Title: Credit Card Fraud Detection

About Team Members

Member1: Shubhangi Kundu

Email: [sharmamn2000@gmail.com](mailto:sharmamn2000@gmail.com)

Member2: Sagar Sharma

Email: [sharmamn2000@gmail.com](mailto:sharmamn2000@gmail.com)

Member3: Samarth Swaroop Saxena

Email: [saxenasamarth531@gmail.com](mailto:saxenasamarth531@gmail.com)

Member4: Harsh Kumar Pandey

Email: [sharmamn2000@gmail.com](mailto:sharmamn2000@gmail.com)

Under the guidance of: Dr. Yogesh Pal Sir

Abstract:

As we saw in the modern days, the influence of online banking is increased because now-a-days mostly pupil are hesitant to carry cash in fear of robbery. The concept of online banking eases the way of transactions and abolishes the role of the middle person that comes between the mode of transactions. It leads to many pros and cons, the main demerit of using online banking in which people are reluctant to use online banking is the fraud that happened between the transactions. To overcome from this demerit, many precautions were taken to avoid this thing from happening but it all were not effective that much. One of the effective way to get rid of credit card fraud in my opinion is to use machine learning. In machine learning many models like Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine were introduced to solve the problem of frauds that are happened by the credit cards. In our research, our motive is to go through various models and choose the best suitable model whether it is specific or hybrid model and we also focused on the whole process of credit card transaction to avoid fraud at any aspect. We take the dataset of the credit card users and their transaction details to find out the loop hole where the fraudster make their move that leads to frauds. Future studies should focus on studying different types of hybridization and algorithms in the credit card domain.

Keywords: Machine Learning, Fraud Detection, Logistic Regression, Decision Tree, Random Forest.

**Introduction:**

**As the events in the world become more digitalised, cybercrimes like credit card or debit card on increase. In today's time, people are making large transactions using credit cards like in businesses, online shopping, card swiping in malls and stores, trading etc. We are facing many kinds of frauds in these respective fields. It simply means credit card fraud can be defined as " Unauthorized account activity by a person for whom the account was not belonged". After emerging of various numbers of fraud in the market various technologies were introduced to detect the fraud and the vital task is to use the proper technology, which is easy to implement, smooth to use and fast to detect the anomaly in the transactions which leads to fraud. These methods are based on neural network, genetic algorithm, Hidden Markov model, Bayesian network, decision tree, clustering method, Support Vector Machine, etc.**

Outcomes of process

Initialization of transaction by customer

Proper Transaction

Connecting service and making process of transaction

**II**

Fraud

Logistic Regression

Random Forest

Decision Tree

Sample Dataset

Customer Dataset

**Problem Statement:**

**Credit card fraud poses a significant threat to financial institutions and cardholders, leading to substantial financial losses and compromised security. The increasing sophistication of fraudulent activities requires advanced detection systems to safeguard against unauthorized transactions. The aim of this project is to develop a robust credit card fraud detection system that can accurately identify and prevent fraudulent transactions while minimizing false positives.**

**Key Challenges:**

**Class Imbalance: The dataset exhibits a substantial class imbalance, with legitimate transactions vastly outnumbering fraudulent ones. This imbalance must be addressed to prevent the model from being biased towards the majority class.**

**Dynamic Patterns: Fraudulent activities are dynamic and evolve over time. The detection system must be capable of adapting to new patterns and techniques employed by fraudsters.**

**Real-time Processing: Credit card transactions occur in real-time, requiring the fraud detection system to make quick and accurate decisions to prevent unauthorized transactions.**

**Interpretability: In addition to accurate predictions, the system needs to provide insights into the features and patterns contributing to its decisions. This is crucial for trust and interpretability, especially in a financial context.**

.

**Objectives:**

**identifying fraudulent credit card transactions.**

**Address the class imbalance through appropriate sampling techniques or algorithmic adjustments to ensure the model's effectiveness for both classes.**

**Implement real-time Develop a machine learning model or ensemble of models capable of accurately processing capabilities to enable swift decision-making during credit card transactions.**

**Enhance the interpretability of the model, providing clear insights into the factors influencing its predictions.**

**Continuously monitor and update the system to adapt to emerging fraud patterns and maintain optimal performance.**

**Methodology**

This chapter discusses the methodology adopted in this study to classify the non-fraudulent transactions from the fraudulent transactions. Figure 1 shows the steps used in this work. However, before we discussed the different steps of the methodology used in this work, we first discussed the dataset.

ML Model Random Forest

Decision Tree

Logistic Reg.

