JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-18, No.-9, September (2023) pp 1-9 ISSN (Print) 0973-8975

IDENTIFYING FRAUD IN ONLINE TRANSACTIONS

Sneha Sen[1], Megha Adhikari[2], Dilip Kumar Gayen[3]

[1, 2] Student member, Department of Computer Science & Engineering

[3]Professor, Department of Computer Science and Engineering, College of Engineering & Management, KTTP Township, Purba MedinipurWest Bengal, India.

Corresponding Author : **Dr. Dilip Kumar Gayen**

Email: dilipgayen@cemk.ac.in, dilipgayen@yahoo.com

https://doi.org/10.26782/jmcms.2023.09.00001

(Received: June 23, 2023; Accepted: September 05, 2023)

Abstract

Fraudulent credit card transactions must be when customers are charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Identifying Fraud in Online Transactions. The Identifying Fraud in Online Transactions problem includes modelling past credit card transactions with the data of the ones that turned out to be a fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 99.99% of the fraudulent transactions while minimizing the incorrect fraud classifications. Identifying Fraud in Online Transactions is a typical sample of classification. In this process, we have focused on analyzing and pre-processing data sets by using a Random Forest Algorithm.

Keywords : Frauds Classification, Online Transactions, credit card transactions.

I. Introduction

Fraud in online transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, online transaction fraud or credit card fraud 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 [IV]. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive, or avoid objectionable behaviour, which consists of fraud, intrusion, and defaulting. 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 [I-III]. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time.

II. Objective

The paper aims to design and develop a user-friendly, efficient computerized system with the main objective of eliminating data redundancy and enhancing data security through a login and password mechanism. This project primarily focuses on predicting the authenticity of credit card transactions based on transaction amounts, achieved by detecting anomalies within transaction data. The system provides a range of functionalities, including comprehensive search capabilities based on factors such as credit cards, datasets, files, and predictions. It effectively manages customer details, transaction information, credit card data, and files, while also improving efficiency in credit card and transaction management. With features like information monitoring, editing, updating, and resource management, the Credit Card Fraud Detection System ensures streamlined operations, offering an integrated solution to enhance accuracy and security across all records and predictions. Utilizing the Random Forest Algorithm, this study focuses on Credit Card Fraud Detection, aiming to promptly identify and categorize fraudulent credit card transactions, thereby preventing unjust charges to cardholders for unauthorized purchases. The model's design emphasizes simplicity and speed to swiftly pinpoint and label anomalies as fraudulent transactions. Techniques to address data imbalance are thoughtfully applied, and user privacy is safeguarded through dimensionality reduction. Furthermore, an emphasis is placed on using reliable data sources that undergo thorough verification, especially during model training. By maintaining a simplistic and interpretable model, the adaptability of scammers prompts the quick development and deployment of new models as a countermeasure. Empirical findings validate the efficacy of the Random Forest Model in accurately and efficiently predicting credit card anomalies. The dataset, obtained from the Kaggle website, encompasses transactions conducted by European credit cardholders in September 2013, containing around 284,808 records of real values with no null entries. This ‘credit card' dataset was subjected to analysis using the Random Forest Algorithm [V-VI].

III. Algorithm

In the realm of credit card fraud detection, a range of algorithms such as Logistic Regression, Decision Tree, and Random Forest are available. However, our project has honed in on the Random Forest Algorithm. This machine learning technique, classified under supervised learning, is versatile for both Classification and Regression tasks in the field of Machine Learning (ML). At its core lies ensemble learning, a method that amalgamates multiple classifiers to address complex issues and enhance model performance. Named after its composition, the Random Forest Classifier consists of several decision trees formed on diverse subsets of the dataset, and through averaging, it heightens predictive accuracy. Rather than relying solely on one decision tree, the algorithm aggregates predictions from each tree, culminating in the final output determined by the majority vote. By increasing the number of trees within the forest, accuracy is elevated, and overfitting concerns are mitigated.

