Original Article

Credit Card Fraud Detection Using Machine Learning

AMIT KUMAR, ANANT JAIN, Mohd Ariz, Nitin Kumar

Department of Computer Science and engineering, Meerut Institute of Engineering and Technology, Meerut, U.P., India
{Amit.kumar.cs.2019, Anant.jain.cs.2019} @miet.ac.in

DOI: 10.47750/pnr.2022.13.S10.701

Abstract

Fraud detection pertained to industries such as retails, banking services, financial services, healthcare, etc. Fraud detection is a campaign undertaken toprevent acquisition of illegal money or property under false belief. As the number of crimes were increasing, it was very difficult to detect online fraud . This research aims to examine some suitable ways to identify the credit card fraud activities that impact negatively on financial institutions. To compare present machine learning algorithms with ML techniques which already existed, we performed a comparative analysis and determined which algorithmwould best predict the fraud transactions by recognizing the pattern that is differentiable from other patterns. Our algorithms were trained above two methods(oversampling and undersampling) of the dataset. The algorithm which performs the best is chosen in the study . AUC score was used to study the results of the algorithm. The results concluded are as follows :

1. The study affirmed that the performance of the algorithms like Decision Trees, K-Means, Random Forest,Neural Network, Xgboost and ,Logistic Regression was better than other algorithms.
2. The tree algorithms like Decision Trees , Xgboost and Random Forest were the best models to predict the frauds with AUC scores of 1.01%, 0.98% and 1.00% respectively.

Keywords: Machine Learning Algorithms, Re-sampling Methods, Banking andFinancial Sector, Machine Learning Classifiers, Fraud Detection..

1. Introduction:

The things in this world are becoming more digitized so cybercrime like debit card or credit card fraud are increasing. The Bureau of Consumer Financial Protection of the Consumer Credit Card Market presented a report in 2019 stating “fraud is still a constant and an expensive reality of credit card market.”This pitiful situation is negatively affecting the public and private organizationsall over the globe . International transactions on the credit card that are flowing above some specific limits are used to mark some of the transactions as fraudulent. Still, it was also noted that 65 % such transactions were reported wrong, which resulted in downfall of sales of merchants. This study investigates techniques that are to be accepted in identifying the credit cards fraud and how the suggested solutions could be helpful in solving this fraudulent activity [1][2][3].

2. LITERATURE REVIEW

Great work had been done in this field. Here, a scrutiny will be done on some articles to recognize the works that is already done. Here we discuss ML. For example Decision Tree, Random Forest, Logistic Regression, XGBoost, (unsupervised methods) such as K – Means Clustering. Researchers like Manirajet al. (2019), Dornadula (2019), Shirgave et al. (2019), Awoyemi et al. (2017), Azhan (2020), Sadineni et al. (2020),Joshi et al. (2020), Priya & Saradha (2021), has identified unsupervised and supervised methods as the maximum used methods [4][5][6].

Maniraj et al. (2019) illustrates modeling of a data set using machine learning with Credit Card Frauds Detection. Maniraj et al detected transactions that were 100% fraudulent as the incorrect fraud classification are minimized.. The approach step is about to analyzer, preprocess dataset and to deploy multiple anomaly detection algorithm like the Local Factor Isolation Forest algorithm onthe PCA transformed Credit Card Transaction Data [7][8][9].

Shirgave te al investigated credit cards fraud detection by Machine Learning. He examined various frauds detection paradigm by Machine Learning and corelate them using methods like precision, accuracy and distiction. He proposed a FDS that used a supervised Random Forest paradigm [10][11][12].

Sadinen (2020) also examined the researches includes Machine Learning paradigm. These investigations considered several machine learning paradigm like Artificial Neural Network (ANN), Decision Trees, SVM , Logistic Regression and Random Forest for identify the fraudulent activities which are done by credit card. The presentation of this paradigm concludes the precision accuracy.

