DOI: 1061294/jiaps2023.2739 ISSN 0974 - 9373

Vol. 27 No. 3 (2023) **J. Int. Acad. Phys. Sci.** pp. 295-303

Performance Analysis of Different Machine Learning Algorithms on Credit Card Fraud Detection

Amanpreet Kaur, Vansh Sachdeva, Abhijot Singh, Ayush Jaiswal, Niyati Aggrawal, Archana Purwar

Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India

Email: amanpreet.kaur1410@gmail.com, vksvs25@gmail.com, abhijotbatra@gmail.com, jaiswalayush131313@gmail.com, niyati.aggrawal@jiit.ac.in, archana.purwar@jiit.ac.in

(Received April 27, 2023, Accepted August 16, 2023)

Abstract: Machine learning (ML) is a logical investigation of various algorithms and factual models that PCs utilize to carry out particular operations that are not clearly programmed. This paper aims to statistically analyze different machine learning algorithms, and compare and contrast their performance for credit card fraud detection. Algorithms used are Artificial Neural Networks(ANN), Sample Vector Machine (SVM), and Kth Nearest Neighbour (KNN), Decision Tree, Logistic Regression and Random Forest. All these above mentioned algorithms are compared on basis of performance measures. It is deduced that the random forest algorithm is the best algorithm.

Keywords: Machine learning, artificial neural networks, sample vector machine, random forest, accuracy, precision, recall, F1 score.

1. Introduction

Machine learning algorithms can be applied for extensive range of applications (data mining, image processing, and predictive analytics, etc). For example, consider a web search engine such as Google that acts as a crawler. It uses learning algorithms that decides ranking of web pages. The machine learning algorithm has an advantage that if it has learned once how to process data, then it can perform its job without human intervention.

In this paper, we will do comparison of different machine learning algorithms such as ANNs, sample vector machine, decision tree, Logistic regression, Kth nearest neighbour, naive bayes, random forest classification on credit card fraud detection. The performance metrics for comparison used are Accuracy, Recall, Precision and F1-Score. Thus, the main contribution of

this paper is comparison of various well recognized machine learning algorithms based on accuracy for data of credit card frauds of duration between 1st Jan 2019 and 31st Dec 2020 which was generated via using Sparkov.

2. Related Work

Zamini et al. proposed an independent charge card Fraud area structure using auto encoders with three mystery layers based gathering. The method has been tested on 284807 transactions from European dataset[1]. Sadgali et al. presented an approach that applies Artificial Intelligence (AI) computation for blackmail recognizing evidence in card trades[2]. Makki et al. research portrays that the Visa coercion causes gigantic money related mishap. By far most of the researchers have been working on this to give an inventive methods of obliterating this mishap and a huge part of the open procedures are excessive, monotonous and work inspiring power task. It has been observed that the imbalanced portrayal of dataset is the essential legitimization behind some unacceptable results after various exploratory assessments. It has been deduced that Linear Regression (LR), Sample Vector Machine (SVM), C5.0 decision tree computation and Artificial Neural Networks (ANN) are best estimations in terms of precision, area under the precision-recall curve (AUCPR) and affectability. The fair dataset have been applied to set up these models[3]. Sohony et al. proposed a gathering learning technique for Credit Card blackmail distinguishing proof as the extent of deception to standard trade is somewhat appropriate. It has been proved[4] that Random forest area is generally most appropriate for recognizing coercion events. Many methods have been attempted with the tremendous authentic Visa trades [4]. Kumar et al. made a review on all methods used to recognize Credit Card coercion area using AI computations and survey the display with the estimations. Enormous heaps of assessments occurred over this space. It has been observed that a more successful structure is required for better performance in every situation[5]. Taha et al. depicted that up gradation in Internet based business and correspondence development have made charge card use an all the more notable strategy for portion. Also the deception related with trades is similarly extending. In this procedure, two plans of genuine open dataset comprising of phony and non-counterfeit trades is used. Taking into account the assessment with various strategies, the proposed system beat similar to precision. The proposed structure conveys the 98.40% precision, 92.88% district under authority working characteristics twist (AUC) and 56.95% F1-score[6]. Prusti et al. arranged an application with

