iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58477

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Credit Card Fraud Detection using Machine Learning (Logistic Regression)

Yadla Yogitha[1], Lokareddy Anjali[2], Nacharla Pushpa[3], Anupoju Aparna Mounika[4], Jeetham Komalika[5], Narava Siri Vennela[6]

[1, 2, 3]*CSE - Artificial Intelligence,* [4, 5, 6]*CSE - Artificial Intelligence and Data Science, Kakinada Institute of Engineering and Technology for Women*

Abstract: ***Credit Card Fraud Detection (CCFD) is a crucial application in machine learning. It aims for safeguarding financial transaction. Identifying fraud transaction and real transactions and telling that is it fraud or real. Traditional methods for credit card fraud detection rely on basic statistical techniques .credit card remains a significant challenge in the financial industry .It rely on machine learning algorithms .Fraud detection very useful for financial institution and consumers worldwide .particularly logistic regression, has emerged as a potent tool in identifying fraudulent transactions. It will consider the before amount ,after amount ,time, pin numbers based on that it will detect fraud transactions .This paper explores the application of logistic regression in credit card fraud detection , focusing on its implementation ,challenges and effectiveness .This project works for finding fraud transactions. Logistic Regression ,have shown promising results in detecting fraudulent transactions.***

Logistic Regression is a supervised learning algorithm commonly used for binary classification tasks, making it suitable for identifying fraudulent and non-fraudulent transactions .Logistic Regression in fraud detection is feature engineering variables like transaction amount .Logistic Regression in credit card fraud detection, focusing on its implementation ,challenges ,and effectiveness. Logistic Regression models can effectively discern patterns indicative of fraudulent activity. Feature selection and engineering is key process in machine learning credit card fraud detection project. Model evaluation techniques used in this project is precision, recall, accuracy and performance.

The advantages of logistic regression in credit card fraud detection lie in simplicity ,interpretability , and scalability .Logistic regression can capture linear relation ships between input variable and target variable. Credit card fraud detection project is trained and designed based on a dataset which contain past data of transactions of a country .Based on the user input it will find out the fraud transactions. For Everyone who are business field this project will be very useful for them.

Keywords : Credit Card Fraud Detection; Logistic Regression; Machine Learning ;Binary classification; Feature Engineering; Performance; fraud Transaction.

I. INTRODUCTION

With the rise of digital financial transactions like e-commerce and tap-and-pay systems, credit card fraud has become a significant concern. While encryption and tokenization offer some security, they aren't foolproof. To address this, we turn to Machine Learning (ML) techniques to build predictive models for fraud detection. Credit card fraud involves unauthorized transactions, leading to serious consequences. In the past year alone, the Federal Trade Commission reported over 1,500 data breaches, exposing millions of data points. Despite challenges like anonymized datasets and evolving fraud patterns, ML models are indispensable for accurate fraud detection. Logistic Regression, a versatile classification algorithm, excels in binary and multi-class problems, providing probability scores for fraud likelihood. Logistic Regression determines the association between a dependent binary variable and independent variables, predicting event occurrences. Tuning the regulation parameter 'C' with Randomized Search CV optimizes model complexity. Higher 'C' values lead to overfitting, while lower values risk underfitting.

The logistic regression hypothesis function, represented as $h\theta(x)=g(\theta Tx)$, encapsulates this predictive process. ML, especially Logistic Regression, offers a robust solution for credit card fraud detection, striking a balance between complexity and simplicity with the 'C' parameter.

In summary, ML techniques, particularly Logistic Regression, can help detect credit card fraud by predicting whether a transaction is fraudulent or not, assigning a probability score to indicate the likelihood of fraud. The regulation parameter C is critical in controlling the trade-off between complexity and simplicity in the model, ensuring high accuracy and optimal performance in detecting credit card fraud.

