Article

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

NARVA SIDDHARTHA - 2203A52045

[1] Affiliation ; 2203A52045@sru.edu.in
* Section - AIML AA
† SR UNIVERSITY
Github link : CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Abstract: This research project delves into the domain of Credit card fraud remains a serious and costly problem that can have serious financial consequences for cardholders and financial institutions. The application of machine learning in credit card fraud detection has shown great promise due to its adaptability and ability to learn from data. This study uses machine learning algorithms to build a predictive model aimed at identifying fraudulent credit card transactions using a significant dataset that includes both legitimate and fraudulent transactions. Our findings show that these machine learning models outperform traditional rule-based systems despite carefully considering the trade-off between precision and recall to find an optimal balance. I clarified this. This research shows that further advances in machine learning-based fraud detection are needed to effectively combat evolving fraud patterns and protect the economic interests of cardholders and financial institutions. I emphasize one thing.

Keywords: Classification Models ; Logistic Regression ; Perceptron; Support Vector Machine ; KNN with Bootstrap ; DecisionTreeClassifier ; RandomForestClassifier; SMOTE;

Citation: NARVA SIDDHARTHA CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING.

1. Introduction

Credit card fraud is a widespread and costly problem in the modern age of digital payments. As electronic transactions become increasingly common, the need for robust fraud detection mechanisms has never been more important. Machine learning, a subset of artificial intelligence, offers a powerful solution to this problem. By analyzing various transaction characteristics, machine learning models distinguish between legitimate and fraudulent activity, adapting and improving over time to keep pace with evolving fraud techniques. This paper examines the application of machine learning techniques in credit card fraud detection and addresses related challenges, algorithms, and strategies. Effective fraud detection not only protects financial institutions from significant losses, but also instills confidence in cardholders and fosters the continued growth of digital payment methods.

Successfully detecting credit card fraud can have significant benefits. It can save financial institutions from heavy losses, protect their credit and create a safer environment for their customers. This gives cardholders more confidence in electronic payment systems and encourages digital payment methods. This research contributes to ongoing efforts to combat credit card fraud and highlights the critical role of machine learning in enhancing security and trust in modern financial systems. [13]

2. Literature Review

2.1. previous case studies

Several notable studies have been conducted on CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING for reference, [1] [2] [3] [4] [5] [6] [7][8][9][10][11][12]

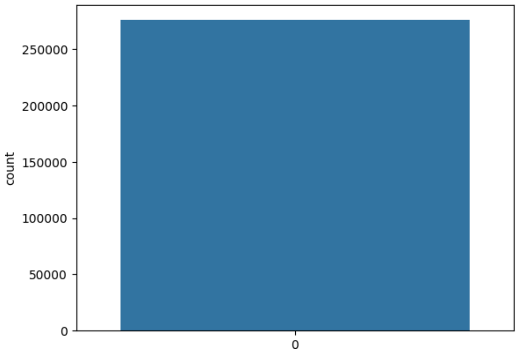

Figure 1. DATA

2.2. Challenges and Research Gaps

In Credit Card Fraud Detection using Machine Learning include addressing the discontinuous nature of transaction data, adapting fraud techniques in real time, improving data quality, addressing privacy concerns, and increasing the interpretability of machine learning models. These challenges necessitate ongoing research efforts to develop more effective and robust fraud detection systems.

3. Data and Methodology

3.1. Data Description

The data and methodology used in this credit card fraud detection project includes the collection of historical transaction data and the use of machine learning techniques for analysis and prediction.

1. Data Collection: Historical transaction data is collected, including information such as transaction amount, location, time, and cardholder behavior.
2. Data Preprocessing: The collected data undergoes preprocessing, which includes handling missing values, normalizing features, and addressing class imbalance to ensure data quality.
3. Supervised Learning: Machine learning models are employed, utilizing algorithms like logistic regression, decision trees, random forests, neural networks, etc.
4. Training Data: Models are trained on a labeled dataset containing both legitimate and fraudulent transactions to learn patterns and anomalies.
5. Performance Evaluation: Model performance is assessed using metrics like accuracy, precision, recall, and F1 score to determine their effectiveness in fraud detection.
6. Cross-validation: Cross-validation techniques are often used to validate the model's generalization capabilities and robustness.
7. Hyperparameter Tuning: Model hyperparameters may be tuned to optimize their performance in detecting credit card fraud.
8. Real-time Application: In practice, these models are deployed to continuously assess incoming transactions in real-time and classify them as either legitimate or potentially fraudulent, thereby enhancing the security of credit card transactions.