The Random Forest algorithm operates in two distinct phases. The first phase involves the creation of the random forest by amalgamating N decision trees, while the second phase entails making predictions for each tree generated in the first phase. The algorithm's functioning can be elucidated through the following sequential steps and diagram:

- Step 1: Randomly select K data points from the training set.
- Step 2: Construct decision trees corresponding to the chosen data points (subsets).
- Step 3: Determine the desired number N of decision trees to be created.
- Step 4: Repeat Steps 1 and 2.
- Step 5: When dealing with new data points, acquire predictions from each decision tree, and allocate the new data points to the category that garners the majority of votes.

For a more profound understanding of the algorithm's intricacies, let's delve into an illustrative example: Envision a dataset encompassing an array of images depicting diverse fruits. This dataset is inputted into the Random Forest classifier, wherein distinct subsets of the dataset are allocated to individual decision trees. As the training phase unfolds, each decision tree generates a predictive outcome. Upon the emergence of a novel data point, the Random Forest classifier undertakes the task of forecasting the ultimate decision, drawing upon the collective majority verdict gleaned from the diverse individual trees. This exemplar exemplifies how the Random Forest algorithm leverages ensemble learning to enhance prediction accuracy and minimize overfitting concerns, making it a robust choice for a wide range of classification and regression problems in machine learning.

IV. Methodology

The methodology applied in this research embodies a systematic sequence of steps meticulously designed to tackle the intricate challenge of credit card fraud detection. Each of these carefully orchestrated steps contributes to the holistic analysis and construction of a robust and efficient fraud detection model.

Step 1: Data Collection

The inception of this methodology involves the procurement of a comprehensive dataset from the reputable Kaggle platform. This dataset encompasses a wealth of credit card transactions carried out by European cardholders during September in the year 2013.

Step 2: Data Preparation

With the dataset in hand, the foundation is laid by importing vital libraries, such as pandas and NumPy. These libraries play an instrumental role in streamlining the subsequent data handling and analysis processes.

Step 3: Data Loading and Exploration

The dataset is ingested into the research environment utilizing the versatile pandas library. This initial interaction with the data paves the way for an extensive exploration, unearthing crucial insights into the inherent characteristics and intricate structure of the dataset.

Step 4: Data Description

A concise yet informative snapshot of the dataset's dimensionality is captured through the display of its shape. Furthermore, a comprehensive descriptive summary of the dataset's statistical attributes is generated, offering a deeper understanding of its composition.

Step 5: Data Imbalance Assessment

A meticulous examination of the dataset uncovers a striking imbalance, with an astonishingly low 0.17% of transactions classified as fraudulent. Acknowledging this data asymmetry, the initial stages of model training ensue without any form of data balancing. Should the accuracy of the model be compromised, considerations will be given to implementing data balancing techniques.

Step 6: Transaction Amount Analysis

The methodology delves into a thorough analysis of transaction amounts associated with both fraudulent and legitimate transactions. This analysis exposes a conspicuous pattern, revealing that fraudulent activities are characterized by notably higher average transaction amounts.

Step 7: Correlation Matrix Visualization

The intricate interplay between various features is unveiled through an adept application of correlation matrix analysis. This visual representation, illustrated via a heatmap, uncovers underlying correlations, thereby augmenting the depth of data comprehension.

Step 8: Data Division

A pivotal stage is reached as the dataset is systematically divided into distinct input parameters and corresponding output values. This initial partitioning sets the stage for the subsequent separation into training and testing subsets, thereby laying the groundwork for comprehensive model evaluation.

Step 9: Model Development and Evaluation

The heart of the methodology resides in the development of a potent Random Forest model, meticulously constructed through the scikit-learn library. The model's efficacy is rigorously assessed through the computation of essential evaluation metrics, encompassing accuracy, precision, and the Matthews correlation coefficient. This comprehensive assessment paints a vivid picture of the model's performance.

Step 10: Confusion Matrix Visualization

The research methodology culminates in the visualization of the confusion matrix, which provides an insightful depiction of the model's performance across various categories, including true positives, true negatives, false positives, and false negatives.

This intricate tapestry of systematic steps weaves together to form a robust and cohesive research methodology, one that not only comprehensively addresses the multifaceted realm of credit card fraud detection but also serves as a foundation for insightful analyses and informed decisions. Through each step, the methodology propels the research endeavor toward a deeper understanding of the subject matter and the development of an effective predictive model.

