E-commerce Return Rate Reduction Analysis: Final Project Report

1. Introduction

High rates of product returns are a major problem in the quickly expanding e-commerce sector. Returns raise inventory and logistics expenses in addition to decreasing profit margins. In order to help businesses lower return rates and improve customer happiness, this initiative focuses on analyzing the trends behind consumer returns and forecasting high-risk products.

2. Abstract

An e-commerce platform's product return data is analyzed in this project to find important trends in returns across brands, categories, regions, and traffic sources. The likelihood of a product return was forecasted by logistic regression modeling and data preparation using Python. High-risk goods, return trends, and actionable information were all graphically represented in a Power BI dashboard. An interactive dashboard, a predictive Python model, and a CSV file of high-risk products are among the final deliverables that assist companies in managing returns proactively.

3. Tools Used

- **Python:** Data cleaning, feature engineering, and logistic regression modelling
- **Libraries:** Pandas, NumPy, Scikit-learn, imbalanced-learn, Matplotlib
- **Power BI:** Data visualization and interactive dashboard creation
- **MySQL:** Data aggregation and time-series grouping for line chart insights
- **Excel:** Minor data exploration and formatting

4. Steps Involved in Building the Project

1. Data Collection & Preparation:

- Extracted and combined multiple CSV files (orders, order_items, products, users, events)
- Created the return_flag binary label and merged all datasets
- Handled missing values, removed duplicates, cleaned categorical columns

2. **Feature Engineering:**

   o Calculated new fields: price_margin, discount, is_discounted, order_day

   o Extracted customer behavior signals using traffic_source, state, gender, etc.

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

   o Computed return rates by category, brand, department, and distribution center

   o Generated charts to explore relationships using Power BI (e.g., pie charts, KPI cards, map, line chart)

4. **Model Building:**

   o Used Logistic Regression with SMOTE and OneHotEncoding to classify returns

   o Evaluated the model using accuracy, precision, recall, F1-score, and ROC curve

   o Extracted feature importance using model coefficients

5. **Output Generation:**

   o Identified high-risk return products (top 5%) and saved to high_risk_products.csv

   o Created a return_summary.csv for further dashboard analysis

   o Cleaned and exported the master dataset full_ecommerce_data_cleaned.csv

---

5. **Conclusion**

The project effectively identified patterns in customer returns and built a logistic regression model to predict return risks. The Power BI dashboard provides a comprehensive view of return behaviour, empowering businesses to make informed decisions. This solution not only enhances business efficiency but also contributes to customer retention by addressing core issues behind product returns.