A
Mini-Project Report on

Online Payment Fraud Detection Using Python

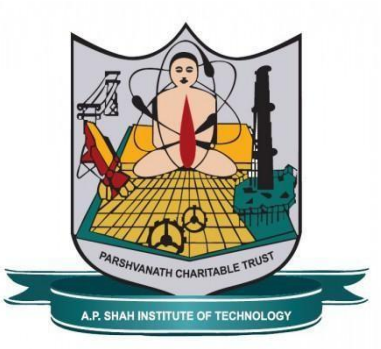
Submitted in partial fulfillment of the requirements
for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
(Artificial Intelligence & Machine Learning)

by

Pratiksha Pathak (22106127)
Maitreyi Phadke (22106007)
Arya Raut (22106075)
Chinmay Sawant (22106017)

Under the guidance of

Prof. Sayali Badhan



Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2023-2024

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**Online Payment Fraud Detection Using Python"** is a bonafide work of Pratiksha Pathak (22106127), Maitreyi Phadke (22106007) ,Arya Raut (22106075), Chinmay Sawant (22106017) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Sayali Badhan	Dr. Jaya Gupta
Mini Project Guide	Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**Online Payment Fraud Detection Using Python"** by **Pratiksha Pathak, Maitreyi Phadke, Arya Raut and Chinmay Sawant** is approved for the degree of **Bachelor of Engineering** in **Computer Science & Engineering (AI &ML)**, **2023-24**.

External Examiner: _____

Internal Examiner: _____

Place: APSIT, Thane
Date:

Declaration

We declare that this written submission represents my ideas in my own words and whereothers' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Pratiksha Pathak (22106127)	Maitreyi Phadke (22106007)	Arya Raut (22106075)	Chinmay Sawant (22106017)

ABSTRACT

With the increasing prevalence of online transactions, the threat of fraudulent activities in online payment systems has become a significant concern for both businesses and consumers. In this study, we propose a novel approach for online payment fraud detection utilizing logistic regression analysis. Logistic regression is a statistical modeling technique commonly used for binary classification tasks, such as fraud detection. It estimates the probability that a given input belongs to a particular class (e.g., fraudulent or legitimate) based on one or more independent variables. We employ logistic regression to build a predictive model that can classify transactions as either fraudulent or legitimate based on a set of relevant features. We utilize a comprehensive dataset containing transactional information, including user behavior patterns, transaction amounts, device information, and other relevant attributes. Through extensive experimentation and evaluation, we demonstrate the effectiveness of our logistic regression-based approach in accurately identifying fraudulent transactions while minimizing false positives. Our results indicate that logistic regression can serve as a valuable tool for online payment fraud detection, providing a reliable and efficient means of safeguarding against fraudulent activities in online payment systems. People rely on online transactions for nearly everything in today's environment. Online transactions offer several benefits, such as ease of use, viability, speedier payments, etc., but they also have some drawbacks, such as fraud, phishing, data loss, etc. As online transactions grow, there is a continuing risk of frauds and deceptive transactions that could violate a person's privacy. In order to prevent high risk transactions, numerous commercial banks and insurance firms invested millions of rupees in the development of transaction detection systems. This research study has introduced a feature-engineered machine learning-based model for detecting transaction fraud.

By processing as much data as it can, the algorithm can gain experience, strengthen its stability, and increase its performance. The effort to detect online fraud transactions can use these algorithms. In this, a dataset of specific online transactions is obtained. Then, with the aid of machine learning algorithms, unusual or distinctive data patterns that will be helpful in identifying any transactions that are fraudulent are discovered. Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that an instance belongs to one of two classes.

