Institute of Information Technology

Jahangirnagar University

Savar, Dhaka-1342

Credit Card Fraud Detection Using Machine Learning

Submitted To:

Md. Mahmudur Rahman

Lecturer
Institute of Information Technology
Jahangirnagar University

Submitted By:

Group - 23

Name	Roll
Md. Shakil Hossain	2023
Mahbubur Rahman	2024
Nahidul Islam	2028

ABSTRACT

Credit card fraud is a significant problem, with billions of dollars lost each year. Machine learning can be used to detect credit card fraud by identifying patterns that are indicative of fraudulent transactions. Credit card fraud refers to the physical loss of a credit card or the loss of sensitive credit card information. Many machine-learning algorithms can be used for detection. This project proposes to develop a machine-learning model to detect credit card fraud. The model will be trained on a dataset of historical credit card transactions and evaluated on a holdout dataset of unseen transactions.

INTRODUCTION

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Fraud has been increasing drastically with the progression of state-of-art technology and worldwide communication. Credit cards are one of the most prevalent fraud objectives but not the only one. Credit card fraud is the wide-ranging term for theft and fraud committed or any similar payment mechanism as a fraudulent resource of funds in a transaction. Credit card fraud has been an expanding issue in the credit card industry. Detecting credit card fraud is difficult when using normal processes, so developing credit card fraud detection models has become essential in academic or business organizations. Fraud can be avoided in two main ways: prevention and detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Detection happens once the prevention has already failed. Therefore, detection helps identify and alert as soon as a fraudulent transaction is triggered.

Machine learning is this generation's solution, which replaces such methodologies and can work on large datasets, which is impossible for human beings. Machine learning techniques fall into two main categories: supervised and unsupervised. Fraud detection can be done either way and can only be decided when to use according to the dataset. Supervised learning requires prior classification of anomalies. During the last few years, several supervised algorithms have been used in detecting credit card fraud. The data used in this study is analyzed in two main ways: categorical data and numerical data. The dataset initially came with categorical data. The raw data can be prepared by data cleaning and other basic preprocessing techniques. First, categorical data can be transformed into numerical data, and then appropriate techniques are applied for the evaluation. Secondly, categorical data is used in machine learning techniques to find the optimal algorithm.

This project consists of selecting optimal algorithms for fraud patterns through an extensive comparison of machine learning such as Logistic Regression, KNN Neighbors, and Decision Trees—techniques via an effective performance measure for detecting fraudulent credit card transactions. The rest of this paper is presented as follows. Section 2 offers the literature review. Section 3 provides the experimental methodology, including results. Finally, conclusions and discussions of the paper are presented in Section 4.

LITERATURE REVIEW

In earlier studies, many approaches have been proposed to bring solutions to detect fraud from supervised methods and unsupervised techniques to hybrid ones, which makes it a must to learn the technologies associated with credit card fraud detection and to have a clear understanding of the types of credit card fraud. With the analysis of various detection models, past researchers have found many problems regarding fraud detection. Classical algorithms such as Support Vector Machines (SVM), Decision Tree (DT), LR and RF have proven useful.

In the paper [1], a European dataset was also used, and a comparison was made between the models based on LR, DT and RF. Among the three models, RF proved to be the best, with an accuracy of 95.5%, followed by DT with 94.3% and LR with an accuracy of 90%.

According to [2] and [3], k-nearest neighbors (KNN) and outlier detection techniques can also be efficient in fraud detection. They have proven helpful in minimizing false alarms and increasing fraud detection rates.

KNN algorithm also performed well in the experiment for paper [4], where the authors tested and compared it with other classical algorithms. The article [5] discussed commonly used supervised techniques, and they have provided a thorough evaluation of supervised learning techniques. Also, they have shown that all algorithms change according to the problem area.

The fraud detection system presented in the paper [6] is built to handle class imbalance, form labelled and unlabeled, and process large datasets. The proposed method was able to overcome all the challenges.

