CREDIT CARD FRAUD DETECTION

BY

DWAIPAYAN JANA

&

DEBAJYOTI ROY

BATCH:2023

INSTITUTE NAME: ASUTOSH COLLEGE

PROJECT GUIDE: KOUSTAB GHOSH

SUBMITTED TO: INSTITUTE OF DATA ENGINEERING, ANALYTICS AND SCIENCE FOUNDATION, ISI KOLKATA

ABSTRACT

Due to the increasing number of customers as well as the increasing number of companies that use credit cards for ending financial transactions, the number of fraud cases has increased dramatically. Dealing with noisy and imbalanced data, as well as with outliers, has accentuated this problem. In this work, fraud detection using artificial intelligence is proposed. The proposed system uses logistic regression to build the classifier to prevent frauds in credit card transactions. To handle dirty data and to ensure a high degree of detection accuracy, a pre-processing step is used. The pre-processing step uses two novel main methods to clean the data: the mean-based method and the clustering-based method. Compared to two well-known classifiers, the support vector machine classifier and voting classifier, the proposed classifier shows better results in terms of accuracy, sensitivity, and error rate.

INTRODUCTION

According to the definition of fraud, the aim of fraud is to achieve personal or financial gain through deception. Based on this, fraud detection and prevention are the two significant methods for avoiding loss due to fraud. Fraud prevention is the proactive technique for avoiding the occurrence of fraudulent acts, and fraud detection is the technique for the detection of fraudulent transactions by fraudsters. A variety of payment cards, including credit, charge, debit, and prepaid cards, are currently widely available. They are the most popular means of payment in some countries. Indeed, advances in digital technologies have paved the way for changes in how we handle money, especially for payment methods that have changed from being a physical activity to a digital activity using electronics means. This has revolutionized the landscape of monetary policy, including the business strategies and operations of both large and small companies. Credit card fraud is the fraudulent use of credit card details to buy a product or service. These transactions can be physically or digitally performed. In physical transactions, the credit card is physically present. On the other hand, digital transactions take place over the internet or telephone. A cardholder normally provides their card number, card verification number, and expiration date through a website or telephone call. With the rapid rise in e-commerce over the past few years, credit card use has increased tremendously.

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.

It contains only numerical input variables which are the result of a PCA transformation.

Due to confidentiality issues, there are not provided the original features and more background information about the data.

Features V1, V2, ... V28 are the principal components obtained with PCA; The only features which have not been transformed with PCA are Time and Amount. Feature Time contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature Amount is the transaction Amount, this feature can be used for example- dependent cost-sensitive learning. Feature Class is the response variable and it takes value 1 in case of fraud and 0 otherwise

OBJECTIVES

The objectives of a credit card fraud detection project using machine learning typically include:

1. **Detecting Fraudulent Transactions**: Implementing models that can accurately identify transactions that are likely to be fraudulent.
2. **Reducing False Positives**: Minimizing the number of legitimate transactions that are mistakenly flagged as fraudulent.
3. **Improving Accuracy**: Developing models with high precision and recall to correctly identify fraudulent activities while minimizing errors.
4. **Real-time Detection**: Creating systems capable of detecting fraud in real-time or near real-time to prevent financial losses promptly.
5. **Handling Imbalanced Data**: Addressing the challenge of imbalanced datasets where fraudulent transactions are much less frequent than legitimate ones.

These objectives collectively aim to create a robust and effective system that protects both consumers and financial institutions from the risks associated with credit card fraud.

METHODOLOGY

Data Pre-processing:

Pre- processing of credit card fraud detection data for machine learning involves several steps to ensure that the data is cleaned, transformed, and prepared in a way that enhances the performance of the models. Here's a detailed approach to pre-process credit card fraud detection data using machine learning techniques:

1. **Data Cleaning:**

The credit card dataset was imported using the python import command, and the data cleaning process was done. During data cleaning we perform removal of null values, and missing values.

