Fraud Detection in Credit Card Transactions Using Logistic Regression

Samir Singh(22051013) , Pratik Dash(22053451), Shashank Pratyush(22051792), Ashutosh Singh(22051760)

CASE STUDY DONE UNDER VIJAY KUMAR MEENA, KIIT UNIVERSITY

***Abstract*-This study aims to develop a robust model for detecting fraudulent credit card transactions using a Logistic Regression model. The highly imbalanced nature of the dataset presents a challenge, which is addressed by employing techniques such as SMOTE for balancing the classes. The model is evaluated using performance metrics like precision, recall, F1-score, and AUC-ROC. The results indicate that the proposed model can effectively distinguish between fraudulent and non-fraudulent transactions, with the chosen threshold maximizing the F1-score.**

***Keywords-* Fraud Detection, Logistic Regression, SMOTE, Imbalanced Dataset, AUC-ROC.**

I. INTRODUCTION

Fraudulent activities in credit card transactions pose significant risks to financial institutions and customers alike. With the increasing volume of online transactions, detecting fraudulent transactions has become a critical area of research in financial security. The need for accurate and efficient fraud detection systems is paramount as fraudsters continuously evolve their tactics. This study aims to enhance the detection of fraudulent transactions by applying a Logistic Regression model. The primary objective is to implement and evaluate the performance of the Logistic Regression model to accurately identify fraudulent transactions in a highly imbalanced dataset.

II. LITERATURE REVIEW

Fraud detection in credit card transactions has been an area of extensive research due to the growing threat it poses to financial institutions and consumers. Traditional supervised learning methods, such as decision trees and logistic regression, have been commonly used to build classification models for fraud detection. However, these methods often struggle with the highly imbalanced nature of fraud datasets, where fraudulent transactions are vastly outnumbered by legitimate ones. Various studies have suggested that without proper handling, such imbalances can lead to models that are biased towards the majority class, resulting in poor detection of fraud cases [1].

To address the imbalance, techniques like Synthetic Minority Over-sampling Technique (SMOTE) have been introduced. SMOTE works by generating synthetic samples for the minority class, thus balancing the class distribution and improving the model's ability to detect fraudulent transactions [2]. While SMOTE has proven effective, it is not without limitations. For instance, generating synthetic data can lead to overfitting, where the model becomes too finely tuned to the training data and fails to generalize well to unseen data [3].

Anomaly detection methods, such as Isolation Forest and Local Outlier Factor (LOF), have also been explored extensively for fraud detection. These unsupervised techniques are particularly useful when labeled data is scarce, and the primary task is to identify outliers that are likely fraudulent. Isolation Forest isolates anomalies by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature [4]. LOF, on the other hand, identifies outliers by comparing the local density of a data point to that of its neighbors, assigning higher anomaly scores to points that have a significantly lower density than their neighbors [5].

More recent studies have started to combine both supervised and unsupervised techniques in hybrid

models, seeking to leverage the strengths of each approach. These hybrid models often use unsupervised methods to identify potential anomalies, which are then further analyzed using supervised classification techniques. This two-step process can enhance the accuracy of fraud detection systems, particularly in real-time scenarios where speed and accuracy are critical [6]. The combination of SMOTE with anomaly detection methods like Isolation Forest and LOF has also been proposed as a way to improve model performance on imbalanced datasets by ensuring that the model is both sensitive to minority classes and robust in detecting outliers [7].

III. DATA AND METHODS

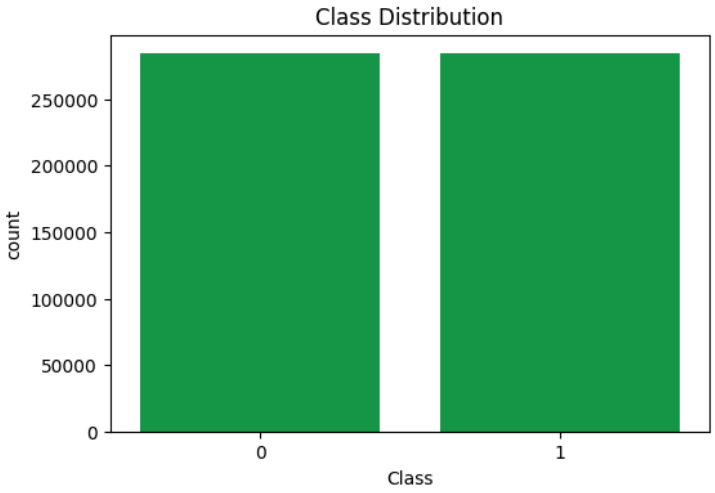
The **Credit Card Fraud Detection Dataset 2023 [8]** includes anonymized transaction details of credit card usage, with features representing various transactional behaviors. The dataset contains a significant class imbalance, with the majority of transactions being legitimate. The data was extracted, transformed, and loaded into a data warehouse using an ETL process, which included handling missing values, normalizing the data, and addressing the class imbalance using techniques like SMOTE.

