# A comparison between statistical and Data Mining methods for credit scoring in case of limited available data

Hassan Sabzevari
Department of Risk Management, Karafarin Bank, Tehran, Iran
Tel: (+98) 21 88660418-20
Fax: (+98) 21 88664600
Email: h.sabzevari@karafarinbank.com

Mehdi Soleymani
Department of Sciences
Shahid Beheshti University
Tehran, Iran
Tel: (+98) 21 88660418
Fax: (+98) 21 88664600
Email: soleymani_mehdi@yahoo.com

Eaman Noorbakhsh
Department of Risk Management, Karafarin Bank
Tehran, Iran
Tel: (+98) 21 88660418-20
Fax: (+98) 21 88664600
Email: i.noorbakhsh@karafarinbank.com

## Abstract
Credit scoring is a method used to estimate the probability that a loan applicant or existing borrower will default or become delinquent. There are two types of methods used for scoring: Traditional statistics models like Probit and Logistic regression and Data Mining models such as Classification and Regression Trees (CART), Multivariate Adaptive Regression Splines (MARS) and Bagging. In this paper, we have examined the performance of different models in credit scoring on real data of a bank and compared two approaches above. We found out that Bootstrap Aggregating (Bagging) model in Data Mining approach and the Logit regression in traditional statistical methods performs better than other methods in credit scoring application. We also present an empirical study on the machine learning feature selection methods, which provide an automatic technique for reducing the feature space. The study illustrates how these methods help to improve the performance of scoring models.

## Keywords
Credit Scoring, Probit and Logit Regression, CART, Neural Network, MARS, Bagging, Feature selection.

## 1. Introduction
Credit scoring is a statistical method used to predict the probability that a loan applicant or existing borrower will default or become delinquent [11]. The method, introduced in the 1950s, is now widely used for consumer lending, especially credit cards, and is becoming more commonly used in mortgage lending. Credit scoring has gained more and more attention as the credit industry can benefit from improving cash flow, insuring credit collections and reducing possible risks. Hence, banks and researchers have developed many different useful techniques, known as credit scoring models, in order to solve the problems involved in the evaluation process. The objective of credit scoring models is to assign credit applicants to either a “good credit” group that is likely to repay its financial obligation completely or a “bad credit” group who has high possibility of defaulting on its financial obligation. Therefore, credit scoring problems are in the scope of the more general and widely discussed classification problems [12].

Modeling techniques like statistical analyses and Data Mining techniques have been developed in order to successfully attack the credit scoring tasks. Statistical methods like Probit and Logit regression and Data Mining methods such as Classification and Regression Tree (CART), Neural Network (NN), Bootstrap Aggregating (Bagging) and Multivariate Adaptive Regression Splines (MARS) due to their associated memory characteristic and generalization capability are becoming very popular in credit scoring models. Some researchers have criticized statistical credit scoring techniques due to their model assumptions. In contrast, Data Mining approach is becoming a common alternative for making credit scoring models due to its associated memory characteristic, generalization capability, and outstanding credit scoring capability. However, Data Mining is also being criticized for its long training process, inability to identify the relative importance of potential input variables, and certain interpretative difficulties [12].

One of the challenges that researchers face when they use classification algorithms building credit-scoring models, is which features to select. The real-world data used for scoring models contains not only many observations, but also a large number of features. Some of the features may be irrelevant to credit risk; some of them may be redundant due to their high inter-correlation. With so many irrelevant and redundant features, most of the classification algorithms suffer from extensive computation time, possible decrease in model accuracy and decrease in scoring interpretation [13].

The objective of the proposed study is to explore the performance of credit scoring using two commonly discussed approaches: Regression and Data Mining techniques given our limited sample size. We are aware that our conclusions are based on a very limited database. We would appreciate if others could follow this study and verify if their results confirm ours.

To demonstrate the effectiveness of credit scoring using approaches above, we have performed credit-scoring task on corporate clients of Karafarin Bank of Iran. We conducted an experiment on these real data sets using the following classification algorithms: Probit and Logistic regression, CART, NN, Bagging and MARS.

