ISSN No. 0976-5697

VOLUME 12 SPECIAL ISSUE 2, JUNE 2021

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CREDIT CARD FRAUD DETECTION

Kumud Sharma
Student
Computer Science& Engineering
Reva University, Bangalore,India
kumudsharma9816@gmail.com

Manish Kumar
Student
Computer Science & Engineering
Reva University, Bangalore,India
manishpankaj72319@gmail.com

Shalini Tiwari
Assistance Professor
Computer Science& Engineering
Reva University, Bangalore,India
shalini.tiwari@reva.edu.in

Malpani Vijay Rakesh
Student
Computer Science Engineering
Reva University, Bangalore, India
vijaymalpani100@gmail.com

Mahamad Shafeek Yalagi
Student
Computer Science & Engineering
Reva University, Bangalore,India
mdshafeek248@gmail.com

Abstract---In the current economic scenario, credit card use has become extremely important. They enable the user to perform transactions of large sums of money without the requirement to carry cash for payments. They have revolutionized the path of making cashless transactions and have made it easy in making payments convenient for the buyer. This digitized form of payment is extremely beneficial but comes with its own set of shortcomings. With constant increase in number of users, credit card frauds are also increasing at a commensurate pace.Billions of dollars of losshaveresulted every year by illegitimate credit card payments. The development of effective and efficient fraud detection models is key to reducing these losses, and more algorithms depend on advanced machine learning methods to help fraud investigators. As the obtainable credit card fraud data is highly imbalanced. In this paper we are overcoming this deficiency by balancing out the data and bringing out the best algorithm that segregates the transaction efficiently.

Keywords---transaction, credit card,fraud detection

I. INTRODUCTION

Fraud is an intentional deceit carried out for monetary gain, It is an unfair practice. There has been an upsurge in electronic payment methods this has in turn led to an increase in credit card frauds. Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit or debit card. These are used in both offline and online modes to carry out transactions. In recent times, bank customers have grappled with increasing attempts to defraud the customer. To overcome this hindrance many algorithms have and are being developed. Various detection procedures are being worked to resolve this issue most effectively.

Credit card transactions are common nowadays and are most preferred payment methods but they also come with their own set of problems. Many hurdles are faced during fraud detection. With the rapid growth in technology transactions are carried out in seconds thus they provide very short time frame for fraud detection. So, the detection has to be extremely quick and efficient. Numerous simultaneous transactions make it difficult to monitor each transaction individually. Thus, an efficient fraud detection system must be put in place to categorize between a genuine and fraud transaction. Such a system works by learning user-specific and usage behavior. Hence existing approaches of machine learning techniques are applied on the data.

The goal of this paper is to evaluate balanced out dataset

with various machine learning models to determine the best suited model for detecting the credit card frauds. We are using Logistic Regression, KNN, SVC and Decision Tree classifier to evaluate the dataset on the predefined criteria.

II. LITERATURESURVEY

S P Mani raj et.al have Analyzed and Pre-processed data sets

for deployment of outlier detection modules i.e., Local Outlier factor and Isolation forest algorithm [1]. In the next paper by Rotulasail Usha et. al Random forest stands out in comparison to Ad boost with respect to accuracy and precision [2]. Samidha Khatri et.al have concluded that Decision tree performed well based on sensitivity, precision and time[3]. SahilDhankad etc. Al showed that Logistic regression was moreeffective after a comprehensive performance evaluation[4]. Olawale Adepoju et.al depicted that Logistic regression worked more efficiently compared with other algorithms [5].

It is observed that the efficiency of machine learning dataset is hindered due to skewness of available imbalanced dataset. To overcome this obstacle the unbalanced datasets need to be converted to balanced ones. The data pre-processing plays an important role in balancing out the unbalanced dataset and serves the purpose for training the model in a better way. The approach through the use of under-sampling, oversampling and hybrid techniques results in the availability of better dataset for the model.

Requisite for resampling the dataset is needed to train the model to obtain an integrated outcome. Training the machine learning model with a balanced dataset will strengthen the efficiency of the trained model to effectively detect the frauds within the short period of time. To keep up with the recent trends the model has to be regularly trained based on the latest available data to further broaden the scope of detection of frauds in real time. This will be helpful in identifying and resolving the illicit transactions. Thus these techniques may considerably improve the working of certain models as they function on amore balanced dataset.

II. MODELS USE

A. Logistic Regression

It is primarily a statistical model which uses a logistic function to model a binary dependent variable. It is mainly used where there is a possibility of occurrence of a binary classification issue. It performs well on linearly separable classes[4]. The odds ratio is one concept using which we will also define the logit function. It is the probability of an event occurring.

$$Odds\ Ratio = p / (1 - p) \quad (1)$$

where, p = probability of the event

The logit function is that the logarithm of the odd ratio. It takes input within the range of [0,1] and transforms them to values over the real-number range.

