LOGISTICS DELIVERY DELAY EDA AND PREDICTION

R AISWARIYA

PROBLEM STATEMENT

Late deliveries cause loss of revenue and customer dissatisfaction.

Manual tracking is inefficient and reactive.

Businesses need a **data-driven prediction model** to proactively manage delivery timelines.

PROJECT OBJECTIVES

- Analyze real-world logistics data to identify delay patterns.
- Engineer predictive features from shipment, product, and time-based attributes
- Build a machine learning model to predict delivery outcomes (early/ on-time / delayed).
- Enable interactive visualization and prediction through a Streamlit app.

DATASET OVERVIEW

Source: Kaggle – Logistics Data Containing Real-World Data

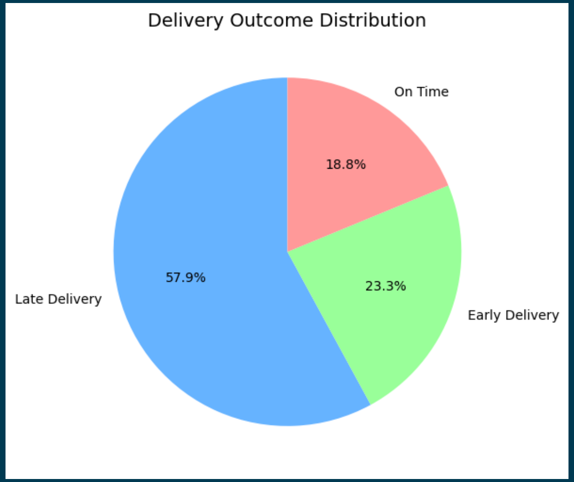
Key Columns:

- Shipment details (Mode, Cost, Weight)
- Dates (Order, Ship, Delivery)
- Product category & department
- Delivery outcome (Target variable)

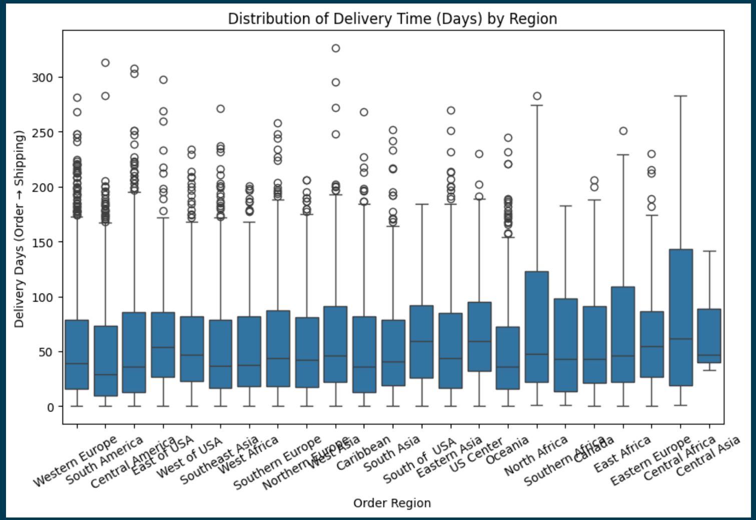
Total rows & columns: 15549 rows × 41 columns

Data cleaning steps: missing values, type conversion, outlier handling

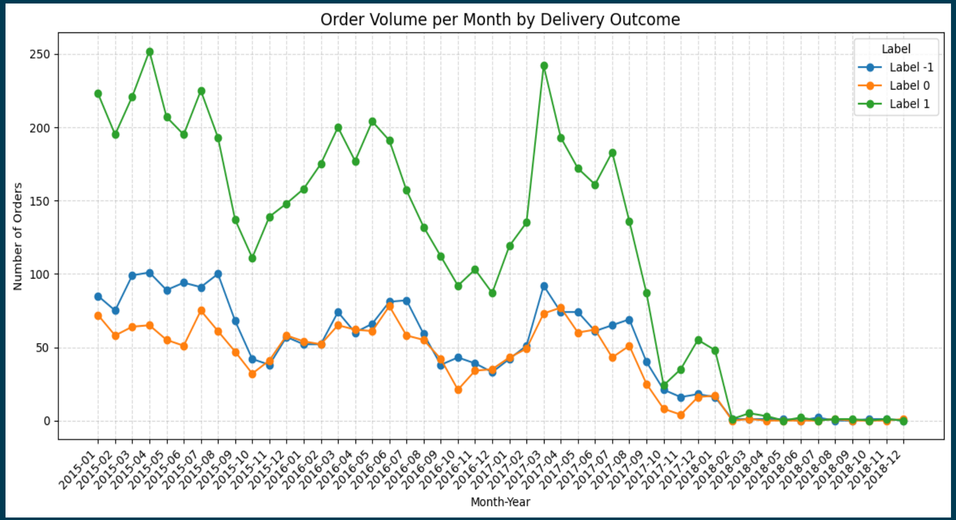
EXPLORATORY DATA ANALYSIS (EDA)



The classes are imbalanced – late delivery dominates the dataset.