This paper explores the most widely used methods for credit scoring, and then briefly explains the methods, which we have applied in our work. We have compared and discussed the performance, advantages and disadvantages of each method. We continue by describing feature selection methods, presenting the process of feature selection and discussing the results of the experiment. Finally, the conclusion and related future work are provided.

## 2. Related works

In terms of the theories and methods used to date, credit scoring models may be divided into two large groups:

**Parametric credit scoring models**

- Linear regression model;
- Probit and Logit models;
- Discrimination analysis-based models;

**Non-parametric credit scoring models**

- Classification trees (recursive partitioning algorithms);
- Neural Networks;
- Multivariate Adaptive Regression Splines;
- Bagging;
- Mathematical programming;
- Analytical hierarchy process;
- Nearest neighbors model;
- Expert systems[9].

The foundations of the 60-plus-year history of credit scoring were laid by FISHER's article published in 1936, which examined the distinctiveness of groups in a factory population based on various measured characteristics. As far as we know, it was DUNHAM in 1938 who first mentioned a system for the evaluation of credit applications, in which he used five criteria. DUNHAM wanted to know which parameters lenders found important and which characteristics were significant statistically [9].

DURAND was the first to use discrimination analysis, based on FISHER's results. DURAND's model is the first formalized version of lending knowledge gained from experience that lending companies could use as a decision making algorithm, without the continuous involvement of credit/lending experts [9].

In 1963 MYERS and FORGY recommended the application of multivariate discrimination analysis to these parameters. Usually, both models are estimated using the maximum likelihood method thus making their computerized implementation and application relatively simple and inexpensive. Those that deserve to be mentioned include the articles of CHESSER, who first recommended the logit model, SRINIVASAN and KIM (logit), STEENACKERS and GOOVAERTS (logit), and BOYES (probit) [9].

Data Mining, sometimes referred to as knowledge discovery in database (KDD), is a systematic approach to find underlying patterns, trend, and relationships buried in data. Data Mining has drawn serious attention from both researchers and practitioners due to its wide applications in crucial business decisions. The classification problems, where observations can be assigned to one of several disjoint groups, have long played important roles in business decision making due to their wide applications in decision support, financial forecasting, fraud detection, marketing strategy, process control, and other related fields [12].

The RPA (recursive partitioning algorithms) method has been examined and used successfully by many researchers. FRYDMAN, ALTMAN and KAO analyzed the classification problem of companies in difficult situations, comparing the efficiency of the RPA with that of processes based on discrimination analysis. MARAIS, PATELL and WALFSON studied the usability of RPA and probit models in commercial lending. KIM and SRINIVASAN examined the applicability of the RPA in industrial lending [9].

The development of the analytical hierarchy process (AHP) is attributable to SAATY. It is based on the principle that when we decide on a given matter, we actually consider a lot of information bits and factors, and there is an information hierarchy [9].

The evaluation of credit/loan applications and customer rating, is also modeled` by expert systems. For example, HOLSAPPLE and his associates and PAU examined the applicability of expert systems in financial management, primarily in lending [9].

TAM, for example, compared the efficiency of neural network models with that of classic methods in relation to a corporate and banking bankruptcy forecasting exercise. MCLEOD and his associates examined lending applications in 1993, ROBINS and JOST summed up the initial experience gained in the application of neural networks in marketing. They agree that the model is capable of more accurate forecasts than previous scoring systems [9].

## 3. A brief review of used methods

In this section, we will briefly review the literature of credit scoring models, which we have used in the present study. Thus, a brief outline of probit and logistic regression, CART, neural network, Bagging, and MARS are presented.

### *3.1.1 Probit and logit models*

Several statistical methods are used to develop credit-scoring systems, including linear probability, Probit and Logit models. If we assume that a standard normal distribution expresses probability, this yields the probit model. Cumulative distribution function in probit constitutes a transformation that puts the probability value in the [0; 1] interval, while maintaining the monotonic property (i.e. they are either monotonic increasing or monotonic decreasing functions) [9].