The paper [7] highlighted fraud detection cost and lack of adaptability as challenges in the fraud detection process. The cost of fraudulent behavior and prevention cost should be considered when considering a system. Lack of adaptability occurs when the algorithm is exposed to new types of fraud patterns and standard transactions.

PROPOSED METHODS

In this project, we will use three Algorithms to determine whether a card is actual or fraudulent. Description of these Algorithms are given below:

Logistic Regression:

This statistical classification model based on probabilities detects fraud using a logistic curve. Since the value of this logistic curve varies from 0 to 1, it can be used to interpret class membership probabilities. The dataset fed as input to the model is classified for training and testing the model. Post-model activity is tested for some minimum threshold cut-off value for prediction. Based on some threshold probabilities, the logistic regression can divide the plane using a single line and divide dataset points into exactly two regions.

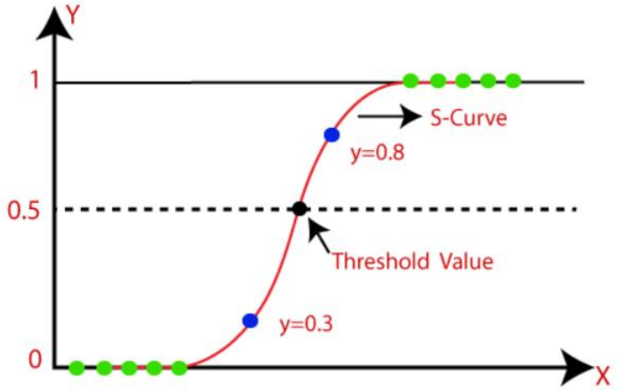


Fig: The logistic regression model

K-Nearest Neighbor (KNN):

This supervised learning technique achieves consistently high performance compared to other fraud detection techniques of supervised statistical pattern recognition [24]. Three factors majorly affect its performance distance to identify the least distant neighbors. There are some rules to deduce a categorization from the k-nearest neighbor & the count of neighbors to label the new sample. This algorithm classifies transactions by computing the least distant point to this particular transaction. If this least distant neighbor is classified as fraudulent, the latest marketing is also labelled as fraudulent. Euclidean distance is an excellent choice to calculate the distances in this scenario. This technique is fast and results in fault alerts. Its performance can be improved by distance metric optimization.

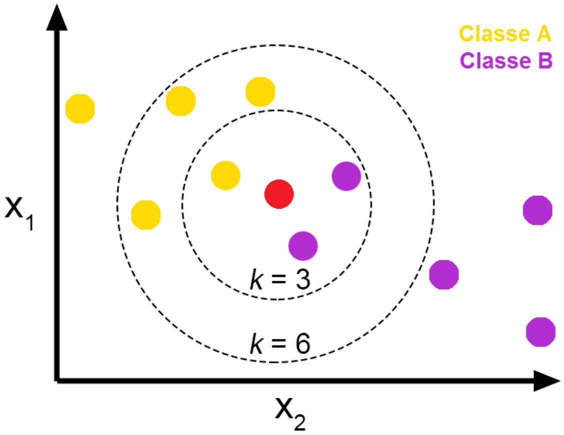


Fig: Pros and Cons of K-Nearest Neighbors - From The GENESIS

Decision Tree:

A supervised learning algorithm is a decision tree in the form of a tree structure consisting of the root node and other nodes split in a binary or multi-split manner further into child nodes, with each tree using its algorithm to perform the splitting process. With the tree growing, there may be possibilities of overfitting the training data with possible anomalies in branches, some errors or noise. Hence, pruning is used for improving classification performance of the tree by removing specific nodes. Ease in use and the flexibility that the decision trees provide to handle different data types of attributes make them quite popular.

