ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

ONLINE FRAUD PAYMENT DETECTION USING BALANCED ML ALGORITHM

[1]VARDHINEEDI LAKSHMI PRASANNA, [2]V.BHASKARA MURTHY

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Professor & Hod**,** Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

Credit card fraud has become a major concern in digital transactions, leading to significant financial losses for individuals and organizations. This project focuses on developing a machine learning-based system to detect fraudulent transactions using a highly imbalanced dataset containing approximately 28,000 records. The dataset consists of two classes: fraud and non-fraud transactions, where non-fraud cases significantly outnumber fraudulent ones. To address this imbalance, the system applies the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples and balance the dataset, improving model performance. The system implements multiple machine learning algorithms such as Decision Tree, Random Forest, and other classifiers to detect fraud. Models are trained on both original and balanced datasets to compare performance. Experimental results show that applying SMOTE significantly improves detection accuracy and reduces bias toward the majority class. Among the models, Random Forest with SMOTE achieves the highest accuracy, exceeding 99%, demonstrating its effectiveness in handling imbalanced data. The system also uses confusion matrix and graphical visualizations to evaluate model performance and highlight correct and incorrect predictions. A user-friendly interface allows users to load datasets, balance data, train models, and test new transaction data for fraud detection. The system predicts whether a transaction is fraudulent or normal in real time. Overall, this project demonstrates the importance of handling imbalanced data and the effectiveness of machine learning techniques in fraud detection.

Keywords: Fraud Detection, Machine Learning, SMOTE, Random Forest, Imbalanced Dataset, Classification, Credit Card Fraud, Predictive Analytics.

I.INTRODUCTION

With the rapid growth of digital payment systems and online transactions, credit card fraud has become a major challenge for financial institutions and customers. Fraudulent

activities not only result in financial losses but also reduce trust in digital banking systems. Detecting fraud is a complex task because fraudulent transactions are rare compared to legitimate ones, making the dataset highly imbalanced. Traditional rule-based systems often fail to detect new and evolving fraud patterns, leading to the need for more intelligent and adaptive solutions. This project focuses on applying machine learning techniques to analyze transaction data and predict fraudulent activities effectively.

Machine learning provides powerful tools for detecting patterns and anomalies in large datasets. In this project, classification algorithms such as Decision Tree and Random Forest are used to distinguish between fraudulent and non-fraudulent transactions. However, due to the imbalance in the dataset, models may become biased toward the majority class (non-fraud). To overcome this issue, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples of the minority class, thereby balancing the dataset. This improves the model's ability to detect fraud cases accurately. Additionally, performance metrics such as accuracy, confusion matrix, and graphical analysis are used to evaluate the effectiveness of the models.

The system is designed with multiple modules, including dataset loading, data balancing, model training, and fraud detection. Users can upload datasets, apply SMOTE, train machine learning models, and test new transaction data through a user-friendly interface. The results show that balancing the dataset significantly improves prediction accuracy, with Random Forest achieving the best performance. Despite its effectiveness, challenges such as evolving fraud patterns and real-time detection remain. This project highlights the importance of machine learning and data balancing techniques in building reliable fraud detection systems.

II SURVEY OF RESEARCH

The study by J. Han, M. Kamber, and J. Pei (2011) [1] introduced data mining techniques for extracting meaningful patterns from large datasets. The methodology includes classification, clustering, and anomaly detection methods. Results showed that data mining can effectively identify hidden patterns in financial transactions. However, handling highly imbalanced datasets remains a challenge. This research provides a foundation for applying data mining techniques in fraud detection systems.

The work by N. V. Chawla et al. (2002) [2] proposed the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance problems. The methodology involves generating synthetic samples for the minority class to balance the dataset. Results

demonstrated significant improvement in classification accuracy for imbalanced datasets. However, SMOTE may introduce noise if not applied carefully. This study is directly relevant to improving fraud detection performance in the proposed system.

