Asian Journal of Research in Computer Science

Volume 17, Issue 7, Page 41-54, 2024; Article no.AJRCOS.118279
ISSN: 2581-8260

A Logistic Regression-based Model for Identifying Credit Card Fraudulent Transactions

Abdulrashid Sani [a]*, Zahriya Lawal Hassan [a] and Anas Tukur Balarabe [a]

[a] *Department of Computer Science, Sokoto State University, Sokoto State, Nigeria.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajrcos/2024/v17i7476

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/118279

Original Research Article

Received: 05/04/2024
Accepted: 09/06/2024
Published: 15/06/2024

ABSTRACT

The rapid evolution of technology has significantly transformed payment methods, with a notable shift towards online platforms. However, this transition has also witnessed a concurrent increase in fraudulent activities, particularly within online credit card transactions. In response to the escalating occurrences of fraudulent online credit card transactions, this study proposes the development of a robust fraud detection model utilizing machine learning algorithms implemented in Python. Leveraging credit card transaction data sourced from Kaggle, the research utilizes Logistic Regression for both training and testing datasets to identify fraudulent transactions. The efficacy of the model is evaluated using separate test data, resulting in an impressive accuracy rate of 99.87% in detecting previously unseen fraudulent transactions. Further scrutiny of the test data reaffirms this high accuracy, registering a similar rate of 99.8%, thus underscoring the model's adeptness in handling novel data instances. The findings are succinctly represented visually, elucidating the

**Corresponding author: E-mail: abdulrashid.sani@ssu.edu.ng;*

Cite as: *Sani, Abdulrashid, Zahriya Lawal Hassan, and Anas Tukur Balarabe. 2024. "A Logistic Regression-Based Model for Identifying Credit Card Fraudulent Transactions". Asian Journal of Research in Computer Science 17 (7):41-54. https://doi.org/10.9734/ajrcos/2024/v17i7476.*

model's efficacy in bolstering online transaction security. By amalgamating advanced machine learning techniques with Python programming, this research contributes to the ongoing efforts aimed at enhancing security measures surrounding online credit card transactions by identifying legit and fraudulent transactions. Such endeavors are paramount in mitigating the adverse impacts of fraudulent activities on both financial stakeholders and consumers.

Keywords: Credit card fraud; machine learning algorithms; logistic regression; fraudulent transactions; legit transactions; accuracy; model; test data; training data; python; library; Jupiter note book.

1. INTRODUCTION

Credit card fraud poses an increasingly pressing and disconcerting challenge in our rapidly evolving technological landscape. As technology advances at an unprecedented pace, security measures must equally evolve and fortify themselves against imminent threats. The surges in digital payment methods and government-driven initiatives promoting plastic money use have exacerbated this issue's complexity [1].

Credit card fraud, defined as the unauthorized use of another person's credit card or credit card information for fraudulent transactions, inflicts substantial financial losses upon both victims and credit card companies. In response to this pervasive issue, credit card fraud detection systems have become indispensable to the credit card industry [2].

Credit card fraud can occur in several ways, including skimming, phishing, counterfeiting, and identity theft. Skimming is particularly widespread, with fraudsters using small devices called skimmers to illegally obtain credit card details from unsuspecting individuals.

Phishing is another common tactic involving deceptive emails or websites designed to trick victims into disclosing their credit card information [3].

Traditional credit card fraud detection software employs diverse techniques, including pattern recognition, anomaly detection, and predictive modeling. While these systems analyze copious transaction data, they often fail to identify cunning fraud attempts that may appear innocuous on the surface but pose significant financial risks internally. Consequently, conventional fraud detection systems struggle to efficiently detect sophisticated fraud schemes due to inherent limitations in their design and functionality [4].

To address this challenge, our project adopts machine learning algorithms, leveraging Python libraries such as Pandas, NumPy, MatLab, and MatPlotLib for data analysis and visualization. Linear Regression Machine learning algorithms are employed in this study; the algorithm possesses the unique capability to scrutinize vast datasets and unveil patterns indicative of illicit behavior, making it highly adept at uncovering credit card fraud. It excels in spotting transaction irregularities, such as those occurring at unusual times or locations and involving atypical amounts, which often elude traditional detection methods [5].

In essence, our endeavor aims to join the power of machine learning to develop a robust and accurate fraud detection system, ensuring the security of financial transactions in an ever-evolving digital landscape.

1.1 Problem Definition

Because of advancements in e-commerce systems and communication technology, credit cards have emerged as one of the most prevalent payment methods for both every day and online transactions. Unfortunately, this widespread adoption has led to a significant surge in associated fraud. Each year, illicit credit card transactions result in substantial losses for both businesses and individuals. Fraudsters have adeptly leveraged technology to siphon funds from unsuspecting victims, necessitating a proactive response to thwart their malicious activities.

When a credit card is duplicated or stolen, the ensuing transactions are classified as fraudulent. Detecting and preventing these illicit transactions in a timely manner is of utmost importance, as the resultant financial losses can be substantial. With the increasing ubiquity of credit card usage, the financial toll inflicted by credit card fraud continues to mount. Simultaneously, fraudsters continually explore new technological avenues to perpetrate their illicit schemes.

The primary aim of this study is to develop a highly accurate and efficient model utilizing logistic algorithms to detect fraudulent credit card transactions. By leveraging advanced analytical techniques, our goal is to create a robust model capable of identifying suspicious activities with precision and timeliness. Through this endeavor, we seek to tackle the prevalent challenges associated with credit card fraud, thereby safeguarding the interests of both individual consumers and businesses. By enhancing fraud detection mechanisms, we aim to minimize financial losses, protect sensitive information, and uphold trust within the financial ecosystem.

2. LITERATURE REVIEW

This chapter delves into the existing body of knowledge surrounding credit card fraud detection model, offering a concise synthesis of key findings and insights from previous research. Through this exploration, this study is aim to contextualize within the broader scholarly discourse and identify gaps for further investigation [6].

a. What is Credit Card Fraud?

