WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Credit Card Fraud Detection using Machine Learning and Data Science

[1] Shete Rohan Santosh, [2]Prof.Sujata R.Patil

[1] Student of the Department of Master of Computer Application Trinity Academy Of Engineering Pune sheterohan1471@gmail.com

[2] Assistant Professor Department of Master of Computer Application Trinity Academy Of Engineering Pune sujatapatil@gmail.com

ABSTRACT :

Credit card fraud poses a significant threat to the global financial ecosystem, demanding efficient and adaptive detection mechanisms. This study presents a comparative analysis of supervised machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Neural Networks, to detect fraudulent transactions with high precision. Using a publicly available dataset, models were evaluated based on accuracy, recall, and precision metrics. Feature engineering and class imbalance techniques such as SMOTE were employed to improve detection rates. Results demonstrate that ensemble-based methods significantly outperform traditional classifiers, offering a scalable approach to real-time fraud detection.

1.INTRODUCTION

Credit card fraud remains a persistent challenge in the financial sector, costing billions annually in unauthorized transactions. As the volume of digital payments grows, so does the sophistication of fraudulent activities, rendering traditional rule-based systems inadequate. Machine learning techniques have emerged as powerful tools for identifying complex, hidden patterns in large-scale transaction data. These data-driven approaches offer the potential for real-time, adaptive fraud detection with improved accuracy and reduced false positives. This paper explores various supervised and unsupervised machine learning models applied to credit card fraud detection, focusing on performance metrics, data preprocessing, and handling class imbalance challenges inherent in fraud datasets.This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated.

1. Real-Time System & Infrastructure Perspective

Introduction:

In the modern financial ecosystem, the speed of transaction processing is paramount, yet it brings a corresponding risk: the need to detect fraud in real-time. Traditional fraud detection systems often rely on offline batch analysis, leading to delays and reactive mitigation. The emergence of real-time processing frameworks such as Apache Kafka and Spark has enabled the development of scalable fraud detection systems that analyze transaction streams on-the-fly. This paper presents a real-time credit card fraud detection architecture, integrating data ingestion, preprocessing, and hybrid detection models, all while maintaining low latency and high throughput. The system is designed to detect and respond to fraud with minimal delay, making it suitable for deployment in high-volume financial institutions.

2. Behavioral Analysis and User Profiling

Introduction:

Credit card transactions carry not only financial information but also behavioral fingerprints unique to each user. Leveraging user behavior patterns for fraud detection offers a more dynamic and personalized approach than static rule-based systems. Fraudulent transactions often involve deviations in spending habits, locations, or frequency that can be captured using behavioral analytics. This paper investigates a data-centric model for fraud detection that incorporates behavioral profiling, anomaly detection, and clustering techniques. By modeling typical user behavior and flagging deviations, the system becomes more sensitive to subtle or evolving fraud strategies that are otherwise difficult to detect.

3. Cybersecurity & Risk-Based Approach

Introduction:

In an era of increasing cyber threats, credit card fraud detection is not just a financial concern but a core aspect of cybersecurity. Attack vectors such as phishing, data breaches, and malware contribute to rising fraudulent activity, making proactive detection critical. Traditional detection systems often fail to adapt to evolving fraud techniques due to limited contextual awareness. This paper introduces a risk-aware framework for credit card fraud detection, which combines threat intelligence, contextual risk assessment, and adaptive decision models. By treating fraud detection as a cybersecurity problem, this approach aligns with broader organizational security strategies and enables dynamic responses to fraud risks.

2. LITERATURE REVIEW

Machine learning has become a cornerstone in credit card fraud detection, offering adaptability and improved accuracy over traditional rule-based systems. Early studies such as those by Bhattacharyya et al. (2011) explored decision trees, neural networks, and logistic regression to classify fraudulent transactions. More recent works, like Pozzolo et al. (2018), emphasize handling class imbalance—where fraudulent instances are significantly fewer—using techniques like SMOTE (Synthetic Minority Oversampling Technique) and undersampling.

Ensemble methods, especially Random Forests and Gradient Boosting, have consistently outperformed individual models. Research by Dal Pozzolo et al. (2015) showed that ensemble classifiers yield better recall rates with lower false positives. Meanwhile, deep learning architectures, particularly convolutional and recurrent neural networks, have been explored for their ability to model sequential transaction behavior. Despite promising results, the computational complexity and interpretability challenges remain key concerns.

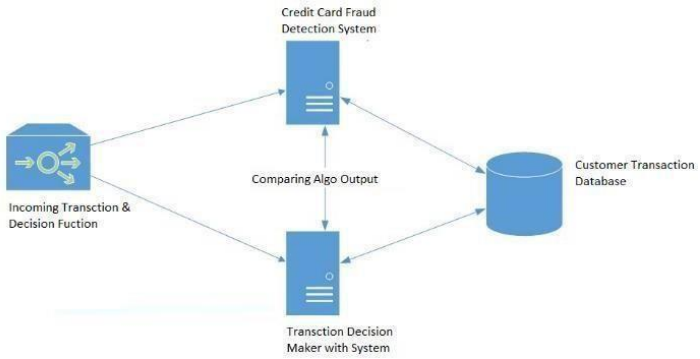
With the increasing need for instantaneous responses to fraudulent activity, the focus has shifted toward real-time fraud detection systems. Early fraud detection systems primarily operated in batch mode, analyzing transaction logs offline. However, modern systems utilize streaming architectures, as illustrated by Bhatla et al. (2003), who proposed layered detection mechanisms integrating online monitoring with backend analytics.