If the logistic distribution function is selected to express the probability of approval, it will lead to the logit model. Logistic regression is a widely used statistical modeling technique in which the probability of a dichotomous outcome is related to a set of potential independent variables. The logistic regression model does not necessarily require the assumptions of some other regression models, like assuming that the variables are normally distributed in discriminant analysis [13]. Usually, both models (probit and logit) are estimated using the maximum likelihood method. As these models are widely used, a large number of studies have been released on their application and the experience gained in consumer, commercial and agricultural lending [9].

### *3.1.2 Classification and regression tree*

Classification tree is a nonparametric method to analyze categorical dependent variables as a function of metric and/or categorical explanatory variables [5]. As the name suggests, CART is a single procedure that analyzes either categorical or continuous data using the same method. The methodology outlined in Breiman et al (1984)[1]. has three stages. The first

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees, Wadsworth, Pacific Grove, CA.

stage involves growing the tree using a recursive partitioning technique to select variables and split points using a splitting criterion. Several criteria are available for determining the splits, including Gini, Twoing and Ordered Twoing. For a more detailed description of the mentioned criteria, one can refer to Breiman et al. In addition to selecting the primary variables, one can choose surrogate variables, which relate closely to the original splits, and use them in classifying observations having missing values for the primary variables [12].

After a large tree is identified, the second stage of the CART methodology uses a pruning procedure that incorporates a minimal cost complexity measure. The result of the pruning procedure is a nested subset of trees starting from the largest tree grown and continuing the process until only one node of the tree remains. Cross-validation or a testing sample will be used to provide estimates of future classification errors for each subtree. The last stage of the methodology is to select the optimal tree, which corresponds to a tree yielding the lowest cross-validated or testing set error rate. Trees in this stage have been identified as unstable. To avoid this instability, trees with smaller sizes, but comparable in classification accuracy, will be chosen as an alternative. This process is referred to as the one standard error rule and can be tuned to obtain trees of varying sizes and complexity [12].

### *3.1.3 Neural network*

Neural networks are artificial intelligence algorithms that allow for some learning through experience to discern the relationship between borrower characteristics and the probability of default and to determine which characteristics are most important in predicting default [11]. Research into neural networks began in 1943 with the publication written by MCCULLOCH and PITT. According to the proclaimed principle, a mathematical model had to be developed that could simulate the natural operation of the neuron. For us, the important parts of a neuron are the dendrites, through which the neuron receives signals, and the axons, which help to forward processed information to other neurons. Synapses play a significant part in processing information. It is through them that axons connect to the dendrites of other neurons [9].

An artificial neural network tries to mimic the brain, which consists of numerous neurons [1]. The operation of the mathematical neuron model is relatively simple. Using a given function, they process the information received from dendrites, and if the incoming signal exceeds a so-called stimulus threshold, they forward the information via axons. The most important property of the neuron is that it is continuously changing its operation (i.e. its internal function) based on the data received – it is 'learning'. Synapses play an important role in this learning process, as they are able to amplify or subdue the signals coming from other neurons. In the learning process, signal amplification factors change on synapses (referred to as 'weights' in the model in light of their function). In the neuron model, the change or modification of these weights means learning [9].

### *3.1.4 Bagging*

Bagging predictors is a recent and successful computationally intensive method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Bagging is extremely useful for large, high dimensional data set problems where finding a good model or classifier in one step is impossible because of the complexity and scale of the problem. Breiman (1996a) introduced Bagging to reduce the variance of a predictor. It has attracted much attention and has been frequently applied and tested on real and simulated data sets. The results show that Bagging can give substantial gains in accuracy [1].