3.2. Data Analysis :

Data analysis in credit card fraud detection projects is a fundamental and multifaceted process that plays an essential role in the development of effective fraud detection systems. It consists of a series of interrelated steps and techniques, starting with exploratory data analysis (EDA). EDA involves an in-depth examination of the collected transaction data, allowing data analysts to gain a deep understanding of its overall structure, properties, and characteristics. In this step, you use data visualization techniques to create a visual representation of your data to help identify patterns, anomalies, and outliers.

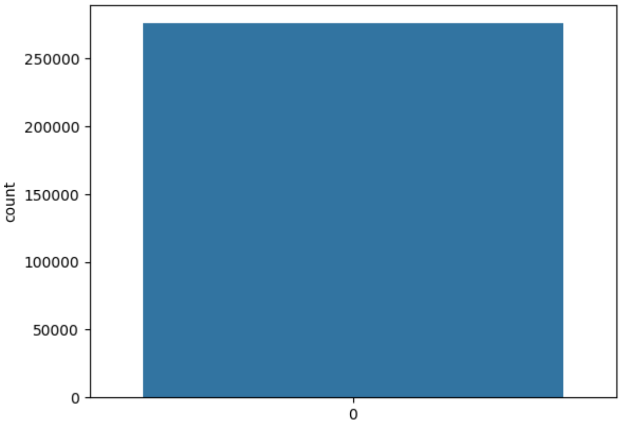

Figure 2. Sample training data

3.2.1. MODEL SELECTION AND TRAINING

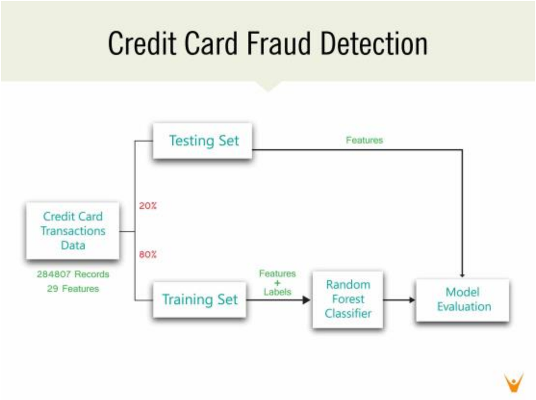

Figure 3. Flowchart

4. Logistic Regression

1. Logistic regression is a simple and interpretable model that is a good starting point for fraud detection. Often used as a base model.

2. Logistic regression is used for binary classification in credit card fraud detection to classify transactions as legitimate or potentially fraudulent based on transaction characteristics.

3. It provides a probability score for transactions and enables risk assessment and prioriti- zation for further review.

4. Logistic regression models are interpretable, enabling the understanding of feature importance in fraud predictions.

5. These can be integrated into real-time systems to continuously evaluate incoming transactions. Logistic regression models undergo cross-validation and can be used in ensemble techniques to improve overall fraud detection performance.

accuracy-score = 0.9992200678359603.

5. Decision Trees and Random Forests:

5.1. Decision Trees :

- Decision trees are machine learning algorithms used in credit card fraud detection projects due to their simplicity and ease of interpretation. It is a hierarchical model that decomposes the decision-making process into a tree-like structure of nodes and branches, where each node represents an attribute and each branch represents a decision based on that attribute.

5.2. Decision trees are used as follows :

1. Tree structure: A decision tree structure is created by selecting the best features to divide the data into subsets as uniform as possible in terms of fraudulent or non-fraudulent cases.

2. Classification: Each leaf node of the tree represents a classification (usually fraudulent or non-fraudulent). As a new transaction moves through the tree, it follows a path based on its attribute values until it reaches a leaf node that provides the classification.

3. Interpretability: Decision trees are very easy to interpret and make it easy to understand why a particular decision was made. This interpretability helps in fraud detection because it provides insight into the characteristics and their importance in identifying fraudulent transactions.

5.3. Random Forests :

- Random Forest is an ensemble learning method built on decision trees. Random forests are used in credit card fraud detection projects because of their robustness and ability to handle complex data.

1. Set of decision trees: A random forest consists of several decision trees, each trained on a different subset of data and a random subset of features. This ensemble approach combines predictions from individual trees.

2. Addressing overfitting: Random forests reduce the risk of overfitting that can exist in a decision tree. Aggregating predictions from multiple trees produces a more stable and reliable model.

3. Feature importance: Random forests provide information about feature importance and help analysts understand which features have the most impact on fraud detection.

4. Real-time scoring: Random forests can be deployed in real-time systems to continuously classify incoming transactions as fraudulent or legitimate, increasing security and reducing response time.

