Project Report - by Devansh

E-commerce Returns & Refund Cost Analysis Using Logistic Regression

Abstract:

In the modern e-commerce ecosystem, product returns represent a major operational and financial challenge. High return volumes lead to increased refund costs, reverse logistics expenses, and customer dissatisfaction. This project analyzes an e-commerce returns dataset to identify patterns and drivers behind high-cost refunds. Exploratory Data Analysis (EDA) and Logistic Regression were used to classify high-cost returns and interpret the influence of operational factors such as processing time, return reason, and refund method. The analytical findings were transformed into an interactive Power BI dashboard to support data-driven decision-making for reducing refund-related losses.

Introduction:

E-commerce platforms operate in a highly competitive environment where customer satisfaction and operational efficiency are critical for long-term success. While flexible return policies help attract customers, they also increase operational complexity and refund costs. Returns due to damaged products, late deliveries, or customer dissatisfaction directly impact profitability and supply chain efficiency. The purpose of this project is to analyze return and refund data to understand key cost drivers, identify inefficiencies in return processing, and predict the likelihood of high-cost returns. By combining data analysis, machine learning, and visualization techniques, the project simulates a real-world data analyst workflow and provides actionable business insights.

Tools Used:

- Python: Pandas for data manipulation, NumPy for numerical computation, Matplotlib and Seaborn for exploratory data visualization.
- Scikit-learn: Logistic Regression model for classification and model evaluation.
- SQL: Data validation, aggregation, and consistency checks.
- Power BI: Interactive dashboards, KPIs, and business-focused visualizations.
- Jupyter Notebook: End-to-end analysis and model development environment.
- CSV Dataset: E-commerce returns and refund transaction data.

Steps Involved in Building the Project:

1. **Data Understanding and Preparation:** The dataset was loaded and examined to understand its structure, data types, and missing values. Columns related to refund amount, processing time, return reason, and refund method were identified as key analytical features.

2. **Data Cleaning and Feature Engineering:**

   Incorrect data types were corrected, missing values were handled, and duplicate records were removed. New features such as a high-cost return flag and processing time buckets were created to support analysis and modeling.

3. **Exploratory Data Analysis (EDA):**

   EDA was performed to analyze refund distributions, return reasons, and processing delays. Visualizations revealed that longer processing times and specific return reasons were strongly associated with higher refund amounts.

4. SQL-Based Analysis:

SQL queries were used to validate aggregate metrics such as total refunds by reason, average refund cost, and processing time trends, ensuring analytical consistency.

5. Machine Learning Modeling:

A Logistic Regression model was trained to classify high-cost versus normal-cost returns. Model evaluation metrics such as confusion matrix, precision, recall, and F1-score were used to assess performance.Model coefficients were analyzed to interpret feature influence.

6. Power BI Dashboard Development:

Key KPIs and visualizations were built to present insights to business stakeholders.
The dashboard highlights refund cost drivers, processing inefficiencies, and high-cost return patterns.

Conclusion:

This project demonstrates how data analytics and machine learning can be applied to address real-world e-commerce challenges. While the Logistic Regression model achieved baseline predictive performance, it successfully provided interpretable insights into the factors influencing high refund costs. Processing delays and specific return reasons emerged as the most significant contributors to high-cost returns. The Power BI dashboard effectively translates analytical findings into actionable insights, enabling stakeholders to monitor return trends, optimize operations, and reduce refund-related losses. Overall, the project showcases a complete data analyst workflow and serves as a strong portfolio example for entry-level data analyst roles.