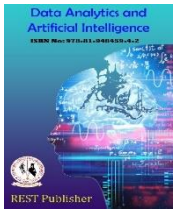
Data Analytics and Artificial Intelligence

Vol: 3(3), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/



Credit Card Fraud Detection Using Machine Learning Algorithms

***Sampriti. S, Rajyasri. S**

Adhiyamaan College of Engineering, Hosur,Tamilnadu,India

*Corresponding Author's mail id: sampritisatish@gmail.com

Abstract: *The Banking industry is the most essential functional unit for a country's economy. When we come across the use of credit card, it came into play in the banking sector in the year 1958 in California by BOA. Credit card is the timeliest mode of payment in terms of pre-set credit limit. When we see with augmentation of technology, cashless payments have touched the peak. As the banking sector has also entered the scenario of digitalization, credit card usage is also increasing. Due to this expansion in technology, credit card fraudulent crimes are proliferating. It is important to take up new measures and stringent actions to limit the crime rates. Fraudulent activities impact the consumers, merchants and the bank organization as well. Thus, it is important to identify the fraud activities by classifying suspicious data from the transactional data. To detect the credit card fraud, various data mining techniques could be used. Some of the data mining techniques discussed in this paper are Support vector machine and Decision tree. Comparison of Support vector Machine and Decision Tree using python and detailed experimental results are proposed in this paper. After analysis, the support vector machine model with kernel as RBF works on an average accuracy of 95%.*

Keywords: *Classifiers, Support Vector Machine, Decision Tree, Gini Index, Gaussian function, Kernel SVM, Radial Basis function,*

1. INTRODUCTION

Data mining has captivated a large deal of attention in the technology of information and society in recent years. This is due to broad availability of amounts of data. With the hike use of financial technology, the utilization of online transactions has increased by years and the rapid increment of E-commerce industry increased the usage of credit cards for purchases in online which has led to more fraud activities related to it. Credit card frauds are categorized into application fraud, internal and offline fraud, bankruptcy fraud etc. Online systems are providers for new opportunities to the fraudsters. For banks it has become burden for detecting the fraud in the card system. These type of frauds in credit cards cause problems and which leads to losses to financial groups and as well as to individuals. Digital Transaction can be processed over phone or on internet. For accomplishing a transaction basic information is required like card number, expiry date, card verification number etc. Therefore, we need a system which can detect the fraudulent transactions and elevate the notification as soon as the transaction is done it can be notified to the credit card provider and take necessary actions and reduce the risk of the loss. In this analysis the dataset which is performed is from Kaggle. It contains Credit Card Transactions which are made by the customers. By observing the behavior of the transactions, they are divided into two categories fraudulent and non-fraudulent. Anomalies are generated based upon these classes and used machine learning algorithms to detect the fraud transactions. These algorithms are analyzed and compared to check which algorithm is best. Rest of the paper is arranged as follows: Discussion of related work in section 2, Section 3 gives the methodology details used in this work, section 4 involves that experimental results and implementation and finally section 5 gives the conclusions from this work.

2. RELATED WORK

[*In1*], in this work various techniques to detect credit card fraud has been compared, the author used quantitative measurements to compare the techniques like Support Vector Machine (SVM), Artificial Neural Network (ANN), Bayesian network, K-nearest Neighbor (KNN), Fuzzy logic-based system, Decision Trees and Logistic regression.

[*In2*], this paper implemented Machine Learning algorithms such as Local Outlier Factor and Isolation Forest. The implementation has been done using python. On analysis, Local outlier factor gives an accuracy of 97% while Isolation Forest gives 76%.

