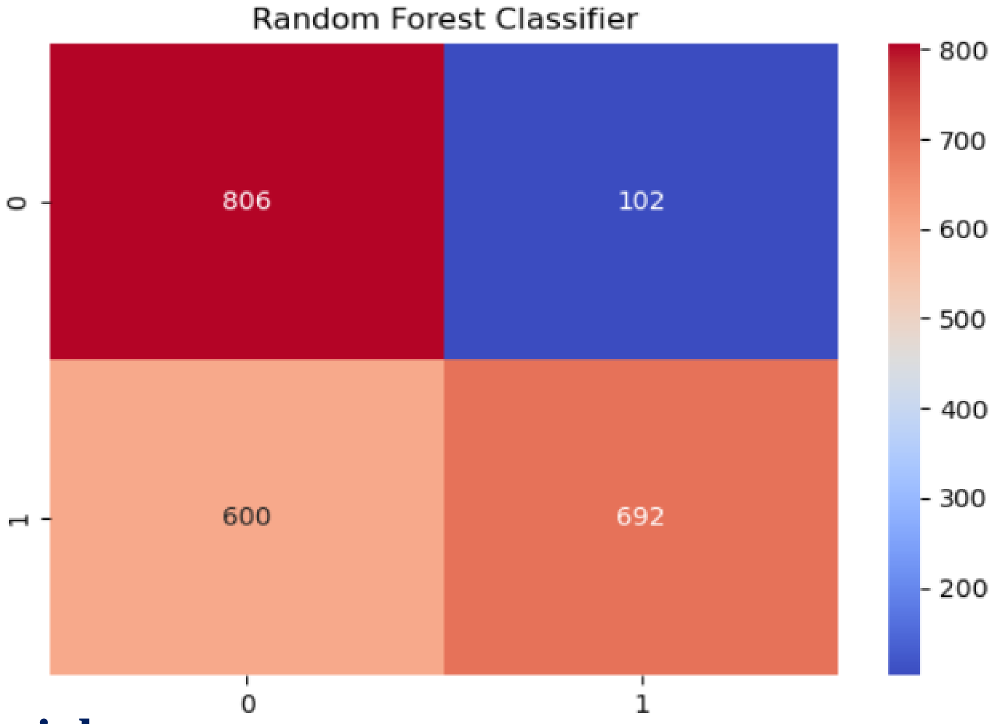
4.KNN Classifier:

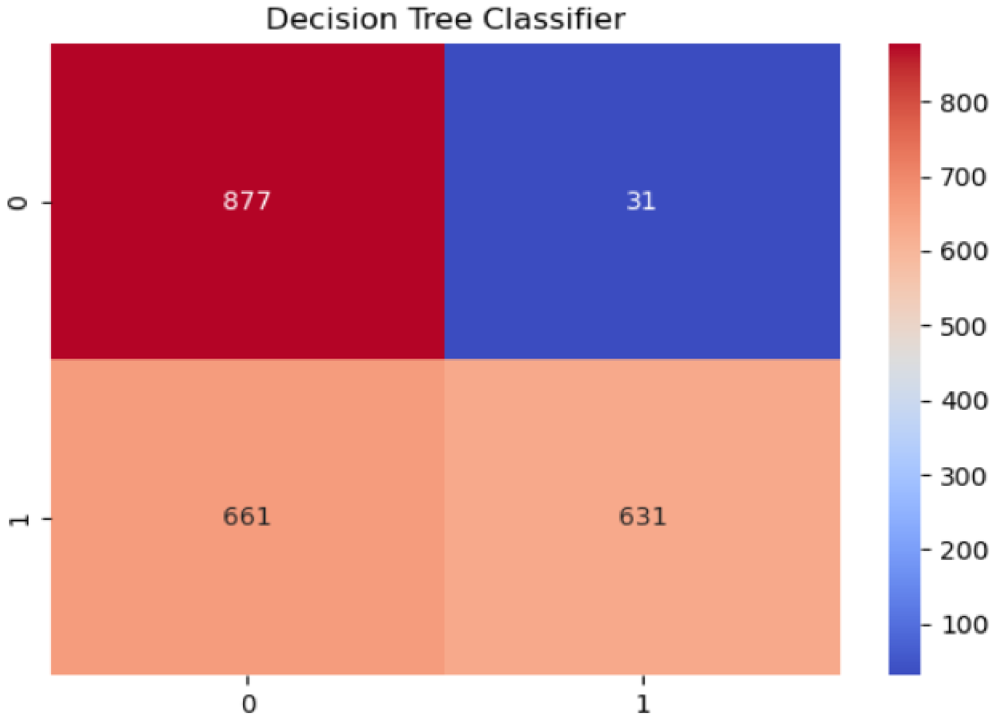
- Scores (Precision, Recall, F1) are lower for both classes (Accuracy: 0.65), indicating it struggles without parameter tuning.
- It may need optimization to improve results



```
Logistic Regression:
              precision    recall  f1-score   support

           0       0.54      0.57      0.56       908
           1       0.69      0.66      0.67      1292

    accuracy                           0.62      2200
   macro avg       0.62      0.62      0.62      2200
weighted avg       0.63      0.62      0.63      2200

KNN Classifier:
              precision    recall  f1-score   support

           0       0.58      0.61      0.59       908
           1       0.71      0.68      0.70      1292

    accuracy                           0.65      2200
   macro avg       0.65      0.65      0.65      2200
weighted avg       0.66      0.65      0.66      2200
```

