IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud Detection Using Machine Learning

Mr. Ansari Kareem R.
UG Student,
Department of ComputerEngineering
SNDCOE & RC, Yeola

Miss. Tanpure Ashwini N.
UG Student,
Department of ComputerEngineering
SNDCOE & RC Yeola

Miss. Nirmal Divya L.
UG Student,
Department of ComputerEngineering
SNDCOE & RC, Yeola

Prof. Pawar U.B.
Assistant Professor, Department of Computer Engineering
SNDCOE & RC, Yeola

Miss. Pathade Sonali T.
UG Student,
Department of ComputerEngineering
SNDCOE & RC,Yeola

ABSTRACT

The usage of credit cards for online and regular purchases is exponentially increasing and so is the fraud related with it. A large number of fraud transactions are made every day. Various modern techniques like artificial neural network Different machine learning algorithms are compared, including Logistic Regression, Decision Trees, Random Forest, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, and K-means clustering etc. are used in detecting fraudulent transactions. This paper uses genetic algorithm, and neural network which comprises of techniques for finding optimal solution for the problem and implicitly generating the result of the fraudulent transaction. The main aim is to detect the fraudulent transaction and to develop a method of generating test data. This algorithm is a heuristic approach used to solve high complexity computational problems. The implementation of an efficient fraud detection system is imperative for all credit card issuing companies and their clients to minimize their losses.

Keywords: Machine learning, Credit card, Electronic commerce, Fraud detection.

1. INTRODUCTION

A credit card is a thin handy plastic card that contains identification information such as a signature or picture, and authorizes the person named on it to charge purchases or services to his account - charges for which he will be billed periodically. Today, the information on the card is read by automated teller machines (ATMs), store readers, bank and is also used in online internet banking system. They have a unique card number which is of utmost importance. Its security relies on the physical security of the plastic card as well as the privacy of the credit card number. There is a rapid growth in the number of credit card transactions which has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging term for theft and fraud committed

using a credit card as a fraudulent source of funds in a given transaction. Generally, Most of the credit card fraud detection systems are based on artificial intelligence, Meta learning and pattern matching.

2. OBJECTIVE

The Main Objective is to detect online fraud detection when using online financial transaction.
Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to end technology and exploit human vulnerabilities.

3. LITERATURE SURVAY

1] Vimala Devi. J et al. To detect counterfeit transactions, three machine-learning algorithms were presented and implemented. There are many measures used to evaluate the performance of classifiers or predictors, such as the Vector Machine, Random Forest, and Decision Tree. These metrics are either prevalence-dependent or prevalence-independent. Furthermore, these techniques are used in credit card fraud detection mechanisms, and the results of these algorithms have been compared.
2] Popat and Chaudhary. supervised algorithms were presented Deep learning, Logistic Regression, Nave Bayesian, Support Vector Machine (SVM), Neural Network, Artificial Immune System, K Nearest Neighbour, Data Mining, Decision Tree, Fuzzy logic based System, and Genetic Algorithm are some of the techniques used. We compared machine-learning algorithms to prediction, clustering, and outlier detection.
3] Deepa and Akila . For fraud detection, different algorithms like Anomaly Detection Algorithm, K-Nearest Neighbor, Random Forest, K-Means and Decision Tree were used. Based on a given scenario, presented several techniques and predicted the best algorithm to detect deceitful transactions. To predict the fraud result, the system used various rules and algorithms to generate the Fraud score for that certain transaction.
4] Kibria and Sevkli. Using the grid search technique, create a deep learning model. The built model's performance is compared to the performance of two other traditional machine-learning algorithms: logistic regression (LR) and support vector machine (SVM). The developed model is applied to the credit card data set and the results are compared to logistic regression and support vector machine models.
5] Borse Suhas and Dhotre Machine learning's Naive Bayes classification was used to predict common or fraudulent transactions.
6] Asha R B et al. have proposed a deep learning-based method for detecting fraud in credit card transactions. Using machine-learning algorithms such as support vector machine, k-nearest neighbor, and artificial neural network to predict the occurrence of fraud.

