CREDIT SCORING USING DATA MINING ALGORITHMS

Abbas Heiat, College of Business, Montana State University-Billings, 1500 University Drive

Billings, MT, 406-657-1627, aheiat@msubillings.edu

ABSTRACT

In this study I investigate the relation between account activity and variables that seems effective in providing information on the default risk of credit card holders. Using Artificial Neural Network, MLP ANN perform relatively well in predicting defaulters based on the dataset used in this study. Decision Tree algorithm used to create a set of decision-making rules based on MLP ANN output.

INTRODUCTION

According to The Balance research site Credit card default rates are commonly around 29.99%. However, this investment and finance research site have forecasted that the average credit card balance per consumer and default rate is going to rise in coming years [1]. Therefore, the information about consumers' account activity and behavior is of great importance for lending institutions which issue credit cards.

In this study I investigate the relation between account activity and variables that seems effective in providing information on the default risk of credit card holders. In the literature this approach to measuring the credit card default risk is called credit scoring. Credit scoring is defined as a method that helps lenders determine credit worthiness of the applicants with respect to the applicants' financial and demographic status such as, income, job, account balance, age and marital status, etc. [2].

Many quantitative techniques have been applied for credit scoring. Linear and quadratic discriminant analysis are among the most commonly used traditional statistical techniques. In this study I am going to use and determine the most efficient data mining algorithms for establishing the credit risk of credit card holders.

REVIEW OF LITERATURE

Many classical statistical methods like discriminant analysis and logistic regression have been used to develop models for prediction of credit risk and credit card default [3]. With the advent of machine learning, new classification and prediction algorithms were being used to predict credit risk [4]. According to Baesens et al, and Desai, Crook, & Overstreet predicting probability of default is an important problem. However, to forecast probability of default is a challenge facing practitioners and researchers, and it needs more future investigation [5][6].

There has been research work on credit card default datasets. Abbas Keramati et.al have done a literature survey on studies with credit card datasets using a variety of data mining algorithms [7]. However, he did not analyze the performance of any prediction or classification algorithms. Similarly, Adela Ioana et.al study application of clustering analysis on a credit card dataset without analyzing the effectiveness of this method [8]. The study of Simona Vasilica Oprea et.al evaluates few classification algorithms [9]. None of the previous research considers the efficiency of these algorithms with respect to feature selection. Feature selection is the process of determining the best input variables that have ability to make prediction of the target variables most efficient. In this study, based on feature selection, I am evaluating and determining the best efficient algorithms for predicting credit card default.

METHODOLOGY

Data Mining may be defined as the process of finding potentially useful patterns of information and relationships in data. As the quantity of clinical data has accumulated, domain experts using manual analysis have not kept pace and have lost the ability to become familiar with the data in each case as the number of cases increases. Improved data and information handling capabilities have contributed to the rapid development of new opportunities for knowledge discovery. Interdisciplinary research on knowledge discovery in databases has emerged in this decade. Data mining, as automated pattern recognition, is a set of methods applied to knowledge discovery that attempts to uncover patterns that are difficult to detect with traditional statistical methods. Patterns are evaluated for how well they hold on unseen cases. Databases, data warehouses, and data repositories are becoming ubiquitous, but the knowledge and skills required to capitalize on these collections of data are not yet widespread. In this research As a First step I used Auto-Classification tool in SPSS Modeler which applies 11 different algorithms shown in Figure 1. In Figure 2 he most efficient algorithms with highest accuracy rates are displayed based on current data set used for analysis.

