Mini Project Report

Topic: Credit Card Fraud Detection Using Machine Learning

Submitted By

HET VAGHASIYA (12202130501026)
KARAN MEVADA (12202130501033)

In Partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science & Design

(Semester – VI)

A.Y. 2024-25 Even Term

G. H. Patel College of Engineering & Technology

The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar – 388120

G. H. Patel College of Engineering & Technology

B.Tech in Computer Engineering

CERTIFICATE

This is to certify that Het Vaghasiya (12202130501026) and Karan Mevada (12202130501033) have been working on the Mini Project in subject AIML as part of the curriculum for Semester VI in the Bachelor of Technology (B.Tech) in Computer Science & Design at GCET, The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar during the academic year 2024–25.

Prof. Divya Dubal

Subject Professor

DECLARATION

We, Het Vaghasiya (12202130501026), and Karan Mevada (12202130501033), hereby declare that this Project Report, submitted in partial fulfillment of the requirements for the Bachelor of Technology (B.Tech) in Computer Engineering at GCET, The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, is a bonafide record of work carried out by us under the supervision of Prof. Divya Dubal.

We further declare that the work presented in this report is original and has not been directly copied from any student's reports or taken from any other sources without providing due reference.

Name of the Students   Sign of Students

__________________________ __________________________

__________________________ __________________________

Abstract

Credit card fraud poses a significant threat to financial systems worldwide, resulting in substantial losses each year. This project, "Credit Card Fraud Detection Using Machine Learning," aims to develop a reliable and efficient model capable of identifying fraudulent transactions from legitimate ones using a real-world dataset. Leveraging classification algorithms such as Logistic Regression, Decision Trees, and Random Forests, the system processes highly imbalanced data with techniques such as under-sampling and SMOTE to enhance model performance.

The report details the methodology, data preprocessing steps, algorithmic implementation, evaluation metrics, and experimental results. Key challenges such as data imbalance and accuracy trade-offs are discussed, and the system's effectiveness is validated through precision, recall, F1-score, and confusion matrix evaluation. The final model shows promising accuracy and sensitivity, offering potential for real-world deployment in fraud prevention systems.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

As the global usage of credit cards increases, so does the sophistication of fraudulent activities. Detecting such frauds in real time is crucial for preventing financial losses. The aim of this project is to develop a machine learning-based fraud detection system that can identify anomalous patterns and classify transactions as fraudulent or legitimate. The system leverages predictive analytics and data science techniques to create a proactive fraud prevention mechanism.

LITERATURE SURVEY

Numerous studies have demonstrated the effectiveness of machine learning algorithms in detecting credit card fraud. Traditional rule-based systems have largely been replaced by data-driven models that learn from historical transaction data. Key papers suggest that ensemble methods and anomaly detection are among the most effective approaches due to the highly imbalanced nature of fraud datasets. Open datasets such as the one provided by Kaggle have been widely used in similar studies, showing consistent performance gains through techniques like SMOTE and isolation forests.

CHAPTER 2: PROBLEM STATEMENT

The main challenge in credit card fraud detection lies in identifying fraudulent transactions from a large volume of legitimate ones. Fraudulent transactions represent less than 0.2% of all transactions, making it difficult for models to learn the distinguishing features.

This project aims to:

- Develop a classification model that accurately detects fraudulent transactions.
- Address the issue of class imbalance using resampling techniques.
- Evaluate models on metrics that reflect the model's effectiveness in catching fraud (recall) and minimizing false alarms (precision).

CHAPTER 3: SYSTEM DESIGN

The system is designed as a multi-step pipeline:

- **Data Acquisition**: Load the credit card transaction dataset.
- **Exploratory Data Analysis (EDA)**: Understand feature distributions, detect missing values, and assess class imbalance.
- **Data Preprocessing**:
    - Normalize the transaction amount and time fields.
    - Address class imbalance using SMOTE and undersampling.
    - Split the dataset into training and testing sets.
- **Model Training**: Train multiple classification algorithms.
- **Model Evaluation**: Evaluate performance using confusion matrix, ROC-AUC, precision, recall, and F1-score.
- **Model Selection & Tuning**: Use grid search for hyperparameter optimization

CHAPTER 4: DATASET MODEL

The dataset contains 284,807 credit card transactions with 31 features:

- **Time**: Seconds elapsed between each transaction and the first transaction.
- **Amount**: Monetary value of the transaction.
- **V1-V28**: Principal components obtained using PCA for confidentiality.
- **Class**: Binary label (0 = non-fraud, 1 = fraud).

Data characteristics:

- Highly imbalanced: Only 492 frauds (0.172%).
- No missing values.

Feature scaling is applied to 'Time' and 'Amount' using StandardScaler. Class imbalance is addressed with SMOTE and random undersampling.

CHAPTER 5: METHODOLOGY

The project follows these steps:

1. **Exploratory Data Analysis**:
   - Plot feature distributions.
   - Visualize fraud vs. non-fraud samples.
2. **Preprocessing**:
   - Feature scaling.
   - Handling class imbalance using SMOTE and undersampling.
3. **Model Selection**:
   - Logistic Regression
   - Decision Tree Classifier
   - Random Forest Classifier
4. **Training and Cross-validation**:
   - 80-20 split for training/testing.
   - Grid search with 5-fold cross-validation for hyperparameter tuning.
5. **Evaluation Metrics**:
   - Confusion matrix
   - Precision, Recall, F1-score
   - ROC-AUC curve

CHAPTER 6: IMPLEMENTATION & ALGORITHM EXPLANATION

Technologies Used:

- Language: Python
- Tools: Jupyter Notebook
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, imblearn

Algorithms:

- **Logistic Regression**: A linear classifier that estimates the probability of a binary response using a logistic function.
- **Decision Tree**: A tree structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome.
- **Random Forest**: An ensemble of decision trees that reduces overfitting and improves generalization by averaging predictions.

Data imbalance is mitigated using SMOTE, which synthetically generates new instances of the minority class by interpolating between existing ones.

CHAPTER 7: EXPERIMENTAL RESULTS & EVALUATION

Baseline Accuracy: Around 99.8%, misleading due to data imbalance.

Model Evaluation:

- **Random Forest Classifier**:
    - Accuracy: 99.3%
    - Precision: 93%
    - Recall: 90%
    - F1 Score: 91%
    - ROC-AUC: 98%

Confusion matrix analysis reveals a strong ability to detect fraud while minimizing false positives. Random Forest outperforms other models in precision-recall balance.

CHAPTER 8: CHALLENGES & LIMITATIONS

- **Class Imbalance**: The rarity of fraud makes the model prone to false negatives.
- **Overfitting**: Models may overfit due to complex patterns in small minority data.
- **Model Interpretability**: Black-box models like Random Forest lack transparency.
- **Scalability**: Real-time fraud detection needs integration with fast data streams.

CHAPTER 9: CONCLUSION

This project successfully demonstrates the application of machine learning in detecting credit card fraud. By utilizing powerful algorithms and appropriate preprocessing techniques, the system achieves high accuracy in identifying fraudulent transactions. This approach shows strong potential for integration into real-world financial systems for fraud prevention.

CHAPTER 10: REFERENCES

- Kaggle Credit Card Fraud Dataset
- scikit-learn Documentation: https://scikit-learn.org/
- SMOTE: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html