ML-Based Online Transaction Fraud Detection

[1.]'Sindhu S L' [2.]'Sushmitha K M'
1. Assistant Professor ,Department of MCA,BIET,DVG
2. Student, Department of MCA,BIET,DVG

ABSTRACT

This paper focuses on fraud detection and the measures necessary to fully automate this process, which has become crucial for banks. With fraud on the rise, it poses significant threats and can result in substantial damages to financial institutions. Transaction data presents unique challenges for fraud detection due to the lack of short-term processing capabilities. The primary objective is to conduct a feasibility study on the selected fraud detection methods. Using various models, we aim to test each transaction individually and proceed accordingly.Initially, we define the detection task, outlining the dataset attributes, metric choices, and techniques to manage unbalanced datasets. This analysis helps identify underlying patterns within the dataset, such as how cardholders' purchasing habits may evolve over time and how fraudsters may adapt their tactics. We then explore various methods used to derive sequential features from credit card transactions. Financial fraud, the practice of gaining financial benefits through deceitful and unlawful means, has become a significant threat to businesses and organizations. Despite numerous efforts to combat financial fraud, it continues to inflict substantial economic and societal harm, with daily losses reaching significant amounts. Traditional methods for fraud detection, introduced years ago, were predominantly manual, making them time-consuming, costly, and prone to errors. Although more research is being conducted, current approaches have not effectively reduced the financial losses associated with fraud. Conventional fraud detection methods, such as manual verifications and inspections, are inefficient, costly, and inaccurate.

Keywords: Transaction data, cardholders' purchasing.

1.INTRODUCTION

Machine Learning is a system of computer algorithms that can learn and improve from experience without needing explicit programming from a human. It is a subset of artificial intelligence that integrates data with statistical tools to predict outcomes, which can then be used to derive actionable insights The key innovation in machine learning is that a machine can independently learn from data (i.e., examples) to generate accurate results. Machine learning is closely associated with data mining and Bayesian predictive modeling, where the machine receives data as input and employs an algorithm to generate answers. A common machine learning task is providing recommendations. For instance, Netflix uses machine learning to recommend movies and TV series based on a user's viewing history. Tech companies utilize unsupervised learning to enhance user experience by personalizing recommendations. Additionally, machine learning is applied to various tasks such as fraud detection, predictive maintenance, portfolio optimization, and task automation.

2. LITERATURE REVIEW

Mobile payment is becoming a major payment method in many countries. However, the rate of payment fraud with mobile is higher than with credit card. One potential reason is that mobile data is easier to be modified than credit card data by fraudsters, which degrades our data-driven fraud detection system. Supervised learning methods are pervasively used in fraud detection. However, these supervised learning methods used in fraud detection have traditionally been developed following the assumption that the environment is benign; there are no adversaries trying to evade fraud detection system[1]. In this paper, we took potential reactions of fraudsters into consideration to build a robust mobile fraud detection system using adversarial examples. Experimental results showed that the performance of our proposed method was improved in both benign and adversarial environments. This paper presents a thorough analysis of 30-minute data sets of KSA residential digital meters to identify all possible discrepancies in the data sets and devise statistical techniques best suited to remove these discrepancies as per the nature of each discrepancy. The analysis is performed through a program that was developed in Python-Pandas. The program parses through three month's meter measurements of 3,283 consumers throughout KSA and detects data inconsistencies, duplicates, missing and outlier values and other issues in the data sets. Statistical techniques that are part of the program are then implemented to correct for these issues[2]. A validation process was developed and included in the program to ensure the adjustment process produces the best reliable outcomes. Analysis indicates that smart masters data have issues that need preprocessing to be used for other applications. The outcome of the program developed shows that smart meters measurement outcome data set could be considered as a valid and trusted, which can be used for smart grid applications such as behavioral analysis of the electricity consumers. Credit card fraud is a serious and growing problem with the increase in e-commerce and online transactions in this modern era. With this identity theft and loss of money, such mischievous practices can affect millions of people around the world. Criminal activity is a rising threat to the financial sector with-reaching implications. Information extraction seemed to have assumed a basic job in recognition of online payment fraud, fraud detection efficiency in credit card purchases is significantly affected by the data set measuring strategy, the choice of variable and the detection techniques used. This publication inspects execution of, Support Vector Machine, Naive Bayes, Logistic Regression and K-Nearest Neighbor on exceptionally distorted data on credit card fraud. The execution of these techniques is assessed dependent on accuracy, sensitivity, precision, specificity. The outcomes show an ideal accuracy for logistic regression, Naive Bayes, k-nearest neighbor and Support vector machine classifiers are 99.07%, 95.98%, 96.91%, and 97.53% respectively. The relative outcomes demonstrate that logistic regression performs superior to other algorithms[3].In today's economic scenario, credit card use has become extremely commonplace. These cards allow the user to make payments of large sums of money without the need to carry large sums of cash. They have revolutionized the way of making cashless payments and made making any sort of payments convenient for the buyer. This electronic form of payment is extremely useful but comes with its own set of risks. With the increasing number of users, credit card frauds are also increasing at a similar pace. The credit card information of a particular individual can be collected illegally and can be used for fraudulent transactions. Some Machine Learning Algorithms can be applied to collect data to tackle this problem. This paper presents a comparison of some established supervised learning algorithms to differentiate between genuine and fraudulent transactions[4].Credit cards are very commonly used in making online payments. In recent years' frauds are reported which are accomplished using credit cards. It is very difficult to detect and prevent the fraud which is accomplished using credit card[5] Machine Learning(ML) is an Artificial Intelligence (AI) technique which is used to solve many problems in science and engineering. In this paper, machine learning algorithms are applied on a data set of credit cards frauds and the power of three machine learning algorithms is compared to detect the frauds accomplished using credit cards. The accuracy of Random Forest machine learning algorithm is best as compared to Decision Tree and XGBOOST algorithms[6].

