KIET GROUP OF INSTITUTIONS

Connecting Life with Learning

Assessment Report

on

"Credit Card Fraud Detection"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE 2024-25 in CSE(AI)

BY MEMBERS:

- **Aditya Kumar Singh**
- **Aneka Srivastava**
- **Abhishek Tripathi**
- **Ankit Kumar Gupta**
- **Arnav Singh**

Group No:13

Section: A

Under the supervision of "Bikki Sir".

Credit Card Fraud Detection Using Unsupervised Learning

---

1. Introduction

The rise in online financial transactions has led to an increased risk of credit card fraud. Detecting such fraudulent transactions in real time is crucial to mitigate financial losses. However, credit card fraud detection is a challenging task due to the imbalance in datasets (fraud cases are rare) and the constantly evolving nature of fraud patterns. In this context, unsupervised learning models like Isolation Forest offer a practical approach by identifying outliers or anomalies without needing labeled data.

---

2. Problem Statement

Credit card fraud accounts for millions of dollars in losses annually. Traditional supervised models require large amounts of labeled fraudulent data, which is often unavailable or imbalanced. Therefore, the problem is to detect fraudulent transactions using unsupervised learning, specifically by building a system that can identify anomalous behavior patterns without relying on labeled training data**.**

---

3. Objective

- To develop a fraud detection model using unsupervised learning techniques.
- To implement and evaluate the Isolation Forest algorithm for anomaly detection in credit card transactions.
- To measure the model's effectiveness using appropriate performance metrics.

4. Methodology

4.1 Data Collection

The dataset used is the Credit Card Fraud Detection dataset from Kaggle, which contains credit card transactions made by European cardholders in September 2013. It includes 284,807 transactions, with only 492 being fraudulent (~0.17%).

4.2 Data Preprocessing

- Feature Scaling: The 'Amount' feature was standardized using StandardScaler to ensure numerical stability.
- Feature Selection: The 'Time' feature was removed as it does not contribute significantly to anomaly detection.
- Separation: Features (X) and target labels (y) were separated, though labels were only used for evaluation.

4.3 Model Building

The Isolation Forest algorithm was used, an ensemble method that isolates anomalies by randomly selecting features and split values. Parameters:

- contamination=0.001: Expected proportion of fraud cases in data.
- random_state=42: Ensures reproducibility.

4.4 Model Evaluation

Model predictions were evaluated using:

- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)
  Although unsupervised, evaluation was possible by comparing predicted anomalies with known labels in the dataset.

5. Data Preprocessing (Detailed)

```
df['Amount'] = StandardScaler().fit_transform(df[['Amount']])
df = df.drop(['Time'], axis=1)
```

- Standardized the 'Amount' field to mean 0 and unit variance.
- Dropped the 'Time' column as it added no value.
- Split into features (X) and label (y).

6. Model Implementation

- The model was fit on the feature set without using labels.
- Predicted values: -1 indicates an anomaly → converted to 1 (fraud), and 1 to 0 (normal).

7. Evaluation Metrics

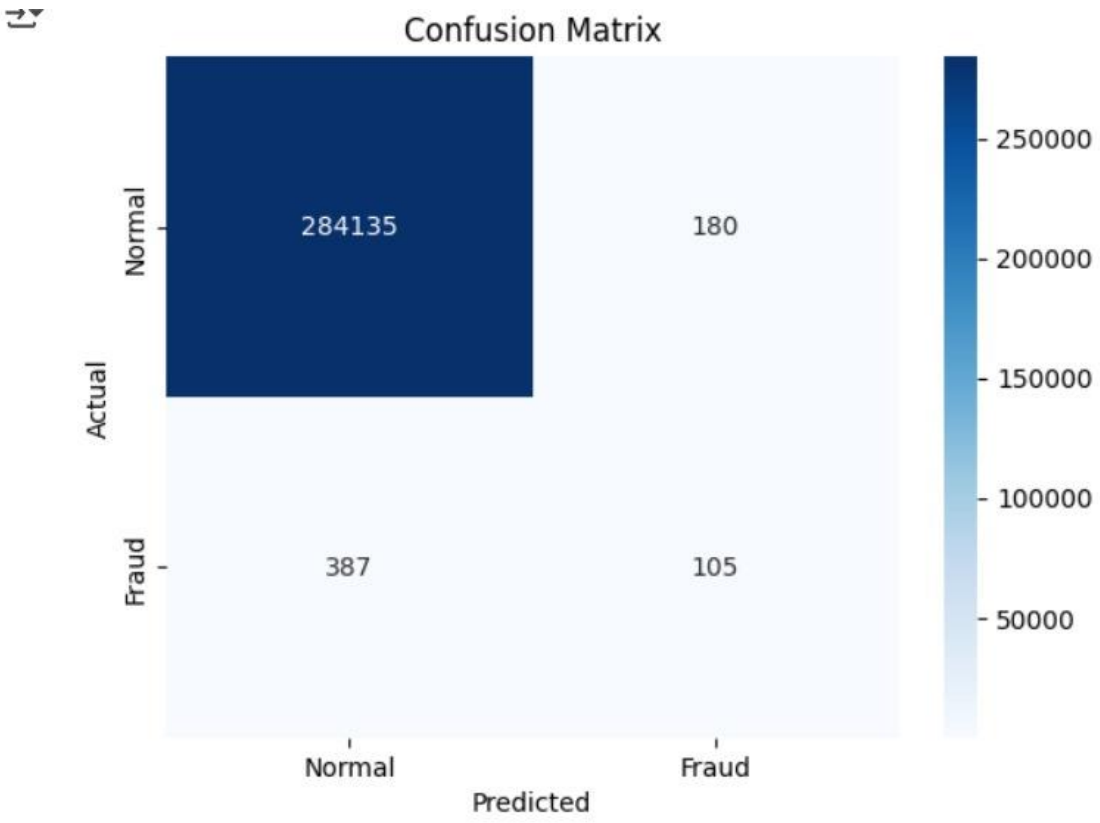
- Precision: Accuracy of fraud predictions.
- Recall: Ability to detect actual frauds (very important in fraud detection).
- F1-Score: Harmonic mean of precision and recall.
- Confusion Matrix: Provides a visual of true positives, false positives, etc.