Model Evaluation

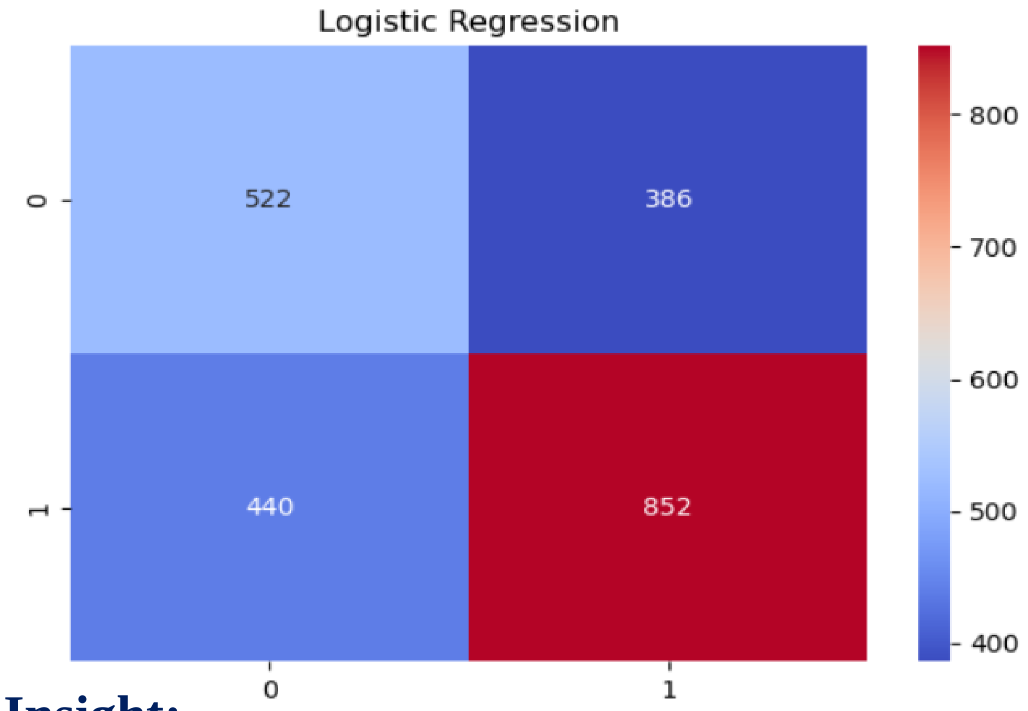


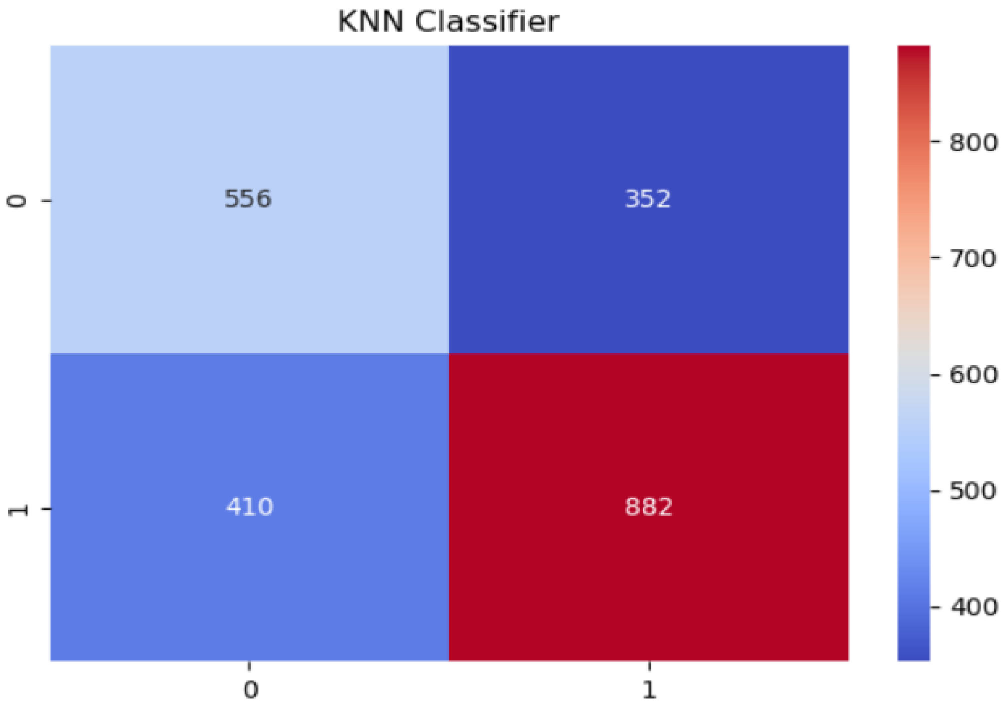


Insight:

1.Decision Tree: Best performance with minimal misclassifications.
2.Random Forest: Good, but slightly more errors than Decision Tree.

