**E-commerce Return Rate Analysis**

**Introduction**

In the e-commerce industry, understanding customer return behavior is crucial for minimizing losses and improving satisfaction. This project focuses on identifying why customers return products and predicting which products are more likely to be returned. By analyzing sales and return data, companies can take preventive measures to reduce high return rates and optimize operations.

**Abstract**

The goal of the project is to forecast the likelihood of returns for products based on past orders and returns history. The data contains customer information, product information, and order details. Python was utilized for data cleaning, data transformation, and model building. Logistic Regression was employed to label products as "likely to be returned" or "not likely to be returned." The result of this model assists firms in identifying high-risk products and optimizing return management practices.

**Tools Used**

* Python – For coding as well as data analysis
* Pandas – To load and clean the data
* NumPy – To use for numerical computations
* Scikit-learn – To perform machine learning (Logistic Regression, preprocessing)
* Jupyter Notebook – To run codes and document
* Microsoft Excel / CSV – To store input and output data

**Steps Involved in Building the Project**

Data Loading

* The data (ecommerce\_returns\_synthetic\_data.csv) was loaded using Pandas.
* Columns like Order\_Date, Return\_Date, Product\_Category, Payment\_Method, and User\_Age were checked for uniformity.

Data Cleaning and Preprocessing

* Missing numeric values were filled in with median values.
* Missing categorical values were replaced with "Missing."
* Dates were changed into proper datetime format and new features like Days\_to\_Return were created.
* Categorical features like Product\_Category, Payment\_Method, and User\_Location were encoded with one-hot encoding.

Feature Engineering

* Added a new column Return\_Flag (1 if product was returned, 0 otherwise).
* Extracted new insightful columns like Age\_Bucket and reduced high-cardinality columns.

Model Development (Logistic Regression)

* Data was divided into training and testing sets.
* Preprocessing pipeline was formed with ColumnTransformer to process both numeric and categorical data.
* Logistic Regression model was trained on preprocessed data to predict the probability of returns of products.

Model Evaluation

* Assessed the model on Accuracy, Precision, Recall, and F1-Score.
* Created a list of products with high predicted probabilities of return and exported them to a CSV file for further investigation.

**Conclusion**

The project effectively employed machine learning to estimate product return probabilities from order and customer information. Logistic Regression yielded intuitive and efficient insights for the identification of high-risk products. Such insights can assist e-commerce websites in improving product descriptions, modifying marketing campaigns, and minimizing the overall return rate. The methodology can also be extended with more sophisticated models or combined into a recommendation system for more intelligent decision-making.

**DOCUMENTS FILES**

https://drive.google.com/file/d/1FCTiG1xtkWC4VoubSxPhJN7agM\_jrd9f/view?usp=drive\_link

https://drive.google.com/file/d/1Ow1Yiu13VFjPzwRkaqzyDn68ZdgyTL24/view?usp=drive\_link