Credit Cards Fraud Detection Using Logistic Regression

[1]Mr.Irfan Bhagawan
Department of Cyber Security &Data Science
MLR Institute of Technology
Hyderabad, India
irfanbagwan4898@mlrit.ac.in

[2] Bhanusri Vattimalla
Department of Cyber Security &Data Science
MLR Institute of Technology
Hyderabad, India
banusrivattimalla@gmail.com

[3] K.V.G Arjun Prasad
Department of Cyber Security &Data Science
MLR Institute of Technology
Hyderabad, India
karjunprasad118@gmail.com

[4] V.Lukman
Department of Cyber Security &Data Science
MLR Institute of Technology
Hyderabad, India
21r21a6260@mlrinstitutions.ac.in

[5] Govindam Sudeep
Department of Cyber Security &Data Science
MLR Institute of Technology
Hyderabad, India
22r25a6201@mlrit.ac.in

ABSTRACT :-

This project utilizes logistic regression to detect credit card fraud, effectively addressing the challenge of imbalanced datasets through under sampling techniques. The solution is implemented as an interactive web application using streamlit, enabling users to perform real-time fraud analysis and predictions. The application features exploratory data analysis tools, such as data summaries, distribution visualizations, and correlation heatmaps, while providing predictions based on user-input transaction details. Model evaluation is conducted using confusion matrices and classification reports to ensure reliability. Custom CSS enhances the application's visual appeal and user experience. This project demonstrates the practical application of machine learning in financial fraud prevention, offering a user-friendly and efficient tool to mitigate risks and enhance transaction security.

Keywords: Credit Card Datasets, Streamlit, Machine Learning, Logistic Regression, Pandas, Numpy, Seaborn, Fraud Detection, Imbalanced Datasets, Undersampling, Exploratory Data Analysis (EDA), Real-time Predictions, Confusion Matrix, Classification Report, AI-driven Solutions, Financial Fraud Prevention

1. INTRODUCTION

Credit card fraud continues to be a pervasive issue, leading to significant financial losses for both consumers and financial institutions worldwide. Detecting such fraudulent activities in real-time is vital for minimizing these risks and ensuring the security of cardholders. Traditional fraud detection methods, based on rule based systems, are often slow to adapt to new frauds patterns, making it challenging to keep up with evolving tactics used by fraudsters. To address this, machine learning techniques have gained prominence as they can automatically learn and adapt to

A key machine learning approach used in this context is **logistic regression**, a statistical method ideal forbinary classification tasks. In credit cards fraud detections, logistic regression is employed to differentiate between legitimate and fraudulent transactions based on various features, such as the amount of the transaction, the time it occurred, the transaction's geographical location, and historical user behavior. By processing vast amounts of transaction data, logistic regression models can detect anomalies and predict fraud more effectively than traditional rule- based systems.

These databases typically consist of features such as transaction ID, user ID, transaction amount, time, merchant information, location, and more. This data is used to train the machine learning model, allowing it to learn the patterns of legitimate versus fraudulent activity. With the application of techniques like under sampling or oversampling, the model can handle imbalanced datasets where fraudulent transactions are much fewer than legitimate ones, improving its detection accuracy.

Once trained, the logistic regression model can be integrated into a fraud detection system, which can evaluate incoming transactions in real-time, flagging those with a high probability of fraud for further review. The system's effectiveness is evaluated using performance metrics such as accuracy, precision, recall, and the F1-score, providing valuable insights into its ability to identify fraud without generating too many false positives.
fraudulent transactions, protecting both consumers and their business operations.

2. LITERATURE SURVEY

Han, J., "Data Mining: Concepts and Techniques" (3rd Edition), Morgan Kaufmann Publishers This comprehensive textbook provides foundational knowledge on data mining techniques, including anomaly detection and classification algorithms, which are essential for developing effective fraud

detection systems. The book covers various data mining techniques, such as classification, clustering, association rule learning, and anomaly detection. It explains how classification methods (e.g., decision trees, neural networks) categorize transactions, clustering algorithms identify unusual patterns, association rule learning uncovers relationships between features. Data preprocessing steps like cleaning and transformation are also discussed, along with evaluation methods like cross-validation.

