Name: Yogesh Chauhan
Program: MSc Data Science

Capstone Project on Detecting Credit Card Fraud

Introduction:

In the realm of banking, the primary objective for many institutions is to retain highly profitable customers. However, banking fraud poses a significant threat to this goal across various banks. This issue is a cause for concern due to its potential for substantial financial losses, as well as its impact on trust and credibility, affecting both banks and their customers. The advent of digital payment channels has only exacerbated this problem, leading to an increase in fraudulent transactions through novel methods.

Within the banking sector, the utilisation of machine learning for credit card fraud detection has evolved from being merely a trend to becoming an essential strategy. Implementing proactive monitoring and fraud prevention mechanisms has become imperative for banks. This case study aims to develop machine learning models to aid financial institutions in mitigating the challenges posed by time-consuming manual reviews, expensive chargebacks and fees, and the denial of legitimate transactions.

Problem Statement:

The banking industry has long acknowledged the necessity of developing a dependable machine learning model for detecting fraudulent credit card transactions. However, many banks continue to grapple with the persistence of credit card fraud, often stemming from ineffective identification and management of constraints. Failure to accurately identify constraints during suspicious transactions, thereby categorizing them as alarming, inevitably leads to subsequent frauds.
As the incidence of credit card frauds rises, the manual identification of such transactions becomes laborious and susceptible to human error. In essence, there exists a pressing need for a deeper comprehension of constraints in banking transactions and a systematic approach to identifying and modelling these constraints to effectively curb credit card fraud. This case study endeavours to pinpoint fraudulent credit card transactions utilising various machine learning models.

Data Exploration: Initially, we'll load the dataset and delve into understanding its various features:
a) We'll examine the shape and data types provided.
b) We'll assess the distribution or spread of the data.
c) Data cleaning will involve scrutinizing for NULL values and rectifying any discrepancies in data types, if detected. This process aids in selecting features essential for our final model.

Exploratory Data Analysis (EDA):

a) During this phase, we'll conduct univariate and bivariate analyses, and consider feature transformations as needed.
b) Since Gaussian variables are utilized in the current dataset, scaling won't be performed.
c) Addressing skewness in the dataset will involve employing log transformation or Box-Cox transformation.

d) Outliers, if any, will be addressed in this stage.

Train/Test Split:
a) This stage involves dividing the data into training and testing sets to evaluate model performance on unseen data.
b) For validation purposes, we'll employ the k-fold cross-validation method. Given data imbalance, stratified k-fold method will be utilized.

Model Development/Hyperparameter Tuning:
a) This final phase entails building different models and fine-tuning their hyperparameters to achieve the desired performance level on the dataset.
b) Various sampling techniques will also be explored to address data imbalance within this section.

Model Assessment:
Within this segment, we will assess the models utilizing suitable evaluation metrics. Given the presence of data imbalance, metrics such as precision, recall, F1-score, and AUR-RUC Score will be employed to gauge the model's performance.