Credit card fraud can be defined as the intentional and unauthorized manipulation or exploitation of credit card information, payment mechanisms, or transactional processes, undertaken with the aim of illicitly acquiring financial gain or benefits at the expense of legitimate cardholders, financial institutions, or merchants [7]. It encompasses a spectrum of deceptive practices, including but not limited to identity theft, card-present fraud, and card-not-present fraud, often facilitated by sophisticated techniques such as phishing, skimming, or data breaches. At its core, credit card fraud represents a breach of trust and integrity within the financial ecosystem, posing significant economic, regulatory, and social challenges that necessitate proactive detection, prevention, and mitigation strategies to safeguard against its deleterious effects [8].

b. Related Work

Renuka Devi has addressed the pressing issue of credit card fraud in the digital era. They assert that the proliferation of online payments and the heightened reliance on credit cards post-pandemic have exacerbated the challenge of fraud detection. Traditional fraud detection mechanisms, they argue, are hampered by inherent limitations, particularly in identifying sophisticated fraudulent activities. In response to this, Devi and Ray propose a credit card fraud detection model leveraging machine learning and convolutional neural networks, recognized for their efficacy in predictive analysis. Their model integrates simple yet potent technologies to ensure robust and accurate fraud detection, encompassing techniques such as pattern recognition, anomaly detection, and predictive modeling. However, they caution that while these methods are commonly utilized in fraud detection software, they may inadvertently overlook subtle fraudulent transactions, thereby exposing organizations to significant financial risks. However, the proposed model, as outlined by Devi and Ray, entails preprocessing techniques, weighted average calculations, and training utilizing machine learning algorithms like Logistic Regression, SVM, and K-Nearest Neighbor. Nonetheless, they acknowledge that the complexity of data preprocessing techniques discussed in their paper, including outlier rectification and feature extraction, may present challenges in practical implementation, particularly for users with limited technical expertise. Furthermore, Devi and Ray highlight that the implementation and maintenance of the AI/ML/CNN model for fraud detection could necessitate substantial computational resources and expertise, potentially rendering it less accessible for smaller financial institutions or organizations with constrained resources [9].

The research paper by Kolli Nikhil et al. proposes a CatBoost-based system for detecting credit card fraud, with the aim of accurately identifying fraudulent transactions while minimizing false positives to maintain customer satisfaction. CatBoost, a machine learning algorithm, is highlighted for its proficiency in handling categorical features and unbalanced datasets, rendering it suitable for credit card fraud detection. The evaluation of the model's efficiency in spotting fraudulent transactions is conducted using various performance indicators such as precision, recall, and F1-score. The paper underscores the importance of robust credit card fraud detection models in light of the significant financial losses associated with credit card theft, emphasizing the potential of machine learning algorithms like CatBoost in effectively addressing this issue. However, the research paper does not explicitly discuss specific limitations or challenges encountered during the implementation or evaluation of the CatBoost-based credit card fraud detection system. While it acknowledges the effectiveness of CatBoost in managing categorical features and unbalanced datasets, it does not delve into potential drawbacks or areas where the algorithm may not perform optimally. The focus of the paper is primarily on presenting the proposed CatBoost-based system for credit card fraud detection and evaluating its efficiency using performance indicators, without discussing potential limitations or areas for improvement in the model. Furthermore, the paper does not address any external factors or real-world constraints that could impact the practical implementation of the proposed fraud detection system using CatBoost. Overall, the limitations of the research paper lie in the lack of discussion on specific challenges faced during the study, potential drawbacks of the CatBoost algorithm in this context, and considerations for real-world application and scalability of the proposed system [10].

The study conducted by Dhwanir Shah and Lokesh Kumar Sharma emphasizes the importance of implementing a secure credit card fraud detection system to mitigate financial losses stemming from fraudulent transactions. Notably, Decision trees and Random Forest algorithms are singled out for their efficacy in dataset analysis and accurate identification of fraudulent transactions. In their research, Shah and Sharma employ data preprocessing techniques such as OneHotEncoding and Target Guided Mean encoding to optimize the dataset for classification tasks. They present a performance evaluation of the Decision Tree classifier, both before and after parameter tuning, using confusion matrices to demonstrate the model's enhanced accuracy and effectiveness in fraud detection. However, it is noted that the dataset utilized in the study is simulated, potentially lacking the complexity and variability of real-world credit card transaction data. Shah and Sharma acknowledge that this simulated dataset may result in classifiers achieving 100% accuracy, which might not accurately reflect their performance in a more realistic setting. Additionally, the paper does not extensively delve into the computational complexity or scalability of the proposed fraud detection system, aspects crucial for real-time applications or handling large-scale datasets. Furthermore, the evaluation metrics employed to gauge the models' performance are not thoroughly discussed, potentially limiting the comprehensive understanding of the model's effectiveness beyond accuracy [11].

Sandhya et al. discussed the application of machine learning techniques in credit card fraud detection. They evaluated algorithms including Naive Bayes, Bernoulli, and Random Forest, focusing on metrics such as accuracy, recall, and F1-score. The study demonstrated the efficacy of these algorithms in analyzing customer transaction data streams for detecting fraudulent activities, with Random Forest exhibiting superior performance in accuracy and precision for fraud detection. The classification report, delineating class 0 as valid transactions and class 1 as fraudulent transactions, was provided. Moreover, the authors highlighted the limitations of using accuracy from the confusion matrix for unbalanced categorization, proposing computation of accuracy score and precision by comparing false positives generated by the code to actual occurrences [12].

The study conducted by Varun Kumar K S et al employed various machine learning algorithms for fraud detection. Logistic Regression was used for classification, Decision Trees for both classification and regression, and K Nearest Neighbor (KNN) algorithm was explored as well. Logistic Regression was supplemented with synthetic minority oversampling to handle data imbalance. However, the paper lacked in-depth discussions on crucial aspects such as dealing with skewed data, class imbalance, and handling categorical data in fraud detection, which are vital in real-world scenarios. Additionally, there was limited exploration on the interpretability of the models, scalability, computational complexity, generalizability to different datasets, adaptability to evolving fraud patterns, and potential drawbacks of the techniques used, which are all essential considerations for deploying effective fraud detection systems [3].

3. METHODOLOGY

The methodology adopted for this research was qualitative in nature, chosen specifically to delve into and comprehend the intricacies of the dataset. Given the primary objective of scrutinizing and distinguishing between legitimate and fraudulent transactions, a qualitative approach emerged as the most fitting strategy [13].

Qualitative research was deemed essential as it allowed for a nuanced exploration of the dataset's parameters, facilitating a deeper understanding of the underlying patterns and anomalies within [14]. By immersing ourselves in the data, we were able to discern subtle nuances that may have eluded a purely quantitative analysis.

Moreover, the complexity of the task at hand left us with no alternative but to employ qualitative methods. Unlike quantitative techniques, which primarily focus on numerical data and statistical analysis, qualitative research offered the flexibility to probe into the contextual nuances and subjective factors that often play a pivotal role in discerning fraudulent activities [15].

In essence, the decision to utilize qualitative research methodology was driven by the need to explore the multifaceted nature of the dataset comprehensively. By adopting this approach, we were able to gain deeper insights into the dynamics of legitimate and fraudulent transactions, thereby enhancing the effectiveness of our analysis and decision-making processes.

a. Model Architecture

The Research Architecture depicted below illustrates the sequential flow of the research process. It commences with data acquisition from the Kaggle website, focusing on transactional data earmarked specifically for detecting credit card fraud. Following data collection, the next phase involves data preprocessing, aimed at cleansing and formatting the data to facilitate analysis and modeling [16]. Subsequently, data analysis ensues, involving statistical manipulations and calculations to inform the development of a logistic regression model.