The dataset contains 284807 transactions in total. There were no null values in the dataset. Also, our dataset does not have any missing value.

2. **Encoding categorical variables:**

After cleaning the dataset, we convert any categorical features to a numeric value as most machine learning algorithms perform better with numeric inputs. There are few ways to convert categorical values into numeric values with each approach having its own trade offs and impact on the feature set. In the study, we have used One-Hot Encoder to convert the categorical variables to numeric values. For a feature with two categories, the categories are assigned a numeric value of 1 or 0.

3. **Feature scaling:**

This is another stage of the data preprocessing method used to normalize the range of independent variables within a dataset. Depending on the adopted scaling technique, it is centred around 0 or in the range of 0 and 1. If input variables have tremendous values applicable to the additional input variables, these large values can overlook or skew some machine learning algorithms. We have performed feature scaling. Scaling can be achieved by calculating the median 50th percentile, the 25th, and 75th percentiles.

4. **Dataset re-sampling:**

Data resampling is a technique of inexpensively using a data sample to improve the accuracy and measure the unpredictability of a population

variable. The nested resampling method has been used to carry out dataset resampling. The dataset used for this study was highly imbalanced; that is why we have carried out resampling methods like

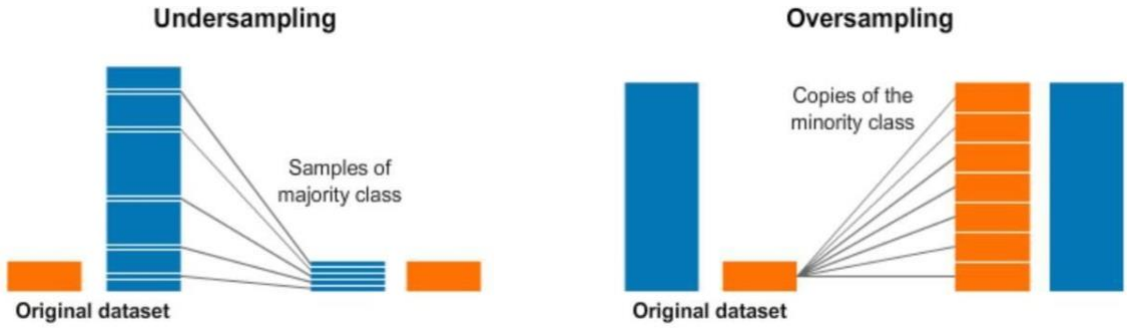
Undersampling and **Oversampling.**

Undersampling:

Undersampling is a technique used in the context of imbalanced classification problems, where the number of instances of one class (usually the minority class) is significantly lower than the number of instances of the other class (majority class). The goal of undersampling is to balance the class distribution by reducing the number of instances in the majority class to match the number of instances in the minority class.

Oversampling:

Oversampling is a technique used to address the issue of imbalanced datasets in machine learning, where one class (typically the minority class) is underrepresented compared to another class (majority class). The goal of oversampling is to increase the number of instances in the minority class to balance the class distribution and improve the performance of models, especially in tasks like fraud detection, medical diagnosis, or anomaly detection where the minority class is of particular interest.



5. **Feature correlation and selection:**

Each of the features we obtain in the dataset might not be beneficial in building a machine learning model to execute the necessary prediction. Using some of the features might improve the prediction accuracy. So, feature correlation performs a tremendous purpose in creating a better machine learning model. Features with high correlation are more likely to be linearly dependent and have almost the same impact on the dependent variable. Therefore, when two features produce a high correlation, we can drop one of

the two features. The heatmap for the correlation of the original dataset is shown below. It can be observed that the heatmap is not revealing too much information because it's a huge dataset, and that is why we performed feature selection to help select the important features. Feature selection is one of the important stages in data preprocessing, and it is known as a path to capture relevant features for use in the implementation of the machine learning model to improve the learning interpretability and decrease the model over-fitting when there are many unnecessary features contributing no more helpful information than the current subset of variables.

