# Finalize Attributes and using Specific Way to Find Fraudulent Transaction

**Heta Naik[1], Prashasti Kanikar[2]**

*[1]Student, Dept of computer Engineering, MPSTME NMIMS, Mumbai, India*
*[2] Professor, Dept. of Computer Engineering, MPSTME NMIMS, Mumbai, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**ABSTRACT -** *Online transactions are increasing because of they make our life easy. Other hand parallel side fraud transactions are also increasing and that is not good. For finding fraudulent transaction, need some strong mechanism or algorithm. In this paper first shortlisting algorithms from Literature review. Their 6 shortlisted algorithms. Then finalise one online transaction dataset. That dataset contains 20 attributes. That all attributes related to transaction directly or indirectly. For finding attributes priority or effect of attributes use Information gain and gain ratio techniques. After this result finalise 4 category 20, 14, 10 and 7 attributes. This number of attributes category apply on WEKA with shortlisted algorithms. On basis of this WEKA result shortlisted 4 algorithms for Implementation. AdaBoost, Logistic regression, J48 and Naïve Bayes algorithms are implemented in python. And final conclusion of research AdaBoost is better for finding fraudulent transactions compare to other algorithms.*

**Keywords:** Fraud, Credit card, Algorithms, machine learning, attribute selection, Information gain, gain ratio, WEKA, AdaBoost

## INTRODUCTION

Now a day's number of frauds are increasing. Fraud has many ways to attempt such as Internet Fraud, hacker, Credit card fraud, Computer crime, etc. Specific credit card fraud occurred via many ways i.e. Skimming, Balance transfer, Account takeover, online transactions. Etc. Day by day online transactions are increasing because of they makes our life easy. Ticket booking, glossary shopping, pay online bills. Etc. are done easily online. Other side fraudulent transactions are also increasing. Find or classify the fraudulent transactions need some mechanism or algorithms. Finding fraudulent transactions KNN, logistic regression, J48, AdaBoost, outlier, Decision tree, etc. algorithms can be used. There are some parameters to classify the fraudulent i.e. Time, amount, transaction frequency, etc. Finding fraud need some information such as Card holder name, address, age, balance, transaction history, expiry date, transaction time, amount, etc. Analyse fraud or not they need passed analysed results and passed data to analyse and classify the transactions.

## LITERATURE REVIEW

Dr. Sanjay Kumar Dwivedi, et.al [14] In this paper they describe Web mining in detail. In web mining there are 3 types of mining: Web content mining, web usage mining and web structure mining. Web content mining is applicable on Text, Audio, Video, Images and Structured record. Web usage mining apply on Web server logs, Application server logs and application logs. And 3rd one Web structure mining is apply on Hyperlinks, Document structure and application level. For the data mining data pre-processing is very important. Data pre-processing have different phases such as data collection, data cleaning, session identification, user identification and path completion. Web data is inconsistent, noisy and irrelevant by the nature so that this data can't be used for analysis or mining. For this data first of all they need to pre-process on dataset after it can be used for mining and analysis.

Aman Srivastava, et.al [15], In this paper they describe what online transaction is and how they work. They compare different detection techniques such as Neural networks, Rule Induction, Case-based reasoning, Genetic algorithm, Inductive logic programming Expert System and Regression. Neural network gave better result. Neural network compare with human brain and human techniques. Human always try to learn something new and learn from experience. Human not follow step by step instruction same as neural network. Neural network also lean from past result and training datasets. This algorithm learn from itself. So, that compare to others Neural network is best to other. This paper suggest one system or flow to find the fraudulent transactions. This system try to find fraud from merchant side. Payment gateway have some details such as credit or debit card number, expiry date time, etc. Merchant pass shipping address, amount etc. then payment gateway send some necessary detail to Fraud detection system. And this Fraud detection system made by Neural network for train data itself and generate the output. That output is decision of transaction i.e. fraudulent or not. This decision output send to payment gateway. In the fraud detection system 2 phase such as Learning and testing. Neural network learn itself from past result and test as Non- fraudulent, Doubtful, Suspicious or Fraudulent transaction. This technique is very sufficient because merchant know very well about transaction so that number of fraudulent transaction less compare to others.