Figure 1. Auto-Classification's Algorithms

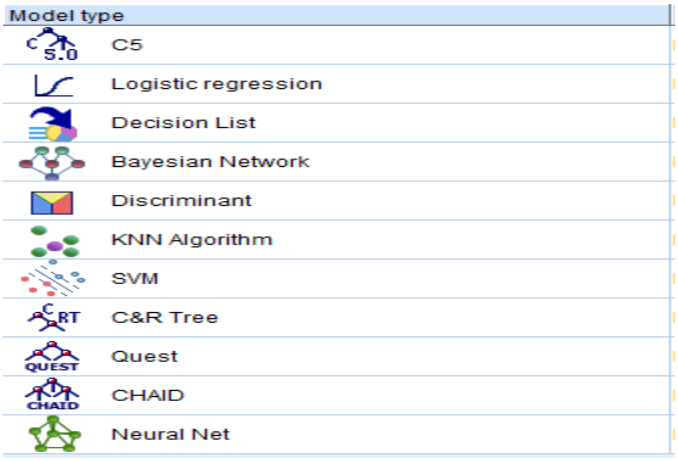
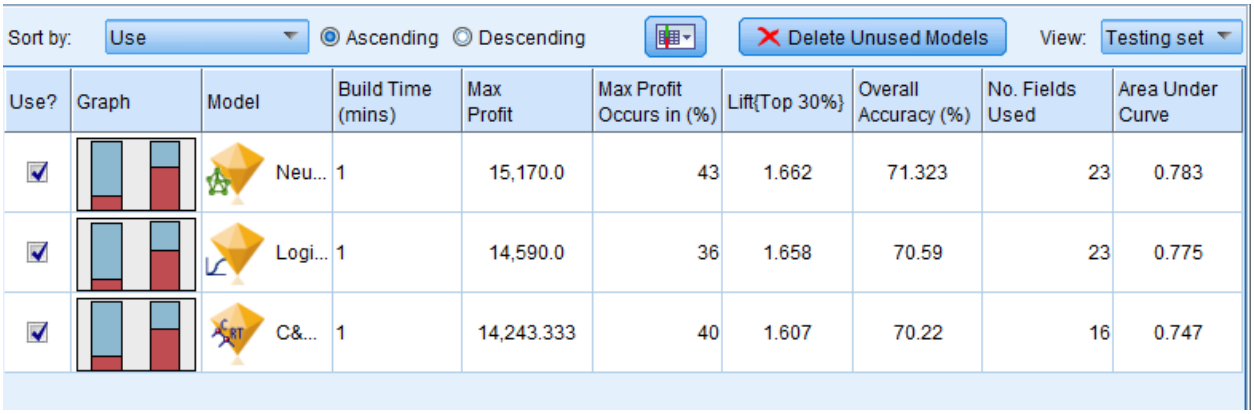


Figure 2. The most accurate algorithms.



Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		Neu...	1	15,170.0	43	1.662	71.323	23	0.783
☑		Logi...	1	14,590.0	36	1.658	70.59	23	0.775
☑		C&...	1	14,243.333	40	1.607	70.22	16	0.747

The following is a brief description of the Artificial Neural Network (ANN) algorithm suggested and displayed by Auto-Classification as the most accurate models as shown in Figure 2.

Artificial Neural Networks (ANN) - Artificial neural networks are defined as information processing systems inspired by the structure or architecture of the brain (Caudill & Butler, 1990). They are constructed from interconnecting processing elements, which are analogous to neurons. The two main techniques employed by neural networks are known as supervised learning and unsupervised learning. In unsupervised learning, the neural network requires no initial information regarding the correct classification of the data it is presented with. The neural network employing unsupervised learning can analyze a multi-dimensional data set to discover the natural clusters and sub-clusters that exist within that data. Neural networks using this technique can identify their own classification schemes based upon the structure of the data provided, thus reducing its dimensionality. Unsupervised pattern recognition is therefore sometimes called cluster analysis [10,11,12].

Supervised learning is essentially a two-stage process; firstly, training the neural network to recognize different classes of data by exposing it to a series of examples, and secondly, testing

how well it has learned from these examples by supplying it with a previously unseen set of data. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. It provides projections given new situations of interest and answers "what if" questions.

