Data Science: Concern for Credit Card Scam with Artificial Intelligence

Neha Tyagi, Ajay Rana, Shashank Awasthi, and Lalit Kumar Tyagi

1 Introduction

As of now, Unlawful use of a cc and obtaining its private information without the owner's knowledge is considered cc fraud [2]. Credit card fraud is increasing day by day with the development of technology leading to the loss of billions of dollars of customers around the world every year. Hence, Fraud detection involves identifying fraudulent activity scattered across many legal transactions as quickly as possible. This is a widespread event problem called outlier analysis, anomaly detection, exception retrieval, rare class retrieval, unbalanced data retrieval, and so on the amount of fraud transitions is much less than the total number of transitions, so the accurate detection of the fraud transition is very difficult and questionable for this, we must use very efficient methods and algorithms. Therefore, in this article we try to collect and integrate the whole series of research in the literature and to analyze it from various aspects. According to the World Payment Report, in 2016 total non-cash transactions increased 10.1% from 2015 for a total of 482.62 billion transactions. Actually, the credit card fraud techniques are mainly categorized into two category application fraud and the behavioral fraud. Application fraud occurs when scammer request new cards from the bank or issuing companies using the wrong information or obtain the wrong information from the other [3]. However, multiple requests can be made by a single user with the same user details or by a different user with identical details. On the other hand, behavioral fraud has four main types: mail theft, fake card, stolen/lost card, and the cardholder has no fraud [5]. Although Due to the

N. Tyagi (✉) · S. Awasthi · L. K. Tyagi
G.L Bajaj Institute of Technology & Management, Greater Noida, India

A. Rana
Amity University, Noida, India
e-mail: ajay_rana@amity.edu

R. Agrawal et al. (eds.), *Cyber Security in Intelligent Computing and Communications*,
Studies in Computational Intelligence 1007,
https://doi.org/10.1007/978-981-16-8012-0_10

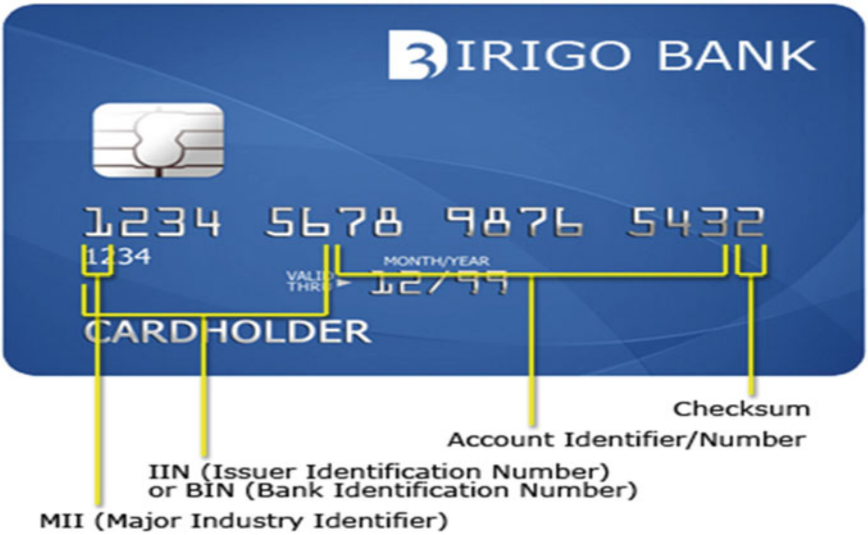

IRIGO BANK
1234 5678 9876 5432
1234
VALID THRU
MONTH/YEAR
12/99
CARDHOLDER
Checksum
Account Identifier/Number
IIN (Issuer Identification Number)
or BIN (Bank Identification Number)
MII (Major Industry Identifier)

Fig. 1 The data of credit card that needs to be confidential by the owner of the credit card in the credit card to protect it from the hacker [6]

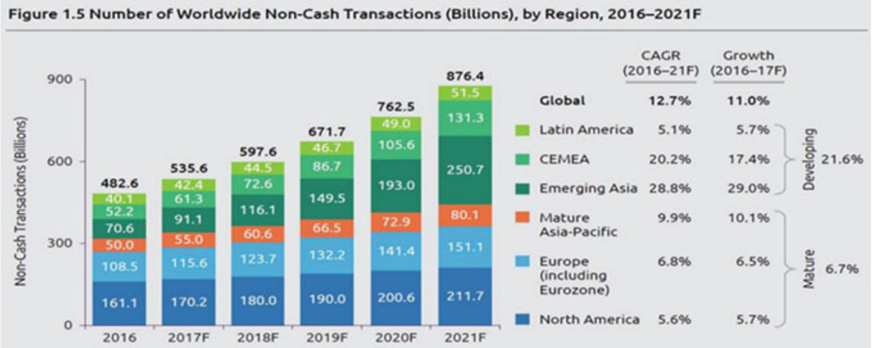


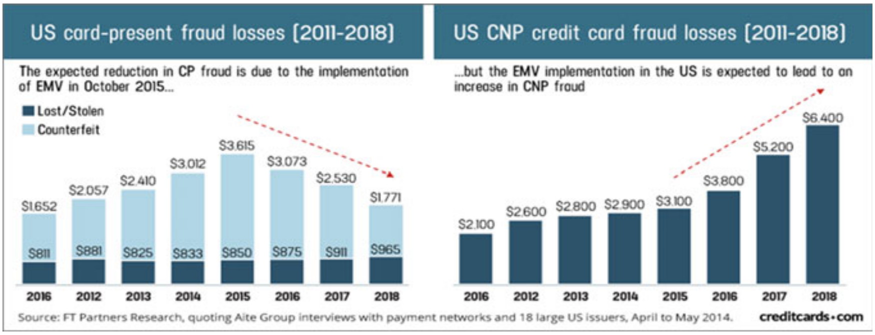
Fig. 2 The steady growth of the non-cash transactions [2]

fraudment of the credit card the flip side fraudment rise we move to the EVM smart chip-based card for security purposes (Figs. 1, 2 and 3).