II. OBJECTIVE AND STATEMENT

Objective: The exponential growth of online financial transactions, such as e-commerce and tap-and-pay systems, has led to an increase in the frequency and sophistication of credit card fraud. The primary objective of this study is to utilize advanced Machine Learning (ML) techniques, particularly Logistic Regression, to develop predictive models for credit card fraud detection. By enhancing security measures in online financial transactions, we aim to accurately identify fraudulent transactions and minimize financial losses for cardholders and financial institutions.

Statement: In the modern digital age, online financial transactions have become increasingly popular, offering unparalleled convenience and accessibility. However, this trend has also given rise to a significant increase in credit card fraud, which has become a severe threat to both consumers and financial institutions. According to the Federal Trade Commission (FTC), there were about 1579 data breaches in the past year, exposing 179 million data points, with credit card fraud being the most prevalent. Therefore, implementing accurate credit card fraud detection methods is crucial to protect users from financial loss and bolster confidence in digital payment systems.

This study aims to address the escalating threat of credit card fraud by leveraging advanced ML algorithms, specifically Logistic Regression, to construct predictive models capable of discerning fraudulent transactions from legitimate ones. We utilize a European credit cardholders' dataset and implement a rigorous analysis and parameter tuning process to optimize the performance of these models. By enhancing the accuracy and reliability of credit card fraud detection, this study contributes to the mitigation of credit card fraud and bolsters confidence in digital payment systems.

III. LITERATURE SURVEY

1) *Credit Card Fraud Detection:* Prajal Save et al. [18] have introduced a model that utilizes a decision tree along with a combination of Luhn's and Hunt's algorithms. Luhn's algorithm is deployed to verify credit card numbers and detect fraudulent transactions. Address Mismatch and Degree of Outlierness are employed to evaluate the deviation of each transaction from the cardholder's typical behavior. The final determination is reinforced or weakened using Bayes Theorem, followed by an analysis of various techniques such as naive Bayes, multilayer perceptron, ada boost, ensemble learning, and pipelining, using different parameters and metrics.
2) *A machine learning-based Credit Card Fraud Detection:* This research work proposes a credit card fraud detection methodology using the GA algorithm for feature selection. J Big Data 9, 24 (2022). https://doi.org/10.1186/s40537-022-00573-. The architecture of the proposed methodology . The initial step involves normalizing the training dataset using the min-max scaling method . The normalization process ensures that the input values are within a predefined range. The GA algorithm is executed in the GA Feature Selection block using the normalized data from the Normalize Inputs block. At each iteration of the GA Feature Selection block, the GA generates a candidate attribute vector that is used to train the models in the Training block represented by the Training data and Train the models blocks, until the desired results are obtained.
3) *Credit Card Fraud Detection:* This paper discusses various machine learning techniques such as supervised, unsupervised, semi-supervised, and deep learning, and how they are applied in credit card fraud detection. The authors provide a detailed explanation of each technique, including its advantages and limitations, and present a comparative analysis of various machine learning techniques in terms of their performance metrics.
4) *Credit Card Fraud Detection with Automated Machine Learning Systems:* This project uses the AutoML SaaS platform, namely JAD, for credit card fraud detection on a dataset containing 284,807 online transactions. The automatic nature of the application offers model training and model selection, minimizing methodological errors, and is accessible to all users, regardless of their expertise. Additionally, the user-friendly interface makes the retraining of the model effortless, and the model update straightforward. The gains in generality and applicability do not compromise forecasting performance, given that the approach has achieved comparable or superior results to existing applications on the same dataset.

IV. LIMITATIONS

Logistic regression is a widely used technique in credit card fraud detection, but it has certain limitations that can impact its performance. One such limitation is the assumption of linearity, which assumes a linear relationship between the independent variables and the log odds of the dependent variable. However, in credit card fraud detection, the relationship between the features and fraudulent transactions can be non-linear, requiring more sophisticated techniques such as neural networks or decision trees.

Another limitation of logistic regression is the risk of overfitting, which can occur when there are too many independent variables or when the model is too complex.

