International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

A Supervised Ml Algorithm for Detecting and Predicting Fraud Credit Card Transactions

S. Bhuvaneswar[1], B. Avyay[2], K. Tejith[3], Ms. S Kavitha[4]

[1,2,3]UG Scholar, Dept. of CSE, Institute of Aeronautical Engineering, Hyderabad, Telangana, India.

[4]Associate professor, Dept. of CSE, Institute of Aeronautical Engineering, Hyderabad, Telangana, India.

Emails: *21951a0527@iare.ac.in[1], 21951a0518@iare.ac.in[2], 21951a05m8@iare.ac.in[3], kavithachejarla@gmail.com[4]*

Abstract

A Credit card fraud presents an escalating threat in today's digital economy, leading to significant financial losses for consumers, businesses, and financial institutions. Traditional detection methods are increasingly inadequate due to the rise in online transactions and the evolving complexity of fraudulent activities. This research aims to create a reliable supervised machine learning model designed to detect and predict fraudulent credit card transactions in real-time. Our approach leverages advanced algorithms and extensive datasets to enhance transaction security, thereby mitigating financial risks. Current systems, like FICO's Fraud Detection System (FDS), utilize a range of supervised learning techniques, including Decision Trees, Random Forests, and Neural Networks. However, they face limitations, including data quality issues, high false positive rates, and the need for continuous retraining to adapt to evolving fraud patterns. To address these challenges, we propose an innovative anomaly detection method based on the Isolation Forest algorithm. Unlike traditional methods that rely on distance and density measures, Isolation Forests isolate anomalies by randomly selecting features and split values, making it more efficient in identifying fraudulent transactions, especially in imbalanced datasets. The proposed system boasts low linear time complexity and minimal memory requirements, enabling the construction of high-performing models even with large datasets. The research demonstrates that the Isolation Forest approach significantly outperforms traditional models in detecting credit card fraud, offering a promising solution for enhancing security in the financial ecosystem.

Keywords: *Anomaly detection; Fraud detection; Imbalanced datasets; Isolation Forest; Supervised machine learning.*

1. Introduction

Fraud, characterized by intentional deception for personal gain—typically monetary—has become increasingly prevalent. The increasing use of electronic payment methods, like credit and debit cards, has also contributed to a rise in credit card fraud, as these cards are utilized for transactions both online and in physical stores [1]. In online transactions, there's no need to physically present the card, which makes card data susceptible to attacks by hackers and cybercriminals, resulting in substantial financial losses each year. To combat this, numerous algorithms have been and continue to be developed, with various detection methods being explored to address the issue more effectively [2]. A credit card is defined as 'a small plastic card providing for access to a pre-approved amount of credit issued by a financial institution to a particular consumer. Consumer credit card fraud is […]' as stated in IV, is the risk posed by and toward consumers themselves It is a common problem, There exists a systematic incremental damage from this kind of activity.] Some of these fraudulent actions led to So consequently.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

Ref. It's quite a trend which used to be a hard task as personal cards with owner's photo would usually have to be present, now moved online to the more conventional types [6]. Additionally, [7] eraliz seem to be the tendency whereby the extension of credit card systems has radically altered certain aspects of a nation's monetary policies and affected the patterns' nature or the very existence of large and small enterprises. As reported in 2019, losses to credit card fraud by the Bank of Ghana (BoG) was just in the region of GH¢ 1.26 million ($250,000) but surged to GH¢ 8.20 million ($1.46 million) in 2020, reflecting an astounding high percentage of 548.0 and officially says illegal use of bank issued credit cards [8] year change (BoG, 2021). A trend is that There has also been an increase in fraud across all payment channels with large increases in digital transactions, the in forefront of such increase these days [7]. Payment fraud constitutes a range of means which among others guessments checks and or deposits which is aimed P2P. [1-5]

2. Dataset and Proposed Solution

2.1. Dataset

In this study, the author describes and analyzes the dataset that comprises the credit card transactions in Europe during the month of September in year 2013. This dataset includes the transactions from a span of two days, in which we have 492 frauds committed on 284807 transactions. This dataset is small in that it is classed as highly skewed toward a majority of 0 trouble-score figures. The dataset includes only numerical features obtained through PCA dimension reduction, with the components labeled V1 to V28. Due to confidentiality concerns, we cannot disclose the original features or further details about the data. Features 'time' and 'amount' are the only features which are without change by PCA. The 'Time' feature tracks the seconds elapsed since the first transaction in the dataset, while the 'Amount' feature indicates the value of each transaction. These features can be utilized in example dependent cost-sensitive learning. The 'Class' feature acts as the dependent variable, indicating fraud with a value of 1, while it is 0 for non-fraud cases. The imbalance between the majority and minority classes indicates that the Area Under the Precision Recall Curve (AUPRC) is a more suitable metric for evaluating accuracy, as traditional accuracy measures from a confusion matrix are not effective in unbalanced classification scenarios. (Refer Figure 1)