There are disadvantages in using ANN. No explanation of the results is given i.e. difficult for the user to interpret the results. They are slow to train due to their iterative nature. Empirical studies have shown that if the data provided does not contain useful information within the context of the focus of the investigation, then the use of neural networks cannot generate such information any more than traditional analysis techniques can. However, it may well be the case that the use of neural networks for data mining allows this conclusion to be reached more quickly than might ordinarily be the case.

Multilayer Perceptron and Radial Basis Function n**eural networks-**The Multilayer Perceptron (MLP) is one of the most widely implemented neural network topologies. In terms of mapping abilities, the MLP is believed to be capable of approximating arbitrary functions. This has been important in the study of nonlinear dynamics, and other function mapping problems. MLPs are normally trained with the back-propagation algorithm. Two important characteristics of the Multilayer Perceptron are:

1. Its smooth nonlinear Processing Elements (PEs). The logistic function and the hyperbolic tangent is the most widely used.
2. Their massive interconnectivity i.e. any element of a given layer feeds all the elements of the next layer.

The Multilayer Perceptron is trained with error correction learning, which means that the desired

response for the system must be known. Back propagation computes the sensitivity of a cost function with respect to each weight in the network and updates each weight proportional to the sensitivity [13].

The Radial Basis Function (RBF) network is a popular alternative to the MLP which can offer advantages over the MLP in some applications. An RBF network can be easier to train than an MLP network. The RBF network has a similar form to the MLP in that it is a multi-layer, feed-forward network. However, unlike the MLP, the hidden units in the RBF are different from the units in the input and output layers. They contain the Radial Basis Function, a statistical transformation based on a Gaussian distribution from which the neural network's name is derived. Like MLP neural networks, RBF networks are suited to applications such as pattern discrimination and classification, pattern recognition, interpolation, prediction and forecasting. In the hidden layer of an RBF, each hidden unit takes as its input all the outputs of the input layer *xi*. The hidden unit contains a basis function which has the parameters center and width. The center of the basis function is a vector of numbers, ***ci*,** of the same size as the inputs to the unit and there is normally a different center for each unit in the neural network. The first computation performed by the unit is to compute the radial distance, *d*, between the input vector xi and the center of the basis function, typically using Euclidean distance:

$$d = SQRT\left((x^1 - c^1)^2 + (x^2 - c^2)^2 + \ldots (x^n - c^n)^2\right)$$

The unit output, ***a*,** is then computed by applying the basis function *B* to this distance divided by the width *w*: $a = B(d/w)$

Decision Trees- Decision trees and rule induction are two most commonly used approaches to discovering logical patterns within medical data sets. Decision trees may be viewed as a simplistic approach to rule discovery because of the process used to discover patterns within data sets.

Decision tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially, you start with a training set in which the classification label (say, "productive" or "non-productive") is known (pre-classified) for each record. All the records in the training set are together in one big box. The algorithm then systematically tries breaking up the records into two parts, examining one variable at a time and splitting the records based on a dividing line in that variable (say, FP> 30 or FP<=30). The object is to attain as homogeneous set of labels (say, "productive" or "non-productive ") as possible in each partition. This splitting or partitioning is then applied to each of the new partitions. The process continues until no more useful splits can be found. The heart of the algorithm is the rule that determines the initial split rule [14].

The process starts with a training set consisting of pre-classified records. Pre-classified means that the target field, or dependent variable, has a known class or label: "productive" or "non-productive". The goal is to build a tree that distinguishes among the classes. For simplicity, assume that there are only two target classes and that each split is binary partitioning. The splitting criterion easily generalizes to multiple classes, and any multi-way partitioning can be achieved through repeated binary splits. To choose the best splitter at a node, the algorithm considers each input field in turn. Each field is sorted. Then, every possible split is tried and

considered, and the best split is the one which produces the largest decrease in diversity of the classification label within each partition. This is repeated for all fields, and the winner is chosen as the best splitter for that node. The process is continued at the next node and, in this manner, a full tree is generated.

