PROJECT REPORT

On

Ecommerce Return Rate Reduction Analysis

Submitted by

SAANVI SHARMA

as a

Data Analyst Intern

- ## INTRODUCTION:

  E-commerce businesses face significant challenges from product returns, impacting profit margins and customer experience. Accurately identifying high-risk return products can help reduce return rates and optimize inventory strategies. This project aims to analyze e-commerce order data, identify return patterns using SQL, and predict the probability of return using logistic regression in Python. The outputs are visualized using a dynamic Power BI dashboard.

- ## ABSTRACT:

  The project explores a comprehensive pipeline for return risk analysis, starting with SQL-based data exploration to uncover return trends by product, city, category, and time. Machine learning (logistic regression) is applied to predict return probabilities using engineered features such as product, quantity, and city. The final insights are exported and visualized via Power BI, empowering decision-makers with interactive risk segmentation and high-return-product views. This end-to-end solution is both scalable and practical for operational deployment.

- ## TOOLS USED:

- **SQL**: Data summarization, return rate computation
- **Python**: Data preprocessing, logistic regression modeling, probability prediction
- **Pandas, Scikit-learn, Matplotlib**: Data manipulation and modeling
- **Power BI**: Interactive visualization dashboard
- **Excel**: Intermediate export for model results

- ## STEPS INVOLVED IN BUILDING THE PROJECT:

1. **Data Cleaning & Preprocessing**
    a. Cleaned column names and handled missing values
    b. Created is_returned flag based on shipped = 0
    c. Simulated return cases due to class imbalance
2. **SQL-Based Exploratory Analysis**
    a. Computed total return rates overall, by product, and by city
    b. Identified high-return categories and seasonal spikes using:
    c. GROUP BY, HAVING, and DATE_FORMAT functions
    d. Heatmaps and bar charts via Python
3. **Predictive Modeling with Logistic Regression**
    a. Selected key features: product, quantity, orderid, city
    b. One-hot encoded categorical variables
    c. Trained logistic regression model with class_weight='balanced'
    d. Evaluated with Accuracy: 0.93, AUC: 0.79 (on simulated returns)
4. **Power BI Dashboard Creation**
    a. Imported model predictions and SQL summaries
    b. Created visuals:
        - Bar Chart: Return Rate by Product (Count of shipped by products and shipped)
        - Line Chart: Monthly Return Trend (Count of shipped and Sum of quantity as per year and month)
        - Column Chart: Return Rate by City (Count of shipped by city)
        - Pie Chart: Count of city by product and shipped
    c. Enabled slicers for category, city, and month filters

- ## CONCLUSION:

This project successfully integrates data engineering, machine learning, and BI visualization to offer a powerful return risk dashboard. By predicting return probabilities and visualizing them across multiple dimensions, decision-makers gain actionable insights to refine product strategy, manage supplier relations, and minimize return-related losses. Future enhancements could include real-time data integration and multi-model ensemble testing.