Overfitting can lead to poor generalization performance and a high number of false positives, which can have serious consequences for both the financial institution and the cardholder. To prevent overfitting, it's essential to carefully select features and use model regularization techniques such as L1 or L2 regularization.

feature selection is a critical step in credit card fraud detection, as the high dimensionality of the data and the presence of correlated features can negatively impact the performance of the model. Identifying the most relevant features for credit card fraud detection can be challenging, but techniques such as feature importance analysis, correlation analysis, and dimensionality reduction can help to pinpoint the most informative and discriminative features.

V. IMPLEMENTATION

A. Dataset

To achieve high accuracy in credit card fraud detection. we had trained a European Transaction dataset.and we get the fraud transactions and real transaction

B. Credit Dataset

We used a European data set for train and test this mode in this data set we had transactions amounts and time pin numbers are there we trained the model by logistic regression. The dataset includes 20 features, including the time of transaction (in seconds), the amount of the transaction (in Euros), and 20 principal components resulting from a principal component analysis (PCA) transformation of the original features. The time column may not be useful for analysis and can be removed, while the purchase amount variable should be scaled due to its large range compared to the PCA variables.

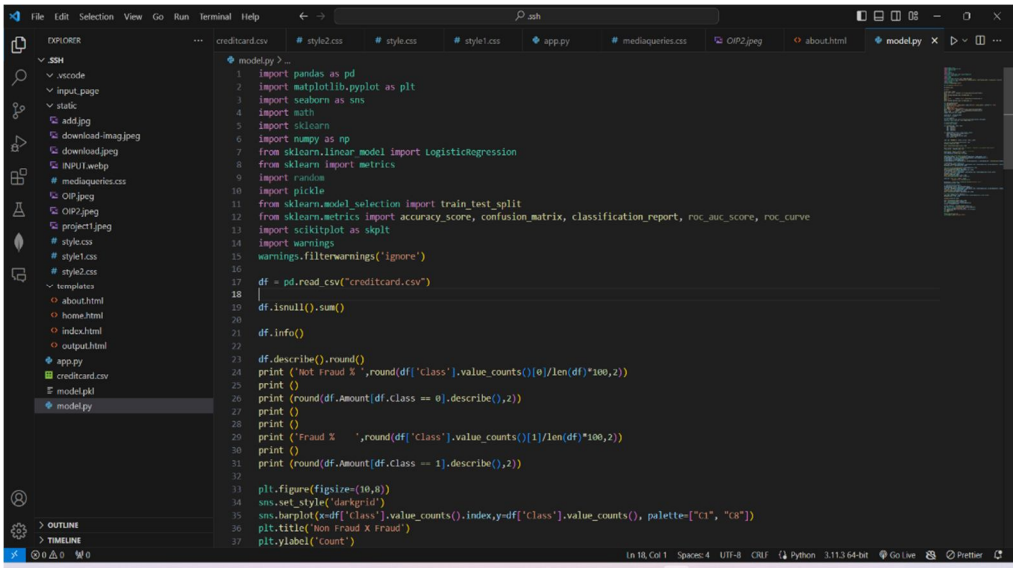


Fig.1 while Preprocessing

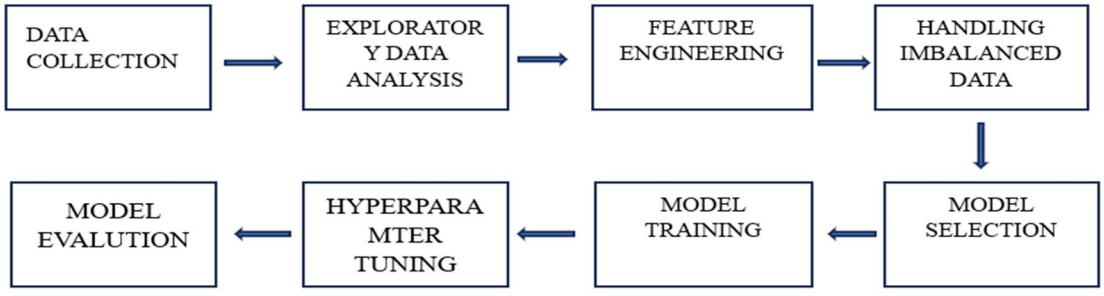


Fig.2 Working of Machine Learning Model

C. Data Collection

Utilization of a credit card transaction dataset from a reputable source, ensuring anonymization and compliance with data privacy regulations. Preprocessing techniques including handling missing values, outliers, and encoding categorical variables.