3. METHODOLOGY

1. **Data Set Used:**

○ **Description:** Choose a dataset that includes transactional data with labeled fraud/non-fraud transactions.

○ **Source:** Common sources include Kaggle datasets, financial institutions' data (with permissions), or synthetic datasets.

- **Features:** Typical features include transaction amount, time, type, location, etc.

2. **Data Processing:**

- **Data Cleaning:** Handle missing values, outliers, and inconsistencies.
- **Feature Engineering:** Create new features if necessary (e.g., transaction frequency, average transaction amount).
- **Normalization/Standardization:** Scale numerical features to ensure uniformity.
- **Handling Imbalanced Data:** Address class imbalance (more non-fraudulent transactions than fraudulent ones) using techniques like oversampling (SMOTE) or under sampling.

3. **Algorithm used:**

- **Selection:** Choose appropriate algorithms like:
  - **Logistic Regression:** Simple and interpretable, suitable for baseline.
  - **Decision Trees/Random Forest:** Handles non-linear relationships well.
  - **Support Vector Machines (SVM):** Effective in high-dimensional spaces.
  - **Gradient Boosting Methods (XGBoost, LightGBM):** Ensemble methods for improved accuracy.
- **Model Training:** Split data into training and validation sets (cross-validation) to train and tune hyperparameters.

4. **Evaluation and Results:**

- **Metrics:** Use metrics like precision, recall, F1-score, and ROC-AUC to evaluate model performance.
- **Confusion Matrix:** Illustrate true positives, false positives, true negatives, and false negatives.
- **Feature Importance:** Identify key features contributing to fraud detection.
- **Visualization:** Display results using plots (ROC curve, precision-recall curve).

5. **Deployment and Monitoring:**

- **Deployment:** Implement the model in a production environment using APIs or batch processing.
- **Monitoring:** Continuously monitor model performance for drift and retrain periodically.

4. ALGORITHM USED

Random Forest Algorithm:

A Brief Overview

Random Forest is a powerful and versatile ensemble learning algorithm used for classification, regression, and other tasks. It works by building multiple decision trees during training and outputting the mode of the classes (for classification) or the average prediction (for regression) of the individual trees. Here's a brief explanation of how **Random Forest works:**

1.***Ensemble Method & Ensemble Learning:*** Random Forest is an ensemble method that combines multiple decision trees to create a stronger overall model. The idea is that multiple trees, working together, will correct each other's mistakes, leading to better generalization and accuracy.

