ROYAL UNIVERSITY OF PHNOM PENH
FACULTY OF ENGINEERING

Department: IT Engineering

Major: Data Science and Engineering

Course: Financial Technology

Lecturer: Mr. Ky Soklay

Class: DSE Cohort 1

Prepared by: Leng Koemhort

Date: 09.01.2024

Topic: Credit Card Fraud Detection

I. Executive Summary

Credit card fraud causes significant financial losses and reduces customer trust. Detecting fraud is particularly challenging because fraudulent transactions are extremely rare, making up only 0.172% of all transactions. Our solution is a fraud detection system that uses **XGBoost Classifier.**. We want to see and compare XGBoost Classifiers with Logistic Regression and Decision Tree Classifiers. In this study we will see those model performance for identifying fraudulent transactions in highly imbalanced datasets. By focusing on **precision** and **recall**, the system ensures accurate detection with minimal disruption to genuine customers. With credit card fraud costing businesses, our system addresses a critical need in the financial industry. It offers a reliable and cost-effective way for banks, credit card companies, and fintech providers to combat fraud.

II. Business Challenge

The financial industry faces an ongoing battle against credit card fraud, which results in substantial financial losses, reputational damage, and decreased customer trust**.** For instance, A false positive can harm your business's revenue and reputation. If you have a system in place to block a potentially fraudulent transaction, a false positive means the customer would not be able to complete their purchase. And they may end up going to a competitor and lose faith in your business. There is no one-size-fits-all approach to fraud detection. With fraud attempts evolving every day, we must be agile and adapt our fraud

detection protocols.A stringent fraud detection protocol can create friction in your customers' journey**,** slowing access to your services and driving customers towards competition.

III. Solution Description

Our proposed solution is to apply a supervised machine learning model on an imbalance dataset which is Logistic Regression model to see its performance and compare it to Decision Tree Classifiers.

1. Datasets

For our dataset we will use a credit card fraud dataset from Kaggle**.** It contains the transactions that occurred in two days made by credit cards in September 2013 by European cardholders. Since it is a credit fraud, it normally is an imbalance dataset that contains 492 frauds out of 284,807 transactions. It is a PCA transformation dataset due to confidentiality issues, the provider cannot provide the original features and more background information about data. Features V1-V28 are the features after PCA transformation, Only 'Time', 'Amount' and Class are keeped as original to use in fraud detection.

2. Methodology

In this study we will be employed using the XGBoost Classification. The XGBoost aims to minimize an objective function that combines Loss Function and Regularization.

2.1. Objective Function

Loss Function is used to measure the difference between predicted probability and the actual class label. The following is the loss function for Binary Classification.

$$L(y_i, p_i) = -\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

Regularization is used to penalizes model complexity (e.g., the number of leaves in the tree and the weights of each leaf) to prevent overfitting.

$$\Omega = \gamma T + \frac{1}{2}\lambda O^2$$

So the ***Objective Function*** to minimize loss is

$$L = [\sum_{i=1}^{n} L(y_i, p_i + O_{value})] + \Omega$$

$$L = [\sum_{i=1}^{n} L(y_i, p_i + O_{value})] + \frac{1}{2}\lambda O^2_{value}$$

We omit the γT because pruning takes place after the full trees are built, and it plays no role in derive the ***Optimal Output Values*** (O_{value})**.** XGBoost uses a **Second Taylor approximation** to approximate the loss function. The first derivative of loss function is ***Gradient*** (g)***)*** and the second derivative of loss function called ***Hessian*** (h).

$$g_i = \frac{d}{dp_i} L(y_i, p_i) = \widehat{y_i} - y_i$$

$$h_i = \frac{d^2}{dp_i^2} L(y_i, p_i) = \widehat{y_i}(1 - \widehat{y_i})$$

$\widehat{y_i}$ is the predicted probability and y_i is the actual label. So the loss function to find the output values is

$$L = [\sum_{i=1}^{n} L(y_i, p_i) + g_i O_{value}] + \frac{1}{2} h_i O_{value}^2$$

2.2. Building Decision Trees

Each decision tree is constructed to minimize the loss by focusing on the residuals. The data is split into branches based on feature values that minimize the objective function. Each leaf in the tree is assigned a weight (w_j). It is the weight of the leaf to which the data point i belongs.

$$w_j = -\frac{\Sigma g_i}{\Sigma h_i + \lambda}$$

When splitting a node into two nodes, the reduction in loss is calculated as Gain. And it chooses the split with the highest gain.

$$\text{Gain} = \frac{1}{2}\left(\frac{(\sum G_{\text{left}})^2}{\sum H_{\text{left}} + \lambda} + \frac{(\sum G_{\text{right}})^2}{\sum H_{\text{right}} + \lambda} - \frac{(\sum G_{\text{total}})^2}{\sum H_{\text{total}} + \lambda}\right) - \gamma$$

Once a tree is built, the model updates predictions for each data point by adding the leaf weights. η is a learning rate.

$$\hat{y}_i^{(new)} = \hat{y}_i^{(old)} + \eta \cdot w_j$$

The process is repeated iteratively to build decision trees to minimize residuals, stopping when loss improvement stalls or a maximum number of trees is reached.