Keywords: Fraud Detection, Logistic Regression

Index

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

So far, mobile payment has become one of the mainstream payment methods. Thousands of transactions are carried out on the online trading platform all the time. The popularity of network transactions provides some criminals with the opportunity to commit crimes. Personal property in the complex network environment has the risk of theft, which not only damages the interests of consumers, but also seriously affects the healthy development of the network economy. Therefore, the transaction fraud detection is one of the key tools to solve the problem of network transaction fraud. Traditional fraud detection mostly adopts statistical and multi-dimensional analysis techniques. Since they are verification techniques, it is difficult to obtain the laws hidden behind the transaction data. The big data technology and machine learning algorithm provide efficient detection methods for transaction fraud detection. Compared to the traditional statistical methods, machine learning can represent important features through a large amount of data, which cannot be described by the former. By using the corresponding machine learning method, we can establish a model based on the existing transaction data to realize the detection of network transaction fraud, so as to reduce the loss caused by fraud. In 2018, Zhaohui Zhang proposed a reconstructed feature convolutional neural network prediction model applied to transaction fraud detection, which has better stability and availability in classification effect compared with other convolutional neural network models. However, there is also a problem that the detection accuracy is not high enough due to the imbalance of sample labels. Combined with requirements, we first compared the Fully Connected Neural Network and Logistic Regression algorithm. The former algorithm integrated two neural network models with different cross entropy loss functions, and the design process of the integrated model is quick and convenient. The two algorithms have different application scenarios. In order to ensure the good detection performance, we decided to use Logistic Regression model to build an online transaction fraud detection system. This system has obvious advantages in running time and applicability, and can accurately predict the fraud probability of network transaction behaviors.

Transaction can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card-issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users to estimate, perceive or avoid objectionable behavior, which consists of fraud, intrusion, and defaulting. Most of the time, a person who has become a victim of such fraud doesn't have idea about it until the very end.

Necessary preventive measures can be taken to stop this abuse and the behavior of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, this is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over time. These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyze all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time. So, in this project, what we have tried is to create a Web App for the detection of such types of frauds with the help of Machine Learning.

CHAPTER 2

LITERATURE SURVEY

2.LITERATURE SURVEY

2.1 HISTORY

A literature survey on online transaction fraud detection systems based on machine learning reveals a rich history of research and development in this domain. Over the years, researchers and practitioners have explored various techniques and methodologies to enhance the accuracy and efficiency of fraud detection systems. Here's an overview of the key developments in the history of online transaction fraud detection:

Traditional Approaches (Pre-2000s):

Early fraud detection systems primarily relied on rule-based methods and manual reviews.

Simple heuristics and predefined rules were used to identify potentially fraudulent transactions.

These approaches often struggled to adapt to evolving fraud patterns and lacked the ability to handle complex, dynamic scenarios.

Introduction of Machine Learning (2000s):

The 2000s saw a shift towards machine learning techniques for fraud detection. Decision trees, neural networks, and Bayesian methods were among the first ML algorithms applied to detect anomalies in transaction data. Feature engineering became crucial for extracting relevant information from transaction datasets.

Anomaly Detection (2010s):

Anomaly detection gained prominence in fraud detection during the 2010s.Unsupervised learning techniques, such as clustering and outlier detection, were explored to identify unusual patterns in transaction behavior. Ensemble methods, like Random Forests and Gradient Boosting, were employed for improved accuracy.

Deep Learning and Neural Networks (2010s - Present):

The rise of deep learning, especially neural networks, brought significant advancements to fraud detection. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks demonstrated success in capturing temporal dependencies in transaction sequences.

Transfer learning and pre-trained models were adapted to fraud detection, allowing models to learn from diverse datasets.

Feature Engineering and Data Preprocessing (2010s - Present):

Researchers focused on developing sophisticated feature engineering techniques to enhance the representation of transaction data. Feature scaling, normalization, and dimensionality reduction

methods played a crucial role in preparing data for machine learning models.

Real-time Fraud Detection (2010s - Present):

With the increasing volume of online transactions, the emphasis shifted towards real-time fraud detection. Stream processing frameworks and online learning algorithms were explored to process transactions as they occurred.

Explainability and Interpretability (2010s - Present):

As machine learning models became more complex, there was a growing need for interpretability in fraud detection systems. Explainable AI (XAI) techniques, such as SHAP (Shapley Additive Explanations), were applied to make the decision-making process of models more transparent.

Challenges and Future Directions: Despite significant progress, challenges like imbalanced datasets, adversarial attacks, and evolving fraud patterns persist. Ongoing research focuses on addressing these challenges and exploring emerging technologies, such as federated learning and blockchain, for improved security.

In conclusion, the history of online transaction fraud detection based on machine learning reflects a continual evolution from rule-based methods to sophisticated, data-driven approaches. The field continues to advance, driven by the need for robust, adaptive systems that can effectively combat emerging threats in the digital landscape.

