IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF CREDIT CARD FRAUD USING MACHINE LEARNING

[1] Mrs. T. Sujatha Jayakrishnan [2] Mr. M. Kannan
[1,2]Assistant Professor
[1,2]Department of Computer Science,
[1,2]SRM Arts and Science College,Chennai

Abstract- Credit card fraud has become a major problem worldwide. Due to the fact that credit cards are the most widely used payment method for both traditional and online purchases, there are an increasing number of fraud incidents. Credit card firms must be able to identify fraudulent transactions in order to prevent charging customers for goods they did not buy. There are several ways that credit card fraud can occur, but the most frequent ones are lost, stolen, non-existent, and card skimming.Several machine learning models are applied to each fraud instance, and the most effective strategy is chosen after assessment.. The set of data serves as the algorithm's input. Training is done using the sample data. Furthermore, the mathematical values for classification are dealt with in the suggested model's training.

Keywords- K-Nearest Neighbor, Naive Bayes,, SupportVector Machine, Random Forest

I INTRODUCTION

In recent years, online payment methods havebecome widespread due to the rapid growth of cashless electronic payments. One of the most popular electronic payment methods utilized today is the credit card. My banks deal with credit card fraud on a global scale, which may be broadly classified into four categories: lost, non-existent, skimming, and hacking cards.

To complete the transaction, debit or credit card information was required. Here, your card information suffices to complete online transactions; a password is not required. We employ the following techniques to stop this: To accept or deny the transactions, we employ the fingerprint or device pattern tool.. We're also able to search the fraud card blacklist

Details; if they are already available, we do not endorse them. By monitoring anomalies, we can identify transactions for which there is no favorable policy by using the equipment. Your account's transactions that prevent a fraudster from making a sizable transaction can be marked by them. This generally takes place.Fraudsters cannot employ card wide variety software program generator to reach this barrier and attempt

to take a look at the transaction speed by restricting the amount of transactions that can be attempted. This serves as a layer of customer safety.

Credit card fraud can be found in a number of ways. The discussion that follows is relevant to the topic of machine learning (ML), as that was the primary focus of this study. The term engineering was first introduced in 1959 by Arthur Samuel, who uses training data to create a mathematical model that makes predictions. We use algorithms based on supervised learning. SVM, Naive Bayes, KNN, Random Forest, and Logistic Regression are the algorithms. This paper is divided into seven sections.

Section II provides a literature review of previous work. Section III presents the approaches used to systematically select elementary studies. Section IV. Several popular credit card fraud detection techniques are briefly presented. Section V provides an analysis and comparison of several algorithms. The debate and results are summarized in Section IV. Lastly, conclusions and references are provided in Section VII.

II. LITERATURE SURVEY

Title: Supervised machine learning algorithms for credit card fraudulent transaction details: a comparative study Year: 2018

Author: SahilDhakhad; EmadMohammmed; BehrouzFar

Methodology

Finding hidden patterns and applying them to regular decision-making in a range of contexts is the aim of data analytics. Today's scammers are increasingly focusing on credit card fraud as a target. In recent years, credit card fraud has cost the financial sector billions of dollars, making it a severe issue. Because there is a dearth of real transaction data and an imbalance in publicly available datasets, developing fraud detection algorithms is a challenging challenge. In order to detect fraudulent transactions, this paper employs numerous supervised learning approaches on a real dataset. To detect fraudulent credit card transactions, we employ a crucial variable. Furthermore, we evaluate how well the super classifier used in this work performs in comparison to different algorithms from the literature. They used supervised system detection algorithms on an actual global dataset, applied the algorithms to build a first-class classifier through supervised learning, and then contrasted the overall effectiveness of the supervised algorithms with their exceptional classifier implementation. Ten supervised learning algorithms were employed, including Random Forest, KNN, Choice Tree, Gradient Boosting, XGB Classifier, Random Forest, Logistic Regression, MLP Classifier, SVM, and Naïve Bayes. They contrasted the outcome of their super classifier—which was used in this work—with the accuracy, precision, and confusion matrix.

