# Delivery Performance Analysis Report

*Date: 01/07/2025*

## 1. Executive Summary

This report presents the results of an analytical project aimed at predicting on-time delivery of e-commerce shipments. Using historical logistics and customer data, we developed and evaluated multiple machine learning models to determine the factors most associated with delays. The final model achieved an accuracy of 86%, offering a valuable decision-support tool for improving delivery performance and customer satisfaction.

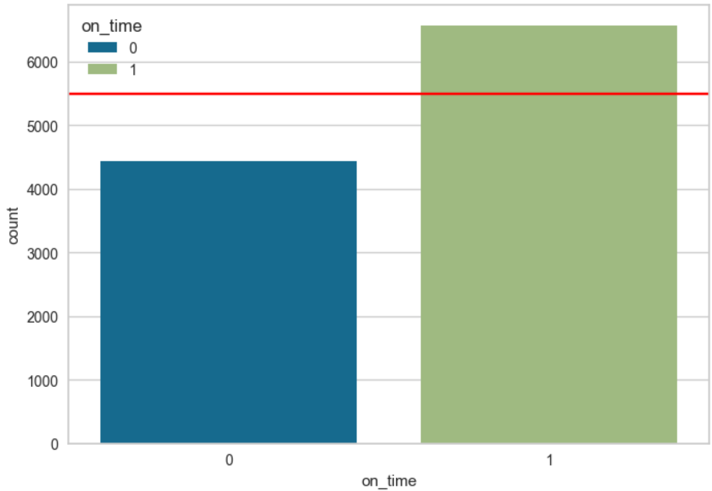
## 2. Objective

The objective of this analysis is to build a predictive model capable of classifying whether a shipment will be delivered on time, based on available features such as transportation mode, product characteristics, customer interaction, and commercial conditions.

## 3. Dataset Overview

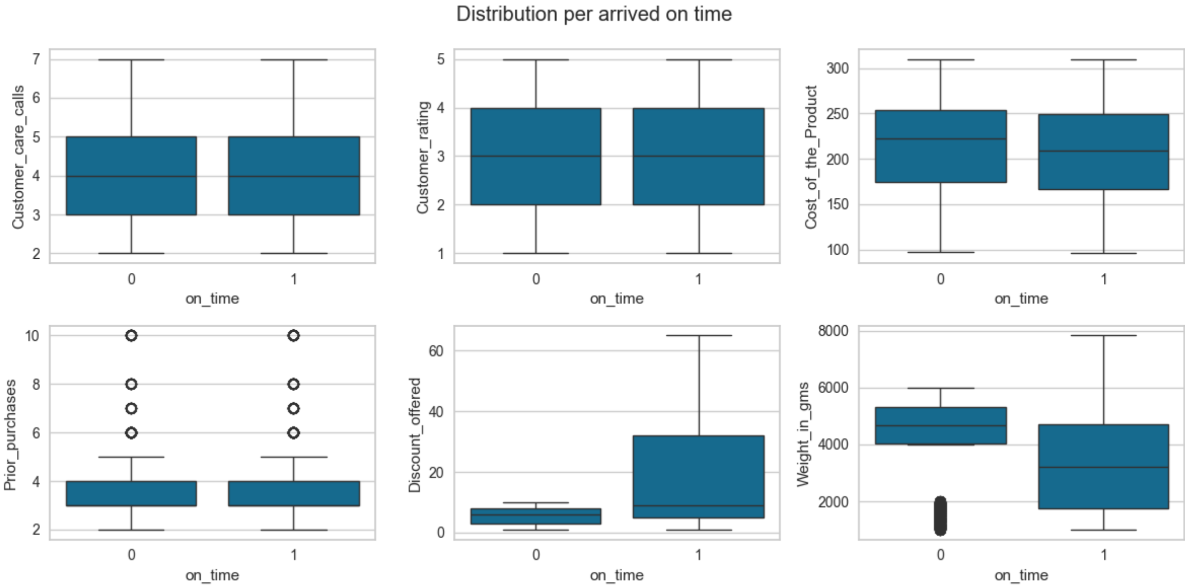
- **Size:** 11,099 records
- **Features:**
  - Shipping method, warehouse block
  - Customer service calls
  - Product cost, discount offered
  - Weight in grams
  - Customer rating
- **Target variable:** `Reached.on.Time_Y.N` (1 = On time, 0 = Delayed)