1.1 *Categories of Credit–Card Fraud Finding*

There are two kinds of main findings:

a. Finding of misuse


US card-present fraud losses [2011-2018]

The expected reduction in CP fraud is due to the implementation of EMV in October 2015...

US CNP credit card fraud losses [2011-2018]

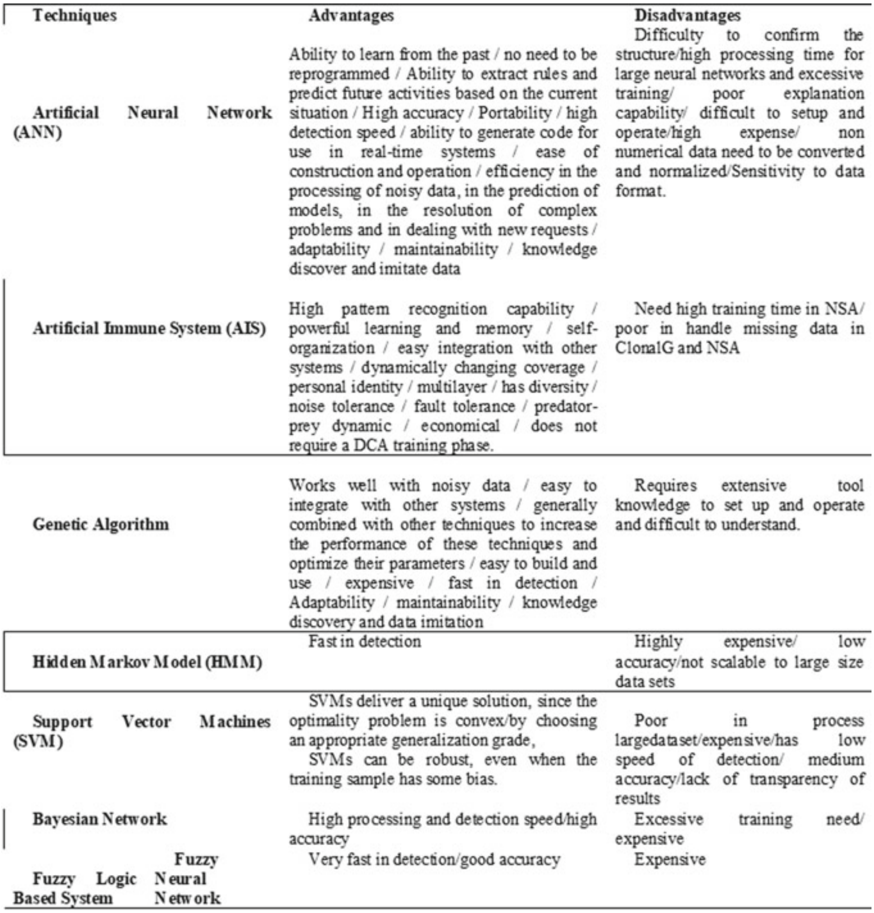
...but the EMV implementation in the US is expected to lead to an increase in CNP fraud

Source: FT Partners Research, quoting Aite Group interviews with payment networks and 18 large US issuers, April to May 2014.

Fig. 3 Due to the fraudment of the credit card the flip side fraudment rise we move to the EVM smart chip-based card for security purposes [5]

b. Finding of Anomaly.

Extortion of misuse identification is a procedure that manages directed grouping movement at the exchange level. In these strategies, exchanges are hailed as extortion or ordinarily dependent on verifiable information from the client's past progress model [4]. This dataset is utilized to make the grouping model that can foresee the state (typical or false) of new records. There are numerous models for making techniques for the common two-class grouping task, for example, rule enlistment, choice trees, and neural systems. This methodology has demonstrated to be the dependable identification of the majority of the misrepresentation recommendations that have been seen previously. It is otherwise called misuse location. Utilization conduct examination misrepresentation (abnormality recognition) manages solo location techniques dependent on account conduct [3]. In this strategy, an exchange is identified as extortion in the event that it clashes with typical client conduct. This is on the grounds that we don't anticipate that con artists should act similarly as the record holder or to know about the proprietor's example of conduct [5]. To do this, we have to remove the real client conduct model, for example, the client profile for each record and afterward recognize false exercises dependent on it by contrasting new practices with agreeing with this model, various exercises that are adequately extraordinary are viewed as misrepresentation. Profiles may contain data about record movement, for example, trader types, sum, spot, and season of exchanges. This strategy is otherwise called irregularity detection [5]. In the table beneath, a few strategies are quickly introduced speaking to some current extortion discovery methods that are applied to the assignments of the charge card misrepresentation recognition framework. It additionally speaks to the favorable position and detriment of each approach [5, 5] (Fig. 4).