4. PROBLEM DEFINITION

There are lots of issues that make this procedure tough to implement and one of the biggest problems associated with fraud detection is the lack of both the literature providing experimental results and of real-world data for academic researchers to perform experiments on. The reason behind this is the sensitive financial data associated with the fraud that has to be kept confidential for the purpose of customer's privacy. Now, here we enumerate different properties a fraud detection system should have in order to generate proper results. The system should be able to handle skewed distributions, since only a very small percentage of all credit card transactions is fraudulent. There should be a proper means to handle the noise. Noise is the errors that is present in the data, for example, incorrect dates. This noise in actual data limits the accuracy of generalization that can be achieved, irrespective of how extensive the training set is. Another problem related to this field is overlapping data. Many transactions may resemble fraudulent transactions when actually they are genuine transactions.

5. SYSTEM ARCHITECTURE

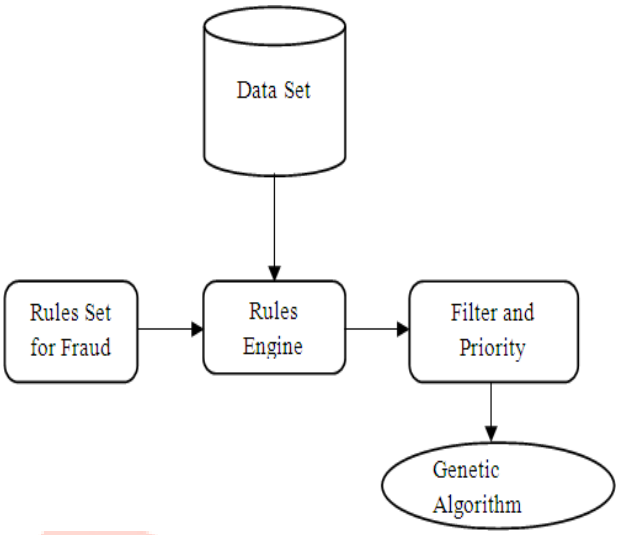


Fig. System Design

The above architectural design describes the work structure of the system. The data warehouse contains the customer data. this customer data is subjected to the rules engine and again ,the rules engine comprises of the rules set.
The filter and priority module sets the priority for the data and hence. Plays a very important role in the system. then the filter data is sent to the genetic algorithm module which performance its functions and generates the output.

6. ALGORITHM

6.1] Random Forest

Classifier fined decision trees in a subset of the data and then aggregates their information to that to get the full dataset's predictive power. Rather than relying on a single decision tree.
The RF takes the predictions from each tree and forecasts the final output based on the majority votes of forecasts. Using a huge number of trees in the forest improves precision and eliminates the issue of over fitting. It predicts output with high precision, and it runs efficiently even with large datasets. It can also keep accuracy when a large proportion of data is lost.
Random Forest can handle both classification and regression tasks. It can handle large datasets with high dimensionality. It improves the model's accuracy and avoids the over fitting problem. We use two-step training techniques in the process of tree-based Random Forest: First, we generate the random forest by mixing N trees together, and then we estimate for each of the trees we generate in the first phase.

6.2] Logistic Regression

An algorithm that can be used for both regression and classification tasks, but it is most commonly used for classification.' Logistic Regression is used to predict categorical variables using dependent variables. Logistic Regression employs a more complex cost function, this cost function is known as the

Sigmoid Function or the Logistic Function.LR also does not require independent variables to be linearly related, not does it require equal variance within each group, making it a less stringent statistical analysis procedure.

6.3] Naive Bayes

The naive Bayes families of statistical algorithm are some of the most used algorithm in text classification and test analysis. Naive Bayes Algorithm is extremely fast relative to other classification algorithms.
Naive Bayes classifier is a collection of classification algorithm based on Bayes Theorem.
It is not a single algorithm but a family of algorithm where all of them share a common principle.

