

E-commerce Return Rate Reduction Analysis

Abstract

This project aims to analyze product return behavior on an e-commerce platform to identify patterns and predict future returns. By combining machine learning, SQL-based analytics, and interactive visualizations, the goal was to pinpoint high-risk products, understand return behavior across customer segments, and provide actionable business recommendations to reduce operational costs and improve customer satisfaction.

Introduction

Product returns are a major challenge in the e-commerce industry, impacting profits, logistics, and customer loyalty. Understanding why returns happen — and which products, customers, or regions are high-risk — enables better product decisions and targeted interventions. This project tackles that challenge by leveraging structured data, statistical modeling, and business intelligence tools to reduce return rates.

Tools Used

- **Python (Jupyter Notebook):** Logistic regression model, EDA, and CSV export
- **MySQL Workbench:** Return trend analysis via queries
- **Power BI:** Dashboard development and visual storytelling
- **Libraries:** pandas, matplotlib, seaborn, scikit-learn

Steps Involved in Building the Project

◆ 1. Data Preparation

- Used `ecommerce_return_enhanced.csv` (preprocessed file).
- Key fields included: `Product_ID`, `Category`, `Region`, `Price`, `CLV`, `Discount`, `Return_Status`.
- Engineered features: `Return_Label`, `Return_Probability`, `Days_to_Return`, `Order_Weekday`.

◆ 2. Python (Modeling & Insights)

- Trained a logistic regression model to predict `Return_Probability`.
- Exported products with return probability > 0.5 into `high_return_risk_products.csv`.

- Observed that most predictions were between 0.5–0.6, indicating model caution.
- Found that returns were more likely in certain categories (Clothing, Electronics) and when high discounts were applied.

◆ 3. SQL Segment (Pattern Analysis)

- Analyzed returns using group-by queries on ecommerce_return_enhanced table.
- Identified high-return product categories (Clothing, Books), high-risk users, and trends by region and shipping method.
- Found that:
 - Urban areas and express delivery had slightly higher return rates.
 - Social media-driven orders had higher returns, suggesting impulsive buys.
 - Most returns were due to "Defective" or "Wrong Item".

◆ 4. Power BI Dashboard (Interactive Reporting)

Page 1: Product & Return Risk Overview

- Category-wise return analysis
- Region vs Margin matrix
- Key Insight: Overall return rate at 50.52%; Books & Clothing highest in volume.

Page 2: Customer Insights Dashboard

- Customer-level segmentation with CLV, Tenure, Risk Flag
- Age 18–25 showed highest returns
- Ribbon chart showed South & East regions with higher CLV
- Funnel chart: 30% of all orders flagged high-risk.

Page 3: Operational Insights

- Monthly return trends, weekday analysis, return reasons
- Next-day shipping linked to faster returns
- Monday saw highest returns, suggesting post-weekend regret
- Payment methods had even return distribution; gift cards slightly higher.

Conclusion

This project successfully integrated data modeling, SQL analytics, and dashboarding to explore and predict product returns.

Key Takeaways (Insights-Driven)

- **Books, Clothing, and Electronics** emerged as the most return-prone categories, often due to issues like sizing errors or defective items — highlighting a need for better product descriptions and quality control.
- **Customers aged 18–25** and users acquired through **social media marketing** showed the highest return rates, indicating possible impulsive buying behavior and the need for more targeted, informed engagement.
- **Express and next-day shipping** methods correlated with faster returns, suggesting that urgency-driven purchases may lead to higher dissatisfaction or regret.
- **Returns peaked on Mondays**, hinting at post-weekend return behavior and possible weekend impulsive shopping trends.
- Regions like the **South and East** showed higher average CLV but also higher return rates, indicating that high-value customers are still experiencing dissatisfaction — a critical signal for customer retention efforts.
- A significant portion of returns stemmed from **“Defective” and “Wrong Item”** reasons, pointing toward issues in fulfillment and logistics workflows.
- **Nearly 30% of all orders were flagged as high-risk**, allowing for the development of targeted interventions (e.g., extra checks before shipment, personalized messaging, or stricter return policies).