The logit function are often defined as follows:

$$Logit\ (P) = \log P/(1-p) \quad (2)$$

In this model, the sigmoid function is also applied effectively.

$$y' = \frac{1}{1-e^{-z}} \quad (3)$$

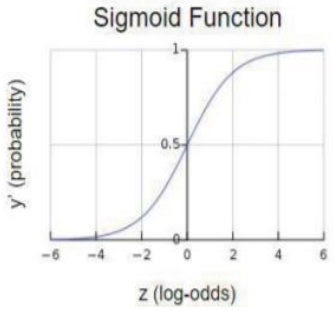

Figure.1 Graph of logistic regression sigmoid function

$$d\left(x^{(t)}, x^{(f)}\right) = \sqrt[P]{\Sigma_k \left|x_k^{(i)} - x_k^{(J)}\right|^{\rho}} \quad (4)$$

$$z = b + m1\ x1 + m2x2 + m3x3 + \ldots\ldots\ldots\ldots + m_n x_n \quad (5)$$

where b is that regression intercept, m is that the weighted values and bias, and x is that the values featured. The chance of an exact outcome is foreseen by the sigmoid function.metric

1.) From k nearest neighbor the space rule is employed to derive.

2.) the full of neighbors accustomed classify the distinct sample.

A suitable price for 'n' ought to be chosen. A relevant distance metric is additionally a demand. Sometimes, the 'Minkowski' distance could also be used. It's a generalization of the Euclidean and Manhattan distance. Mathematically, it's often described as:

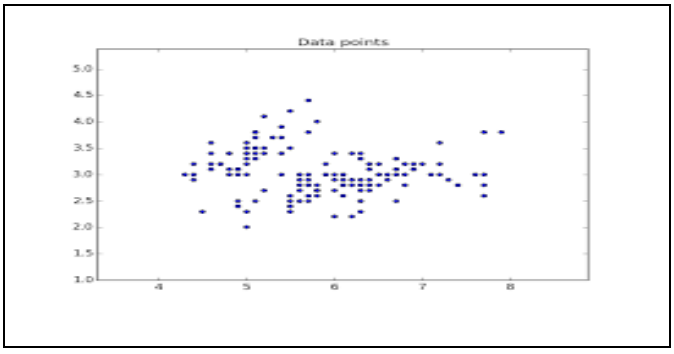

Figure.2 Graph of KNN Classifier

C. Support Vector Classifier

"Backing Vector Machine" (SVM) is an administered AI calculation which can be utilized for both grouping or relapse difficulties. In any case, it is generally utilized in arrangement issues. In the SVM calculation, we plot every information thing as a point in n-dimensional space (where n is number of highlights you have) with the worth of each element being the worth of a specific organize. At that point, we perform characterization by tracking down the hyper-plane that separates the two classes well indeed. Backing Vectors are just the co-ordinates of individual perception.

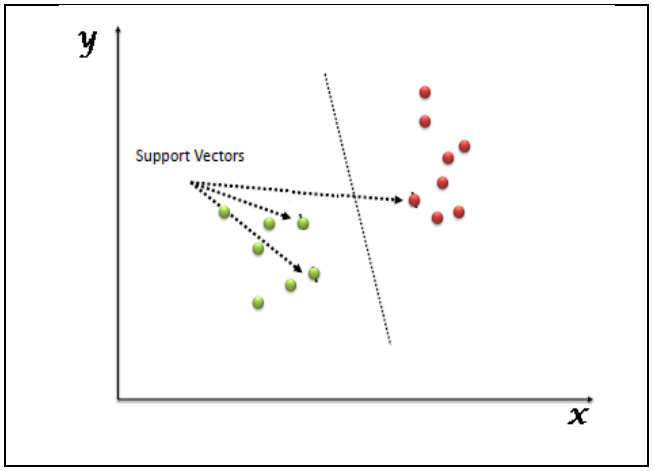

Figure 2. Graph of KNN classifier data points

Backing Vectors are essentially the co-ordinates of individual perception. The SVM classifier is an outskirt which best isolates the two classes (hyper-plane/line).