In our project we use **Logistic Regression Model**.

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyse the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modelling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

Advantages of Logistic Regression:

1. **Easier to implement machine learning methods**: A machine learning model can be effectively set up with the help of training and testing. The training identifies patterns in the input data (image) and associates them with some form of output (label). Training a logistic model with a regression algorithm does not demand higher computational power. As such, logistic regression is easier to implement, interpret, and train than other ML methods.

2. **Suitable for linearly separable datasets**: A linearly separable dataset refers to a graph where a straight line separates the two data classes. In

logistic regression, the y variable takes only two values. Hence, one can effectively classify data into two separate classes if linearly separable data is used.

3. **Provides valuable insights**: Logistic regression measures how relevant or appropriate an independent/predictor variable is (coefficient size) and also reveals the direction of their relationship or association (positive or negative).

Disadvantages of Logistic Regression:

1. **Limited to Linear Relationships**: Logistic regression assumes a linear relationship between the independent variables (predictors) and the log odds of the dependent variable (outcome). If the true relationship is non-linear, logistic regression may underperform unless non-linear transformations of predictors are included.

2. **Assumption of Independence of Errors**: Logistic regression assumes that observations are independent of each other. If there is dependence among observations (e.g., time series data, spatial data), this assumption may be violated, leading to biased estimates and inaccurate predictions.

3. **Not Suitable for Complex Relationships**: It may not perform well with complex relationships where multiple interactions or higher-order terms among variables are important. In such cases, more flexible models like decision trees or neural networks might be more appropriate.

Model creation:

In this section, we present the specifications on model creation. Following preprocessing the dataset, data are split into training and test. The training

data is used to define the parameters for the models while the test set is used to evaluate our models.

Splitting of data into training and test-

The main objective of the machine learning model is to learn from previous experience and its ability to make use of the information to generate new instances. Performance evaluation of the model is usually done on the subset of the whole dataset by training on it, and the remaining dataset can be used to evaluate the model's performance. In this study, our dataset was split into a 80:20 ratio; that is, 80% of the dataset is used for training the model and the remaining 20% to evaluate the model's performance. Parameters, often called hyperparameters of the model, are determined during model training, and these hyperparameters also helped find the best model fit for a machine learning model.

DATA ANALYSIS AND RESULTS

To evaluate the performance of our model, we adopted the use of a metric called AUC score and other metrics to evaluate the performance of our model. The Metrics of each model will be shown based on how they have performed with our original, then we present a comparative study to determine which of our model is the best for predicting of credit card fraud.

METRICES:

Evaluating the performance of the machine learning algorithms is an essential part of any project work. This will show how each of the algorithms performed and to know which gives satisfactory or unsatisfactory results. We often use accuracy to weigh the model performance in classification algorithms, although it is not the only true way to judge the model. In this study, evaluation metrics like F1-Score, Precision, Recall, Confusion matrix, Accuracy, and ROC AUC Score are used.

ACCURACY:

Accuracy is the ratio of the correct prediction number to the total number of input samples. It functions admirably just if there are an equivalent number of samples having a place with each class. For instance, consider 98% examples of class A and 2% examples of class B in our training set. Then, at that point, our model can undoubtedly get 98% accuracy by basically anticipating each training sample to be allied to class A. When a similar model is tried on a test set with 60% examples of class A and 40% examples of class B, then, at that point, the test accuracy would be reduced to 60%. Classification Accuracy is extraordinary; however, it gives us the misguided feeling of accomplishing high precision.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

RECALL:

Recall can be calculated when the correct positive number results are divided by the number of all samples, which should have been recognized as a positive value.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

PRECISION:

Precision is dividing the correct positive number results by the number of positive results that the classifier predicted.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

F1-SCORE:

F1-score is used to evaluate the test's accuracy. It is the consonant mean between recall and precision. It allows a report on how precise the Classification is and how strong it can be. If a result gives high precision but low recall, it means we have incredibly high accuracy but note; it may miss a very high number of possibilities that are hard to classify. In short, it means the higher the F1 score, the best the model performed. It can be calculated using

$$F1 = 2 \times \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

CONFUSION MATRIX:

Confusion Matrix gives us a complete breakdown of the model performance in terms of matrix output. It evaluates well, especially when working with a binary classification where we have samples that belong to two classes: TRUE or False, YES or NO.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

The four important terms we have are True Positives, True Negatives, False Positives, and False Negatives.

- True Positives: this is the case where the algorithm predicted YES, and the true output came out YES.
- True Negatives: this is the case where the algorithm predicted NO, and the true output came out NO.
- False Positives: this is the case where the algorithm predicted YES, and the true output came out NO.
- False Negatives: this is the case where the algorithm predicted NO, and the true output came out YES. The accuracy of the confusion matrix can be calculated by

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

ROC AUC Score:

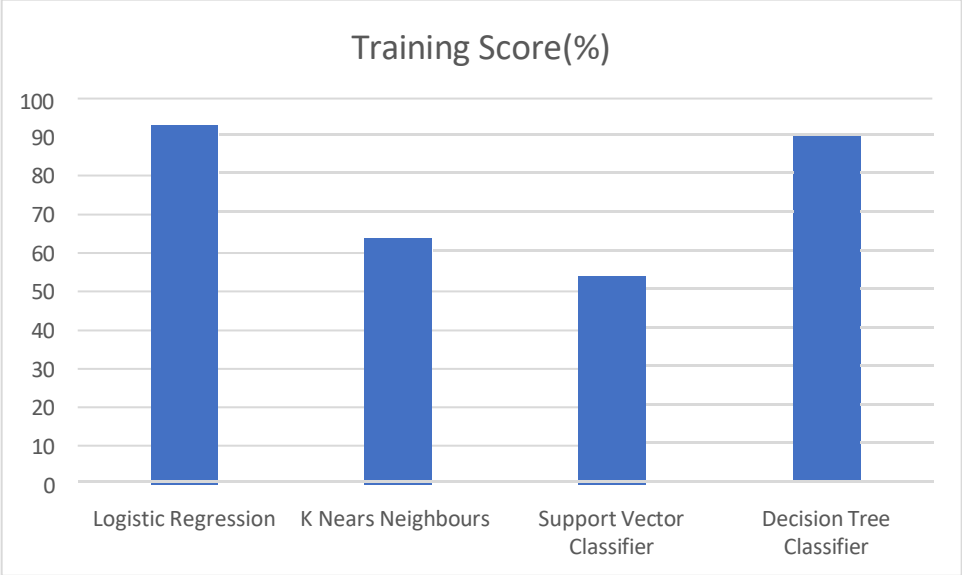
ROC (Receiver Operating Characteristics) AUC (Area Under Curve) is a widely used metric for model evaluation. AUC is the degree of measurement for separability, which reports how the model can differentiate between classes.