### *3.1.5 Multivariate adaptive regression splines*

Multivariate Adaptive Regression Splines (MARS) was introduced by Friedman. The modeling procedure is inspired by the recursive partitioning technique governing CART and generalized additive modeling, resulting in a model that is continuous with continuous derivatives. It excels at finding optimal variable transformations and interactions, the complex data structure that often hides in high-dimensional data, and hence can effectively uncover important data patterns and relationships that are difficult, if not impossible, for other methods to reveal. Adaptations of MARS include parallel computing algorithms and an automatic stopping rule for the forward selection algorithm. The original MARS version provides a smoothing option to guarantee a continuous first derivative, and Chen et al developed a modification [12].

The MARS model is essentially a linear statistical model with a forward stepwise algorithm to select model terms followed by a backward procedure to prune the model. The approximation bends to model curvature at "knot" locations, and one of the objectives of the forward stepwise algorithm is to select appropriate knots [12].

The optimal MARS model is selected in a two-stage process. First, MARS constructs a very large number of basic functions (BF), which are selected to overfit the data initially, where variables are allowed to enter as continuous, categorical, or ordinal, the formal mechanism by which variable intervals are defined, and they can interact with each other or be restricted to enter in only as additive components. In the second stage, basis functions are removed in the order of least contribution using the generalized cross-validation (GCV) criterion [12].

## 4. Variables and feature selection methods

Essentially, there are three main categories as possible model input for credit scoring: accounting variables, market based variables such as market equity value and so-called soft facts such as the firm's competitive position or management records. Historically banks used to rely on the expertise of credit advisors who looked at a combination of accounting and qualitative variables to come up with a subjective assessment of the client's credit risk. During the last decades, especially larger banks switched to quantitative models. However, in addition to the selected accounting ratios they included other factors into the model building process [6].

To select the most predictive independent features and reduce redundancy, some scholars and practitioners have applied the following methods. For example, some used univariate analysis to evaluate the effect of each independent feature on the class feature, some used statistical correlation analysis to detect the correlation between different independent features and eliminate highly correlated ones. In addition, sometimes the process of feature selection incorporates into the classification algorithms. For example, this method combines with stepwise

statistical procedures in discriminant analysis and in feature selection while building decision trees [6].

Feature selection has become the focus of a lot of researches for different applications in which there are datasets with tens of variables. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost effective predictors, and providing a better understanding of the underlying process that generated the data [4].

## 4.1 Forward vs. backward selection

Backward elimination methods were applied to check whether all chosen accounting ratios added statistical significance to the group or whether the Logit model could be reduced to a lower dimension. Backward elimination is one possible method of statistical stepwise variable selection procedures. It begins by estimating the full model and then eliminates the worst covariates one by one until all remaining input variables are statistically significant, i.e. their significance level is below the chosen critical level. For this study, the analysis was based on the precise likelihood ratio test, where the significance level was set at 95 percent. Of course, some critics argue that statistical stepwise selection procedures are kind of data mining methods, as they can yield theoretically implausible models and select irrelevant variables [6]. It is often argued that forward selection is computationally more efficient than backward elimination to generate nested subsets of variables. However, the defenders of backward elimination argue that weaker subsets are found by forward selection because the importance of not included variables is not assessed yet. Backward selection method may perform better than forward selection by eliminating the variables that don't provide the best separation [4].

## 4.2 Data Mining feature selection for credit scoring models

Many machine-learning researchers have developed various feature selection methods. Whether these methods are efficient for the feature selection problem in credit scoring models is rarely discussed. Feature selection using automatic Data Mining is defined as 'the process of finding a best subset of features, from the original set of features in a given data set, optimal according to the defined goal and criterion of feature selection (a feature goodness criterion)'. Many researchers have developed various feature selection algorithms using different evaluation criterions and searching strategies. Some feature selection methods use a measure to evaluate the goodness of individual features. For example, information measures, distance measures and dependence measures. Features are ranked according to their values on this measure. One can simply choose the first X features as the selected feature subset. X is decided according to some domain knowledge or a user-specified threshold value [13].

