Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 385–395

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

Suraj Patil*, Varsha Nemade, PiyushKumar Soni

Dept of Computer Engineering, Mukesh Patel School of Technology Management and Engineering , NMIMS, Shirpur Campus, India

Abstract

The finance and banking is very important sector in our present day generation, where almost every human has to deal with bank either physically or online [10]. The productivity and profitability of both public and private sector has tremendously increased because of banking information system. Nowadays most of E-commerce application system transactions are done through credit card and online net banking. These systems are vulnerable with new attacks and techniques at alarming rate. Fraud detection in banking is one of the vital aspects nowadays as finance is major sector in our life. As data is increasing in terms of Peta Bytes (PB) and to improve the performance of analytical server in model building, we have interface analytical framework with Hadoop which can read data efficiently and give to analytical server for fraud prediction. In this paper we have discussed a Big data analytical framework to process large volume of data and implemented various machine learning algorithms for fraud detection and observed their performance on benchmark dataset to detect frauds on real time basis there by giving low risk and high customer satisfaction.

Keywords: Hadoop; SAS; PB; CCFD; Logistic Regression; Decision Tree.

1. Introduction

The credit card and online net banking fraud is an international problem in banking domain. In 2014, global fraud accounted for loss of $16.31 billion and this figure is increasing day by day as fraudster are developing new analytical techniques to alter the normal working behavior of credit card fraud detection system(CCFD) [1]. Also as data is increasing in terms of volume, velocity and variability the performance of machine learning algorithm is becoming bottleneck. The main challenge for today's CCFD system is how to improve fraud detection accuracy with growing number of transactions done by user per second. The increase in number of users and online transactions has brought heavy workloads to these systems.

*Correspondance author. Tel:+919765720693, Email:suraj.patil@nmims.edu

10.1016/j.procs.2018.05.199

The size of day to day transaction and past historical transactions has increased to several PB in recent years [8][9]. In this case, processing of data for model building, model training and predictive modelling on incoming transaction with minimum delay is very hard to achieve for current CCFD system. So as to overcome this problem we proposed a solution of interfacing SAS with Hadoop framework and building self-adaptive analytical framework model on top of it for fraud detection. We have also compared the performance of various algorithm of fraud detection and tune the analytical server with most optimal model for fraud detection.

2. Related Work

In the development of modern technology financial frauds are increasing significantly and hence fraud detection is very important area. Fraud detection is very important to save the financial losses for the banks as they issue credit cards to customer. Without knowledge of card holder use of the card information is a credit card fraud. There are two types of fraud detection approaches: misuse detection and anomaly detection [1]. In misuse detection, the system trains on normal and fake transactions, it will identify the known frauds. In anomaly detection system normal transactions are used for training so it has potential to identify novel frauds. Leila et.al had proposed a method which used anomaly detection techniques which extracts the inherent pattern of aggregated daily purchases of cardholder from a credit card time series and uses this pattern for earlier fraud detection [2]. Sanchez ´ et al. [3] described the method for fraud detection from transactional databases using fuzzy association rule mining in extracting knowledge .This method is very effective and optimizes the execution time and reduces the excessive generation of rules. Panigrahi et al. [4] proposed the new approach using rule-based filtering, Dempster-Shafer theory, transaction history database and Bayesian learner. Initial belief about each incoming transaction can be computed by coming multiple evidences from the rule based component by using Dempster's rule. S. Maes et al. [5] described the fraud detection system by using BBN and ANN. They found that BBN gives better result than ANN. In BNN training period is short as compared to ANN. Chen et al. [6][11] suggested the new method for fraud detection in which QRT data is collected by using online questionnaire. A support vector machine (SVM) is used to train the data and develop the QRT models which are used to predict about new transactions that whether it is fraud or not.

The problem of protection of a passive RFID network from intruders was investigated by Tsiropoulou et al.[12]. They have suggested association of a utility function with RFID tags with two goals, one is proper demodulation of signal by the reader and second is to categorize them as normal or intruder tags. In [13] Dai et al. have analyzed the advantages of using block chain in cyber security and summarized the current research and application. In [14] Berkowsky and Hayajneh have discussed about the security issues with certificate authority model. They have discussed about problems both in its design and implementation.