Further along, the dataset undergoes division into training and testing subsets, crucial for model development through iterative training and evaluation. Here, logistic regression is employed to train the model using the training data, distinguishing between legitimate and fraudulent transactions [17]. Finally, the model undergoes evaluation by testing its performance on the unseen test dataset, providing insights into its efficacy and robustness.

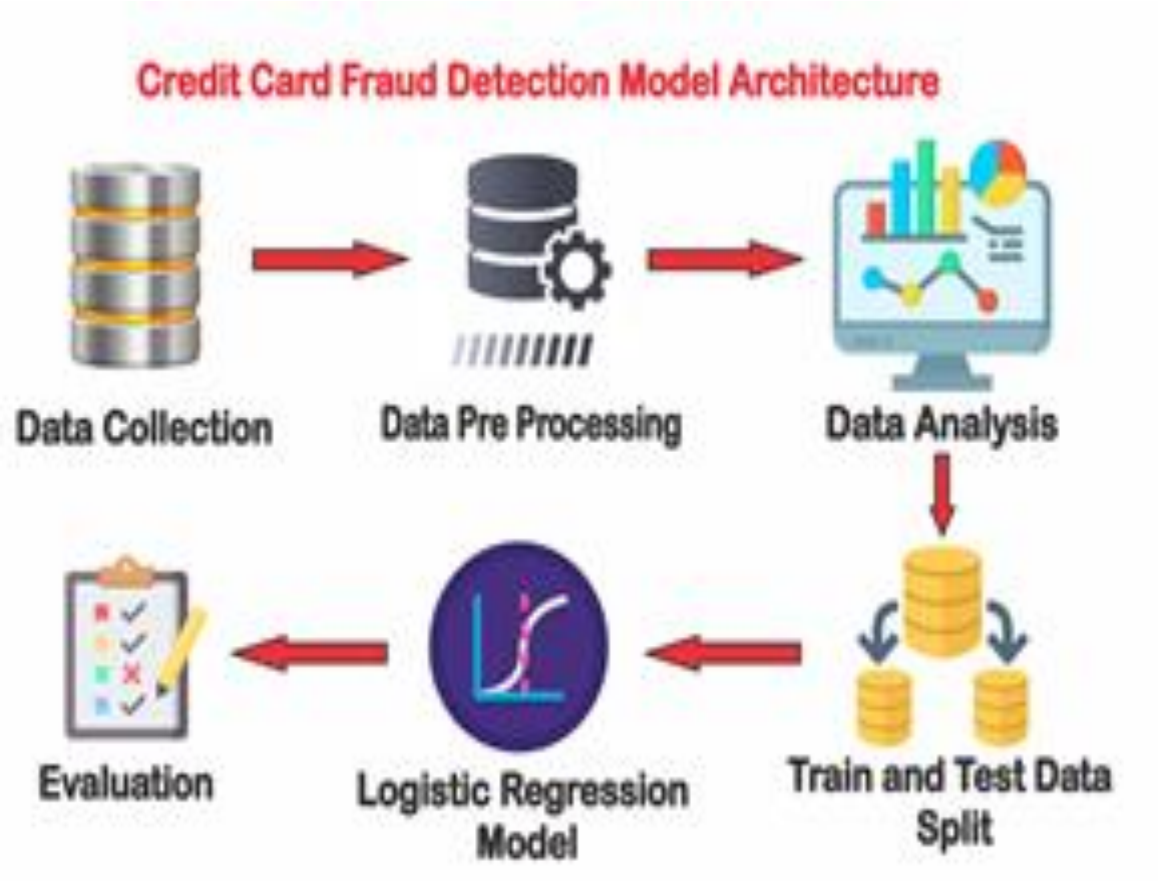

Credit Card Fraud Detection Model Architecture
Data Collection
Data Pre Processing
Data Analysis
Train and Test Data Split
Logistic Regression Model
Evaluation

Fig. 1. Research architecture

b. Data Collection

For this research, the data was sourced from the Kaggle website using Jupyter Notebook in Python. The dataset obtained comprises credit card transactions made by European cardholders in the year 2023. With over 550,000 records, the dataset has undergone anonymization (refer to the process of removing or altering personally identifiable information from data sets, thus making it anonymous) to safeguard the identities of the cardholders. The main purpose of utilizing this dataset is to support the development of algorithms and models aimed at detecting fraudulent transactions effectively. By leveraging this dataset, researchers can explore patterns and characteristics indicative of potential fraud, thus enhancing the accuracy and efficiency of fraud detection mechanisms [18].

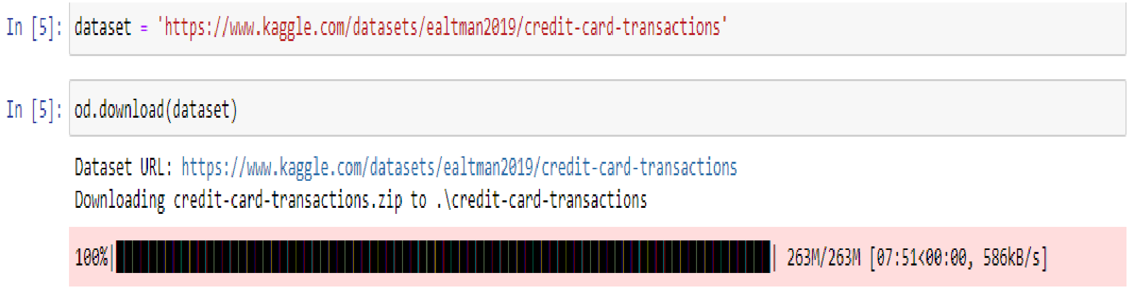
The following screenshot depicts a Jupyter Notebook displaying Python code used to download a dataset from the Kaggle website. Notably, the variable 'dataset' contains the web address where the dataset is located on the Kaggle website. Later, the 'opendatasets' library, imported as 'od', is utilized to directly download the dataset to the specified location.

c. Sampling technique

This research utilizes the Credit Card Transactions Dataset from Kaggle, comprising a comprehensive record of credit card transactions over a specified period, including both genuine and fraudulent transactions. With a total of 568,630 entries, it's essential to balance the dataset, ensuring equal representation of fraudulent and legitimate transactions, resulting in 284,315 instances for each. The data is then divided into four categories: X_train: This contains the features (input variables) used for training the machine learning model. Features could include things like transaction amount, location, time of transaction, etc. Y_train: This contains the corresponding labels or target values for the training set. In this case, it would indicate whether each transaction is fraudulent or not. Typically, 0 might indicate a legitimate transaction, while 1 might indicate a fraudulent one. X_test: This contains a separate set of features that are used for testing the trained model's performance. It's important that the model doesn't see this data during training, as it should evaluate how well it generalizes to unseen data. Y_test: This contains the corresponding labels or target values for the test set. Like Y train, it indicates whether each transaction is fraudulent or not, but it's used for evaluating the model's performance [19].

