Enhancing Credit Card Fraud Detection: A Machine Learning Approach

Mohamed Affan M   
School of Advance Sciences  
VIT UniversityVellore,India  
mohamedaffan.m2022@vitstudent.ac.in

Mohammed Saqui T  
School of Advance Sciences  
VIT UniversityVellore,India  
mohammedsaqui.t2022@vitstudent.ac.in

Dr. Khadar Babu SK  
School of Advance Sciences  
VIT UniversityVellore,India  
khadarbabu.sk@vit.ac.in

***Abstract—In the rapidly evolving digital business landscape, the surge in online activities has been paralleled by a sharp increase in fraudulent incidents. Traditional rule-based fraud detection systems, while effective to an extent, fall short in addressing the dynamic and sophisticated nature of contemporary online fraud. This research paper delves into the realm of machine learning-driven fraud detection, focusing on the deployment of Amazon SageMaker as a comprehensive solution. The primary objective of this study is to develop a fraud detection framework that possesses adaptability, self-improvement, and scalability, all crucial aspects for businesses operating in the online sphere. The paper showcases the integration of SageMaker into the fraud detection pipeline, highlighting its role in fostering an iterative approach to model enhancement and adaptation. By harnessing SageMaker's capabilities, businesses can proactively combat fraud, swiftly respond to evolving tactics, and scale their defenses according to the demands of their online operations. As businesses continue to migrate online, the proposed approach presents a promising avenue for fortifying their security measures against the ever-evolving landscape of online fraud.***

***Keywords: digital fraud detection, machine learning models, Amazon SageMaker, adaptive and dynamic, online business security,* scalability, iterative enhancement.**

1. INTRODUCTION

Customer transactions are increasingly conducted electronically as a result of the digital revolution in society. More and more payment systems and services are available online, reducing the need to visit a physical financial institution. Which increase the risk of fraudulent activities and credit card fraud is a significant concern in the financial industry, causing substantial losses and eroding trust in the safety of electronic transactions. The evolving tactics of fraudsters continuously outpace the capabilities of traditional rule-based fraud detection systems[2]. In this research paper, we introduce an innovative approach to address this persistent issue by harnessing the potential of Machine Learning (ML). Our strategy involves the utilization of Amazon Web Services (AWS) SageMaker, a potent cloud-based ML platform. Machine Learning enables computers to learn and adapt from data, and AWS SageMaker offers a scalable and versatile environment for this purpose. It comes equipped with pre-built ML models and robust tools for data preprocessing, model training, and deployment. Our research investigates the implementation of AWS SageMaker to develop a more robust and effective credit card fraud detection system. The primary objective is to enhance the system's accuracy while minimizing false alarms. By achieving this, we can reduce financial losses and bolster the security of electronic transactions, thus enhancing trust and confidence in financial operations.

1. LITERATURE REVIEW

The paper "Deep Learning Methods for Credit Card Fraud Detection," authored by Thanh Thi Nguyen, Hammad Tahir, Mohamed Abdelrazek, and Ali Babar, discusses the increasing rate of credit card fraud and the challenges involved in detecting and preventing it. It explores the use of machine learning techniques, particularly deep learning, to improve fraud detection systems. The authors compare the performance of deep learning methods with traditional machine learning algorithms on different financial datasets and find that deep learning approaches show great promise for real-world credit card fraud detection systems[1].

The article discusses the need for fraud detection in payment systems and proposes using machine learning and a cloud-based automated system for detection. The authors describe the structure and setup of the system, using AWS as the platform. They also highlight the importance of developing a specific methodology for implementing the software product. The article includes a literature review on fraud detection in the banking industry and discusses the challenges and objectives of the project. The authors present their research results, including the choice of model and deployment of the model as automated technology. They also discuss the benefits of using AWS for the implementation[8].

The article “Credit Card Fraud Detection using Machine Learning Methodology” authored by Hamzah Ali Shukur and Sefer Kurnaz discusses the use of machine learning methodology such as a logistic regression model, a k-means clustering model, and a neural network for credit card fraud detection. It highlights the prevalence of fraudulent e-card activities and the need for effective detection systems. Data processing techniques and machine learning classifiers are used to identify suspicious transactions. Preprocessing the dataset improves accuracy. The article compares different machine learning models and finds that logistic regression performed the best. However, further research and experimentation could be done to optimize the neural network model. The article concludes by mentioning the importance of detecting fraud to protect legitimate customers[4].