[*In3*], in this paper the authors have identified the different types of fraud and have discussed about detection measures. The techniques discussed are pair-wise matching, decision trees, clustering techniques, neural networks and genetic algorithms.
[*In4*],in this paper authors have developed credit card fraud detection model using Machine Learning Algorithms such as Logistic regression, Artificial Neural Networks(ANN) and Support vector machines. The results are compared using performance metrics such as accuracy, precision, recall, specificity, F1-score, Matthews's correlation coefficient (MCC), and Receiver operating characteristic curve (ROC). The authors have conducted that SVM has better performance.
[*In5*], in this paper various types of credit card frauds has been discussed. The authors have also discussed the techniques used in credit card fraud such as Logistic Regression, outlier detection, neural networks, decision tree and genetic algorithms.
[*In6*], in this paper the authors have discussed about the Clustering Model and Bayesian network model. In addition the probability density estimation method has been used.
[*In7*], this paper implements Machine Learning algorithms like Logistic regression, decision tree and Random Forest are used for the detection of credit card fraudulent activities. Performance metrics like sensitivity, specificity, accuracy and error rate are used to compare the models. The results shows that random forest is more efficient than other two models.
[*In8*],in this work the authors have discussed about how data mining techniques like Association, Classification, Clustering, Decision tree, Prediction and Neural Networks .
[*In9*], the authors have discussed about how data mining techniques are useful in banking sectors. It covers applications of data mining like Fraud Detection, Marketing, Risk management and Customer Relationship Management and how data mining is important in terms of decision making.
[*In10*], in this paper the authors implemented Naïve Bayesian Classifier for prediction of fraud activities. Performance metrics such as accuracy, recall and precision along with ROC analysis are used to know about the model's performance.
[*In11*], in this paper, data mining techniques such as neural networks. Decision Trees, Genetic Algorithms and Rule Extraction are discussed and it is concluded that Artificial Neural Networks (ANN) has more scope in problem solving since it could outperform other techniques in terms of characteristics like parallel processing, self-adaptive, robustness etc.
[*In12*], the authors have discussed about various outlier detection techniques like supervised, semi-supervised, unsupervised, neural networks and rule based outlier detections. Principal Component Analysis (PCA) has been applied for data reduction to enhance the performance of the outlier detection system.

3. METHODOLOGIES

We analyze the dataset and assort the transactions as fraud or non-fraud. The methods chosen in this work are Decision tree and Support Vector Machine with two different kernel functions- Linear and Radial Basis Function (RBF) or Gaussian Function for our dataset for detecting frauds in credit card transactions. Comparisons are done for these algorithms to decide which algorithms gives better results and which can be used to identify frauds in credit card transactions.
Support vector machine: The first methodology we have used in the work is Support Vector Machine (SVM).The Support Vector Machine (SVM) is one of the algorithms used for classifying linear and non-linear training datasets. The objective of this methods is to first hyperplane. There could be several planes of separation however, this method focuses on searching the optimal hyperplane. Support vectors are the points closest to the hyperplane and is useful in the prediction of classes of new data points. For classifying linear data points, the SVM searches the optimal hyperplane with larger margin and it is called maximum marginal hyperplane. The new data point is substituted in the hyperplane equation and then it is classified based on which side the point falls on the vector space. For classifying non-linear data points, the SVM follows two steps. It first transforms the input data to higher dimension data using non-linear mapping and then searches for optimal linear hyperplane and is computed using the same steps as in linear SVM formulation. To transform data with low dimensionality to a higher dimensional space, Kernel function is useful. Some of them are

1. Polynomial kernel of degree h: $K(Yi, Yj) = (Yi. Yj + 1)h$

2. Gaussian radial basis function kernel: $K(Yi, Yj) = e - \| Yi - \frac{Yj\|2}{2\sigma 2}$

SVM classifies the fraud data by classifying it into the appropriate classes based on above mentioned methods
Decision tree: The Second methodology we have used in this work is Decision tree. Normally decision tree transacts with continuous data. Decision tree look alike in shape of a tree which will have connecting lines to the nodes. Each node will have their respective connections to other nodes or for a single leaf node. Decision tree is one of the supervised learning algorithms which is mostly utilized in classification of problems. There will be two or more than two homogeneous sets in tree. It detaches the obscure problems into tiny units or small units and solves the problem with separating and resolving

in a strategic approach. The topmost node can be said as root node. We can classify the tree by two types where one is categorical variable decision tree which will have categorical target variable and the other is continuous variable decision tree which will have continuous target value for the decision tree. The mechanics of decision tree consists of a root node, splitting, decision node, leaf node, pruning, branch or sub-tree, parent or child node. The construction of decision tree is a hierarchy and it does not need scaling of information. With comparison with other techniques decision tree is said to be the simple method for data mining. We have prominent attribute selection measures.

1. Information gain:

   This measure is based on information theory, which scrutinize the information content. Let's assume N is a node that represents the tuples of partition D. The highest information gain of the attribute is chosen for N node. Then this attribute is curtailed the information to classify the tuples in the partitions. The proposed information is given by

$$Info(D) = -\sum_{i=1}^{m} pi \ \log 2(pi)$$

   Where, Info (D) is the entropy, pi is the probability that an arbitrary tuple in D.

