**MINOR-1 PROJECT**

**SYNOPSIS**

**on**

Credit card fraud detection using ML models

Submitted By:

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **Branch** |
| Antra Chauhan | R2142220442 | CSE AIML(Hons.) |
| Sumit Verma | R2142220184 | CSE AIML(Hons.) |

**Under the guidance of**

Dr. Achala Shakya

A blue letter on a black background

Description automatically generated

School of Computer Science

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007 2023**

**Approved By**

**Project Guide**  **Cluster Head**

Dr. Achala Shakya Dr. Anil Kumar

**School of Computer Science**

University of Petroleum & Energy Studies, Dehradun

|  |  |  |
| --- | --- | --- |
| Synopsis Report (2024) | | |
| SNo. | **TOPIC** | **PageNo.** |
|  | Abstract | 3 |
|  | Introduction | 3 |
|  | Literature Review | 4 |
|  | Problem Statement | 6 |
|  | Objective | 6 |
|  | Methodology | 7 |
|  | System Requirements (Software/Hardware) | 9 |
|  | References | 10 |

**Project Title: Credit card fraud detection using ML models**

**Abstract**

Credit card fraud is a significant and growing challenge in the financial industry, causing substantial financial losses and threatening consumer trust. The complexity and volume of transactions require advanced detection mechanisms to prevent fraudulent activities effectively. This project focuses on developing a robust credit card fraud detection system using machine learning models, specifically the Random Forest Classifier and AdaBoost Classifier.

The proposed system leverages these models to identify fraudulent transactions by analyzing complex patterns and distinguishing anomalies within transaction data. The Random Forest Classifier is employed for its ensemble learning capabilities that enhance prediction accuracy through multiple decision trees, while the AdaBoost Classifier is used for its ability to improve weak classifiers through iterative learning, boosting performance on challenging datasets.

The system's effectiveness will be validated using performance metrics such as accuracy, precision, recall, F1-score, and particularly the ROC AUC score, which provides a comprehensive measure of the model's ability to discriminate between genuine and fraudulent transactions. This will offer deeper insights into its real-world applicability in safeguarding financial transactions and ensuring a high detection rate with minimal false positives.

**Introduction**

The rise of online transactions and digital payment systems has revolutionized the financial industry, making transactions faster and more convenient. However, this shift has also brought about an increase in credit card fraud, posing a major risk to consumers and financial institutions alike. Fraudulent transactions can lead to substantial financial losses and erode trust between businesses and their customers. Therefore, developing effective fraud detection systems is crucial to ensuring the integrity and security of financial operations.

Traditional fraud detection methods often rely on rule-based systems that struggle to adapt to the evolving strategies employed by fraudsters. Machine learning (ML) has emerged as a powerful tool for enhancing fraud detection capabilities due to its ability to learn complex patterns and make data-driven predictions. This project focuses on leveraging advanced ML models, specifically the Random Forest Classifier and AdaBoost Classifier, to build a credit card fraud detection system that can accurately identify fraudulent transactions.

The Random Forest Classifier is chosen for its ensemble approach, which combines multiple decision trees to improve the overall prediction accuracy and robustness. On the other hand, the AdaBoost Classifier is utilized for its boosting technique that strengthens the performance of weak learners through iterative training. By combining these models, the system aims to achieve a balance between high detection rates and minimal false positives.

To evaluate the effectiveness of the proposed system, performance metrics such as accuracy, precision, recall, F1-score, and the ROC AUC score will be employed. These metrics will provide a comprehensive analysis of the models' capabilities in distinguishing fraudulent activities from legitimate transactions, demonstrating their real-world applicability and potential to enhance fraud prevention measures.

# Literature Review

Credit card fraud detection is an area of extensive research due to its significance in ensuring the security of financial transactions. This section reviews key methodologies and approaches in the field, highlighting the progression from traditional techniques to advanced machine learning models, such as the Random Forest Classifier and AdaBoost Classifier, which are central to this project.

**1. Traditional Fraud Detection Techniques**

Initially, credit card fraud detection systems relied on rule-based and statistical models like logistic regression and decision trees. While these models were relatively simple to implement, they lacked the adaptability needed to respond to the rapidly changing strategies employed by fraudsters. Consequently, rule-based systems often resulted in higher false positive rates and reduced detection efficacy as fraudulent tactics evolved.

**2. Emergence of Machine Learning Approaches**

The limitations of traditional methods paved the way for machine learning (ML) algorithms that could model complex, non-linear relationships within transaction data. Early ML approaches included Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and neural networks, which improved detection accuracy and adaptability. However, these methods required significant preprocessing and fine-tuning to perform optimally, presenting challenges for practical deployment.

