Capstone Project on Credit Card Fraud Detection- using Machine Learning

Introduction: Credit card fraud is a significant issue in the financial industry, causing billions of dollars in losses each year. With the increasing number of transactions and the rise of online shopping, detecting fraudulent activities has become more challenging. Machine learning provides a powerful toolset to identify and prevent fraudulent transactions by analysing patterns and anomalies in transaction data.

In today's world, we are on the express train to a cashless society. According to the World Payments Report, in 2016 total non-cash transactions increased by 10.1% from 2015 for a total of 482.6 billion transactions! That's huge! Also, it's expected that in future years there will be a steady growth of non-cash transactions.

Now, while this might be exciting news, on the flip-side fraudulent transactions are on the rise as well. Even with EMV smart chips being implemented, we still have a very high amount of money lost from credit card fraud.

This is now becoming a serious problem since most of the time, a person who has become a victim of this fraud don't have any idea about what has happened until the very end. So in this project, what we have tried is to create a Web App for the detection of such type of frauds with the help of Machine Learning. In the following sections, we will be explaining about the creation and importance of both a good Machine Learning model.

Objectives:

The primary objectives of this report are:

- To understand the nature of credit card fraud.
- To explore the dataset used for fraud detection.
- To develop and evaluate machine learning models for detecting fraudulent transactions.

Steps That are Carried in this Project

1-Data Sourcing: The Data was taken from the document of the capstone project

Features of the data set:

- Time: The seconds elapsed between this transaction and the first transaction in the dataset.
- V1 to V28: Principal components obtained with PCA to protect user identities and sensitive features.
- Amount: Transaction amount.
- Class: Response variable (1 for fraud, 0 for non-fraud).

Understanding the data and related constraints:

1-Since the data for this project is very unbalanced due to the fact that number of cases of Fraud transactions are very low in comparison to number of cases of Valid transactions makes the model training a bit hectic.

2-So, now, consider the fact that if our data have 98% of the values to be valid while only 2% to be frauds, if our model predicts all values to be valid, it will eventually achieve 98% accuracy at the end of the day, but the model will be an absolute wastage.

2-Preprocessing data: Data preprocessing is a crucial step in the machine learning pipeline, especially for credit card fraud detection. It involves transforming raw data into a clean and structured format suitable for model training. Key steps in data preprocessing include:

1. Data Cleaning: Identifying and handling missing, duplicate, or inconsistent data to ensure accuracy and completeness.
2. Data Transformation: Normalizing or scaling numerical features to ensure they are on a comparable scale, which helps improve model performance.
3. Encoding Categorical Variables: Converting categorical data into numerical format using techniques like one-hot encoding or label encoding.
4. Handling Imbalanced Data: Addressing the class imbalance between fraudulent and non-fraudulent transactions through methods such as oversampling, undersampling, or using specialized algorithms.
5. Feature Selection and Engineering: Identifying and creating relevant features that can improve the predictive power of the model while reducing dimensionality.

By meticulously preprocessing the data, we enhance the model's ability to accurately detect fraudulent transactions and ensure robust, reliable performance in real-world applications.

3- Model Selection: Choose appropriate machine learning algorithms such as Logistic Regression. Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome

4-Splitting Data into Train and Test Data: Train-test splitting is a fundamental step in the machine learning workflow, aimed at evaluating the model's performance on unseen data. This process involves dividing the dataset into two subsets: one for training the model and one for testing

1. Purpose:
    - Training Set: Used to train the machine learning model. The model learns patterns and relationships within this data.
    - Test Set: Used to evaluate the model's performance. This set simulates how the model will perform on new, unseen dat a
    - Splitting Ratio: Typically, the dataset is split into 70-80% for training and 20-30% for testing. Common ratios include 70/30 or 80/20 splits.

- Stratification: For imbalanced datasets, such as fraud detection where fraudulent transactions are rare, stratified splitting ensures that both training and test sets maintain the same proportion of classes as the original dataset.
- By performing a train-test split, we can train the model on one subset of the data and validate its performance on another, helping to ensure that the model generalizes well to new, unseen data and doesn't overfit to the training data.

5- Model Evaluation: Model evaluation aims to define how well the model performs its task. The model's performance can vary both across use cases and within a single use case, e.g., by defining different parameters for the algorithm or data selections. As our Model's Accuracy score was 0.92 which is 92% out of 100 our model has performed very well and will be accepted

6- Conclusion: Credit Card is a great tool to pay money easily, but as with all the other monetary payment tools, reliability is a issue here too as it is subjected to breach and other frauds. To encounter this problem, a solution is needed to identify the patterns in the transactions and identify the ones which are fraud, so that finding such transactions beforehand in future will be very easy. Machine Learning is a great tool to do this work since Machine Learning helps us in finding patterns in the data. Machine Learning can help producing great results if provided enough amount of data. Also, with further advances in the technology, Machine Learning too will advance with time, it will be easy for a person to predict if a transaction is fraud or not much more accurately with the advances.

Challenges Faced During This Project

- The challenge is to recognize fraudulent credit card transactions so that the customers of credit card companies are not charged for items that they did not purchase.

Main challenges involved in credit card fraud detection are:

- Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.
- imbalanced Data that is most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones
- Data availability as the data is mostly private.
- Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.

How to tackle these challenges?

- The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible.
- Imbalance can be dealt with by properly using some methods which we will talk about in the next paragraph
- For protecting the privacy of the user the dimensionality of the data can be reduced.
- A more trustworthy source must be taken which double-check the data, at least for training the model.
- We can make the model simple and interpretable so that when the scammer adapts to it with just some tweaks we can have a new model up and running to deploy.