[1] Yan, X., "Credit card fraud detection using machine learning algorithms, " Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-7 This paper investigates Publishers This comprehensive textbook provides foundational knowledge on data mining techniques, including anomaly detection and classification algorithms, which are essential for developing effective fraud detection systems. The book covers various data mining techniques, such as classification, clustering, association rule learning, and anomaly detection. It explains how classification methods (e.g., decision trees, neural networks) categorize transactions, clustering algorithms identify unusual patterns, association rule learning uncovers relationships between features, and anomaly detection identifies transactions that deviate from normal patterns. Data preprocessing steps like cleaning and transformation are also discussed, along with evaluationmethods like cross-validation.

[2] Yan, X., "Credit card fraud detection using machine learning algorithms," Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-7 This paper investigates.

This project utilizes logistic regression to detect credit cards fraud, effectively addressing the challenge of imbalanced datasets through under sampling techniques. The solution is implemented as an interactive web application using Streamlit, enabling users to perform real-time fraud analysis and predictions. The application features exploratory data analysis tools, such as data summaries, distribution visualizations, and correlation heatmaps, while providing predictions based on user-input transaction details. Model evaluation is conducted using confusion matrices and classification reports to ensure reliability. Custom CSS enhances the application's visual appeal and user experience. This project demonstrates the performance metrics such as accuracy, precision, recall, and the F1-score, providing valuable insights into its

practical application of machine learning in financial fraud prevention, offering a user-friendly and efficient tool to mitigate risks and enhance transaction security.

Keywords: Credit Card Datasets, Streamlit, Machine Learning, Logistic Regression, Pandas, Numpy, Seaborn, Fraud Detection, Imbalanced Datasets, Undersampling, Exploratory Data Analysis (EDA), Real-time Predictions, Confusion Matrix, Classification Report, AI-driven Solutions, Financial Fraud Prevention.

[3] Credit cards fraud continues to be a pervasive issue, leading to significant financial losses for both consumers and financial institutions worldwide. Detecting such fraudulent activities in real-time is vitalfor minimizing these risks and ensuring the security of cardholders. Traditional fraud detection methods, basedon rule-based systems, are often slow to adapt to new fraud patterns, making it challenging to keep up with evolving tactics used by fraudsters. To address this, machine learning techniques have gained prominence as they can automatically learn and adapt to these patterns

A key machine learning approach used in this context is **logistic regression**, a statistical method ideal for binary classification tasks. In credit card fraud detection, logistic regression is employed to differentiate between legitimate and fraudulent transactions based on various features, such as the amount of the transaction, the time it occurred, the transaction's geographical location, and historical user behavior. By processing vast amounts of transaction data, logistic regression models can detect anomalies and predict fraud more effectively than traditional rule-based systems.

Databases play an essential role in this process as

they store the vast amounts of transactional data needed for analysis.

These databases typically consist of features such as transaction ID, user ID, transaction amount, time, merchant information, location, and more. This data is used to train the machine learning model, allowing it to learn the patterns of legitimate versus fraudulent activity. With the application of techniques like under sampling or oversampling, the model can handle imbalanced datasets where fraudulent transactions are much fewer than legitimate ones, improving its detection accuracy.

Once trained, the logistic regression model can be integrated into a fraud detection system, which can evaluate incoming transactions in real-time, flagging those with a high probability of fraud for further review.

The system's effectiveness is evaluated using

ability to identify fraud without generating too many false positives.

The combination of machine learning techniques like logistic regression and well-structured databases for storing transactional data offers an efficient solution to combat credit card fraud. By automating the detection process, financial institutions can significantly reduce the risks associated with fraudulent transactions, protecting both consumers and their business operations.

3.1 LITERATURE SURVEY

[4] Han, J., "Data Mining: Concepts and Techniques" (3rd Edition), Morgan Kaufmann Publishers This comprehensive textbook provides foundational knowledge on data mining techniques, including anomaly detection and classification algorithms, which are essential for developing effective fraud detection systems. The book covers various data mining techniques, such as classification, clustering, association rule learning, and anomaly detection. It explains how classification methods (e.g., decision trees, neural networks) categorize transactions, clustering algorithms identify unusual patterns, association rule learning uncovers relationships between features, and anomaly detection identifies transactions that deviate from normal patterns. Data preprocessing steps like cleaning and transformation are also discussed, along with evaluation methods like cross-validation.