7] EXPERIMENTAL SETUP

In this section, we report our experimental study that we performed with selected machine learning algorithms and imbalance classification approaches. First, we provide a detailed description of the design of experiments followed by the results and a discussion. Finally, we discuss some critical shortcomings we discovered in our experiments.

Design of Experiments

This section briefly presents the workflow of our experiments, the dataset used, the selection of target variables and performance measure.

A] Workflow of Experiments

Our experimental study is organized as follows. The experiments are presented and discussed in two phases. In the first phase, eight classification methods are compared. The comparison was carried out with respect to three parameters including the following: accuracy, sensitivity, and the Area under Precision-Recall Curve (AUPRC). This comparison results in selecting the most suitable algorithms including the SVM and ANN.
In the second phase, the selected algorithms are used in comparing selected imbalance classification approaches such as Random Oversampling, One-Class Classification and Cost Sensitive. Then, the SVM is used as a binary classification tool, and compared to the One-Class Classification SVM and Cost Sensitive SVM. Also, the ANN is applied and compared to the Auto-Associative Neural Network.

B] Dataset and Variable Selection

The dataset used in our experiment contains credit card fraud labeled data. It contains ten million credit card transactions described by 8 variables listed here:

- Cust ID is an auto increasing integer value that represents the customer ID: This variable is removed later as it has no relevance for detecting fraud.
- Gender: represents the customer's gender.
- State: represents the state in which the customer lives in the United States.
- Card holder: is the number of cards that the customer holds (maximum 2).
- Balance : indicates the balance on the credit card in USD.
- Num Trans: is a discrete variable that represents the number of transactions made to date.
- NumInt Trans: is a discrete variable representing the number of international transactions made to date.
- Credit Line: denotes the customer's credit limit.
- Fraud Risk: the binary target variable, taking the values 0 denoting legitimate transaction, and 1 denoting fraudulent transaction.

8] APPLICATIONS

- net banking
- e-commerce

9] ADVANTAGES

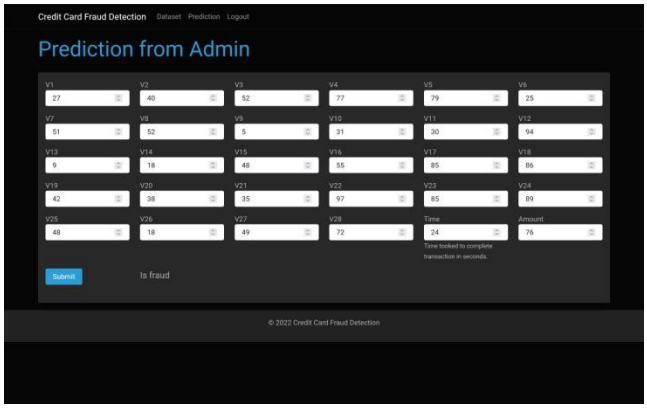
- Higher accuracy of fraud detection.
- Less manual work needed for additional verification.
- Fewer false declines.
- Ability to identify new patterns and adapt to changes.

