**< Fraud Detection in Online Transactions>**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**(E23CSEU1595) BHAVYA SODHI**

**(E23CSEU1591) NIKSHAY RATHAUR**

Submitted to

**DR. ASHIMA YADAV**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr.No | Content | Page No |
| 1 | **Abstract** |  |
| 2 | **Introduction** |  |
| 3 | **Related Work** |  |
| 4 | **Methodology** |  |
| 5 | **Hardware/Software Required** |  |
| 6 | **Experimental Results** |  |
| 7 | **Conclusions** |  |
| 8 | **Future Scope** |  |
| 9 | **GitHub Link of Your Complete Project** |  |

**Abstract**

This project investigates the detection of fraudulent transactions using two machine learning models: Decision Tree Classifier   
and Logistic Regression.

The dataset comprises millions of transactions with features such as amount, origin, and destination.

Exploratory Data Analysis (EDA) and preprocessing steps were performed to clean and prepare the data. Both models were trained   
and evaluated using appropriate metrics.

This report details the methodologies, results, and future scope of the project.

**Introduction**

Fraud detection is a critical challenge in the domain of financial services. With the growing volume of online transactions,   
traditional rule-based systems are becoming insufficient to detect sophisticated fraud patterns.

Machine learning offers an intelligent, automated approach to identify fraudulent activities.

This project implements Logistic Regression and Decision   
Tree Classifier models to classify transactions as fraudulent or legitimate. The report describes the dataset, data preparation,   
model building, evaluation metrics, and results.

In this project, the focus is on utilizing supervised machine learning models, specifically Logistic Regression and Decision Tree Classifier, to classify transactions. These models are selected due to their simplicity, effectiveness, and interpretability in solving classification problems. The dataset used includes millions of financial transactions, with features such as transaction type, amount, sender and receiver details, and fraud indicators.

Furthermore, this project addresses challenges such as class imbalance, feature preprocessing, and model evaluation. By employing techniques like under sampling, oversampling, and feature encoding, the models are fine-tuned to detect fraudulent transactions with high accuracy. This study aims to provide a comprehensive understanding of how machine learning can contribute to financial fraud prevention, ultimately leading to safer and more reliable transaction systems.

**Related Work**

Various machine learning models, such as Random Forest, Support Vector Machines, and Neural Networks, have been extensively   
studied for fraud detection.

However, these models often require significant computational resources. This project focuses on relatively simpler models, namely Logistic Regression and Decision Tree Classifier, to achieve high accuracy with moderate   
resource consumption.

Previous studies indicate that feature engineering and data balancing significantly impact the performance of fraud detection models.

Also, the techniques used in related work are under sampling which is reducing the size of the data set to a large extent.

I am providing some links for related work:

<https://youtu.be/e2gzNgI5DOs?feature=shared>

<https://youtu.be/HHh8qNGnc3I?feature=shared>

**Methodology**

The methodology for this project involves several key steps:

1. The dataset includes transaction details with features like amount, type, origin, destination, and fraud indicators.

2. Exploratory Data Analysis (EDA):  
 - Visualized the distribution of transaction types and fraud cases.  
 - Identified missing values and performed data imputation where necessary.  
 - Detected and handled outliers using the IQR method.

3. Data Preprocessing  
 - Encoded categorical variables using Label Encoding and One-Hot Encoding.  
 - Applied feature scaling to normalize numerical data.  
 - Balanced the dataset using:  
 - Under sampling for Logistic Regression.  
 - Oversampling for Decision Tree Classifier.

4. Model Building  
 - Logistic Regression - A linear model suitable for binary classification.  
 - Decision Tree Classifier - A non-linear model offering better interpretability and performance on imbalanced datasets.

5. Model Evaluation  
 - Used metrics like accuracy, precision, recall, and F1-score.  
 - Analyzed confusion matrices and classification reports to assess model performance.

**Hardware/Software Required**

- Hardware  
 - Processor: Intel i5 or higher  
 - RAM: Minimum 8GB  
 - Storage: 20GB free disk space

- Software  
 - Operating System: Windows/Linux/MacOS  
 - Python 3.8 or higher  
 - Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Imbalanced-learn

**Experimental Results**

Exploratory Data Analysis  
- Transaction types include CASH\_OUT, PAYMENT, CASH\_IN,

TRANSFER, DEBIT.  
 - The dataset contains 8,213 fraudulent transactions out of a total of 63 lakh  
 - Outliers were capped to ensure data consistency.  
  
Logistic Regression Results  
 - Training Accuracy: 99%  
 - Testing Accuracy: 98%  
 - Confusion Matrix indicates balanced predictions across classes.

Decision Tree Classifier Results  
 - Achieved almost 100% accuracy on oversampled data.  
 - Classification report shows perfect precision, recall, and F1-score.

**Conclusions**

This project highlights the effectiveness of machine learning models in fraud detection. Logistic Regression offers a robust baseline model, while Decision Tree Classifier excels when provided with balanced data.

The preprocessing steps, especially data balancing, were critical to improving model performance.

The project demonstrates the critical role machine learning plays in detecting fraudulent transactions, a key challenge in modern financial systems. By implementing Logistic Regression and Decision Tree Classifier models, the study highlights how supervised learning techniques can effectively identify fraudulent patterns in large datasets.

The results underline the importance of data preprocessing, including handling class imbalance and encoding categorical features, which significantly impact model performance. While Logistic Regression provided robust baseline results, the Decision Tree Classifier excelled in detecting fraud when trained on a balanced dataset, achieving near-perfect performance.

This study also emphasizes the need for continuous improvement in fraud detection methods, as fraudulent behaviours evolve over time. Incorporating real-time data analysis, advanced algorithms, and interpretability techniques can further enhance the reliability and scalability of such systems.

Ultimately, this project serves as a foundation for future research in applying machine learning for fraud prevention, paving the way for more efficient, and real-time detection systems.

**Future Scope**

1. Implement advanced ensemble models like Random Forest and Gradient Boosted Trees.

2. Integrate real-time fraud detection systems using streaming data platforms.

3. Explore model explainability techniques for better interpretability and regulatory compliance.

4. Experience with feature selection and dimensionality reduction techniques to optimize model performance.

5. The field of fraud detection offers immense opportunities for further research and development. While this project focused on Logistic Regression and Decision Tree Classifier, exploring advanced machine learning models like Random Forests, Gradient Boosted Machines, or even deep learning techniques can provide better accuracy and adaptability to complex fraud patterns. These models can also handle high-dimensional datasets more effectively.

6.Real-time fraud detection is another crucial area for future work. Integrating these models with real-time data processing frameworks, such as Apache Kafka or Spark Streaming, can enable financial institutions to identify and block fraudulent transactions instantly. Additionally, creating a pipeline for continuous learning, where models are updated as new data becomes available, will ensure that the system adapts to evolving fraud tactics.

7.Explainable AI (XAI) can also play a vital role in the future. Implementing model interpretability techniques will help stakeholders understand the decision-making process, which is particularly important for regulatory compliance and building trust with users.

8.Finally, the inclusion of external data sources, such as user behaviour analytics or geographic data, can enhance the models' ability to detect fraud. Collaborative efforts between financial institutions to share anonymized fraud data may also lead to more robust detection systems, benefiting the entire industry.

GitHub link of Project: <https://github.com/2bhavyasodhi7/Transaction_fraud_detection_using_ML>