3. Methodology

The proposed system is used to detect the frauds on real time basis by analyzing incoming transactions. The system design consists of two components for fraud detection

3.1 Designing a framework for data pre-processing

This component is responsible for processing of big data efficiently and gives it to the analytical server for predictive modelling as shown in fig1. The system design mainly consists of Hadoop network which stores data in HDFS which is coming from different sources. The data from Hadoop is read by SAS using data step and proc hadoop step and converted into raw data file. The fields in a raw data file are separated by some delimiter. The raw data file is given to analytical model for building data model. These makes system highly scalable and helps to build strong self-learning analytical model on real time basis.

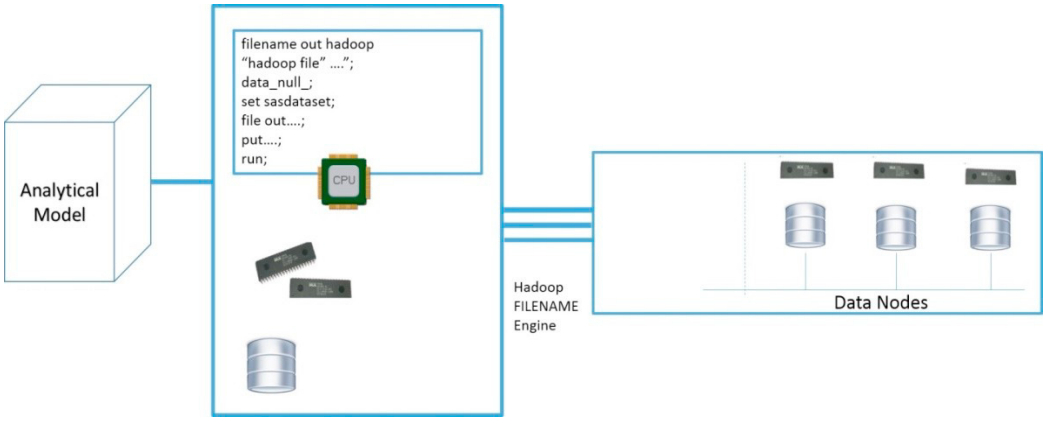

Fig. 1. (a) Analytical framework for data pre-processing

3.2 Designing analytical model for fraud prediction

The analytical model is used to check whether incoming transaction is legimate transaction or not. The logistic regression and decision tree machine learning models are implemented for fraud detection. The model is built on credit card banking data set. Here we are using two models for fraud detection classification.

3.2.1 Logistic regression

We are using Logistic Regression for the classification of fraud detection. Logistic regression is a type of probabilistic statistical classification model and uses logistic curve for fraud detection. The formula for univariate logistic curve is

$$p = \frac{e^{(c_0 + c_1 x_1)}}{1 + e^{(c_0 + c_1 x_1)}} \quad (1)$$

The logistic curve gives a value between 0 and 1, so it can be interpreted as the probability of class membership. To perform the regression, the logarithmic function can be applied to logistic function as given shown below.

$$loge\left(\frac{p}{1-p}\right) \tag{2}$$

Here *P* is the probability of tuple being in class and 1-*P* is the probability of tuple not in class. However the model chooses values of coefficient $_{C0}$ and $_{C1}$ that maximizes the probability of incoming transaction.

3.2 .2 Decision Tree

It is a type of supervised learning algorithm. The decision tree uses ID3 technique for building decision tree by considering entropy of dataset. The entropy is used to measure the amount of uncertainty in set of data. The splitting criteria in design of decision tree are decided by calculating entropy of each attribute. The entropy of the different state can be calculated by equation as

$$H(p1, p2 \ldots\ldots p_s) = \sum_{i=1}^{s} \left(pi log\left(\frac{1}{pi}\right)\right) \tag{3}$$

Where $P_1, P_2, \ldots Ps$ are the probabilities of the attributes of dataset.