V. Results and Discussions

We present the snapshots of the user interface that highlight the outcomes of our study. These snapshots provide a visual representation of the key findings and interactions within the developed system.

Snapshot 1: Prediction Visualization

The first snapshot (Fig. 1) showcases the prediction visualization, where the system's algorithm predicts the likelihood of a credit card transaction being fraudulent. The user interface displays color-coded indicators to represent the predicted outcomes, with green indicating a non-fraudulent transaction and red denoting a potentially fraudulent one. The snapshot captures the dynamic nature of real-time predictions and serves as an informative tool for users to assess transactions quickly.

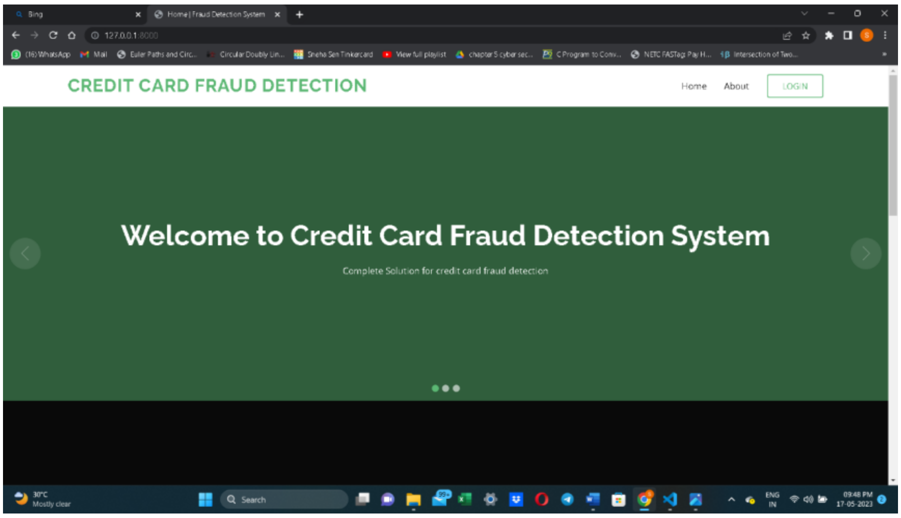
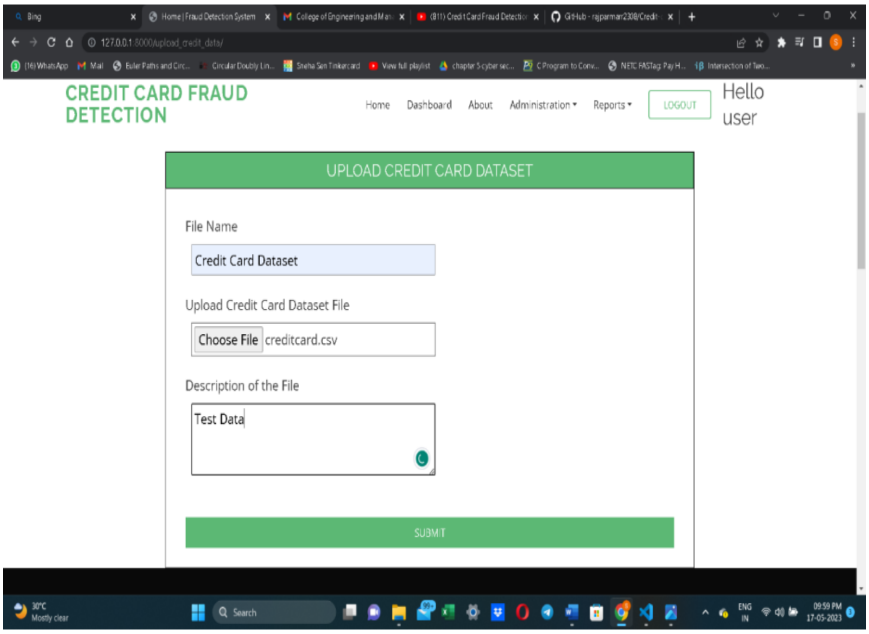

Home | Fraud Detection System
127.0.0.1:8000
CREDIT CARD FRAUD DETECTION
Home
About
LOGIN
Welcome to Credit Card Fraud Detection System
Complete Solution for credit card fraud detection
Search

Fig. 1. User interface

Snapshot 2: Data Distribution Analysis

Snapshot 2 (Fig. 2) illustrates the data distribution analysis. It presents a histogram depicting the distribution of transaction amounts for both normal and fraudulent transactions. The distinct patterns in the two distributions can be visually compared, aiding in understanding the inherent differences between genuine and potentially fraudulent activities.