### *4.2.1 First method*

Relief algorithm is a feature-ranking algorithm first proposed by Kira and Rendell, and was enhanced by Kononenko. The key idea of Relief is to rank the quality of features according to how well their values distinguish between the cases that are near to each other. It is reasonable to expect that a useful feature should have different values between cases from different classes and have the same value for cases from the same class [4].

In this method, value of variable X of each observation is compared to X value of all other observations with the same response value so the mean of k-nearest value computed. Then the same procedure is done but all observations with the different response value are used. This method is done for all observations and the same X variable. By repeating the procedure for all independent variables, we can rank the explanatory variables. The following algorithm shows the first method of feature selection, which is used in this paper, and is very similar to RELIEF algorithm [4].

*Set M1 (feature A) = 0;*

*For r=1 to n do:*

*Begins;*

*Select case R;*

*Find its K-nearest neighbors from the same class (case S);*

*Find its K-nearest neighbors from a different class (case D);*

*M1=M1-S+D;*

*End;*

### *4.2.2 Second method*

Our second chosen method is the Wrapper feature selection algorithm. Wrapper algorithms, the evaluation measure for the goodness of feature subsets is the classification accuracy provided by a target learning algorithm. The classification accuracy is estimated with a test data set in the wrapper algorithm in this paper:

*MWRP= classification accuracy= 1 - test error rate*

The wrapper algorithm searches for the feature subset that generates the lowest error rate in the test data set. If a decrease of the classification error rate is the foremost concern in building a scoring model, the wrapper feature selection approach is appropriate. The wrapper approach can generally give better results in model accuracy when using the target-learning algorithm. However, due to extensive computation time required, this approach cannot be used for large data sets and time-consuming classification algorithms [4]. In this empirical study, we used above Data Mining feature selection methods in corresponding credit scoring models.

# 5. Empirical study

In order to assess the effectiveness of credit scoring using statistical and Data Mining methods, we used the corporate client data of a small private bank in Iran. There are totally 719 customers in the data set. Out of which, 558 have good records, while 161 have defaulted. These observations include client's financial statements, client's specifications, type and amount of loan and client's performance in loan repayment.

The loan is considered as default if the client does not repay part of his due obligations more than 90 days. Usually the number of records for credit scoring is above 10000, but since we had restrictions on our available data, we had to find an approach that works best in case of limited available data. However, we reduced the number of variables to less than 10 using the feature selection approaches explained above. We can use these variables for all of described models. We tested and compared the performance of following models in classifying customers.

## 5.1 Feature or variable selection

We used the following methods (described above) to provide for a data mining technique in order to reduce the feature space of various models.

### *5.1.1 First method (Rellif)*

We have run the first algorithm by selected Software. Through using the search method described above, Features can be ranked as follows:

e.year, bdet.det, ldet.ast, roe, logast.gdp, det.ast, wcap.ast, bdet.ast, prof.ast, cast.totd, aprof.ast, equ.ast, csh.ast, bdet, ldet.equ, opr.ast.

### *5.1.2 Second method (Wrapper)*

According to this method, we can reduce the candidate variables to below variables:

Equ.ast, det.ast, l.perd, r.year, e.year, hist.cod, logast.gdp

## 5.2 Estimating desired models

In this section, we present the performance result of each model briefly.

### *5.2.1 Logit Model*

The stepwise variables selection procedure (both forward and backward) is used in building the logistic regression credit scoring model. The intercept and seven significant independent variables including the ratio of Accumulated profit to Total Assets (Aprof.ast), History of Bad Debt (hist.cod), Loan period (L.perd), the ratio of Net profit to shareholders equity (ROE), Number of years as a bank's customer (R.year), the ratios of total debt to total assets (Det.ast) and Bank debt to total assets (Bdet.ast) were included in the final regression model.

The output of the model is a number lying between zero and one, which corresponds to the probability of default of that customer. The model also finds the best cut-off point at which you can distinguish good customers from bad ones. In our experiment, the cut-off point was 0.25, which means that those customers, who have a score more than 0.25, are classified in "bad credit" group.