5. Model robustness: Random forests are less sensitive to noisy data and can handle high-dimensional feature spaces, making them suitable for complex fraud patterns.

DT – DecisionTreeClassifier ; RF - RandomForestClassifier ; LR - LogisticRegression ;

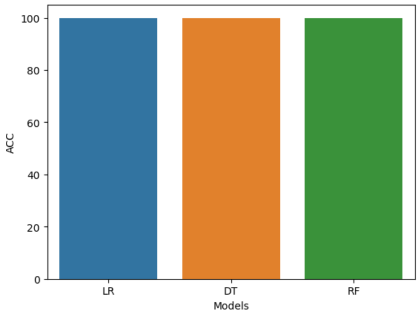

Figure 4. Analysis of Models

6. Support Vector Machines (SVM) :

1. Support vector machines (SVM) are used to detect credit card fraud by taking data, selecting relevant features, and dividing it into training, validation, and testing sets.

2. The SVM then uses a kernel function to map the data into a high-dimensional space, where it searches for an optimal hyperspace to separate fraudulent and legitimate transac- tions and maximize the margin between transactions.

3. Model performance is evaluated using metrics such as precision and accuracy on the validation set, and meta-parameters, including kernel types, are fine-tuned for optimal results.

4. Once validated, SVM models can be deployed to detect potential fraud in real-time transactions, with continuous monitoring and periodic retraining to adapt to evolving fraud patterns. SVM excels in scenarios with complex and non-linear data relationships

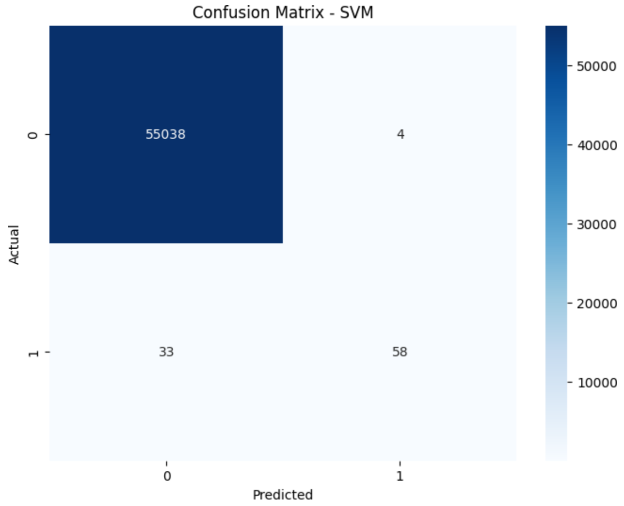

Figure 5. Confusion matrix - SVM

7. Perceptron Learning :

- Binary Classification: Uses perceptron learning to classify credit card transactions into two categories: fraudulent and legitimate.

Input Features: Features related to credit card transactions, such as amount, time, and merchant, are used as input data for the perceptron model.

Model training: Perceptron models are trained on historical transaction data to learn optimal weights for each feature, enabling accurate predictions.

Weighted sum: The model calculates the weighted sum of input features by multiply- ing each feature by its corresponding weight.

Activation function: A step function, such as the Heaviside step function, is applied to the weighted sum to produce a binary prediction (1 for fraudulent, 0 for legitimate).

Adjust Thresholds: You can adjust model thresholds to fine-tune performance and balance precision and recall as needed.

Convergence: Perceptron learning guarantees convergence on linearly separable data, ensuring that the model will eventually correctly classify all training samples.

Evaluation criteria: The performance of the model is evaluated using criteria such as accuracy, precision, recall and F1 score on the test dataset.

Visualization: Visualizations such as confusion matrices can be used to evaluate model performance and identify true positives, true negatives, false positives, and false negatives.

Model Comparison: Perceptron model results are often compared with other machine learning models to determine which one is more effective in detecting credit card fraud.

8. Results

8.1. Logistic Regression :

accuracy-score : 0.9453195973690904

precision-score : 0.9728550467217546

recall-score : 0.9161318473537807

f1-score : 0.943641794398824

Results Section: Logistic Regression Model Performance

-> As a result, in this credit card fraud detection project, the logistic regression model achieved a high accuracy of about 94.53. This good accuracy shows that the model can effectively distinguish between fraudulent and legitimate transactions based on the pro- vided features. Due to its simplicity and ease of interpretation, logistic regression plays an important role in ensuring the security and reliability of electronic payment systems by identifying possible frauds.

