Things I did for data preprocessing:

Dimensionality Reduction - Pearson Correlation:

This part of the code is designed to compute the Pearson correlation matrix of the features and display it as a heatmap. High correlations between features can indicate redundancy, and such features might be candidates for removal to reduce dimensionality, simplify the model, and potentially reduce overfitting.

Dropping Unnecessary Columns:

The column labeled 'step' is dropped from the dataset. This indicates that the 'step' feature, which maps a unit of time in the real world, is not considered relevant for the subsequent analysis or model training. I also dropped columns that were highly correlated with others.

Feature Scaling (Z-score Standardization):

You apply Z-score standardization to specific numeric features: 'amount', 'oldbalanceOrg', 'newbalanceOrig', 'oldbalanceDest', and 'newbalanceDest'. Standardization adjusts the values so that they have a mean of 0 and a standard deviation of 1.

Custom Encoding for 'nameOrig' and 'nameDest':

You create two new binary features, 'customerTypeOrig' and 'customerTypeDest', based on whether 'nameOrig' and 'nameDest' start with 'C' (0) or 'M' (1). This is a form of custom encoding that simplifies these categorical variables for the model.

One-Hot Encoding for 'type':

The 'type' feature, which includes categorical data like CASH-IN, CASH-OUT, etc., is transformed using One-Hot Encoding. This creates binary columns for each category, ensuring that the model can interpret these categorical values correctly.

Feature Extraction:

New features, 'balanceDiffOrig' and 'balanceDiffDest', are created. These represent the differences in balance before and after the transaction for the originator and the recipient, respectively. Such feature engineering can help in highlighting patterns that might be predictive of certain behaviors, like fraudulent transactions.

Train-Test Split:

The dataset is split into training and testing sets with an 80-20 split.

Oversampling with SMOTE:

To address class imbalance (typically in cases where fraudulent transactions are much fewer than legitimate ones), SMOTE (Synthetic Minority Over-sampling Technique) is used. This technique generates synthetic samples for the minority class, thereby balancing the dataset and potentially improving model performance on minority classes.