Techniques	Advantages	Disadvantages
Artificial Neural Network (ANN)	Ability to learn from the past / no need to be reprogrammed / Ability to extract rules and predict future activities based on the current situation / High accuracy / Portability / high detection speed / ability to generate code for use in real-time systems / ease of construction and operation / efficiency in the processing of noisy data, in the prediction of models, in the resolution of complex problems and in dealing with new requests / adaptability / maintainability / knowledge discover and imitate data	Difficulty to confirm the structure/high processing time for large neural networks and excessive training/ poor explanation capability/ difficult to setup and operate/high expense/ non numerical data need to be converted and normalized/Sensitivity to data format.
Artificial Immune System (AIS)	High pattern recognition capability / powerful learning and memory / self-organization / easy integration with other systems / dynamically changing coverage / personal identity / multilayer / has diversity / noise tolerance / fault tolerance / predator-prey dynamic / economical / does not require a DCA training phase.	Need high training time in NSA/ poor in handle missing data in ClonalG and NSA
Genetic Algorithm	Works well with noisy data / easy to integrate with other systems / generally combined with other techniques to increase the performance of these techniques and optimize their parameters / easy to build and use / expensive / fast in detection / Adaptability / maintainability / knowledge discovery and data imitation	Requires extensive tool knowledge to set up and operate and difficult to understand.
Hidden Markov Model (HMM)	Fast in detection	Highly expensive/ low accuracy/not scalable to large size data sets
Support Vector Machines (SVM)	SVMs deliver a unique solution, since the optimality problem is convex/by choosing an appropriate generalization grade, SVMs can be robust, even when the training sample has some bias.	Poor in process largedataset/expensive/has low speed of detection/ medium accuracy/lack of transparency of results
Bayesian Network	High processing and detection speed/high accuracy	Excessive training need/ expensive
Fuzzy Logic Based System Fuzzy Neural Network	Very fast in detection/good accuracy	Expensive

Fig. 4 Due to the fraudment of the credit card the flip side fraudment rise we move to the EVM smart chip-based card for security purposes

2 Methodology

User comes and selects the transaction methods and after that the process is matched with the stored datasets and is seen that is there any issues in the process of the transaction if the datasets find any issues then the transitions are rejected and the fraud is confirmed [6]. For further if there is any issues like the user isn't fraud then security questions are checked and if the security question can't be answered then it is a complete fraud and the system exits the user from the transaction. The flow chart diagram of the Credit Card Fraud detection technique that is used by us is of the user behavioral model that is shown in the figure below. In this the new transition is matched with the past transition pattern such as amount of transition pattern, location of the transition and the type of purchasing the transition if the

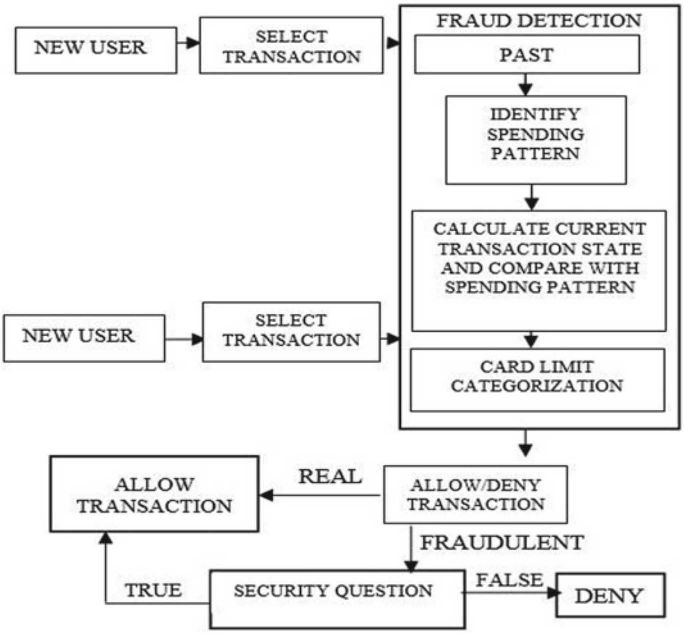


Fig. 5 Block diagram of credit card fraud detection system

system found it fraud than it raises the security question if it answered than transition further proceed otherwise the system will abort the transition and report to the holder about the fraudment of the transition [10] (Fig. 5).

2.1 *Stages of CC Fraud Detection Classification Methods*

The **Credit card Fraud Detection** system is an 8 step process in which we will use different machine learning algorithms and use them to analyze data import dataset plot different graphs use ANN that is artificial neural network to recognize patterns will see different curves.

1. Importing Datasets
2. Data Exploration
3. Data manipulation
4. Data Modeling
5. Fitting Logistic Regression Model
6. Fitting Decision Model
7. Artificial Neural Network
8. Gradient Boosting.

In the very first step import the library and different datasets and then when it is successful then move to our second step and which is data exploration. After that there is a need to explore the imported data that will manipulate it with step called data manipulation and after that Model the data. The Data Modelling is an important part as it will compare with many states and transactions with this modeled data and if there is any issue in modeled data then the system won't work properly [].

Now from the next step start using different ML algorithms, at first do fitting logistic Regression Model. Then apply fitting Decision Model after that apply ANN artificial neural networks and at the end do gradient boosting [].

a. **Importing Datasets**: Here import the datasets that contain exchanges made by Visas or check card holders. We have utilized Kaggle.com which has the premade informational collections which contained 285,500 lines of information and 32 sections. Out of the apparent multitude of segments, the main ones that are the most sense were Time, Amount, and Class (misrepresentation or not extortion). The other 29 sections were changed by utilizing what is by all accounts a PCA dimensionality decrease so as to ensure the client personality [11].

b. **Data Exploration**: In this we will investigate the information that take a gander at various fields' tables and different subtleties like kind of charge card utilized generally the territory which has more number of fakes. We investigate the information that is contained in the MasterCard information outline. We will continue by indicating the charge card utilizing the head () and the tail () work.

