Report: Classification Using Raw vs Pre-Processed Dataset

Objective

Build a machine learning model to classify customers based on their demographic and behavioral data. Two approaches were tested:

1. Using the raw dataset with minimal processing
2. Using a fully pre-processed version of the dataset

1. Classification Using Raw Dataset

Steps:

- Dropped rows with missing values.
- Dropped the 'Legacy_Customer_ID' and 'Customer_Feedback' columns.
- Categorical columns were encoded using Ordinal Encoding.
- Data was split into training and test sets (70% training, 30% testing), using stratification.
- A Random Forest Classifier was trained.
- Evaluated with accuracy, precision, recall, F1 score, and ROC AUC.

Notes:

- Simple and fast to implement.
- Dropping rows can lead to information loss.
- Ordinal encoding may not be suitable if there is no natural order.
- Lacks feature scaling and advanced handling.

2. Classification Using Pre-Processed Dataset

Motivation:

Most real-world datasets have missing values and varying data types. Preprocessing is needed.

Preprocessing Strategy:

- Numerical: impute with mean, scale with StandardScaler
- Categorical: impute with most frequent, One-Hot Encode
- Drop ID and text columns

Implementation:

- Used ColumnTransformer for different types
- Built a pipeline to include preprocessing and classifier

- Used Random Forest again

Advantages:

- Retains data with imputation
- One-hot encoding properly handles categories
- Scaling improves performance
- Suitable for production

Conclusion

Raw Dataset:

Pros:

- Simple, fast, and easy to implement

Cons:

- Discards data with missing values
- May misrepresent categorical data
- Lower accuracy compared to a pre-processed model

Pre-Processed Data:

Pros:

- More accurate and reliable results
- Handles missing values and categorical variables properly
- Suitable for production or real-world scenarios

Cons:

- Requires more setup and understanding of data preprocessing techniques