Credit card Fraud Detection using Predictive Modeling: a Review

Varre.Perantalu[1], K. BhargavKiran[2]
[1]*PG Scholar, CSE, Vishnu Institute of Technology, Bhimavaram, A.P, India*
[2]*Assistant Professor, CSE, Vishnu Institute of Technology, Bhimavaram, A.P, India*

***Abstract*– In this paper author proposed that fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions. This paper presents detection of frauds in credit card transactions, using data mining techniques of Predictive modeling, logistic Regression, and Decision Tree. The data set contains credit card transactions in September 2013 by European cardholders. This data set present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The data set is highly unbalanced, the positive class(frauds) Account for 0.172% of all transactions.**

***Index Terms*— Credit card fraud detection, Data *Mining*, Predictive modeling, Logistic Regression, Decision Tree.**

I.INTRODUCTION

Banks collect a lot of historical records corresponding to millions of customer's transactions. They are credit card and debit card operations, but unfortunately, only a small portion, if any, is open access. Fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions [1]. The proposed method consists of the Predictive modeling and Logistic Regression. Now a day's bank transactions as well as credit card frauds increased. One of the most target frauds are credit card fraud, the fraud can occur any type of credit products, such products are retail, home loan and personal loan. During the last few decades as technology has changed, dramatically the face of fraud also changed. To detect credit card fraud, data mining techniques- Predictive modeling and Logistic Regression are used. In prediction model to predict the continuous valued functions. Credit card of CSV files will be analyzed to predict the outcome.

In this paper, we propose to detect credit card transaction using available data set and data mining techniques of predictive modeling, Decision tree, and Logistic Regression. Predictive modeling splits the data into two partitions 70% of testing and 30%of training check output class distribution to predict the outcome. The decision tree to get the result as a tree with root node describes the best predictor in the data, the combination of two or more branches is denoted by decision node (non leaf nodes) and each branch represents a value for the attribute which is tested. The leaf node may be 1 in the case of fraud and 0 otherwise. Logistic regression or logistic model is a regression model, where the dependent variable is categorical of a linear generalized model.The rest of the paper is organized as explained. Section II describes fraud detection methods. Section III explains Dataset description for credit card transaction. Section IV consists of experimental results of fraud detection methods, and finally, the conclusion of this work included in section V.

II. FRAUD DETECTION METHODS

Predictive modeling

Predictive modeling is used to analyze the data and predict the outcome. Predictive modeling used to predict the unknown event which may occur in the future. In this process, we are going to create, test and validate the model. There are different methods in predictive modeling. They are learning, artificial intelligence and statistics. Once we create a model, we can use many times, to determine the probability of outcomes. So predict model is reusable. Historical data is used to train an algorithm. The predictive modeling process is an iterative process and often involves training the model, using multiple models on the same dataset.

The Process of Predictive Modeling

- Creating the model:
- To create a model to run one or more algorithms on the data set.
- Testing a model:
  The testing is done on past data to see how the best model predicts.
- Validating a model:
  Using visualization tools to validate the model.
- Evaluating a model:
  Evaluating the best fit model from the models used and choosing the model right fitted for the data.

Decision Trees

Decision trees are used to choose between several courses of action. It provides effective structure to investigate the possible outcomes. Decision trees use tree structure to build classification or regression model. A decision tree is a flowchart like tree structure, where non leaf node denotes a test on attribute. In the results, the decision tree will have a decision node and leaf nodes. A decision node is a combination of two or more branches; each branch represents a value for the attribute which is tested. The leaf node holds a class label; the top most nodes in the decision tree are called as root node. Which corresponds to the best predictor in the data? Decision trees can be used to analyses the categorical data and numerical data. One of the algorithm is used to build a de is decision tree is ID3 which is developed by J. Ross Quinlan. This algorithm uses top down approach and greedy search. The top down approach is recursive divide-and –conquer method. Backtracking is not used in this algorithm.