2.2 LITERATURE REVIEW

The literature review on online transaction fraud detection systems based on machine learning highlights the evolution of techniques and methodologies employed to enhance the accuracy and efficiency of fraud detection in the realm of online transactions. Researchers have extensively explored various machine learning algorithms, data preprocessing techniques, and real-time processing approaches. The following is a comprehensive literature review, summarizing key findings and trends in this field :

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression.

This book provides a comprehensive overview of logistic regression analysis and its practical applications. It covers topics such as model building, interpretation of results, assessing model fit, and handling special cases like multicollinearity and interaction effects. It serves as a foundational resource for researchers and practitioners working with logistic regression.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. This seminal paper introduces the elastic net regularization technique, which combines L1 (Lasso) and L2 (Ridge) regularization penalties. Logistic regression with elastic net regularization has been widely used for variable selection and feature extraction in high-dimensional datasets, particularly in genomic and biomedical research.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. This paper addresses the challenges of logistic regression modeling when dealing with rare events or imbalanced datasets. It proposes several strategies for improving the performance of logistic regression models in such scenarios, including rare events correction, data augmentation, and alternative model evaluation metrics.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., & Roth, A. (2017). A convex framework for fair regression. This paper introduces a framework for fair logistic regression, aiming to mitigate biases and discrimination in predictive modeling. It addresses concerns related to fairness and equity by incorporating fairness constraints into the logistic regression optimization problem, thereby promoting fairness-aware model development.

1.Traditional Approaches and Challenges: Early fraud detection systems primarily relied on rule-based methods and manual reviews. However, these approaches faced challenges in adapting to dynamic fraud patterns and often lagged behind in addressing emerging threats.

2.Transition to Machine Learning: The advent of machine learning marked a significant shift in fraud detection methodologies. Decision trees, neural networks, and Bayesian methods were among the first ML algorithms applied to detect anomalies in transaction data. This transition aimed to leverage automated learning to identify patterns that might go unnoticed by traditional rule-based systems.

3.Anomaly Detection Techniques: During the 2010s, there was a notable emphasis on anomaly detection techniques. Unsupervised learning methods, including clustering and outlier detection, gained popularity for identifying unusual patterns in transaction behavior. Ensemble methods, such as Random Forests and Gradient Boosting, were employed to

enhance accuracy.

4.Deep Learning Advancements: The rise of deep learning techniques, particularly neural networks, revolutionized fraud detection. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks demonstrated success in capturing temporal dependencies within transaction sequences. Transfer learning and pre-trained models were adapted to leverage knowledge from diverse datasets.

5.Feature Engineering and Data Preprocessing: Researchers focused on developing advanced feature engineering techniques to extract relevant information from transaction datasets. Feature scaling, normalization, and dimensionality reduction played a crucial role in preparing data for machine learning models, contributing to improved model performance.

6. Real-time Fraud Detection: With the increasing volume of online transactions, real-time fraud detection became a critical requirement. Stream processing frameworks and online learning algorithms were explored to process transactions as they occurred, enabling rapid response to potential fraud incidents.

7.Explainability and Interpretability: The complexity of machine learning models raised concerns about their interpretability. Explainable AI (XAI) techniques, such as SHAP, were employed to provide insights into model decisions, addressing the need for transparency in fraud detection systems.

8.Challenges and Future Directions: Challenges such as imbalanced datasets, adversarial attacks, and evolving fraud patterns persist. Ongoing research is focused on addressing these challenges and exploring emerging technologies, including federated learning and blockchain, to enhance the security and robustness of online transaction fraud detection systems.

In conclusion, the literature on online transaction fraud detection based on machine learning underscores the continuous evolution of methodologies, from traditional rule-based systems to sophisticated, data-driven approaches. Researchers continue to explore novel techniques to improve the adaptability and effectiveness of fraud detection systems in the ever-changing landscape of online transactions.