Methodology:

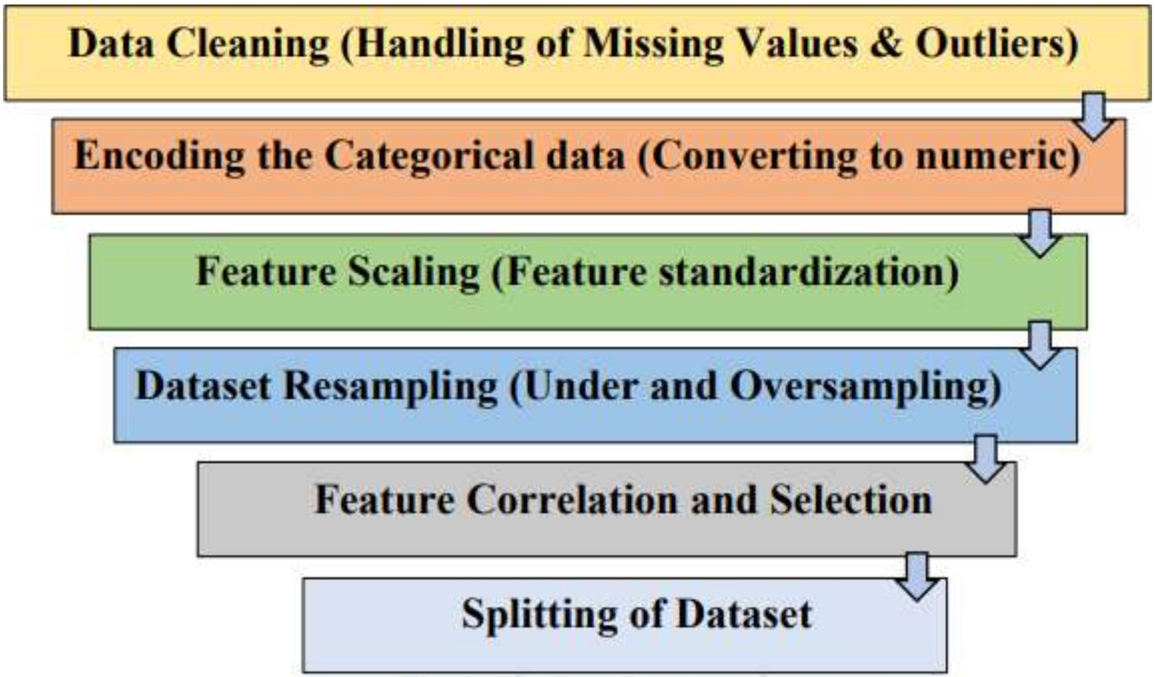

Data Cleaning (Handling of Missing Values & Outliers)
Encoding the Categorical data (Converting to numeric)
Feature Scaling (Feature standardization)
Dataset Resampling (Under and Oversampling)
Feature Correlation and Selection
Splitting of Dataset

Fig. 3: The Data Preprocessing steps

1. DATA CLEANING

This dataset which we are using to detect the fraud is imported by using the python's import command. After that the cleaning of data is achieved. When the data is being cleaned we have to carry out two chores ; 1. Pull out all the missing and null values 2. Handle all outliers. Our dataset encloses 1048574 transactions .Null values in our dataset were not observed [13][14][15]. Moreover, no missing values were there in the dataset. Next step is to find outliers. They are the observations which are numerically different from the other data. The technique which is used to detect the existence of outliers is called boxplot [16][17][18].