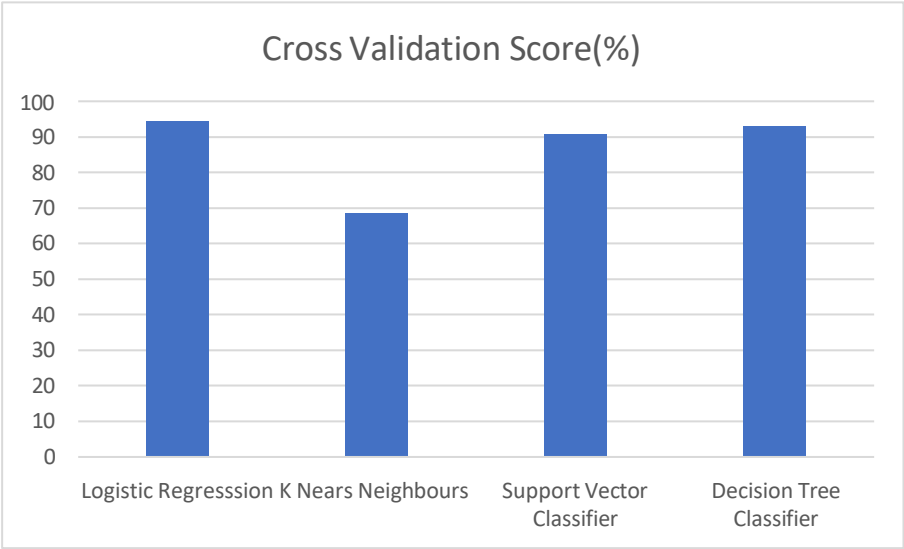
Classification problems should measure performance with different thresholds been set. A better model can predict 0 classes as 0 and 1 classes as 1, while this can be confirmed if the AUC score is high. ROC is the curve probability.

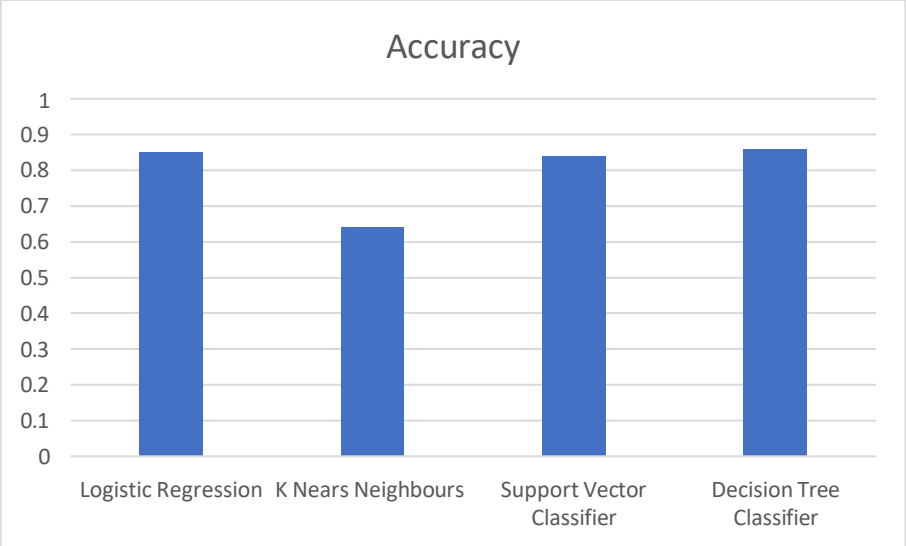
COMPARITIVE ANALYSIS:

In this section, a comparative analysis of our model was made based on the types of datasets and the result of the metrics used to measure how each algorithm has performed. Based on the training score, Cross Validation Score & accuracy we have pick the best overall model.

Methods	Training Score(%)	Cross Validation Score(%)	Accuracy
Logistic Regression	93	94.28	0.85
K Nears Neighbours	64	68.49	0.64
Support Vector Classifier	54	90.72	0.84
Decision Tree Classifier	90	92.88	0.86