IV. Solution Justification

Fraud detection is important because of the costs and consequences businesses face without it. In addition to financial losses, fraudulent activities can cause reputational damage, business interruptions and lost productivity. Firms that don't provide fraud protection also risk negative customer experiences that can affect loyalty and lead to turnover.

Beyond the business benefits, fraud detection may also be required by law. Insurance providers, financial institutions and others can face regulatory mandates to detect and prevent fraud. Noncompliance could bring penalties and fines. For example, US federal regulators fined the Bank of America USD 225 million for a faulty fraud detection system during the COVID-19 pandemic.

V. Benefits Justification:

Based on a Literature review called "Financial fraud detection through the application of machine learning techniques" (published on 3 September 2024) analyzes 104 articles from 2012 to 2023. Key models mentioned include Random Forest (RF) with 34 mentions, Logistic Regression (LR) with 32, Support Vector Machine (SVM) with 29, Decision Trees (DT) with 29, Naive Bayes (NB) with 19, Artificial Neural Networks (ANN) with 17, K-nearest Neighbors (KNN) with 14, XGBoost with 13, Autoencoders with 10, and Long-Short Term Memory (LSTM) with 7 mentions.

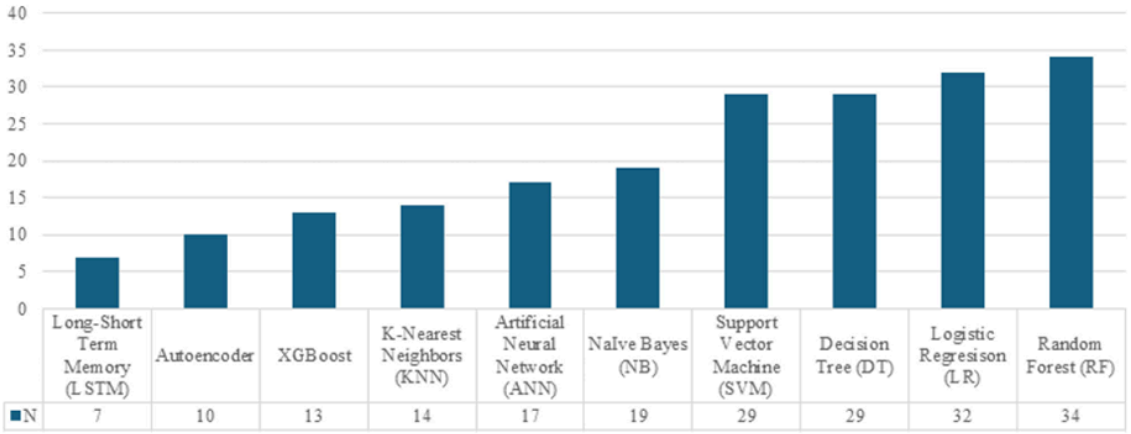

Fig. 7 Main machine learning models used for financial fraud detection. Illustrates the most common machine learning models in financial fraud detection. Authors' own elaboration.

XGBoost, combining multiple weak decision tree models, offers an efficient solution for classification and regression tasks. SVM, Decision Trees (DT), Logistic Regression (LR), and Random Forest (RF) are widely used for various applications, including external fraud (e.g., credit card fraud) and internal fraud (e.g., financial statement fraud). The most common machine learning techniques are supervised learning (56.73%), unsupervised learning (18.29%), a mix of supervised and unsupervised learning (15.38%), and a combination of supervised and deep learning (2.88%).

