Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Credit Card Fraud Detection using Machine Learning Algorithms

Varun Kumar K S, Vijaya Kumar V G, Vijay Shankar A, Pratibha K
Department of Electronics and Communication
RV College of Engineering, Bangalore.

***Abstract*- Now a days Using technology like phishing technique to do internet banking fraud means transferring and removing the money from banker account without the permission of the banker. credit card frauds are happening in large amount and also some banking companies and the companies giving service to banks are facing problem. In this project trying build the model which predict fraud and nonfraud transaction in best way using machine learning algorithms and neural networks. Objective of the project is to predict the fraud and fraud less transaction with respect to the time and amount of the transaction using classification machine learning algorithms and statistics and calculus (differentiation ,chain rule etc) and linear algebra in building of the complex machine learning models for prediction and understanding of the data set . we have achieved accuracy of 94.84% using logistic regression and 91.62% using naive Bayes and 92.88% using decision tree and we step into deep learning, we used ANN achieved better accuracy then all other algorithms of 98.69%.**

I. INTRODUCTION

Credit card is a small thin plastic or fiber card that contains information about the person such as picture or signature and person named on it to charge purchases and service to his linked account charges for which will be debited regularly. Now a day's card information is read by ATM's, swiping machines, store readers, bank and online transaction. Each card as a unique card number which is very important, its security is mainly relies on physical security of the card and also privacy of the credit card number.

There is rapid increase in the credit card transaction which as led to substantial growth in fraudulent cases. Many data mining and statistical methods are used to detect fraud. Many fraud detection techniques are implemented using artificial intelligence, pattern matching. Detection of fraud using efficient and secure methods are very important.

Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Now days Internet or online transaction growing as new technology are coming day by day. In these transaction Credit card holds the maximum share. In 2018 Credit card fraud losses in London estimated US dollar 844.8 million. To reduce these losses prevention or detection of fraud must be done. There are different types of frauds occurring as technology is growing rapidly. So there are many machine algorithms are used to detect fraud now days hybrid algorithms, artificial neural network is used as it gives better performance.

II. PROBLEM STATEMENT

Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Every year due to fraud Billions of amounts lost. To analyze the fraud there is lack of research. Many machine learning algorithms are implemented to detect real world credit card fraud. ANN and hybrid algorithms are applied.

III. OBJECTIVES

The objectives of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction.

IV. RELATED WORKS

In previous studies, many methods have been implemented to detect fraud using supervised, unsupervised algorithms and hybrid ones. Fraud types and patterns are evolving day by day. It is important to have clear understanding of technologies behind fraud detection. Here discuss machine learning models, algorithms and fraud detection models used in earlier studies.
In [1] data mining techniques are discussed and these methods take time dealing with huge data. Overlapping is another problem with credit card transaction data preparation. Imbalanced data distribution is overcome using sampling methods.

In [2] discuss about skewed data that is Fraud transaction are quite a less compared to normal transaction. When normal transaction looks like fraudulent or fraud transaction appear as legitimate. Also discuss about difficulties in dealing categorial data. Many machine learning algorithms will not support categorial data. Discuss about the detection cost and adaptability as a challenge. Prevention cost and cost of fraudulent behavior are taken into consideration.

In [3] discuss about class imbalance and how to handle it and also discuss how to work on large dataset. The implemented work was overcome these challenges.

In [4] many models are implemented for fraud detection. In every model different algorithm are used. Detection of credit card fraud for new frauds will be problematic if new data has drastic changes in fraud patterns. Replacing the model is risky as machine learning algorithm take much time for training rather than predicting.

In [5] Logistic Regression algorithm (LR) is implemented to sort the classification problem. Using Gaussian Mixture Models fraudulent cases are discretized. To balance data synthetic minority oversampling is used. Sensitivity analysis is used calculate economic value.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

In [6] this Risk Based Ensemble model is used this model can give good results for data with issues and to remove implicit noise in transaction Naıve Bayes algorithm is used.