**3. Ensemble Learning and Its Advantages**

Ensemble learning methods, such as Random Forests and boosting algorithms, have gained prominence for their ability to enhance predictive performance by combining multiple models. The Random Forest Classifier, developed by Breiman (2001), leverages an ensemble of decision trees to create a robust model that reduces overfitting and handles imbalanced data effectively. Its use in fraud detection has been recognized for delivering high accuracy and resilience across various datasets.

AdaBoost (Adaptive Boosting), introduced by Freund and Schapire (1997), is another ensemble technique that iteratively improves the performance of weak classifiers to build a stronger predictive model. Research has shown AdaBoost's effectiveness in handling high-dimensional data and adapting to imbalanced datasets, making it well-suited for complex problems like fraud detection.

**4. Comparative Analyses**

Comparative studies in the literature indicate that ensemble models often outperform individual algorithms in fraud detection tasks. For instance, Bahnsen et al. (2016) found that ensemble techniques like Random Forests and boosting methods surpassed models such as SVM and logistic regression in terms of precision, recall, and the ROC AUC score. Additionally, Carcillo et al. (2018) demonstrated the advantage of AdaBoost's adaptive learning in evolving fraud patterns, contributing to its robust performance in real-world applications.

**5. Key Challenges and Performance Metrics**

Fraud detection systems face challenges, including the severe class imbalance present in transaction data, where legitimate transactions vastly outnumber fraudulent ones. This imbalance can skew model training and reduce the effectiveness of detection. To counteract this, techniques such as data resampling (oversampling and undersampling) and synthetic data generation (e.g., SMOTE) have been employed to create balanced training sets. Performance metrics such as accuracy, precision, recall, F1-score, and the ROC AUC score are essential for evaluating a model’s overall effectiveness in distinguishing fraudulent transactions.

**Conclusion**

The literature emphasizes the importance of leveraging advanced, ensemble-based ML models to build effective fraud detection systems. Random Forest and AdaBoost Classifiers have proven their capability to process complex datasets and achieve high accuracy while minimizing false positives. This project incorporates these models to develop a comprehensive credit card fraud detection system, validated through key performance metrics including the ROC AUC score, underscoring its real-world applicability and reliability.

# Problem Statement

The increasing volume and complexity of financial transactions have made credit card fraud a significant and persistent challenge for financial institutions and consumers alike. Fraudulent activities not only result in substantial financial losses but also compromise customer trust and pose risks to the integrity of financial systems. Traditional fraud detection methods, such as rule-based systems and basic statistical models, struggle to keep up with the evolving tactics of fraudsters, leading to high false positive rates and reduced detection accuracy.

The primary challenge in credit card fraud detection lies in the imbalance of data, where legitimate transactions vastly outnumber fraudulent ones. This imbalance can hinder the effectiveness of detection models, as they may become biased towards predicting non-fraudulent transactions. Additionally, identifying complex and subtle patterns indicative of fraud requires robust algorithms capable of learning from highly imbalanced and noisy datasets.

This project aims to address these challenges by developing an advanced credit card fraud detection system that leverages the strengths of machine learning, specifically the Random Forest Classifier and AdaBoost Classifier. By applying these ensemble learning techniques, the project seeks to create a model that not only improves the detection rate of fraudulent transactions but also minimizes false positives, ensuring a balance between precision and recall. The effectiveness of the proposed system will be validated using comprehensive performance metrics, including accuracy, precision, recall, F1-score, and the ROC AUC score, to demonstrate its real-world applicability in safeguarding financial transactions.

# Objective

# The main objective of this project is to develop an effective credit card fraud detection system using advanced machine learning models. The specific objectives include:

# 1. Develop and Implement Machine Learning Models: To design and implement credit card fraud detection models using the Random Forest Classifier and AdaBoost Classifier to identify fraudulent transactions accurately.

# 2. Enhance Detection Performance: To optimize the models to achieve a balance between high detection rates and minimal false positives, ensuring the models can effectively distinguish fraudulent transactions from legitimate ones.

# 3. Address Class Imbalance: To tackle the challenge of class imbalance in transaction data by incorporating techniques such as resampling and synthetic data generation (e.g., SMOTE) to improve model training and accuracy.

# 4. Evaluate Model Performance: To validate the effectiveness of the models using comprehensive performance metrics, including accuracy, precision, recall, F1-score, and the ROC AUC score, for a thorough analysis of the system's predictive capability.

# 5. Ensure Real-World Applicability: To demonstrate the robustness and reliability of the proposed fraud detection system, ensuring that it is applicable in practical financial scenarios and capable of adapting to evolving fraud patterns.