The Fig. 3 presents the count of legitimate transactions and fraudulent transactions, both totaling 284,315. Each dataset consists of 31 columns.

Likewise, the Fig. 4 illustrates the sampled data for X_train, X_test, Y_train, and Y_test, utilized in training and testing the logistic regression model.



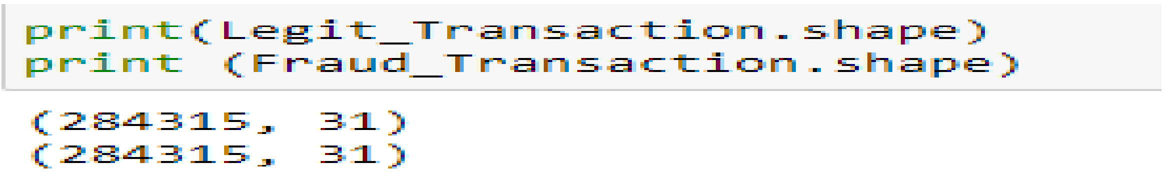
```
In [5]: dataset = 'https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions'

In [5]: od.download(dataset)

        Dataset URL: https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions
        Downloading credit-card-transactions.zip to .\credit-card-transactions

        100%|████████████████████████████████████████| 263M/263M [07:51<00:00, 586kB/s]
```

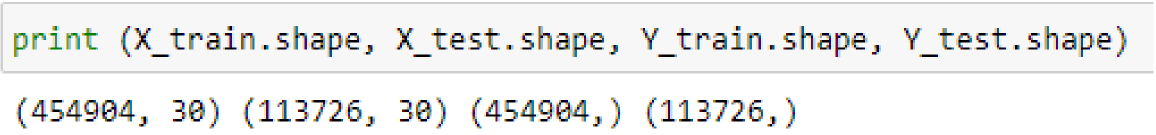
Fig. 2. Downloading the datasets



```
print(Legit_Transaction.shape)
print (Fraud_Transaction.shape)

(284315, 31)
(284315, 31)
```

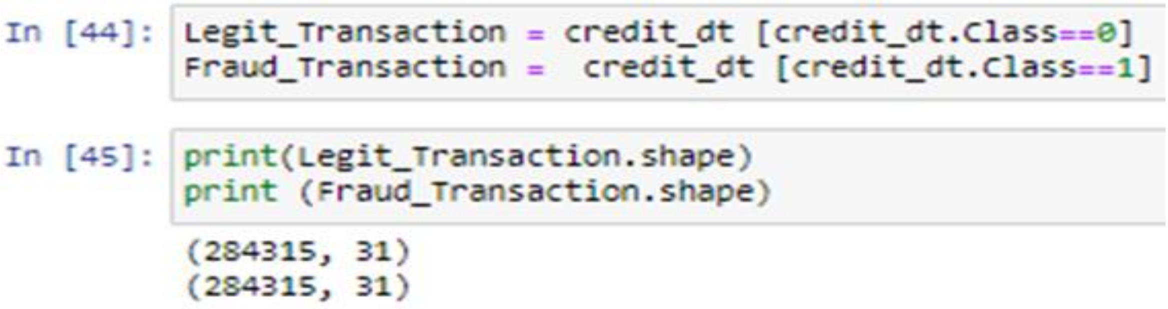
Fig. 3. Legit and fraud transaction sampling



```
print (X_train.shape, X_test.shape, Y_train.shape, Y_test.shape)

(454904, 30) (113726, 30) (454904,) (113726,)
```

Fig. 4. Sampling for training and test datasets

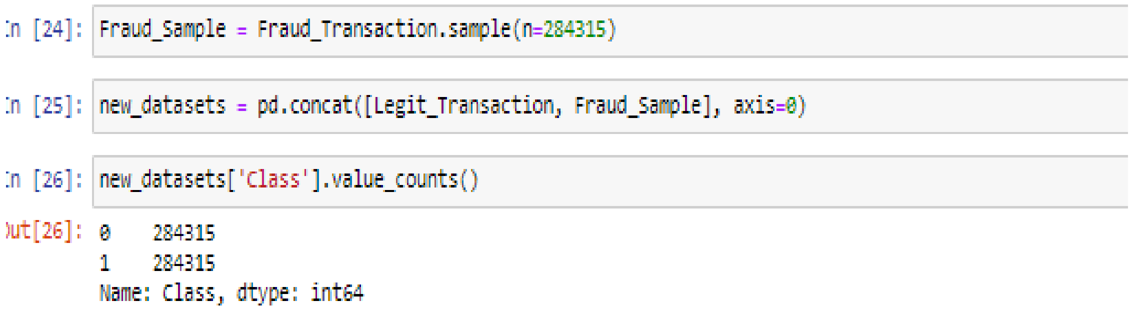


```
In [44]: Legit_Transaction = credit_dt [credit_dt.Class==0]
         Fraud_Transaction =  credit_dt [credit_dt.Class==1]

In [45]: print(Legit_Transaction.shape)
         print (Fraud_Transaction.shape)

         (284315, 31)
         (284315, 31)
```

Fig. 5. Variable for legit and fraud transactions



```
:n [24]: Fraud_Sample = Fraud_Transaction.sample(n=284315)

:n [25]: new_datasets = pd.concat([Legit_Transaction, Fraud_Sample], axis=0)

:n [26]: new_datasets['Class'].value_counts()

)ut[26]: 0    284315
         1    284315
         Name: Class, dtype: int64
```

Fig. 6. New datasets containing equals number of legit and fraud transaction

d. Data Preprocessing

Before analysis and visualization, it is essential to preprocess a dataset to align it with a usable pattern. This includes checking for any null values in rows or columns and ensuring a balance between legitimate and fraudulent transactions [20]. In this study, the dataset underwent the following preprocessing steps before analysis:

To start, the data's first five rows and last five rows are displayed to provide an overview of the dataset, aiding in understanding its structure and other key parameters. The dataset undergoes further scrutiny using the Isnull function to detect any missing values that could potentially compromise the accuracy of the results.

The data is additionally processed by segregating it into two variables: the first variable stores records of legitimate transactions, defined as transactions with a class equal to 0, while the second variable stores fraudulent transactions, identified by a class equal to 1. Following this segmentation, the frequency of each occurrence and its corresponding column number are retrieved.