THE DATASET

The dataset is obtained from UCI Machine Learning Repository credit card defaulter [15]. It is a newly published dataset obtained in 2015. The attribute details in the dataset according to UCI Repository are given as follow:
This dataset employs a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. It is tracked the past monthly payment records (from April to September 2005) as follows: X6 = the repayment status in September 2005; X7 = the repayment status in August 2005; . . .; X11 = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September

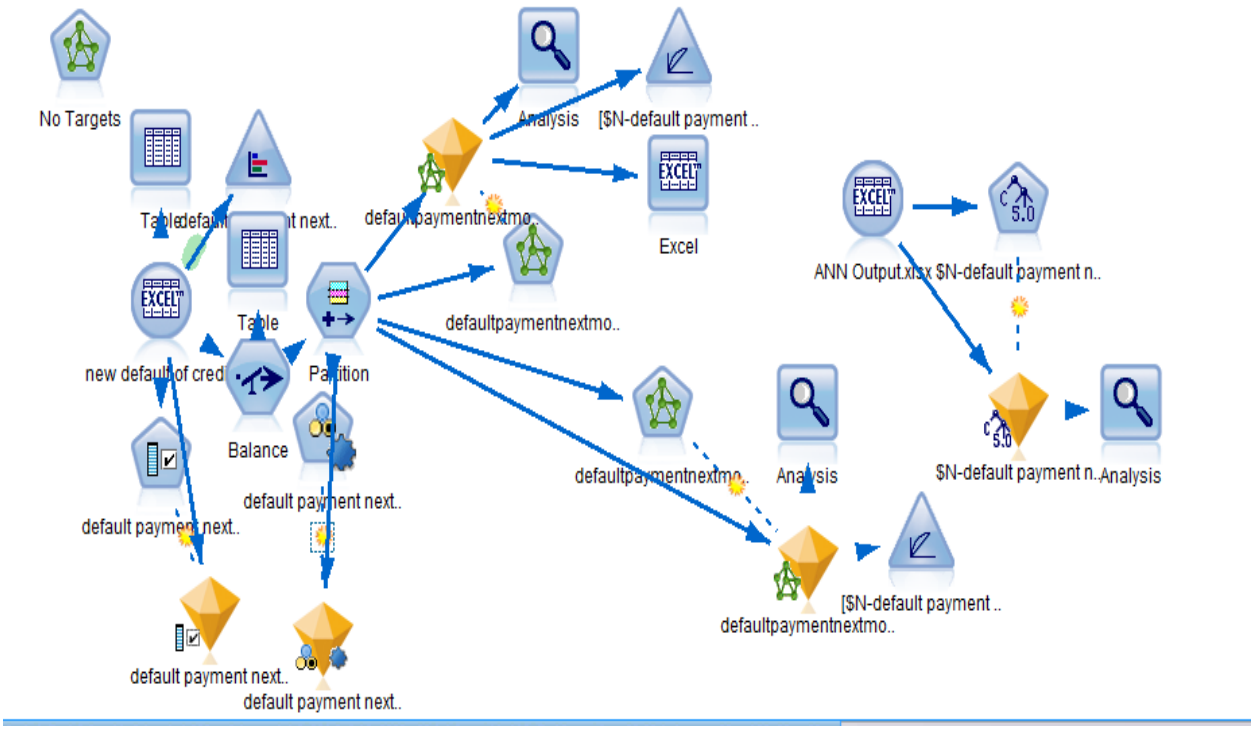
2005; X13 = amount of bill statement in August 2005; . . .; X17 = amount of bill statement in April 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005; . . .; X23 = amount paid in April 2005.

