As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 7(12): 129-133, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0712022
Available online at: http://www.ijmtst.com/vol7issue12.html

Credit Card Fraud Detection System

Aashish Jha | Vasudha Bahl | Nidhi Sengar| Amita Goel

Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India.

To Cite this Article
Aashish Jha, Vasudha Bahl, Nidhi Sengar and Amita Goel. Credit Card Fraud Detection System. *International Journal for Modern Trends in Science and Technology* 2021, *7* pp. 129-133. https://doi.org/10.46501/IJMTST0712022

Article Info
Received: 31 October 2021; Accepted: 05 December 2021; Published: 08 December 2021

ABSTRACT

Fraud is one major issue in credit card industry. It is essential to distinguish between authentic and fraudulent transactions and to handle such issues, we could use machine learning and artificial intelligence. This main aim to, firstly, to identify the different types of threats, and to analyze the other alternative techniques that are being used in this industry. This paper aims to deliver model that can be used to check if a transaction is fraudulent or not, with maximum efficiency and accuracy. The proposal declared in this paper are likely to have valuable in terms of time efficiency and cost minimization. Although, there's always a scope of improvement and resolve the issue where genuine customers are misclassified as fraudulent.

KEYWORDS: Credit Card Fraud, Credit Card Detection Techniques, Credit Card Fraud Detection using ML.

INTRODUCTION

In this 21st century, one thing that's growing rapidly is the digitalization. The internet is at boom and soaking the whole world in it. Everyone is using the internet in one way or another. The main reason behind this is ease in accessing data, though we are not aware that it could also make us vulnerable. The digitalization in banking sector has been improvising day by day, a user can easily transfer money to one another in many ways through internet. To give an estimate, there were roughly around 369 billion transactions happened for good and services worldwide in 2018, that means nearly over 1 billion transactions every day. Though the internet made our lives easier, it could be destructive sometimes. We all know with ease of data comes vulnerabilities, security threats, we are heading towards a time internet is everything but also, we cannot unveil our privacy in wrong hands. With these transactions, the cyber attacking is also rising, fraud related to banking transactions increasing day by day. 'Fraud' as the term conveys, unauthorized usage of an account of someone another without their grant. These frauds happen with credit card, debit, and in many ways with the intention of stealing money. These fraud causes billions of dollars damage every year. The credit card companies are improvising their security day by day but there's a very chance of happening these frauds. With the advancement in technology and machine learning, now there are several advance ways in which we can stop these fraud transactions before happening.

Fraud detection is a method which consists examining the past activities, analyze the user behavior, learn through these activities and predict the future fraud transactions to avoid these intruding, obtrusive activities from happening. With advance machine learning and artificial intelligence, today's computer can learn to objectify between these transactions, and with use of it, we can lower the amount of these fraud transactions even before happening. In order to protect their customers, companies rely on these fraud detection and prevention software to analyze credit card purchases.

Now, the detection system softwares are improvising day by day but there's always a high learning curve and scope of improvement in thesesystems. These systems highly dependent on several factors like past data, no. of valid transactions, their patterns, no. of fraudulent transactions, their statistical properties over the course of time.

The software used to identify these frauds comprises of different algorithms, models, there's already been significant improvement in these. To effectually detect credit fraud, we need to know the various technologies, and types of credit card frauds. Different companies have different technologies, and all have their own advantages and accuracy. Many software uses one or more machine learning models like Linear Regression, Random Forest Algorithm, K-nearest neighbour, Neural Network and much more… These algorithms try to filter the data on their ability of logic. In this learning, we are trying to also trying to create a model which can be used to identify these fraudulent transactions before successfully transact and prevent it from happening. We are going to used different machine learning techniques, and models to achieve the maximum accuracy and with efficient performance.

LITERATURE REVIEW

Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. They generally used data mining techniques, adversarial detection and automated fraud detection techniques. Even though these methods and algorithms achieved an unforeseen success in some areas, they failed to provide an enduring and reliable solution to fraud detection. A similar research domain was presented by Wen-Fang YU, Na Wang where they used Outlier mining, OM and Distance sum algorithms to precisely predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. It is a field of data mining which is basically used in internet and monetary fields. It is generally used for detecting objects that are detached from the main system i.e., the transactions that aren't authentic. They have taken traits of customer's behavior and based on the value of those tra they've calculated that distance between the observed value of that attribute and its predetermined value.

In order to get a proper detailed analysis, we need to know the types of challenges that we could face during our experiment.

A. Challenges in these fraud detection techniques

- Imbalanced data: Generally, most of data (99.8) are non-fraudulent which makes it very difficult for a machine to detect fraudulent ones.
- Data Availability: The data we need to use to train our model are private data which cannot be found easily as these are generally private.
- Huge Data: As we talked above, there are billions of transactions every year and in order to process these data we need a highly efficient and fast processing model.
- High advanced techniques are used by scammers and they know how to breach into the system by finding loopholes.
- Misclassified Data: Data that we get are not 100% accurate and there's a very chance of not covering every fraudulent transaction that is caught or reported.