2. Gini Index:

   Gini Index is utilized in CART. The impurity of D which is the data partition can be measured by Gini index which is defined as

$$Gini(D) = 1 - \sum_{i=1}^{m} pi^2$$

4. EXPERIMENTS AND RESULTS

The research study implements machine learning algorithms for credit card fraud detection. The results are compared with each other. The sample dataset has been taken from Kaggle. The dataset we have taken in this work is of CSV format-creditcard.csv. This dataset contains the Credit transactions made by the customers of Europe in the year 2013. The dataset's original features have not been enclosed for the security and confidential reasons. The dataset has been transformed using Principal Component Analysis (PCA). The resulting features are numerical variables that are principal component. They are V1, V2 … V28. Along with these, the features 'Time' and 'Amount' has been included, however these two features have not undergone PCA. The features 'Class' acts as a variable to denote Fraud and Non-Fraud data, it takes the value 0 for non-fraud and 1 for fraud The data preprocessing steps has been carried out for improving the quality of the data for better classification. The dataset is split into training set of 70% of data and test set of 30% data.

The Decision tree classifier has been designed based on the given features. The python code as follows:

```
from sklearn.tree import DecisionTreeClassifier
clf =  DecisionTreeClassifier()
clf  =  clf.fit(X_train,Y_train)
```

Likewise, the SVM algorithm for the classification has been designed for the given features. The linear SVM model and kernel SVM model with Radial basis function has been implemented with the following python code:

For linear SVM model:

```
from sklearn.svm import SVC
svclassifier = SVC(kernel= 'linear')
svclassifier.fit(X_train,Y_train)
```

For Kernel SVM model:

```
from sklearn.svm import SVC
svclassifier  = SVC(kernel= 'rbf')
svclassifier.fit(X_train,Y_train)
```

PERFORMANCE EVALUATION: The performance of the models is compared by using evaluation metrics such as:

- Precision
- Recall
- F1-Score
- Support

Precision: The precision is the ratio between true positive predictions and total positive values and is calculated as follows:

Precision=TP/TP+FP

Recall: The recall is the ratio between true positive predictions and total positive values that are classifier

Recall=TP/TP+FN

F1 Score: The F1-Score is the mean of two performance measures-precision and recall and is calculated as follows:

F1 Score=2 x recall x precision/recall + precision

Support: The Support is the number of occurrences of true response in appropriate target values. Where, The True Positive- TP is the number of transactions that are fraud and classified as fraud by the classifier. The True Negative- TN is the number of transactions that are non-fraud and classified as non-fraud by the classifier. The False Positive- FP is the number of transactions that are non-fraud and classified as fraud by the classifier. The False Negative- FN is the number of transactions that are fraud and classified as non-fraud by the classifier.

TABLE 1

Model	Target Class	Precision	Recall	F1 score	Support
Decision Tree	0	0.86	0.88	0.91	103
	1	0.87	0.93	0.90	86
Linear SVM	0	0.89	0.97	0.93	96
	1	0.96	0.87	0.92	94
Kernel SVM(RBF)	0	0.90	0.99	0.94	92
	1	0.99	0.90	0.94	97

Receiver Operating Characteristic Curve (ROC): In case of imbalanced datasets, precision recall might be biased. Although we have balanced while preprocessing steps, the performance metrics may not be accurate and could be misleading since precision recall values focus more on one class. Thus, in this study we have considered the Receiver Operating Characteristic curve which is useful in evaluating the performance of binary classifiers. The ROC is plotted using two parameters. True Positive Rates (TPR) which is the same as recall and False Positive Rate (FPR).

TPR= TP/TP+FN

FPR=FP/FP+TN

The graph is plotted in a space of the range 0 to 1 with TPR in Y-axis and FPR in X-axis. The entire area under the ROC curve is measured and is called Area under curve (AUC-ROC).