File Name

Credit Card Dataset

Upload Credit Card Dataset File

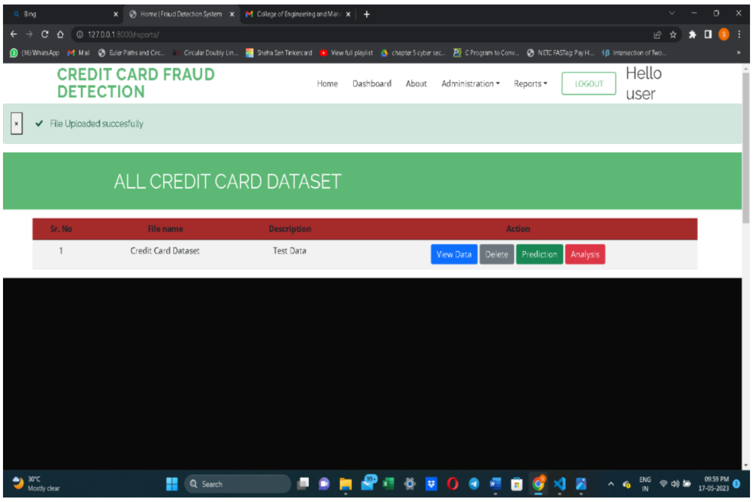
Choose File creditcard.csv

Description of the File

Test Data

Fig. 2. User interface 2

Fig. 3 focuses on the display of model evaluation metrics. Bar charts depict metrics such as accuracy, precision, recall, and F1-score. Users can easily gauge the overall performance of the model through these visual representations, enabling them to make informed decisions about deploying the system for fraud detection. Incorporating these snapshots into the paper enhances the reader's comprehension of the research outcomes and the practical implications of the developed user interface. The graphical representations provide an intuitive insight into the system's effectiveness in detecting credit card fraud and empower users to make informed decisions based on real-time predictions and analyses. Please adjust the descriptions as needed and consider adding relevant captions to each snapshot for clarity and context in this paper.



CREDIT CARD FRAUD DETECTION

Hello user

✓ File Uploaded succesfully

ALL CREDIT CARD DATASET

Sr. No	File name	Description	Action
1	Credit Card Dataset	Test Data	View Data Delete Prediction Analysis

Fig. 3. User interface 3

The methodology employed in this study comprises a series of systematic steps aimed at effectively addressing the task of credit card fraud detection. Each step contributes to the comprehensive analysis and development of a robust model. This meticulous methodology encompasses data preparation, exploration, analysis, model development, and evaluation to devise an effective Random Forest model for credit card fraud detection. The proposed approach leverages various techniques to enhance model accuracy, thereby contributing to improved results in combating fraudulent transactions. The dataset at hand encompasses a collection of transactions spanning a duration of two days, totaling approximately 284,808 entries. Within this dataset, a specific subset of transactions, consisting of 492 cases, is identified as fraudulent. However, the distribution of these fraudulent transactions within the dataset reveals a significant imbalance. To be precise, the instances of fraud, representing the positive class, constitute a mere 0.172% of the overall transaction count.

This imbalance in class distribution has crucial implications for the analysis and modeling of the dataset. When such a vast majority of transactions belong to the negative class (i.e., legitimate transactions), while only a minute fraction falls under the positive class (i.e., fraudulent transactions), it can pose challenges for predictive modeling and analysis. Traditional machine learning algorithms tend to favor the majority class due to its prevalence, leading to suboptimal performance in detecting the minority class.