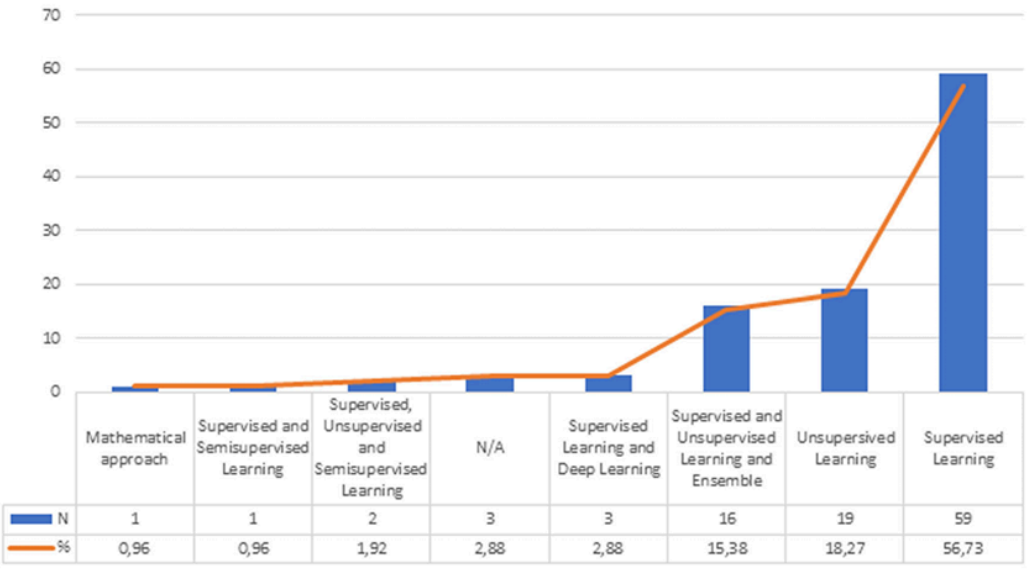

Fig. 8 Approaches used in the experiments. Shows the different experimental approaches used in the study. Authors' own elaboration.

This figure indicates the number of times each type of technique is applied. Some articles applied several ML methods, in which the algorithms are mainly classified according to the learning methods.

After implementation of applying XGBoost within PCA dataset and compare it with Logistic Regression and Decision Tree, We get the result as the table below

Here's the table ranked by accuracy, from highest to lowest:

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	0.9995	0.9459	0.7692	0.8484
Logistic Regression	0.9992	0.8906	0.6263	0.7354
Decision Tree	0.9989	0.6600	0.7252	0.6910

XGBoost performs best in accuracy, precision, recall, and F1 score. Its high accuracy (0.9995), precision (0.9460), and recall (0.7692) yield the best F1 score (0.8485). **Logistic Regression** has high accuracy (0.9993) but lower recall (0.6264), potentially missing more fraudulent cases. It offers good precision (0.8906), reducing false positive errors. **Decision Tree Classifier has** strong performance but low precision (0.6600) and F1 score (0.6911), making it less reliable in distinguishing fraud cases and prone to false positives. In conclusion XGBoost strikes a balance between accuracy and precision-recall, making it the most suitable choice for fraud detection.

VI. Case Study

As a case study, SPD Technology developed an AI-powered fraud prevention solution for an e-commerce and financial services company, leveraging advanced machine learning models, including **XGBoost**, as part of the credit fraud detection system. The platform, which

supports mobile money and card payments, faced a surge in fraudulent transactions, prompting the need for a robust solution to protect customers and reduce fraud.

The project presented several challenges. The dataset was highly imbalanced, with only 6% of 140,000 transactions classified as fraudulent. Detecting fraud for users with limited transaction histories required innovative approaches such as **few-shot learning**. Additionally, the team generated and evaluated over 700 features to identify the most relevant ones for effective fraud detection.

To overcome these challenges, SPD Technology utilized XGBoost, alongside other classification models like CatBoost and LightGBM, which outperformed traditional anomaly detection methods. Techniques such as **oversampling** and **synthetic data generation** were employed to handle the imbalanced dataset effectively. A three-tier fraud detection system was implemented: transactions with a fraud probability below 10% were approved, those between 10% and 80% required additional authentication, and those above 80% were frozen for manual review. The model incorporated transaction details, client behavior, and contextual features to make precise predictions.

The XGBoost-powered solution delivered significant results. It reduced fraud-related operational costs, improved customer satisfaction with streamlined authentication for low-risk transactions, and enhanced platform security. Over 140,000 transactions were analyzed annually, resulting in fewer reported fraud cases and increased trust among users.

Overall, the AI-driven fraud prevention system, with **XGBoost at its core**, transformed the platform into a safer and more efficient service. Continuous updates to the model ensure adaptability to emerging fraud patterns, maintaining the system's effectiveness and the platform’s reliability for its users.

VII. Summary

In this study, we implemented and compared three machine learning models—XGBoost, Logistic Regression, and Decision Tree Classifiers—to detect credit card fraud within a highly imbalanced dataset. The dataset, sourced from Kaggle, contains 492 fraudulent transactions out of 284,807, with features transformed via PCA. Our results show that XGBoost outperforms both Logistic Regression and Decision Tree classifiers in terms of accuracy, precision, recall, and F1 score, making it the most reliable model for fraud detection. The XGBoost model strikes an optimal balance between precision and recall, achieving an accuracy of 99.95%, a precision of 94.59%, and a recall of 76.92%.

We also reviewed the broader impact of fraud detection systems, highlighting how their implementation is essential for businesses to prevent financial loss, protect customer trust, and comply with regulatory requirements. The case study further demonstrated how AI-powered fraud prevention, particularly using XGBoost, can successfully address challenges in detecting fraudulent transactions, even in highly imbalanced datasets. This approach has proven effective in reducing operational costs, improving user experience, and maintaining the platform’s security.

VIII. References

- https://doi.org/10.1057/s41599-024-03606-0
- https://arxiv.org/abs/1603.02754
- https://youtu.be/OtD8wVaFm6E?si=nVhon8ivCgDOyYjJ