# **Credit Card Fraud Detection - Imbalanced Data Handlin**g

## **Overview**

This code addresses the issue of imbalanced data in a credit card transaction dataset to enhance the performance of a machine learning model. The dataset contains information about legitimate and fraudulent credit card transactions.

## **Code Explanation**

1. Loading the Dataset:

* The credit card transaction data is loaded from a CSV file into a Pandas DataFrame.

2. Data Exploration:

* The distribution of legitimate and fraudulent transactions is inspected using value\_counts().

3. Data Separation:

* The dataset is split into two DataFrames: one for legitimate transactions and one for fraudulent transactions.

4. Data Sampling and Concatenation:

* A random sample of legitimate transactions is taken, and both the sample and the fraudulent transactions are concatenated to create a new balanced dataset.

5. Feature and Target Variable Definition:

* Features are obtained by dropping the 'Class' column, and the target variable is set as the 'Class' column.

6. Stratified Train-Test Split:

* The dataset is split into training and testing sets using train\_test\_split with the stratify parameter to maintain the class distribution in both sets.

7. Logistic Regression Model Training:

* A logistic regression model is created and trained on the balanced training dataset.

8. Model Evaluation:

* The accuracy of the trained model is evaluated on the training data.

## **Dealing with Imbalanced Data**

The code employs several strategies to handle imbalanced data:

* Under-sampling: A random sample of legitimate transactions is taken to balance the class distribution.
* Creating a Balanced Dataset: The under-sampled legitimate transactions are concatenated with the fraudulent transactions to create a new dataset with a balanced class distribution.
* Stratified Train-Test Split: When splitting the dataset, the stratify parameter is used to ensure that the class distribution is maintained in both the training and testing sets.

These techniques are essential for preventing model bias towards the majority class and improving the overall performance on imbalanced datasets.