The issue of class imbalance warrants careful consideration, as a model trained on such imbalanced data might lead to a high false negative rate for the minority class, rendering it ineffective in accurately identifying and flagging fraudulent transactions. To address this challenge, various techniques can be employed, such as resampling the dataset to balance class distributions, utilizing different evaluation metrics that account for class imbalance, or employing specialized algorithms designed to handle imbalanced data.

Given the rarity of fraudulent transactions, detecting them becomes a critical task, especially in financial systems where such incidents can result in substantial financial losses. The utilization of appropriate techniques and algorithms becomes paramount to ensure a robust fraud detection model. Therefore, practitioners and researchers must adopt a comprehensive approach that takes into account the inherent class imbalance, enabling the development of models that can effectively distinguish between legitimate and fraudulent transactions.

The dataset's presentation of transactions occurring over two days, with only 492 frauds out of nearly 285,000 transactions, underscores the highly imbalanced nature of the dataset. This imbalance requires careful consideration and specialized techniques to develop accurate and reliable fraud detection models, ensuring the effective identification of fraudulent transactions despite their rarity in the dataset.

VI. Conclusion

In conclusion, this study has delved into the intricate realm of credit card fraud detection, employing the potent Random Forest algorithm as the focal point of investigation. The meticulous methodology encompassed various stages, starting from data collection and preparation to model development and evaluation.

The dataset, procured from the Kaggle website, served as the foundation for this study. With 492 fraud cases out of approximately 284,808 transactions, the dataset's highly imbalanced nature posed a significant challenge. The rarity of fraudulent instances underscored the need for a robust algorithm capable of accurately detecting them amidst the overwhelming majority of legitimate transactions.

The Random Forest algorithm emerged as a prominent choice due to its ensemble nature, combining multiple decision trees to enhance predictive accuracy. The step-by-step journey through the methodology allowed for data exploration, correlation analysis, and model evaluation. Visualization tools like confusion matrices and correlation heatmaps further enriched the analysis, aiding in the comprehension of the algorithm's performance and feature relevance.

The study's findings reinforce the importance of specialized approaches to address class imbalance, especially in fraud detection scenarios. The model's accuracy, precision, and Matthews correlation coefficient were pivotal metrics to assess its efficacy, and the visualization of the confusion matrix offered a tangible representation of its performance.

Overall, this study sheds light on the challenges and opportunities in credit card fraud detection. The presented methodology can serve as a blueprint for tackling similar imbalanced datasets and enhancing the performance of fraud detection algorithms. In a digital landscape where financial security hinges on the ability to swiftly identify and counteract fraudulent transactions, the pursuit of effective and accurate detection mechanisms remains paramount. This research contributes to this endeavor by showcasing the potential of the Random Forest algorithm and demonstrating the importance of specialized techniques in addressing class imbalance challenges.

Conflict of Interest:

The author declares that there was no conflict of interest regarding this paper.

References

I. A. A. Taha and S. J. Malebary, : "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine", *IEEE Access*, 8, 25579–25587, 2020.

II. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, : "Credit Card Fraud Detection - Machine Learning methods", 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 1–5, 2019.

III. J. I-Z. Chen, K.-L. Lai, : "Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert", *Journal of Artificial Intelligence and Capsule Networks*, 03(02), 101–112, 2021.

IV. Mehria Nawaz, Twinkle Agarwal, Dilip Kumar Gayen. : : 'ONLINE SKILL TEST PLATFORM'. *J. Mech. Cont. & Math. Sci.*, Vol.-17, No.-11, November (2022) pp 46-53. 10.26782/jmcms.2022.11.00003

V. R. B. Sulaiman, V. Schetinin, P. Sant, : "Review of Machine Learning Approach on Credit Card Fraud Detection", *Human-Centric Intelligent Systems*. volume 2, 55–68, 2022.

VI. Q. Han, C. Gui, J. Xu, G. Lacidogna, : "A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm", *Construction and Building Materials*. 226, 734-742, 2019.

VII. Y. Chen, W. Zheng, W. Li, Y. Huang, : "Large group activity security risk assessment and risk early warning based on random forest algorithm". *Pattern Recognition Letters*, 144, 1–5, 2021.