Credit Card Fraud Detection

1. Domain Background

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications with much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. fraud detection models can help the companies and the banking tracks to improve their services and gain the customer's trust, moreover, that will play a main role in the country's economy. It's like fraud crimes on the earth, now with the digital transformation as the technologies grow, the criminal will find a way to reach victims on a larger scale by using the wrong side of the technologies , from here comes the role of technical engineers to reduce and detect fraud crimes in the best way. As the world knows now, the role of machine learning and its many uses that are countless so far, it is one of the best ways to detect fraud crimes.

2. Problem statement

The Credit Card Fraud Detection Problem includes modeling past credit card transactions with the knowledge of the ones that turned out to be a fraud. This model is then used to identify whether a new transaction is fraudulent or not. Our aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

3. Dataset and inputs

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 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The dataset includes 31 features and 2 classes . Also, the dataset contains imbalanced classes with 87 for class 1(fraud cases) and 56874 for class 0 (genuine cases). In addition the dataset needs to be preprocessed since there are some columns contained some NaN that need to be cleaned before building the model . otherwise the features are all numerical values.

4. Solution statement

This model aims to reduce cases of credit card fraud, by machine learning classification algorithm which is an algorithm that divides the data into two classes, and by using some techniques to clean and analyze the dataset. Then determine the extent of the correlation of the features in detecting fraud cases and then building the model using the classification algorithm, training it, then testing it on an unseen dataset. Finally evaluating it and discovering its effectiveness and efficiency using criteria for evaluating classification algorithms.

5. Benchmark Model

Building a support vector machine model with a linear kernel has register accuracy 97.5% . Our goal is to build a supervised model with a better result than SVM [1] .

6. Evaluation Metrics

Deciding the best model for the used dataset is determined by measuring the accuracy, precision, and recall of the model, and since the dataset contained imbalanced classes the cross-validation method will be used to split the data into testing and training set k times.

7. Project Design

- **Software and Hardware**

The model will be build Python (version 3) and Jupyter notebook. For the pre-process of the dataset the project will use the Pandas and Numpy libraries, and matplotlib.pyplot and scikit-learn libraries for data visualization and machine learning, respectively.

- **Data Pre-processing**

During the preprocessing phase and since the dataset contains some symbols, these values will be replaced by the mean for each numeric feature or the most repetition values for the binary features. Then, the dataset will be splitted using the cross validation stratified k fold into training and testing sets to handle the imbalanced classes.

- **Model Design**

Three models will be build: Logistic Regression . Support Vector Machine (SVM). and Random Forest (RF). After that, the model with high score will be tuned in order to enhance the model and get a better result.

8. References

[1] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” *Stat. Sci.*, vol. 17, no. 3, pp. 235–255, 2002, doi: 10.1214/ss/1042727940.