The entropy of each attribute in dataset is calculated and gain is found by subtracting entropy of entire dataset with entropy of splitting attribute. The attribute which as highest gain is selected as root node and accordingly decision tree is designed. The ID3 calculates the gain of a particular split by the following equation as

$$\text{Gain}(D, S) = \text{H(D)} - \sum_{i=1}^{S} p(Di) H(Di) \tag{4}$$

3.2 .2 Random Forest Decision Tree

The random forest tree is supervised learning machine learning technique used to solve regression as well as classification problem. To perform prediction of fraudulent transaction the random forest algorithm uses the below pseudocode.

1. Extract the **test features of incoming transaction** and use the rules of each randomly created decision tree to predict the result and stores the predicted result (target)
2. Calculate the **votes** for each predicted target output.
3. Evaluates the **high voted** predicted target from different decision tree as the **final prediction output.**

4. Experiment Result

To build analytical model, German credit card fraud dataset is taken consisting of 20 attributes out of which 7 are numerical attributes and 13 are categorical attributes and almost 1000 transactions [7]. After doing preliminary data exploratory analysis on dataset for all attributes, the Table1 shows sample distribution proportion of fraud transaction and genuine transaction in dataset.

Table1. Exploratory Analysis of Credit Card Dataset

Attribute: Status Of Checking Account	Fraud	Genuine	Total
'<0'	135	139	274
'0<=X<200	105	164	269
'>=200'	14	49	63
'No checking'	46	348	394
Grand Total	300	700	1000
Attribute: Credit History			
'all paid'	28	21	49
'critical/other existing credit'	50	243	293
'delayed previously'	28	60	88
'existing paid'	169	361	530
'no credit/all paid'	25	15	40
Grand Total	300	700	1000
Attribute: Property			
'Car'	102	230	332
'Life Insurance'	71	161	232
'No known Property'	67	87	154
'Real estate'	60	222	282
Grand Total	300	700	1000

4.1 Logistic Regression Analytical Model

For Logistic Regression Model building the input dataset is divided into trainData and testData. Below is code of model building.

```
Model <- glm (Fraud ~., data = trainData, family = binomial)
Predict <- predict (model, type = 'response', data=testData)
```

In this glm, logistic analytical model is used for fraud prediction. The fraud is responses variable and all other are predictors. Once model gets trained it is tested with the test data with threshold cut-off of 0.5 for prediction. The model is tuned by choosing most significant variables as shown below

```
Function: model<-glm (Fraud~ Status.of.existing.checking.account
                      +Duration.in.months
                      +Savings.account.bonds
                      +Present.employment.since
```

```
                +Other.installment.plans
                , data = trainData, family = binomial)
Summary (model)
Optcutoff<-OptimalCutoff (trainData$Fraud, predictTest)
table (traindata$Fraud,Predict>Optcutoff)
```

To improve the performance of the model optimal cut-off is decided with the help of OptimalCutoff function. The optimal cut-off 0.18 is used by the above model with most significant variable and it gives best performance compared to threshold cut-off 0.5 with logistic model with all variables as shown in ROC curve Fig.2.