Amount to be paid next month

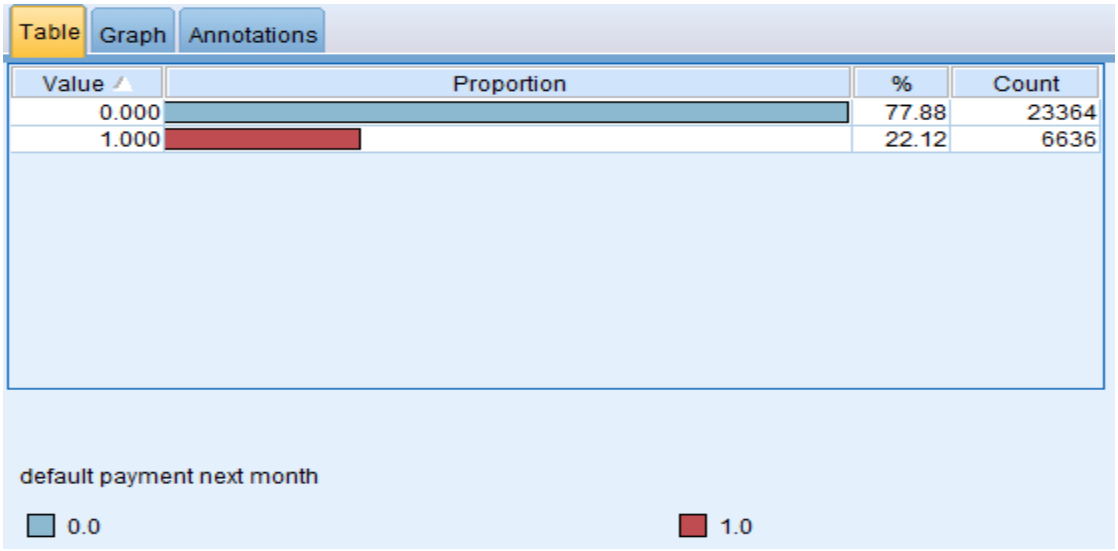
The dataset took payment data in October 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank. Among the total 25,000 observations, 5529 observations (22.12%) are the cardholders with default payment.

Figure 3. The credit card default model



Since data is unbalanced in terms of number of records for customers with credit card default and non-default customers, a balance node is used to make the distribution of the two categories almost equal. Figure 4 displays the original distribution.

Figure 4. Distribution of customer records with default and non-default



The data was partitioned into two groups, training including 70% of records and validation/testing including the 30% of the records.

Analysis Results for Artificial Neural Network.

In this study I am reporting the results of the most efficient algorithm according to Auto-Classification feature of IBM SPSS Modeler. Figures 5 and 6 shows the accuracy rates of RBF neural network and MLP neural network.

Figure 5. RBF Confusion Matrix

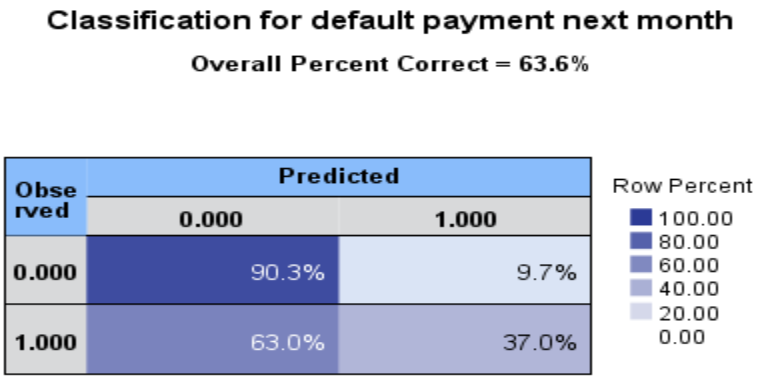
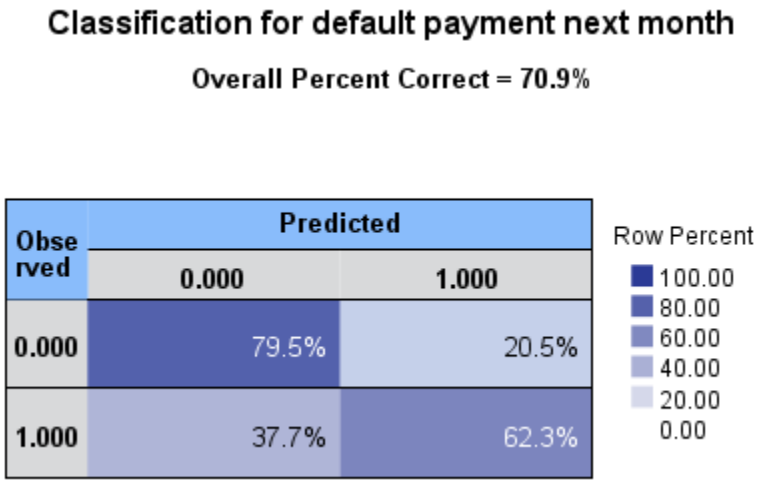


