Credit card Fraud Detection based on Machine Learning Algorithms

Heta Naik
Student of M.Tech Compuer Engineering
NMIMS University, Bombay, India

Prashasti Kanikar
Faculty of M.Tech Compuer Engineering
NMIMS University, Mumbai, India

ABSTRACT

Now a days online transactions have become an important and necessary part of our lives. As frequency of transactions is increasing, number of fraudulent transactions are also increasing rapidly. In order to reduce fraudulent transactions, machine learning algorithms like Naïve Bayes, Logistic regression, J48 and AdaBoost etc. are discussed in this paper. The same set of algorithms are implemented and tested using an online dataset. Through comparative analysis it can be concluded that Logistic regression and AdaBoost algorithms perform better in fraud detection.

Keywords

Credit card, Fraud detection, Machine learning, supervised learning, Naïve Bayes, Logistic regression, J48, AdaBoost

1. INTRODUCTION

Due to rise and acceleration of E- Commerce, there has been a tremendous use of credit cards for online shopping which led to High amount of frauds related to credit cards. In the era of digitalization the need to identify credit card frauds is necessary.

Fraud detection involves monitoring and analyzing the behavior of various users in order to estimate detect or avoid undesirable behavior. In order to identify credit card fraud detection effectively, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds.

Algorithm can differentiate transactions which are fraudulent or not. Find fraud, they need to passed dataset and knowledge of fraudulent transaction. They analyze the dataset and classify all transactions.

2. LITERATURE REVIEW

You Dai, et. al [2] In this paper, they describe Random forest algorithm applicable on Find fraud detection. Random forest has two types, i.e. random tree based random forest and CART based random forest. They describe in detail and their accuracy 91.96% and 96.77% respectively. This paper summarise second type is better than the first type.

Suman Arora [3] In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree, naïve Bayes and KNN algorithms compare each other i.e. 94.59%, 95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summarise of this paper, SVM has the highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335.

Kosemani Temitayo Hafiz [4] In this paper, they describe flow chart of fraud detection process. i.e. data Acquisition, data pre-processing, Exploratory data analysis and methods or algorithms are in detail. Algorithms are K- nearest neighbour (KNN), random tree, AdaBoost and Logistic regression accuracy are 96.91%, 94.32%, 57.73% and 98.24% respectively.

3. WORK DONE

Find fraud detection need transaction dataset and for finding or classifying need some algorithms. There are plenty of algorithms for finding fraudulent transaction, so first select some better algorithms from Literature review. And Implement better algorithms in python for classifying fraudulent and non-fraudulent transaction.

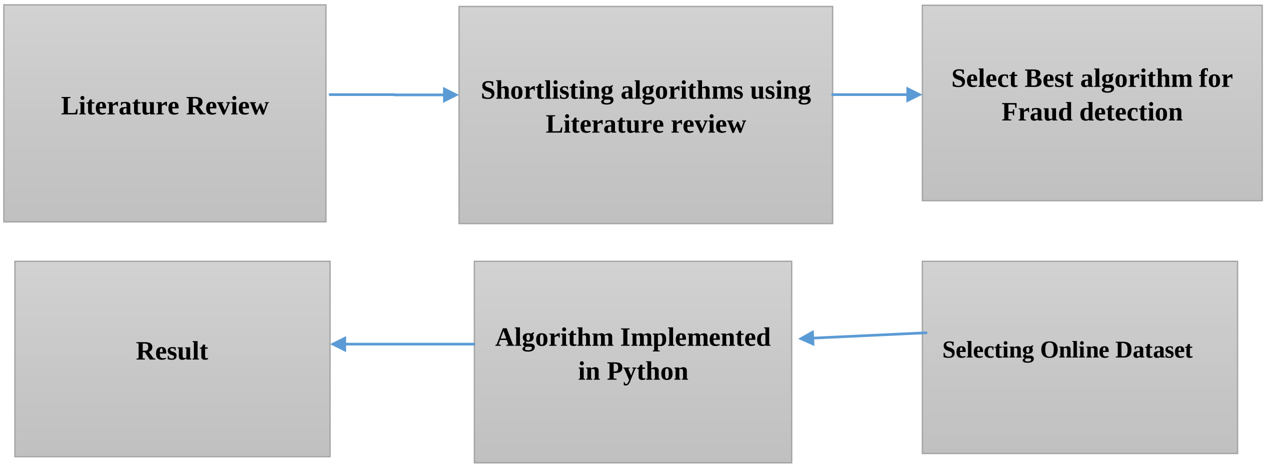
Literature Review
Shortlisting algorithms using Literature review
Select Best algorithm for Fraud detection
Selecting Online Dataset
Algorithm Implemented in Python
Result

Fig. 1 Flow of finding fraud detection

4. SHORTLISTED ALGORITHMS

There are many algorithms which can be used in credit card fraud detection. But these algorithms are more Powerful in this fraud detections. So, compare all the algorithms with their advantages and disadvantage

Table 1. Shortlisted algorithms in detail

Algorithm	Accuracy	Advantages	Disadvantages	Related to Database
K – nearest neighbor	97.69%	• No use of predictive model before classification. • Compare algorithms and their power methods KNN is best	• In KNN algorithm accuracy depends upon neighbours distance • It cannot detect the fraud at the time of transaction.	• KNN works well with a small number of input variables • Better if all data has same scale
Naïve Bayes	97.92% / 70.13	• High processing and detection speed/high accuracy	• Excessive training need / expensive	• It good if dataset has plenty of input but small number of records
Random Tree	94.32%	• It can handle thousands of input variables without variable deletion	• It is over fit for classification /regression tasks with noisy dataset	• Easily work on large databases
Logistic Regression	54.86%	• This algorithm gives simple formula for classification. • Work better on linear dataset.	• Not preferable on non-linear data • It is not capable of handling fraud detection at the time of transaction	• This algorithm wants dependent and independent attributes • This algorithm return values between 0 end 1
Outlier		• Using less memory • Computation is required • Works fast and well on online large datasets	• It can handle thousands of input variables without variable deletion	• Good in large datasets
AdaBoost	57.73%	• It is a powerful classifier that works well on both basic and more complex recognition problems	• It can be sensitive to noisy data and outliers.	• This algorithm use weighted dataset
J48	93.50%	• This algorithm use weighted dataset	• This algorithm can be payoff but there is chances to get different decision	• Give decision tree as a result

5. SELECT ONLINE DATASET

Find a fraud detection choose a sample dataset. In the dataset, given Credit card usage, purpose, Current balance in credit card, Average credit balance, Holder of a credit card, Holder status, CC age, Holder's Property, Housing, Job, Employment, Location, Own telephone, Foreign worker etc. Credit Card and Holder's Detail. In this dataset total 1000records are there and that was pre-processed.

6. SELECTED ALGORITHM FOR IMPLEMENTING

On the Literature review many algorithms are applied on Fraud detection. On the survey bases Naïve Bayes, Logistic regression, J48 and AdaBoost are better than other algorithms for fraud detection.

1) Naïve Bayes

Naïve Bayes is a classification algorithm. This algorithm depends upon Bayes theorem. This is simple and very powerful algorithm.

- Bayes theorem: Bayes theorem find probability of event occurring given probability of another event that has been already occurred.

P (A/B) = (P (B/A) P (A)) / P (B)

Where, P (A) – Priority of A

P (B) – Priority of B

P (A/B) – Posteriori priority of B

- Naïve Bayes algorithm is easy and fast. This algorithm need less training data and highly scalable

2) Logistic Regression

- This algorithm similar to linear regression algorithm. But linear regression is used for predict / forecast values and Logistic regression is used for classification task.
- Linear regression classified as
  - ➢ Binomial – 2 Possible types (i.e. 0 or 1) only
  - ➢ Multinomial – 3 or more Possible types and which are not ordered
  - ➢ Ordinal – Ordered in category (i.e. very poor, poor , good, very good)
- This algorithm easy for binary and multivariate classification task.

3) J48

- J48 algorithms used to generate a decision tree and it is for classification task.
- J48 is an extended of ID3 (Iterative Dichotomieser 3). This algorithm has some special features such as, rules derivation, continues value range, decision tree pruning, etc.
- J48 algorithm is most extensively analyzed area in machine learning. They analyze based on generated decision tree and understandable rules.
- This algorithms works on constant and categorical variables. .

4) AdaBoost

- AdaBoost is a machine learning algorithm. Mainly developed for binary classification. This algorithm is used to boost the performance of decision tree. .
- For AdaBoost, Each instance in the training dataset is weighted. Initial weight is set To: Weight (xi) = (1/n)Where, xi – ith training instance

  n – Number of training instance
- This algorithm mainly for classification rather than regression. So that AdaBoost algorithm is used in fraud detection because this classify the transaction which transactions are fraudulent and non-fraudulent.

7. IMPLEMENTED ALGORITHMS IN PYTHON

Table 2. Result of Implemented algorithms

Name	Naïve Bayes	Logistic Regression	J48	AdaBoost
Accuracy	83.00%	100.00%	69.93%	100.00%
Time Duration	10.0	38.1	46.24	2.80
Method	Classification method	Machine Learning	Supervised Learning	Machine Learning
Training : Testing	70 : 30	70 : 30	70 : 30	70 : 30
Inbuilt Packages	Gaussian NB	Logistic Regression	Decision Tree Classifier	AdaBoost Classifier

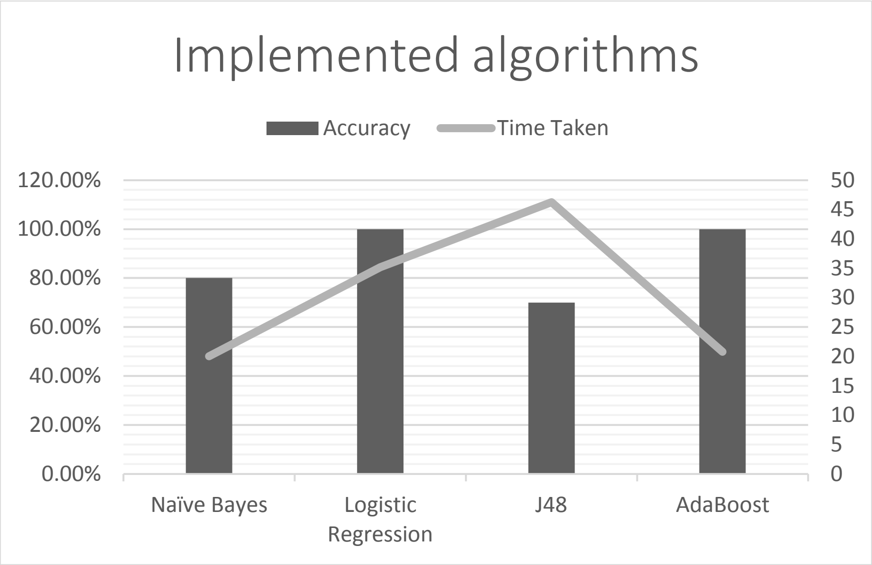

Fig. 2 Analysis chart on Implemented algorithms

This four algorithms are implemented in python using their library and packages. All algorithms need some amount of training dataset. Implemented algorithms gave result as an accuracy, Time duration and classify fraud transactions.

8. CONCLUSION

After implementing algorithm, highest accuracy gave Logistic Regression and AdaBoost respectively 100% and 100%. And taken very low time is AdaBoost. So, concluding that for fraud detection AdaBoost algorithm is better than other algorithms.

9. REFERENCES

[1] Heta Naik, “Credit card fraud detection for Online Banking transactions”, International Journal for Research in Applied Science & Engineering Technology, pp 4573-4577, 2018https://www.ijraset.com/fileserve.php?FID=16732

[2] You Dai, Jin Yan, Xiaoxin Tang, Han Zhao and Minyi Guo, "Online Credit CardFraud Detection: A Hybrid Framework with Big Data Technologies", IEEE TrustCom/BigDataSE/ISPA , pp 1644 -1651, 2016

[3] Suman Arora , "Selection of Optimal Credit Card Fraud Detection Models Using a Coefficient Sum Approach" , International Conference on Computing, Communication and Automation (ICCCA2017), pp 482 - 487, 2017

[4] Kosemani Temitayo Hafiz, Dr. Shaun Aghili and Dr. Pavol Zavarsky, “The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada”,

[5] N.Malini and Dr.M.Pushpa , "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection" , 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17) , 2017

[6] Anusorn Charleonnan , "Credit Card Fraud Detection Using RUS and MRN Algorithms" , The 2016 Management and Innovation Technology International Conference (MITiCON-2016) , pp 73 - 76, 2016

[7] John Richard D. Kho and Larry A. Vea, “Credit card Fraud detection based on transaction Behavior”, IEEE Region 10 Conference (TENCON), Malaysia, pp 1880 – 1884 , November 2017

[8] Fahimeh Ghobadi and Mohsen Rohani, "Cost Sensitive Modeling of Credit CardFraud Using Neural Network Strategy" , IEEE ICSPIS 2016, Dec 2016

[9] S Md. S Askari and Md. Anwar Hussain, "Credit Card Fraud Detection Using Fuzzy ID3" , International Conference on Computing, Communication and Automation (ICCCA2017), pp 446 - 452 , 2017

[10] Sarween Zaza and Mostafa Al-Emran, "Mining and Exploration of Credit Cards Data in UAE", Fifth International Conference on e-Learning , pp 275-79 , 2015

[11] Krishna Keerthi Chennam and Lakshmi Mudanna, “Privacy and Access Control for Security of Credit Card Records in the Cloud using Partial Shuffling”, IEEE International Conference on Computational Intelligence and Computing Research, 2016

[12] Rajeshwari U and Dr B Sathish Babu, “Real-time credit card fraud detection using Streaming Analytics”, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp 439 – 444, 2016

[13] John O. Awoyemi, Adebayo O. Adetunmbi and Samuel A. Oluwadare, “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis”, IEEE , 2017

[14] Mukesh Kumar Mishra and Rajashree Dash, “A Comparative Study of Chebyshev Functional Link Artificial Neural Network, Multi-Layer Perceptron and Decision Tree for Credit Card Fraud Detection”, International Conference on Information Technology, pp 228 -233, 2014

[15] Pornwatthana Wongchinsri and Werasak Kuratach, “A Survey - Data Mining Frameworks in Credit Card Processing”, IEEE, 2016

[16] Yufeng Kou, .et. al. , “Survey of Fraud Detection Techniques”, International Conference on Networking, Sensing & Control, pp 749 – 754, 2004

[17] John O. Awoyemi, et.al., “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis” IEEE , 2017

[18] Shiyang Xuan, et.al., “Random Forest for Credit Card Fraud Detection”, IEEE – 2018

[19] *Sahil Dhankhad, et.al., “*Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study” , IEEE International Conference on Information Reuse and Integration for Data Science, pp 122-125, 2018

[20] R. Brause, et.al., “Neural Data Mining for Credit Card Fraud Detection”

[21] Zahra Kazemi and Houman Zarrabi, “Using deep networks for fraud detection in the credit card transactions”, IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), pp 0630 – 0633, 2017

[22] Samaneh Sorournejad, et.al., “A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective”, pp 1 - 26

[23] http://weka.8497.n7.nabble.com/file/n23121/credit_fruad.arff