Model Evaluation


Logistic Regression

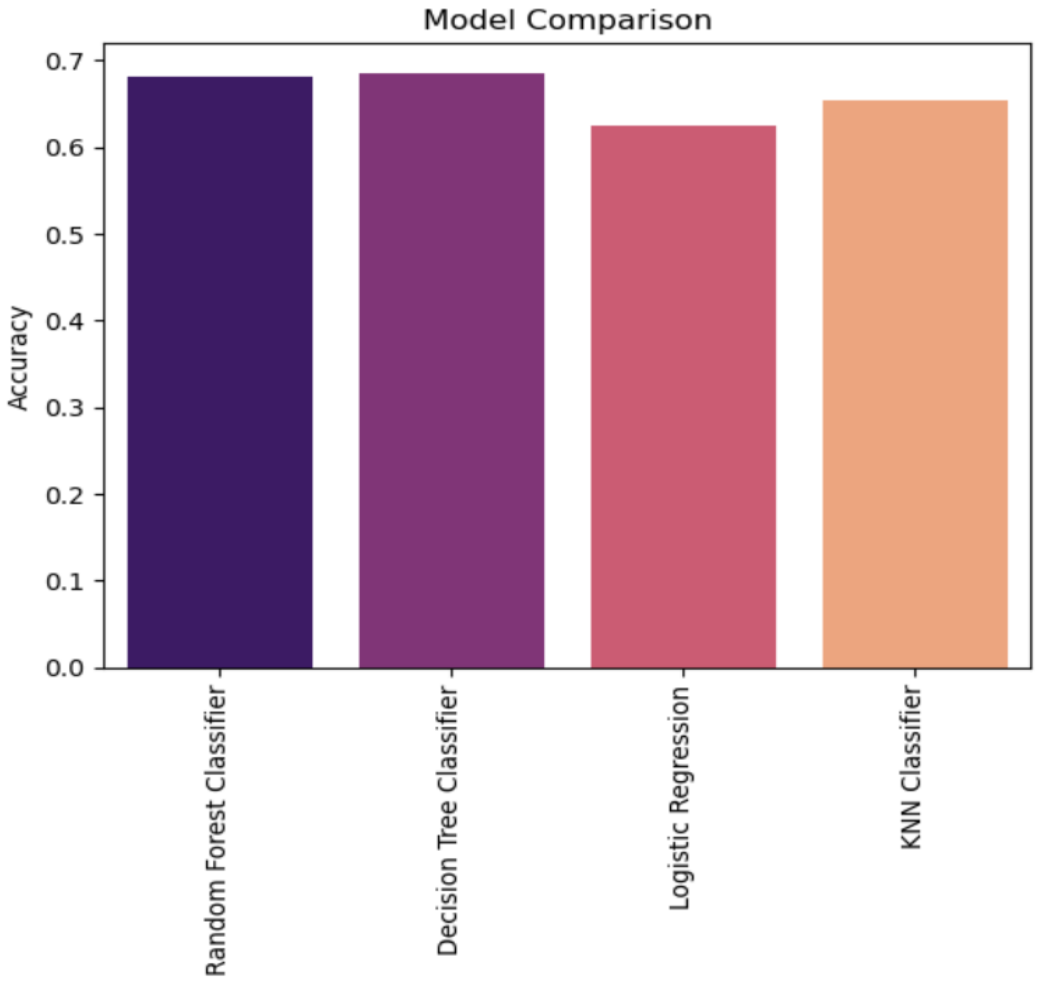

KNN Classifier

Insight:

3.Logistic Regression: Balanced performance

4.KNN Classifier: Higher misclassification; may need tuning for better results.

Model Comparison



Insight:

The Random Forest and Decision Tree classifiers achieved the highest accuracy among the models tested, closely followed by Logistic Regression and KNN Classifier. This indicates that ensemble-based and tree-based approaches may be more effective for this dataset. for high-cost

Conclusion

1. **The project aimed to predict on-time delivery of e-commerce products and analyze factors affecting delivery times and customer behavior.**

2. **Key findings:**
   - Products weighing 2500-3500 grams and priced below $250 were more likely to arrive on time.
   - Warehouse F, near a seaport, handled a significant number of shipments.
   - More customer care calls often indicated delivery delays.
   - Loyal customers with multiple purchases experienced more punctual deliveries.
   - Products with higher discounts (>10%) were delivered on time more frequently.

3. **Model Performance:**
   - Decision Tree Classifier: 69% accuracy (best performer).
   - Random Forest Classifier: 68% accuracy.
   - KNN Classifier : 65% accuracy
   - Logistic Regression : 62% accuracy (least accurate).

Any Questions ?

Thank You!