2. Decision Trees Building Blocks: Each decision tree in the forest is a basic model that partitions the data based on feature values to make decisions. Decision trees are prone to overfitting, but when used in a random forest, this tendency is mitigated.

3.Bootstrapping and Aggregation (Bagging)Bootstrapping: Random Forest creates each decision tree using a different bootstrap sample of the data. A bootstrap sample is a random sample taken with replacement from the training dataset, meaning some data points may be repeated while others are left out.

Aggregation: After training all the trees, their predictions are aggregated to form the final output. For classification tasks, it uses majority voting, and for regression tasks, it uses averaging.

4. Random Subset of Features

Random Feature Selection: To build each tree, Random Forest randomly selects a subset of features at each split in the tree. This random selection helps in reducing correlation between trees, leading to more diverse models that improve the ensemble's performance.

5. Construction of Trees

Tree Building: Each tree is constructed by:

Selecting a random sample of the data with replacement (bootstrapping).

At each node, a random subset of features is chosen.

The best split for the selected feature subset is determined.

The process is repeated until a stopping criterion is met (e.g., maximum depth, minimum number of samples per leaf).

6. Majority Voting/Averaging

Classification: For classification tasks, each tree votes for a class label, and the class with the most votes is the final prediction.

Regression: For regression tasks, each tree predicts a numerical value, and the final output is the average of these values.

7. Benefits of Random Forest

Reduced Overfitting: By averaging multiple trees, Random Forest reduces the risk of overfitting compared to individual decision trees.

Robustness to Outliers and Noise: The ensemble method makes Random Forest more robust to outliers and noise in the data.

Feature Importance: Random Forest provides measures of feature importance, which can be useful for feature selection and understanding the data.

8. Hyperparameters

Number of Trees (estimators): More trees usually improve performance, but with diminishing returns after a certain point.

Number of Features (max_features): Controls the number of features to consider at each split. Common choices include the square root of the total number of features or a logarithmic function.

Tree Depth (max_depth): Limits the depth of the individual trees, balancing the trade-off between bias and variance.

Summary

Random Forest is a robust ensemble learning algorithm that combines multiple decision trees to make predictions. By averaging or voting across a large number of uncorrelated trees, it improves prediction accuracy and reduces overfitting. Its ability to handle large datasets, high dimensionality, and provide insights into feature importance makes it a popular choice for both classification and regression tasks.

This brief overview captures the essence of how Random Forest works and why it is a valuable tool in machine learning.

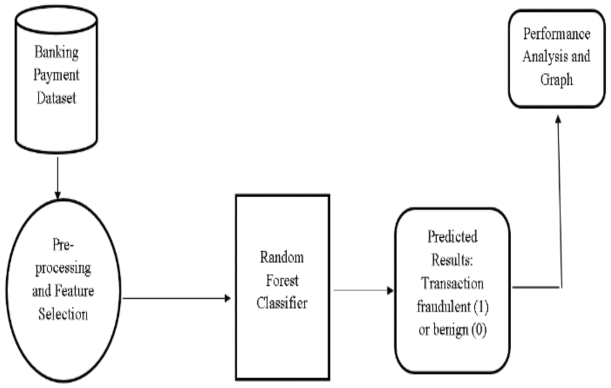
How does the algorithm work?

Select random samples from a given dataset. Construct a decision tree for each sample and get a prediction result from each decision tree. Perform a vote for each predicted result. Select the prediction result with the most votes as the final prediction Finding important features Random forests also offer a good feature selection indicator. Scikit-learn provide an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction.

It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses Gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

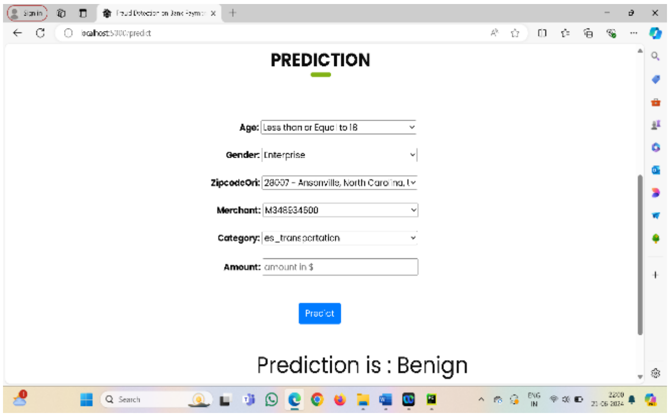
5. SYSTEM ARCHITECTURE:


Banking Payment Dataset
Performance Analysis and Graph
Pre-processing and Feature Selection
Random Forest Classifier
Predicted Results: Transaction fraudulent (1) or benign (0)

6. RESULTS

Financial fraud is a pervasive issue affecting various sectors including corporate, banking, insurance, and taxation.