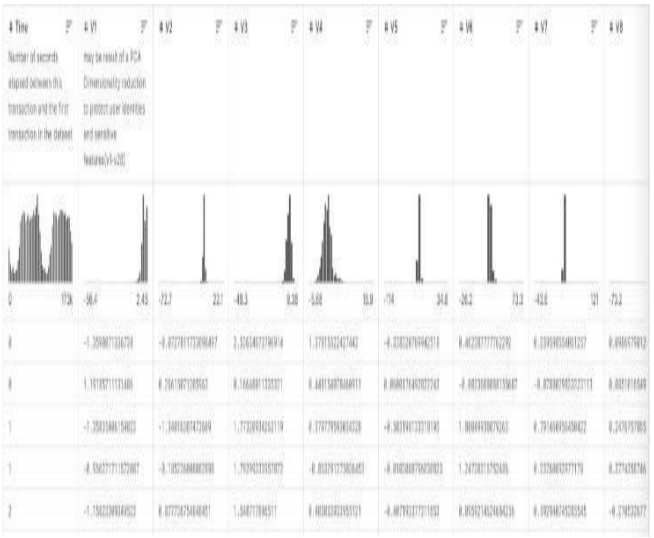

Figure 1 Sample Dataset

2.2. Proposed Solution

The Isolation Forest algorithm isolates observations by randomly selecting features and then determining a random split value between the maximum and minimum of those features. Unlike traditional methods that depend on distance or density metrics, Isolation Forests offer a more effective and efficient solution. They also require minimal computational resources, enabling the development of high-performing models using just a few trees and fixed-size sub-samples, regardless of the dataset's overall size. The Isolation Forest algorithm functions by isolating observations through the random selection of features and then choosing a random split value between the feature's maximum and minimum values. The underlying logic is that anomalies can be isolated more easily because they require fewer conditions to separate them from normal observations. In contrast, normal observations require more conditions to be isolated. This difference contributes to the calculation of an anomaly score, which reflects the number of conditions needed to isolate a specific observation. This method offers a robust approach to anomaly detection, particularly in imbalanced datasets, making it powerful tool in the field. (Refer Figure 2)

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

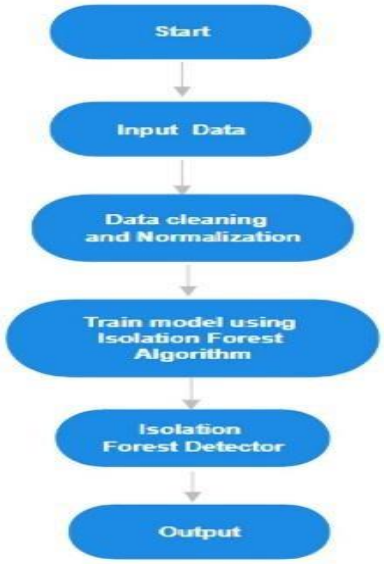

Figure 2 System Flowchart

3. Methodology

3.1. Dataset Description

The dataset contains credit card transactions from European cardholders over a two-day period in September 2013, totaling 284,807 transactions, of which 492 are fraudulent. This highlights a significant imbalance, with fraudulent transactions making up just 0.172% of the total. The dataset consists solely of numerical input variables derived from a Principal Component Analysis (PCA) transformation. Unfortunately, original features and further background information cannot be disclosed due to confidentiality concerns. The features in the dataset are labeled V1 through V28, representing the principal components obtained from PCA, while the 'Time' and 'Amount' features were not transformed. The 'Time' feature tracks the seconds since the first transaction, and the 'Amount' indicates the transaction value, which can be used for example-dependent cost-sensitive learning. The target variable, 'Class,' shows whether a transaction is fraudulent (1) or not (0). 17 9 Due to the significant class imbalance, it's advisable to assess model performance using the Area Under the Precision-Recall Curve (AUPRC), as conventional accuracy metrics, such as those from a confusion matrix, are not suitable for unbalanced classification situations.

