**📄 Project Summary: E-commerce Return Rate Prediction (Logistic Regression)**

**🔹 Objective**

To identify key factors driving product returns and build a logistic regression model to predict whether an order is likely to be returned.

**📊 Dataset Overview**

* **Source**: E-commerce order dataset
* **Records**: 70,052 rows
* **Features**: 19 columns (IDs, transaction details, quantities, revenues, return metrics)

**🔍 Data Preparation**

* Cleaned data using SQL Server (SSMS):
  + Added Is\_Returned flag based on quantity/refund logic
  + Standardized date format as Order\_Date
* Exported cleaned data to CSV for further analysis in Python (Jupyter Notebook)

**🧠 Modeling in Python**

* **Model**: Logistic Regression (Binary Classification)
* **Target Variable**: Is\_Returned (0 = not returned, 1 = returned)
* **Train-Test Split**: 80/20
* **Steps**:
  + Dropped identifiers (IDs, codes, names)
  + Converted categorical features (Category, Version) to dummies
  + Normalized structure for model training

**✅ Model Performance**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 100% |
| Precision (both classes) | 1.00 |
| Recall (both classes) | 1.00 |
| F1-Score (both classes) | 1.00 |

📌 *Note*: Such high scores suggest excellent feature separation — but also possibility of data leakage or very strong features like refund quantity directly indicating returns.

**📌 Top Features Influencing Returns**

| **Feature** | **Coefficient** | **Interpretation** |
| --- | --- | --- |
| Refunded\_Item\_Count | -1.36 | More refunds → more likely returned |
| Final\_Quantity | -1.22 | Negative/zero quantity → return indicator |
| Refunds | -0.62 | Actual refund amount correlates with return |
| Category\_Product H | +0.47 | Higher chance of return in this category |
| Overall\_Revenue | -0.43 | Lower revenue → higher return risk |

🔍 *Features with large negative coefficients push the prediction toward return.*