Figure 6. MLP Confusion Matrix



As you can see from Figure 5 RBF ANN is very efficient predicting non-defaulters (90.3%). However, it has a poor performance predicting the defaulters (37%). Figure 6 shows MLP ANN is much better than RBF ANN in predicting the defaulters (62.3%). However, it is not as good as RBF in predicting non-defaulters (79.5%). In general, we are more interested in identifying defaulters, and overall accuracy of MLP ANN (70.9%) is higher than RBFANN (63.6%).

Decision Tree algorithm-Since Artificial Neural Network does not explain how it arrives at its' prediction using input variables and does not create a set of rules for making decision about who is a defaulter and/or non-defaulter, I used a decision tree algorithm using output of MLP ANN to create set of decision-making rules. Figures 7 and 8 displays partial rules for predicting defaulters and non-defaulter.

Figure 7. Partial rules for defaulters

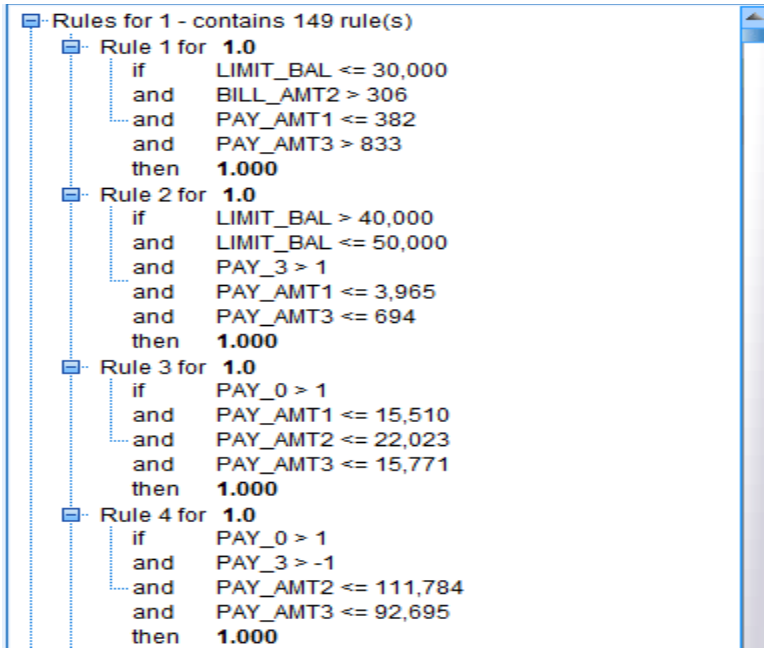
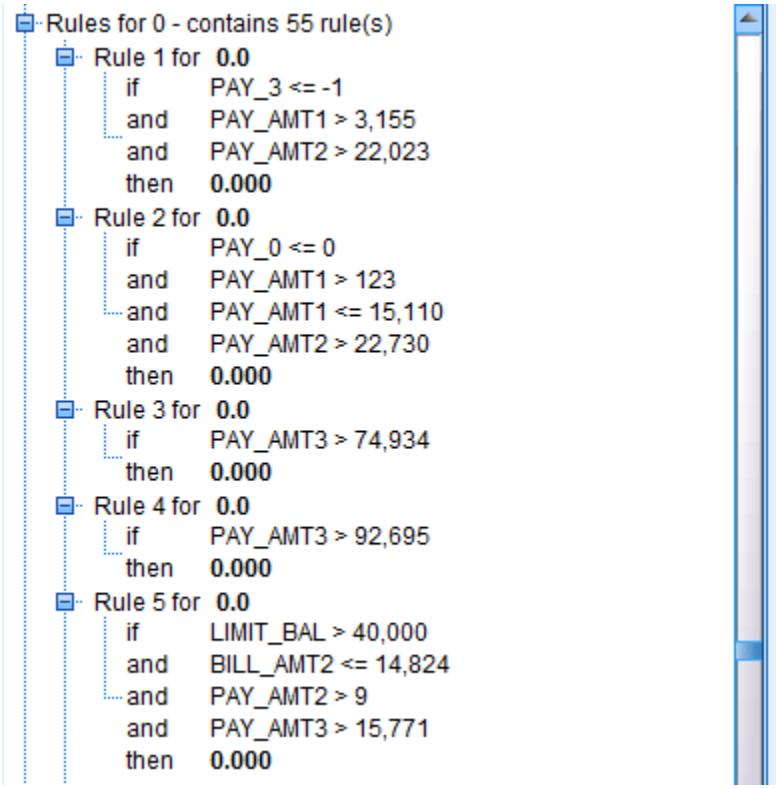


