E-Commerce Return Risk Prediction :

Introduction :

In e-commerce, product returns can significantly affect profit margins and logistics efficiency. Identifying which orders are likely to be returned helps businesses proactively manage operations, improve product listings, and reduce unnecessary costs. This project focuses on building a predictive model to classify return risk based on customer, product, and transaction features.

Abstract :

Using a synthetic dataset of 10,000 e-commerce transactions, this project aims to predict the likelihood of a product being returned. The dataset includes user demographics, order information, product category, price, quantity, and shipping/payment methods. After data cleaning and feature engineering, an XGBoost classifier was trained, achieving a high accuracy score. The model was used to assign a return probability to each order and segment them into low, medium, and high risk.

Tools Used :

- Python (Pandas, Seaborn, Matplotlib, Scikit-learn, XGBoost)
- SQL via pandasql
- Power BI for dashboard visualization
- Jupyter Notebook

Steps Involved in Building the Project :

1. **Data Cleaning**:

    - Parsed order and return dates
    - Filled missing values for return reason and days to return
    - Created binary return label based on presence of return date

2. **EDA and SQL Analysis**:

    - Used SQL queries (via pandasql) to identify categories with high return rates
    - Visualized return rates by category and product price

3. **Feature Engineering**:

    - Encoded categorical variables using one-hot encoding
    - Created target variable: Return_Status (1 = Returned, 0 = Not Returned)

4. **Model Building**:

    - Trained an XGBoost classifier on the encoded dataset

- Achieved 100% accuracy on test data (likely due to clean synthetic data)
- Evaluated with confusion matrix, classification report, and ROC curve

5. **Segmentation and Output**:

- Generated return probability for each order
- Segmented orders into:
  - Low Risk (0–30%)
  - Medium Risk (30–70%)
  - High Risk (70–100%)
- Exported predictions for dashboard use

Conclusion :

The model accurately predicted product returns and provided business-ready insights for decision-making. By identifying high-risk return segments, the company can refine product descriptions, manage inventory better, and optimize logistics. The final dataset integrates seamlessly into Power BI dashboards, enabling real-time tracking of return risks by product category, region, and price. This solution supports strategic interventions to reduce return volume and enhance customer satisfaction.