   Now the data that use to run through a few initial comparisons between the three columns that Time, Amount, and Class (Fig. 6).

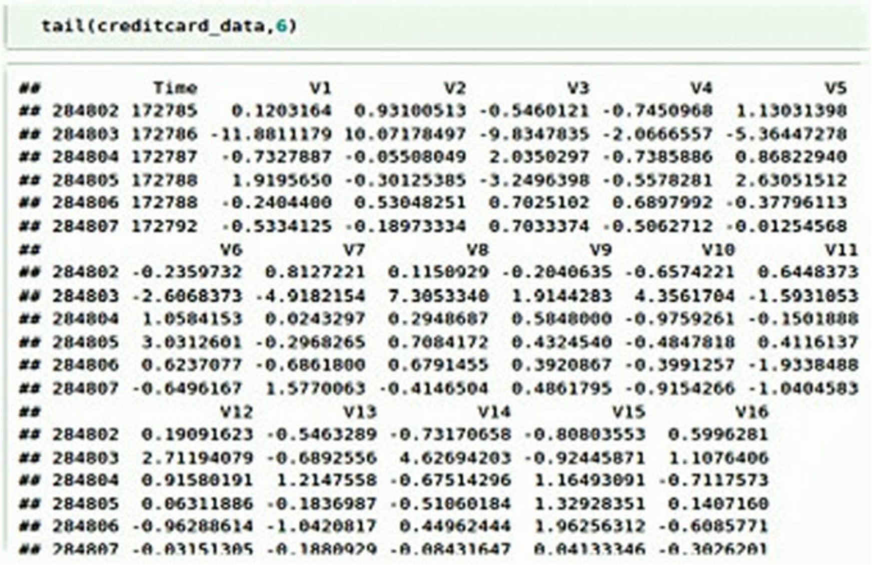
   Now we have the data that we want to run through a few initial comparisons between the three columns that **Time, Amount, and Class**.

   Figure 7 is depicting that, while most of the big deals are very small, this distribution is also planned. Most of the time, day-to-day transactions aren't overly expensive (most are <$50), but they're probably the most fraudulent transactions to happen too.

   Figure 8 is depicting that the visible distribution is valid data for two days. It is for regular consumers for most purchases made during the day and when people leave work/school and go home, purchases decrease overnight. Figure 9 is illustrating the Class (Fraud/Not Fraud).

   In the dataset, there is only 493 fraud transactions, i.e., Only the approx. 0.179% of all of the transactions in this dataset.

c. **Data Manipulation**: The information control is applied to the sum part of Visa informational collection. In this we do Scaling. Scaling is otherwise called the exceptional normalization by the assistance of which the scaling of information is organized as indicated by a particular range. When we've normalized our whole dataset, we'll split our dataset into a preparation set and a test set with a split proportion of 0.83. This implies 83% of our information is given to preparing information while 17% of information is given to test information [12] (Fig. 10).



```
tail(creditcard_data,6)
```

```
##               Time          V1          V2         V3         V4          V5
## 284802 172785   0.1203164  0.93100513 -0.5460121 -0.7450968  1.13031398
## 284803 172786 -11.8811179 10.07178497 -9.8347835 -2.0666557 -5.36447278
## 284804 172787  -0.7327887 -0.05508049  2.0350297 -0.7385886  0.86822940
## 284805 172788   1.9195650 -0.30125385 -3.2496398 -0.5578281  2.63051512
## 284806 172788  -0.2404400  0.53048251  0.7025102  0.6897992 -0.37796113
## 284807 172792  -0.5334125 -0.18973334  0.7033374 -0.5062712 -0.01254568
##                V6         V7         V8         V9        V10        V11
## 284802 -0.2359732  0.8127221  0.1150929 -0.2040635 -0.6574221  0.6448373
## 284803 -2.6068373 -4.9182154  7.3053340  1.9144283  4.3561704 -1.5931053
## 284804  1.0584153  0.0243297  0.2948687  0.5848000 -0.9759261 -0.1501888
## 284805  3.0312601 -0.2968265  0.7084172  0.4324540 -0.4847818  0.4116137
## 284806  0.6237077 -0.6861800  0.6791455  0.3920867 -0.3991257 -1.9338488
## 284807 -0.6496167  1.5770063 -0.4146504  0.4861795 -0.9154266 -1.0404583
##                V12        V13         V14         V15        V16
## 284802  0.19091623 -0.5463289 -0.73170658 -0.80803553  0.5996281
## 284803  2.71194079 -0.6892556  4.62694203 -0.92445871  1.1076406
## 284804  0.91580191  1.2147558 -0.67514296  1.16493091 -0.7117573
## 284805  0.06311886 -0.1836987 -0.51060184  1.32928351  0.1407160
## 284806 -0.96288614 -1.0420817  0.44962444  1.96256312 -0.6085771
## 284807 -0.03151305 -0.1880929 -0.08431647  0.04133346 -0.3026201
```

