PREDICTION OF CREDIT CARD FRAUD

A credit card is one of most used financial product for online payments and transactions. Though the credit cards can be a convenient way to manageyour finances,they can also be risky.

Credit card fraud is the unauthorised use of someone's credit card or credit card information to make purchases or withdraw cash.
It is important that credit card companies can be able to recognize fraudulent credit card transactions,so that customers are not charged for items that they did not purchase.

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days,where we have 492 fraudulent transactions out 2,84,807 transactions. The dataset is highly unbalanced.

The aim of this project is to build a classification model to predict a transaction is fraudulent or not.

Approach

Step 1: Data Understanding and Preparation

Load the data to understand it.Data understanding is critical since we will select the subset of features to carry out the model training.

The targer variable is "Class" that we need to predict ,

"0" means normal transaction
"1" means Fraudulent Transaction

Step 2: Exploratory Data Analysis

Histogram is used to compare each variable's data distribution pattern.
Make necessary data adjustments to avoid any problems while we train the model. Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations
This might include standardization, handling the missing values and outliers in the data.

Step 3: Class Imbalances

This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building .Here we have used two methods to balance the data

1. Under sampling : Balancing the data set by reducing the size of the abundant class.This method is using when quantity of data is less.

2. Over sampling : Balancing the dataset by increasing the size of the rare samples.This method is using quantity of data is large.

Here we have tried these two methods for balancing the dataset and comparing which technique is more efficient to balance the data.

Step 4: Data Modeling and Model selection

We will start building the model with the train test split,used 70-30 ratio to split the data.

We need to find which ML model works good with the imbalanced data and have better results on the test data.

We have used mainly three models here for evaluation.
1. Logistic Regression
2. Random Forest Classifier
3. Decision Tree

In these three models Random Forest classifier is more accurate and no need to improve model using hyper tuning .
Comparison of accuracy rates of these models are given below.

0]:

	Model	RUS	ROS
0	LR	93.918919	94.817316
1	RF	94.932432	99.995310
2	DT	87.837838	99.980067

RUS -Random Under Sampling Technique

ROS - Random Over Sampling Technique

Step 5: hyper parameter tuning with the model

At this time ,we will have the best understanding of the type of data we have and what kind of model we were going to build.After model building our next step will be hyper parameter tuing.This is step is essential to improve accuracy.
For Hyper tuning here we have used K-Fold cross validation method for evaluating the performance when the data is split into ‘k’ groups.
But here the best model is Random Forest which has a very good accuracy. So, do not need to do hyper tuning the same.
So, For understanding I have used the Hyper tuning parameter in Logistic Regression Model.

The Accuracy rates after Hyper tuning as shown below.

```
final_data_new
```

[117]:

	Model	RUS	ROS
0	LR	94.932432	94.732892
1	RF	94.932432	99.995310
2	DT	87.837838	99.980067

Step 5: Model Evaluation

Here we have evaluated the above three models for this unbalanced data set .The quantity of fraudulant transactions are very small compared to normal transactions. So for balancing the data set we can use oversampling and under sampling methods. But, when we use under sampling method, some datas may be missed. So, we can use oversampling for the same.

In this data set, We will get more accuracy when using Random Forest model as shown in the above table. This type of model evaluation will be key for banks to represent the business strategy to bring down the Fraud.