1. METHODOLOGY USED:
2. Data Collection:

Transactions by European cardholders with online credit cards are included in the dataset. During the 2-day period, there were 491 fraudulent transactions out of 284,808 total transactions. As a result Fig.2, fraud occurs in approximately 1 out of 579 transactions, or 0.172% of all transactions. In addition to 28 anonymized variables, there are also two named variables "Time" and "Amount." The masked variables are a result of transforming the original data for confidentiality purposes using a Principal Component Analysis (PCA). Amount refers to the transaction amount and the time variable contains the seconds taken between each of the transactions and the first transaction in the dataset.

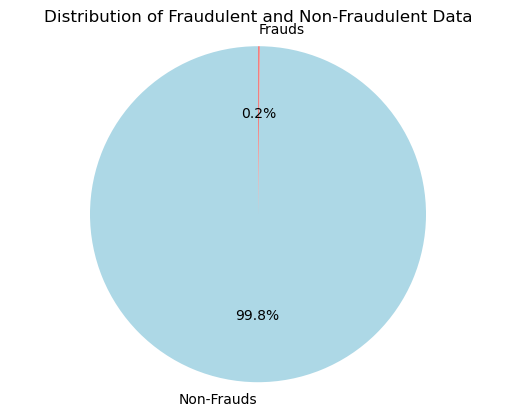


Figure 2 Distribution of data

1. Data Pre-Processing

In the initial stages of our research, data preprocessing is a pivotal step Fig.2. We primarily focus on two critical tasks: splitting the data into training and testing datasets and uploading these datasets to an Amazon S3 bucket.

1. *Data Splitting:* Our dataset is divided into two subsets, with 80% allocated for training and 20% for testing. The training data helps the model learn from historical transaction data, while the test data remains unseen for model evaluation.
2. *Uploading to S3 Bucket:* We use Amazon S3 for storing and managing our datasets. Uploading data to an S3 bucket streamlines data accessibility for AWS SageMaker, ensuring efficient model training and evaluation.