The learning of decision trees from training tuples using ID3 and CART (Classification and Regression Trees) algorithms were invented independently of one another around the same time. The ID3 and CART algorithms are used to generate decision tree induction. These algorithms also follow top down approach in recursive manner. Decision tree is built based on training tuples are recursively partitioned into smaller subsets.

Decision tree Induction Algorithm:(the algorithm used from this paper [3])

Generating decision tree from data partition D, of training tuples.

Algorithm: Generate decision tree

Input: Data partition, D, is the set of training tuples and their associated class labels.

Attribute list, the set of candidate attributes.

Attribute selection method, procedure to determine the best partitions that the tuples into individual classes of splitting criterion. This criterion consists of splitting attribute or splitting subset.

Output: a decision tree.

Method:

i. create a node N;
ii. if tuples in D are belongs to same class, C, then
iii. return N as a leaf node and labeled with the class C;
iv. if attribute list is empty then
v. return N as a leaf node with majority of class in D;
vi. apply attribute selection method (D, attribute list) to determine best splitting criterion;
vii. label node *N* with *splitting criterion;*
viii. *if splitting attribute is discrete value and multiway splits allowed, then*
ix. *attribute list =splitting attribute*
x. for each outcome *j* of *splitting criterion*
xi. *let Dj be the set of data tuples in D satisfying outcome j;*
xii. if *Dj* is empty then
xiii. attach a leaf labeled with the majority class in *D* to node *N*;
xiv. **else** attach the node returned by Generate decision tree(*Dj , attribute list*) to node *N*; end for;
xv. return N;

Attribute Selection Measures

Attribute Selection Measures is splitting the data into training tuples of classes. They decide how the tuples at a given node are split such that are also called as splitting rules. For each attribute, Attribute Selection measures provides ranking to describing the given training tuples. The best score for measure of particular attribute choose as the splitting attributes of training tuples.

The Attribute Selection Measure also describes information gain, gain ratio and Gini index, notations used in that are let D be the class labeled training tuples of data partitioned, attribute has m distinct values defining m distinct classes, C_i (for i= 1, ..., m). Let C_i, D be the set of tuples of class C_i in D.

Information Gain

Information Gain is sub part of Attribute selection measure they are used in ID3. Which attribute has the highest information gain is chosen as splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions, the expected number of tests needed to classify given tuple that simply a tree is found.

The expected information needed to classify a tuple in *D* is given by

$$\text{Info (D)} = -\sum_{i=1}^{m} pi \log_2 (\text{pi}) \quad \ldots\ldots\ldots (1)$$

In above formula 1, *pi* is the nonzero probability that an arbitrary tuple in *D* belongs to class c_i, and Info (D) also known as Entropy of D.

Logistic Regression

Logistic regression is similar to linear regression but interpret curve is using natural logarithm of the "odd" of the target variable which is developed by statistician David cox in 1958.To predict the probability of an outcome which has two values either zero or one, yes or no and false or true. The prediction is based on the use of one or several predictors; logistic regression produces logistic curves, which are two values of zero and one.

Linear regression model is used to predict binary target variables. Binary targets variables either 0 or one. The linear regression equation

$$Y=\beta_0+ \beta_1 + \sum_i \quad \ldots\ldots\ldots\ldots\ldots (2)$$

In equation (2) the actual value of Y is binary variable, then the predicted Y can be less than zero or greater than one. Logistic Regression or logit model is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors. The Logistic Regression can be a binomial, ordinal or multinomial, ordinal Logistic Regression deals with dependent variables that are ordered. In multinomial Logistic Regression where the outcomes can have three or more possible types are not in order. The Logistic Regression is used to determine probability of an event occur over the probability of an event not occurred, and then predicted variable may be continuous or categorical.

The Logistic curve

The method of logistic regression is fitting a regression curve, y= f(x) when y consists of binary code (0, 1—failure, success) data. The response variable is a binary variable and x is numerical. In equation (3), the relationship between x and y is determined by logistic curve, to fit curve using logistic regression. The shape of logistic curve is an s_ shaped or sigmoid curve. A logistic curve starts with slow, linear growth, followed by exponential growth, which then slow again to a stable state.

