CREDIT CARD FRAUD DETECTION

ABSTRACT

The purpose of this project is to detect Credit Card Fraud by recognising fraudulent transactions. we have developed a credit card fraud detection system using Big Data Analytics to solve the problem of financial losses encountered by customers and financial institutions. Detection of fraudulent transactions will be made using big data technologies such as Hadoop, and Spark for real-time data processing and analysis. The dataset will be processed using extraction techniques to identify patterns and anomalies in the data. Algorithms like Graph analysis, and neural networks will be used to determine the fraud behaviour.

INTRODUCTION

With the rise of digital transactions, credit card fraud has become an increasingly prevalent problem, leading to significant financial losses for consumers and financial institutions. Detecting credit card fraud is challenging due to the constantly evolving nature of fraud techniques and the large volume of transactions processed daily. According to Credit card statistics in 2021, the number of people using credit cards around the world was 2.8 billion in 2019, in addition, 70% of those users own a single card at least. Therefore, there is a need for effective fraud detection methods that can detect fraudulent transactions in real-time to minimize financial losses. To combat this problem, we have developed a credit card fraud detection system using Big Data Analytics.

RELATED WORK

Several studies have been performed on credit card fraud detection. Some studies have used rule-based systems, while others have used machine learning algorithms. Machine learning-based approaches have shown better results than rule-based systems. Some studies have used logistic regression, while others have used decision trees, random forests, and neural networks. However, there is still room for improvement in terms of accuracy and efficiency. We emphasise the strengths and limitations of these techniques and identify the research gaps that our proposed method aims to address.

METHODOLOGY

In this section, we present our proposed methodology for credit card fraud detection. The proposed technique uses a dataset of credit card transactions, which includes both fair and fraudulent transactions. The dataset is preprocessed to remove any missing values and outliers and then will be processed using extraction techniques to identify patterns and anomalies in the data. The dataset is then split into training and testing sets. From the analyzed data we developed a histogram for Time and Amount features considering the column that represents the frauds. Since the data is very imbalanced with a majority of transactions being non-fraudulent.Using the imbalanced dataset as the basis for predictive models and analysis can lead to errors and overfitting of algorithms. So, we have developed a model that can accurately detect patterns, despite the imbalanced nature of the dataset. To identify the features that heavily influence whether a transaction is fraudulent or not, it is crucial to use the correct Data Frame for analysis. This helps us to determine which features have a high positive or negative correlation with fraudulent transactions.

EXPERIMENTAL DISCUSSION

In this section, we present the experimental results of our proposed method. We have used the "Credit Card Fraud Detection" dataset available on Kaggle. The dataset consists of credit card transactions that have been labelled as fraudulent or non-fraudulent. The dataset contains a total of 284,807 transactions, of which only 492(0.17%) are labelled as fraudulent. We provided large volume card transaction data and leverage it in big data technologies such as Hadoop, and Spark for real-time data processing and analysis. The dataset will be processed using extraction techniques to identify patterns and anomalies in the data.

To prevent overfitting issues due to unbalanced data, we use decision trees and set them to a maximum depth. To determine the optimal depth, we generate models with varying depths and evaluate their performance. By selecting the model with the best results, we can identify the optimal depth and use it to create and assess the final decision tree model. The best accuracy is 0.9902 with depth =3.

Then we trained the model using underbalanced data and it resulted in a graph as shown in fig 1.1.

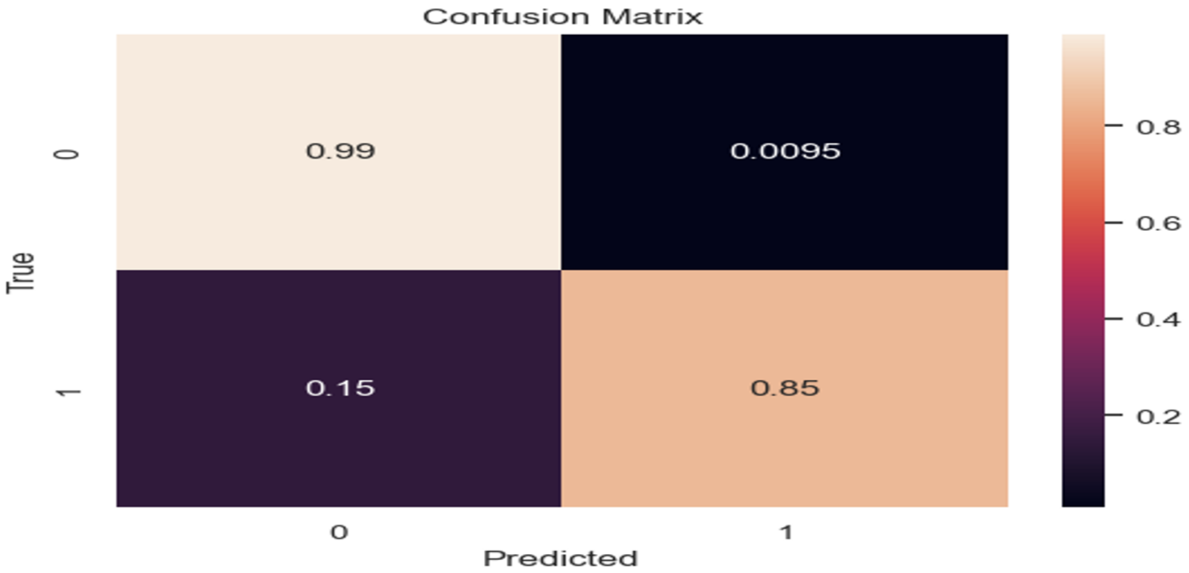

Fig 1.1

This level of accuracy demonstrates the effectiveness of the decision tree model in detecting fraudulent transactions in credit card data. The KNN model was implemented with a value of k=10 and achieved an accuracy of 0.9891, which is slightly lower than the accuracy achieved by the decision tree model with a depth of 3. The model can be used as a reliable tool for detecting fraudulent transactions, which is crucial for preventing financial losses for both consumers and financial institutions.

CONTRIBUTION

Mounica Chirravuru: Final Project Report
Niralee Kothari: Data Cleaning and presentation
Prachi Bhosale: Debugging and provided the required support during the coding phase
Pragati Gunai: Developed Algorithms
Sachin Lade: Coding and graphical representation

CONCLUSION:

The project uses the "Credit Card Fraud Detection" dataset available on Kaggle, which consists of credit card transactions labelled as fraudulent or non-fraudulent. We have implemented decision trees and k-NN algorithms to train machine learning models to detect fraudulent transactions. We plan to leverage big data technologies such as Hadoop and Spark for real-time data processing and analysis.

Overall, this project aims to develop a robust and scalable credit card fraud detection system using big data analytics that can process and analyse a large volume of transactions in real-time, detecting and preventing fraudulent activity.

REFERENCES

1. Kaggle dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
2. Scikit-learn.(2021).KNeighborsClassifier.Revisedfrom https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
3. Some concepts revised from: Big Data Made Accessible. 2020. Anil Maheshwari.