In [7] they discuss about huge real time date is a main issue. Real life data contains privacy and sensitive so it is difficult to analyze and implement algorithms.

In [8] They were evaluated using both benchmark and real-world data. A summary of the strengths and limitations of the methods were evaluated. The Matthews Correlation Coefficient metric (MCC) has been taken as the performance measure. To evaluate the robustness of the algorithms noise was added to the data. Also, they have proved that the majority voting method was not affected by the added noise.

In [9] this paper K nearest Neighbor algorithm shows good results with respect to performance parameters specificity, sensitivity but accuracy of results are not good in KNN algorithm.

In [10] discuss mainly supervised techniques. They made comparison study of all the algorithms and they showed algorithms behave differently for different situation of problems.

V. TRANSACTION DATABASE

Dataset contains the transaction from Europe card owners during September 2013. In this 492 out of 2,84,807 are fraud transactions. Data is not balanced because less amount of fraud cases as compared to huge transaction data. Dataset is converted PCA transformation and contains only numeric values. Due to privacy and confidentiality many background information is not provided only PCA transformed data is given. Only time and amount are not transformed to PCA all other given values v1, v2, v3 . . . .. v28 are PCA transformed numeric values. Feature class contains 1 for fraud and 0 for normal transaction.

Fig. 5.1 credit card transaction dataset

VI. DESIGN AND IMPLEMENTATION OF ALGORITHMS

The procedure which we followed to predict the result are understanding problem statement and data by performing statistical analysis and visualization then checking whether the data is balance or not, In this data set the data is imbalanced, balanced by using oversampling, then scaling the data using standardization and normalization and testing data with different ML algorithms For any data science project some package are very important such as Numpy that is numeric python And pandas and for visualization of the data, matplotlib and seaborn is used which build on matplotlib with some extra features.

Anaconda navigator is used as it is having several IDE's installed in it python programming language is used to implement machine learning algorithms as it is easy to learn and implement. In this project Jupyter notebook is used to process the complete code where the code can be viewed as block of codes and running each section and identifying the errors is easier. User interface to train and test the algorithms is implemented using python Tkinter module. Test and train buttons are given to train or test the data.

A. Machine learning algorithms

1)Logistic Regression:
Logistic regression works with sigmoid function because the sigmoid function can be used to classify the output that is dependent feature and it uses the probability for classification of the dependent feature.

This algorithm works well with less amount of data set because of the use of sigmoid function if value the of sigmoid function is greater than 0.5 the output will 1 if the output the sigmoid function is less than 0.5 then the output is considered as the 0. But this sigmoid function is not suitable for deep learning because the if deep learning when we back tracking from the output to input we have to update the weights to minimize the error in weight update. we have to do differentiation of sigmoid activation function in middle layer neuron then results in the value of 0.25 this will affect the accuracy of the module in deep learning.

2) Decision Tree:
Decision tree can be used for the classification and regression problems working for both is same but some formulas will change. Classification problem uses the entropy and information gain for the building of the decision tree model. entropy tell about how the data is random and information gain tells about how much information we can get from this feature.

Regression problem uses the gini and gini index for the building of the decision tree model. In classification problems the root node is selected by using information gain that the root node t id selected by using is having the high information again and low entropy. In Regression problems the root node is selected by using gini , the feature which is having the less gini is selected as the root here Depth of the tree can be determined

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

by using hyper parameter optimization, this can be achieved by Using grid search cv algorithm.

3) Random Forest:
The random forest randomly selects the features that is independent variables and also randomly selects the rows by row sampling and the number of decision tree can be determined by using hyper parameter optimization. For classification problem statement the output is the maximum occurrence outputs from each decision tree models inside the random forest. This is one the widely used machine learning algorithm in real word scenarios and in deployed models. And in most of the Kaggle computation challenges this algorithm is used to solve the problem statement.