Fig. 6 Comparison train data of time, amount and class

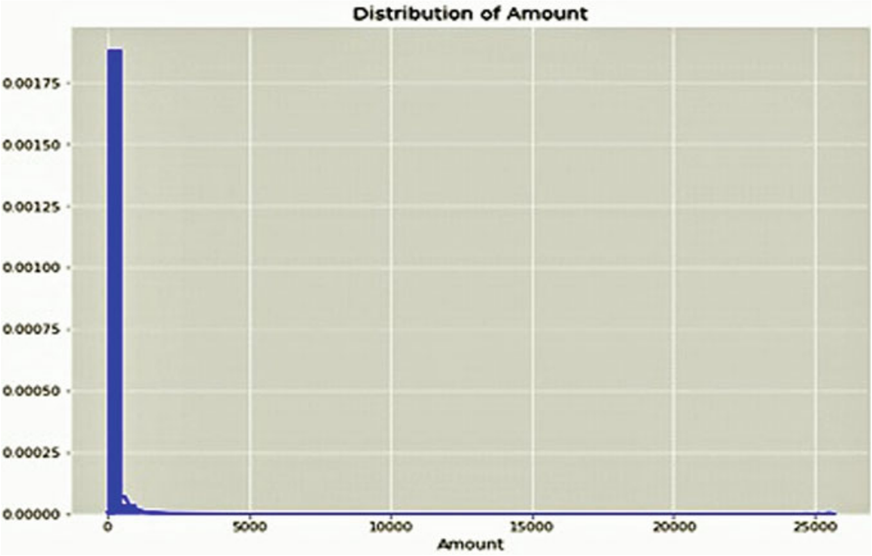


Fig. 7 Distribution of amount of credit card data

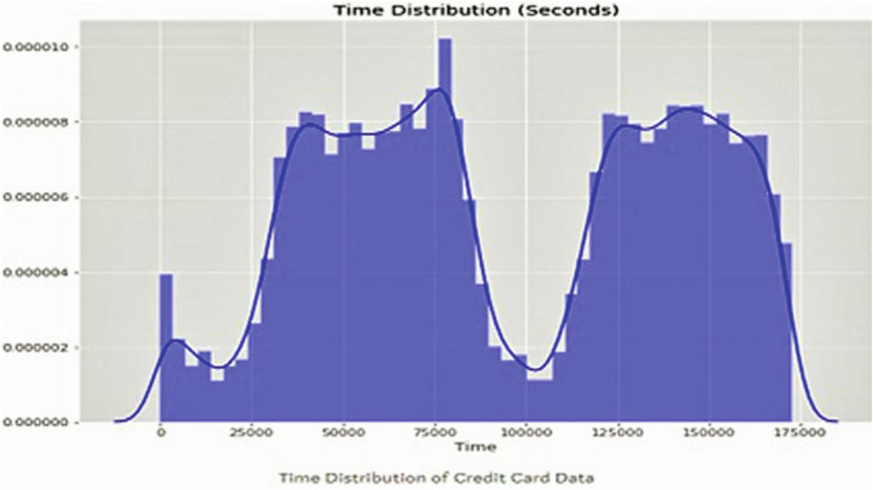


Fig. 8 Time distribution of credit card data

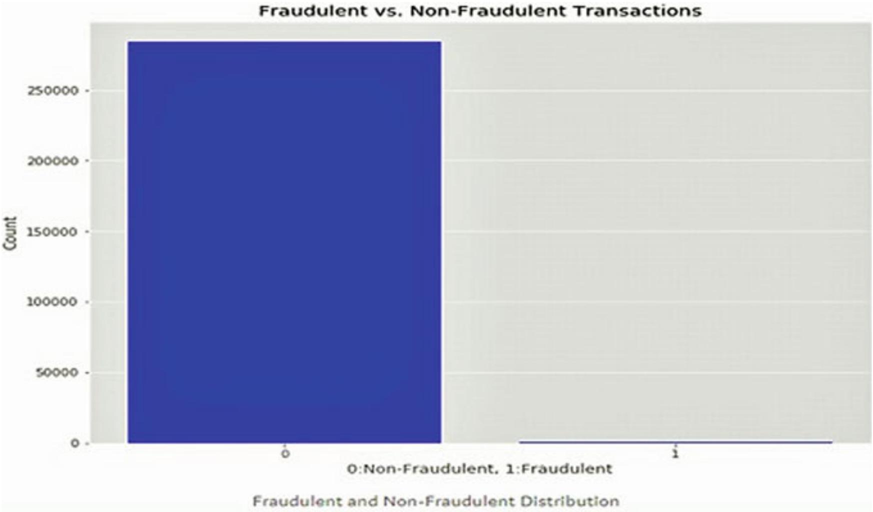
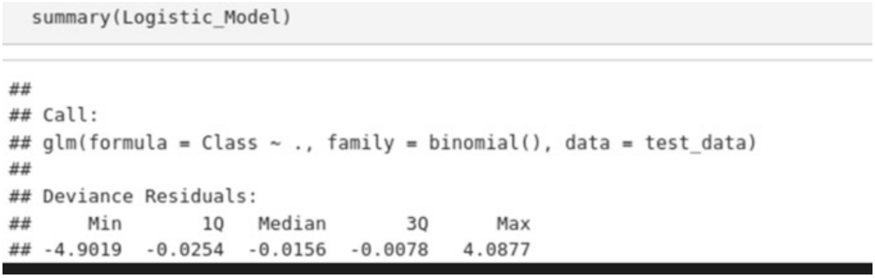


Fig. 9 Fraudulent and Non-Fraudulent distribution of credit card data

d. **Data Modeling**: Later we have normalized our whole dataset, we will partition our dataset into preparing set just as test set with the split proportion of 0.83 which implies that 83% of the information is ascribed to the preparation information though 17% of the information is given to the test information [9, 9].