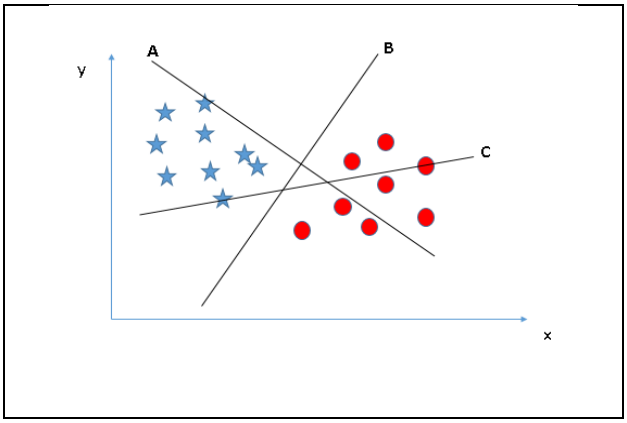

Figure 4: Identify the hyper-plane

At the point when we take a gander at the hyper-plane in unique information space it would seem that a circle:

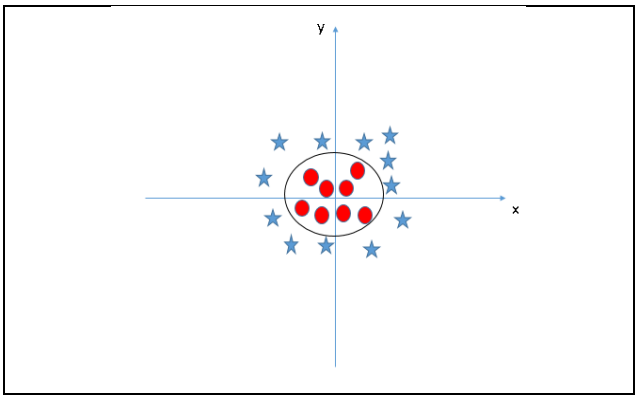

Figure 5: Hyper-plane in original input space

D. Decision Tree

Among all Predictive displaying approaches this one is utilized all the more every now and again/generally. As it is showing up from the name "Choice Tree", construction of the model is readied comparable like tree. It tends to be utilized in Multi-Dimensional investigation when we have various classes. Past Vector or in another words Past information is utilized to make a Model which can be utilized to get the yield esteem dependent on the gave input. Every one of the various hubs of the tree address distinctive Vector. At long last, the end point of the Tree is leaf hub which addresses the conceivable outcome or result.

E. Datasets

The datasets contain exchanges made by Visas in September 2013 by European cardholders. This dataset presents exchanges that happened in two days, where we've 492 cheats out of 284,807 exchanges. The dataset is incredibly unequal, the positive class (fakes) represent0.172% of all exchanges It contains just mathematical information factors which are the aftereffect of a PCA change. Sadly, because of classification issues, we can't give the first highlights and more foundation data about the information. Highlights V1, V2, … V28 are the important segments gotten with PCA, the solitary highlights which have not been changed.

PCA are 'Time' and 'Sum'. Highlight 'Time' contains the seconds slipped by between every exchange and the principal exchange in the dataset. The component 'Sum' is the exchange Amount, this element can be utilized for instance subordinate expense delicate learning. Highlight 'Class' is the

reaction variable and it takes esteem 1 if there should arise an occurrence of misrepresentation and 0 in any case.

VII.OBSERVATIONS

We used the balanced dataset and got the following results:

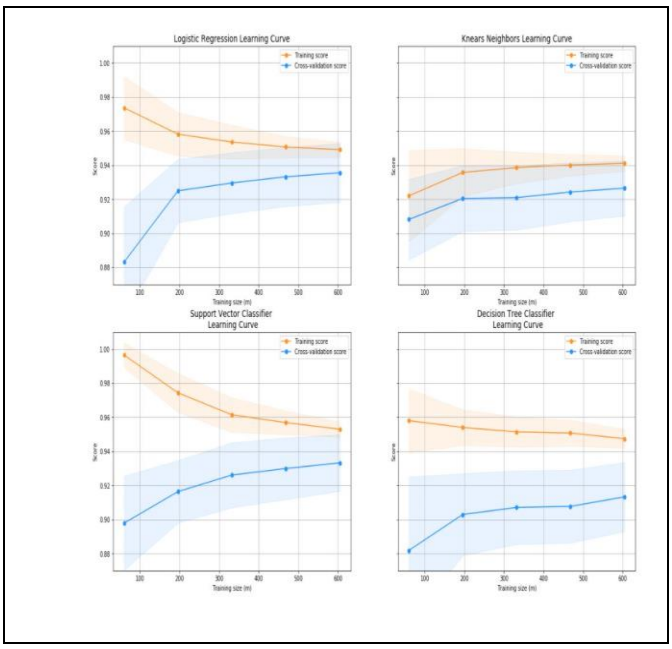

Figure 6: Learning curves

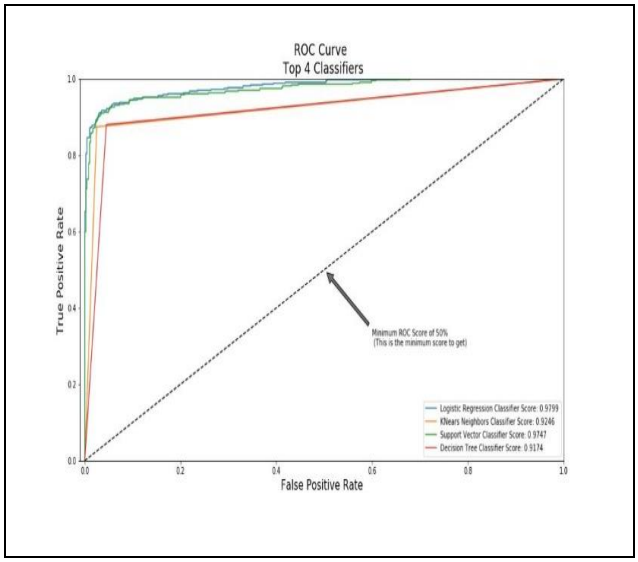

Figure 7: Validation Curves for each model

ROC(Receiver Operating Characteristic) curve shows that the Logistic regression performs well at the classification thresholds.

Logistic Regression

Data	Precision	Recall	F1-score	support
Genuine	0.90	0.99	0.94	91
Fraud	0.99	0.90	0.94	99

KNN

Data	Precision	Recall	F1-score	support
Genuine	0.87	1.00	0.93	91
Fraud	1.00	0.86	0.92	99

SVC

Data	Precision	Recall	F1-score	support
Genuine	0.88	0.99	0.93	91
Fraud	0.99	0.88	0.93	99

Decision Tree

Data	Precision	Recall	F1-score	support
Genuine	0.87	0.99	0.93	91
Fraud	0.99	0.87	0.92	99

The tables bring out the broad view of performance attributes of each model and it is easily understood by the data that Logistic Regression resulted in better detection of the case

TABLE-1ACCURACY OF THE MODEL

MODEL	ACCURACY
Logistic Regression	95.0%
KNN Classifier	93.0%
Support Vector Classifier	92.0%
Decision Tree Classifier	88.0%

We used the balanced dataset through sampling techniques and found that the Logistic Regression algorithmperformed well when compared with other models. Thus percentage of correct predictions were high in case of Logistic Regression.

TABLE 2
CROSS VALIDATION SCORE OF MODELS

MODELS	CVS
Logistic Regression	94.05%
KNN Classifier	92.73%
Support Vector Classifier	93.79%
Decision Tree Classifier	91.41%

.

Here too it is the Logistic Regression models that stands out from the rest of the group. Cross validation technique involves training the model on a subset of data and evaluating on the complementary subset of the data.

VIII. CONCLUSION

Thepaperevaluates Logistic Regression, KNN, SVC and Decision Tree classifier with the balanced dataset obtained through the sampling techniques. Here we used these models for predicting the possibility of occurrence of fraudulent credit card transaction out of the available number of transactions. After analyzing the above mention models we came to the conclusion that the best suited model for predicting transactions involving frauds in the Logistic Regression model. The analysis depicts that the accuracy of Logistic Regression is greater. After the comparative analysis of these models we can infer that we need to train the models with all available latest data in the field to better train the model and derive the desired result from them

As the technology grows with time we need to give way to multiple algorithms to perform the task of identifying or detecting the fraud transactions simultaneously in real time so that the predicted result clearly distinguishes between the genuine and fraud transactions carried out through the credit cards.

IX. FUTURESCOPE

1. In this study we analyzed Logistic Regression, KNN, SVC and Decision Tree classifier, there is a scope for inclusion of more such models.
2. Training and evaluating models based on the most recent available data could further improve the performance.
3. Carrying out the evaluation of performance of every algorithm based on multiple attributes would give a better picture.
4. With the rapid growth in technology multiple algorithms can be simultaneously run in real time to detect the frauds.
5. Building the models with various combinations of sampling techniques and relying on the most recent available dataset further balancing it and training the model would yield better results.

X. REFERENCES

[1] PradheepanRaghavan, Neamat El Gaya: Fraud Detection using Machine Learning and Deep Learning [IJERT 2019] International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) December 11–12, 2019, Amity University Dubai, UAE

[2] RuttalaSailusha, V.Gnaneswar, R. Ramesh, G. RamakoteswaraRao:Credit Card Fraud Detection using Machine Learning [IEEE 2020] Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

[3] SamidhaKhatri, AishwaryaArora, ArunPrakashAgrawa: Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison 978-l-7281-2791-0/20/$31.00 ©2020 IEEE

[4] Ashwinkumar.U.M and Dr.Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012

[5] AlaeChouiekha, EL HassaneIbn EL Haj. “ConvNets for Fraud Detection analysis”. Procedia Computer Science 127, pp.133–138. 2018

[6] OlawaleAdepoju, Julius Wosowei, Shiwanilawte, HemaintJaiman:Comparative Evaluation of Credit Card Fraud DetectionUsing Machine Learning Techniques [IEEE 2019].Global Conference for Advancement in Technology

[7] (GCAT) Bangalore, India. Oct 18-20, 2019