The study by L. Breiman (2001) [3] introduced the Random Forest algorithm, an ensemble learning method that combines multiple decision trees. The methodology improves prediction accuracy and reduces overfitting. Results showed that Random Forest performs well on complex datasets and provides high accuracy. However, it requires more computational resources. This research supports the use of Random Forest as the primary model in the system.

The research by T. M. Mitchell (1997) [4] focused on machine learning techniques for predictive modeling. The methodology includes supervised learning algorithms for classification tasks. Results demonstrated the effectiveness of machine learning in detecting patterns and making predictions. However, model performance depends on feature selection and data quality. This study provides the theoretical basis for applying machine learning in fraud detection.

The study by C. Cortes and V. Vapnik (1995) [5] introduced Support Vector Machines (SVM) for classification problems. The methodology uses hyperplanes to separate data into different classes. Results showed high accuracy in classification tasks, especially with high-dimensional data. However, parameter tuning is required for optimal performance. This research highlights alternative models for fraud detection.

The work by V. Mnih et al. (2015) [6] explored deep learning techniques for pattern recognition. The methodology uses neural networks to learn complex data representations. Results showed improved performance in classification and anomaly detection tasks. However, deep learning requires large datasets and high computational power. This study suggests future improvements for fraud detection systems using advanced AI techniques.

III. WORKING METHODOLOGY

The proposed system follows a structured methodology to detect fraudulent transactions using machine learning techniques and data balancing methods. Initially, the process begins with dataset loading and preprocessing. The dataset contains approximately 28,000 transaction records with two class labels: fraud and non-fraud. Since the dataset is highly imbalanced, with a significantly larger number of non-fraud cases, preprocessing is essential. This step includes handling missing values, removing duplicates, and normalizing the data to ensure consistency. Feature selection is also performed to identify the most relevant

attributes for fraud detection. This stage ensures that the data is clean and suitable for training machine learning models, which is crucial for improving prediction accuracy.

In the next phase, the system applies the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance problem. SMOTE generates synthetic samples for the minority class (fraud transactions) to balance the dataset. As a result, the number of training samples increases, and both classes have an equal representation. This balanced dataset is then used to train machine learning models such as Decision Tree and Random Forest. The models are trained on both the original and balanced datasets to compare performance. Evaluation metrics such as accuracy, precision, recall, and confusion matrix are used to measure the effectiveness of each model. Random Forest with SMOTE typically achieves the highest accuracy due to its ability to handle complex patterns and reduce overfitting.

Finally, the trained model is deployed for real-time fraud detection. Users can upload test data, and the system predicts whether each transaction is fraudulent or normal. The results are displayed in a structured format, allowing users to easily interpret predictions. Visualization tools such as graphs and confusion matrices provide insights into model performance and prediction results. Although the system performs effectively, challenges such as evolving fraud patterns and real-time scalability remain. Future improvements can include advanced deep learning models and continuous learning mechanisms to enhance detection accuracy and adaptability in dynamic financial environments.

IV RESULTS EXPLANATIONS

In propose work we are utilizing ML algorithm to detect fraud payments from online transactions. To train ML algorithms we have used same fraud dataset which is given by you and this dataset contains only 1% records as fraud and remaining records are non-fraud which is making dataset highly imbalanced.

In propose work we have used Naïve Bayes and Random Forest algorithms for fraud detection on original and Smote based oversampling data. Smote algorithm will generate synthetic records for under-sampling classes to make dataset balanced.

After balancing dataset both algorithms got little improvement in accuracy. Each algorithm performance is evaluated in terms of accuracy, precision, recall, FSCORE and confusion matrix classification graph. Among all algorithms Random Forest with SMOTE giving high accuracy.