[5] Yan, X., "Credit card fraud detection using machine learning algorithms," Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-7 This paper investigates

machine learning algorithms, focusing on logistic regression and decision trees, for credit card fraud detection. It evaluates various models for their effectiveness in identifying fraudulent transactions using metrics like accuracy, precision, recall, and F1 score. Logistic regression is used for binary classification, while decision trees are used for their interpretability and ability to handle both categorical and numerical features. The study involves data preprocessing, feature selection, and cross-validation to ensure robust results.

[6] Ahmed, A., "Credit cards fraud detection using support vector machines," Journal of Computer Science and Technology, 26(6), 1057-1069 This study explores the application of Support Vector Machines (SVMs) for credit card fraud detection. SVMs are used to manage high-dimensional data and model non-linear relationships between features. The methodology includes preprocessing, normalization, and feature selection, along with kernel functions to handle non- linearity. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1score, with cross-validation used for robustness and generalizability.

3. Objectives of Logistic Regression in Credit Cards Fraud Detection

In this project, the objective is to utilize logistic regression to create a robust system for detecting fraudulent credit cards transactions. Logistic regression, a supervised learning algorithm, is particularly well-suited for binary classification problems such as fraud detection. By leveraging its simplicity, interpretability, and scalability, the following objectives are achieved:

- **Accurate Fraud Identification**:
  The model's primary role is to correctly classify transactions as either legitimate or fraudulent. This requires minimizing false positives (wrongly flagged legitimate transactions) and false negatives (fraudulent transactions that go undetected).
  Achieving high accuracy ensures trust in the system while reducing unnecessary disruptions to users.
- **Real-Time Detection**:
  Fraudulent activities often occur rapidly, necessitating immediate action. Logistic regression's efficiency in computation enables real-time detection, allowing institutions to block fraudulent transactions as they occur, preventing further damage.

Model Transparency:
Unlike complex models, logistic regression provides clear insights into how features (such as transaction amount, location, and user behavior) influence fraud

- predictions. This interpretability fosters stakeholder confidence and ensures compliance with regulatory standards by explaining decision-making PROCESSES.
- **Imbalanced Data Handling**:
  Fraudulent transactions are rare compared to legitimate ones, creating an imbalanced dataset. Logistic regression accommodates this challenge through techniques such as resampling (oversampling fraud cases or undersampling legitimate cases) and cost-sensitive learning, ensuring balanced performance across both categories.

Scalability:
Credit card systems process millions of transactions daily, requiring scalable models. Logistic regression's computational simplicity allows it to handle large volumes of data

- efficiently, making it suitable for environments with high transaction loads.
- **Financial Loss Mitigation**:
  Accurate detection prevents unauthorized transactions, reducing financial losses for customers and financial institutions. This objective directly contributes to enhancing financial security and customer trust.

Cost Efficiency:
Compared to more complex machine learning models, logistic regression is computationally less intensive, reducing the operational costs of fraud detection systems without compromising effectiveness.

4. EXISTING METHODOLOGIES

Credit card fraud detection has advanced over time, leveraging a range of techniques from traditional statistical methods to cutting edge artificial intelligence (AI). Each approach offers unique strengths and weaknesses, requiring careful selection to meet specific system needs.

1. Rule-Based Systems

- **Description**: Operates using predefined rules (e.g., transactions thresholds, unusual locations).
- **Advantages**: Interpretable and scalable for large datasets.
- **Challenges**: Struggles with non-linear relationships and highly imbalanced data, often requiring resampling techniques.

Machine Learning Models

- **Description**: Utilizes algorithms like decision trees, random forests, and support vector machines to learn fraud patterns.
- **Advantages**: Captures complex, non-linear relationships and adapts to new patterns.
- **Challenges**: Requires significant computational resources and may lack interpretability.