# Methodology

The project will be conducted in the following structured steps to ensure comprehensive data analysis and model development for credit card fraud detection:

**1. Read the Data:** The first step involves loading the dataset into the working environment. This ensures that the data is ready for analysis and modeling, forming the basis for subsequent steps.

**2. Check the Data:** Conduct an initial examination of the dataset to verify data integrity. This includes checking the data types, column names, and basic statistics to understand its structure and characteristics.

**3. Glimpse the Data:** Obtain a brief overview of the data through summary statistics and visual inspections, such as using `head()` and `describe()` functions. This step provides insights into the distribution and general patterns present in the dataset.

**4. Check for Missing Data:** Identify and handle any missing or incomplete data points to prevent issues during the modeling process. This involves using techniques such as imputation or removal, depending on the extent and nature of the missing data.

**5. Check Data Imbalance:** Assess the balance of the dataset, particularly the ratio of fraudulent to non-fraudulent transactions. Given that fraud detection datasets are often imbalanced, this step ensures appropriate handling through methods such as resampling or applying synthetic data generation techniques like SMOTE to balance the dataset.

**6. Data Exploration:** Conduct an in-depth exploratory data analysis (EDA) to uncover relationships, correlations, and trends within the data. Visualization tools such as histograms, box plots, and scatter plots will be used to better understand the features and their impact on fraud detection.

**7. Predictive Models:**

- Random Forest Classifier: Build and train a Random Forest model, leveraging its ensemble nature to create a robust predictive model through multiple decision trees. Hyperparameter tuning will be performed to optimize the model's performance.

- AdaBoost Classifier: Develop and train an AdaBoost model to enhance prediction accuracy by combining weak classifiers iteratively. This step will focus on maximizing the model's ability to adapt and improve detection rates.

**8. Model Evaluation:** Validate the performance of both models using key metrics such as accuracy, precision, recall, F1-score, and the ROC AUC score. This comprehensive evaluation ensures that the models are assessed on their effectiveness in detecting fraud and handling class imbalance.

**9. Conclusions:** Summarize the findings and outcomes of the project. Highlight the strengths and weaknesses of the models and provide insights into their potential real-world application. Recommendations for future improvements or further research will also be included.

**System Requirements**

**Software Requirements:**

1. Programming Language:

- Python (preferred for its extensive libraries and community support)

2. Development Environment:

- Jupyter Notebook or any Python IDE (e.g., PyCharm, Visual Studio Code, or Spyder)

3. Libraries and Frameworks:

- Data Analysis: Pandas, NumPy

- Data Visualization: Matplotlib, Seaborn

- Machine Learning: Scikit-learn (for building and evaluating Random Forest and AdaBoost classifiers)

- Imbalance Handling: Imbalanced-learn (for techniques like SMOTE)

- Performance Metrics: Scikit-learn’s `metrics` module for evaluating model performance

- Data Preprocessing: Scikit-learn’s `preprocessing` module for scaling and transforming data

4. Version Control:

- Git (optional, for version control and collaboration)

5. Operating System:

- Windows 10/11, macOS, or Linux (any OS capable of running Python and associated libraries)

6. Others:

- Anaconda Distribution (optional, provides an easy installation of Python and commonly used data science libraries)

**Hardware Requirements:**

1. Processor:

- Minimum: Intel Core i5 or equivalent

- Recommended: Intel Core i7 or AMD Ryzen 5/7 for faster data processing and model training

2. RAM:

- Minimum: 8 GB (sufficient for small to medium datasets)

- Recommended: 16 GB or more (for handling large datasets and faster computation)

3. Storage:

- Minimum: 256 GB HDD/SSD (to store datasets and project files)

- Recommended: 512 GB SSD or more (for better read/write speed and data handling)

4. Graphics Processing Unit (GPU):

- Optional: Dedicated GPU (e.g., NVIDIA GeForce GTX/RTX series) for accelerated model training if using deep learning frameworks or processing very large datasets.

5. Internet Connection:

- Required for downloading libraries, datasets, and cloud-based resources if necessary.

**References**

1. Breiman, L.\*(2001). Random Forests. Machine Learning, 45(1), 5-32.

2. Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119-139

3. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2016). Example-Dependent Cost-Sensitive Decision Trees for Credit Card Fraud Detection. Expert Systems with Applications, 42(10), 5321-5331.

4. Carcillo, F., Le Borgne, Y. A., Caelen, O., et al. (2018). Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Fusion, 51, 258-270.

5. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

6. 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.

7. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

8. Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer. ISBN: 978-3-319-14142-8.