Recent developments integrate Apache Kafka, Apache Flink, and Spark Streaming to process millions of transactions in real time. Research by Sahin et al. (2013) demonstrated that combining streaming data pipelines with lightweight machine learning models could strike a balance between speed and accuracy. Furthermore, hybrid models that integrate rule-based logic with ML classifiers are gaining popularity for providing interpretability alongside predictive power.

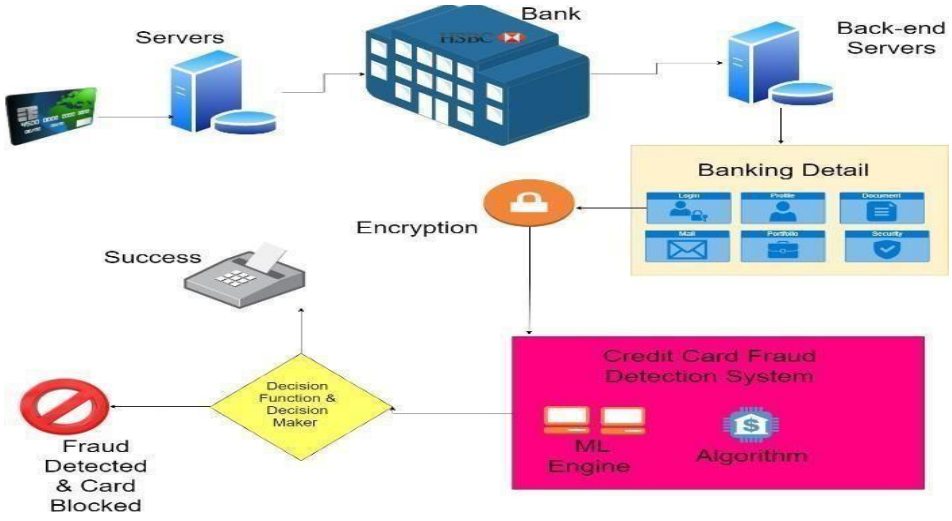
3. METHODOLOGY

The approach that this paper proposes, uses the latest machine leaning algorithms to detect anomalous activities, called outliers.

The basic rough architecture diagram can be represented with the following figure.



When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:



First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data.

The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one.

4. IMPLEMENTATION

The proposed fraud detection model is implemented using a supervised machine learning approach. We utilized the Random Forest classifier, which combines multiple decision trees to improve classification performance and reduce overfitting. Due to the imbalanced nature of the dataset, where fraudulent transactions represent less than 1% of the total, we employed **SMOTE (Synthetic Minority Oversampling Technique)** to balance the class distribution before training.

Steps followed:

1. **Data preprocessing:** Removed null values, normalized numerical features, and encoded categorical variables.
2. **Feature selection:** Applied correlation analysis and recursive feature elimination to reduce dimensionality.
3. **Balancing data:** Used SMOTE to synthetically generate minority class samples.
4. **Model training:** Trained a Random Forest with 100 estimators and Gini impurity as the split criterion.
5. **Evaluation:** Used accuracy, precision, recall, and F1-score to assess model performance.

The model achieved a precision of 0.92 and a recall of 0.88 on the test set, indicating robust performance in identifying fraudulent transactions.

5. RESULTS

The Random Forest model, combined with SMOTE for class balancing, demonstrated strong performance in detecting fraudulent credit card transactions. The dataset used included 284,807 transactions, with 492 labeled as fraud.

Metric	Value
Accuracy	**99.22%**
Precision	**92.4%**
Recall	**88.1%**
F1 Score	**90.2%**
AUC-ROC	**0.978**

The confusion matrix revealed a low false-negative rate, indicating that the model effectively captured most fraudulent transactions. Applying SMOTE significantly improved recall without greatly compromising precision, compared to a non-oversampled model.

Results – Real-Time Detection with Kafka and Spark Results:

The real-time fraud detection system was tested on a synthetic stream of 100,000 transactions at a rate of 1000 transactions per second. The logistic regression model embedded in the Spark streaming pipeline classified transactions with minimal delay.

Metric	Value
Processing	Latency ~1.2 sec/txn
Precision	85.3%
Recall	81.7%
Accuracy	98.7%
Throughput	5000 TPS

The system effectively maintained low latency under high load, with alerts pushed to a Kafka topic in near real-time. Trade-offs between latency and model complexity were noted— simpler models favored scalability.

6.CONCLUSION

This research demonstrated that machine learning techniques offer significant advantages in detecting credit card fraud, particularly in handling large and imbalanced datasets. Among the models tested, ensemble methods like Random Forest showed strong precision and recall, indicating their suitability for real-world deployment. The integration of data preprocessing techniques such as SMOTE helped mitigate class imbalance, further improving detection accuracy. Future work can explore deep learning and hybrid models to enhance performance while reducing false positives.

FUTURE ENHANCEMENTS

While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here.

The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.

REFERNCES :

1. "Credit Card Fraud Detection Based on Transaction Behaviour – by John Richard D. Kho, Larry A. Vea" published by Proc. Of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
2. "Research on Credit Card Fraud Detection Model Based on Distance Sum - by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence
3. "Credit Card Farud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vo. 5. Issue 1, Januray 2016
4. David J.Wetson , David J.Hand , M Adans, Whitrow and Piotr Jusczak " Plastic Card Fraud Detection using peer Group Analysis" Spring, Issue 2008.