applied decision tree (DT), Extreme learning machine (ELM), k-nearest estimation (kNN), support vector machine (SVM) and Multilayer perceptron (MLP) to recognize the precision in coercion ID. A hybrid model of DT, SVM and kNN is proposed. Results show that[7] the hybrid system had higher precision of 82.58%. Jiang et al. proposed a smart cycle with various stages. Firstly, the trades were assembled, then, considering the individual direct principles trades are added up to, next the dataset is organized, further the model is ready. Expecting any uncommon lead arises, an information is given to the structure about the bizarre direct through an analysis framework[8]. Li et al. proposed a blackmail disclosure system through Kernel based oversaw hashing (KSH). The proposed model relies upon gathered nearest neighbor thought. The model is generally appropriate for a colossal dataset with high angle data. Strangely KSH is used for gauge, which performs better contrasted with other existing structures[9]. Tran et al. proposed two new data driven methodologies for distortion trade in Credit Card trades. The two unique approaches are segment limit decision and T2 control chart[10].

3. Proposed Methodology

Figure 1 shows classification of Machine learning algorithms. The algorithms highlighted in dark colour have been implemented.

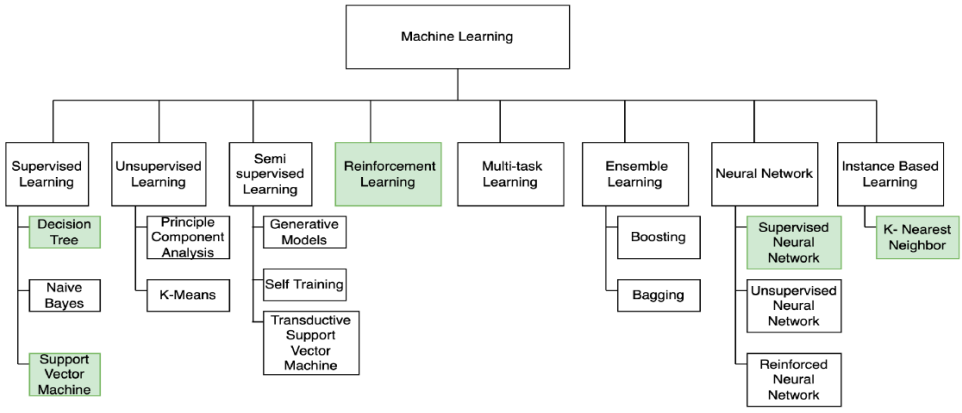

Figure 1: Classification of Machine Learning algorithms

3.1 Artificial Neural Networks(ANN)

ANN enjoy specific benefits that it can learn and show non-immediate and complex associations as various associations among information and yield are non-straight, After planning, ANN can instigate covered associations from subtle data, and consequently it is summarized, Unlike many AI models,

ANN doesn't have limitations on datasets like information ought to be Gaussian circulated or nay other conveyance.

3.2 Sample Vector Machine (SVM)

SVM is a synchronized AI estimation that can be applied for both gathering and backslide challenges. For SVM, all information is plotted as a point in n-dimensional space with the worth of each part being the worth of a specific facilitator. Then, a gathering is performed by obtaining the hyper-plane that isolates the two classes.

The upsides of SVM are it high robustness in view of dependence on help vectors and do not get impacted by outliers. Numeric assumption issues can be overseen by SVM. The disadvantage of SVM is it is a blackbox technique and is inclined to overfitting strategy, extremely thorough calculation.

3.3 Kth Nearest Neighbour (KNN)

KNN can be used for both request and backslide farsighted issues. In any case, it is even more commonly used in gathering issues in the business. KNN works on a standard expecting every datum directly falling in close toward each other is falling in a comparative class. All things considered, it arranges another data point subject to resemblance.

The meaning of K in the KNN estimation works in the way: find a distance between an inquiry and all models (factors) of data, select the particular number of models (say K) nearest to the request, then, pick

- The most successive mark if utilizing for the order based issues, or
- The midpoints the name if utilizing for relapse based issues
  Hence, the calculation tremendously relies on the quantity of k, to such an extent that
- Worth of k – greater the worth of k builds trust in the forecast.
- Choices might be slanted if k has exceptionally enormous worth.