S. Benson Edwin Raj, et.al [18], In this paper they describe what is credit card fraud and how to detect credit card fraud. For detecting credit card fraud give some methods such as Fusion approach using Dempster - Shafer theory and Bayesian learning, BLAST-SSAHA Hybridization, Hidden Markov Model, Fuzzy Darwinian Detection and Bayesian and Neural Networks. Dempster – Shafer theory and Bayesian learning is hybrid approach for credit card fraud. This Fusion approach proposed a fraud detection using Information fusion and Bayesian learning for counter credit card fraud. This FDS system has four components namely, Rules based filter, Dempster – safer adder, Bayesian learner and Transaction history dataset. This technique gave high accuracy and high processing speed. They improve detecting rate and reduce false alarm. BLAST-SSAHA is combination of hybrid BLAST and SSAHA algorithm. BLAST- FDS is a two stage sequence alignment which are profile analyser (PA) and deviation analyser (DA). PA determine similarity of an incoming sequence of transaction and DS determine unusual incoming transactions. This algorithm is good. Gave high accuracy and processing speed also fast enough. Fuzzy Darwinian detection system uses genetic fuzzy logic rules and classify credit card transaction in 'Suspicious' and 'Non- suspicious'. In this techniques first data cluster in 'low', 'medium', and 'high' range. This system gave good accuracy and intelligibility level for real data. Conclusion of this paper was Fuzzy Darwinian and BLAST-SSAHA have very high accuracy in terms of TP and FP. But processing speed is fast enough to enable online detection of credit card fraud detection in BLAST-SSAH.

Samaneh Sorournejad, .et. al [9], A survey of credit card fraud detection techniques as on data basis and techniques oriented perspectives Let, in 21st.century credit card is a very important role plays. For using in daily routines, daily needs, e-shopping and other variants. It should be too many applications for this generation, with using credit card people has buy whatever needs. With the help of credit card we reduced our external affairs or external corridor time which is minimized with smartly using of credit card. Credit card is a good for those who really punctual for its personal economy & finance statistics. At the current situations of the world, banking systems and financial companies are expand to their facilities to valuable clients for their innovative services without using their money and as on credit period they given on their client for the same. Therefore credit card is a very harmful and useful part of person's financial life. From the bank or financial organizations are given credit card to their client on basis of client's financial banking record, government taxation policy's involvement and other liable priorities which should follow or fulfil their criteria that clients are eligible for credit card. It means person who is in actual? That also should be defined on basis of credit card holder or non-holder. Credit card is not get easily to any other person for any uses or any applications. It should be strictly in banking criteria and its civil report basis they has holder of it. Credit card fraud is also a big issues now a days, some fraud peoples are using other credit card fraudily and loses are facing some card holders. Its occurs only for a some lack of knowledge, some mismanagement, and missed their credit card, then fraud people are using benefits of other for personal needs.

## CREDIT CARD FRAUD DATASET

Transactions information are very authenticate. Finding a fraud, need some basic information about credit card and card holder data.

| Attribute Name | Data Type | Description |
|---|---|---|
| **Over_draft** | Qualitative | Status of existing checking account |
| **Credit_usage** | Numerical | Credit usage in month vise |
| **Credit_history** | Qualitative | Credit card history |
| **Purpose** | Qualitative | Purpose of credit card transaction |
| **Current_balance** | Numerical | Credit Amount |
| **Average_credit_balance** | Qualitative | Average credit amount |
| **Employment** | Qualitative | Present employement |
| **Location** | Numerical | Installemt rate in percentage of disposable income |
| **Personal_status** | Qualitative | Personal Status and sex |
| **Other_parties** | Qualitative | Other debtors/ guarantors |
| **Recidence_since** | Numerical | Present residence since now |
| **Property_magnitude** | Qualitative | Property |
| **CC_age** | Numerical | CC age in months |

| | | |
|---|---|---|
| **Other_payment_plans** | Qualitative | Other installment plans |
| **Housing** | Qualitative | Housing |
| **Existing_credits** | Numerical | Number of exidting credit in this bank |
| **Job** | Qualitative | Job |
| **Num_dependent** | Numerical | Number of people being liable to provide maintenance for |
| **Own_telephone** | Qualitative | Telephone is there or not |
| **Foreign_worker** | Qualitative | Foreign workers |

This attributes are related to online transactions. Some attributes are directly and some are indirectly connected to transaction details.

## ATTRIBUTE SELECTION

There are many attributes and that attributes contains transaction related data. So, first find the priority wised attributes from the credit card dataset. Selecting attributes on priority based. Finding attributes priority using two algorithms such as, Gain Ratio and Information Gain.

Information Gain:

- Information gain = information before splitting – Information after splitting
- It works fine for most cases, unless you have a few variables that have a large number of values (or classes)
- Information gain is biased towards choosing attributes with a large number of values as root nodes.

Gain Ratio:

- Gain ratio is modification of information gain that reduces its bias and is usually the best option
- This ratio overcome the problem with information gain by taking into account the number of branches that would result before making the split

Based on both algorithms, finalized attributes from dataset. This two algorithms apply on WEKA. They give attributes priority rank. Based on attributes priority, make 4 categories such as, 20 attributes, shortlisted first 14 attributes, shortlisted first 10 attributes and shortlisted first 7 attributes.

| 20 Attributes (Entire dataset) | Shortlisted 14 Attributes | Shortlisted 10 Attributes | Shortlisted 7 Attributes |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 |
| 4 | 4 | 4 | 4 |
| 5 | 5 | 5 | 5 |
| 6 | 6 | 6 | 6 |
| 7 | 7 | 7 | 20 |
| 8 | 9 | 12 | |
| 9 | 10 | 13 | |
| 10 | 12 | 20 | |
| 11 | 13 | | |
| 12 | 14 | | |
| 13 | 15 | | |
| 14 | 20 | | |
| 15 | | | |
| 16 | | | |
| 17 | | | |

| | | | |
|---|---|---|---|
| 18 | | | |
| 19 | | | |
| 20 | | | |

## SHORTLISTED ALGORITHM USING LITERATURE REVIEW

From literature review there are many algorithms which are used for credit card detection. Such as, K- nearest neighbour, logistic regression, Outliers. J48, AdaBoost, Naïve Bayes, random tree, etc. From literature survey Naïve Bayes has highest accuracy 97.92% and Logistic regression has lowest accuracy of 54.86%.

| Algorithm | Accuracy | Advantages | Disadvantages |
|---|---|---|---|
| K – nearest neighbor | 97.69% | • There is no requirement of predictive model before classification.<br>• Compare to power methods and other known anomaly detection methods, KNN is best | • The accuracy of the method depends on the measure of distance.<br>• It cannot detect the fraud at the time of transaction. |
| Naïve Bayes | 97.92% / 70.13 | • High processing and detection speed/high accuracy | • Excessive training need / expensive |
| Random Tree | 94.32% | • It gives estimates of what variables are important in the classification<br>• It can handle thousands of input variables without variable deletion | • It is over fit for some datasets with noisy classification /regression tasks |
| Logistic Regression | 54.86% | • It produces a simple probability formula for classification.<br>• It works well with linear data for credit card fraud detection. | • It cannot be applied on non-linear data<br>• It is not capable of handling fraud detection at the time of transaction |
| Outlier | | • Using less memory and computation requirements<br>• Works fast and well on online large datasets | • It can handle thousands of input variables without variable deletion |
| 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. |
| J48 | 93.50% | • This algorithm use weighted dataset | • This algorithm can be payoff but there is chances to get different decision |

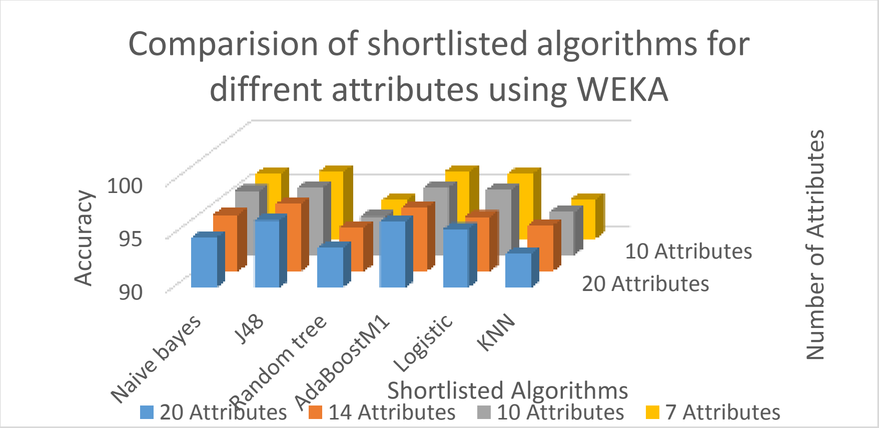
## SHORTLISTED ALGORITHMS USING WEKA

Algorithms from Literature survey i.e. Naïve Bayes, J48, Random tree, AdaBoost, Logistic regression and KNN are implemented on WEKA with Credit card fraud detection.

Selected attributes category wise apply on different algorithms and compare their accuracy.

| Algorithms / Attributes | 20 Attributes | 14 Attributes | 10 Attributes | 7 Attributes |
|---|---|---|---|---|
| Naive Bayes | 94.7 | 95.3 | 96 | 96.1 |
| J48 | 96.3 | 96.3 | 96.3 | 96.3 |
| Random tree | 93.7 | 94.1 | 93.6 | 93.6 |
| AdaBoostM1 | 96.2 | 96 | 96.3 | 96.3 |
| Logistic | 95.5 | 95.1 | 96.1 | 96.1 |
| KNN | 93.2 | 94.3 | 94 | 93.7 |

All algorithms run with 20 attributes, 14 attributes, 10 attributes and 7 attributes. Conclude on WEKA result J48 algorithm has highest accuracy with 20 attributes is 96.3% and lowest accuracy with 20 attributes is KNN 93.2%. From WEKA result analysed that only J48 algorithm gave constant accuracy with different number of attributes and other algorithms are increased or decreased with different number of attributes



Based on WEKA result, Naïve Bayes, J48, AdaBoost and Logistic regression algorithms are shortlisted for implement. Random tree and KNN algorithms has lowest priority with all number of attributes categories.

## IMPLEMENTED ALGORITHMS IN PYTHON

Shortlisted WEKA algorithms are Naïve Bayes, J48, AdaBoost and Logistic regression implemented in Python.

| Name | Naïve Bayes | Logistic Regression | J48 | AdaBoost |
|---|---|---|---|---|
| **Accuracy** | 83.00% | 100.00% | 69.93% | 100.00% |
| **Time Duration** | 1.17 | 3.81 | 4.62 | 2.80 |
| **Training Testing** | 70 : 30 | 70 : 30 | 70 : 30 | 70 : 30 |
| **Inbuilt Packages** | Gaussian NB | Logistic Regression | Decision Tree Classifier | AdaBoost Classifier |

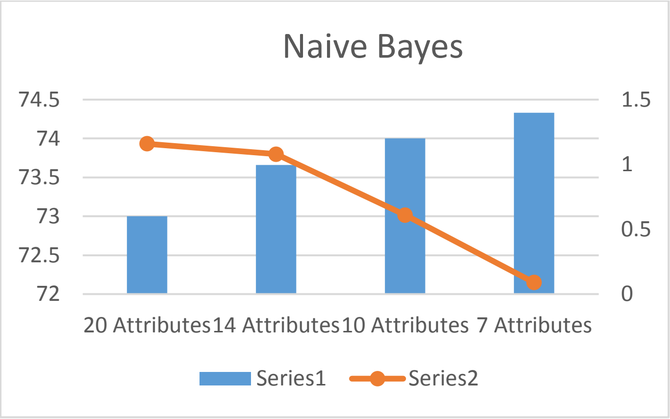
Result of implemented algorithms, AdaBoost and Logistic regression has highest accuracy 100% and J48 has lowest accuracy 69.96%. AdaBoost and Logistic regression has same accuracy but time duration was different. AdaBoost takes 2.80 seconds and Logistic

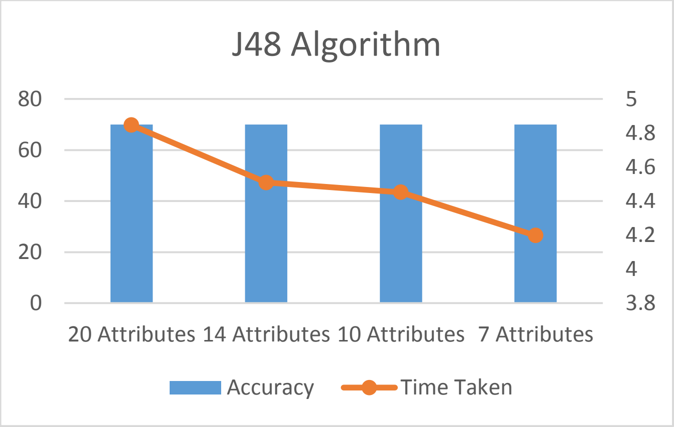
Regression takes 3.81seconds. So that, from our Fraud detection flow and Credit card dataset AdaBoost algorithm is more preferable from others.

## ANALYSIS & CONCLUSION

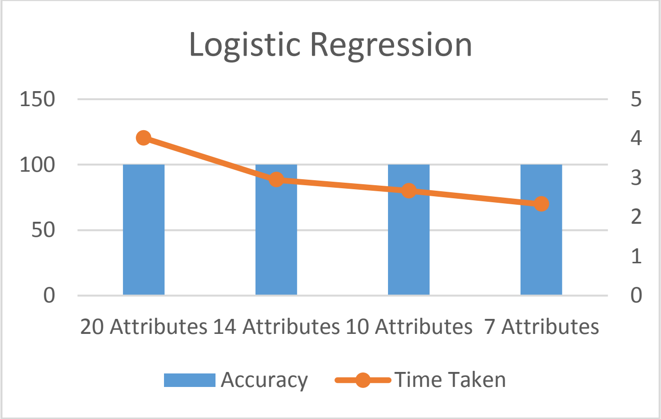
Implemented algorithm with different number of attributes such as 20, 14, 10 and 7 attributes.

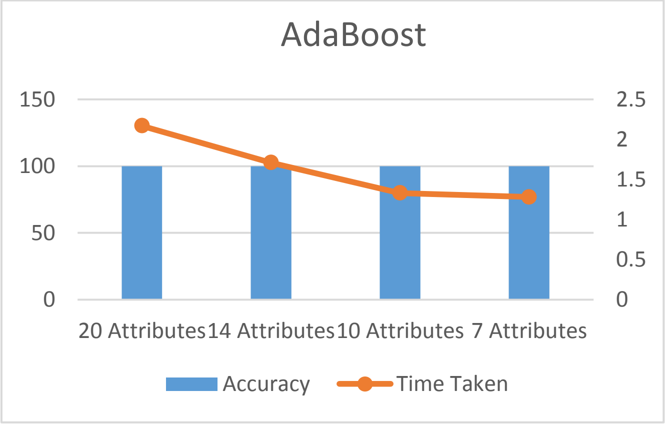
This Analysis by Algorithm. After Analysis conclude that AdaBoost algorithm is better than other algorithms.


Naive Bayes
74.5
74
73.5
73
72.5
72
1.5
1
0.5
0
20 Attributes
14 Attributes
10 Attributes
7 Attributes
Series1
Series2


J48 Algorithm
80
60
40
20
0
5
4.8
4.6
4.4
4.2
4
3.8
20 Attributes
14 Attributes
10 Attributes
7 Attributes
Accuracy
Time Taken

s


Logistic Regression
150
100
50
0
5
4
3
2
1
0
20 Attributes
14 Attributes
10 Attributes
7 Attributes
Accuracy
Time Taken


AdaBoost
150
100
50
0
2.5
2
1.5
1
0.5
0
20 Attributes
14 Attributes
10 Attributes
7 Attributes
Accuracy
Time Taken

## REFERENCES

[1] 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

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

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

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

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

[6] 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

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

[8] 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

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

[10] P.Sreenivas and Dr.C.V.Srikrishna, “An Analytical approach for Data Pre-processing”

[11] Mine Isik, M.hamdi, et.al, “Improving a credit card fraud detection system using genetic algorithm”, International Conference on Networking and Information Technology, pp 437 – 440, 2010

[12] Kuldeep Randhawa, Chu Kiong Loo, et.al, “Credit card fraud detection using AdaBoost and majority voting, IEEE Access, pp- 1 – 8, vol XX, 2017

[13] Dr. Sanjay Kumar Dwivedi and Bhupesh Rawat, “A Review Paper on Data pre-processing: A Critical Phase in Web Usage Mining Process”, International Conference on Green Computing and Internet of Things (ICGCIoT), pp 506-510, 2015

[14] Aman Srivastava, Mugdha Yadav, et.al, “Credit Card Fraud Detection at Merchant Side using Neural Networks”, International Conference on Computing for Sustainable Global Development (INDIACom), pp- 667- 670, 2016 IEEE

[15] Chun-Hua JU and Na Wang, “Research on Credit Card Fraud Detection Model Based on Similar Coefficient Sum”, First International Workshop on Database Technology and Applications, IEEE computer Society, pp 295 – 298, 2009

[16] Krishna Modi and Reshma Dayma, “Review On Fraud Detection Methods in Credit Card Transactions”, International Conference on Intelligent Computing and Control (I2C2'17), 2017

[17] S. Benson Edwin Raj and A. Annie Portia, “Analysis on Credit Card Fraud Detection Methods”, International Conference on Computer, Communication and Electrical Technology – ICCCET, pp 152 – 156, March - 2011

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