ISSN : 0973 - 8355

www.ijmmsa.com

INTERNATIONAL JOURNAL OF MATHEMATICAL, MODELLING, SIMULATIONS AND APPLICATIONS

IJMMSA

E-MAIL
editor.ijmmsa@gmail.com
editor@ijmmsa.com

INTERNATIONAL JOURNAL OF MATHEMATICAL, MODELLING, SIMULATION AND APPLICATIONS
IJMMSA

ISSN; 0973-8355 www.ijmmsa.com

Vol 16 Issue 4 Oct 2023

Credit Card Fraud Detection Using Machine Learning

jwala A. Bongale[1] , Uzma Naaz M. Rangrez[2]

[1,2]Department of Electronics and Telecommunications,
N. K. Orchid College of Engineering &
Technology, Solapur, Maharashtra, India
ujwalabongale@orchidengg.ac.in, uzmarangrez28@gmail.com

Abstract - Credit card fraud is a significant problem in the financial industry, with billions of dollars lost annually due to fraudulent activities. To combat this issue, Credit Fraud Detection Systems have been developed to detect fraudulent transactions and alert the appropriate parties. In this project, we propose a Credit Fraud Detection System that utilizes machine learning algorithms to identify potentially fraudulent transactions. The system utilizes a combination of supervised and unsupervised learning techniques to analyze transaction data and identify patterns that are indicative of fraud. The proposed system takes in transactional data such as transaction date, amount, merchant name, card number, and other details, which are preprocessed to extract relevant features. The preprocessed data is then fed into the machine learning model, which has been trained on a dataset containing both fraudulent and non-fraudulent transactions. The model is designed to learn the patterns and characteristics of fraudulent transactions and classify new transactions as either fraudulent or non-fraudulent.

Keywords— *Credit Card Fraud, Machine Learning, Data Science, Isolation Forest Algorithm, Local 0utlier Factor, Automated Fraud Detection.*

INTRODUCTION

'Fraud' in credit card 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, 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. Fraud detection involves monitoring the activities of populations of users 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 technologies 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, and the transaction patterns which often change their statistical properties over time [1].

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 analyse 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.

Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies.

These frauds are classified as:

- Credit Card Frauds: Online and Offline
- Card Theft
- Account Bankruptcy
- Device Intrusion
- Application Fraud
- Counterfeit Card
- Telecommunication Fraud

Some of the currently used approaches for the detection of such fraud are:

- Artificial Neural Network
- Fuzzy Logic
- Genetic Algorithm
- Logistic Regression
- Decision tree

ISSN; 0973-8355 www.ijmmsa.com
Vol 16 Issue 4 Oct 2023

- Support Vector Machines
- Bayesian Networks
- Hidden Markov Model
- K-Nearest Neighbour

Fraud acts as unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule, or policy to attain unauthorized financial benefit.

Numerous kinds of literature about anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates has revealed that techniques employed in this domain include data mining applications, automated fraud detection, and adversarial detection [2]. In another paper, Suman, a Research Scholar, at GJUS & T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection [3]. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection.

A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining, and Distance sum algorithms to accurately predict fraudulent transactions in an emulation experiment of a credit card transaction data set of one certain commercial bank [4]. Outlier mining is a field of data mining that is used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that aren't genuine. The attributes of a customer's behavior were taken and based on the value of those attributes the distance between the observed value of that attribute and its predetermined value was calculated.

Unconventional techniques such as hybrid data mining/complex network classification algorithms can perceive illegal instances in an actual card transaction data set, based on a network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium-sized online transactions.

There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transactions.

In case of a fraudulent transaction, the authorised system would be alerted and feedback would be sent to deny the ongoing transaction.

Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problems with variable misclassification costs.

II SYSTEM DESCRIPTION

First, the credit card dataset is taken from the supply, and cleaning and approval are executed on the dataset which joins disposal of excess, filling void territories in sections, and changing imperative variables into components or exercises then actualities are divided into 2 sections, one is preparing dataset and another is checking data set.