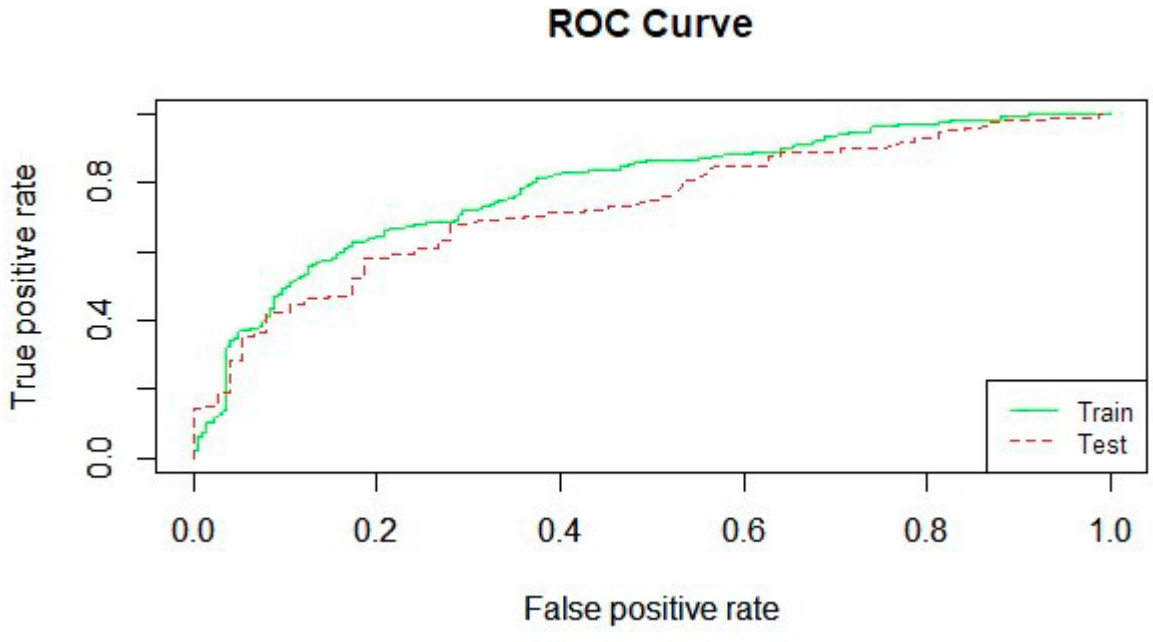

Fig. 2. (b) ROC curve for Logistic Regression.

4.2 Decision Tree Analytical model

We have used ID3 algorithm for designing of decision tree which can classify whether incoming transaction is fraud or not. The decision tree is generated by running the package rpart which takes Fraud as response variable and other variables are predictors that classifies incoming transaction to class good (not fraud) and bad(fraud).

Function: rpart(Fraud ~ ., method= "class", data= trainData**)**

To improve the performance of decision tree, most significant variable are taken from trained model and the model is tuned with those most significant variables and decision tree is design to predict class of transaction as good or bad as shown in Fig3.

```
Function: rtree_fit <- rpart(Fraud ~ Status.of.existing.checking.account
                        +Duration.in.months
                        +Savings.account.bonds
```

```
                    +Purpose
          , trainData)

predTrain <- predict(rtree_fit, newdata=trainData, type="class")
table(pred, trainData$Fraud)
pred <- predict(rtree_fit, newdata=testData, type="class")
pred.prob <- predict(rtree_fit, newdata=testData, type="prob")
table(pred, testData$Fraud)
```