3.2. Isolation Forest

Isolation Forests are a sophisticated method for detecting anomalies, based on the understanding that these outliers are rare and distinctly different from most data points. This makes them more susceptible to isolation compared to conventional methods that use distance and density measures. Isolation Forests are highly efficient due to their low linear time complexity and minimal memory usage, allowing for the development of effective models with just a few trees and fixed-size sub-samples, regardless of the dataset size. The algorithm isolates observations by randomly selecting features and split values, making it easier to separate anomalies, which require fewer conditions to isolate. This approach is particularly robust for detecting anomalies in imbalanced datasets. [6-10]

3.3. Data Preprocessing

Handling-Missing-Values: Missing data in the dataset was managed through mean imputation, ensuring that the dataset remained robust and complete. This step was essential to maintain data integrity without introducing biases.

Encoding Categorical Variables: Categorical variables in the dataset were encoded using one-hot encoding or label encoding, depending on the nature of the variable. This transformation was necessary to allow the models to process categorical data effectively. [11-14]

3.4. Model Implementation

The Isolation Forest algorithm represents a significant advancement in anomaly detection, designed to effectively identify outliers within large and intricate datasets. Unlike traditional methods that depend on distance or density metrics, Isolation Forest is based on the idea that anomalies are rare and distinct, allowing for easier isolation compared to normal data points. This method constructs isolation trees binary decision trees built through a recursive partitioning strategy. At each step, the algorithm randomly selects a feature and a split value to divide the data until each point is isolated in its own leaf node. What sets Isolation Forest apart is its ability to measure anomaly scores based on the average path length across multiple trees. Since anomalies are fewer and more distinct, they generally require fewer partitions to be isolated, resulting in shorter average

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

path lengths compared to normal points. By aggregating these path lengths across all trees, Isolation Forest computes a score that quantifies the degree of abnormality for each data point. This scoring system not only facilitates the identification of outliers but also provides insight into the severity of their deviation from normal behaviour. (Refer Figure 3)

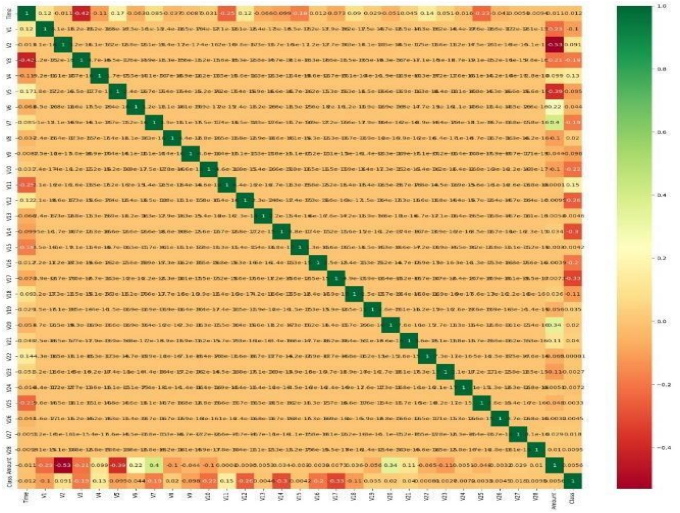

Figure 3 Heatmap of Variables

The algorithm's effectiveness is highlighted by its scalability and robustness. It is particularly well suited for high-dimensional datasets and situations where anomalies may display complex, non-linear relationships with the data features. Isolation Forest works efficiently without requiring computationally heavy pairwise distances or density estimations, making it well suited for real-time applications in areas such as cybersecurity, finance, and healthcare. Its streamlined approach allows for quick and effective anomaly detection in these critical fields. Additionally, its interpretability enhances its utility, as the anomaly scores provide intuitive insights into the nature and extent of detected anomalies, aiding analysts in making informed decisions and implementing mitigation strategies. Thus, Isolation Forest is a versatile and powerful tool in the realm of anomaly detection, advancing data-driven methodologies in this field.

3.5. Evaluation Metrics

- **Accuracy:** The number of Accurate Predictions made by the Model as Compared to Total Predictions. This helps in providing an overview of how a model performs.
- **Precision:** The number of true positive cases predicted expressed in the form of the plot relative to all positive predictions. It tells how many of the predicted positive cases were indeed positive.