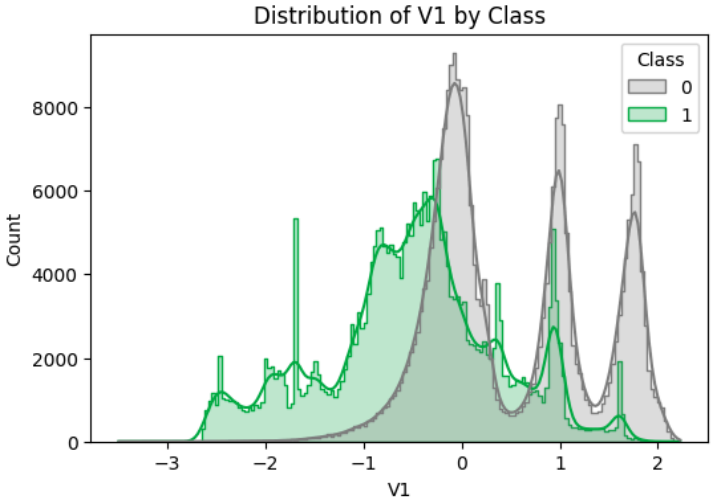
A star schema was used for organizing the data, with a central fact table containing transaction records linked to dimension tables representing various attributes of the transactions. A Logistic Regression model was implemented to predict the likelihood of a transaction being fraudulent. The model was trained and evaluated using metrics such as precision, recall, F1-score, and AUC-ROC to measure its effectiveness in detecting fraud. Additionally, the classification threshold was adjusted to maximize the F1-score, ensuring a balance between precision and recall.

1. Analysis of Dataset:

An initial analysis was conducted to understand the distribution of the dataset, including class distribution, correlations between features, and feature distributions. The following images provide insight into these aspects:

1. Class Distribution: Visual representation of the imbalance in the dataset.
2. Feature Distributions: Depicts how various features are distributed across fraudulent and non-fraudulent transactions for the column V1 as an example.





IV. RESULTS

The Logistic Regression model achieved a precision of **97%**, recall of **97%**, and an AUC-ROC score of **99.35%**, indicating its capability to detect fraudulent transactions effectively. The application of SMOTE significantly improved the recall rate, ensuring that

more fraudulent transactions were detected. This improvement was critical in reducing false negatives, which are particularly costly in fraud detection. A threshold of **0.45** was identified as optimal, balancing precision, recall, and F1-score.

2. Classification Report and ROC AUC Score:

The classification report provides a detailed breakdown of the model's performance across various metrics, including precision, recall, and F1-score. Additionally, the ROC AUC score reflects the model's overall ability to discriminate between the positive and negative classes. This is illustrated in the following table:

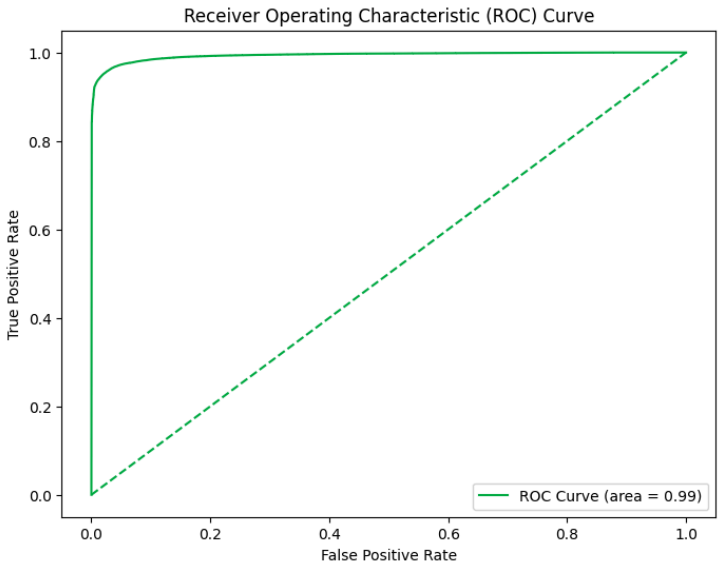
Classification Report

Classification	Precision	Recall	F1-Score	Support
0	0.95	0.98	0.97	85149
1	0.98	0.95	0.97	85440
accuracy	null	null	0.97	170589
macro avg	0.97	0.97	0.97	170589
weighted avg	0.97	0.97	0.97	170589

