Credit Card Fraud Detection with Machine Learning Techniques

Farhat Lamia Barsha
Department of Computer Science, Tennessee Technological University
fbarsha42@tntech.edu

***Abstract*— In an era of increasing digitalization, cyber fraud poses a significant threat to both individuals and organizations. Cyber fraud is a rising crime with the intent to corrupt individuals' personal information. As malicious actors continue to evolve their techniques, the need for effective fraud detection systems becomes more critical than ever. Here, I am going to briefly discuss the paper "Cyber Fraud Predication with Supervised Mchine Learning Techniques". This research focuses on the application of supervised machine learning techniques to enhance cyber fraud prediction capabilities. After replicating this research paper, I will provide a comparison between their result and my results along with the strengths and weaknesses of my approach.**

I. Paper Overview

Zhoulin Li and his co-authors published the paper "Cyber Fraud Prediction with Supervised Machine Learning Techniques" in ACMSE 2020 [1]. This study focuses on the identification of credit card fraud transactions using publicly accessible datasets. The researchers employ three different machine learning algorithms, including Naive Bayes, logistic regression, and artificial neural networks, to address cyber-security classification challenges. The experimental findings indicate that logistic regression outperformed Naive Bayes and neural network techniques in terms of accuracy when applied to datasets with a more balanced distribution for categorization. In the dataset characterized by a highly unbalanced distribution, the Naive Bayes algorithm exhibited superior performance compared to logistic regression and neural networks. The balance accuracy metric is employed to assess the performance of three supervised algorithms.

They used a publicly available dataset from Kaggle that contains credit cards transactions in September 2013 by some European cardholders [2]. The dataset includes 284,807 transactions that took place over two days, among which 492 are frauds which is 0.172% of all transactions. The original features and details of background information about the data are transformed into PCA due to confidentiality issues. Only "time" and "Amount" are not converted by PCA. The function "time" contains the seconds between each transaction and the dataset's first transaction. The "amount" in the dataset is the transaction amount, enabling cost-sensitive learning. Feature 'Class' is the response variable, reporting 1 for fraud and 0 otherwise.

The dataset is partitioned into a training set (75% of the total sample) and a test set (25% of the total sample). Additionally, they conducted tests with a split of 20% for testing and 80% for training, as well as a split of 30% for testing and 70% for training. They used a balanced accuracy to evaluate an algorithm's efficacy and quality because when dealing with an imbalanced dataset, accuracy may not be enough to accurately represent model performance. From the balanced accuracy comparison, they found that Naive Bayes outperforms Logistic Regression and Neural Network. They also got that Naive Bayes is performing better from ROC curve.

TABLE I: Balanced Accuracy and Comparison with Naive Bayes, Logistic Regression, and Neural Network

Dataset Split	Naive Bayes	Logistic Regression	Neural Network
(75%/25%)	0.90600293	0.78742966	0.88324189
(70%/30%)	0.90772206	0.75843305	0.82304718
(80%/20%)	0.91497069	0.77715737	0.84149622

As future work, they proposed that the hyperparameters of the proposed Neural Network, like- the activation function, number of hidden layers, and number of nodes in the hidden layers can be tuned to achieve better performance for the dataset. To reduce the overfitting issue in the training of Logistic Regression, Lasso, and Ridge regularization can be applied to the dataset to improve the performance.

II. Dataset

The dataset utilized in this study was obtained from the Kaggle website and comprises around 150,000 e-commerce transactions. The features encompass sign-up time, purchase time, purchase value, device ID, user ID, browser, and IP address. A novel functionality was implemented to assess the temporal disparity between the registration and acquisition events, given that the duration of an account's existence frequently serves as a significant factor in the identification of fraudulent activities. The dataset has a total of 151,112 transactions, of which 14,151, or 9.36%, are identified as fraudulent transactions. All of the data provided is synthetic in nature.

III. Methodology

In this work, I have applied two machine learning algorithms Logistic Regression and Gaussian Naive Bayes, and three deep learning algorithms Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN).

A. Logistic Regression:

Logistic regression is a well-established machine learning technique commonly employed for binary classification tasks. It aims to estimate the likelihood that a given instance belongs to a specific category.

B. Gaussian Naive Bayes:

The Gaussian Naive Bayes algorithm is a probabilistic machine learning technique that is well-suited for classification tasks. It operates under the assumption that the features are independent of each other, given the class.