Neural Networks

- **Description**: Deep learning models extract intricate patterns from large datasets, excelling in detecting sophisticated fraud schemes.
- **Advantages**: Effective for complex fraud scenarios; automatically processes raw data.
- **Challenges**: Demands high computational power, large labeled datasets, and suffers from interpretability issues.

Ensemble Methods

- **Description**: Combines multiple models (examples bagging, boosting, stacking) to improve detection accuracy.
- **Advantages**: Enhances robustness and reduces overfitting.
- **Challenges**: Increases system complexity and computational requirements.

By achieving these objectives, the project ensures a secure, scalable, and cost-effective solution for detecting credit card fraud, aligning with both technical and business goals.

- **Advantages**: Simple to implement and easy to understand.
- **Challenges**: Limited adaptability to evolving fraud patterns; prone to high false-positive rates.

2. Statistical Models

- **Description**: Employs techniques like logistic regression to predict fraud based on historical data (e.g., transactions amount, time, location).

Anomaly Detection Techniques

- **Description**: Identifies deviations from normal behavior using methods like clustering and outlier detection.
- **Advantages**: Effective for rare or novel fraud types.
- **Challenges**: Sensitive to noise and false positives if "normal" behavior is poorly defined.

These methodologies can be applied individually or in combination, depending on the credit card fraud detection system's requirements and constraints. The choice of approach must balance accuracy, interpretability, computational needs, and adaptability to evolving fraud trends.

5. EXPERIMENTAL SETUP

In today's digital era, credit card transactions are a vital part of the global economy, enabling seamless and convenient financial exchanges. However, this increased reliance on online payments has also led to a rise in fraudulent activities, posing a significant challenge to businesses, financial institutions, and consumers alike. The motivation for developing a credit cards fraud detection system using logistic regression stems from the following key factors:

□ Rapid Growth of Online Transactions: The surge in e-commerce and online banking has created new opportunities for fraudsters.

□ Financial Losses: Credit card fraud results in billions of dollars in losses globally every year, affecting not just financial institutions but also individual users.

□ Need for Proactive Detection: Early detection and prevention of fraud are essential to mitigate financial damage and protect users.

Privacy Concerns: Preventing fraud also helps in securing sensitive user data, ensuring that personal and financial information is not misused

□ Sophistication of Fraudsters: Fraudsters use advanced techniques, making it difficult for traditional methods to detect unusual patterns.

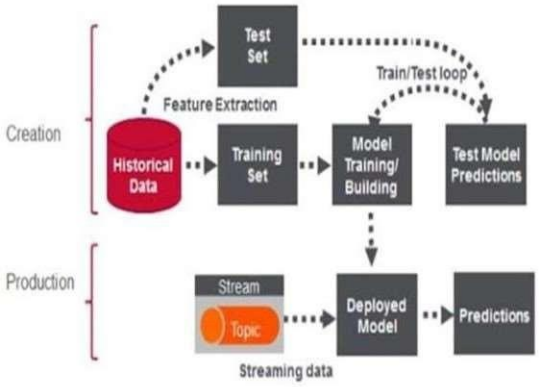


Fig.3.2.1 System Architecture

5.1 System Architecture

Data Flow:

1. **Input Data**:
   - A dataset containing transaction details, such as transaction amount, time, location, and whether the transaction is fraudulent or legitimate.
   - Real-time transaction inputs during deployment.
2. **Preprocessing Layer**:
   - Handles data cleaning, normalization, and feature selection to ensure high-quality inputs for the model.
3. **Model Layer**:
   - The logistic regression model is trained on historical data to predict the probability of a transaction being nice.

1. **Data Cleaning**:
   - Handle missing or incorrect values.
   - Remove duplicate entries to avoid the model.
2. **Feature Scaling**:
   - Normalize or scale numerical features to bring them within a uniform range (e.g., using Min Max Scaler or Standard Scaler from scikit-learn).
3. **Data Splitting**:
   - Split the dataset into:
     - **Training Set**: 70%-80% of fraudulent.
   - Outputs a probability score for each transaction.
4. **Decision Layer**:
   - Compares the probability score againsta predefined threshold (e.g., 0.5).
   - Classifies the transaction as either **fraudulent** or **legitimate**.
5. **Output Layer**:
   - Displays the classification and probability in real-time to users.
   - Provides visualizations and detailed reports.