3.4 Decision Tree

Decision Trees (DT) is a machine learning technique utilized for characterization and relapse problems. In decision trees, each node checks for true condition and in case it is, it goes to the child hub having that choice.

3.5 Logistic Regression

It is an order model rather than a relapse model. Strategic relapse is a basic and more proficient technique for parallel and straight grouping issues. It is a characterization model, which is extremely simple to acknowledge and accomplishes awesome execution with directly distinguishable classes. It is a broadly utilized calculation for arrangement in industry. The calculated relapse model is a measurable strategy for twofold characterization that can be summed up to multiclass grouping.

3.6 Random Forest

A Random Forest is used to tackle relapse and order issues. It applies gathering realizing, procedure that constructs various decision trees on different data subsets. Final outcome is obtained by taking the mean of the yield from different trees.

4. Simulation and Results

The whole code can be executed on Windows 10 or above platform. The language used is Python. Our whole code execution is finished utilizing Jupyter Notebook programming. which depends on Anaconda Distribution which is utilized for Programs dependent on data science and related Advanced Python projects It is an open-source IDE extraordinarily intended for the python language.

Some of its features used in our work are :

1) Its editor is used for code completion, editing and highlighting syntax Editing of variables and exploring them using GUI.
2) Its file explorer, variable explorer and help features were of great use. Linkage with various libraries is helpful in writing code easily.
3) In some parts during implementation Codebooks was also used to run the code in C++ to find errors and display the output.
4) Microsoft excel has also been used to maintain the dataset used in the entire project.

Hardware Requirements are 8 GB RAM, Processor with speed 2 GHz Intel Core i7 processor (10th Generation) and System Type should be 64-bit Operating System.

```
trans_date_trans_time     0
cc_num                    0
merchant                  0
category                  0
amt                       0
first                     0
last                      0
gender                    0
street                    0
city                      0
state                     0
zip                       0
lat                       0
long                      0
city_pop                  0
job                       0
dob                       0
trans_num                 0
unix_time                 0
merch_lat                 0
merch_long                0
is_fraud                  0
age                       0
dtype: int64
False
```

Figure 2: Showing The Number of Empty Cell Count Of Each Row Post in data preprocessing

The data set utilized for training and testing of supervised learning and deep learning algorithms have been taken from *www.kaggle.com*.
This dataset contains genuine and fraud credit card transactions from the period from 1st Jan 2019 till 31st Dec 2020. It includes 1000 customers doing transactions via credit card and 800 merchants. In total the data have more than 1.2 million records for training purposes and more than 5 Million Records for testing the Models trained under various algorithms.

The source of simulation was generated using Sparkov Data Generation tool. The files were downloaded and changed according to a standard format. Later, the data was pre-processed in which we checked if there were any empty rows and columns and we found out that data did not have any missing values. So we did not have to interpolate the values in the dataset.