Presently k crease move approval is done that is the special example is arbitrarily divided into k same and equivalent measured subsamples. Of the k subsamples an individual subsample is held because the approval actualities for looking at the model and the last k-1 subsamples are utilized as instruction data, designs are made for Logistic relapse, choice tree, SVM, Random woodland after which exactness, affectability,

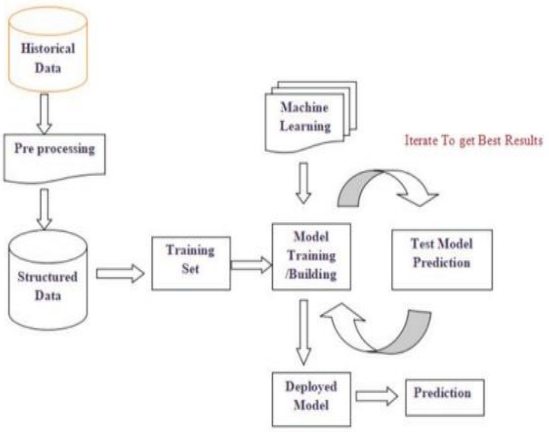

Fig 1: System Architecture

explicitness, accuracy are determined and a differentiation is made. The dataset is sourced from the ULB framework contemplating association. The dataset incorporates Visa exchanges and the event of exchanges that occurred in days is exhibited on this dataset, comprehensive of 284,786 exchanges. The dataset is outstandingly unequal and skewed towards the magnificent tastefulness and extortion cases make up 0.173% of the exchanges certainties. It conveys the handiest numerical (constant) enter factors that are a result of a Principal Component Analysis (PCA) trademark decision change coming about to twenty-eight basic added substances. Furthermore, generally speaking, 30 input highlights are connected. The social element of the card is demonstrated by the method for a variable of each profile usage speaking to the spending behavior of the clients alongside days of the month, hours of the day, topographical areas, and diverse kinds of clients. where the exchange happens in the area. Later those factors are used in making renditions that separate deceitful exercises. The subtleties and the verifiable past certainties about the capacity are not issued because of security issues. The time factor monitors the number of seconds that have been conveyed in each

ISSN; 0973-8355 www.ijmmsa.com
Vol 16 Issue 4 Oct 2023

exchange alongside the exchange that is done toward the start of the dataset. The overall system diagram is shown in Figure 1.

The following software packages are used for data exploration, pro processing, and for random forest algorithms

- NumPy: For simple arrays.
- Pandas: For reading the file.
- SciKit: Learn- for pre-processing.
- Matplotlib or Seaborn: For plotting and representing confusion matrix color format.
- Tensor flow: For matrix format.

The complete system is implemented in the following Modules

- Data collection
- Data pre-processing
- Data Visualization
- Feature extraction
- Evaluation model

A. Data Collection

Data used here is a set of product reviews collected from credit card transaction records. This step is concerned with selecting the subset of all available data that will be working with. ML problems start with data preferably, lots of data (examples or observations) for which the target answer is known. Data that is already known as the target answer is called labeled data to classify fraudulent and legitimate credit card transactions by supervised learning algorithms such as Random Forest. This helps customers to get awareness about the fraud without losing any financially. Selected data is shown in Figure 2.

B. Data Pre-processing

Pre-processing is the process of three important and common steps as follows:

• *Formatting:* It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. The most recommended format is .csv files.

• *Cleaning:* Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming categories and so on. For most data scientists, Data Cleaning continues for 80% of the work.

• *Sampling:* This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behavior and pattern of data in an integrated way. It is shown in Figure 3.

C. Data visualization

Data Visualisation is the method of representing the data graphically and pictorially, data scientists depict a story by the results they derive from analysing and visualising the data. The best tool used is Tableau which has many features to play around with data and fetch wonderful results.

Fig 2:Data Collection

3.3.3.1 Pre-processing with python commands

STEP 1:

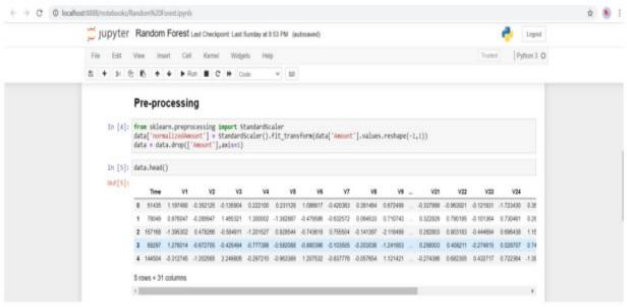

Fig 3: Pre -Pre-processing with Step

D. Feature extraction

Feature extraction is the process of studying the behavior and pattern of the analyzed data and drawing the features for further testing and training. Finally, models are trained using

ISSN; 0973-8355 www.ijmmsa.com
Vol 16 Issue 4 Oct 2023

the Classifier algorithm. The classify module is used on the Natural Language Toolkit library in Python. Then the gathered labeled dataset is used. The rest of our labeled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random Forests. These algorithms are very popular in text classification tasks.