Data Processing

Performance

Evaluation and model comparision

Best Model

Classification Methodology

DATASET:

The dataset for this research work is obtained from Kaggle. The dataset is a

simulated credit card transaction containing legitimate and fraudulent transactions. It covers the credit card of several customers doing transactions. The transactions presented by this dataset have 284806 transactions in total, and the number of fraudulent transactions was recorded to be 2574 out of the total number of transactions. The dataset is highly imbalanced; the positive class (frauds) account for a tiny percentage of about 0.52 of the complete transactions. The dataset contains 22 features such as" Amount," "V1," "Class," and so on, comprising different data types. It also includes both numerical and categorical features. Each transaction recorded per transaction time is contained in the feature "time" column. The 'Amount' feature column includes the transaction amount carried out, while the last feature in this dataset called "Class" is the response variable that shows whether a transaction is a fraud or not. It takes 1 as a value if it is fraud and 0 if it is not. 'V1-V23' defines the various kinds of data provided by the different persons. The dataset is available at…

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

A screenshot of a computer

Description automatically generated

Data Preprocessing:

Data preprocessing is a vital step before implementing machine learning algorithms, especially when dealing with large datasets. Various models often have distinct requirements for predictors, and the quality of data preparation significantly influences predictive performance. The primary goals of data preprocessing are to enhance data cleanliness, address missing values, and reduce variability. Given that datasets typically comprise both numerical and categorical features, encoding categorical data is imperative before modeling. Outlier detection and removal have been executed to ensure data integrity. To normalize the range of independent variables and mitigate the impact of outliers, feature scaling has been applied. Additionally, to address feature skewness on a larger scale, a box-cox transformation has been employed . Resampling methods, such as undersampling and oversampling, were crucial steps, especially considering the imbalanced nature of the original dataset. This helps prevent bias and overfitting during the training of our machine learning model, particularly important for large datasets.

Graphs:

To show the fraud detection graphically with the use of several plots-

A: Count plot

A graph of a diagram

Description automatically generated with medium confidence

B: Histogram

A graph of a number of money

Description automatically generated with medium confidence

C: Scatter Plot

A graph of blue dots

Description automatically generated

D: Heatmap

A screen shot of a chart

Description automatically generated

Model’s Introduction:

Decision Tree- A decision tree is a supervised machine learning algorithm that partitions the dataset into subsets based on the values of different features. It makes decisions by traversing the tree from the root to the leaves, where each internal node represents a decision based on a feature, and each leaf node corresponds to the predicted outcome.

Decision trees can naturally identify anomalies in the data. Outlying branches or leaves with fewer instances may signal potential anomalies or unusual patterns that could be indicative of fraud. Decision trees can naturally identify anomalies in the data. Outlying branches or leaves with fewer instances may signal potential anomalies or unusual patterns that could be indicative of fraud.

X\_train\_pred = DTC.predict(X\_train)

train\_accuracy = accuracy\_score(X\_train\_pred,y\_train)

train\_accuracy

1.0

X\_test\_pred = DTC.predict(X\_test)

test\_accuracy = accuracy\_score(X\_test\_pred,y\_test)

test\_accuracy

0.9390862944162437

The outcome of the Decision tree model gave us an train accuracy (1.0) and test accuracy(0.9390)

Logistic Regression-A decision tree is a supervised machine learning algorithm that partitions the dataset into subsets based on the values of different features. It makes decisions by traversing the tree from the root to the leaves, where each internal node represents a decision based on a feature, and each leaf node corresponds to the predicted outcome.

While logistic regression is a linear model, it can capture complex relationships between features and the log-odds of the response variable. Nonlinearity in the data can be addressed through feature engineering or by introducing polynomial features.

training\_accuracy

0.9339263024142312

X\_test\_pred = LR.predict(X\_test)

testing\_accuracy = accuracy\_score(X\_test\_pred,y\_test)

testing\_accuracy

0.9746192893401016

The outcome of the Logistic Regression model gave us an train accuracy (0.9339) and test accuracy(0.9746)

Random forest -A Random Forest model is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. In the context of credit card fraud detection, Random Forests offer flexibility, robustness, and the ability to handle complex relationships in the data. Here's a description of how a Random Forest model can be applied to the pre-processed data.

Random Forests are capable of capturing complex, nonlinear relationships in the data due to the diverse set of decision trees. This makes them suitable for scenarios where fraud patterns may not be linear.

X\_train\_pred = RFC.predict(X\_train)

train\_accuracy = accuracy\_score(X\_train\_pred,y\_train)

train\_accuracy

1.0

X\_test\_pred = RFC.predict(X\_test)

test\_accuracy = accuracy\_score(X\_test\_pred,y\_test)

test\_accuracy

0.9644670050761421

The outcome of the Random Forest model gave us an train accuracy (1.0) and test accuracy(0.9644)

Conclusion

As we described earlier about the fraud happening now-a-days from credit card transactions takes a major turn to become a large and short-time taken scam by which people can lose money in a blink of an eye.. To overcome from this problem, we experimented a model in which we used three algorithms that are i.e., Decision tree, Logistic Regression, Random forest in which we used three classifiers to predict the model how far it is accurate. Our experiment shows 96% accuracy with precision remains only at 1%. Our motive is to provide the best search case to show defect the fraud quickly and efficiently. Our future target is to represent this experiment in the form of software aligned with latest technologies like Machine learning, Artificial Intelligence, and Deep Learning.

REFRENCES:

[1] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, “Customer churn prediction in telecommunication industry using data certainty,” *J Bus Res*, vol. 94, pp. 290–301, Jan. 2019, doi: 10.1016/j.jbusres.2018.03.003.

[2] T. A. R. Akbar and C. Apriono, “Machine Learning Predictive Models Analysis on Telecommunications Service Churn Rate,” *Green Intelligent Systems and Applications*, vol. 3, no. 1, pp. 22–34, Jun. 2023, doi: 10.53623/gisa.v3i1.249.

[3] A. Aldalan, A. M. Aldalan, and A. Almaleh, “Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector.” [Online]. Available: https://www.researchgate.net/publication/36911377