10] OVERVIEW OF PROJECT MODULES

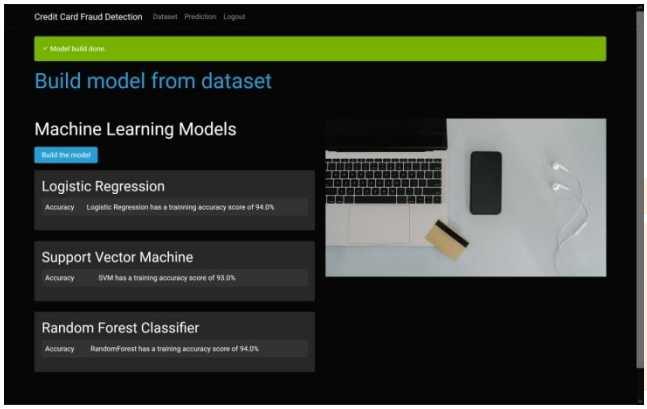
User login

Welcome to credit card fraud detection

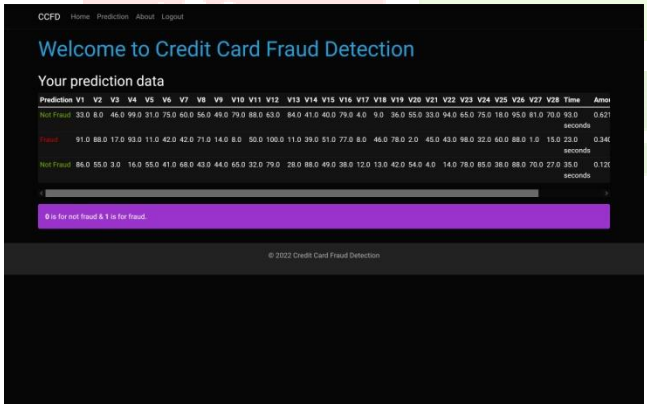
Prediction from Admin



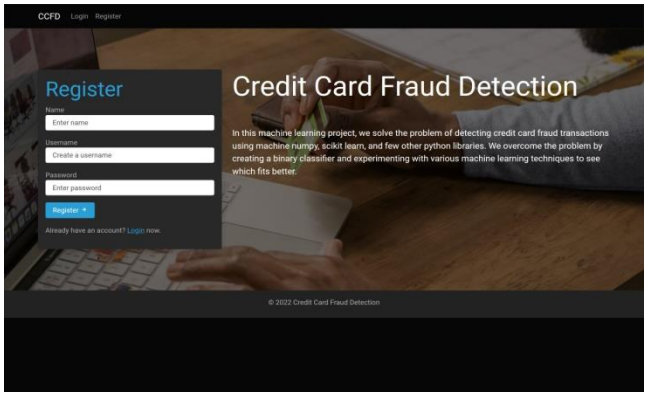
Build model from dataset



Welcome to credit card fraud detection



User register



11] IMPLEMENTATION PART

• In this module, we'll gather all of the credit card data and save it to a folder. The dataset would then be subjected to descriptive analysis.

• After reviewing the dataset, we must clean the data in the next phase. Both redundant values and null Values in the dataset will be deleted during this cleaning step, and a new dataset will be created. •

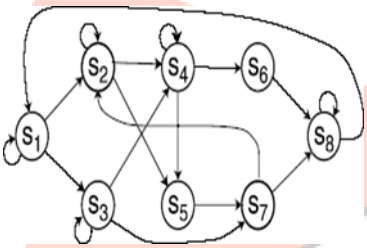
The cleaned dataset will be reprocessed in this module, with the dataset being grouped by volume and transaction period.

• The dataset will be split into two parts in this module: qualified dataset and testing dataset. The Random Forest Algorithm is used after the data has been partitioned. Finally, a confusion matrix is obtained after using the Random Forest Algorithm.

Evaluation Now that the resulting data in the form of an uncertainty matrix has been obtained, it can be analyzed using a graphical representation, which provides greater precision.

12] MATHEMATICAL MODEL

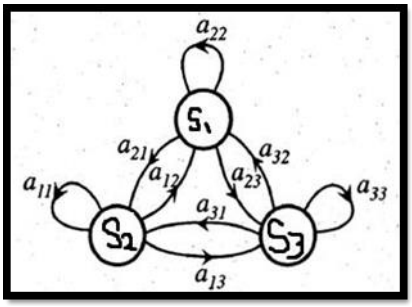
i] N is the number of states in the model. The set of states is denoted as S={S1 ; S2 ; . . . SN} N is an individual state. The state at time instant t is denoted by qt



In the above diagram we have a HMM with 8 stated state S1 to state S8.

ii] M is the number of distinct observation symbols per state. The observation symbols correspond to the physical output of the system being modeled. We denote the set of symbols V= {V1 ; V2 ; . . . VM}

Eg:- in the above diagram we can say that the observation symbol of the state S1 is V1 state S2 is V2 and so on till state S8 is V8



The above diagram shows a hmm with 3states and aij as the state transition probabilities.