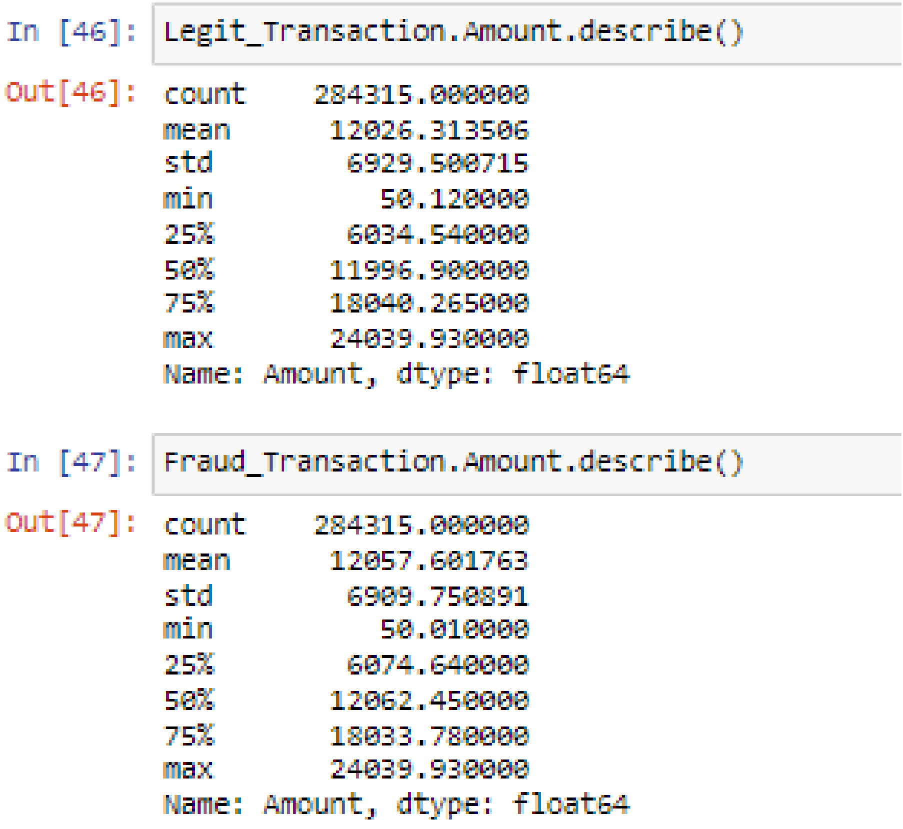
Here, a new variable named "new_dataset" is created to accommodate the concatenated sample of legitimate and fraudulent transactions. This new dataset is utilized from this stage onward for easier access to the datasets. It's evident from the screenshot below that the values of each occurrence in the balance have an equal number, ensuring accurate results.

e. Data Analysis

Data analysis involves the systematic examination and interpretation of data to uncover patterns, trends, relationships, and insights that address research questions or objectives. It plays a crucial role in transforming raw data into meaningful information, facilitating decision-making, hypothesis testing, and drawing conclusions [21].

Eventually, various statistical constraints of the datasets are identified to ensure proper analysis of the data. The Fig. 7 illustrates the mean, count, standard deviation, minimum, and maximum amounts of transactions made for both legitimate and fraudulent transactions.

The dataset further analyzed by grouping data using credit_dt.groupby('Class').mean(). This is to classify data as fraud (Class 1) versus those that are not (Class 0). credit_dt.groupby('Class'): This part groups the data by the 'Class' column, which typically contains binary values indicating whether a transaction is fraudulent or not. So, this groups the data into two groups: one for transactions classified as fraudulent (Class 1) and the other for legitimate transactions (Class 0). After grouping the data, .mean() calculates the mean value for each numerical column within each group. By doing so, it gives you the average values of various features (such as transaction amount, time of transaction, etc.) for both fraudulent and non-fraudulent transactions separately.



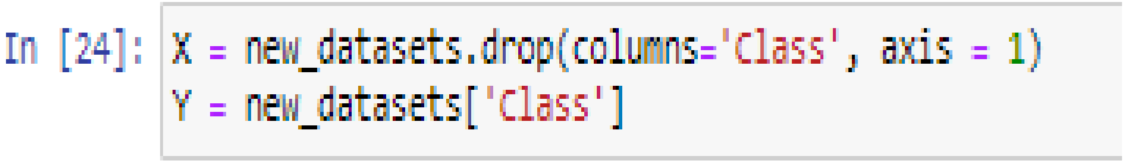
```
In [46]: Legit_Transaction.Amount.describe()

Out[46]: count    284315.000000
         mean      12026.313506
         std        6929.500715
         min          50.120000
         25%        6034.540000
         50%       11996.900000
         75%       18040.265000
         max       24039.930000
         Name: Amount, dtype: float64

In [47]: Fraud_Transaction.Amount.describe()

Out[47]: count    284315.000000
         mean      12057.601763
         std        6909.750891
         min          50.010000
         25%        6074.640000
         50%       12062.450000
         75%       18033.780000
         max       24039.930000
         Name: Amount, dtype: float64
```

Fig. 7. Statistical calculation base on amount of transaction



```
In [24]: X = new_datasets.drop(columns='Class', axis = 1)
         Y = new_datasets['Class']
```

Fig. 8. Variables stores status of transactions and other table schema

This allows us to compare the average values of different features between fraudulent and non-fraudulent transactions. Discrepancies in these averages can sometimes highlight patterns or characteristics that are indicative of fraudulent activity, which can then be used to build better fraud detection models.

To prepare the dataset for logistic regression algorithms, the parameters are divided into two variables. Variable 'X' contains all the figures schemas except for the 'class' attribute, which distinguishes between legitimate and fraudulent transactions. Meanwhile, variable 'Y' exclusively stores the 'class' schema, representing '0' for legitimate transactions and '1' for fraudulent ones. This step is essential as a prerequisite for training the dataset in logistic regression algorithms.

f. Model Training

Modeling using logistic regression involves using the logistic regression algorithm to build a predictive model that can classify transactions as either fraudulent or legitimate based on various features or attributes associated with each transaction [22].

The modeling begins by separating the overall data into training and test data. Below is the overview of the training and the test data;

i. Training Dataset:

The training dataset is a subset of the entire dataset that is used to train the logistic regression model. It consists of historical transaction data, where each transaction is labeled as either fraudulent or legitimate. This dataset is used by the model during the training process to learn the relationship between the input features (e.g., transaction amount, time of transaction, etc.) and the target variable (fraudulent or legitimate) [23].

ii. Test Dataset:

The test dataset is another subset of the entire dataset that is kept separate from the training dataset. It is used to evaluate the performance of the trained logistic regression model. The test dataset also consists of labeled transaction data, but the model has never seen this data during the training process. By evaluating the model's performance on unseen data, we can assess its ability to generalize to new, unseen transactions.