Despite numerous efforts to combat it, financial fraud persists, resulting in substantial economic losses and societal harm. The advent of artificial intelligence (AI) and machine learning (ML) technologies presents new opportunities for detecting fraudulent activities by analyzing large volumes of financial data. This paper presents a detailed study on the application of ML-based technologies for fraud detection. We focus specifically on the Random Forest Classifier methodology, evaluating its effectiveness in identifying fraudulent transactions. By systematically extracting, synthesizing, and reporting results, our study provides a thorough analysis of existing literature and methodologies. Our findings highlight the robustness and accuracy of the Random Forest Classifier in detecting financial fraud, making it a valuable tool for businesses and industries aiming to mitigate the risks associated with fraudulent activities. The study also discusses the broader implications of ML-based fraud detection systems, emphasizing their potential to enhance the security and integrity of financial operations.



localhost:5000/predict

PREDICTION

Age: Less than or Equal to 18

Gender: Enterprise

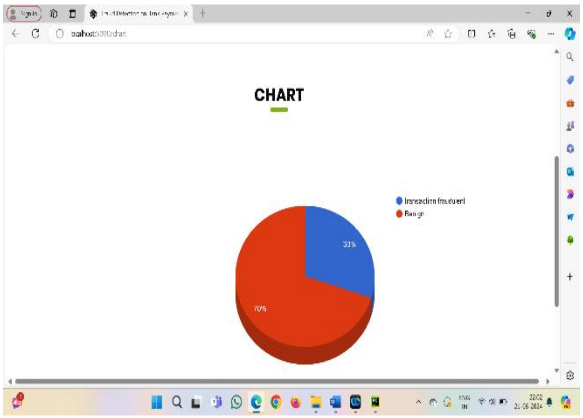
ZipcodeOri: 28007 - Ansonville, North Carolina, U

Merchant: M348934600

Category: es_transportation

Amount: amount in $

Prediction is : Benign



localhost:5000/chart

7.CONCLUSION

Financial fraud can occur across various financial sectors, including corporate, banking, insurance, and taxation. Recently, it has become a growing concern for businesses and industries. Despite numerous efforts to combat it, financial fraud remains prevalent, causing significant daily financial losses that negatively impact both society and the economy. However, with advancements in artificial intelligence, machine learning technologies can now be effectively utilized to detect fraudulent transactions by analyzing large volumes of financial data. In this article, we present a comprehensive study that thoroughly analyzes and summarizes existing research on machine learning-based fraud detection. Our study employs the Random Forest Classifier methodology, which systematically extracts, synthesizes, and reports findings using well-defined procedure.

8. REFERENCE

1. S. Delecourt and L. Guo, “Developing a resilient mobile payment fraud detection system with adversarial examples,” in Proceedings of the 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), IEEE, 2019, pp. 103–106.

2. T. Alquthami, A. M. Alsubaie, and M. Anwer, “Significance of processing smart meter data – a case study of Saudi Arabia,” in Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), IEEE, 2019, pp. 1–5.

3. O. Adepoju, J. Wosowei, S. Lawte, and H. Jaiman, “Comparative assessment of credit card fraud detection through machine learning techniques,” in Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT), IEEE, 2019, pp. 1–6.

4. S. Khatri, A. Arora, and A. P. Agrawal, “Comparative study of supervised machine learning algorithms for credit card fraud detection,” in Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 680–683.

5. V. Jain, M. Agrawal, and A. Kumar, “Performance evaluation of machine learning algorithms for credit card fraud detection,” in Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, 2020, pp. 86–88.

6.A.Thennakoon,C.Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, “Real-time detection of credit card fraud using machine learning,” in Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2019, pp. 488–493.