The observation sequence O = O_1, O_2, O_3,... OR, where each observation O_t is one of the symbols from V, and R is the number of observations in the sequence. It is evident that a complete specification of an HMM requires the estimation of two model parameters, N and M, and three probability distributions A, B, and π.We use the notation λ = (A, B, π) to indicate the complete set of parameters of the model, where A, B implicitly include N and M. An observation sequence O, as mentioned above, can be generated by many possible state sequences. Consider one

such particular sequence $Q = q_1, q_2, \ldots\ldots, q_R$
Where q_1 is the initial state
where q_1 is the initial state

13] CONCLUSION AND FUTURE SCOPE

Credit card fraud becomes a serious concern to the world. Fraud brings huge financial losses to the world. This urged Credit card companies have been invested money to create and develop techniques to reveal and reduce fraud. The prime goal of this study is to define algorithms that confer the appropriate, and can be adapted by credit card companies for identifying fraudulent transactions more accurately, in less time and cost. Different machine learning algorithms are compared, including Logistic Regression, Decision Trees, Random Forest, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, and K-means clustering. Because not all scenarios are the same, a scenario-based algorithm can be used to determine which scenario is the best fit for that scenario

14] ACKNOWLEDGMENTS

A very firstly we gladly thanks to my project guide **Prof. Pawar U.B.** For his valuable guidance for implementation of proposed system. We will forever remain a thankful for their excellent as well as polite guidance for preparation of this report. Also we would sincerely like to thank to **HOD Pawar U.B.** and other staff for their helpful coordination and support in project work.

15] REFERENCES

1] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong -''A survey on enhanced subspace clustering,'' Data Mining Knowledge Discovery, vol. 26, no. 2, pp. 332–397, 2019.
2] S. Mcskimming -''Trade-based money laundering: Responding to an emerging threat,'' Deakin Law Rev, vol. 15, no. 1, 2020.
3] Nitu Kumari, S. Kannan and A. Muthukumaravel - "Credit Card Fraud Detection Using Genetic-A Survey" published by Middle-East Journal of Scientific Research , IDOSI Publications, 2014
4] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey - "A Tool for Effective Detection of Fraud in Credit Card System", published in International Journal of Communication Network Security ISSN: 2231 – 1882, Volume-2
5] S.H. Projects and W. Lovo , ―JMU Scholarly Commons Detecting credit card fraud : An analysis of fraud detection techniques,‖ 2020. [2] S. G and J. R. R, ―A Study on Credit Card Fraud Detection using Data Mining Techniques,‖ Int. J. Data Min. Tech. Appl., vol. 7, no. 1, pp. 21–24, 2018, doi: 10.20894/ijdmta.102.007.001.004
6] V. N. Dornadula and S. Geetha ―Credit Card Fraud Detection using Machine Learning Algorithms, Procedia Comput. Sci., vol. 165, pp. 631–641, 2019, doi: 10.1016/j.procs.2020.01.057.
7] A. H. Alhazmi and N. Aljehane - A Survey of Credit Card Fraud Detection Use Machine Learning, 2020 Int. Conf. Computes. Inf. Technol. ICCIT 2020, pp. 10–15, 2020, doi: 10.1109/ICCIT-144147971.2020.9213809
8] M. Kanchana, V. Chadda , and H. Jain - Credit card fraud detection,‖ Int. J. Adv. Sci. Technol., vol. 29, no. 6, pp. 2201–2215, 2020, doi: 10.17148/ijarcce.2016.5109. [14] A. RB and S. K. KR, ―Credit Card Fraud Detection Using Artificial Neural Network,Glob. Transitions Proc, pp. 0–8,2021,doi: 10.1016/j.gltp.2021.01.006