4) Naïve Bayes:
Naïve Bayes is the machine learning algorithm for classification problem, which work on the property of Bayes theorem. It can be implemented by using features in data set independent feature as input and dependent feature as a output, the same thing what is behind the Naïve Bayes theorem is applied here to calculate probability of the dependent feature with respect to independent features.

5)ANN Model:
Artificial neural networks in deeps learning can be used to replace the machine learning algorithms for better prediction, ANN is having different types of layers such as input layer, number middle layers having activation function for the action of neurons and the output layer having some kind of activation function like sigmoid and weight initialization and reinitialization in backward propagation for reducing the error between actual and predicted values.

VII. RESULTS AND DISCUSSION

The figure 7.1 shows the user interface for test and train the data. Train and Test buttons are given to the user where using train the algorithms are trained and then o predict the fraud by clicking predict button it will take to another window where the input is given and output is seen as fraud or nonfraud.

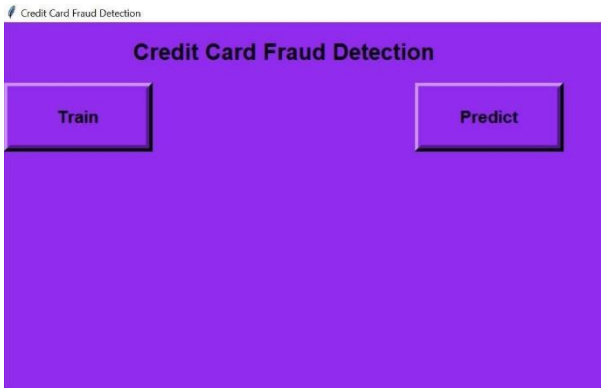

Fig. 7.1: User interface for train and test data

The figure 7.2 shows detection of fraud or nonfraud transaction. when predict button is clicked it will take to another window where it asks for data which is input to the machine learning algorithms and in the predict it will give output as fraud or nonfraud. comma separated 30 values are given including amount and time. Predicted result is displayed as fraud after providing the data. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction

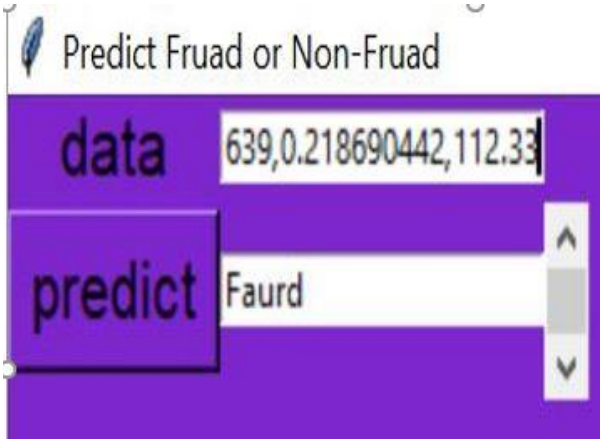

Fig. 7.2: Detection of fraud or normal transaction

1)Confusion matrix for Logistic regression Algorithm:
Fig 7.3 represents confusion matrix for Logistic regression algorithm. Which contains True Positive, True Negative, False Positive, False Negative. False positive value is lesser which shows fraud not detected cases are low. For logistic regression algorithm accuracy, recall, precision achieved are 94.84, 92.00, 97.58 respectively.