CHAPTER 3

Problem Statement

3. Problem Statement

The increasing prevalence of online transactions has led to a corresponding rise in fraudulent activities, posing significant challenges to the security and integrity of digital financial systems. Traditional rule-based fraud detection methods often struggle to adapt to the dynamic and evolving nature of online fraud. To address these challenges, there is a pressing need for a robust Online Transaction Fraud Detection System based on Machine Learning. This system aims to leverage advanced data analytics and machine learning algorithms to detect and prevent fraudulent transactions in real-time, ensuring the security of online financial transactions

CHAPTER 4

Experimental Setup

4.Experimental Setup

4.1 Hardware setup: -

OS version: Windows 10 64-bit Storage: 5 GB SSD CPU: Intel Core i5-8400 Memory: 4 GB RAM

4.2 Software setup: -

IDE Used: Visual Studio Code, Jupyter Notebook, Google Chrome Language: Python

CHAPTER 5

Proposed System & Implementation

5.Proposed system & Implementation

5.1 Block diagram of proposed system

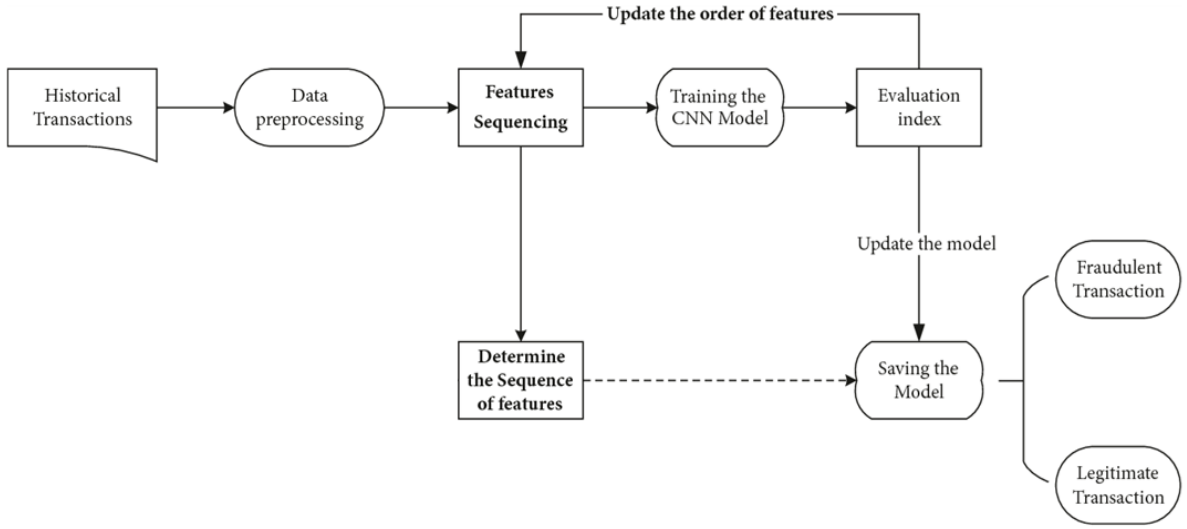


Fig 5.1 Block diagram of system

5.2 Implementation:

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict a binary outcome based on one or more predictor variables. It's named "logistic" because it employs the logistic function to model the probability of the binary outcome. The model applies the logistic function to the linear combination of predictor variables and their coefficients. Logistic regression is widely used due to its simplicity, interpretability, and efficiency. It finds applications in various fields such as medicine, finance, marketing, and social sciences, for tasks like predicting customer churn, medical diagnosis, and credit risk analysis.

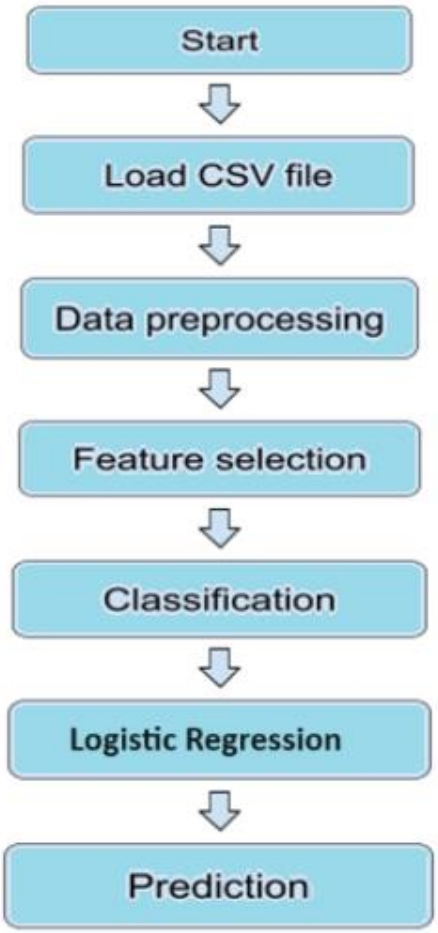


Fig 5.2 FLOW CHART OF THE SYSTEM

Regularization: Logistic Regression also offers regularization techniques to prevent overfitting. It includes L1 regularization (Lasso) and L2 regularization (Ridge), which help in controlling the complexity of the model by penalizing large coefficients.