**Expectation-Prediction (Classification) Table**

The following table (table 3) shows correct and incorrect classification based on a user specified prediction rule and on expected value calculations. Lying between zero and one, the predicted default probability would classify each observation based on a cutoff point (here 0.25).

"Correct" classifications are obtained when the predicted probability is less than or equal to the cutoff and the observed Dep[2]=0, or when the predicted probability is greater than the cutoff and the observed Dep=1. In our experiment above, 413 of the Dep=0 observations and 105 of the Dep=1 observations are correctly classified by the estimated model. Overall, the estimated model correctly predicts 72% of the observations (74% of the Dep=0 and 65% of the Dep=1 observations).

[2] Dependent

**Table 1. All candidate variables for modeling**

| Row | Variable Name | Variable Full Name |
|---|---|---|
| 1 | E.year | Years of Establishment |
| 2 | R.year | Years of Relation With Bank |
| 3 | L.perd | Length of Loan |
| 4 | Prof.cod | Profit Code |
| 5 | Hist.cod | History Code |
| 6 | Det.ast | Debt to Asset ratio |
| 7 | Equ.ast | Equity to Asset ratio |
| 8 | Ldet.ast | Long term Debt to asset ratio |
| 9 | Bdet.ast | Bank Debt to Asset ratio |
| 10 | Bdet | Bank Debt to (Assets – Bank Debt) ratio |
| 11 | Bdet.det | Bank Debt to Debt ratio |
| 12 | Ldet.equ | Long Term Debt to Equity ratio |
| 13 | Liqu | Liquidity Ratio ratio |
| 14 | Cast.totdet | Current Asset to Total Debt ratio |
| 15 | Wcap.ast | Working Capital To Asset ratio |
| 16 | Csh.ast | Cash to Asset ratio |
| 17 | Wcap.sal | Working Capital to Sale ratio |
| 18 | Csh.sal | Cash to Sale ratio |
| 19 | Cast.sal | Current Asset to Sale ratio |
| 20 | QuAst.sal | Quick Asset Sale ratio |
| 21 | Csh.cDet | Cash to Current Debt ratio |
| 22 | Wcap.cDet | Working Capital to Current Debt ratio |
| 23 | Qui | Quick ratio |
| 24 | Racc.sal | Receivable Account to Sale ratio |
| 25 | Racc.Det | Receivable Account to Debt ratio |
| 26 | Pacc.sal | Payable Account to Sale ratio |
| 27 | Sal.Ast | Sale to Asset ratio |
| 28 | Opr.Ast | Operational Profit to Sale ratio |
| 29 | Ebt.sal | Earning Before Sale ratio |
| 30 | Gprof.ast | Gross Profit to Asset ratio |
| 31 | Prof.ast | Profit to Asset ratio |
| 32 | Prof.sal | Profit to Sale ratio |
| 33 | Aprof.ast | Accumulate Profit to Asset ratio |
| 34 | Ast.CPI | Asset to CPI ratio |
| 35 | Sal.CPI | Sale to CPI ratio |
| 36 | Log.ast.GDP | Long Term Asset to GDP ratio |
| 37 | ROE | Return On Equity ratio |

**Table 2. Outputs of Logit**

| Dependent variable | coefficient | Std | p- value |
|---|---|---|---|
| Intercept | -2.547 | 0.718 | 0.0004 |
| Aprof.ast | 2.40 | 0.86 | 0.005 |
| hist.cod (0) | -0.77 | 0.129 | 0.0001 |
| hist.cod (1) | 0.09 | 0.14 | 0.539 |
| L.perd | 0.021 | 0.006 | 0.0009 |
| ROE | -0.712 | 0.23 | 0.0024 |
| R.year | -0.17 | 0.06 | 0.005 |
| Det.ast | 1.7 | 0.79 | 0.02 |

| | | | |
|---|---|---|---|
| Bdet.ast | 1.16 | 0.44 | 0.009 |

**Table 3. Logit classification table**

| | Good Loan | Bad Loan | Total |
|---|---|---|---|
| P(DEP=1)<=./25 | 413 | 145 | 558 |
| P(DEP=1)>./25 | 56 | 105 | 161 |
| Total | 469 | 250 | 719 |
| Correct | 413 | 105 | 518 |
| % Correct | 74 | 65 | 72 |
| % incorrect | 26 | 35 | 28 |

### *5.2.2 CART model*

CART has become a standard tool for developing credit scoring systems, since they are easily interpretable and may be demonstrated graphically. Therefore, we would like to discuss this method, more in detail.