C. Artificial Neural Network (ANN):

ANN is a deep learning model that draws inspiration from the structure and functioning of the human brain. It possesses the ability to learn intricate patterns and is extensively utilized in a multitude of tasks, such as image identification and natural language processing.

D. Convolutional Neural Network (CNN):

CNN is specifically designed for the purpose of picture and pattern identification. It utilizes convolutional layers to automatically extract hierarchical features, hence facilitating the analysis of visual data with exceptional performance.

E. Generative Adversarial Networks (GANs):

GANs are a deep learning system that consists of two networks, namely the generator and the discriminator. These networks are trained in an adversarial manner, where they compete against each other to generate realistic data. GANs have been widely utilized in the fields of image synthesis and data production.

IV. Experimental setup

This section will provide a comprehensive analysis of the experimental setup employed in this project.

1) **Data Loading and Exploration:** The initial step of the code is the utilization of the pandas package to load the dataset which contains 151,112 transactions, with 14,151 (9.36%) flagged as fraudulent.
2) **Feature Engineering:** The columns that have been considered unnecessary for the analysis, namely user_id, signup_time, device_id, source, browser, sex, age, and ip_address, have been excluded from the dataset. The purchase_time field is transformed into distinct temporal attributes, including year, month, day, hour, minute, and second.
3) **Undersampling:** The technique of undersampling is employed on the non-fraudulent subset in order to equalize the amount of fraudulent transactions, hence achieving a balanced class distribution.
4) **Data Preprocessing:** The dataset is partitioned into a collection of independent variables (X) and a single dependent variable (y). The dataset is again partitioned into training (80% data) and testing (20% data) sets through the utilization of the train_test_split function.
5) **Model Construction:** Logistic regression, Naive Bayes, ANN, CNN and GANs are applied to the dataset.
6) **Training the Model and Making Predictions:** The model undergoes training using the training set, and subsequently, the trained model is employed to provide predictions on the test set.
7) **Model Evaluation:** At the end the model's performance is evaluated by accuracy, precision, recall, f1-score and confusion matrix entities.

V. Result Analysis

I evaluate the dataset with five algorithms. In the case of credit card fraud, true positive means positive cases are correctly assigned to the positive class which means that fraud transactions are correctly classified as fraud. True negative means negative cases are correctly assigned to the negative class, which means that non-fraud transactions are correctly classified as non-fraud or real transactions.

Table II represents confusion metric entity of Logistic Regression, Naive Bayes, ANN, CNN and GANs. Logistic regression exhibits a high rate of false positives, indicating a tendency to incorrectly identify negative cases as positive and a relatively low rate of false negatives, indicating its superior ability to accurately identify negative instances. The Naive Bayes algorithm exhibits a balanced distribution of false positives and false negatives and overall inferior performance in comparison to logistic regression. In ANN, the number of false positives is remarkably low whereas the number of false negatives is considerably significant. CNN exhibits a balanced distribution of both false positives and false negatives, similar performance to Naive Bayes. The GANs exhibits a balanced occurrence of false positives and false negatives. The overall performance is outstanding, exhibiting a better level of precision.

Table III presents comparison with Logistic Regression, Naive Bayes, ANN, CNN and GANs based on accuracy, precision, recall and f1-score. The logistic regression model has relatively low accuracy, but it shows decent precision and high recall. This suggests that it is effective at capturing fraudulent transactions but may also generate a significant number of false positives. The Naive Bayes model performs slightly better than logistic regression with balanced precision and recall. The ANN model has the highest accuracy among the models, with very high precision but a lower recall. It excels in correctly identifying non-fraudulent transactions but may miss some fraudulent ones. The CNN model performs similarly to Naive Bayes, with balanced precision and recall. The GAN model shows good overall performance, with high precision and decent recall. The accuracy is also relatively high, indicating that the model is effective in distinguishing between positive and negative cases.

The following paragraph presents the key findings derived from this work:

1) Deep learning algorithms (ANN, CNN, GANs) perform better than traditional Machine learning algorithms (LR, NB) in credit card fraud detection.