To implement this project we have designed following modules

1) New User Sign up: using this module user can sign up with the application

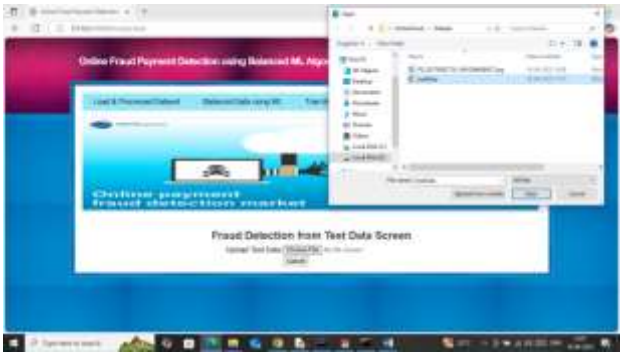
2) User Login: user can login to system
3) Load & Processed Dataset: after login user can run this module to load and normalized dataset using Standard Scaling and then convert all non-numeric data to numeric values using Label Encoder class. This module will split dataset into train and test where application using 80% data for training and 20% for testing and then will plot fraud and non-fraud transaction to show data imbalance
4) Balanced Data using ML: all processed training data will be input to SMOTE algorithm balanced dataset and then plot graph of balanced dataset class labels
5) Train ML Models: 80% training data will be input to both ML models with and without smote training data and then apply trained model on 20% test data to calculate prediction accuracy
6) Detect Fraud: using this module user can upload test data and then best ML model will be applied to predict weather test data contains normal or fraud payments

In above screen dataset loaded and can see dataset contains nearly 28000 records and can see available class labels as 'Fraud and Non-Fraud'. In graph can see dataset contains more number of 'Non-Fraud' and less number of fraud payments which make dataset highly imbalance. Now click on 'Balanced Data using ML' link to balanced dataset and then will get below page

In above screen can see training records before smote is 22570 and after applying smote data size increased to 31886 and in graph can see both class labels have equal number of records. Now click on 'Train ML Models' link to train ML algorithms with and without smote and then calculate prediction accuracy

In above screen in table format can see ML algorithm accuracy with and without smote and in above screen can see Random Forest with

smote got more than 99% accuracy. In confusion matrix graph x-axis represents 'Predicted Labels' and y-axis represents True Labels and then yellow and light green boxes in diagonal represents correct prediction count. Both blue boxes got incorrect prediction count which are very few. In bar graph showing comparison between all algorithms where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In all algorithms Random Forest with smote got high accuracy. Now click on 'Detect Fraud' link to get below page



In above screen selecting and uploading 'testData.csv' file and then click on 'Open and submit' button to load test data and then will get below prediction output

In above screen in first column can see 'Test Data Values' and then in second column can see predicted payments as 'Fraud or Normal'.

V.CONCLUSION

The proposed system for credit card fraud detection using machine learning and SMOTE demonstrates an effective approach to handling imbalanced datasets and improving prediction accuracy. By analyzing transaction data and applying preprocessing techniques, the system prepares the dataset for efficient model training. The use of the Synthetic Minority Over-sampling Technique (SMOTE) successfully balances the dataset by generating synthetic samples for the minority class, enabling the models to learn fraud patterns more effectively. Machine learning algorithms such as Decision Tree and Random Forest are applied, with Random Forest showing superior performance due to its ensemble learning capability and robustness against overfitting.

The experimental results indicate that models trained on the balanced dataset significantly outperform those trained on the original imbalanced dataset. Random Forest with SMOTE achieves an accuracy of over 99%, with very few false predictions as observed in the confusion matrix. The system also provides graphical visualizations that help in understanding model performance and class distribution. This demonstrates the importance

of data balancing techniques in improving the reliability of fraud detection systems.

Despite its effectiveness, the system has certain limitations. The performance depends on the quality and diversity of the dataset, and there is a possibility of overfitting due to synthetic data generation. Additionally, real-time fraud detection and adaptability to new fraud patterns remain challenges. Future work can focus on integrating deep learning models, real-time data processing, and continuous learning mechanisms. Overall, this project highlights the potential of machine learning in building accurate and efficient fraud detection systems.

RE.FERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[2] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[6] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[9] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2010.

[10] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*. Cambridge Univ. Press, 2014.

[11] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2020.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[13] G. James et al., *An Introduction to Statistical Learning*. Springer, 2013.

[14] R. Kohavi, "A study of cross-validation and bootstrap," in *Proc. IJCAI*, 1995.

[15] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[16] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines*. Cambridge Univ. Press, 2000.