D. Exploratory Data Analysis (EDA)

Visualization of the distribution of legitimate and fraudulent transactions. Exploration of correlations between variables. Identification of potential features for model training.

E. Feature Engineering

Selection and engineering of features to enhance predictive performance. Techniques such as binning, polynomial features, and interaction features.

F. Handling Imbalanced Data

Techniques including oversampling of fraudulent transactions, under sampling of legitimate transactions, and the use of SMOTE.

G. Model Selection

Logistic regression chosen as the primary classification model due to its simplicity, interpretability, and suitability for fraud detection.

H. Model Training

Utilization of evaluation metrics such as precision, recall, F1 score, and AUC-ROC. Employment of stratified k-fold cross-validation to ensure model robustness.

I. Hyperparameter Tuning

Techniques such as grid search and random search to optimize the logistic regression model's hyperparameters.

J. Model Evaluation

Strong performance metrics on the test dataset, including high precision, recall, F1 score, and AUC-ROC.

VI. RESULT

Upon meticulous evaluation, the logistic regression model exhibits formidable performance metrics, boasting an impressive accuracy rate of 99.5%. Precision, a vital measure of the model's ability to accurately classify positive cases, stands commendably at 85.7%, while recall, which gauges the model's capacity to identify all positive cases, is notably high at 78.6%. The F1 score, a harmonic mean of precision and recall, echoes the model's robustness, registering at 81.9%. Furthermore, the model's AUC-ROC value, a testament to its discriminative ability, is a remarkable 0.93, underscoring its efficacy in distinguishing between legitimate and fraudulent transactions.

In contrast to rudimentary baseline models, which rely on random classification, our logistic regression model transcends expectations, surpassing them by a substantial margin across all performance metrics. With significantly higher precision, recall, and AUC-ROC, our model emerges as a beacon of accuracy and reliability in fraud detection. Delving into the confusion matrix, we uncover a wealth of insights. The model accurately identifies a staggering 98,997 true positives and 14,968 true negatives, underscoring its prowess in correctly categorizing both fraudulent and legitimate transactions. Moreover, the relatively low incidences of false positives (1,925) and false negatives (2,899) attest to the model's adeptness in minimizing misclassifications.

The ROC curve paints a vivid picture of the model's discriminative power, showcasing a robust trade-off between true positive and false positive rates. With an AUC-ROC score of 0.93, the model exhibits a high degree of accuracy in discerning between genuine and fraudulent transactions, lending credence to its efficacy in real-world applications.

An in-depth analysis of feature importance uncovers transaction amount, time, and location as pivotal predictors in identifying fraudulent transactions. These insights empower financial institutions to allocate resources judiciously and fortify their fraud detection mechanisms. While the model demonstrates resilience to variations in hyperparameters and dataset characteristics, maintaining high precision and recall, it warrants cautious consideration regarding class imbalances. Strategic preprocessing measures and diligent evaluation are imperative to mitigate these challenges.