The area under curve for decision tree classifier is:

- Decision Tree: 0.8957

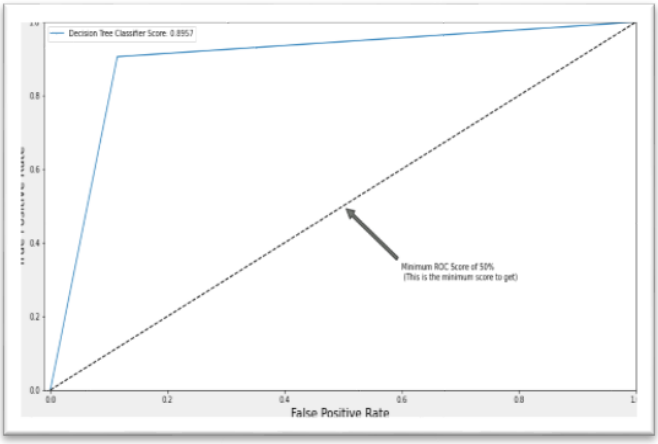


FIGURE 1. ROC curve for Decision Tree

The area under curve for rbf and linear SVM is:

- RBF SVM : 0.98
- Linear SVM : 0.9766

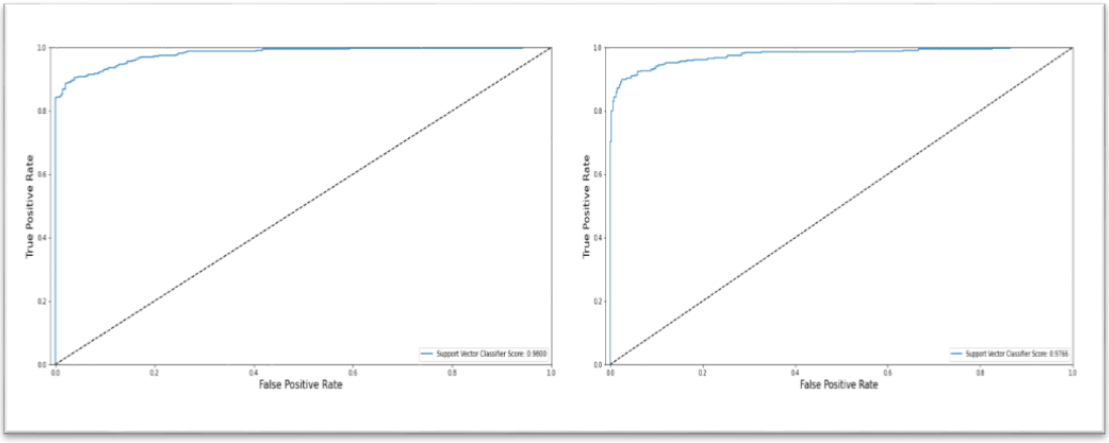


FIGURE 2. ROC curve for RBF SVM

FIGURE 3. ROC curve for linear SVM

TABLE 2. Comparison of various Algorithms

Algorithm	Accuracy
Decision Tree	87.89%
Linear SVM	92.63%
Kernel SVM(RBF)	95.26%

This table shows the comparison results of Decision tree, Linear SVM and Kernel SVM (RBF), among these three algorithms Linear SVM with 92.63% accuracy, Decision tree with 87.89% whereas Kernel SVM (RBF) with 95.26% accuracy. Hence Kernel SVM (RBF) gives the better prediction among these algorithms.

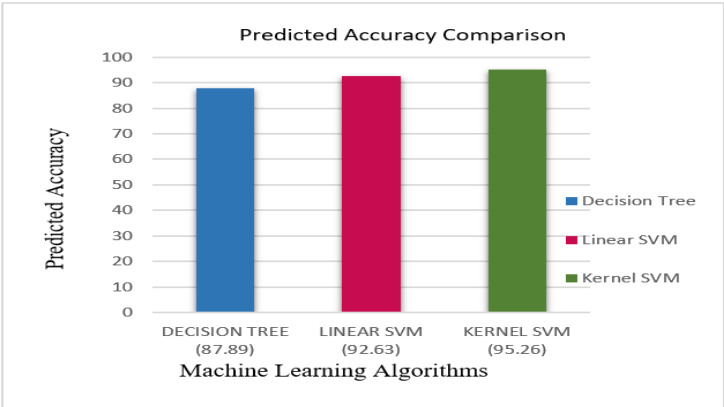


FIGURE 4. Predicted accuracy of these three algorithms

In this Figure we can clearly say that Kernel SVM predicts the better results for the analysis when compared to Decision Tree and Linear SVM. To enhance the better perfromance in figuring out the fraudulent transactions ,this comparison has been made to check which algorithm produces the better results. In this experimental implementation we can say that the Kernel SVM gets 95.26 accuracy in predicting the results