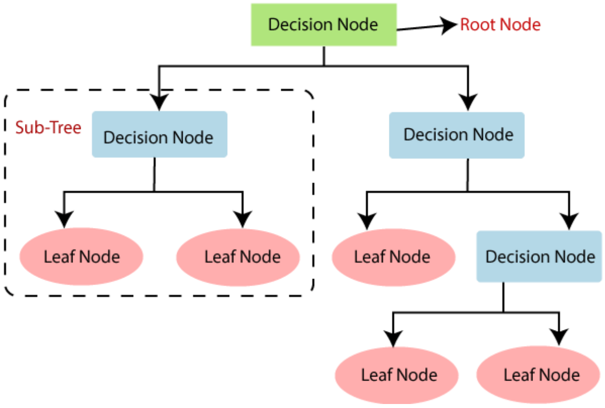


Fig: Decision Tree Algorithm in Machine Learning

Support Vector Machine:

Support vector machines or SVMs are linear classifiers, as stated in that work in high dimensionality because, in high dimensions, a non-linear task in input becomes linear. Hence, this makes SVMs highly useful for detecting fraud. Its two most important features that is a kernel function to represent the classification function in the dot product of input data point projection and the fact that it tries finding a hyperplane to maximize separation between classes while minimizing overfitting of training data; it provides a very high generalization capability.

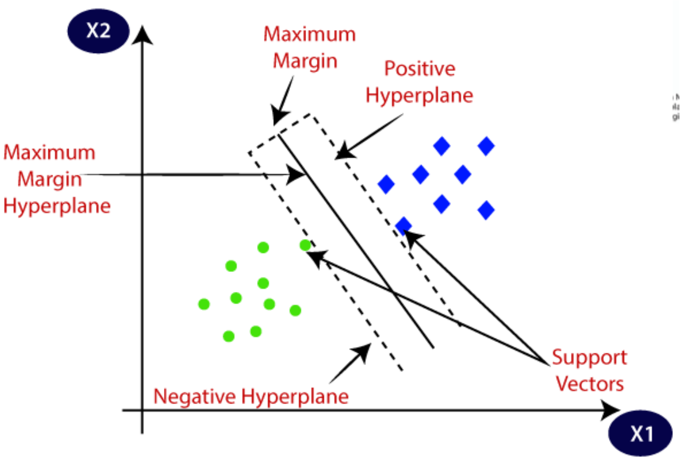


Fig: Support Vector Machine algorithm.

Dataset:

The Credit Card Fraud Detection dataset was used in this research, which can be downloaded from Kaggle [8]. This dataset contains transactions that occurred in two days and were made in September 2013 by European cardholders.

[Credit Card Fraud Detection]

PROJECT PLAN

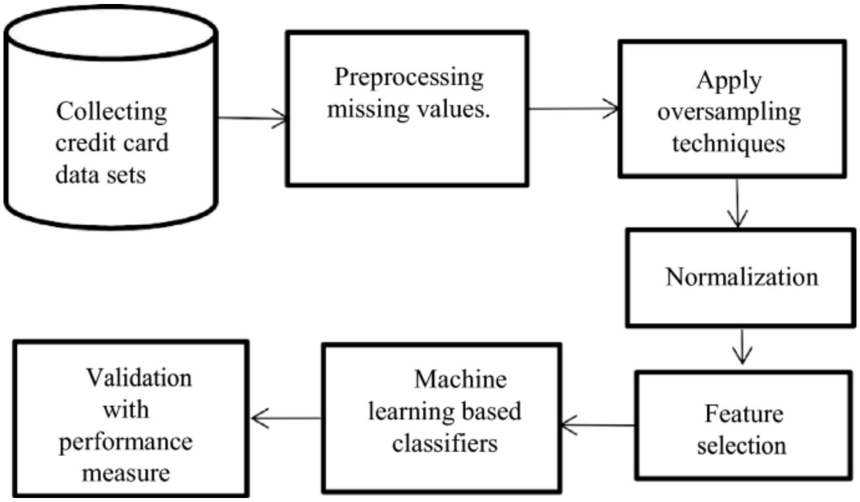

Collecting credit card data sets
Preprocessing missing values.
Apply oversampling techniques
Normalization
Feature selection
Machine learning based classifiers
Validation with performance measure

Fig: Project plan.

The project will be completed in different phases:

Data collection:

The first phase will involve collecting a dataset of historical credit card transactions. The data will be collected from various sources, including banks, credit card companies, and merchants.