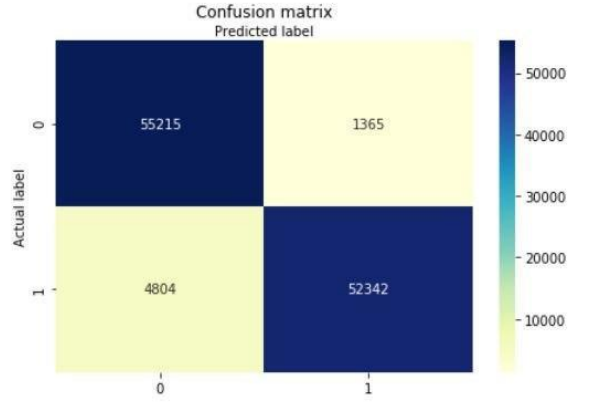

Fig. 7.3: Confusion matrix for Logistic regression

.2) Confusion matrix for Naive Bayes Algorithm:
Fig 7.4 represents confusion matrix for Naive Bayes algorithm. Which contains True Positive, True Negative, False Positive, False Negative. False positive value is lesser which shows fraud not detected cases are low. For Naive Bayes algorithm accuracy, recall, precision achieved are 91.62, 84.82, 97.09 respectively.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

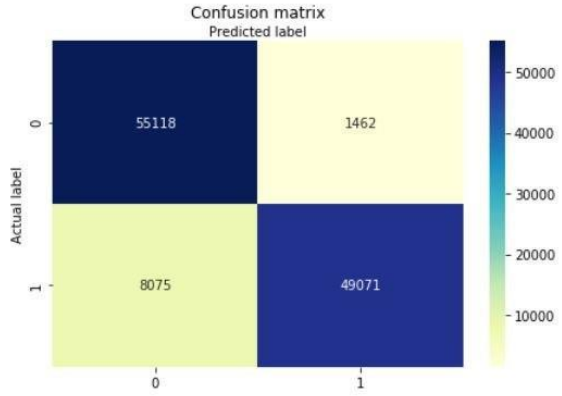

Fig. 7.4: Confusion matrix for Naive Bayes

3)Confusion matrix for Decision Tree Algorithm:
Fig 7.5 represents confusion matrix for Decision Tree algorithm. Which contains True Positive, True Negative, False Positive, False Negative. False positive value is lesser which shows fraud not detected cases are low. For Decision Tree algorithm accuracy, recall, precision achieved are 92.88, 98.98, 99.48 respectively.

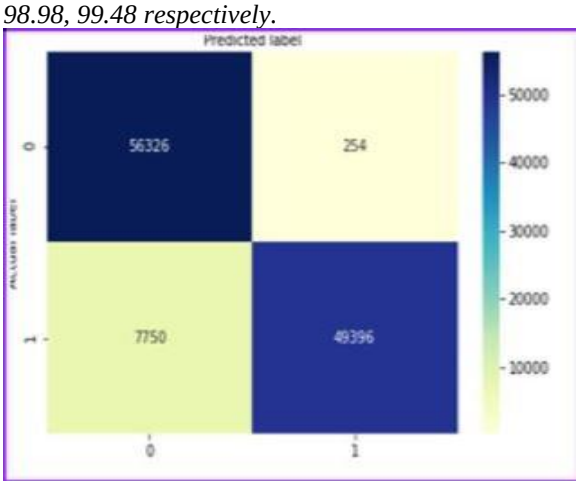

Fig. 7.5: Confusion matrix for Decision Tree

4)Confusion matrix for ANN model:
Fig 7.6 represents confusion matrix for ANN model. Which contains True Positive, True Negative, False Positive, False Negative. False positive value is lesser which shows fraud not detected cases are low. For ANN model algorithm accuracy, recall, precision achieved are 98.69, 98.98, 98.41 respectively.

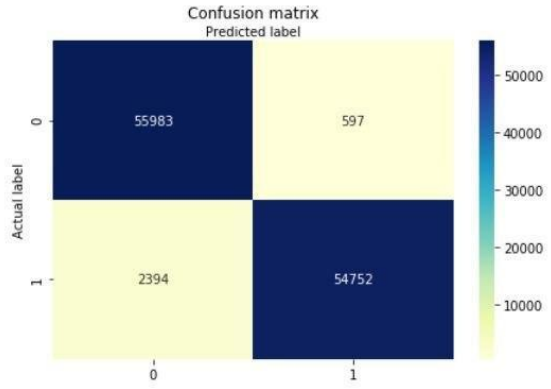

Fig. 7.6: Confusion matrix for ANN

Comparison of algorithms:
Table 7.1 represents the comparison table made using results obtained using simulation. Factors compared are accuracy, precision, recall. From table we can conclude that ANN model as best accuracy, precision and recall.

