# E-Commerce Return Risk Prediction Project Report

## INTRODUCTION

E-commerce has revolutionized the way consumers shop, offering convenience, accessibility, and competitive pricing. However, one major challenge that online retailers face today is the increasing rate of product returns. Returned products not only reduce profit margins but also increase logistical costs, impact inventory management ,and lower  customer satisfaction .This project aims to analyze product return patterns using a real-world e-commerce dataset and develop a predictive model to identify products with a high probability of return. By combining SQL-based business analysis, Python-based machine learning, and Power BI dashboards, this project provides actionable insights that help businesses proactively reduce returns and improvedecision-making. Through this integrated data-driven approach, the goal is to identify high-risk suppliers, categories, and customers, and to predict future return probabilities based on transaction characteristics and customer behavior.

## ABSTRACT

The “E-Commerce Return Risk Analysis and Prediction” project focuses on leveraging data analytics and predictive modeling to understand and minimize product returns in online retail.The project involves multiple stages — starting with data extraction and cleaning, followed by SQL analysis for business insights, predictive modeling using logistic regression in Python, and data visualization using Power BI. SQL queries were used to compute key performance indicators such as return percentages by category, supplier, and region. These insights helped identify which product groups contributed most to returns. Next, Python was used to build a logistic regression model that predicts the likelihood of a product being returned, using features like price, category, vendor, and delivery issues. Finally, Power BI was utilized to create an interactive “Return Risk Dashboard” that displays return rates, high-risk suppliers ,and predicted probabilities of returns.

The end result is a comprehensive data pipeline that not only explains past behavior but also predicts future risk — enabling business teams to act strategically.

## TOOLS USED

This project integrates multiple tools across the data analysis pipeline:

1. **SQL (PostgreSQL )** – Used for data cleaning, filtering, and performing aggregations such as total orders, returned products, and return percentages per category and supplier.
2. **Python (Google Colab)** – Used for machine learning tasks including data preprocessing, feature engineering, logistic regression modeling, and prediction of return probabilities.
3. **Power BI** – Utilized for creating interactive dashboards that visualize the results and allow drill-down filters by product, category, region, and supplier.
4. **Pandas & NumPy** – For data manipulation and numerical analysis in Python.
5. **Scikit-learn** – For implementing the logistic regression model and evaluating its accuracy through metrics such as ROC-AUC score and classification report.
6. **Matplotlib & Seaborn** – For visualizing relationships and patterns in the dataset.
7. **Joblib & CSV** – For saving and exporting model results, high-risk product lists, and probability scores for Power BI integration.

Together, these tools provide a complete workflow from raw data to intelligent business decisions.

## STEPS INVOLVED IN THE PROJECT

### 1️⃣ Data Collection & Cleaning

- Imported the raw e-commerce dataset containing order details, supplier information, and return flags.
- Removed duplicates, handled missing values, and standardized column formats (e.g., dates, numeric values).
- Created derived features such as order month, price range flag, and delivery issue indicators.
- Verified the dataset integrity using SQL queries and summary statistics in Python.

### 2️⃣ Data Analysis (SQL & Python)

- Used SQL queries to calculate return percentages for each product category and supplier.
- Identified high-return categories and vendors contributing most to product returns.
- Created KPIs like “Return Rate by Category”, “Supplier-Wise Return %”, and “Regional Order Distribution”.

### 3️⃣ Predictive Modeling (Python)

- Split the dataset into training and testing sets using train_test_split().
- Preprocessed categorical and numerical columns using OneHotEncoder and StandardScaler.
- Trained a Logistic Regression model with calibrated probabilities to predict the chance of return for each product.
- Evaluated performance using ROC-AUC, precision, recall, and accuracy metrics.
- Generated a “Return Probability” column and tagged products as *High Risk* or *Low Risk* based on a defined threshold (e.g., 0.25).

### 4️⃣ Result Export & Visualization (Power BI)

- Exported the model output (CSV files) containing product risk scores and probability values.
- Designed a Power BI dashboard showing:
    - Category-wise return distribution
    - Supplier performance overview
    - High-risk products and return probability ranking
    - Drill-through filters by region, supplier, and price range
- Created a composite “Return Risk Score” metric that integrates prediction results with historical return data.

### 5 Insight Generation & Business Recommendations

- Discovered that certain categories like *Electronics* and *Fashion* had significantly higher return rates.
- Identified specific suppliers with poor fulfillment or high defect rates.
- Suggested business actions such as improving quality checks, optimizing pricing, and offering targeted return policies.

## CONCLUSION

The E-Commerce Return Risk Prediction Project demonstrates how data analytics and predictive modeling can transform raw transactional data into powerful business insights. By integrating SQL-based exploratory analysis, Python-based logistic regression, and Power BI visualizations, this end-to-end project successfully identifies products and suppliers with a higher likelihood of returns. The logistic regression model provided a reliable prediction accuracy, while the Power BI dashboard translated the outcomes into a user-friendly visualization layer for stakeholders. This approach empowers business teams to anticipate return behavior, minimize operational losses, improve supplier accountability, and enhance customer satisfaction.Overall, this project not only delivers a scalable technical solution but also showcases the strategic value of combining **data engineering, analytics, and visualization** in solving real-world business challenges.