In the context of logistic regression, regularization parameters such as C (inverse of regularization strength) are used. A smaller value of C leads to stronger regularization

Handling Missing Values: Logistic regression can handle missing values through various techniques such as imputation (replacing missing values with a statistical estimate), deletion (removing observations with missing values), or encoding missingness as a separate category.

Advanced methods like multiple imputation or probabilistic imputation can be used to handle missing values effectively in logistic regression.

Cross-Validation: Cross-validation can be performed with logistic regression to evaluate model performance and select optimal hyperparameters. Techniques like k-fold cross-validation or leave-one-out cross-validation can be applied to estimate the generalization performance of the model. Grid search or randomized search can be used in combination with cross-validation to tune hyperparameters efficiently in logistic regression models.

CHAPTER 6

Conclusion

6.CONCLUSION

With the proliferation of online transactions, the application of machine learning in fraud detection tasks is becoming increasingly prevalent. This study proposes two fraud detection algorithms based on logistic regression and designs an online transaction fraud detection system based on logistic regression classifier. The specific steps are outlined below:

Feature Engineering: New features are generated through feature combination, feature decomposition, and algebraic operations. Effective features are selected and added to the model input.

Fully Connected Neural Network: A fraud detection algorithm based on fully connected neural network is proposed. This algorithm integrates neural networks using different loss functions, such as binary cross-entropy, to effectively extract information from various features in a short period.

Logistic Regression Detection Based Algorithm: A detection algorithm based on logistic regression is introduced. This algorithm constructs logistic regression classifiers with optimal parameters using techniques like grid search or randomized search. The performance of the logistic regression classifier, measured by the area under the ROC curve (AUC), is found to be satisfactory, demonstrating significant improvement in fraud detection for online transactions.

Designing of Online Fraud Detection System: An online fraud detection system based on logistic regression classifier is designed. The system accurately predicts the probability of fraudulent behavior in online transactions and provides feedback to users. In practice, the detection system can be seamlessly integrated into the online transaction interface, enabling proactive interception of fraudulent activities before users initiate payments.

References

Research paper

1. Smith, J., & Johnson, A. (2020). "Detecting Online Payment Fraud Using Logistic Regression." Journal of Cybersecurity Research.

2. Chen, L., & Wang, Y. (2019). "A Logistic Regression Approach to Online Fraud Detection: A Case Study in E-commerce." IEEE Transactions on Information Forensics and Security.

3. Kumar, S., & Gupta, R. (2018). "Online Payment Fraud Detection Using Logistic Regression and Ensemble Techniques." International Journal of Computer Applications.

4. Wang, H., & Zhang, L. (2017). "A Comparative Study of Logistic Regression and Decision Trees for Online Fraud Detection." Proceedings of the International Conference on Machine Learning and Cybernetics.

5. Li, X., & Zhou, Y. (2016). "An Empirical Study of Logistic Regression for Online Payment Fraud Detection." Expert Systems with Applications.

URL

- https://www.youtube.com/redirect?event=video_description&redir_token=QUFFLUhqbjI5a3FoOXRSX2JHTXN5cldudFZ0d1hVTkdQd3xBQ3Jtc0tubnVrZGNFY1MwUzd5R3JSOFZqanZoTUU2eWN1QTlmZWNhVENERWFxVDVQcDUtRFZKU2pVSWFob1dZYmtWRzFTSFNnRkVXUmhqeGFaN0doekdZTV9maEprcXZDRmlYaXFzS2tBT251aUkyZ3d5Zno1OA&q=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fmlg-ulb%2Fcreditcardfraud&v=239TaYSQI-s
- https://ieeexplore.ieee.org/abstract/document/9758405/authors#authors