E. Evaluation model

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and overfitted models. To avoid overfitting, evaluation methods such as hold-out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

I. ALGORITHMS

A. Random Forest

Random forest is a supervised machine learning algorithm based on ensemble learning. Ensemble learning is an algorithm where the predictions are derived by assembling or bagging different models or similar models multiple times. The random forest algorithm works similarly and uses multiple algorithms i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest".

Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor.

Random forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in a random forest classifier, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

B. Linear Regression

Linear regression is one of the most popular and simple machine learning algorithms that is used for predictive analysis. Here, predictive analysis defines the prediction of something, and linear regression makes predictions for continuous numbers such as salary, age, etc

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1.

A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

C. Logistic Regression

Logistic regression is the supervised learning algorithm, which is used to predict categorical variables or discrete values. It can be used for classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc. Logistic regression is similar to linear regression except for how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and predict the discrete values.

Logistic regression is a supervised machine learning algorithm mainly used for binary classification where we use a logistic function, also known as a sigmoid function that takes input as independent variables and produces a probability value between 0 and 1. For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

D. Decision tree

ISSN; 0973-8355 www.ijmmsa.com
Vol 16 Issue 4 Oct 2023

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the tests are performed based on features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions**.** It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. To build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

II. SOFTWARE DESCRIPTION

Python Language

- Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation via the off-side rule.
- Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented, and functional programming.
- Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000. Python 3.0, released in 2008, was a major revision not completely backward compatible with earlier versions. Python 2.7.18, released in 2020, was the last release of Python 2. 5.2

Machine Learning

- Machine learning is the study of computer algorithms that learn without human input.
- ML has countless applications, from natural language processing to computer vision, neural networks, predictive analytics, and more.
- Lower-level languages (like R, C++, or Java) offer greater speed but are harder to learn.
- Higher-level languages (like JavaScript and Python) are easier to use but slower to execute.
- Python is a key language for machine learning and data analytics. For speed-to-competence and breadth of application, it's probably the best one for beginners.
- Nevertheless, the right language hinges on the problem you're solving, your expertise, and your programming experience. Don't limit yourself.

V. RESULT DISCUSSION

Figure 4 shows the output of the Logistic Regression algorithm which shows it selects a large amount of transaction. Confusion Matrix due to "Logistic Regression Algorithm" where y-axis is True Transaction and x-axis is Predictable Transaction. It shows a maximum transaction value of 56852 for the Decision Tree Algorithm.

Figure 5 shows a maximum transaction value of 56836 for the Decision Tree Algorithm. Confusion Matrix due to "Decision Tree Algorithm" where y-axis is True Transaction and x-axis is Predictable Transaction.

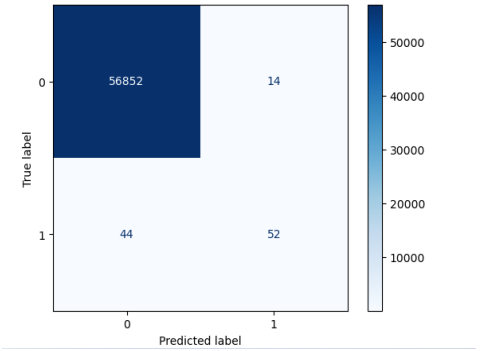

Fig 4: Output for Logistic Regression

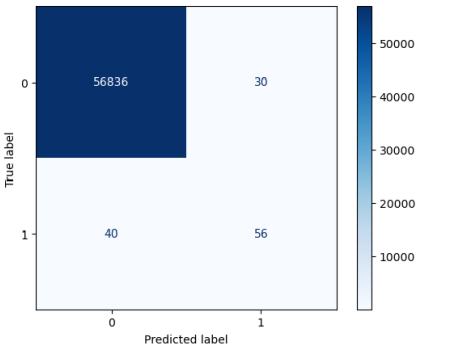

Fig 5: Output for Decision Tree

ISSN; 0973-8355 www.ijmmsa.com

Vol 16 Issue 4 Oct 2023

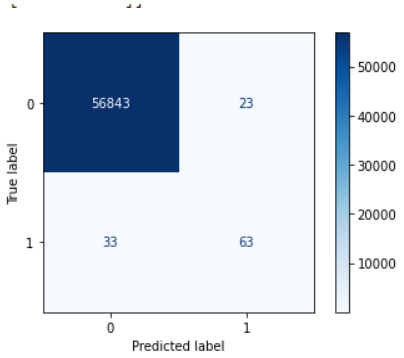

Fig 6: Output for random forest

Figure 6 shows a maximum transaction value of 56843 for Confusion Matrix due to "Random forest Algorithm" where the y-axis is True Transaction and the x-axis is Predictable Transaction.