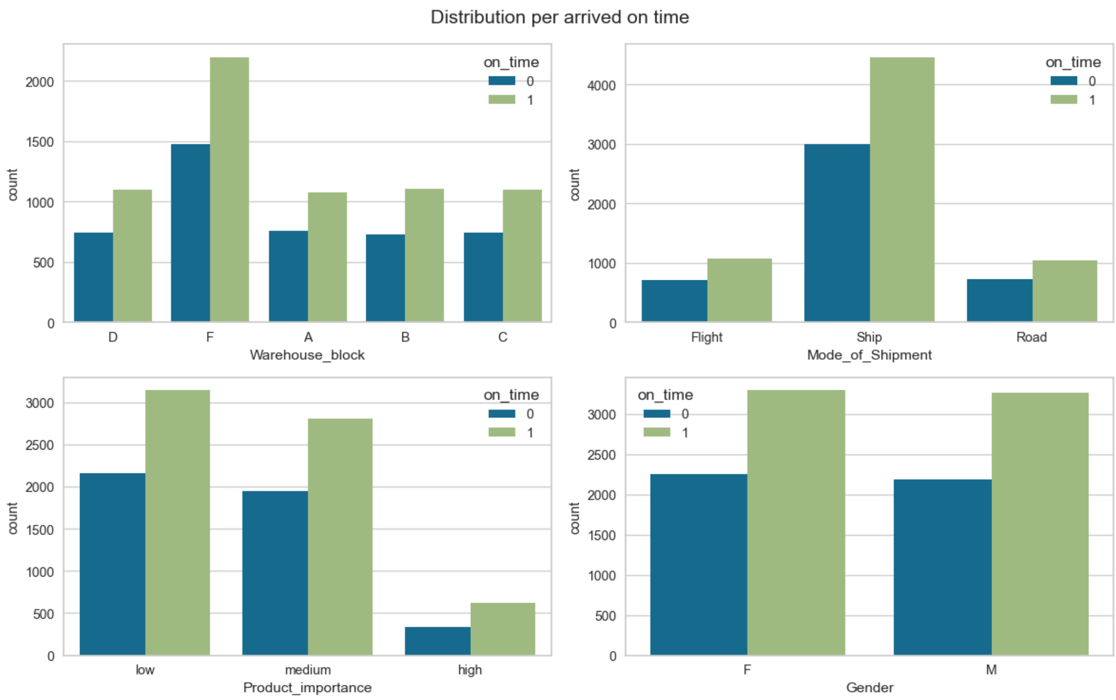
The dataset contains both categorical and numerical variables, with a slight imbalance in the target distribution (~60% on-time, ~40% delayed).



## 4. Data Analysis & Key Insights

- Shipments with higher discounts and greater weight tend to be delayed more frequently.
- Road shipment was the most common delivery method and also the one with the highest delay ratio.
- A higher number of customer care calls was correlated with late deliveries.
- Correlation and visual analyses supported the creation of new derived features (feature engineering).



## 5. Data Preparation

Key steps:

- Removed irrelevant identifier columns, using Spearman Relation to see relations.
- Detected and marked outliers using Local Outlier Factor (LOF).
- Scaled numerical variables using StandardScaler.
- Encoded categorical features using LabelEncoder.
- Created new variables to improve prediction, such as:
  - `discount_per_cost`
  - `cost_per_weight`
  - `calls_per_weight`
  - `purchase_per_weight`
- Checked for multicollinearity using Variance Inflation Factor (VIF) and removed redundant variables.

## 6. Model Development

Multiple models were trained and evaluated:

- Naive Bayes
- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Random Forest
- Support Vector Machine
- Neural Network (MLPClassifier)

Hyperparameter tuning was performed using GridSearchCV, and performance was evaluated using 30-fold cross-validation.

## 7. Results

The Neural Network (MLPClassifier) showed the best performance:

| Model | Accuracy | F1-Score | $R^2$ Score | Std Dev (CV) |
|---|---|---|---|---|
| Naive Bayes | 0.771 | 0.77 | 0.08 | 0.228 |
| Logistic Regression | 0.779 | 0.78 | 0.22 | 0.220 |
| K-Nearest Neighbors | 0.828 | 0.79 | 0.31 | 0.171 |
| Decision Tree | 0.831 | 0.80 | 0.32 | 0.169 |
| Random Forest | 0.841 | 0.81 | 0.36 | 0.158 |
| Support Vector Machine Classifier | 0.860 | 0.82 | 0.44 | 0.139 |
| Neural Network (MLPClassifier) | 0.863 | 0.83 | 0.45 | 0.136 |

The model was saved using joblib for potential deployment or reuse.

## 8. Business Impact

- The model can help logistics teams identify high-risk shipments before delays occur.
- Operations can be prioritized based on delivery risk, improving efficiency and customer trust.
- Cost-saving opportunities exist through better resource allocation.

## 9. Recommendations

- Focus on orders with high discounts, large weights, and multiple customer service calls.
- Investigate potential bottlenecks in road shipment logistics.
- Implement model predictions into a dashboard for real-time shipment risk tracking.