According to this paper, the regression algorithm finds the fraud detection level with high prediction when compared to others.

Title: Credit card fraud detection using ML Year: 2020

Author: Ruttala Sailusha; V. Gnaneswar; R.Ramesh; G. Ramakoteswara Rao Methodology

Frauds involving credit cards are simple and easy to target. Numerous online platforms, including e-commerce, have raised the internet

Right now, identifying credit card fraud is the most prevalent problem in the modern world. These have emerged from the growth of e-commerce platforms and online transactions. A credit card is typically fraudulently obtained when it is stolen and used for any illicit purpose, or when the cardholder uses the card information for their own benefit. We are currently dealing with a large number of credit card issues. Credit card fraud detection technology was developed as a result of identifying fraudulent activity. Essentially, the main focus of this work is machine-learning techniques. The confusion matrix is used as the basis for plotting the ROC curve. The best algorithm for detecting fraud is said to be the one that compares the Random Forest and Adaboost algorithms in terms of accuracy, precision, recall, and F1-score. This study claims that bank systems can lower the rate of fraud among clients and increase their perception of our bank's trustworthiness by utilizing these findings..

Title: Utilizing Machine Learning to Identify Credit Card Fraud.

2021 is the year

Author: Yuxin Gao, Shouming Zhang, Jiapeng Lu

Methodology

Due to the problem of credit card fraud, credit card operations are going through an unknown evolution despite the notable improvement of electronic banking. This was the main issue, and the only way to solve it is to have automated systems finish and identify fraud discovery. Our staff is unable to collect all of the covered account data due to the sheer volume of accounts. Since there are thankfully far fewer fraudulent transactions than legitimate ones. Data distribution is unreliable and biased toward non-fraudulent compliances. Numerous learning techniques are available to counteract the outcomes of unstable dataset difficulties; other approaches akin to oversampling or undersampling are also available to improve the delicacy or vaticination.

Building and constructing a novel fraud discovery system for streaming transaction data is the main objective of the study. The ideal result would be the analysis of guests' once-sale data and the detection of recurrent behavioral patterns. where cardholders are divided into various groups according to the amount of sales they have made. Additionally, using the sliding window approach (1), add up the sales made by the cardholders from various groups so that each group's behavioral pattern can be disrupted separately.Latterly different classifiers (3),(5),(6),(8) are trained over the groups independently. Additionally, the classifier with the highest standing score may be selected as one of the fashionable approaches to predict frauds. Thus, in order to address the issue of conception drift, a feedback channel comes next (1). In this work We used a dataset of credit card fraud across Europe.

Title: Autonomous credit card fraud detection usingmachine learning

Year: 2022

Author: **J Femila Roseline, GBSR Naidu, V. Samuthira Pandi, S Alamelu alias Rajasree, Dr. N. Mageswari Methodology**

In recent years, credit card fraud in sensitive commodities has increased as more individuals pay for things with credit cards. This is related to the growth of internet commerce and technological advancements, both of which have led to massive financial losses due to fraud. It is necessary to create and implement an efficient fraud discovery system in order to lower losses of this kind. The use of machine literacy techniques to automatically identify credit card fraud ignores behavioral issues or deception processes, which could result in warnings. The goal of this research is to identify indicators of credit card fraud. A Long Short-Term Memory-intermittent Neural Network (LSTM-RNN) is proposed to characterize the fraud scenario. To improve performance even further, an attention medium has also been added. Models with this structure have proven to be particularly useful when dealing with scenarios such as fraud identification, when the information sequence consists of vectors with closely connected parcels. LSTM-RNN is compared with other classifiers such as Naive Bayes, Support Vector Machine (SVM), and Artificial Neural Network (ANN). Experiments demonstrate that our suggested paradigm yields significant outcomes with a high degree of delicacy.

Proposed Model

The proposed model shows the main steps for pre- processing stage feature extraction and classification. The first process is to pre - process the dataset to remove missing values and null values from the taken dataset.