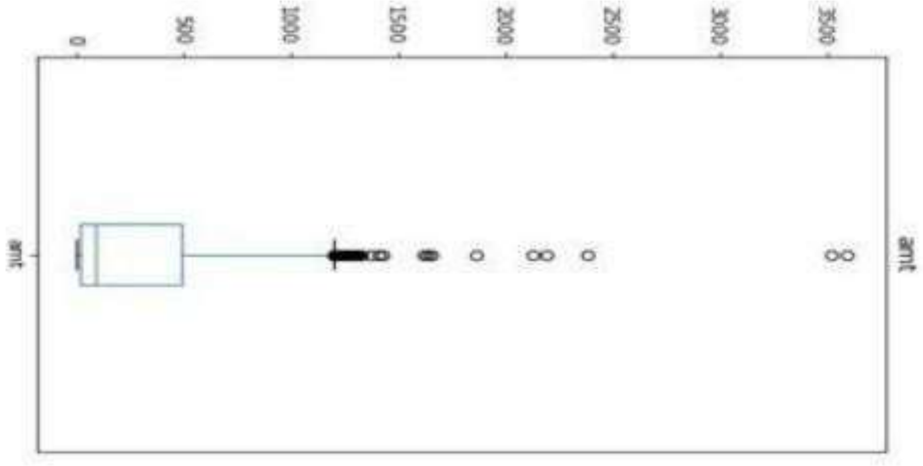
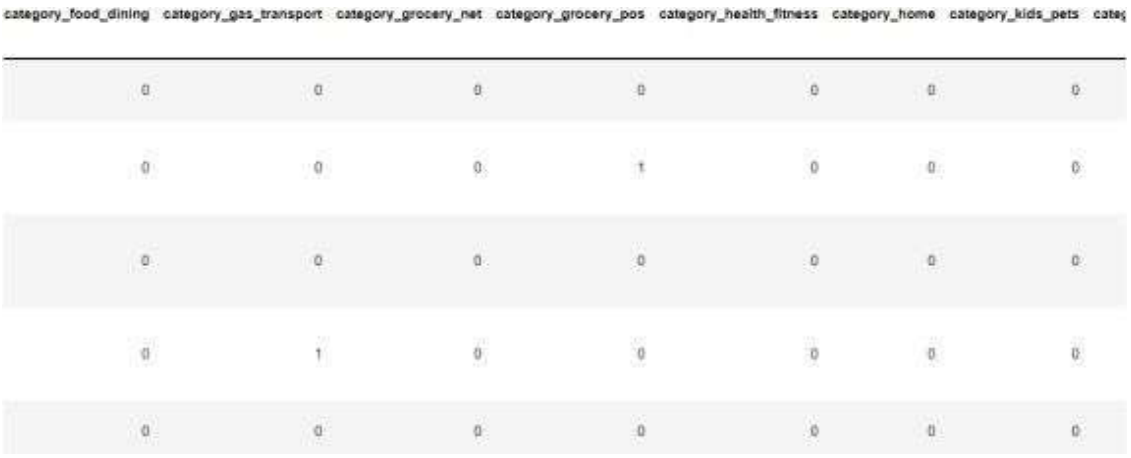


Fig. 4: Boxplot of the amount feature

2. ENCODING CATEGORICAL VARIABLES

After the dataset is cleaned , the categorical attributes are converted into numeral data. This is because maximum ML algorithms performs better with numeral data. Some ways are present to convert categorical data to numerical data with each technique having their own advantages and disadvantages. In this research, One-Hot Encoder was used to convert categorical data to numerical data [19][20][21].



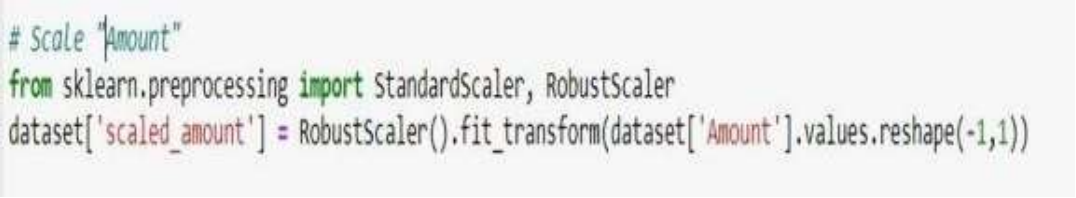
category_food_dining	category_gas_transport	category_grocery_net	category_grocery_pos	category_health_fitness	category_home	category_kids_pets	categ
0	0	0	0	0	0	0	
0	0	0	1	0	0	0	
0	0	0	0	0	0	0	
0	1	0	0	0	0	0	
0	0	0	0	0	0	0	

Fig 5 Sample of absolute features using One Hot Encoder

3. FEATURE SCALING

Feature Scaling was also the step of a data Initialization which is use in normalizing the scale of the non-dependent variables in a given dataset.Depends at adopted scaling techniques, it was set around zero or between 0and 1. The given variables has enormous value that were relevant to additional given variable, the bigger value could be overlooked or change some other machine learning methods. We had also done feature scaling by applying the Robust Scaler method, which was also called as the robust standardization. Scaling shall also be done by calculating , median of 50(th) percentile, the 75(th) and 25(th) percentiles. These values of the variables then having their medians subtracted and were divided by interquartileranges , which was the difference between 75(th) and 25(th) percentile. The figure Below showing a feature scaling process [22].

Featured Scaling using Robestscalingmethods()



```
# Scale "Amount"
from sklearn.preprocessing import StandardScaler, RobustScaler
dataset['scaled_amount'] = RobustScaler().fit_transform(dataset['Amount'].values.reshape(-1,1))
```

4. DATASET RESAMPLING

This is the method of economically using data samples to improvised the precision and calculate the unpredictability of the population variable. Nested resampling methods are using to carry out the datasets resampling easily. Datasets using in this study are highly inappropriate so that's why we carrying out a resampling method like Oversampling and Undersampling.