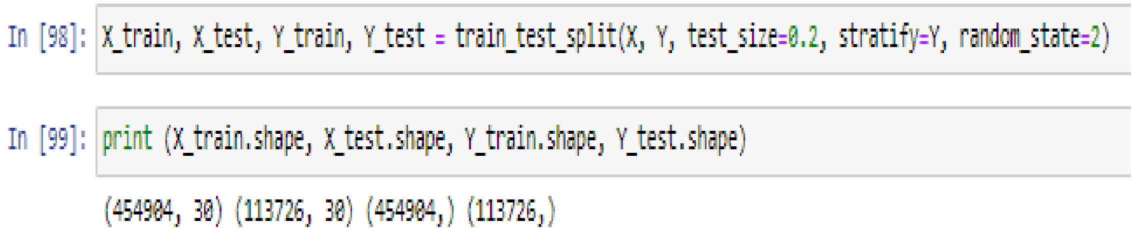
iii. How the training and test datasets works:

Training Phase: During the training phase, the logistic regression model is trained using only the training dataset. The model learns the patterns and relationships in the training data, adjusting its parameters to minimize the prediction error [23].

Evaluation Phase: After training, the model's performance is evaluated using the test dataset. The model makes predictions on the transactions in the test dataset, and these predictions are compared to the true labels (i.e., whether each transaction is fraudulent or legitimate). Evaluation metrics such as accuracy is calculated based on these predictions to assess the model's performance.

By using separate training and test datasets, we can obtain an unbiased estimate of the model's performance on new, unseen data. This helps to ensure that the model is not over fitting to the training data and that it generalizes well to real-world transactions.

The screenshot below illustrates how the overall data is divided into four variables for both the training and test sets. Additionally, the number of data points under each variable is highlighted in the screenshot.



```
In [98]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)

In [99]: print (X_train.shape, X_test.shape, Y_train.shape, Y_test.shape)

         (454904, 30) (113726, 30) (454904,) (113726,)
```

Fig. 9. Splitting dataset; training and test

iv. Employing Logistic Regression Algorithm

Currently, after completing preprocessing steps to clean and organize the data, it is prepared for analysis using logistic regression. In the line of code 'model = LogisticRegression()', we initialize an instance of the logistic regression algorithm, which sets the stage for building a predictive model. This line essentially creates a container named 'model' that holds all the necessary functions and properties of the logistic regression algorithm [22].

Following this initialization, the subsequent line of code from the Fig. 10. signifies the beginning of the training process. Training involves feeding our prepared datasets into the logistic regression algorithm. During this training phase, the algorithm scrutinizes the provided data, examining the features and patterns within each transaction. By adjusting various parameters, such as weights and biases, the algorithm iteratively learns from the dataset, gradually improving its understanding of the relationships between input features and the outcome we're trying to predict.

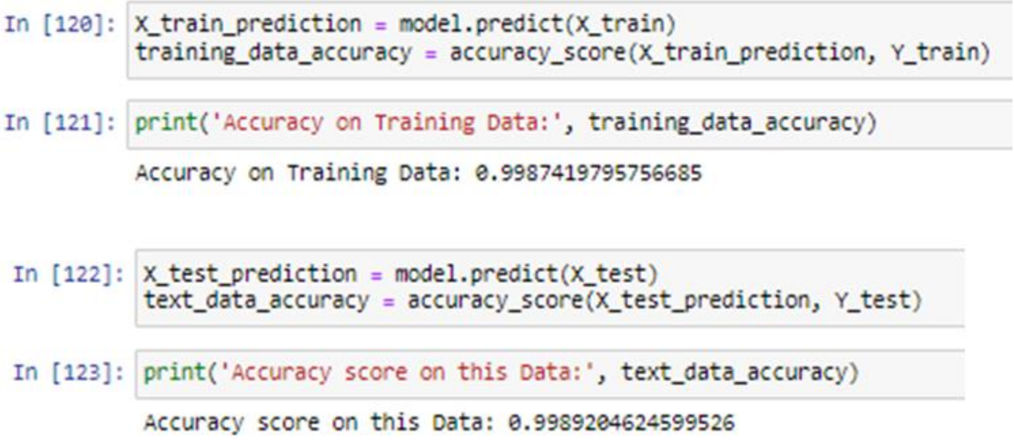
Through this iterative learning process, the algorithm constructs a model that encapsulates the learned relationships between the input variables (features) and the output variable (target). This model serves as a representation of how the algorithm perceives the underlying structure of the data. Ultimately, the goal is to develop a model that accurately predicts the outcome of future transactions based on their features, leveraging the insights gained during the training phase.

v. Evaluation of the Model

In this research, we assess the performance of our developed model through two crucial metrics: accuracy on training data and accuracy on test data [17].

Firstly, we measure the accuracy on the training data by employing the following process: the model is tasked with predicting the outcomes of the X_train dataset, which comprises the input variables used during the training phase. Subsequently, we calculate the accuracy score by comparing these predicted values against the actual outcomes present in Y_train, which encapsulates the corresponding labels or target values for the training set. The resulting accuracy score ranges between 0 and 1, with higher scores indicating a more precise alignment of the model with the training data. Notably, we achieve an accuracy score of approximately 0.999, signifying a high level of success in training the model with the provided data.

Similarly, we evaluate the model's performance on unseen data through the accuracy on the test data. This evaluation entails soliciting predictions from the model for the test data, which it hasn't encountered previously. Remarkably, the model achieves a score of around 0.999 on the test data as well, mirroring the accuracy achieved on the training data. This parity in scores underscores the consistency and reliability of the model's performance. Essentially, the similarity in accuracy scores between the training and test data suggests that the model generalizes well beyond the data it was trained on, exhibiting robust predictive capabilities.



```
In [120]: X_train_prediction = model.predict(X_train)
          training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

In [121]: print('Accuracy on Training Data:', training_data_accuracy)

          Accuracy on Training Data: 0.9987419795756685


 In [122]: X_test_prediction = model.predict(X_test)
           text_data_accuracy = accuracy_score(X_test_prediction, Y_test)

 In [123]: print('Accuracy score on this Data:', text_data_accuracy)

           Accuracy score on this Data: 0.998920462459526
```

Fig. 10. Evaluation

4. RESULTS AND DISCUSSION

The logistic regression model underwent a rigorous training phase, during which it was exposed to a comprehensive dataset containing a wide array of credit card transactions. The primary objective of this training was to equip the model with the capability to discern between legitimate and fraudulent transactions accurately. Following the training phase, the model underwent an extensive evaluation process to assess its performance.

During evaluation, particular attention was paid to the accuracy of the model on both the training and test datasets. This meticulous scrutiny aimed to ensure that the model's performance was consistent across different subsets of data and indicative of its real-world effectiveness.

Upon analyzing the results, it was found that the model exhibited an exceptional accuracy rate of 0.99 when tested on the training data. This high accuracy score underscores the model's proficiency in learning from the provided examples and effectively distinguishing fraudulent transactions from legitimate ones within the training set.