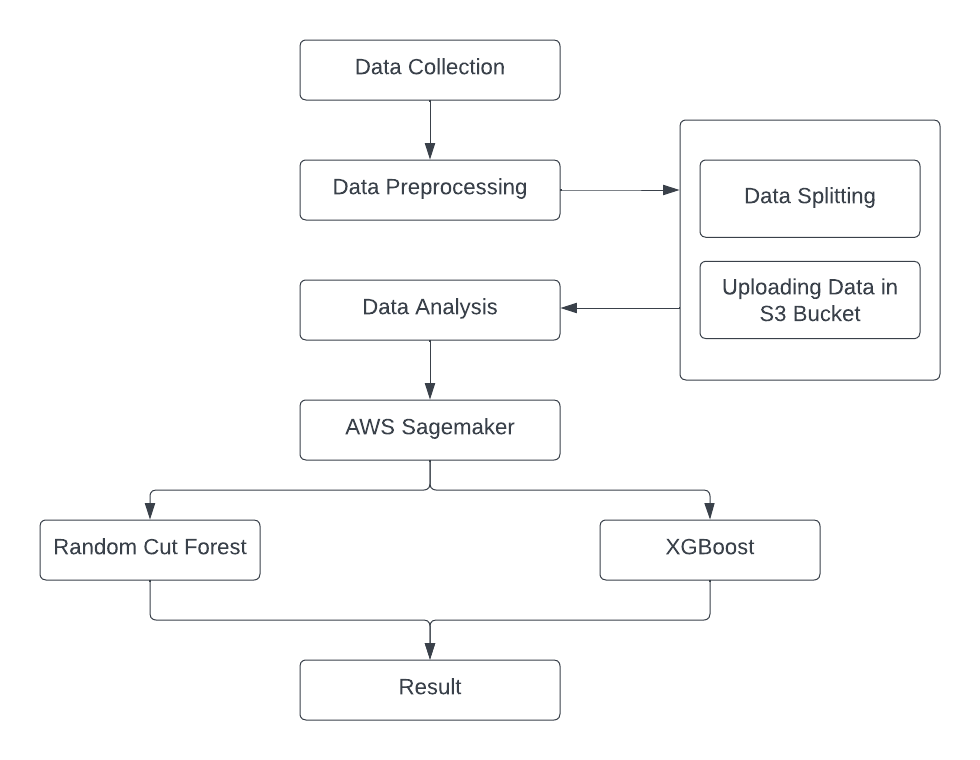


Figure 1 work flow or Architecture diagram

1. IMPLEMENTATION
2. Random-cut forest

Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm designed for anomaly detection across arbitrary-dimensional datasets. It assigns an anomaly score to each data point, allowing identification of divergent observations. RCF excels in handling various types of anomalies, such as unexpected spikes or breaks in periodicity. The algorithm scales efficiently with the number of features, dataset size, and instances, making it versatile for anomaly detection tasks.

1. XGBoost Algorithm

XGBoost the open-source gradient boosted trees approach is particularly well-executed by XGBoost, commonly referred to as eXtreme Gradient Boosting.. This algorithm, a type of supervised.learning, excels in predicting a target variable by combining estimates from a group of simpler models. XGBoost's strength lies in its ability to handle diverse data types, relationships, and distributions. Its flexibility, coupled with the option to fine-tune numerous hyperparameters, makes it a popular choice in machine learning competitions. Whether you're dealing with regression, binary or multiclass classification, or ranking problems, XGBoost proves itself as a versatile and powerful tool.

1. RESULT AND DISCUSSION
2. Training of Random Cut Forest model.

In fraud detection, limited labeled examples prompt the use of anomaly detection for insights from unlabeled data. Leveraging data imbalance, Random Cut Forest, an accurate and scalable algorithm, assigns anomaly scores to identify anomalies. Employing SageMaker's RCF, we train a model on the training set and predict anomalies in the test set.

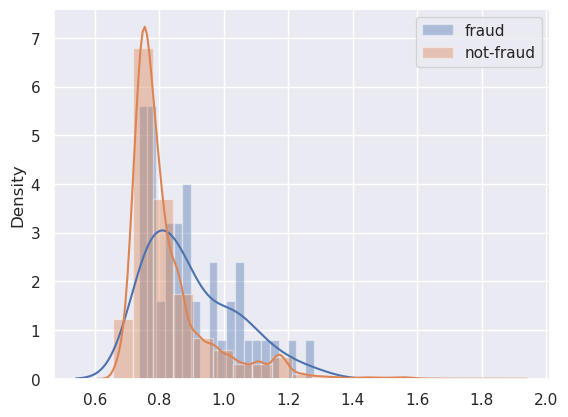


Figure 3 Histogram of Anomaly scores

The graph Fig.3 illustrates a distinct separation in anomaly scores between fraud and non-fraud transactions. Over 90% of transactions have scores below 1.0, with fraud concentrated at higher scores compared to non-fraudulent transactions that generally tends to get lower anomaly scores. Peaks in fraud transactions occur around 1.2 and above 1.6, while non-fraud transactions rarely exceed 1.2. An effective threshold for identifying fraud could be approximately 1.6, prompting further investigation for transactions surpassing this score.

Finally, we assess the classification outcomes against the actual labels and calculate evaluation metrics.

|  |  |
| --- | --- |
|  | RCF |
| Accuracy | 0.7438 |
| F1 Score | 0.1397 |
| Cohen’s Kappa | 0.0718 |

Table Random cut Forest matrices

1. Training an XGBoost a supervised model

After gathering a sizable labelled training dataset, we use supervised learning technique to reveal feature-class correlations. It is imperative to select the XGBoost algorithm because of its reputation for dependability, scalability, and proficiency with handling missing data. In order to avoid the majority class's (non-fraudulent) predominate influence during the learning process, addressing data imbalance is essential at this phase.