Figure 11 is depicting that, here not added any numbers because it would be



```
summary(Logistic_Model)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial(), data = test_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -4.9019   -0.0254   -0.0156   -0.0078    4.0877
```

Fig. 10 Divison of test data and train data

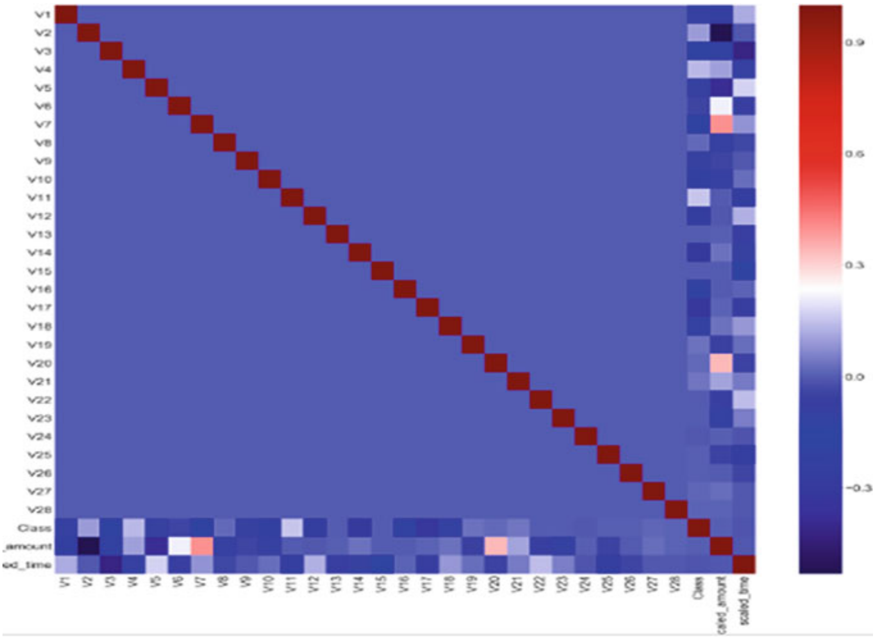


Fig. 11 Correlation among time and amount of credit card fraud

very, very difficult for the reader to see. So look for anything that shows strong correlation.

e. **Logistic Regression and Random Forests**: Calculated relapse is a factual model that endeavors to decrease the expense of an awful gauge. The Random Forests calculation is a gathering of choice trees that all things considered anticipate whether the change is a misrepresentation or not. In the event that progressing is a fake, I run my information through these two models and get extraordinary outcomes. The complete score for calculated relapse and arbitrary woods models is exceptionally encouraging for our dataset. Each model

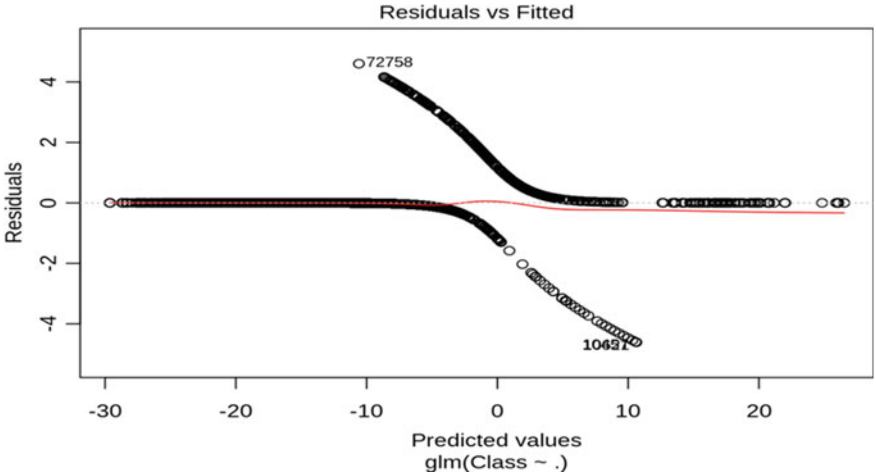


Fig. 12 Logistic regression of credit card fraud/non-fraud transition

has a high evident positive rate and takes into consideration a bogus positive rate which we need here [14].

f. **Fitting Logistic Regression Model**: In this credit card fraud detection system project, we will adapt to our standard model. So let's start with logistic regression. In our case, we are trying to find out if it is a fraud/non-fraud transition. [6, 6] (Fig. 12).

g. **Decision TREE**: In the choice tree—Regression, the choice tree makes relapse or characterization models as a tree structure. We partition the dataset into littler and littler subsets and simultaneously build up a related choice tree. The final product is a tree with choice hubs and leaf hubs. The choice tree can contain both numeric and clear cut information (Fig. 13).

h. **ANN Artificial Neural Network**: ANN's are the sort of AI calculation (ml) methodology displayed based on the human sensory system. ANN models picked up utilizing authentic information and rank dependent on input information. The information imported from the neural system bundle permits us to actualize our ANN. At that point we can plot it utilizing the plot work. Presently, on account of fake neural systems, there is a scope of qualities somewhere in the range of 1 and 0. We characterize limit esteem. 0.5 Otherwise the worth more prominent than 0.5 will be 1 and the rest is 0 [9] (Fig. 14).

3 Proposed Algorithm

Some common methods are used for making the system. Algorithm used: HMM (Hidden Markov Model (HMM)); Semi HMM; Multiple HMM; Optimized HMM, and Advanced HMM.