Furthermore, when subjected to an independent test dataset, the model maintained its impressive accuracy, achieving a similar score of 0.99. This remarkable consistency between the accuracies obtained on the training and test datasets is a crucial indicator of the model's robustness in generalizing well to unseen data—a pivotal characteristic for its practical application in real-world scenarios.

The remarkable congruence in high accuracy rates achieved on both training and test datasets serves as compelling evidence of the logistic regression model's effectiveness in accurately identifying fraudulent transactions. This performance is particularly noteworthy when compared to the findings of previous studies, such as [3], where a lower accuracy of 0.94 was reported using the same algorithm.

To provide stakeholders with a succinct and visually accessible representation of the model's performance, a meticulously crafted bar chart was generated. This graphical illustration offers a clear depiction of the model's consistent and reliable accuracy across diverse datasets, reinforcing confidence in its effectiveness for fraud detection purposes.

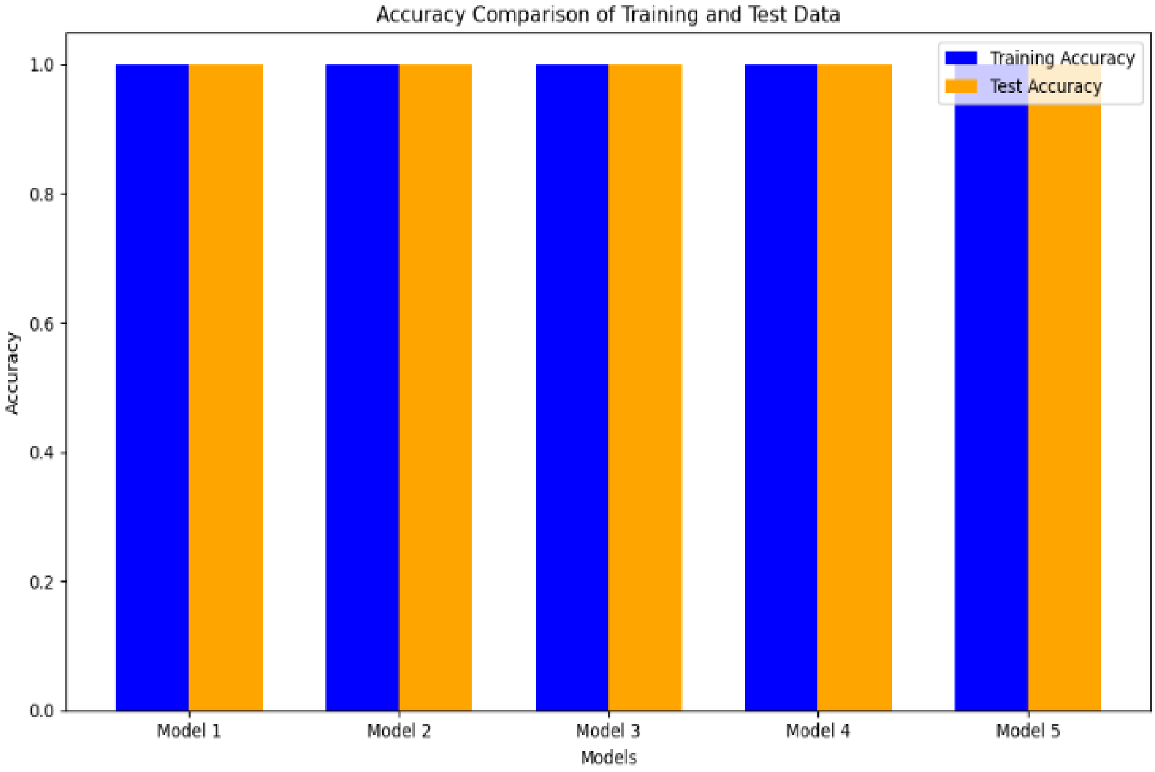


Fig. 11. Accuracy in test and training

4.1 Transaction Status Prediction

In a concerted effort to comprehensively evaluate the efficacy of the logistic regression model, we embarked on an extensive assessment aimed at delving deep into its predictive capabilities, particularly concerning unseen transactions. This thorough evaluation sought to ascertain the model's prowess in accurately discerning the status—whether fraudulent or legitimate—of transactions that had not been encountered during its training phase.

To initiate this evaluation, the logistic regression model was deployed to meticulously analyze a carefully curated dataset consisting exclusively of transactions that had not been previously encountered by the model. Subsequently, the statuses predicted by the model were meticulously compared against the ground truth labels associated with each transaction to precisely gauge the model's predictive accuracy.

The findings of this in-depth evaluation unveiled a commendable performance by the logistic regression model, boasting an impressive accuracy rate of 99.89% in predicting the status of unseen transactions. This remarkable accuracy underscores the model's exceptional ability to navigate through uncharted territory and make accurate predictions. Specifically, the model demonstrated its adeptness by correctly identifying over 55,000 transactions as legitimate, showcasing its proficiency in discerning genuine transactions amidst the noise of the dataset. Moreover, the model exhibited its efficacy in flagging approximately 5,000 transactions as fraudulent, thereby highlighting its capability to identify potentially illicit activities within the dataset with a high degree of precision.

In an effort to offer stakeholders a clear and intuitive understanding of the distribution of predicted transaction statuses, a meticulously crafted bar chart was generated. This graphical representation serves as a visual aid, elucidating the model's predictive prowess by illustrating the distribution of predicted statuses—fraudulent or legitimate—across the unseen dataset. Through this visual depiction, stakeholders can gain insights into the model's predictive performance and its ability to accurately classify transactions, further bolstering confidence in its effectiveness for fraud detection purposes.