5. CONCLUSION

Credit card fraud is an act of crime and it is important to design a model that could clearly classify the fraud and non-fraud transactions Fraudulent actions have cost over 32 billion worldwide in recent years. This may be projected may increase 5 times than the current fraud activities if it is not solved. In this work, the credit card fraud detection model has been designed using Machine Learning algorithms. However, the machine learning approaches have given the better results. The Support Vector Machine with 'RBF' as kernel yields better results than Linear Support Vector Machine and Decision Tree Classifier. The algorithms have been computed thoroughly using different performance evaluation measures. The accuracy of decision tree, linear support vector and kernel support vector machine (RBF) is 87.89, 92.63, 95.26 respectively. The

performance of the models is evaluated using various metrics. The area under curve ROC for each method have been implemented. The AUC-ROC score for decision tree is 89.57%. The AUC-ROC score for linear SVM is 97.66%. The AUC-ROC score for kernel SVM (RBF) is 0.98. The Area Under curve for ROC of Kernel SVM is 0.4% higher than linear SVM and 8.5% higher than decision tree classifier. Thus, on the basis of various performance measures it can be concluded that support vector machine with kernel outperforms the other two models. Using this algorithm credit card fraud transactions can be figured out and this issue can be solved and the financial losses can be avoided.

6. ACKNOWLEDGEMENT

We are very much thankful to our respected Guide Dr. M. Lilly Florence for her continuous support and encouragement. We are always grateful for her patient guidance and useful critiques of this work.

REFERENCES

[1]. Yashvi Jain, NamrataTiwari, ShripriyaDubey, Sarika Jain- " A comparative Analysis of Various Credit Card Fraud Detection Techniques", International Journal of Recent Technology and Engineering (IJRTE) Volume-7 Issue-5S2, January 2019 ISSN:2277-3878

[2]. Hyder John, Sameena Naaz- "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest" , JCSE, International Journal of Computer Sciences and Engineering, , Vol-7, Issue-4, April 2019 E-ISSN: 2347-2693.

[3]. Delamaire, Linda, Abdou, Hussein and Pointon, John (2009) "Credit card fraud and detection techniques: a review. Banks and Bank Systems", 4(2). pp 57-68. ISSN 1816-7403.

[4]. Shahnawaz Khan, Abdullah Alourani, Bharavi Mishra, Ashraf Al, Mustafa Kamal- "Developing a Credit Card Fraud Detection Model Using Machine Learning Approaches", International Journal of Advanced Computer Science and Applications, Vol. 13, No.3, 2022, doi: 10.14569/IJACSA.2022.0130350.

[5]. Khyati Chaudhary, Jyothi Yadav, Bhawna Mallick- "A review of Fraud Detection Techniques Credit Card", International Journal of Computer Applications (0975-8887) Volume 45-No. 1, May 2012, doi: 10.5120/6748-8991.

[6]. V. Dheepa, Dr.R.Dhanapal- "Analysis of Credit Card Fraud Detection Methods", Internatinal Journal of Recent Trends in Engineering, Vol 2, No.3, November 2009. Available now:

[7]. Lakshmi S V S S, Selvani Deepthi Kavila, - "Machine Learning for Credit Card Fraud Detection System", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24(2018) pp, 16819-16824.

[8]. Abdullahi Sidow Osman- "Data Mining Techniques Review", IJDSR, Volume 2, Issue I June 2019, Al-Madinah International University Malaysia. Available now: https://www.academia.edu/77542877/Data_Mining_Techniques_Review.

[9]. Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan- "A Review on Data mining in banking sector", American Journal of Applied Sciences 10 (10):1160-1165, 2013, ISSN: 1546-9239.

[10]. Rekha Bhowmik- "Data Mining Techniques in Fraud Detection", "Journal of Digital Forensics, Security and Law: Vol: No.2, Article 3, doi:https;//doi.org/10.15394/jdfsl.2008.1040.

[11]. Nikita Jain, Vishal Srivastava- "Data Mining Techniques: A Survey Paper", IJRET: International Journal of Research in Engineering and Technology, eISSN: 2319-1163| pISSN: 2321-7308.

[12]. Ms. Amruta D. Pawar, Prof. Prakash N. Kalavedekar, Ms. Swapnali N. Tambe- "A Survey on Outlier Detection Techniques for Credit Card Fraud Detection": IOSR Journal of Computer Engineering (IOSR-JCE) Volume 16, Issue 2, Ver VI (Mar-Apr.2014), PP 44-48, e-ISSN: 2278-0661, P-ISSN: 2278-8727.

[13]. Machine Learning Group, "Credit Card Fraud Detection", Kaggle, Mar, 2018 [Online], Available: "https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud".