Achievement of accuracy is done using different algorithms and Ann model gives the best accuracy. confusion matrix gives visualization of results in the form table and minimum false positive rate is seen in all algorithms which is required results to achieve the objective. finally by providing the numerical data fraud or nonfraud detected using basic user interface design.

Table 5.7: Accuracy, precision, recall comparison table for different ML algorithms

	Accuracy	Precision	Recall
Logistic Regression	94.84	97.58	92.00
Naive Bayes	91.62	97.09	84.82
Decision Tree	92.88	99.48	86.34
ANN model	98.69	98.41	98.98

VIII. CONCLUSION

Credit card fraud is most common problem resulting in loss of lot money for peoples and loss for some banks and credit card company. This project want to help the peoples from their wealth loss and also for the banked company and trying to develop the model which more efficiently separate the fraud and fraud less transaction by using the time and amount feature in data set given in the Kegel. first we build the model using some machine learning algorithms such as logistic regression, decision tree, support vector machine, this all are supervised machine learning algorithm in machine learning.

In feature solving this problem statement using another part of artificial intelligence that is time series analysis, in our present project we used both and time and amount feature mainly for predicting the weather the transaction is fraud or Nonfraud transaction, in time series analysis we can reduce the number of parameters that is feature required for the model and we can achieve this model by using average method ,moving average or window method, naive method and sessional naive methods but all this method have some advantages and disadvantages

IX. ACKNOWLEDGEMENT

The authors thank Ms. Pratibha K for providing excellent guidance in carrying out the work. We also thank Dr. Mahesh A and Dr. Prakash Biswagar, professors of RV College of Engineering for providing their valuable feedback.

X. REFERENCES

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, “Random forest for credit card fraud detection”, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.

[2] 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, Issue-1, 2013.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

[3] Rinky D. Patel and Dheeraj Kumar Singh, "Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm", published by International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[4] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, "Improving a credit card fraud detection system using genetic algorithm", published by International conference on Networking and information technology, 2010.

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

[6] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000.

[7] Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis; "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE.

[8] Soltani, N., Akbari, M.K., SargolzaeiJavan, M., "A new user-based model for credit card fraud detection based on artificial immune system," Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on., IEEE, pp. 029-033, 2012.

[9] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network", Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ Knowledge Based Systems, pages 621-630, 1994. IEEE Computer Society Press.

[10] MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, "Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria", International Journal of Computer Applications (0975 – 8887) Volume 52– No.3, 2012.

[11] Fraud Brief – AVS and CVM, Clear Commerce Corporation, 2003, http://www.clearcommerce.com.

[12] All points protection: One sure strategy to control fraud, Fair Isaac, http://www.fairisaac.com, 2007. [13] Clear Commerce fraud prevention guide, Clear Commerce Corporation, 2002, http://www.clearcommerce.com

[13] Samaneh Sorournejad, Zahra Zojaji , Reza Ebrahimi Atani , Amir Hassan Monadjemi, "*A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective* ", IEEE 2016

XI. BIOGRAPHY

1.
Vijayakumar V G,
BE final year,
Dept. of Electronics and Communication,
RV College of Engineering.

2.
Varunkumar K S,
BE final year,
Dept. of Electronics and Communication,
RV College of Engineering.

3.
Vijayshankar A,
BE final year,
Dept. of Electronics and Communication,
RV College of Engineering.

4.
Ms. Pratibha K
Assistant Professor,
Dept. of Electronics and Communication,
RV College of Engineering.