Data Cleaning:

- Impute the missing values with the column's mean, median, or mode.
- Drop the rows with missing values.
- Use a machine learning model to predict the missing values like isnull() and heatmap().

Normalize the data:

Normalization is scaling the data so that all features have similar values. This can improve the performance of machine learning models by making the parts more comparable.

Model training:

The second phase will involve training the machine learning model on the collected data. The model will be prepared using a supervised learning algorithm like SVM.

Model evaluation:

The third phase will involve evaluating the machine learning model's performance on a holdout dataset of unseen transactions. The model's performance will be evaluated using accuracy, precision, and recall metrics.

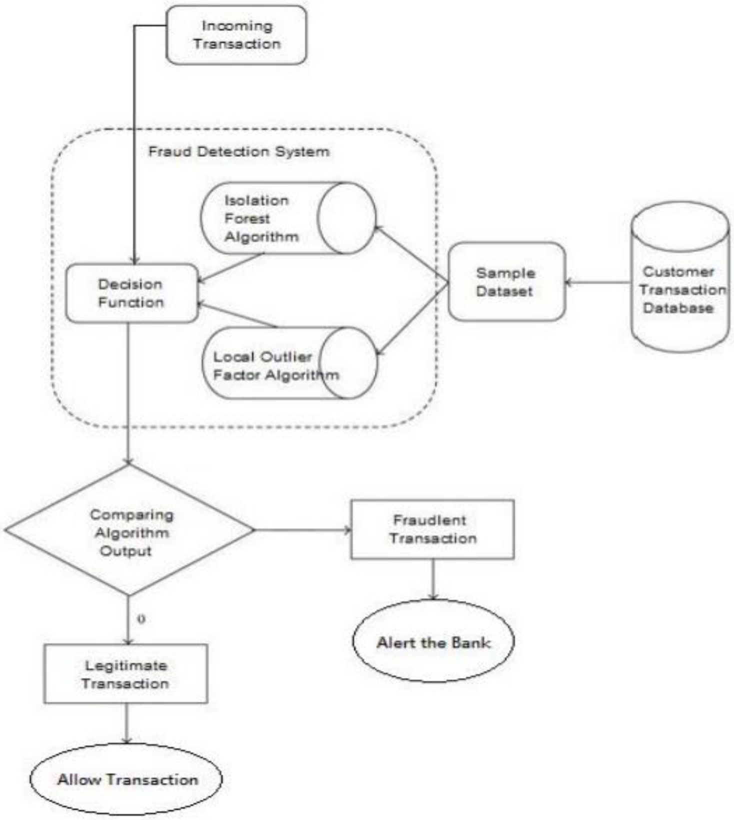
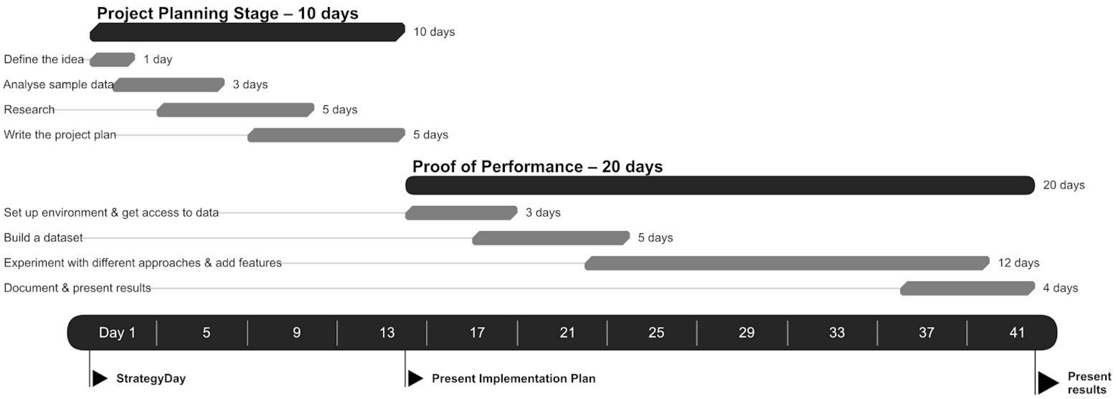


Fig: Working Flow of Credit Card Fraud Detection

Timeline for Our Project:



Results and Evaluations

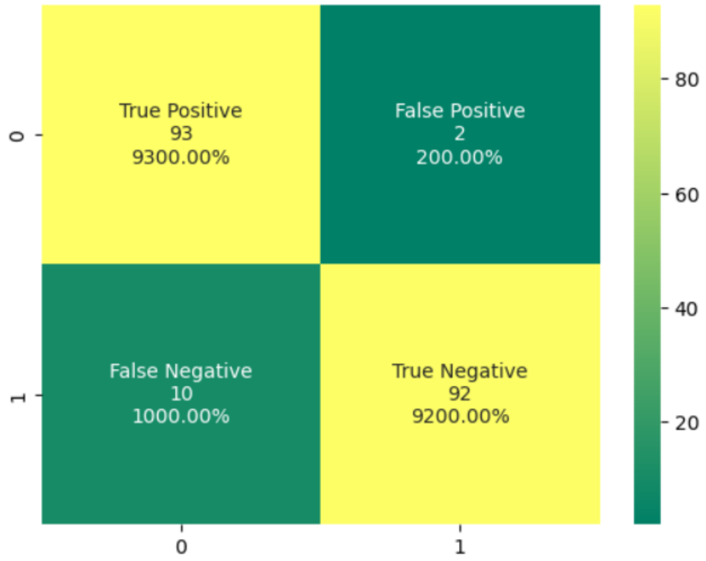
Expected Result:

- A machine learning model that can detect credit card fraud with high accuracy.
- A better understanding of the patterns that are indicative of fraudulent transactions.
- A framework for using machine learning to detect real-time credit card fraud.

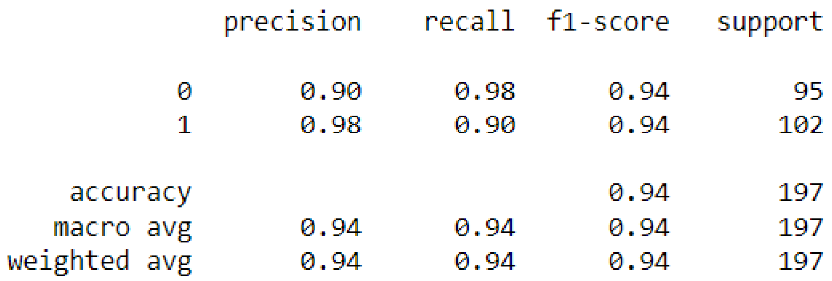
Performance Metrics and Evaluation Methodology:

Confusion Metrics:

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those the machine learning model predicted.



Classification Report:



```
              precision    recall  f1-score   support

           0       0.90      0.98      0.94        95
           1       0.98      0.90      0.94       102

    accuracy                           0.94       197
   macro avg       0.94      0.94      0.94       197
weighted avg       0.94      0.94      0.94       197
```

REFERENCES

[1]. S. V. S. S. Lakshmi, S. D. Kavilla “Machine Learning for Credit Card Fraud Detection System”, unpublished

[2] N. Malini, Dr M. Pushpa, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection ", Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on pp. 255-258. IEEE.

[3] Mrs. C. Navamani, M. Phil, S. Krishnan, “Credit Card Nearest Neighbor Based Outlier Detection Techniques”

[4] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis”, Computing Networking and Informatics (ICCNI), 2017 International Conference on pp. 1-9. IEEE.

[5] R. Choudhary and H. K. Gianey 2017 Int. Conf. Mach. Learn. Data Sci., pp. 3743, 2017.

[6]. G. E. Melo-Acosta, F. Duitama-Muñoz, and J. D. Arias-Londoño, -supervised

Common. Compute. (COLCOM), 2017 IEEE Colomb. Conf., pp. 16, 2017.

[7]. Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented 26, 2016

[8]. Credit Card Fraud Detection dataset: downloaded from Kaggle, September 2013 by European cardholders.