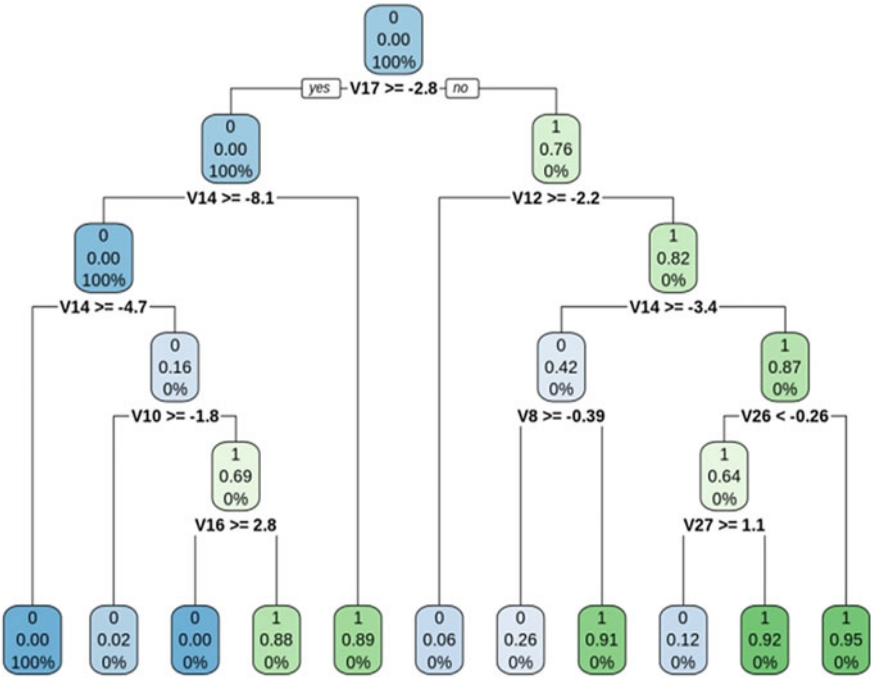


Fig. 13 Decision tree—regression model for credit card fraud/non-fraud transition

The comparison between different algorithms on their accuracy here the optimized HMM is most accurate in the functioning whereas the HMM is lead Functioning. Whereas the dataset trading using R language is also a good accurate algorithm this functionally performs equally to the optimized HMM. Among all the methods of finding the fraud detection technique system build the Hidden Markov Model (HMM) is the one of the best and accurate method for the detection of the fraud transition [7]. HMM method has the very high accuracy among all the methods as shown in the figure below among all the methods of building the fraud detection system techniques (Fig. 15).

3.1 ROC Curve of the Algorithms

A Receiver Operator Characteristics curve (ROC) is a graphical diagram used to show the diagnostic capability of binary classifiers. It was first used in the theory of signal detection but it is now used in many other field such as medicine, radiology, natural hazards, and machine learning [8] (Fig. 16).