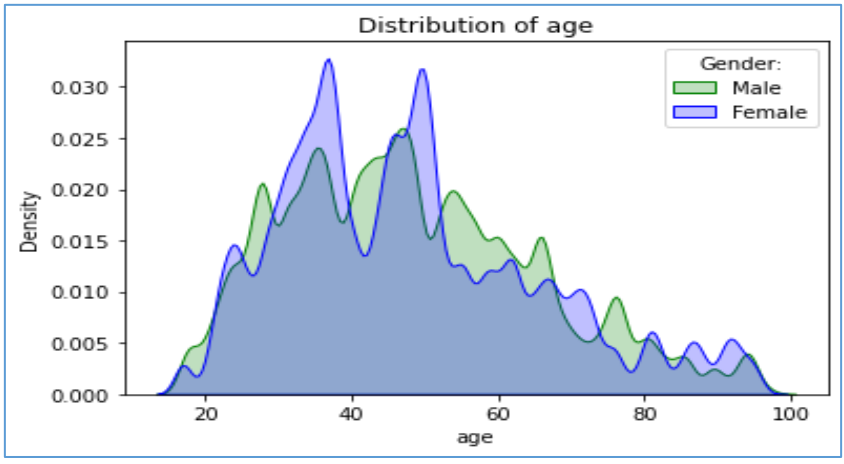

Figure 3: Graphical Bifurcation of Data on the basis of Gender And Age

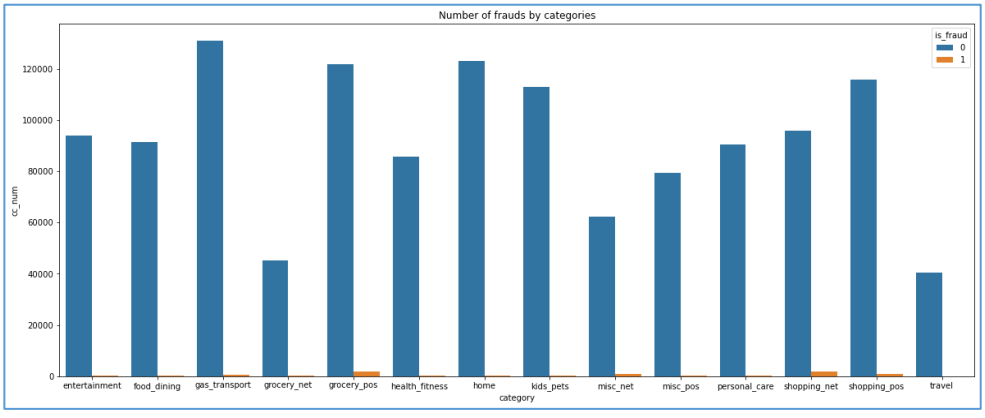

Figure 4: Graphical Representation of where how many authenticate and fraud transactions in various categories

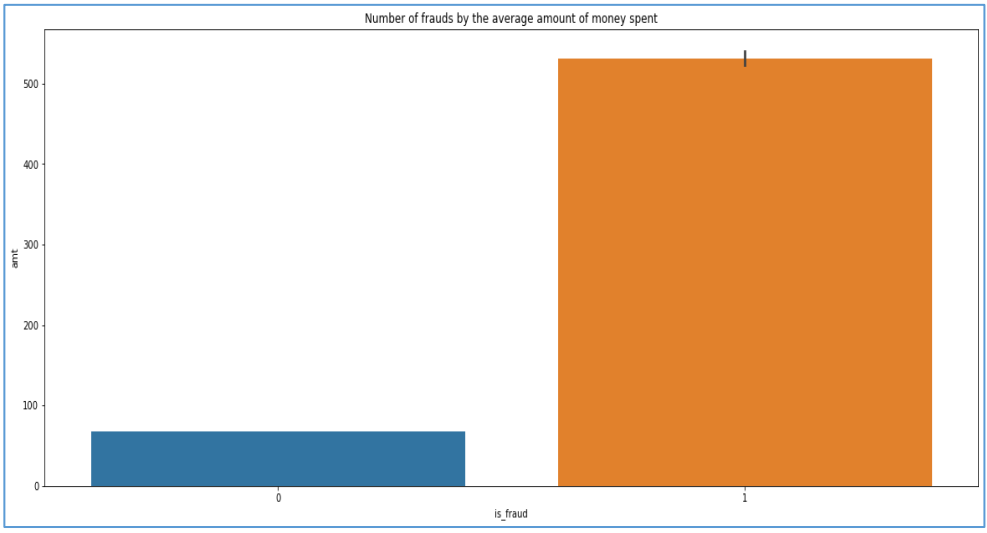

Figure 5:Graphical Bifurcation of data on the basis of being fraud transaction or authenticate to the average amount spent on that transaction

4.1 Result

Table 1 shows the after-effects of the pre-owned calculations on the exhibition measurements like exactness, accuracy and review.

Table-1: Comparison of algorithms

Models	Accuracy	Precision	Recall	F1 Score
Random Forest Tree	0.93	0.96	0.91	0.93
SVM	0.86	0.96	0.75	0.84
Decision Tree	0.89	0.98	0.81	0.88
KNN	0.85	0.86	0.85	0.85
Logistic Regression	0.85	0.95	0.74	0.83
Neural Networks	0.92	0.92	0.92	0.92

In our model, a random forest model gives highest accuracy of approximately equal to 93% for our test models followed by Neural Networks.