5. FEATURE CORRELATION AND SELECTION

Each of these feature that were obtained in these datasets may be not usefulin build up the machine learning models or for implementing these necessary predictions. Applying some of these features may improvised the prediction accuracy. That is why features correlation performing the various purpose in the creation of a better machine learning models.Features that having a high correlation are suitable to linear dependent and having almost the similar affect to a dependent variable. So, one feature is dropped when both features generate the high correlation.

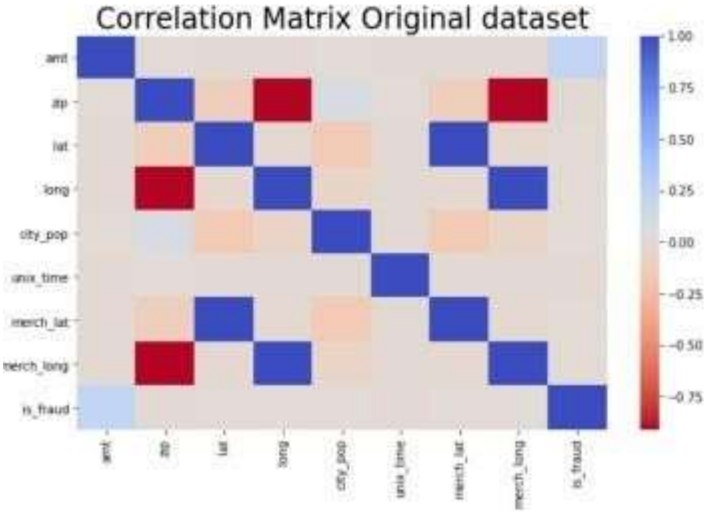


Fig 6 Heatmap for the Original dataset

6 SPLITTING OF DATA INTO TRAINING AND TEST

The objectives of Machine learning paradigm was learnt by earlier experiences or making use of these information to create recent results.

Evaluation about the performance was generally perform on portion of the entire datasets by doing trainings upon sample, and these leftover datasets were used for evaluating the model's presentation. Our studies shows the dataset is divided into the 70 /30 ratio; The 70 percent of datasets was used in model training and the 30% was used to evaluate the model performance. Frameworks that were also called as hyperparameter of models, were determined within the models training, and the models hyperparameters as well used to finding the finest model fit for the machine learning models.

Technology Used:

Machine Learning Methods Used:

Machine Learning is undoubtedly one of the all the foremost powerful technologies that this world will ever see.. It's a component of computer science that is used to develop a capability to systems or programs to find outautomatically i.e. from their past which helps them in enhancing their future performance. Its main focus is on the event of computer programs that accessdata and specific techniques to induce the required results. Machine Learning techniques are wont to find the educational patterns within complex data thatwe might otherwise struggle to find. Then, the hidden patterns and feedback isemployed to predict the specified result.
Following are the used ML algorithms:

Artificial Neural Network (ANN):

ANN may be a model sort of a human brain's nerve system that encompasses asizable amount of nodes connected to every other. Each node comprises two states: 0 is thought to be nonactive and 1 means active. Every node encompasses a positive or negative weight attached to them to regulate the strength of the node. ANN provides samples of information to coach the machine. The trained machine is employed to detect the pattern of hidden date.
The advantages of using ANN are- they need the power to find out and model non-linear and sophisticated relationships. they'll generalize. It doesn't imposeany restrictions on the input variables (like how they ought to be distributed). Its region growing-based segmentation method is improved by using the extracted intensity features from ROIs and applying the ANN to get an adaptivethreshold. Three layer linearisation ANN with 1000 hidden layer nodes was used.
The ANN with 1000 hidden layer nodes generalizes better than networks withalittle number of hidden nodes when trained with backpropagation and "earlystopping". 10-fold cross-validation is employed for training and testing. This ensures that a test sample is rarely used for training.

DECISION TREE

Decision tree method works easily by directing the transaction in a certain direction basis on a features obtained from a data. This follows the elementaryroot question and branches into which details were used to make the particular components thats finally terminating the last branch or the tree's leaves. Decision Trees were semi constant supervise learning models that could be used in classification and regression objective where the continuous division of the data is depends on the certain parameter.

LOGISTIC CLASSIFICATION

This demonstrate chances of the outputs that sometimes either Binomial or multinomial. Its acquires a sigmoid function to describes a data input and the relationship between independent and dependent variables. It will also saw ina present researches to classify the transaction if it is fraud or accurate. It was ideal effective, whereas, this can be overfitted high dimension datasets . It provides better correctness and make no assumptions on the scattering of theclasses in a featured space ,may be others methods use. The limitation is it used the assumption in linearity amongs the independent variables and the dependent variables.