B. Adaptive solutions to tackle these challenges

- We need to use some techniques to balance the data which we could use further in our procedure.
- To protect the privacy and anonymity of users we could reduce the dimensionality of data.
- The model must be efficient and simple to use which could detect and classify the fraudulent and non-fraudulent transactions in less time as possible

We could make our model more interpretable so we a hacker or scammer try to adapt our technology we could re-design our model instantly.

METHODOLOGY

The detection of a fraud is a complex computational task and still there is no system that could predict the future transactions accurately. However, a good model must consist three properties written below:

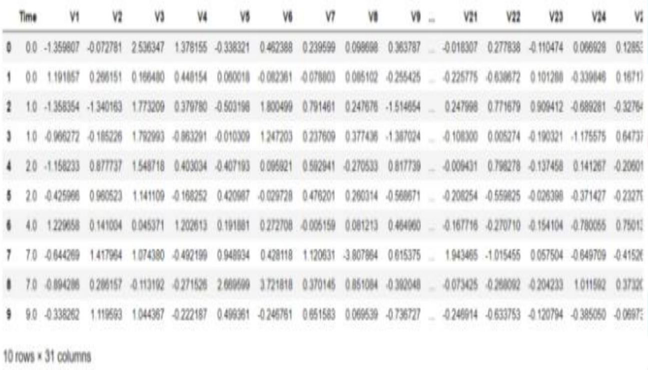
- It should be able to identify accurately.
- It should be able to detect the frauds as quick as possible.

- It should never classify the authentic transaction as fraudulent one.

In this research, we are going to use various predictive model to see how accurate they are, and analyze how different they work in diverse scenarios and trying to find out the best case.

Data Set

The dataset we are going to use is of September2013. The dataset contains transactions made by European countries. This dataset present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly imbalanced, the positive class(frauds) account for 0.17% of all transactions. Due to high imbalance in data, we also need to use some technique to filter out the data. As a security concern, the genuine variables have not been shared publicly but – they have been transformed versions of PCA. As a result, we can find 1 final class column and 29 feature columns.



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.1285
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.1671
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.3276
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.6473
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.2060
5	2.0	-0.425966	0.960523	1.141109	-0.168252	0.420987	-0.029728	0.476201	0.260314	-0.568671	...	-0.208254	-0.559825	-0.026398	-0.371427	-0.2327
6	4.0	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.005159	0.081213	0.464960	...	-0.167716	-0.270710	-0.154104	-0.780055	0.7501
7	7.0	-0.644269	1.417964	1.074380	-0.492199	0.948934	0.428118	1.120631	-3.807864	0.615375	...	1.943465	-1.015455	0.057504	-0.649709	-0.4152
8	7.0	-0.894286	0.286157	-0.113192	-0.271526	2.669599	3.721818	0.370145	0.851084	-0.392048	...	-0.073425	-0.268092	-0.204233	1.011592	0.3732
9	9.0	-0.338262	1.119593	1.044367	-0.222187	0.499361	-0.246761	0.651583	0.069539	-0.736727	...	-0.246914	-0.633753	-0.120794	-0.385050	-0.0697

10 rows × 31 columns

Model Building techniques:

- Decision Tree Model
- K-Nearest Neighbors Model
- Logistic Regression Model
- Identify applicable funding agency here. If none, delete this text box.
- Support Vector Machines Model
- Random Forest Model
- XG Boost Model

We will test different machine learning models one by one and checkout their performance. Defined models are easier, a single line of code can define our model and in the same way a single line of code can fit the model in our data.

IMPLEMENTATION

We are at the stage of performing the different models. But before doing the test we need to filter the data and split them into two sets (one for train model and one for test model). If we look at the dataset, the data is imbalance toward a feature. Its main reason is because the banks have already adopted different kind of security measures – so that it would be harder for scammers/hackers to make such moves. However, no system is perfect and there's will always be a chance of any vulnerabilities in the system that can increase such activities.

In our dataset, we found that only 0.17% of transaction are fraudulent. To filter out more, we can remove the null values from the dataset. Also, we can remove the duplicate transactions.

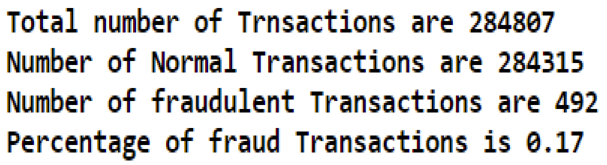
We found out the as per count, we have no null values. To more optimize the results, one can also apply some mechanisms like feature selection which could give better result, however, we are not using this technique in this project. The data which we are using consists 28 features which are transformed versions of PCA but the amount is the real one. And, upon checking their difference we can see a huge difference that can deviate our result.

In this case, its always a good practice to scale this variable for which we can use a standard scalar. It will fix the values and reshape the variable. There's also a external deciding factor which can deviate our result and this would be time – but in our modelling process, we can skip it.