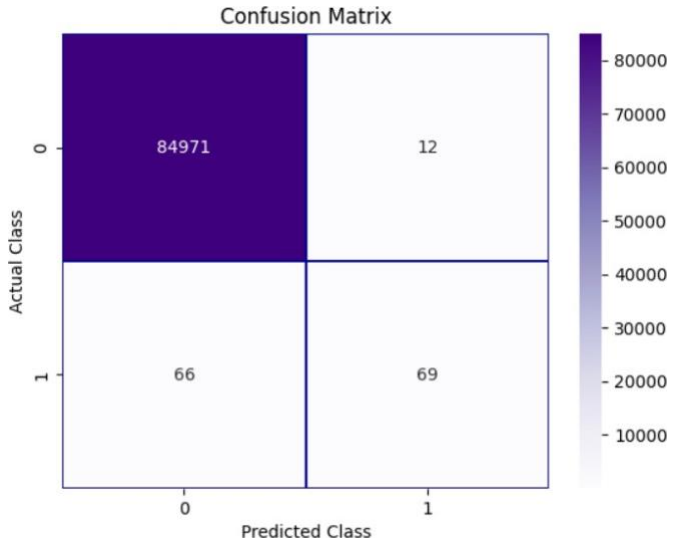


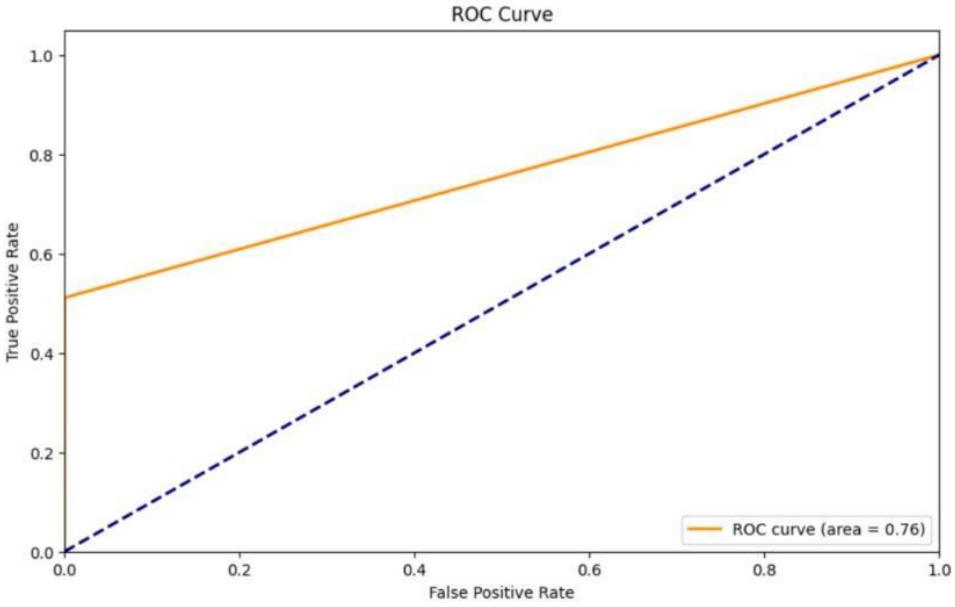




Logistic Regression model performed best in 2 parameters. So it is the best estimator for our data set.

Confusion matrix and ROC curve of Logistic Regression model is given below:





ROC-AUC Score of Logistic Regression is 0.9348

CONCLUSION

The detection of credit card fraud is a vital research field. This is because of the increasing number of fraud cases in financial institutions. This issue opens the door for employing artificial intelligence to build systems that can detect fraud. Building an AI-based system to detect fraud requires a database to train the system (or classifier). The data in reality are dirty and have missing values, noisy data, and outliers. Such issues negatively affect the accuracy rate of the system. To overcome these problems, a logistic regression-based classifier is proposed.

APPENDICES

1. REFERECES

i. Credit Card Fraud Detection Database, Anonymized credit card transactions labelled as fraudulent or genuine, https://www.kaggle.com/mlg-ulb/creditcardfraud

ii. Principal Component Analysis, Wikipedia Page, https://en.wikipedia.org/wiki/Principal_component_analysis

iii. ROC- AUC Characteristics, https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve

2. **COLLAB FILE**
https://colab.research.google.com/drive/1DWj4t4JiEmTU_3QVJ4hLgdDFyNsU-VUZ?usp=sharing#scrollTo=I-7E1Zxbd7-a