A simple logistic function is defined by the formula

$$Y=e^{x}/_{1+} e^{x} = 1 / 1+ e^{-x} \qquad (3)$$

Logistic Regression is a classification method that return the probability of binary dependent variable may be predicted from the independent variables.

III. DATASET DESCRIPTION

The data set is highly unbalanced; it contains credit card transactions in September 2013 by European cardholders occur in two days. Where, we have 492 frauds out of 284,807 transactions. It contains numeric input variables result of PCA transformation. Features V1, V2, V3, and V4……., V28 are the principal components obtained with PCA. Due to confidentiality issues, we cannot provide original features. The other two features 'time' and 'amount' have not been transformed to PCA. The feature 'amount' is used as dependent variable of transaction amount, and feature 'time' contains seconds elapsed between each transaction and first transaction in the data set. Response variable, as the class variable, takes value 1 in case of fraud and 0 otherwise.

IV. EXPERIMENTAL RESULTS

Predictive Modeling

0	1
85295	148

Table1. Predictive modeling of output Class distribution

In Predictive modeling, to split the data set into 70% of testing and 30% of training then check output class distribution as shown above table 1, 148 frauds out of 284,807 transactions. To predict an outcome of frauds in credit card transaction that occurred in two days. We have to analyzed 85295 not frauds in credit card transaction.

Logistic Regression

	FALSE	TRUE
0	85279	16
1	69	79

Table 2. Logistic Regression for credit card fraud detection

Logistic Regression or logit model is a regression model, where the dependent variable is categorical. In case of logistic regression as shown above table 2, 79 frauds out of 85279 transactions. The categorical variable is class variable may be 0 or 1, to generalize a model with 1 that is fraud, otherwise 0.

Decision tree model

Decision tree uses tree structure to build regression model, the final result is a tree as shown Fig .1.with top most node is root node. Root node describes the best predictor in the data, and decision node is a combination of two or more branches, each branch represents a value for the attribute which is tested, leaf node holds class label. The leaf node may be 1 means fraud and 0 otherwise.

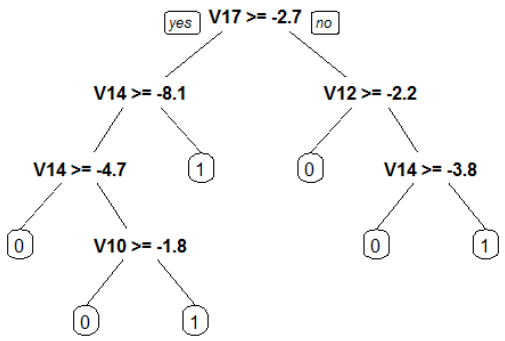

Fig .1. Decision tree for fraud detection

A decision tree is a tree structure. The decision tree can be linearized into decision rules, easy to understand. If the transaction observations V17 as a root node and V14, V12 are decision nodes, leaf node is V10 as shown below fig2. If V17 transaction greater than or equal to -2.7 and V14 greater than or equal to -8.1 then check V14, V10 may be 1 in a case of fraud and 0 otherwise.

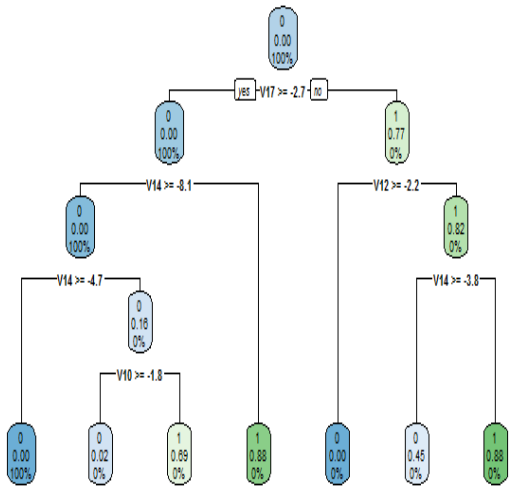

Fig 2. Plot for predicted tree

Rules in Decision Tree

A decision tree can easily be altered to a set of rules. These rules are map from the root node to the leaf nodes one by one. The decision tree can be linearized into decision rules,[4] where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form:

if condition1 *and* condition2 *and* condition3 *then* outcome.

