Credit Card Fraud Detection Using Machine Learning

Name:	**Kartikey Singh**
Registration No./Roll No.:	20146
Institute/University Name:	Indian Institute Of Science Education And Research, Bhopal
Program/Stream:	Electrical Engineering And Computer Science (EECS)
Problem Release date:	January 12, 2023
Date of Submission:	March 01, 2023

1 Problem Statement

With the digitalisation of modern world banking systems, the issue of e-banking frauds have been on constant rise. People often tend to become a victim of these frauds either due to mismanagement of personal information or due to phishing attacks made via third party websites, mails or messages. According to the Global Banking Fraud Survey by **KPMG**, the amount of banking frauds have risen by **35** times in past 10 years. This alarming issue has led to a big bang in the field of development of fraud detection systems in recent years. But such systems are also prone to attacks by hackers and fail to protect users against malicious activities. In this project, we intend to present a quality and sustainable solution to this problem using **Machine Learning**.

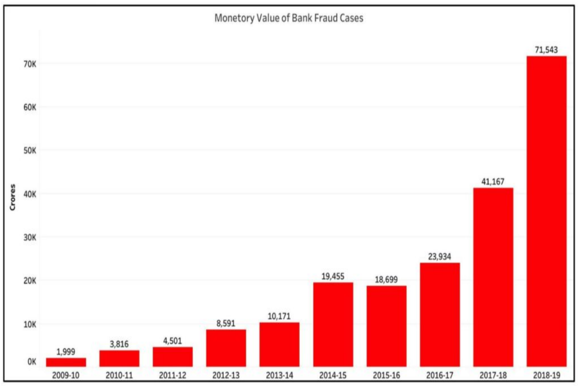

Figure 1: Increase In E-Frauds

2 Approach And Methods

2.1 Exploratory Data Analysis

In this section, we will be performing EDA on our dataset to get important insights and to find important features. The first method used for this purpose are the **correlation heat maps** to find which of the features are highly correlated to the target variable **Y**. For this we have made 3 different heat maps for clarity and tried to choose features which are highly correlated to Y.
To get a much better choice of features, we have used **Extra Tree Classifier**, also known as Extremely Randomized Tree Classifier is an inbuilt library in sklearn that is specifically designed for choosing best features according to a chosen importance function. In our case, we have chosen **Gini Entropy** function which is also taken as default parameter in this library. Next we also have tried to analyse features individually by plotting density and scatter plots for correlation of each of them. The link to the ipynb file is attached here.

2.2 Probable Methods And Techniques

We present the following solutions and techniques to be used in our model. All the below solutions will be tested and trained for their best performance using various cross validation techniques and fine tuning.

- The first solution we present is using simple **Logistic Regression** model.
- The next approach is to use the standard **Random Forest Classifier** model with **XGBoost**.
- Next possible solution we present is to use **Deep Learning i.e using Artificial Neural Network** with regularisation or hyper parameter tuning that can directly weed out irrelevant features and keep the model free from any overfit.
- Another solution we intend to present is using **K Nearest Neighbours**. This would include trying out various variants of this technique.
- The last solution we intend to present is actually a new idea which can be briefly termed as a **Weighted Ensemble of Models (WEM)**. In this approach, we first tabulate the results predicted by each of the above models, their accuracy (as denoted by number of true positives and true negatives) and then pass these metrics through a filteration process that re-evaluates labels of each data instance and finally the new results are given out as final prediction.