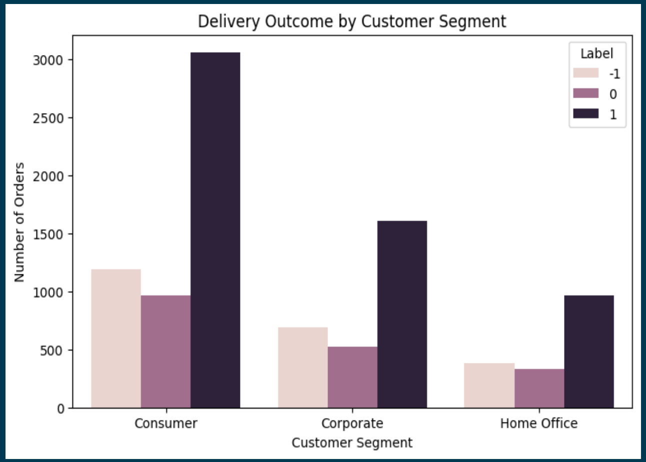


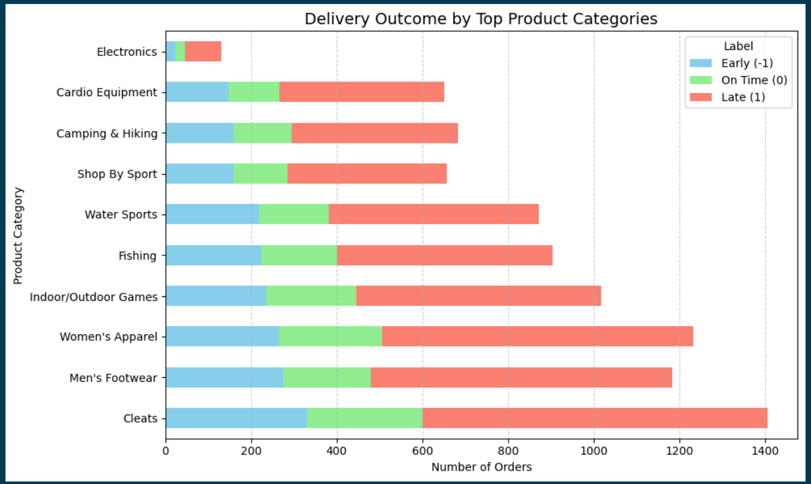
Some regions like Western Europe, South America, Central America, Oceania and Northern Europe show a large no of outliers, indicating presence of cases with unusually large delivery time.



The volume of deliveries always reduce during Aug-Nov window and then gradually increases to a peak in Jan-March.

EXPLORATORY DATA ANALYSIS (EDA)





Standard Class, early and late deliveries occur at similar rates, but it shows the highest likelihood of on-time delivery.
First, Second, and Same-Day modes rarely have early deliveries; while late deliveries are similar across First and Second Class, Second Class performs better for on-time deliveries

The larger number of deliveries are consumers and as expected, they have the highest no of late deliveries, followed by corporate customer segment.

The largest number of orders are placed on Cleats, then Men's Footwear and then Women's Apparel. From the graph, the chances of late delivery follows that order too. But, the chances of on-time delivery slightly greater for Women's Apparel than Men's Footwear

DATA PREPROCESSING & FEATURE ENGINEERING

- Extracted date difference (delivery duration) from order and delivery dates.
- Cleaned data by removing incorrect values and outliers.
- Checked for multicollinearity and removed redundant features.
- Dropped features with near-zero correlation to the target label.
- Applied scaling and normalization to numerical features and label encoding to categorical features.
- Performed train-test split for model development.

MODEL DEVELOPMENT

Model	Accuracy
Gaussian Naïve Bayes	62.1%
KNN	54.6%
SVC	62.2%
Decision Tree	56.5%
Random Forest	62.7%
AdaBoost	62.9%
GradientBoosting	62.5%
XGBoost	61.0%

- Base models achieved ~54–63% accuracy, showing moderate performance.
- Random Forest, AdaBoost, and Gradient Boosting showed relatively better results among all models.
- Class imbalance caused bias toward majority delivery outcomes.
- Applied SMOTE and backward feature selection to boost accuracy and generalization.

RESULTS-01

Model	Accuracy
Gaussian Naïve Bayes	61.7%
KNN	57.1%
SVC	37.8%
Decision Tree	62.2%
Random Forest	71.7%
AdaBoost	64.0%
GradientBoosting	65.9%
XGBoost	72.1%

- Model performance improved notably after SMOTE and feature selection.
- Random Forest and XGBoost achieved the highest accuracies (~72%).
- Class balance correction enhanced minority class prediction.
- Feature optimization reduced noise, leading to better generalization.

RESULTS-02

RESULTS-03
STACKING CLASSIFIER

- Implemented a Stacking Classifier combining XGBoost and Random Forest as base learners with Logistic Regression as meta-model.
- Leveraging multiple models strengths for improved ensemble performance.
- Achieved the highest accuracy of **73.41%**, outperforming all individual models.
- Stacking effectively enhanced generalization and reduced overfitting.
- Final stacked model deployed in Streamlit for real-time prediction.

KEY INSIGHTS

- Delays are strongly influenced by **shipping mode, delay between order and shipping,** and **region**.
- Early shipments are rare; oversampling helped the model learn minority classes.
- The stacking approach outperformed single models in overall F1-score.
- Real-time predictions can improve operational planning.

FUTURE SCOPE

- Integrate **traffic, weather, and holiday data** for better accuracy.
- Implement **real-time API integration** with logistics tracking systems.
- Build an **alert system** for predicted delays.
- Expand the dashboard with trend forecasting.

THANK YOU

STREAMLIT DASHBOARD FOR LIVE PREDICTION:

CLICK 👉 LAUNCH APP