RANDOM FOREST

This is the approach adopted for solving both classification and regression problem. It was the pool of the massive number of distinct decision trees which are called as "forest". Each distinct trees made the class predictions. whatever classes that have the maximum votes was taken for prediction. Then the methods adopts the bagging approach for creating the sets of Decision Trees that would made the forest. Advantages of this paradigm are that the selections of features is not necessary and ,It flows the model fastly and handles the errors wisely. The disadvantage of the paradigm is , It is too sensitive data with the various attributes and with too many values and it caneasily mark that values as a fraud.

XGBOOST CLASSIFIER

XGBOOST defines as the EXTREME GRADIENT BOOSTING. An assemble paradigm that implemented a extreme gradient boosting decision trees was blueprint for the changes in momentum and excellent performance.

It is machine learning paradigm that will used for stop the data science and machine learning complications. It works on interfaces like , Python ,c++, R, Java, kotlin and Command line interface. It was an exceptionally adjustable and tractile structure that works on the Regression, Classification and also handles the ranking problems.

K-MEANS CLUSTERING

It is a unsupervised learning paradigm use for a crucial grouping of data. It collects the random data points into various k clusters. Then this was defined as the unsupervised, as the points has no independent classification. This strategy was the major perceptive strategy in the data mining, the methods for clustering paradigm will effects the result of a clustering straight forwardly.

CONCLUSION

This study mades many important improvements. In this we are talking about online transactions which are done by the credit card, that leads to the credit cards fraud, and this study improvises ML algorithms for fraud detection. At last, we concludes that Random Forest could be the excellent fit for our model.It will be concludes that oversampling works better because the smaller number of observation helping in the training our model effectively.

Oversampling is the ideal sampling method in this real time scenario as this given information contained the pattern is not departed.

REFERENCES

1. Jason Brownlee. (2021). A Gentle introduction to XGBoost for Applied Machine Learning.

2. https://machinelearningmastery.com/gentle-introduction-xgboostapplied-mac hine-learning/.

3. Jason Brownlee. (2021). Bagging and Random Forest for imbalanced Classification.

4. https://machinelearningmastery.com/bagging-and-random-forestfor-imbalance d- classification/

5. Jisha, M.V. & Vimal, D. (2020). Population-based Optimized and Condensed Fuzzy Deep Belief Network for Credit Card Fraudulent Detection. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110970.

6. Joshi, Aruna & Shirol, Vikram & Jogar, Shrikanth & Naik, Pavankumar & Yaligar, Annapoorna. (2020). Credit Card Fraud Detection Using Machine

7. Learning Techniques. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology. 436-442.

8. Mohari, Ankit & Dowerah, Joyeeta & Das, Kashyavee & Koucher, Faiyaz & Bora, Dibya & Bora. (2021). A COMPARATIVE STUDY ON CLASSIFICATION ALGORITHMS FOR CREDIT CARD FRAUD DETECTION.

9. More, Rashmi & Awati, Chetan & Shirgave, Suresh & Deshmukh, Rashmi & Patil, Sonam. (2021). Credit Card Fraud Detection Using Supervised Learning Approach. International Journal of Scientific & Technology Research.

10. Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169.

11. Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.

12. Irfan, Daniyal, et al. "Prediction of Quality Food Sale in Mart Using the AI-Based TOR Method." Journal of Food Quality 2022 (2022

13. Pramanik, Sabyasachi, et al. "A novel approach using steganography and cryptography in business intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.

14. Mohseni, Sina, et al. "Machine learning explanations to prevent overtrust in fake news detection." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 15. 2021.

15. Narayan, Vipul, et al. "To Implement a Web Page using Thread in Java." (2017).

16. Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).

17. Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.

18. Faiz, Mohammad, et al. "Improved Homomorphic Encryption for Security in Cloud using Particle Swarm Optimization." Journal of Pharmaceutical Negative Results (2022): 4761-4771.

19. Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.

20. Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

21. Narayan, Vipul, et al. "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).

22. NARAYAN, VIPUL, A. K. Daniel, and Pooja Chaturvedi. "FGWOA: An Efficient Heuristic for Cluster Head Selection in WSN using Fuzzy based Grey Wolf Optimization Algorithm." (2022).