

E-commerce Return Rate Reduction Analysis: Project Report

1. Project Objective

The primary objective of this project is to identify why customers return products and how return rates vary by category, geography, and marketing channel. This analysis aims to provide actionable insights to reduce e-commerce return rates.

2. Methodology

The project followed a structured approach to analyze return data and build a predictive model:

2.1. Data Preparation and Cleaning

The initial step involved cleaning the return and order datasets. This included merging order and return information to create a comprehensive dataset, identifying returned items, and handling missing values. For instance, missing marketing channel data was imputed with 'Unknown' to ensure completeness for analysis.

2.2. Return Rate Analysis

Return percentages were analyzed per product category and supplier to identify areas with high return rates. This step helps in pinpointing specific product types or suppliers that contribute significantly to returns, allowing for targeted interventions.

2.3. Predictive Modeling (Logistic Regression)

A logistic regression model was employed to predict the probability of a product return. This model leverages various features from the order and product data to

assess the likelihood of a return. The output of this model is a return probability score for each product, which can be used to identify high-risk items.

2.4. Dashboard Creation (Conceptual)

While the direct creation of an interactive Power BI dashboard was outside the scope of the provided environment, the methodology includes the conceptual step of using Power BI to create a return risk score dashboard. This dashboard would visualize the return probabilities and allow for drill-through filters to explore insights by category, geography, and marketing channel.

3. Tools Used

- **Python:** Utilized for data cleaning, preparation, return rate analysis, and predictive modeling (using libraries such as pandas and scikit-learn).
- **Power BI:** (Conceptual) Intended for creating interactive dashboards for data visualization and risk score presentation.
- **SQL:** (Conceptual) Potentially used for data extraction and management from databases, though not directly demonstrated in the provided scripts.

4. Deliverables

Based on the project objective and methodology, the following deliverables were produced or are conceptualized:

- **Interactive dashboard with drill-through filters:** (Conceptual) A Power BI dashboard designed to visualize return risk scores and allow for detailed exploration of return patterns.
- **Python codebase for prediction:** Provided in `returnprediction.py` and `returnrateanalysis.py`, demonstrating the data processing, model training, and prediction logic.
- **CSV of high-risk products:** Generated as `high_risk_products.csv`, containing products identified by the predictive model as having a high probability of return (e.g., probability > 0.7).

5. Visualizations

The `chart.png` provided illustrates the 'Order Distribution by Category', showing the proportion of orders across different product categories. This type of visualization is crucial for understanding the composition of sales and returns.