- Highly effective
- Gives accurate prediction result

III COMPARSION OF DIFFERENTML ALGORITHM'S

The five approaches used in this work are Gaussian NB, Random Forest, K-Neighbour Classifier, Logistic Regression, and Support Vector Machine. In our experiments with these algorithms, we used the grid search algorithm to determine which algorithmic parameters provide the best accuracy for our model.

SVM: Because it performs well in nonlinear classification tasks and can handle the uneven structure of the data, we chose to employ the SVM technique.

Logistic regression: When applied to data with associated qualities, logistic regression performs well. It requires very little in the way of computer resources. Because it is easy to develop, we may use it as a benchmark before going on to other algorithms. It typically yields the finest categorization method outcomes.

K-nearest neighbour classifier: For handling noisy information, it works perfectly. Using both class types—binary grandness and multi-grandness—we may employ this retrieval-based strategy without putting in extra work. We may also employ regression and categories. Prior to every parameter thereafter matching the initial parameter, parameter selection is challenging.

The Random Forest: There is no need to rescale or modify facts when using it. Classification and regression problems can be addressed with it. With each tree having a high variance and sporadic biases, the algorithm separates the datasets according to their features, producing a final result that is extremely good. Fast feature loss and error correction in the dataset are both supported by the approach, which trains the version of the model extremely quickly.

Gaussian NB (Naïve Bayes): Even with real-time data, the algorithm remains accurate because it is fully based on conditional possibilities. It might result in a top-notch recommendation system. Large data sets can be used with it. It determines the conditional possibility using a formula.

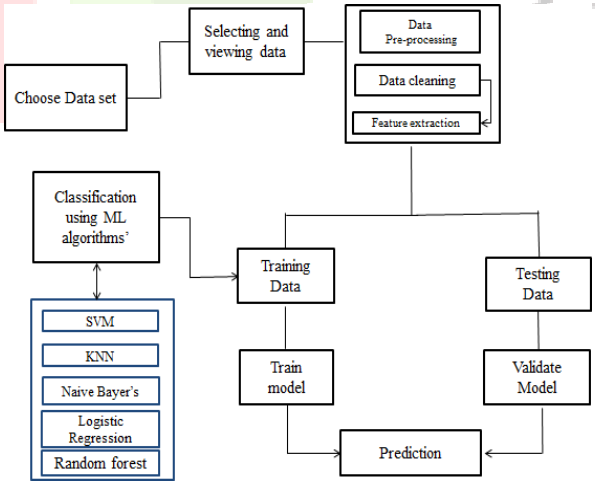
The formulation is

P (A|B) = P ((B|A) * P (A))/P (B),

where P (A|B) = later chance, P (B|A) = earlier chance, P (A) = probability, P (A) = evidence. The result is a probabilistic prediction with a less trained data set. Both continuous and discrete facts can be handled by it.

Steps involved are

- Data set
- Data Preprocessing
- Feature Selection
- Classifying



- Predicting

- Generating results

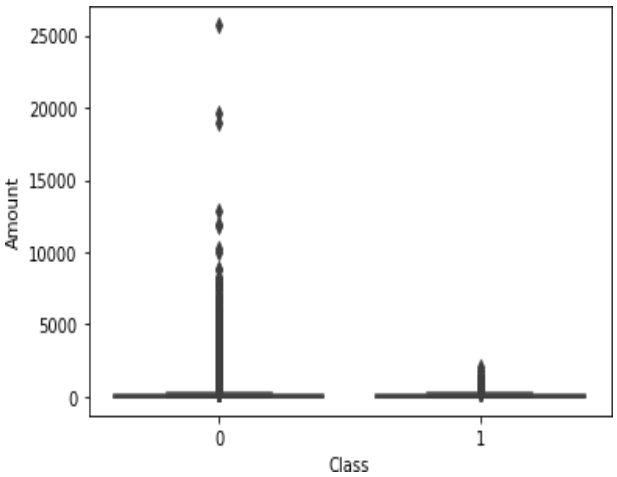


Figure 2: In this case, 0 denotes positive transitions while 1 denotes fraudulent transactions.