**Tree building**

Beginning with the first node, the CART software finds the best possible variable to split the node into two child nodes. In choosing the best splitter, the program seeks to maximize the average "purity" of the two child nodes. The "maximal" tree which is created is generally very overfit. In other words, the maximal tree follows every idiosyncrasy in the learning dataset, many of which are unlikely to occur in a future independent group of customers [7].

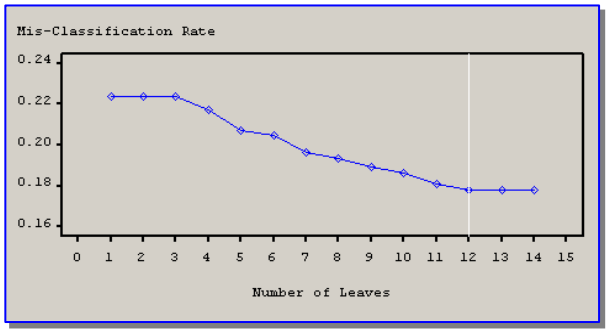


**Figure 1. Number of leaves for the final tree**

**Tree pruning**

In order to generate a sequence of simpler trees, each of which is a candidate for the appropriately-fit final tree, the method of "cost-complexity" pruning is used. This method relies on a complexity parameter, denoted $\alpha$, which gradually increases during the pruning process [7].

**Optimal tree selection**

The maximal tree will always fit the learning dataset with higher accuracy than any other tree. At the beginning, the largest tree is constructed based on train sample and minimum misclassification rate [7]. The constructed tree in this experiment has 12 leaves, which you can see in figure 1.

**Score importance**

These importance measures incorporate information both on the use of variables as primary splitters, and also their relative worth as surrogate variables. A discussion of variable importance is beyond the scope of this paper. In this part, it is shown how customers are classified by the decision tree [7].

| Name | Importance | Role | Rules | Variable Label |
|---|---|---|---|---|
| L_PERD | 1.0000 | input | 2 | L#perd |
| HIST_COD | 0.9425 | input | 1 | hist#cod |
| ROE | 0.9347 | input | 3 | ROE |
| BDET_AST | 0.7741 | input | 2 | Bdet#ast |
| DET_AST | 0.6020 | input | 1 | Det#ast |
| APROF_AST | 0.5020 | input | 1 | Aprof#ast |
| R_YEAR | 0.4684 | input | 1 | R#year |

**Figure 2. The importance of used variables**

### *5.2.3 Neural networks model*

Since Vellido et al. (1999) pointed out that more than 75% of business applications using neural networks will use the BPN[3] training algorithms, this study will also use the popular BPN in building the neural network credit scoring model. As recommended by Cybenko (1989), and Hornik et al. (1989) that three-hidden-layer network is rather appropriate to model complex system with desired accuracy, the designed network model will have three hidden layer as figure 3 shows [7].

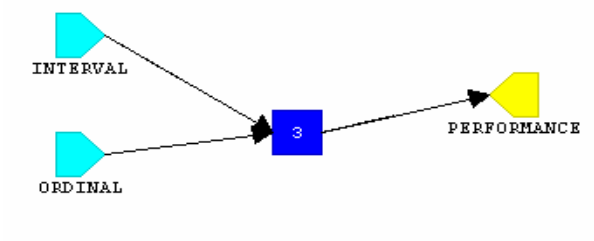


**Figure 3. The structure of final Network**

## 6. Comparison of various models

In the present empirical study, the performance of models can be evaluated with other two important practicability criteria: the speed of models in terms of training time and applying time as well as the comprehension of the final model [10]. In this section, other more important methods will be presented.