Hardware and Software Requirements

Hardware:

- **System Configuration**: A computer with:
  - At least **8 GB RAM**.
  - A **multi-core processor** for faster computations.
- **Cloud Platform (Optional)**:
  - Use cloud services like aws, google cloud, or azure for scalable storage and hosting.

Software:

- **Programming Language**: Python (preferred for machine learning tasks).
- **Development Environment**: Use IDEs such as:
  - **Jupyter Notebook**: For iterative development.
  - **Visual Studio Code**: For larger, modular implementations.
- **Libraries and Tools**:
  - **pandas**: For handling and preprocessing data.
  - **numpy**: For mathematical operations.
  - **scikit-learn**: For machine learning tasks.
  - **matplotlib** and **seaborn**: For data visualization.
  - **streamlit**: For deploying the model as a web application.

3.3 Data

the data for training the model.

- **Testing Set**: 20%-30% of the data for evaluating the model.

4. **Handling Class Imbalance**:
   - Fraudulent transactions are usually rare. Address this imbalance using techniques like:
   - The model uses **binary cross-entropy** (also known as log loss) to measure how well it's The loss function compares the predicted probabilities with the actual labels.

Hardware and Software Requirements
Hardware:

- **System Configuration**: A computer with:
  - At least **8 GB RAM**.
  - A **multi-core processor** for faster computations.
- **Cloud Platform (Optional)**:
  - Use cloud services like AWS, google cloud, or azure for scalable storage and hosting.

Software:

- **Programming Language**: Python (preferred for machine learning tasks).
- **Developmental Environment**: Use IDEs such as:
  - **Jupyter Notebook**: For iterative development.
  - **Visual Studio Code**: For larger, modular implementations.
- **Libraries and Tools**:
  - **pandas**: For handling and preprocessing data.
  - **numpy**: For mathematical operations.
  - **scikit-learn**: For machine learning tasks.
  - **matplotlib** and **seaborn**: For data visualization.
  - **streamlit**: For deploying the model as a web application.

3.5 Data Preprocessing

Steps in Preprocessing:

1. **Data Cleaning**:
   - Handle missing or incorrect values.
   - Remove duplicate entries to avoid bias in the model.
2. **Feature Scaling**:
   - Normalize or scale numerical features to bring them within a uniform range (e.g., using Min Max Scaler or Standard Scaler from scikit-learn).
3. **Data Splitting**:
   - Split the dataset into:
     - **Training Set**: 70 percent -80 percent of the data for training the model.
     - **Testing Set**: 20 percent – 30 percent of the data for evaluating the model.
4. **Handling Class Imbalance**:
   - Fraudulent transactions are usually rare. Address this imbalance using techniques like:
   - The model uses **binary cross-entropy** (also known as log loss) to measure how well it's The loss function compares the predicted probabilities

3.2 How Logistic Regression Works in Credit Card Fraud Detection

Logistic regression used for binary classification tasks, such as distinguishing between legitimate and fraudulent credit card transactions. Here's a detailed breakdown of how it operates in the context of this project:

2. Model Basics

with the actual labels.

• Output: Logistic regression predicts the probability that a given input belongs to a specific class (eg, fraudulent vs. legitimate). The output is a value between 0 and 1.

Sigmoid Function: the range [0, 1]. The formula is P(Y=1 X)=1+ Here, P(Y1X) is the probability of the transaction being fraudulent (1), B is the intercept, and are the coefficients for each feature X1, XX

2. Input Features

Data Preparation: Each transaction is represented by a set of features (transaction amount time of transaction, and other relevant metrics).

Normalization: Features may be scaled or normalized to ensure they contribute equally to the model, especially if they vary in magnitude.

3. Training the Model

Loss Function: The model uses a loss function (usually binary cross-entropy) to measure the difference between predicted probabilities and actual labels (fraudulent or legitimate).

Optimization: An optimization algorithm, such as gradient descent, adjusts the coefficients (8) to minimize the loss function over multiple iterations.