Figure 8. Partial rules for non-defaulters



CONCLUSION

MLP Artificial neural Network perform relatively well in predicting defaulters based on the dataset used in this study. However, data mining algorithms needs to be applied to other credit card datasets to make a more reliable observation on their performance and accuracy.

References

[1] www.thebalance.com, April 26, 20180

[2] Chen, M. C., and Huang, S. H., 2003, "Credit scoring and rejected instances reassigning through evolutionary computation techniques." Expert Systems with Applications, 24(4).

[3] Henley, W. E., & Hand, D. J. (1997). Construction of a k-nearest- neighbor credit scoring system. IMA Journal of Management Mathematics, 8(4), 305–321.

[4] Thomas, J.W., "A Review of Research on Project-Based Learning",

https://scholar.google.com, 2000.

[5] Baesens, Setiono, Mues, & Vanthienen, "Using Neural Network Rule Extraction a Decision Tables for Credit-Risk Evaluation, Management Science, 2003.

[6] Desai, Crook, & Overstreet, " A comparison of neural networks and linear scoring models in the credit union environment", European Journal of Operational Research, 1996.

[7] Abbas Keramati and Niloofar Yousefi, "A Proposed Classification of Data Mining Techniques in Credit Scoring", Proceedings of the 2011, International conference on Industrial Engineering and Operations Management, 2011.

[8] Adela Ioana et.al, "Clustering analysis for credit default probabilities in a retail bank portfolio", Bucharest Acad Econ Stud, Database Syst J, 2012.

[9] Simona-Vasilica OPREA, "Informatics Solutions for Smart Metering Systems Integration", Informatica Economica, 2015.

[10] Arminger, G., D., and Bonne, T., "Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feed- forward networks", Computational Statistics, Vol. 12, pp. 293-310, 1997.

[11] Hand, D. J. and Henley, W. E., "Statistical Classification Methods in Consumer Credit Scoring: A Review", Journal of Royal Statistics Society, Part 3, pp. 523-541, 1997.

[12] Shmueli, G., Patel, N., and Bruce, P., *Data Mining for Business Intelligence*, Wiley, New Jersey, 2010.

[13] Neurosolution Documents, http://www.neurodimension.com/, 2018.

[14] Parsaye, K.A., "Characterization of Data Mining Technologies and Processes", The Journal of Data Warehousing, January 1998.

[15] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php, 2015.