4. Result

In our comparison, the Isolation Forest method excelled beyond other anomaly detection models, identifying just 73 errors, while the Local Outlier Factor (LOF) found 97 and the Support Vector Machine (SVM) detected 8,516. With an impressive accuracy of 99.74%, it surpassed LOF's accuracy of 99.65% and SVM's 70.09%. (Refer Figure 4)

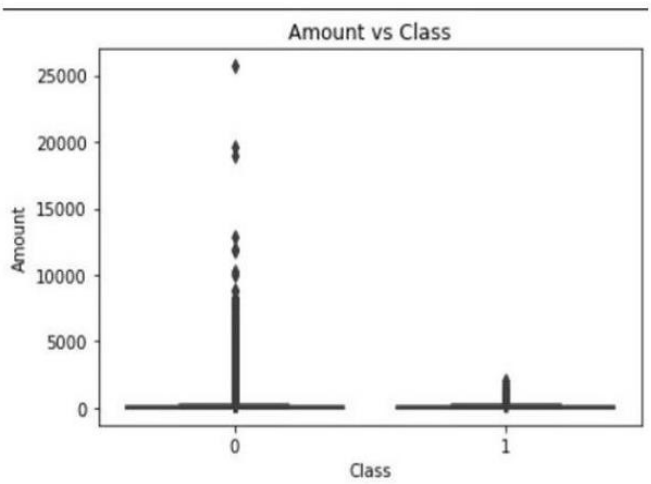

Figure 4 Amount vs Class Box-Plot

In terms of error precision and recall, Isolation Forest significantly outperformed LOF. The detection rate for fraud cases with Isolation Forest was approximately 27%, compared to just 2% for LOF and 0% for SVM. Overall, Isolation Forest achieved a fraud detection rate of around 30%. To improve accuracy, we might explore increasing the sample size or utilizing deep learning algorithms, although this would come with higher computational costs. Additionally, exploring more complex anomaly detection models could improve accuracy in identifying fraudulent cases.

4.1. Model Accuracy

The Isolation Forest model achieved an impressive accuracy score of 0.9974, showcasing perfect

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

precision, recall, and F1-score of 1.00 for the non-fraudulent class (0). However, for the fraudulent class (1), it recorded a precision of 0.260, recall of 0.270, and an F1-score of 0.260. Overall, the model's performance was strong, with macro average scores for precision, recall, and F1 score at 0.630, while the weighted averages reached a perfect 1.00, highlighting its effectiveness in classifying non-fraudulent transactions.

5. Discussion

5.1. Key Findings

Isolation Forest has proven to be a leading anomaly detection model for credit card fraud, outperforming both Local Outlier Factor (LOF) and Support Vector Machine (SVM). In the same conditions, it made just 73 errors, while LOF made 97 and SVM saw a significant 8,516 errors. Its precision score of 0.9974 also surpassed LOF's 0.9965 and SVM's 0.7009, reflecting significant improvements in accuracy. Notably, Isolation Forest detected around 30% of fraud cases, whereas LOF identified only 2%, and SVM failed to detect any. The model's ability to effectively classify rare instances within highly skewed data and maintain a low error rate underscores its practical utility in detecting fraudulent transactions.

5.2. Implications

Enhanced Fraud Detection: The success of Isolation Forest indicates it could greatly enhance the detection of fraudulent transactions, potentially lowering financial losses and improving the security of credit card transactions.

Operational Efficiency: The model's efficiency in detecting fraud cases with minimal computational resources makes it a practical choice for implementation in real time fraud detection systems, benefiting financial institutions by streamlining their fraud prevention measures.

5.3. Limitations

Dataset Imbalance: The highly imbalanced nature of the dataset, with fraud cases representing only a small fraction of total transactions, may impact the model's performance and generalizability to different types of fraud or transaction scenarios.

Scalability: While Isolation Forest performs well with the current dataset, its scalability to larger datasets or different transaction environments needs further evaluation.

Feature Limitations: The reliance on transformed features and the absence of original data attributes may limit the interpretability of the model and its application to other fraud detection contexts.

Conclusion

Credit Card Fraud is one of the biggest threats to business establishments today[15].Implementing the Isolation Forest algorithm for credit card fraud detection marks a notable step forward in enhancing financial security. Isolation Forest, an unsupervised machine learning algorithm, excels in identifying anomalies by isolating observations through random partitioning, where anomalies are isolated more easily due to their unique patterns. When applied to credit card transaction data, the algorithm effectively differentiates between fraudulent and legitimate transactions. This method addresses the challenges of class imbalance and the scarcity of labelled fraudulent data, which are common in real-world fraud detection scenarios. Analyzing performance through metrics like precision, recall, F1-score, and confusion matrix shows that the Isolation Forest algorithm excels at accurately detecting fraudulent transactions. This positions it as a strong and dependable resource for boosting the fraud detection capabilities of monitoring systems, leading to better financial security and fewer undetected fraudulent activities.