Figure 7 shows a maximum transaction value of 56859 for the Confusion Matrix due to the "XG boost Algorithm" where the y-axis is True Transaction and the x-axis is Predictable Transaction.

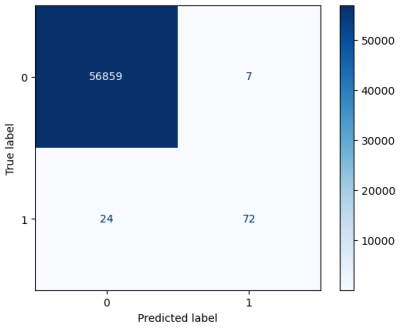

Fig 7: Output for XG boost

Out[90]:

	Model Name	Accuracy	F1-score	ROC
1	XG Boost	0.999456	0.822857	0.978537
0	Logistic Regression	0.998982	0.641975	0.977696
3	Random Forest	0.999017	0.692308	0.963978
2	Decision Tree	0.998771	0.615385	0.921750

Fig 8: Result table

Figure 8 shows the comparative results of different models of X Boost, Logistic Regression, Decision Tree, Random Forest, and XG boost Algorithms. It shows Accuracy, F1-score, and Region of convergence ROC. This table indicates that the XG Boost algorithm gives a more accurate prediction than other algorithms.

VI. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed the most common methods of fraud along with their detection methods and reviewed recent findings in this field. Machine learning can be applied effectively to get better results in fraud detection, along with the algorithm, pseudocode, its implementation, and experimentation results.

While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and the number of genuine transactions.

Since the entire dataset consists of only two days' transaction records, it's only a fraction of the data that can be made available. Machine learning algorithms, the program will only increase its efficiency over time as more data is put into it. Ere Advanced Machine Learning Techniques such as deep learning, anomaly detection, and ensemble methods can be used to identify complex patterns and anomalies in credit card transactions, improving fraud detection accuracy.

There is a scope to use Real-time monitoring systems that can analyze transactions as they occur. This allows for immediate detection and response to fraudulent activities, reducing the potential impact and losses [5].

Also, Feature Engineering continuously refines and enhances the features used in fraud detection models. Additional data sources such as location information, device fingerprints, user behavior patterns, and social network analysis are considered to improve the accuracy of fraud detection [6].

Unsupervised learning algorithms can be used to identify new and emerging patterns of fraudulent behavior [7]. These algorithms can detect previously unknown fraud patterns by identifying deviations from normal behavior, even without labeled examples of fraud.

III. REFERENCES

[1] Randhawa, Kuldeep, et al. "Credit card fraud detection using AdaBoost and majority voting." *IEEE access* 6 (2018): 14277-14284.

[2] Roy, Abhimanyu, et al. "Deep learning detecting fraud in credit card transactions." *2018 systems and information engineering design symposium (SIEDS)*. IEEE, 2018.

ISSN; 0973-8355 www.ijmmsa.com
Vol 16 Issue 4 Oct 2023

[3] Xuan, Shiyang, et al. "Random forest for credit card fraud detection." *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*. IEEE, 2018.

[4] Kazemi, Zahra, and Houman Zarrabi. "Using deep networks for fraud detection in the credit card transactions." *2017 IEEE 4th International conference on knowledge-based engineering and innovation (KBEI)*. IEEE, 2017.

[5] Altman, Erik. "Synthesizing credit card transactions." *Proceedings of the Second ACM International Conference on AI in Finance*. 2021.

[6] Gill, Mohsin Asad, et al. "Detection of credit card fraud through machine learning in banking industry." *Journal of Computing & Biomedical Informatics* 5.01 (2023): 273-282.

[7] Dheepa, V., and R. Dhanapal. "Behavior-based credit card fraud detection using support vector." *ICTACT Journal on Soft Computing* 2.4 (2012): 391-397.