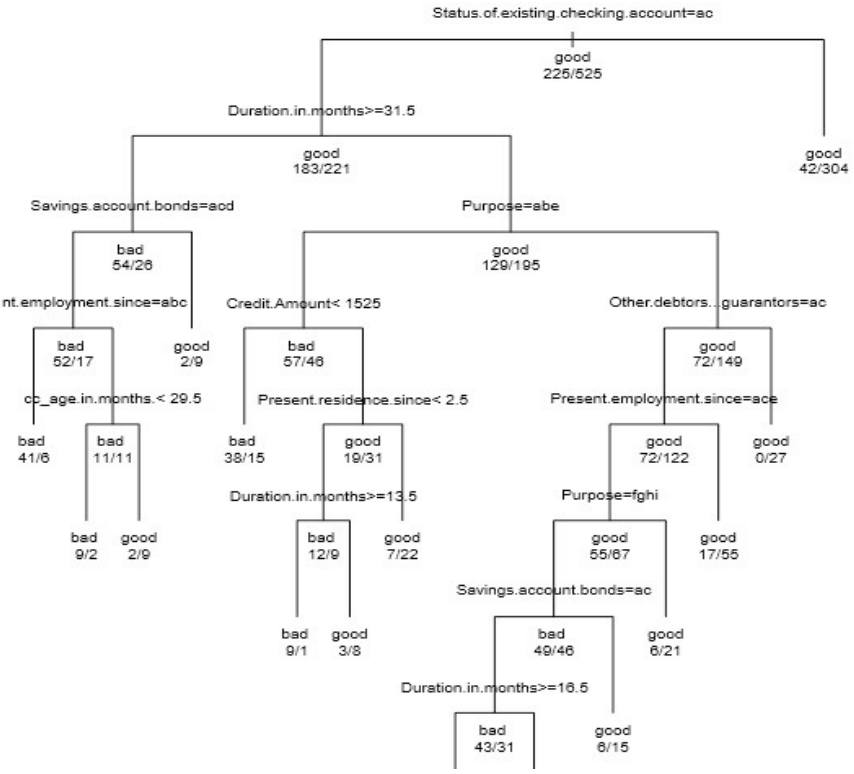

Fig3. (c). Decision Tree

4.2 Random Forest Decision Tree Analytical model

The code snippet of Random Forest Tree designed is given below. The model is build with most significant variables based on feature selection. The model is tested with testing data and confusion matrix is used to evaluate the model.

```
randomModel <-  randomForest(Fraud~Status.of.existing.checking.account
                    +Credit.History+Savings.account.bonds
                    +Present.employment.since
                    +Other.installment.plans,  data=trainData

ntrees=1000, cutoff = c(0.7,1-0.7))
```

```
pred <- predict(randomModel, newdata=trainData, type="class")

pred.prob <- predict(randomModel, newdata=trainData, type="prob")
table(pred, trainData$Fraud)
pred <- predict(randomModel, newdata=testData, type="class")
pred.prob <- predict(randomModel, newdata=testData, type="prob")
table(pred, testData$Fraud)
```

5. Model Evaluation

The three models are run on credit card fraud dataset and accuracy of analytical model is evaluated with help of confusion matrix. The confusion matrix tells how the tuples in training and testing models are correctly classified. The models are evaluated based on parameters such as precession, recall, accuracy and F1 score. The detail description of each parameter is shown in confusion matrix. The precession is the proportion of the predicted non fraudulent transactions that are actually predicted as good and the recall is the proportion of actual non fraudulent transaction that are correctly recall as good. The Table 2 and 3 shows confusion matrix for logistic regression and its accuracy of predicting incoming transaction is fraud or not comes to be 70% were as for the decision tree accuracy of predicting incoming transaction is fraud or not is 72% which is better as compare to logistic regression. If we consider other statistical parameters of model such as precision, prevalence, likelihood ration ,false omission rate etc. as shown in Table 2 and 3, we find decision tree model works better as compared to the logistic regression model because decision tree model bisects the dataset points into smaller and smaller region according to the splitting of attribute based on entropy, whereas logistic regression fits a single line to divide the dataset points exactly into two regions based on threshold probability. If data points are nonlinear then single line can be limiting to the logistic regression as the outlier points are not handle effectively, in that case decision tree performs better. In order to improve overall performance of fraud detection, we have built random forest tree data model. The random forest model was build and tested on same data and it out performs better as compared to logistic regression and decision tree in terms accuracy, precision and recall parameters. The confusion matrix Table 4 of random forest shows better result as compared to confusion matrix of logistic regression and decision tree.

Table 2. Confusion Matrix for Logistic Regression

	Actual Condition	Accuracy		
	Condition +ve	Condition – ve	**72%**	
Predicted Condition +ve	**154**	**48**	Precision	False discovery rate
			76%	**24%**
Predicted Condition – ve	**21**	**27**	False omission rate	Negative predictive value
			44%	**56%**

Prevalence	Sensitivity, Recall TPR	Fallout FPR	+ve likelihood ratio	- ve likelihood ratio
70%	**88%**	**64%**	**1.38**	**0.33**
	Miss Rate FNR	Specificity TNR	F1 Score	
	12%	**36%**	**70**	

Table 3. Confusion Matrix for Decesion Tree

	Actual Condition		Accuracy	
	Condition +ve	Condition – ve	**72%**	
Predicted Condition +ve	**156**	**19**	Precision	False discovery rate
			89%	**11%**
Predicted Condition – ve	**51**	**24**	False omission rate	Negative predictive value
			68%	**32%**
Prevalence	Sensitivity, Recall TPR	Fallout FPR	+ve likelihood ratio	- ve likelihood ratio
83%	**75%**	**44%**	**1.71**	**0.44**
	Miss Rate FNR	Specificity TNR	F1 Score	
	25%	**56%**	**71**	