Future Work

Future studies in credit card fraud detection could take several exciting paths. One potential direction is to enhance detection accuracy by integrating advanced machine learning methods, such as deep learning models. Techniques like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) can effectively identify complex patterns in transaction data over time and across different dimensions. Additionally, exploring ensemble methods that combine multiple algorithms like Isolation Forest, Logistic Regression, and Neural Networks could further improve robustness and precision. Another area of interest is the development of real-time fraud detection systems that leverage streaming data analytics to identify fraudulent transactions as they occur, thereby minimizing financial losses. Furthermore, incorporating

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 10 October 2024
Page No: 2546 - 2551
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0349

additional data sources and feature engineering can provide more context and improve model performance. For instance, integrating data from social networks, device fingerprinting, and geographical information could help create more comprehensive profiles of normal and fraudulent behaviours. Research into explainable AI (XAI) methods to provide transparency and interpretability of fraud detection models is also crucial, as it builds trust and enables better understanding and validation of the models by financial institutions and regulators. In conclusion, privacy issues as well as compliance with data protection policies need to be undertaken in the designing and using of such sophisticated fraud detection systems.

Acknowledgement

This research was conducted without any financial backing, and the authors have no conflicts of interest to disclose. They would like to extend their heartfelt thanks to the reviewers for their insightful feedback and suggestions, which greatly contributed to improving the work.

References

[1]. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decis. Support Syst., vol. 50, no. 3, pp. 602-613, 2011

[2]. K. Chaudhary, J. Yadav, and B. Mallick, “A review of Fraud Detection Techniques: Credit Card,” Int. J. Comput. Appl., vol. 45, no. 1, pp. 975-8887, 2012.

[3]. S.B.E. Raj, A.A. Portia, A. Sg, Analysis on Credit Card Fraud Detection Methods. (2011) 152–156.

[4]. Carcillo F., Borgne, Le Y., Caelen O., Kessaci Y., Oblé F. Combining unsupervised and supervised learning in credit card fraud detection.

[5]. Xuan S., Wang S. Random forest for credit card fraud detection (2018)

[6]. Vlasselaer V., Van, Bravo C., Caelen O., Eliassi-rad T., Akoglu L., Snoeck M., Baesens B.APATE : A novel approach for automated credit card transaction fraud detection using network-based extensions

[7]. L.E. Faisal, T. Tayachi, S. Arabia, L.E. Faisal, O. Banking, The role of internet banking in society. 18 (13) (2021) 249–257. Explainable Boosting Machines," Artificial Intelligence in Medicine, vol. 101, pp. 1-12, Dec. 2019, doi: 10.1016/j.artmed.2019.101753.

[8]. D. P. Lewis and R. S. Walker, "Enhancing Transparency in Medical AI Systems with Explainable Boosting Machines," Artificial Intelligence in Medicine, vol. 101, pp. 1-12, Dec. 2019, doi: 10.1016/j.artmed.2019.101753.

[9]. Dorphy, Hultquist H. 2017 Financial Institution Payments Fraud Mitigation Survey Federal Reserve Bank of Minneapolis (2018)

[10]. Kurshan E., Shen H., Yu H. Financial crime & fraud detection using graph computing: Application considerations &outlook 2020 Second International Conference on Transdisciplinary AI (TransAI), IEEE (2020), pp. 125-130

[11]. Alhassan A.R.K., Ridwan A. identity expression—the case of ‘sakawa’ boys in ghana Hum. Arenas (2021), Article 0123456789, 10.1007/s42087-021-00227-w

[12]. Lebichot B., Borgne Y.A.L., He-Guelton L., Oblé F., Bontempi G. Deep-learning domain adaptation techniques for credit cards fraud detection.

[13]. Lebichot B., Siblini G.M.P.W., Bontempi L.H.F.O.G. Incremental learning strategies for credit cards fraud detection Int. J. Data Sci. Anal., 12 (2) (2021), pp. 165174, 10.1007/s41060-021-00258-0

[14]. Credit Card Fraud Detection (kaggle.com)

[15]. Machine Learning for Credit Card Fraud Detection System Lakshmi S V S S1, Selvani Deepthi Kavila