E-commerce Return Rate Reduction Analysis

Introduction

This project focuses on analyzing and predicting product return behavior in an e-commerce setting. High return rates affect profitability and customer experience. The goal is to identify key drivers of product returns, build predictive models, and suggest strategies to minimize avoidable returns.

Abstract

The dataset consists of 10,000 e-commerce transactions including product, customer, and order details. A binary target variable (Returned / Not Returned) was created from the return status. The data was cleaned, categorical variables encoded, and numerical features standardized. Multiple machine learning models (Logistic Regression, Random Forest, XGBoost) were trained and evaluated. The analysis highlights the most important factors contributing to returns such as product category, discount levels, and shipping methods.

Tools Used

- Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost) - SQL (for aggregations and category-wise return analysis) - Power BI (for interactive dashboards and visual storytelling)

Steps Involved in Building the Project

1. Data Cleaning: Removed irrelevant columns (IDs, return-only data), handled missing values. 2. Feature Engineering: Encoded categorical features, created discount-related features. 3. Train-Test Split: 70-30 split with stratification on target variable. 4. Model Training: - Logistic Regression as baseline (Accuracy ~50%). - Random Forest improved recall and feature interpretability. - XGBoost delivered the best performance with higher accuracy and balanced metrics. 5. Feature Importance: Discount applied, product category, shipping method, and user demographics were strong predictors. 6. Dashboarding: Interactive Power BI visuals to analyze return % by category, supplier, region, and discount level.

Conclusion

The project demonstrates that product returns are heavily influenced by discounts, specific product categories (like fashion), and shipping methods. Tree-based models such as Random Forest and XGBoost outperform linear models, making them better suited for predicting return likelihood. Businesses can reduce return rates by optimizing discount strategies, focusing on quality in high-return categories, and improving logistics. This project not only strengthens technical data science skills but also provides actionable business insights.