Table 4. Confusion Matrix for Random Forest Decesion Tree

	Actual Condition		Accuracy	
	Condition +ve	Condition – ve	**76%**	
Predicted Condition +ve	**163**	**12**	Precision	False discovery rate
			93%	**7%**
Predicted Condition - ve	**48**	**27**	False omission rate	Negative predictive value
			64%	**36%**
Prevalence	Sensitivity, Recall TPR	Fallout FPR	+ve likelihood ratio	- ve likelihood ratio
84%	**77%**	**31%**	**2.51**	**0.33**
	Miss Rate FNR	Specificity TNR	F1 Score	
	23%	**69%**	**61**	

Where

Precision - % of true predictions are actually true?

FDR - % of true predictions are actually false?

FOR - % false predictions are actually true?

NPV - % false predictions are actually false?

LR-ratio of TPE to FPR

Sensitivity - % of actual positive recalled correctly?

Miss Rate - % of actual positive predicted wrongly?

Fallout - % of actual negatives predicted wrongly?

Specificity -% of actual negatives recalled correctly?

F1 Score mean of precision & recall

6. Conclusion

The credit card fraud detection is becoming important topic of research, as different types of attacks are increasing at an alarming rate. In this paper we have proposed a robust framework to process large volume of data, the functionality of framework can be extended to extract real time data from different desperate sources. The extracted data is then used to build strong analytical model. To improve the analytical accuracy of fraud prediction, we have implemented three different analytical techniques. These analytical models are run on credit card dataset and accuracy of analytical model is evaluated with help of confusion matrix. Among the three models, random forest decision tree performs best in terms of accuracy, precision and recall. The only problem with random forest is overfitting of tree in memory as data increases. The future scope of this work is to remove overfitting problem of decision tree and to detect real time fraud transaction for high streaming real time data.

Acknowledgement

Suraj P Patil acknowledged to SVKM's NMIMS Deemed-to-be University and thankful for financial assistance.

References

[1] A. Kundu, S. Sural, and A. Majumdar,(2006) "Two-stage credit card fraud detection using sequence alignment," Information Systems Security, Springer Berlin , Heidelberg, 260-275.
[2] Seyedhossein, Leila, and Mahmoud Reza Hashemi. (2010)"Mining information from credit card time series for timelier fraud detection." Telecommunications (IST), 5th International Symposium on. IEEE,
[3] Sánchez, Daniel, et al. (2009),"Association rules applied to credit card fraud detection." Expert systems with applications 36(2), 3630-3640.
[4]Panigrahi, Suvasini, et al. (2009) "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning." Information Fusion, 10(4), 354-363.

[5]Maes, Sam, et al. (2002)"Credit card fraud detection using Bayesian and neural networks." Proceedings of the 1st international naiso congress on neuro fuzzy technologies.
[6] Chen, Rong-Chang, et al. (2004): "Detecting credit card fraud by using questionnaire-responded transaction model based on support vector machines." Intelligent Data Engineering and Automated Learning–IDEAL 800-806.
[7]http://weka.8497.n7.nabble.com/file/n23121/credit_fruad.arff
[8] You Dai, Jin Yan, et al. (2016) “Online Cedit Card Fraud Detection: A Hybrid Framework with Big Data Technologies”. IEEE Trust Com-BigSE-ISP
[9] Philip K Chaan (2016), “Distributed Data Mining in Credit Card Fraud Detection”, Florida Institute of Technology.
[10] S.N. John, Kennedy O, C. Anele, F. Olajide, Chinyere Grace Kennedy, (2016) “Fraud Detection in the Banking Sector Using Data Mining Techniques Algorithm”, International Conference on Computational Science and Computational Intelligence
[11] Lu Q, Ju C (2011) “Research on credit card fraud detection model based on class weighted support vector machine”, Journal Convergence Information Technology l 6, 62–68.