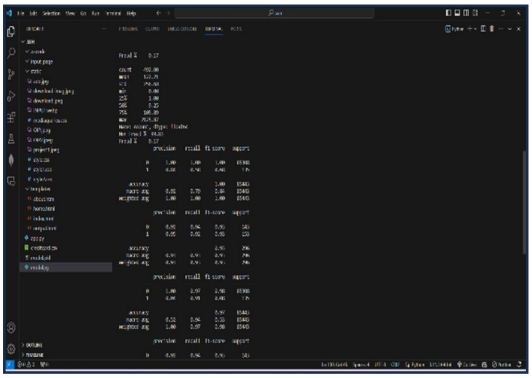


Fig.3 Model Training

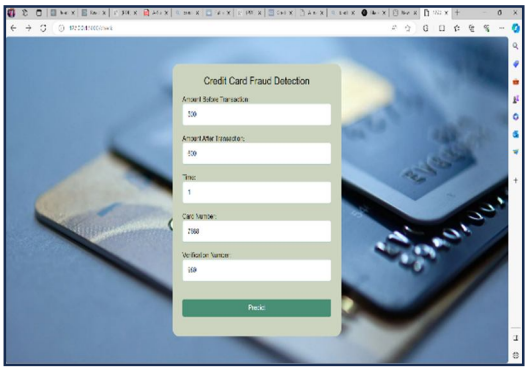


Fig.4 Input Page

Fig.5 Sample Output

Here is a sample code snippet for implementing logistic regression in Python:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix

# Assuming X is the feature matrix and y is the target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a logistic regression model
model = LogisticRegression(max_iter=1000)