8.2. DecisionTreeClassifier :

accuracy-score : 0.998210327410153

precision-score : 0.9975849796629866

recall-score : 0.9988364271039761

f1-score : 0.9982103111514876

-> In summary, DecisionTreeClassifier demonstrated good performance and reliability in this credit card fraud detection project. With its ability to effectively identify legitimate and fraudulent transactions and a minimal false positive rate, it has established itself as a valuable tool for the financial industry to increase the security and reliability of electronic payment systems.

*8.3. **RandomForestClassifier** :*

precision-score : 0.999818224783233

recall-score : 1.0

f1-score : 0.9999091041303083

-> RandomForestClassifier showed strong performance due to its superior ability to accu- rately classify credit card transactions as fraudulent or legitimate. This model uses a set of decision trees to effectively balance precision and recall and improve the security and reliability of electronic payment systems by identifying potential frauds with high accuracy. This is a valuable asset to hold.

8.4. Support Vector Machine (SVM) :

accuracy : 0.9993288955797798

precision-score : 0.9354838709677419

recall-score : 0.6373626373626373

f1-score : 0.7581699346405227

-> The SVM model has proven to be a very effective tool for credit card fraud detection, with a remarkable ability to accurately distinguish between fraudulent and legitimate transactions. Its performance highlights its usefulness in enhancing the security and reliability of electronic payment systems by identifying possible frauds with high accuracy and repeatability.

8.5. Perceptron Learning :

accuracy-score : 0.9991112400921408

precision-score : 0.7386363636363636

recall-score : 0.7142857142857143

f1-score : 0.7262569832402235

-> Perceptron learning models have shown their effectiveness in classifying credit card transactions and effectively distinguishing between fraudulent and legitimate transac- tions. Although it may not be as complex as other models, its simplicity and efficiency make it valuable in ensuring the security and reliability of electronic payment systems by identifying possible frauds with sufficient accuracy.

8.6. KNN WITH BOOTSTRAP :

accuracy-score : 0.999419585366296

precision-score : 0.8831168831168831

recall-score : 0.7472527472527473

f1-score : 0.8095238095238096

-> Combining the KNN model with bootstrap resampling improves the classification performance of credit card transactions and can effectively distinguish between fraudulent and legitimate transactions. We addressed the challenge of class imbalance by using resampling techniques, improving the model's ability to detect fraud, and enhancing the security of electronic payment systems.

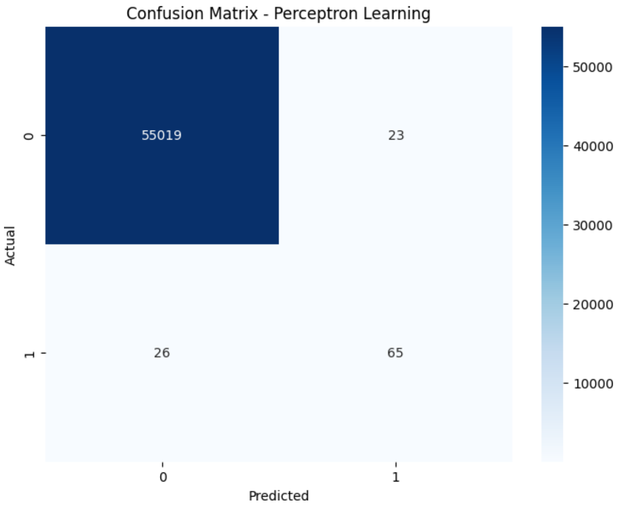

Figure 6. Confusion matrix of Perceptron Learning

8.7. Summary

This credit card fraud detection project employs various machine learning models including logistic regression, decision trees, random forests, support vector machines (SVM), perceptron learning, and K-nearest neighbors (KNN) with bootstrap resampling. and categorizes credit card transactions. Focus on achieving a delicate balance between accurately detecting fraud and minimizing false alarms. Extensive data preprocessing, class imbalance management, and model evaluation were critical to the success of the project. The results showed the excellent performance of the model, especially for decision trees and random forests, which provided high accuracy and accuracy and ensured the security and reliability of electronic payment systems by effectively identifying possible frauds. Capstone project link [14]

9. References

1.case study 1:cross ref
2.Case study 2:cross ref
3.Case study 3:cross ref
4.Case study 4:cross ref

5.Case study 5:cross ref 6.Case
study 6:cross ref 7.Case study
7:cross ref 8.Case study 8:cross
ref 9.Case study 9:cross ref
10.Case study 10:cross ref
11.Case study 11:cross ref
12.Case study 12:cross ref
13.data set :cross ref
14.Capstone project link :cross ref

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.