After all this, we have our values and that is (275663,0) after removal of duplicate transactions which was before (284807, 30). Now that we scaled our data without duplicate, it is time to split our data and move forward to model building.

A. Test & Split

Now that we are the stage of test & split, we should first declare dependent and independent variable. The dependent will be known as X and the independent variable is known as y.



```
Total number of Trnsactions are 284807
Number of Normal Transactions are 284315
Number of fraudulent Transactions are 492
Percentage of fraud Transactions is 0.17
```

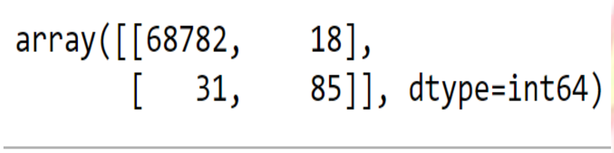
After split our data into train and test data, now, Train data will be used for training the model and the data which is concealed will be used for testing.

B. Model Building

Now, we will be trying different models and check their accuracy. We can also tune these models by altering their parameters. But, if the accuracy is better even with less parameter, then, there's no need to make it complex.

- Decision Tree

Decision Tree usually starts with a single node and then decompose the into additional node to show more possibilities (like next move in chess).



```
array([[68782,    18],
       [   31,    85]], dtype=int64)
```

After implementing the decisions tree model, the accuracy that we got is 0.99928898. Talking about the F1 score, the F1 score is 0.776255707. We could also check its confusion matrix,

Here, the first row represents positive and the second row represents negative. So, we have 68782 as true positive and 18 are false positive. It means, out of 68782 + 18 = 168800, we have 68782 that are classified as a normal transaction and 18 were falsely classifies as normal – but they were fraudulent. But there's more models to try on, let's see their accuracy.

- K-Nearest Neighbors

The K-Nearest Neighbors is generally classified data by determining their neighbor group and their states.

After implementing the K-Nearest neighbors' model, we have got the result, and the accuracy is 0.999506645771664.

The F1 score of this model is 0.8365184615384616.

- Random Forest

Random Forest model is made up of a large number of small decision trees, which each produce their own predictions.

The accuracy that we got after implementing is 0.9993615415868594. The F1 score of this model is less though than above others which is 0.7843137254901961. This model combines the output of different decision tress and tries to reach a single result.

- Support Vector Machines

After implementing the support vector machines model, the accuracy we got is 0.9993615415868594 and the F1 score is 0.7777777777779.

- Logistic Regression

Logistic Regression can be useful in these scenarios. This model predicts a dependent variable by analyzing the relationship between one or more existing independent variables.

- XGBoost

XGBoost is a popular model that implements machine learning algorithms under the gradient boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

After implementing this model in our dataset, the accuracy we got is 0.9995211561901445 and the F1 score of this model is 0.8421052631578947.

RESULT

After analyzing different machine learning models in our dataset, we found out that the most accurate result that we have gotten is from XGBoost model. We have got 99.95% accuracy in our detection. This should not be shocking as our data was balanced towards one variable (non-fraudulent). One more good thing that we noticed is that after applying confusion matrix – our model is not overfitted.

XGBoost is the clear winner in our case. Although the data on we have received the result is for model training. The data feature is the transformed version of PCA. If the actual features follow similar pattern the we are doing great!

FUTURE SCOPE AND CONCLUSION

Credit Card Fraud is obviously one of the most common crimes that happens every day and costs billions to financial banking sector. This paper has pointed one of the techniques which are right now used by the companies to detect and prevent these transactions. This paper comprises of few of the modern machine learning techniques which could be helpful in detection of fraudulent transaction by analyzing their past and present behavior.

We implemented different models and XGBoost has been declared as one of the most accurate model to detect these transaction. The accuracy reached over 99.5%, which is better as the dataset we used contains only 0.17% of fraudulent transactions. This high accuracy is expected due to the high imbalance between the positive and negative variables.

REFERENCES

[1] John Richard D. Kho, Larry A., "Credit Card Fraud Detection Based on Transaction Behavior, "published by Proc. of the 2017 IEEE Region 10 Conference(TENCON), Malaysia, November, 5-8, 2017

[2] Wen-Fang YU and Na Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum", published by International Joint Conference on Artificial Intelligence, 2009

[3] Anshul Singh, Devesh Narayan "A Survey on Hidden Markov Model for Credit Card Fraud Detection". International Journal of Engineering and Advanced Technology (IJEAT), (2012). Volume-1, Issue-3; (49- 52).

[4] Cortes, C. &Vapnik, V "Support vector networks, Machine Learning". . (1995). Vol. 20; (273–297).

[5] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature". Elsevier-Decision Support Systems(2011). 50; (559–569).

[6] Dipti D. Patil, V.M. Wadhai, J.A. Gokhale "Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy". International Journal of Computer Applications, (2010). Volume 11– No.2; (23- 30).