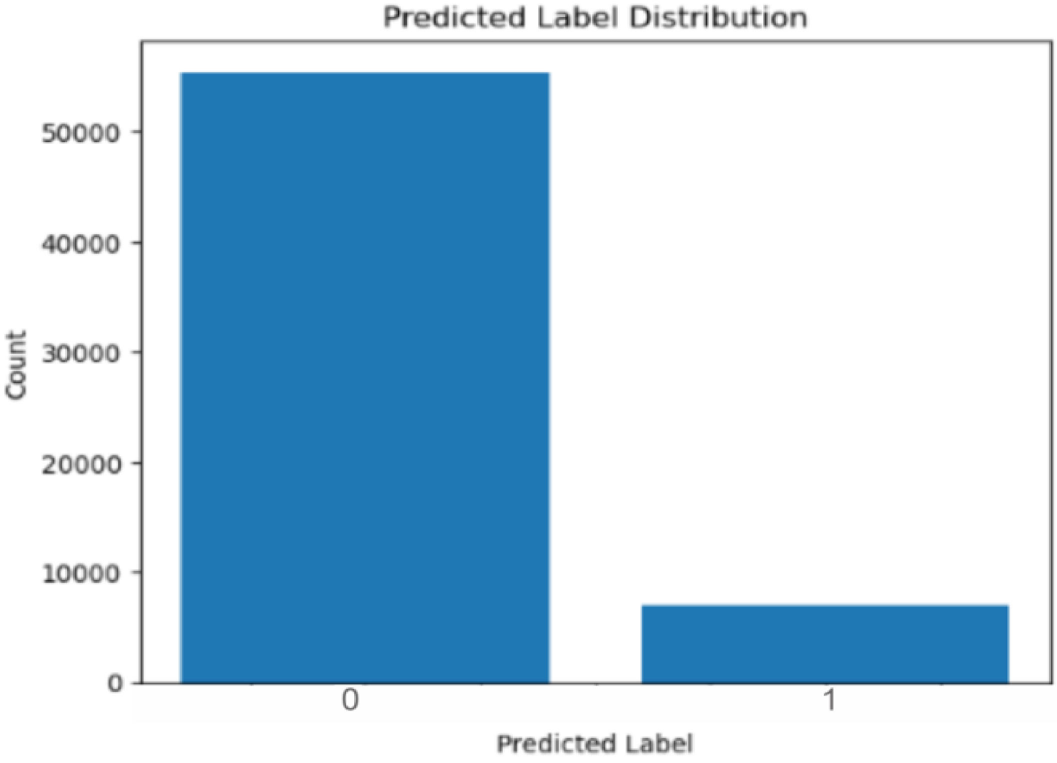


Fig. 12. Result on legit and fraudulent transactions

The obtained results demonstrate the effectiveness of the logistic regression model in detecting credit card fraud transactions. The high accuracies achieved on both the training and test datasets suggest that the model has successfully learned meaningful patterns from the data and can generalize well to unseen transactions.

However, it's important to note that the model's performance may vary depending on the characteristics of the dataset and the features used for training. Further analysis is warranted to identify potential areas for improvement and to assess the robustness of the model across different scenarios.

5. CONCLUSION

In conclusion, the logistic regression model stands as a formidable asset in the continuous endeavor to combat financial fraud. Through its capacity to detect subtle irregularities within credit card transactions and its impressive ability to generalize to unseen data, the model emerges as a cornerstone in the arsenal of fraud detection mechanisms. These findings not only highlight the efficacy of advanced analytical techniques but also underscore the critical importance of fortifying financial ecosystems to preserve the integrity of transactions in an increasingly digitized landscape.

The study's revelations regarding the logistic regression model's proficiency in distinguishing between fraudulent and legitimate transactions align with broader trends in the field of fraud detection. By leveraging sophisticated algorithms and comprehensive datasets, researchers and practitioners alike can enhance their ability to detect and prevent fraudulent activities, thereby safeguarding financial systems and bolstering consumer trust.

Moreover, the model's robust generalization capabilities signify its adaptability to evolving fraud patterns and emerging threats. This adaptability is paramount in an environment characterized by rapid technological advancements and increasingly sophisticated fraudulent schemes.

Looking ahead, further research and development efforts are warranted to continuously refine and improve fraud detection methodologies. Collaborative endeavors between academia, industry, and regulatory bodies can foster innovation and drive the adoption of cutting-edge technologies in the fight against financial fraud.

Ultimately, the findings of this study underscore the pivotal role of the logistic regression model and advanced analytical techniques in fortifying financial ecosystems. By embracing these tools and leveraging data-driven insights, stakeholders can work towards a future where financial transactions are conducted with heightened security and confidence, safeguarding both individual consumers and the broader economy against the pervasive threat of fraud.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Kiruthika S, Sowmyarani CN. Credit Card fraud detection using machine learning and deployment of model in public cloud as a web service. 2020;3878(2): 548–552.
   DOI: 10.35940/ijrte.B3800.079220.
2. Soni KB, Chopade M, Vaghela R. Credit card fraud detection using machine learning approach. 2021;4(2):71–76.
3. Vijaya VKKSKVG, VSA, Pratibha K. Credit card fraud detection using machine learning algorithms. 2020;9(7):1526–1530.
4. Purohit N, Vishwakarma RG. Credit card fraud detection using machine learning algorithms using python technology. 2021;18(6):7995–8006.
5. Alali A, Alali A. Financial fraud detection using machine learning techniques; 2020.
6. Journal C. COOU journal of physical s ciences. 2020;3(1)L493–498.
7. Dekou R. Machine learning methods for detecting fraud in online marketplaces . 2021;1–8.
8. Swarna BP, JAPSS. Credit card fraud detection system using machine learning. 2021;8(7):249–252.
9. Devi RR. Credit card fraud detection using AI / ML / CNN,. 2023;6(9):242–249.

10. Nikhil K, Maharshi BV, Tanooj K. Credit card fraud detection using machine learning algorithms. 2023;14(4):471–485.
11. Shah D, Sharma LK. Credit card fraud detection using decision tree and random forest. 2023;02012.
12. Sandhya G, Abishek M, Kumar SG, RSJ. Kumar. Credit card fraud detection using machine learning algorithms credit card fraud detection using; 2023.
DOI: 10.1007/978-981-19-5221-0
13. Sammelwerksbeitrag PP. Scraping social media data as platform research : A data hermeneutical perspective; 2023.
Available: www.ssoar.info Scraping.
14. Eswari MGS, Malleswari MAS, Sakina MS, Anjali MK, Sindhu MDNS, Visha MJR. International Journal of Engineering Technology Research & Management International Journal of Engineering Technology Research & Management. 02, 2022;9–12.
15. Abuhamda EAA, Ismail IA, English T. Understanding quantitative and qualitative research methods : A theoretical perspective for young researchers understanding quantitativeand qualitative researchmethods : A Theoretical Perspective for Young Researchers. no. February. 2021;70–87.
DOI: 10.2501/ijmr-201-5-070.
16. Hayashi T. Shimizu T, Fukami Y. Collaborative problem solving on a Data Platform. 2021;120(362):37–40.
17. Chowdari B, Parthiban S, International Journal of research publication and reviews credit card fraud detection using logistic regression compared with t-sne to improve accuracy. 2022;3(8):1000–1004.
18. Ramaswamy A, Mulimani ADA, Pal S, Singh AK, HG. Rani, credit card fraud detection using machine learning and deep learning. 2021;775–777.
19. Alenzi HZ, Aljehane NO. Fraud detection in credit cards using logistic regression. 2020;11(12):540–551.
20. Agarwal V. Research on data preprocessing and categorization technique for smartphone review analysis; 2015, 2016.
DOI: 10.5120/ijca2015907309
21. Lakshmi JVN. Machine learning techniques using python for data analysis in performance evaluation Machine learning techniques using python for data analysis in performance evaluation; 2018, 2019,
DOI: 10.1504/IJISTA.2018.10012853
22. Alemad M. Credit card fraud detection using machine learning; 2022.
23. Tech V, Virginia A. Systematic training and testing for machine learning using combinatorial interaction testing a national; 2022.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/118279