IV. VIEW OF THE DATASET

Using a dataset including 284807 records of transactions performed by European cardholders in just two days in September 2013, we are able to identify credit card fraud. 492 transactions out of these are fraudulent, while 284,315 are legitimate. The fraud rate is 0.172% when looking at the full dataset. 31 sensitive features in the dataset are likewise hidden and designated V1 through V28. These attributes are kept private. The data was converted using PCA (principal component analysis), which decreased the dimensionality and produced input data that was only numerical values, making the data more efficient. Time and amount are the only features that PCA did not alter. Our target column is the feature class, where 1 denotes fraudulent transactions and 0 denotes positive transactions.

Figure 2: In this case, 0 denotes positive transitions while 1 denotes fraudulent transactions.

Libraries used

There are 4 libraries used in the test to createthe effects, namely

Pandas: In order to arrange the data into record frames and perform various operations on them, it is utilized to review Excel and CSV files..

Matplotlib: is used to create two-dimensional graphs representing the datasets. With its assistance, we may control the graphs' size and color and create a variety of unique graph styles..

Seaborn: is also used for creating graphs and visualizing data; however, because it offers a variety of distinct graph types, such as boxplots and heat maps, the graphs made with it are somewhat more specialized than those made with Matplotlib.

Scikit-learn: This tool is used to import algorithms in the SVC, KNN, regression, classifier, and other classes.

Implementing Algorithm's

As we apply SVM (assist vector set of rules) to the dataset, we divide it into training records and testing records by using a parameter in our train-observe-break up algorithm. We specify the test length as a parameter, which for the dataset is equal to zero.25. Consequently, the whole material may be split into two sections: 0.75 for educational activities and 0.25, as stated, for testing reasons. We also try to target elegance in distinctive variables, avoiding all of the qualities. After these, we form a model of our SVC technique wherein schooling records are suited to teach the model. as soon as the version is ready it predicts the values for our checking out information. As a result, we get the type record and confusion matrix.

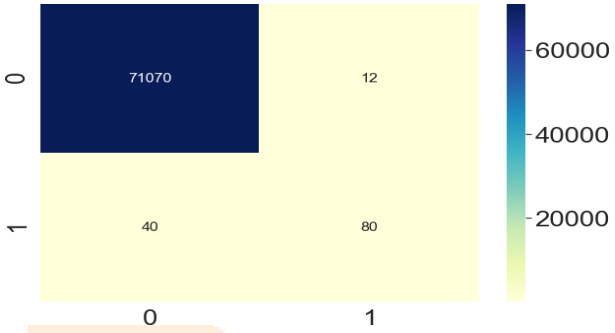


Figure 3: confusion matrix of SVM

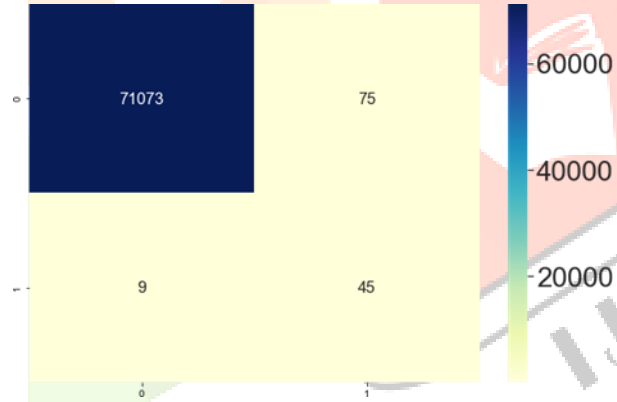


Figure 4: confusion Matrixof Logistic Regression and Naïve Bayes

The SVM algorithm produced the aforementioned matrix, which contains real benefits of 71073, false disadvantages of 75, fake high-quality of 9, and genuine disadvantages of 45. According to our confusion matrix, the correct anticipated values for this model are 71, 118, while the incorrect projected values are 84.

While we utilize the MCC (Matthews Correlation coefficient) for binary class as the standard overall performance dimensionscore, the outstanding result for SVC is 0. 558.

All other algorithms follow the same process, but use different methodologies. Similarly, when using Logistic Regression Gaussian NB, ok-acquaintances Classifier, and Random wooded area Classifier are the employed algorithms, although Naive Bayes, k-nearest neighbor, and Random forest are not. Consequently, we created the Confusion matrix. The Matrix seen above is the outcome of Logistic Regression. Whereby actual exceptional is 71070, fake excellent is forty, fake poor is 12, and true poor is 80. The correct prediction values

in this model are 71,150, whereas the incorrect predictive values are 52. The MCC for the Logistic Regression variant is zero.761. It is far superior to use MCC's first-rate rating of 1 for credit card fraud detection as opposed to MCC's highest score.

For this dataset, the same matrix is produced using both the NaveBayes set of rules and logistic regression. This indicates that, for the specified dataset, the MCC rating obtained with the Nave Bayes algorithm is 0.761. Because of the f1 score and don't forget values that are exclusive in every approach, the algorithms can also yield exclusive findings for any other credit score card fraud detection dataset, but the size of the dataset may vary.

Table: Recall values of all ML algorithms

Models	Recall value
SVM	0.92
Naïve Bayes	0.91
Logistic Regression	0.83
KNN	0.84
Random Forest	0.89

Table: Confusion matrixfor K nearest neighbor

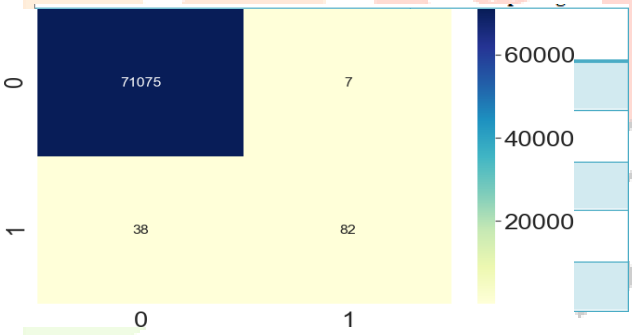


Figure 5: confusion matrixfor Random Forest

The KNN set of rules culminates in the matrix that is displayed above. For this model, the accurate expected values are 71,157, while the incorrect expected values are 45. The real high-quality values are 71075, the false superb values are 38, the fake-poor value is 7, and the genuine negative values are 82. The MCC rating for this KNN version is 0.793.

The matrix shown above is the end result of the KNN set of rules. where the real high-quality values are 71075, the false superb values are 38, the fake-poor value is 7, and the genuine negative values are 82 The correct expected values for this model are 71,157, while the incorrect expected values are 45. This version of KNN has an MCC rating of 0.793.

Table: F1-score values of all ML algorithms

Models	F1-score
SVM	1.00
Naïve Bayes	0.98
Logistic Regression	1.00
KNN	1.00
Random Forest	1.00

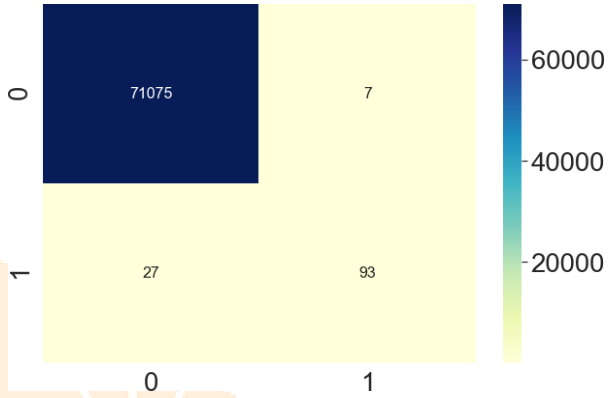


Figure 6: MCC values of all ML algorithms

Looking at the comparison table now, we can see that MCC's good score, which was determined by using Random forest with random parameters, is zero.848. The Random Forest set of rules was then selected, and the optimal parameters were found using the Grid Search method. A model was then built using the revised parameters, and the results were compared.

IV OUTPUT SCREENSHOTS

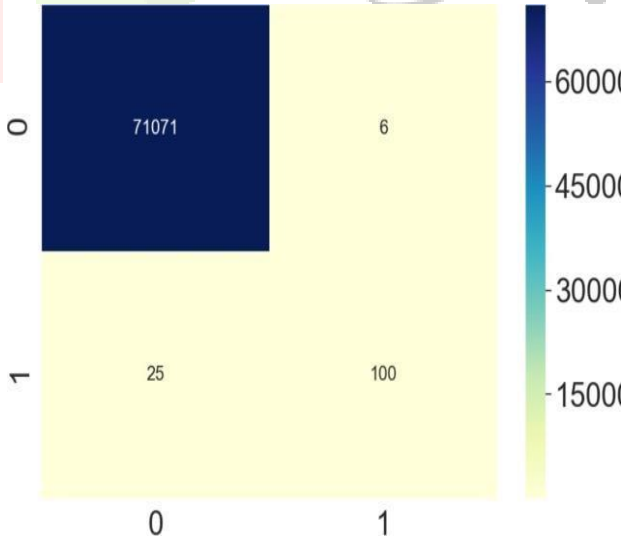


Figure 7 Algorithm classification report

The matrix above displays the Random forest output with the Grid seek parameters. The parameters are as follows: criterion = entropy; max-features = vehicle; max-depth = 10; n estimators = 500. It has 71071 true wonderful values, 6 false awful values, 25 fake terrific values, and 100 actual poor values as a result of the confusion matrix. This indicates that 31, rather than 71,171, are the incorrectly projected values. The MCC rating of the new resulting algorithm is 0.89.

‘

Models	Recall value	F1-score	MCC value
Random Forest with random parameters	0.89	1.00	0.848
Random Forest with Grid Search parameters	0.90	1.00	0.89

We can see from the table above that the nice fee compared to the first-rate rating of MCC is 1 .With new parameters produced by the Grid search algorithm, the Random Wooded Area algorithm yields the closest cost, which comes out to be 0.89. We will now infer that we are getting some better results from where we are.

V CONCLUSION

In this test, the best algorithm for credit card fraud detection is found by utilizing system learning approaches with the credit score card fraud detection dataset. Typically, five approaches are used: SVM, Nave Bayes, Logistic Regression, KNN, and Random Woody Area. Random woodland area and KNN follow in order to offer a suitable score result. The MCC score, which ranges from -1 to 1, is optimal since it is utilized to assess an algorithm's efficacy. Random Woodland's score of 0.848 is the closest to 1 when it comes to MCC assessed values completely. The result of Random Woodland advanced slightly when the Grid search method was used on it and the version was trained again with new parameters. The MCC cost of Random Woodland climbed by 0.848 to 0.89, which is again quite close to the pleasant MCC rating of 1. This leads us to the conclusion that, in terms of credit card fraud detection, the Random Wooded Area set of rules produces the greatest results.

We might include new technologies and mix them with current algorithms to enhance credit card fraud detection algorithms and produce more accurate results. In an attempt to support our early fraud detection and reduction efforts. The total amount of money lost by credit card scammers will decrease as a result.

REFERENCES

1. Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang, Random Forest for Credit Card Fraud Detection,2018 (IEEE).

2. Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla, Credit Card Fraud Detection - Machine Learning methods, Publish in:18th International Symposium INFOTEH-JAHORINA, 20-22 March 2019 (IEEE).

3. Shail Machine, Emad A. Mohamad, Behrouz Far, Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study, 2018(IEEE) computer society, pp.122-125.

4. SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi,A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.

5. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, Ashoke

K. Nandi,Credit Card Fraud Detection Using AdaBoost and Majority Voting, Published in: IEEE Access on 15 February 2018, vol. no.6, pp. 14277 – 14283.

6. N. Sivakumar, Dr.R. Balasubramanian, Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques, Published in (IJCSIT) International Journal of Computer Science and Information Technologies, vol no. 6 (2),2015, pp. 1379-1386.

7. Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, Maheshwar Sharma, Credit card fraud detection using Naïve Bayes model based and KNN classifier, Published in: International Journal of Advance Research, Ideas and Innovations in Technology, Issue no. 3, vol.no. 4, 2018, pp.44-47.