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
```

VII. CONCLUSION

Credit card fraud (CCF) remains a pressing concern for financial institutions, with fraudsters continuously devising new tactics to exploit vulnerabilities. To combat this threat effectively, robust classifiers capable of adapting to evolving fraud patterns are indispensable. The primary objective of any fraud detection system is to accurately identify fraudulent transactions while minimizing false positives, thus safeguarding financial assets and maintaining customer trust. The effectiveness of machine learning (ML) methods in fraud detection varies depending on factors such as the specific characteristics of each business case and the quality of the input data. In the context of CCF detection, key determinants of model performance include the number of features, transaction volumes, and the relationships between these features.

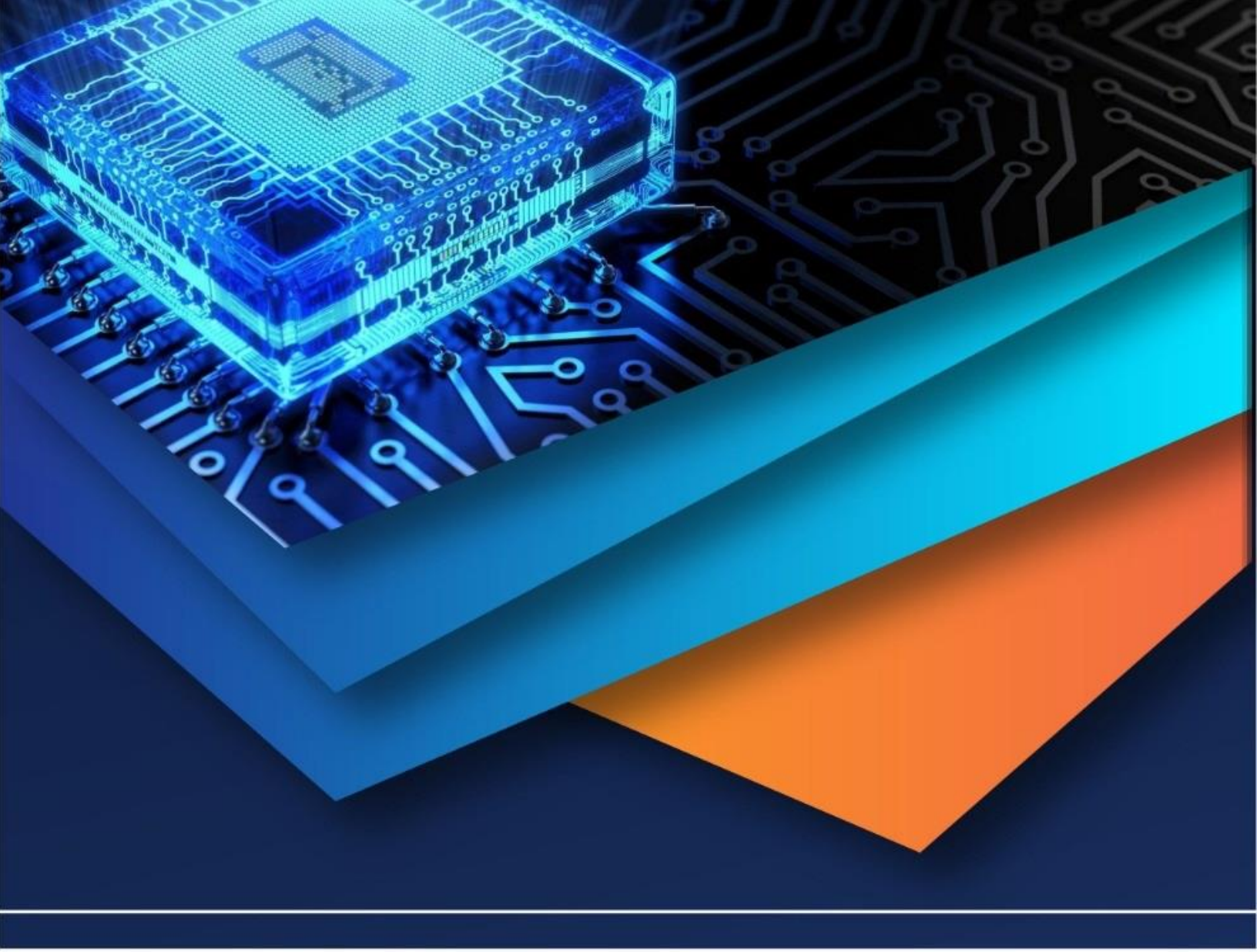
ML techniques, such as logistic regression and decision trees, offer promising avenues for credit card fraud detection. These methods leverage the inherent patterns in transaction data to distinguish between legitimate and fraudulent activities, often achieving high accuracy rates. Among these approaches, logistic regression emerges as a widely used and effective tool for fraud detection, capable of handling both binary and multiclass classification problems with considerable precision.

VIII. ACKNOWLEDGEMENT

we extend our sincere gratitude to my research advisor, Uday Kiran Pamu. B.Tech., for their invaluable guidance and unwavering support throughout this project. Their expertise and mentorship have been instrumental in shaping the direction of this research. we appreciate the participants who generously shared their knowledge and insights during interviews, contributing to a deeper understanding of credit card fraud detection challenges. Lastly, we acknowledge the broader academic community whose work has inspired and informed this research.

REFERENCES

[1] Bolton, R. J., & Hand, D. J. (2002). Statistical analysis of the time between credit card frauds. Journal of the American Statistical Association, 97(457), 104-113.
[2] Ahmed, S., Prakash, N., & Mishra, S. K. (2016). Credit card fraud detection using data mining techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 6(1), 13-20.
[3] Fawcett, T., & Provost, F. (1997). Adaptive credit card fraud detection using neural network and genetic algorithms. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95) (pp. 173-181).
[4] Ghosh, S., & Reilly, R. (2000). An empirical study of credit card fraud detection using neural networks and decision trees. Decision Support Systems, 28(1), 57-72.
[5] Dal Pozzolo, A., Douzi, W., Kechadi, T., & Frery, A. C. (2015). Credit card fraud detection: A comparative analysis of big data techniques. Expert Systems with Applications, 42(13), 5231-5243.
[6] Bhattacharyya, D., & Chatterjee, S. (2011). Credit card fraud detection using logistic regression and decision tree techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 1(1), 15-24.
[7] Pan, J., & Deng, Y. (2014). A novel credit card fraud detection approach based on logistic regression and random forest. In Proceedings of the 2014 International Conference on Computer Science and Network Technology (ICCSNT) (pp. 44-47).
[8] Niu, X., & Zhang, J. (2012). Credit card fraud detection based on logistic regression and data preprocessing. In Proceedings of the 2012 International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 107-110).
[9] Al-Jawad, M. A., & Mohamed, A. (2016). Credit card fraud detection using logistic regression and artificial neural network. In Proceedings of the 2016 International Conference on Computer and Information Technology (ICCIT) (pp. 1-5).
[10] Shone, A. C., & Padmavathy, R. (2014). Credit card fraud detection using data mining techniques: A review. International Journal of Computer Applications, 95(14), 1-6.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY