Credit Card Fraud Detection Using Machine Learning

M. Ummul Safa[1], R. M. Ganga[2]

[1,2]*B.E. Student, Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Erode, India*

***Abstract*: Credit card fraud events take place frequently and then result in huge financial losses. Data mining had played an imperative role in the detection of credit card fraud in online transactions. Credit card fraud detection, which is a data mining problem, becomes challenging due to two major reasons: first, the profiles of normal and fraudulent behaviour's change constantly and secondly, credit card fraud data sets are highly skewed. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection technique(s) used. This paper investigates the performance of naive Bayes, k-nearest neighbour and logistic regression on highly skewed credit card fraud data. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. Three techniques are applied on the raw and pre- processed data. The work is implemented in Python. The performance of the techniques is evaluated based on accuracy, time duration and balanced classification rate. The results shows of optimal accuracy for naive Bayes, logistic regression and k-nearest neighbour classifiers are 83.00%, 97.69% and 54.86% respectively. The comparative results show that logistic regression performs better than naïve Bayes and k-nearest neighbour techniques.**

***Keywords*: machine learning**

1. Introduction

Credit Card Fraud is one of the biggest threats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of modus operandi to commit fraud. In simple terms, Credit Card Fraud is defined as: When an individual uses another individuals credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used. Further, the individual using the card has no connection with the cardholder or issuer, and has no intention of either Contacting the owner of the card or making repayments for the purchases made.

Credit card frauds are committed in the following ways:

- An act of criminal deception (mislead with intent) by use of unauthorized account and/or personal information
- Illegal or unauthorized use of account for personal gain
- Misrepresentation of account information to obtain goods and/or services.

Credit card fraud is divided into two types:

- Offline fraud
- Online fraud.

Offline fraud is committed by using a stolen physical card at storefront or call center. In most cases, the institution issuing the card can lock it before it is used in a fraudulent manner. Online fraud is committed via web, phone shopping or cardholder not present. Only the cards details are needed, and a manual signature and card imprint are not required at the time of purchase.

2. Literature survey

1. You Dai, et. al: 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 summarize second type is better than the first type.
2. Suman Arora: 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. Summarize of this paper, SVM has the highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335.
3. Kosemani Temitayo Hafiz: 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. Problem definition and solution

A. Problem definition

Major problem is that online payment does not require physical card. Anyone who knows the details of the card can make fraud transactions. Card holder comes to know only after the fraud transaction is carried out.

B. Problem solution

Unfortunately, no mechanism is to track the information about the fraud transaction. Logistic Regression is the algorithm used to find out the fraud transaction. It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2 ...V28 are the principal components obtained with PCA; the only features which have not been transformed with PCA are 'Time' and 'Amount'. The feature 'Amount' is the transaction Amount, this feature can be used for example - dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

4. Machine learning and its algorithms

A. Machine learning

Machine learning is a collection of methods that can automatically identify patterns in data, and then use those patterns to predict future outcomes, or to perform other types of decision making below certain conditions. Machine learning introduces various algorithms, those enable machines to understand the current situations and on the basis of that machines can take appropriate decisions. Machine learning works independently and takes decision at its own. The main two types of machine learning are, supervised learning and unsupervised learning Supervised Learning: In supervised learning, the input and its corresponding output is already known. This is called supervised learning because it learns from training data set and creates model from it and when this model applies on new data set it gives predicted results. Decision Tree, naive Bayes etc. are the examples of supervised learning.

Unsupervised Learning: Unsupervised learning is where we have only input data and no corresponding output variable. The main job of unsupervised learning is to build up class labels automatically. The relationship between the data can be found using unsupervised learning algorithms to discover whether the data can characterize to form a group. This group is known as clusters. Unsupervised learning can be also described as cluster analysesǁ. K Means Clustering, KNN etc. are the examples of unsupervised learning.

B. Selected online dataset

In this project, we have used a Kaggle provided dataset of simulated mobile based payment transactions. We analyze this data by categorizing it with respect to different types of transactions it contains. We also perform PCA - Principal Component Analysis - to visualize the variability of data in two dimensional spaces. The datasets contain transactions made by credit cards in September 2013 by European cardholders. These dataset present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise."

C. Selected algorithm for implementing

On the Literature review many algorithms are applied on Fraud detection. On the survey bases Naïve Bayes, Logistic regression, and K – nearest neighbor 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 finds 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

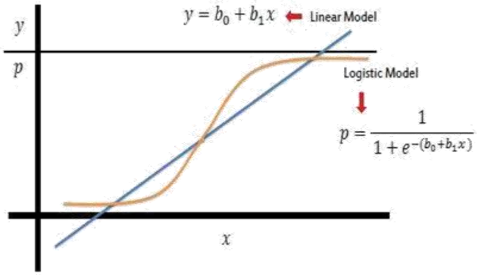

Fig. 1. Logistic regression

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.

1 / (1 + e^-value)

Logistic regression equation:

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b)

3) K-Nearest Neighbor Algorithm

The concept of nearest neighbor analysis has been used in several anomaly detection techniques. One of the best classifier algorithms that have been used in the credit card fraud detection is k- nearest neighbor algorithm that is a supervised learning algorithm where the result of new instance query is classified based on majority of K-Nearest Neighbor category. The performance of KNN algorithm is influenced by three main factors:

- The distance metric used to locate the nearest neighbors.
- The distance rule used to derive a classification from k-nearest neighbor.
- The number of neighbors used to classify the new sample.

5. Implementation

A. Block diagram

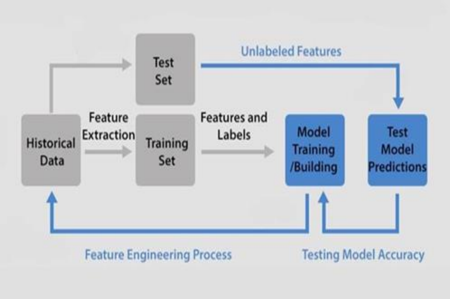

Fig. 2. Block diagram

B. Plotting the variables using subplots

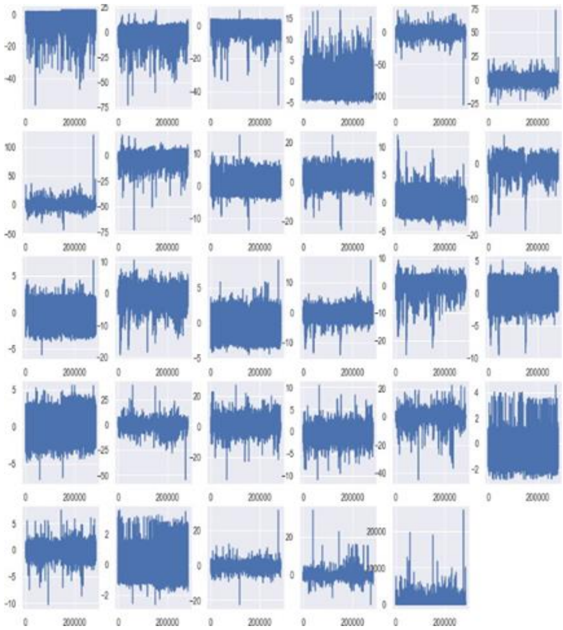

Fig. 3. Plotting of variables

C. Output graph using confusion matrix

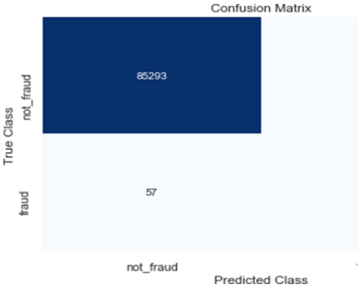

Fig. 4. Output graph

6. Results of implemented algorithms in python

Name	Naïve Bayes	Logistic Regression	K – nearest neighbour
Accuracy	83.00%	97.69%	54.86%
Time Duration	10.0	38.1	46.24
Method	Classification method	Machine Learning	Supervised Learning
Training: Testing	70 : 30	70 : 30	70 : 30
Inbuilt Packages	Gaussian NB	Logistic Regression	Decision Tree Classifier

7. Conclusion

After implementing algorithm, highest accuracy is given by Logistic Regression. The time duration is quite high in logistic regression, but in this case the accuracy is mainly considered for obtaining the results. The results show optimal accuracy for naïve Bayes, logistic regression and k- nearest neighbour classifiers are 83.00%, 97.69% and 54.86% respectively. Therefore, the comparative results show that logistic regression performs better than naïve Bayes and k-nearest neighbour techniques. So, Logistic regression technique can be used for credit card detection

References

[1] Erkin, Erkin et. al., "Privacy-preserving distributed clustering" in EURASIP Journal on Information Security, licensee Springer, 2013.

[2] Ge-Er Teng, Chang-Zheng He, Jin Xiao, Xiao-Yi Jiang, "Customer credit scoring based on HMM/GMDH hybrid model" in, London: Springer-Verlag, 2012.

[3] Ashphak Khan, Tejpal Singh, Amit Sinhal, "Implement Credit Card Fraudulent Detection System Using Observation Probabilistic in Hidden Markov Model", NUiCONE-2012, December. 2012.

[4] Mohammed Ibrahim Alowais, Lay-Ki Soon, "Credit Card Fraud Detection: Personalized or Aggregated Model", Third FTRA International Conference on Mobile Ubiquitous and Intelligent Computing IEEE2012, 2012.

[5] S. Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", International Conference on Computer Communication and Electrical Technology — ICCCET2011, 18th & 19th March.

[6] Divya Lyer, Arti Mohanpurkar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE, 2011.

[7] V. Bhusari, S. Patil, "Study of Hidden Markov Model in Credit Card Fraudulent Detection", International Journal of Computer Applications, vol. 20, no. 5, pp. 0975-8887, April. 2011.

[8] Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE Transactions On Dependable and Secure Computing, vol. 5, no. 1, January-March. 2018.