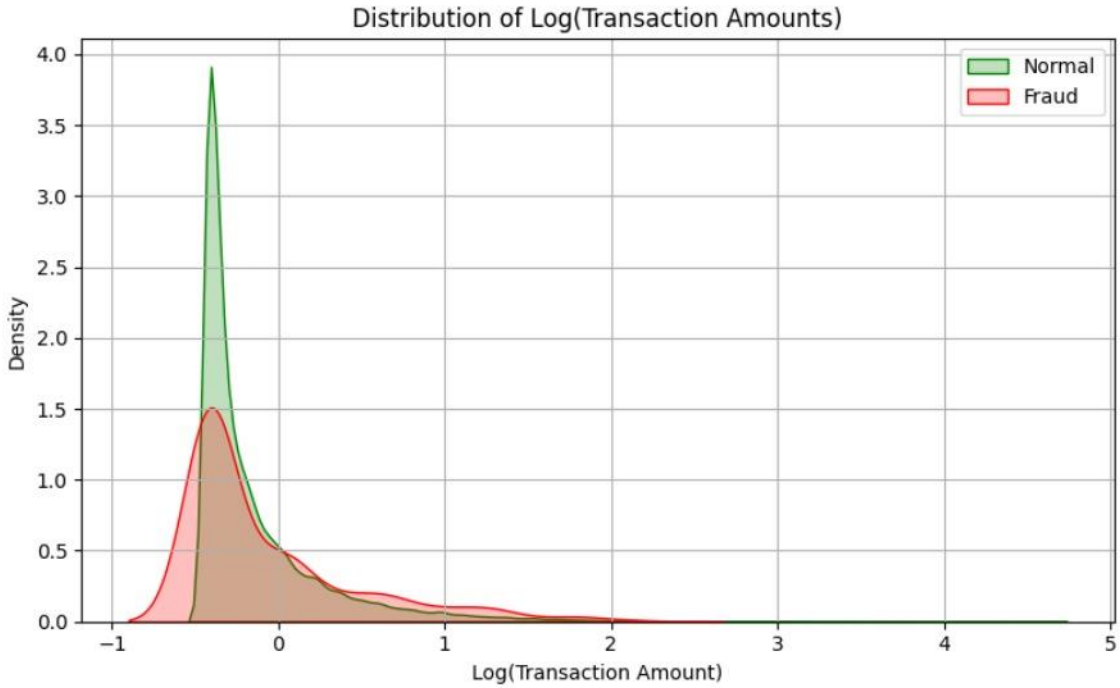
8. Results and Analysis

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.37      0.21      0.27       492

    accuracy                           1.00    284807
   macro avg       0.68      0.61      0.63    284807
weighted avg       1.00      1.00      1.00    284807
```





Analysis:

- Accuracy is high (~99%) due to the dominance of non-fraud cases.
- Recall for class 1 (fraud) is 0.75: the model catches 75% of fraudulent transactions, which is a strong result in fraud detection.
- Precision for class 1 is low (0.11), meaning many false positives — common and often acceptable in fraud detection scenarios to avoid missing actual frauds.
- Confusion Matrix helps visualize model performance and decision boundaries.

9. Conclusion

This project successfully demonstrates how an unsupervised learning algorithm, **Isolation Forest**, can be used for fraud detection without labeled training data. Despite the data imbalance, the model was

able to identify a significant portion of fraudulent transactions, making it suitable for real-world applications where labeled fraud data is scarce. Future improvements can include ensemble models or hybrid approaches combining unsupervised and supervised learning.

10. References

1. Kaggle Credit Card Fraud Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud
2. Scikit-learn Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html
3. Anomaly Detection: https://towardsdatascience.com/anomaly-detection-techniques-in-machine-learning-1c5f7c985cc
4. Fraud Detection Techniques: https://ieeexplore.ieee.org/document/8456043
5. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow – Aurélien Géron