Predictive Return Risk Analysis

Objective:

The primary objective of this project was to establish a proactive system for reducing operational costs and revenue leakage associated with product returns. By applying machine learning to historical sales data, the goal was to develop a predictive **Return Risk Score** that enables intervention before orders are shipped.

Methodology:

The analysis utilized transactional data from `sales_data_sample.csv`. Since a direct return column was unavailable, a synthetic target variable (`is_returned`) was created based on high-risk features, including **high implied discounts** and **order dispute/cancellation statuses**.

A **Logistic Regression Model** was trained on engineered features (e.g., Sales, Discount, Product Line) to generate a **Return Risk Score** (a probability between 0 and 1) for every single order. This score directly measures the likelihood of an order resulting in a return.

Key Findings:

Based on the predictive model, the analysis revealed concentrated risk in specific areas:

1. **Synthetic Baseline Risk:** The analysis established a synthetic baseline return rate of approximately **8.40%** across all orders.

2. **Top Risk Drivers:** The model coefficients highlight that orders with **high implied discounts** and those flagged with **'Disputed'** or **'Cancelled'** statuses are the strongest predictors of a potential return. This suggests

that a combination of impulse purchasing (driven by steep discounts) and internal order processing issues significantly drives risk.

3. **High-Risk Product Concentration:** The predicted return risk is not evenly distributed. The product lines **Motorcycles**, **Classic Cars**, and **Vintage Cars** contain the highest number of products with the top average return risk scores, making them priority areas for intervention.