Next, we assess our model using the three metrics that were previously described in the previous stage.

|  |  |  |
| --- | --- | --- |
|  | RCF | XGBoost |
| Accuracy | 0.7438 | 0.8568 |
| F1-Score | 0.1397 | 0.7551 |
| Cohen’s kappa | 0.0718 | 0.7548 |

Table 2 RCF and XGBoost Matrices

In addition to single-value measurements, metrics that show performance by class are also helpful to examine. More details regarding the effectiveness of our model can be found, for instance, in the confusion matrix, recall, per-class precision, and F1-score. The confusion matrix Fig.4 we found is

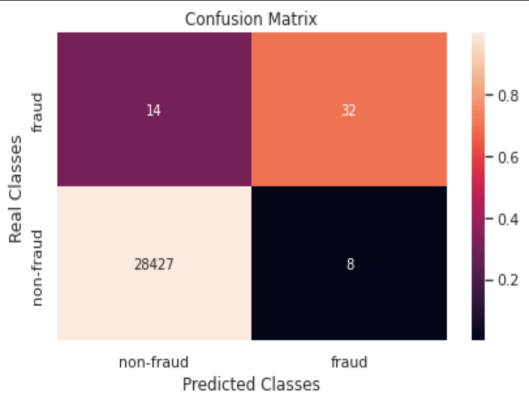


Figure 4 Confusion matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Re-call | F1-score | Support |
| Negative(non-fraud) | 1.00 | 1.00 | 1.00 | 28415 |
| Positive(fraud) | 0.79 | 0.72 | 0.72 | 54 |

Table 3 matrices of each class/Attributes

1. CONCLUSION

Traditional fraud detection systems face challenges in adapting to dynamic fraud tactics, prompting the exploration of machine learning-driven solutions. Leveraging Amazon SageMaker, a comprehensive cloud-based ML platform, we propose a framework that prioritizes adaptability, self-improvement, and scalability. In our research the RCF model, proposing a fraud threshold of 1.6, achieved 74.38% accuracy. XGBoost addressed imbalance, achieving 85.68% accuracy, 75.51.Next step we plan to automate the fraud detection system using AWS API Gateway that is backed by AWS Lambda helping financial institutions to reduce the risk of fraudulent transactions in real time with more accuracy.

REFERENCE

1. A. Agrawal, S. Kumar and A. K. Mishra, "Credit card fraud detection: a case study," 2nd International Conference on Computing for Sustainable Global Development, New Delhi, 2015
2. Ashfaq, T.; Khalid, R.; Yahaya, A.S.; Aslam, S.; Azar, A.T.; Alsafari, S.; Hameed, I.A. A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism. Sensors 2022, 22, 7162. https:// doi.org/10.3390/s22197162
3. A. O. Adewumi and A. A. Akinyelu, "A survey of machine-learning and nature-inspired based credit card fraud detection techniques." International Journal of System Assurance Engineering and Management 8, no. 2 (2017): 937-953.
4. H. A. Shukur and S. Kurnaz, "Credit Card Fraud Detection using Machine Learning Methodology," in International Journal of Computer Science and Mobile Computing, vol. 8, no. 3, pp. 257-260, March 2019.
5. Ileberi, E., Sun, Y. & Wang, Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. J Big Data 9, 24 (2022). <https://doi.org/10.1186/s40537-022-00573-8>
6. Jegadeesan, Karthika & Ayothi, Senthilselvi. (2022). Credit Card Fraud Detection based on Ensemble Machine Learning Classifiers. 1604-1610. 10.1109/ICESC54411.2022.9885649.
7. Ngai, E.W.T.; Hu, Y.; Wong, Y.H.; Chen, Y.; Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis. Support Syst. 2011, 50, 559–569.
8. Nguyen, Thanh & Tahir, Hammad & Abdelrazek, Mohamed & Ali Babar, Muhammad. (2020). Deep Learning Methods for Credit Card Fraud Detection.
9. Pujari, S., Diwanji, K., Malegaonkar, S., Shaikh, S., & Bhosale, A. (2023). Fraud Detection in Credit Card Automated System using ML with AWS SageMaker. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(V), 1867. URL: [www.ijraset.com](http://www.ijraset.com/)
10. Vasilios Plakandaras, Periklis Gogas, Theophilos Papadimitriou & Ioannis Tsamardinos (2022) Credit Card Fraud Detection with Automated Machine Learning Systems, Applied Artificial Intelligence, 36:1, 2086354, DOI: 10.1080/08839514.2022.2086354