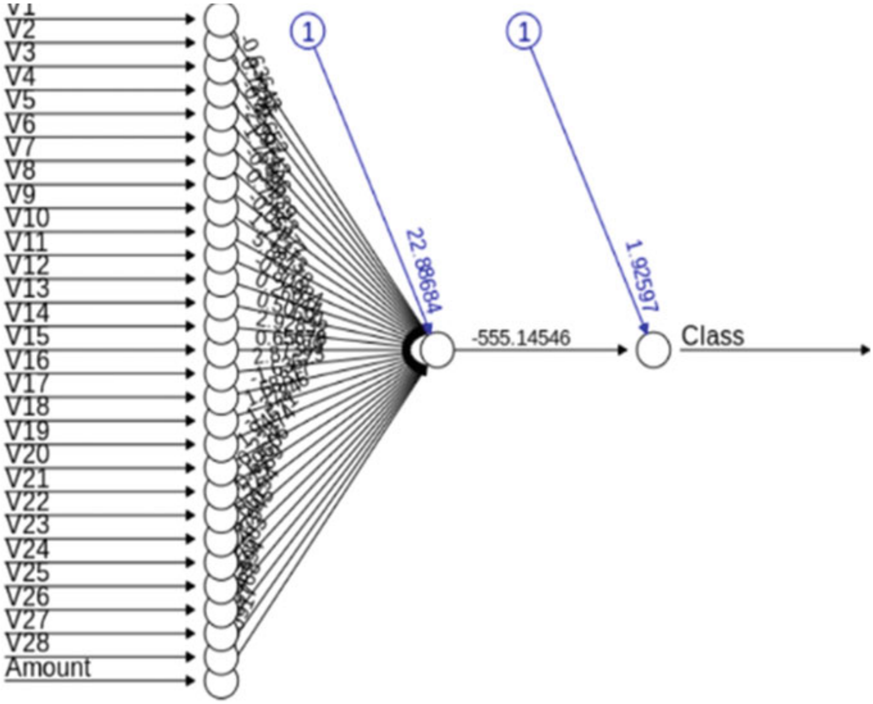


Fig. 14 Artificial neural networks for credit card fraud/non-fraud transition

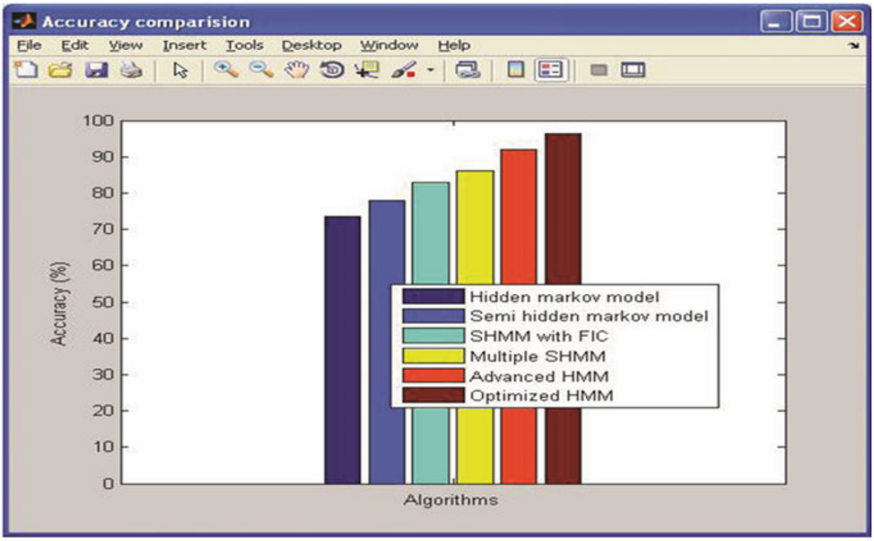


Fig. 15 Accuracy comparison of different methods

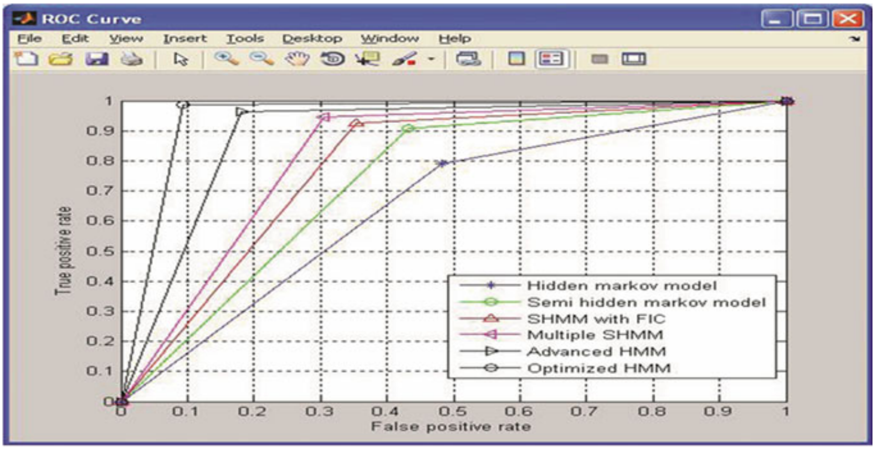


Fig. 16 Receiver operator characteristics curve of various methods

4 Issues: CC Fraud—Detection

Fraud detection system is subject to a lot of trouble and has a lot of problems mentioned below. Effective fraud detection techniques that most should have the ability to eliminate these difficulties to obtain and achieve the best performance.

a. Data Misbalancing: credit card data fraud detection is skewed, which means that a very small percentage of all credit card transactions are fraud. This makes detection difficult and inaccurate fraudulent transactions.
b. Misclassification of data: In a fraud detection system tasks, different misclassifications have different interests. One classifies normal transactions such as fraud harmless as normal detection fraudulent transactions. Because in the first case of misclassification are identified by further investigation [].

5 Conclusion

To find the fraud transactions with credit cards are actually important specifically in digital society. So detection of credit card fraud is very multifarious and important matter that entails the large number of data analysis and a very precise system which will detect fraud in maximum number of cases and also able to handle error and some other inputs. It can be considered as part of machine learning and data science and this type of system ensures the customer and organization that the money transfer is always safe and there is no loss of the money or anything. Future work will focus on more about the fake bin detection which authors covered in the introduction part. This is one of the most important things to look on as there is much fraud using this bin system that authors analyzed. Another thing is the POS hacking or hijacking which is also an important task to add in it. The POS hacking is also a big fraud and we should always look on that to and try our system to tackle this problem also.

References

1. K. Chaudhary, J. Yadav, A review of fraud detection techniques: credit card. Int. J. Comput. Appl. **45** (2012)
2. M. E. Edge, P. R. Falcone Sampaio, A survey of signature based methods for financial fraud detection. J. Comput. Sec. **28**, 381–394 (2009)
3. L. Delamaire, J. Pointon, Credit card fraud and detection techniques: a review. Banks Bank Syst. **4**(2) (2009)
4. S. J. Stolfo, D. W. Fan, W. Lee, A. L. Prodromidis, Credit card fraud detection using meta-learning: issues and initial results. Department of Computer Science Columbia University (1999)
5. S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, Credit card fraud detection using bayesian and neural networks. Vrije University Brussel–Belgium (2002)
6. https://towardsdatascience.com/credit-card-fraud-detection-a122c7e1b75f59
7. Machine Learning Group—ULB, Credit Card Fraud Detection, Kaggle (2019)
8. Visualizing High-Dimensional Data Using t-SNE, J. Mach. Learn. Res. (2013)
9. S. J. Stolfo, W. Fan, W. Lee, A. L. Prodromidis, Cost-based modeling for fraud and intrusion detection: results from the JAM Project. 0-7695-04910-6/99 (1999)
10. N. Tyagi, A. Katiyar, S. Garg, S. Yadav (2017) Methods for protection of key in private key cryptography. Int. J. Innov. Res. (2017)
11. N. Tyagi, A. Agarwal, A. Katiyar, S. Garg, S. Yadav, Information security: a saga of security measures. Int. J. Eng. Comput. (2017)
12. N. Tyagi, Algorithm for protection of key in private key cryptography. Int. J. Eng. Res. Comput. Sci. (2017)
13. N. Tyagi, Security issues, research and challenges in cloud computing. Int. J. Eng. Res. Comput. Sci. (2017)
14. N. Tyagi, Protection of key in private key cryptography. Int. J. Adv. Res. (2017)
15. N. Tyagi, Function codes for protection of key in private key cryptography. J. Emerg. Technol. Innov. (2017)
16. L. K. T. Neha Tyagi, Need of combining symmetric key cryptosystem with public key cryptosystem, in International Conference on Issues & Challenges in Networking, Intelligence (2011)