TABLE II: Basic metric rates comparison with Logistic Regression, Naive Bayes, ANN, CNN and GANs

Metrics	Logistic Regression	Naive Bayes	ANN	CNN	GANs
True Positive Rate	2363	2287	1546	2307	1901
False Positive Rate	1807	1587	51	1583	754
True Negative Rate	994	1214	2750	1218	2047
False Negative Rate	497	573	1314	553	959

TABLE III: Comparison with Logistic Regression, Naive Bayes, ANN, CNN and GANs

	Logistic Regression	Naive Bayes	ANN	CNN	GANs
Accuracy	0.5930047694753577	0.6184419713831478	0.758876523582406	0.6226815050344462	0.6974032856385798
Precision	0.5666666666666667	0.5903458957150233	0.9680651221039449	0.5930591259640103	0.7160075329566855
Recall	0.8262237762237762	0.7996503496503496	0.5405594405594406	0.8066433566433566	0.6646853146853147
F1-score	0.6722617354196301	0.6792396792396793	0.6937401839802558	0.6835555555555556	0.6893925657298277

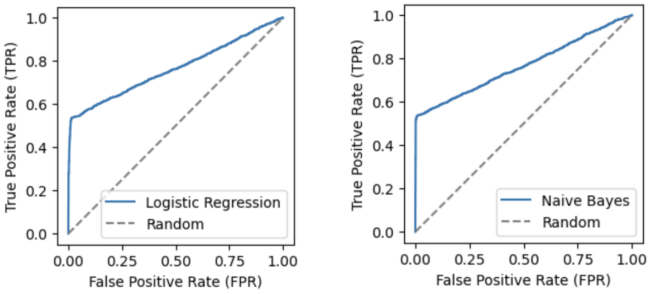

Fig. 1: ROC Curve of Logistic Regression and Naive Bayes

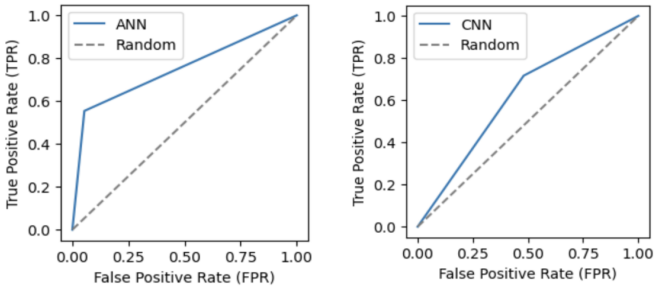

Fig. 2: ROC Curve of ANN and CNN

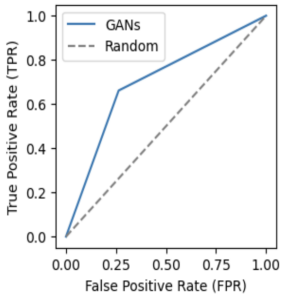

Fig. 3: ROC Curve of GANs

2) The CNN and GAN both provide a good balance between precision and recall, making them potentially suitable for scenarios where missing fraudulent transactions is costly.
3) The objective of reliably detecting credit card fraud, while minimizing the occurrence of false positives, necessitates the use of a model that exhibits high accuracy and a favorable trade-off between precision and recall. The Generative Adversarial Network (GAN) demonstrates considerable potential as a viable option among other models.

VI. Discussion

Within this section, an analysis will be conducted to evaluate the merits and limitations of my research. Additionally, a comparative examination will be undertaken to compare the findings of my study with the selected research article.

Strengths of my research are- (1) Algorithm diversity- I have applied 2 machine learning and 3 deep learning algorithms. (2) Dataset balancing- The dataset was undersampled to address the class imbalance problem and mitigate potential biases towards the majority class in the models. (3) Evaluation metrics- I have placed significant emphasis on the significance of precision and recall as metrics for evaluating model performance, particularly in the context of credit card fraud situations, rather than relying solely on accuracy.

Weaknesses of my research are- (1) Optimizing the hyperparameters can significantly improve model performance. (2) The comprehension of model interpretability is of utmost importance and should be taken into account in order to get insight into the decision-making process.

Comparison between my work and chosen research paper [1] work:

1) The research paper didn't consider the class imbalance issue while I have undersampled my dataset to train the model to equal number of fraud and non-fraud cases to mitigate potential biases towards the majority class.
2) The research paper primarily concentrated on accuracy as the sole metric for evaluating model performance, while I underlined the significance of incorporating precision and recall as additional evaluation metrics.

References

[1] Z. Li, H. Zhang, M. Masum, H. Shahriar, and H. Haddad, "Cyber fraud prediction with supervised machine learning techniques," in *Proceedings of the 2020 ACM Southeast Conference*, 2020, pp. 176–180.
[2] "Credit card fraud detection," https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud, (Accessed on 11/03/2023).