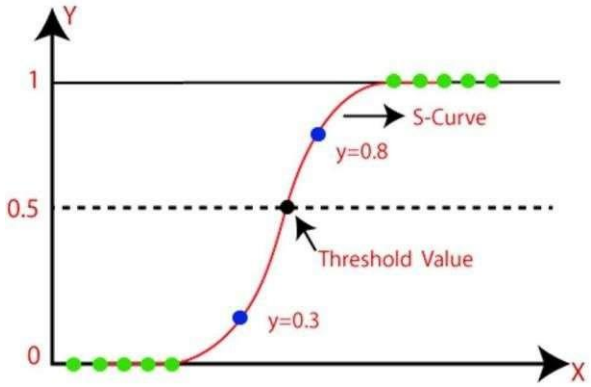
4. Making Predictions

Input New Data: Once trained the model takes new transaction data as input.

Probability Output: it calculates the probability of fraud using the learned coefficients and the sigmoid function.

5. Evaluation:

Performance is calculated using metrics like accuracy, precision, and recall.
In essence, it helps detect fraud by calculating the likelihood of a transaction being fraudulent and making



Nmap stands for Network Mapper and is an open- source application tool that is used for purposes of network discovery and security. This feature is vital when it comes to the disclosure of devices connected to the network, disclosure of a network's open ports, and closure of the services that are running on the specific ports. In a cybersecurity lab, the roles played by Nmap are for the reconnaissance phases of penetration testing where it can be used to create a map of the network topology and discover hosts, services, and possible open ports that can be exploited in other tests.

4. CONCLUSION:

In this project, logistic regression is employed as a powerful and interpretable machine learning technique for credit cards fraud detection. By transforming transaction data into a probability using the sigmoid function, the model classifies transactions as either fraudulent and legitimate based on a set threshold. Key steps in the process include data preprocessing, model training with historical transaction data, and real-time fraud detection.

The logistic regression model's ability to output probabilities makes it suitable for this task, where decision-making needs to be both precise and easily interpretable. It allows for dynamic and scalable fraud detection in real time, helping users make informed decisions. Furthermore, the model's evaluation using performance metrics such as confusion matrix, precision, recall, and F1 score ensures the model is both reliable and effective in distinguishing fraudulent activities.

By combining machine learning with a user-friendly web application, this project offers an efficient and accessible solution for detecting credit card fraud. It ensures that financial transactions are safe, reduces potential financial losses, and offers real-time protection for users, making it a practical tool for fraud

REFERENCES:

1. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques (3rd Edition)*.Morgan Kaufmann Publishers.
2. Yan, X., Liu, Y., & Xie, X. (2019). Credit card fraud detection using machine learning algorithms. In *Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp. 1-7.
3. Ahmed, A., Mohamed, S. N. S. K., & Al- Yahya, K. H. S. (2011). Credit card fraud detection using support vector machines. *Journal of Computer Science and Technology*, 26(6), 1057-1069.
4. Ghosh, S., & Reilly, D. (1994). Credit card fraud detection with a neural network. In *Proceedings of the IEEE International Conference on Neural Networks*, vol. 3, pp. 2215-2220.
5. Alahakoon, H. A., Munasinghe, H. A. S., & Lee, A. J. R. (2022). Credit card fraud detection using deep learning. *Journal of Financial Services Research*, 56(2), 229-247.
6. Iglewicz, L., & Iglewicz, K. C. N. T. K. L. R. (2020). Anomaly detection for credit card fraud using autoencoders. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 325-332.
7. Liu, F., Zhao, K., & Zhang, Y. (2020). Credit card fraud detection using Isolation Forests. *Journal of Machine Learning Research*,21(118), 1-25.
8. Ng, K. W., Tan, L. C., & Chiu, L. Y. (2020). Ensemble methods for credit card fraud detection: A comparative study. *IEEE Access*, 8,91870-91883.
9. Al-Ghamdi, M. T. B., & Ibrahim, S. N. I. (2019). Challenges and future directions in credit card fraud detection. *International Journal of Computer Applications*, 178(4), 28- 34.
10. Wang, Y., Zhang, Z., & Zhang, Y. (2017). Credit card fraud detection using deep learning :An evaluation of recent models. *Journal of Financial Crime*, 24(3), 4