E-Commerce Return Rate Reduction Analysis

Abstract

In today's fast-paced e-commerce industry, product returns pose a persistent challenge for businesses, impacting profitability, customer satisfaction, and supply chain efficiency. This project addresses that issue by identifying patterns behind product returns and predicting high-risk return orders using data analytics and machine learning. By employing Python for data processing and modeling, SQL for filtering, and Power BI for dynamic visualization, the project delivers a comprehensive solution that helps businesses understand, monitor, and act on return trends.

Introduction

Product returns in e-commerce affect not just profits but also customer experience, brand perception, and logistics operations. Understanding why customers return products—whether due to high prices, long shipping times, poor product fit, or regional preferences—can help companies optimize their offerings and processes. The main goal of this project is to analyze historical return data to find trends across various product categories, regions, and customer profiles and to build a predictive model that identifies which future orders are most likely to be returned. This data-driven approach helps stakeholders take informed actions such as adjusting pricing strategies, targeting high-risk regions with better logistics, or flagging high-risk orders for review.

Tools Used

- **Python (Google Colab):** Utilized for loading and cleaning data, exploratory data analysis (EDA), feature encoding, and logistic regression modeling. Key libraries include pandas, pandasql, scikit-learn, and matplotlib.
- **SQL (via pandasql):** Used to query and clean the dataset by filtering out records with missing values.
- **Power BI:** Created an interactive dashboard that visualizes return patterns, predicted risks, and allows drill-through views for detailed insights into individual high-risk orders.

Steps Involved in Building the Project

1. **Data Loading and Cleaning:**
   - Imported a mock dataset simulating real-world e-commerce returns.
   - Used SQL to remove rows where key attributes like price, shipping_days, and customer_rating were missing.

2. **Exploratory Data Analysis (EDA):**
- Analyzed return rates by grouping orders based on product categories and suppliers.
- Plotted return percentage distributions to highlight categories with the highest return issues (e.g., clothing vs. electronics).

3. **Feature Engineering and Model Training:**
- Converted categorical columns (like region and channel) into numerical values using one-hot encoding.
- Scaled numeric features such as price and shipping days to improve model performance.
- Trained a logistic regression model to classify orders as returned or not returned.

4. **Model Evaluation:**
- Assessed performance using metrics such as precision, recall, F1-score, and confusion matrix.
- Model accurately flagged high-return orders, providing a useful probability score for each.

5. **Exporting High-Risk Orders:**
- Predicted return probability for every order.
- Exported those with a probability > 0.4 into a CSV file (high_risk_products.csv) for further review or operational action.

6. **Dashboard in Power BI:**
- Created visuals showing return probabilities by region, product category, and shipping days.
- Included filters and drill-through features that allow users to click on a category and see all related high-risk orders.

Conclusion

The E-commerce Return Rate Reduction Analysis offers a comprehensive approach to solving a real-world business challenge. The predictive model not only highlights where returns are happening but also offers foresight into where they're likely to happen next. Power BI enhances the storytelling aspect by making these insights accessible and interactive for business users. This project proves how integrating machine learning with data visualization tools can lead to smarter decisions, reduced operational costs, and improved customer satisfaction in the e-commerce sector.