5. Conclusion

This paper did performance analysis of different machine learning algorithms on credit card fraud detection data set. Supervised machine learning and deep learning algorithms such as k-Nearest Neighbor, Support vector machine, Decision Trees and Random Forest have been compared. In our model, a random forest model gives highest accuracy of approximately equal to 93% for our test models followed by Neural Networks whose accuracy might increase depending on the number of the hidden layers and

number of iterations between them. The possible future aspects can be use of various other types of data sets and new machine learning or deep learning algorithms or hybrid models. Also, Neural Network algorithm can be optimized by changing the values of the number of hidden layers of the neural network along with the maximum number of iterations.

References

1. M. Zamini and G. Montazer Credit Card Fraud Detection using autoencoder based clustering, *9th International Symposium on Telecommunications (IST)*, (2018), 486-491, doi: 10.1109/ISTEL.2018.8661129.
2. I. Sadgali, N. Sael, and F. Benabbou; Adaptive Model for Credit Card Fraud Detection, *International Journal of Interactive Mobile Technologies (iJIM)*, 14(03) (2020), 54-65. https://doi.org/10.3991/ijim.v14i03.11763.
3. S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid, H. Zeineddine; A trial review with imbalanced characterization approaches for charge card misrepresentation location, *IEEE Access* 7 (2019), 93010-93022. doi:10.1109/ACCESS.2019.2927266.
4. Ishan Sohony, Rameshwar Pratap and Ullas Nambiar; Ensemble learning for credit card fraud detection, *Proceedings of the ACM India joint international conference on data science and management of data*, (2018), 289-294. doi:10.1145/3152494.3156815.
5. P. Kumar, F. Iqbal; Credit card extortion recognizable proof utilizing AI draws near, *Proceedings of the first International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India*, (2019), 1-4. doi:10.1109/ICIICT1.2019.8741490.
6. A.A. Taha, S.J. Malebary; An insightful way to deal with charge card misrepresentation identification utilizing an upgraded light slope helping machine, *IEEE Access*, 8 (2020), 25579-25587.

   doi:10.1109/ACCESS.2020.2971354.
7. Abhishek Kumar, Debachudamani Prusti, Shubham Ingole and Santanu Rath; Real-time SOA based credit card fraud detection system using machine learning techniques, *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, (2021), 1-6. 10.1109/ICCCNT51525.2021.9579598. 44
8. C. Jiang, J. Melody, G. Liu, L. Zheng, W. Luan; Credit card misrepresentation discovery: a clever methodology utilizing total procedure and criticism component, *IEEE Internet Things J.*, (5) (Oct. 2018) 3637-3647, doi:10.1109/JIOT.2018.2816007.
9. Z. Li, G. Liu, S. Wang, S. Xuan, C. Jiang; Credit card misrepresentation identification through kernel based directed hashing, *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and*

Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (*SmartWorld/ SCALCOM/ UIC/ ATC/ CBDCom/ IOP/ SCI*), *Guangzhou*, (2018), 1249-1254. doi:10.1109/SmartWorld.2018.00217.

10. Phuong Hanh Tran, Kim Phuc TRAN, Truong Huong, Cédric Heuchenne, Phuong Hien Tran and Huong Le; Real Time Data-Driven Approaches for Credit Card Fraud Detection, *Proceedings of the 2018 international conference on e-business and applications,* (2018), 6-9.

    Doi: 10.1145/3194188.3194196.

11. S. Akila, U.S. Reddy; Credit card misrepresentation discovery utilizing non-covered danger based sacking outfit (NRBE), *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore*, (2017), 1-4.

    doi:10.1109/ICCIC.2017.8524418.

12. A. Peter, K. Manoj and P. Kumar; Blockchain and Machine Learning Approaches for Credit Card Fraud Detection, *5th International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), (2023), 1034-1041.

13. Y. Jain, N. Tiwari, S. Dubey, S. Jain; A similar investigation of different Visa extortion discovery methods, *Int. J. Late Technol. Eng.*, 7 (2019), 402-407.

14. Abhishek Kumar, Debachudamani Prusti, Shubham Ingole and Santanu Rath; Real time SOA based credit card fraud detection system using machine learning techniques. (2021), 1-6. 10.1109/ICCCNT51525.2021.9579598.

15. Simulated Credit Card Transactions generated using Sparkov, Credit Card Transactions Fraud Detection Dataset link https://www.kaggle.com/datasets/kartik2112/fraud-detection