Decision rules can be generated by construct association rules with the target variable on the right. They can also denote temporal or casual relatives.[5]

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	85275	20
1	44	104

Sub Rule	Variable		Value	Less	Greater
1	L1	V17	-2.703790986	<NA>	-2.703790986
2	L2	V14	-8.097680187	<NA>	-8.097680187
3	L3	V14	-4.661453694	<NA>	-4.661453694
4	L4	V10	-1.847304421	<NA>	-1.847304421
5	L5	V12	-2.188630 73	<NA>	-2.188630 73
6	L6	V14	-3.830215636	<NA>	-3.830215636
7	R1	V17	<NA>	-2.703790986	<NA>
8	R2	V14	<NA>	-8.097680187	<NA>
9	R3	V14	<NA>	-4.661453694	<NA>
10	R4	V10	<NA>	-1.847304421	<NA>
11	R5	V12	<NA>	-2.188630 73	<NA>
12	R6	V14	<NA>	-3.830215636	<NA>

Table 3. Rules for Decision Tree

A confusion matrix is a table that describes the performance and validating a model as shown above table 4. It explains the number of correct and incorrect predictions are summarized with count values and broken down by each class.

Accuracy : 0.9993
95% CI : (0.999, 0.9994)
No Information Rate : 0.9985
P-Value [Acc > NIR] : 2.098e-09

Kappa : 0.7643
Mcnemar's Test P-Value : 0.00404

Sensitivity : 0.9995
Specificity : 0.8387
Pos Pred Value : 0.9998
Neg Pred Value : 0.7027
Prevalence : 0.9985
Detection Rate : 0.9980
Detection Prevalence : 0.9983
Balanced Accuracy : 0.9191

'Positive' Class : 0

Table 4. Confusion Matrix for predicted data

Predicting between pairs produces categorical output:-1, 0, or 1 and counts how many times the predicted category mapped to the various true categories. It determines accuracy for predicted model, sensitivity and precision of predicted data. Confusion matrix is used in the field of machine learning and specifically the problem of statistical classification, is also known as an error matrix

V. CONCLUSION

This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and Logistic Regression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution. The prediction model predicts continuous valued functions. We have to detect 148 may be fraud and other are genuine. In decision tree generate a tree with root node, decision node and leaf nodes. The leaf node may be 1 becomes fraud and 0 otherwise. Logistic Regression is same as linear regression but interpret curve is different. To generalize the linear regression model, when dependant variable is categorical and analyzes relationship between multiple independent variables.

REFERENCES

1.Salazar,Addisson, et al. "Automatic credit card fraud detection based on non-linear signal processing." *Securit Technology (ICCST), 2012 IEEE International Carnahan Conference on*. IEEE, 2012.

2.Delamaire, Linda, H. A. H. Abdou, and John Pointon. "Credit card fraud and detection techniques: a review." *Banks and Bank systems* 4.2 (2009): 57-68.

3.Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

4.*Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies.* ***27*** *(3): 221. doi:10.1016/S0020-7373(87)80053-6.*

5. K. Karimi and H.J. Hamilton (2011), "Generation and Interpretation of Temporal Decision Rules", International Journal of Computer Information Systems and Industrial Management Applications, Volume 3.

6.Aggarwal, Charu C. "Outlier analysis." *Data mining*. Springer International Publishing, 2015.

7. Salazar, Addisson, Gonzalo Safont, and Luis Vergara. "Surrogate techniques for testing fraud detection algorithms in credit card operations." *Security Technology (ICCST), 2014 International Carnahan Conference on*. IEEE, 2014.

8.Ogwueleka, Francisca Nonyelum. "Data mining application in credit card fraud detection system." *Journal of Engineering Science and Technology* 6.3 (2011): 311-322.