ROC AUC Score: 0.9934683620303304

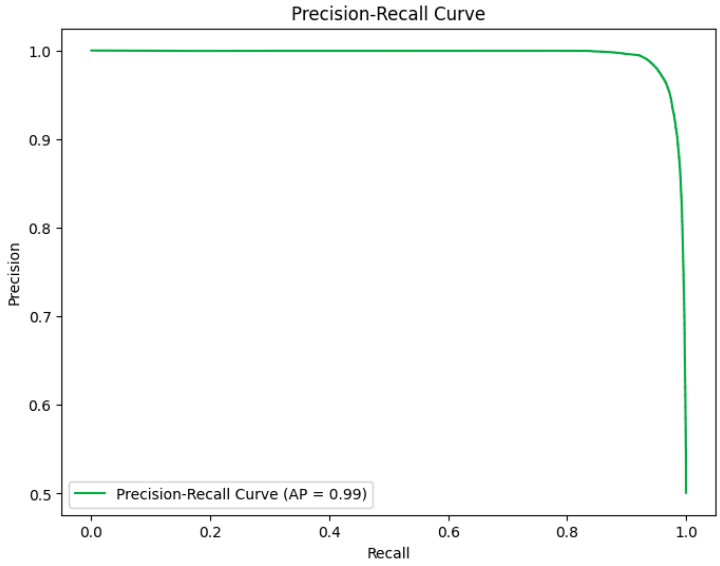
3. Receiver Operating Characteristic (ROC) Curve:

The ROC Curve is a graphical representation of the true positive rate against the false positive rate. It is used to visualize the trade-off between sensitivity (recall) and specificity. The ROC curve for the Logistic Regression model is shown below -



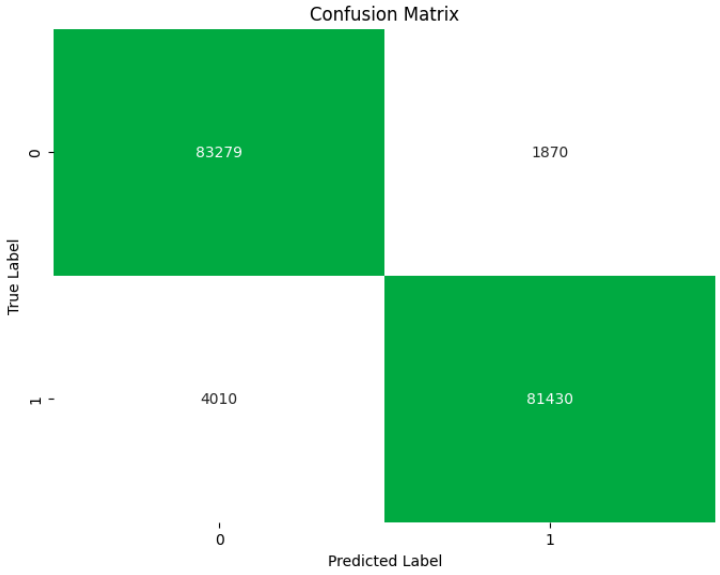
4. Precision-Recall Curve:

The Precision-Recall Curve is particularly useful in evaluating the performance of the model on imbalanced datasets, where the minority class is more important. The curve for the Logistic Regression model is depicted below:



5. Confusion Matrix:

The confusion matrix provides a visual representation of the true positives, false positives, true negatives, and false negatives, helping to understand the model's performance in classification tasks. The confusion matrix for the test dataset is shown below:



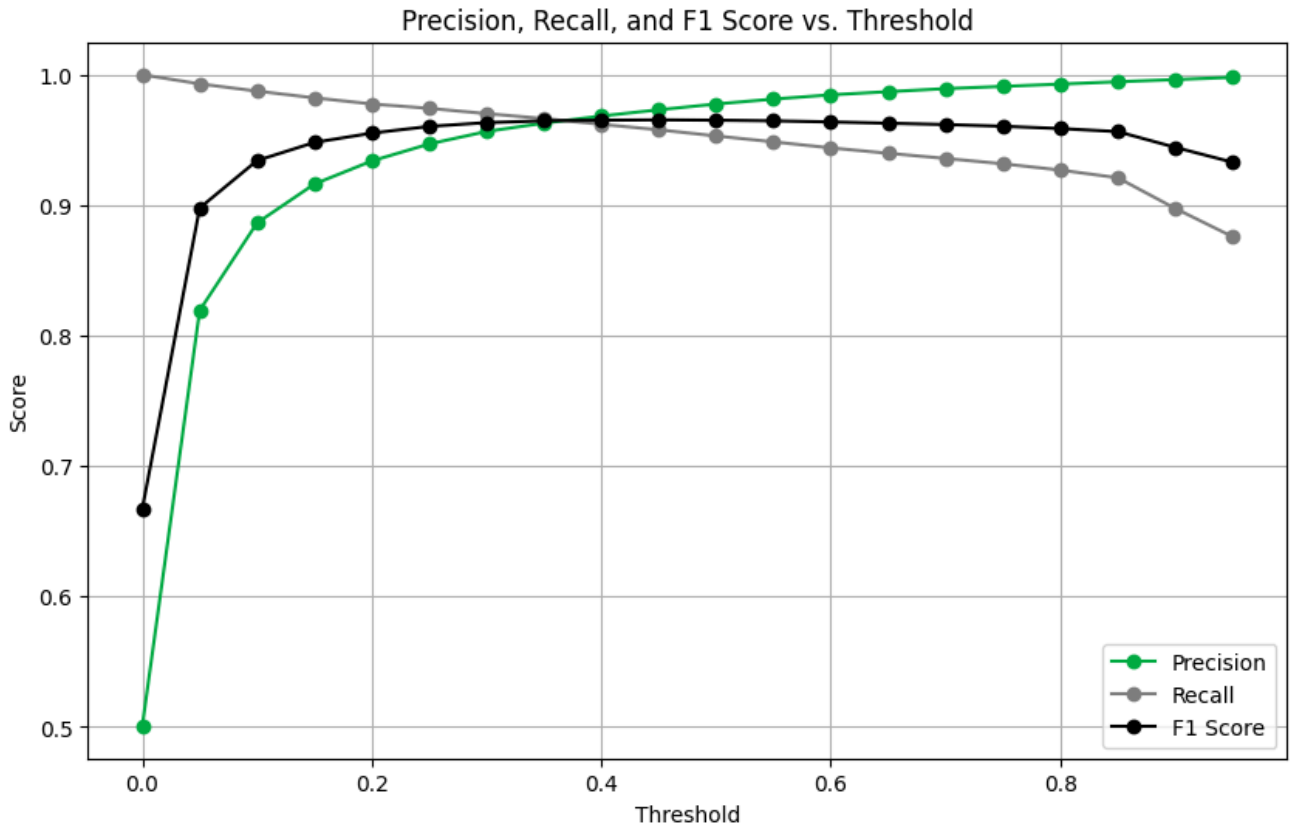
The source code used for the analysis can be accessed and reviewed in the GitHub Repository. **[9]**

V. DISCUSSION

The results suggest that the Logistic Regression model, when combined with SMOTE, is effective in identifying fraudulent transactions in imbalanced datasets. The use of SMOTE was crucial in enhancing the model's performance, especially in improving recall without significantly sacrificing precision. These findings align with previous research that highlights the effectiveness of logistic regression in fraud detection. However, the application of this method in a real-time context may require further optimization. The study's limitations include the potential overfitting due to the synthetic nature of the data generated by SMOTE and the fixed threshold, which may not generalize well across different datasets or in a real-time environment.

6. Precision, Recall, and F1 Score vs. Threshold:

In the evaluation of the Logistic Regression model, varying the decision threshold allows for optimizing the balance between precision, recall, and F1-score. The following plot demonstrates how these metrics change as the threshold is adjusted:



VI. CONCLUSION

This study successfully implemented and evaluated a Logistic Regression model for fraud detection in credit card transactions, highlighting the importance of handling imbalanced datasets. The findings have significant implications for the development of real-time fraud detection systems, emphasizing the need for continuous model refinement to combat evolving fraud techniques. Future work could explore hybrid models combining supervised and unsupervised techniques, as well as the integration of real-time data streams for immediate fraud detection. The exploration of different classification thresholds based on business needs is also recommended.

VII. REFERENCES

1. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review, 34(3), 243-279.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
3. Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science, 3644, 878-887.
4. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest. Proceedings of the 8th IEEE International Conference on Data Mining, 413-422.
5. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. ACM SIGMOD Record, 29(2), 93-104.
6. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications, 38(5), 5947-5957.
7. Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5), 363-387.
8. Kaggle. (2023). Credit Card Fraud Detection Dataset 2023
9. https://github.com/samirdead66/dmdw