## 6.1 Classification accuracy

We should point out that the seven models obtained above, are based on a relatively small sample. In addition, each model is tested with the test sample; the overall view of comparative analysis is presented by following results (table 4). The classification outputs of each model for in-sample and out-of-sample are firstly derived. Outputs of models includes the true class sample, the predicted class, and the predicted score for each case in the test sample.

At first, to compare the performance of different classification algorithms, the introduced seven models are estimated: Their performances are compared with respect to the classification accuracy, both for the whole 719 training observations and 250 balanced (means equal size of good and bad customers) training observations.

[3] Back Propagation Neural network

**Table 4. The accuracy table of in sample or training data set**

| Num of Observations | Model | 0 | 1 | Accuracy |
|---|---|---|---|---|
| 719 Training observations (without Data Mining feature selection) | Logit | 74 | 65 | 72 |
| | probit | 74 | 59 | 70 |
| | CART | 80 | 64 | 77 |
| | Neural Network | 75 | 69 | 73 |
| | MARS | 78 | 49 | 71 |
| | Bagging | 82 | 72 | 79 |
| | Combined | 74 | 70 | 73 |
| 250 Training observations (without Data Mining feature selection) | Logit | 74 | 71 | 73 |
| | probit | 77 | 72 | 76 |
| | CART | 85 | 63 | 76 |
| | Neural Network | 84 | 80 | 79 |
| | MARS | 82 | 53 | 67 |
| | Bagging | 80 | 80 | 80 |
| | Combined | 81 | 80 | 80 |

According to the results above, Bagging method performs better than the other models all together in and out of sample test.

After the completion of the default risk prediction modeling process, we apply the estimated models to the validation samples to produce out-of sample forecasts. Table 5 illustrates the results of out-of sample test for two separate test observations. In the first data set, we have used 70 observation (randomly) chosen form a 320 observations data set, while 250 observations are used above for training and creating different models. As you can observe, the MARS, Bagging, and neural network among Data Mining models and logit regression among two regression models, relatively perform better than others.

**Table 5. The accuracy table of out-of sample or test data set**

| Num of Observations | Model | 0 | 1 | Accuracy |
|---|---|---|---|---|
| 70 Test observations (with Data Mining feature selection) | Logit | 75 | 74 | 75 |
| | probit | 75 | 73 | 74 |
| | CART | 70 | 68 | 69 |
| | Neural Network | 74 | 70 | 72 |
| | MARS | 76 | 71 | 74 |
| | Bagging | 72 | 73 | 72 |
| | Combined | 78 | 76 | 77 |
| 70 Test observations (with Data Mining feature selection) | Logit | 74 | 72 | 73 |
| | probit | 73 | 50 | 57 |
| | CART | 76 | 78 | 77 |
| | Neural Network | 72 | 56 | 60 |
| | MARS | 69 | 82 | 75 |
| | Bagging | 80 | 68 | 77 |
| | Combined | 65 | 60 | 60 |

## 6.2 Model discrimination

Model discrimination power was assessed for each species using the concepts of Receiver Operating Characteristic Curve (ROC), Cumulative Accuracy Profiles (CAP), and Accuracy Ratios (AR). The ROC Curve is a plot of the percentage of the defaulters predicted correctly as defaulters (Hit-Rate) vs. the percentage of the non-defaulters wrongly classified as defaulters using all cut-off values that are contained in the range of the rating scores. In contrast to this, the CAP is a plot of the Hit-Rate vs. the percentage of all debtors classified as defaulters for all possible cut-off values. Besides, as proven in Engelmann, Hayden, and Tasche (2003), the Accuracy Ratio is just a linear transformation of the area below the ROC curve. Hence, both concepts contain the same information and all properties of the area under the ROC curve are applicable to the AR [6].

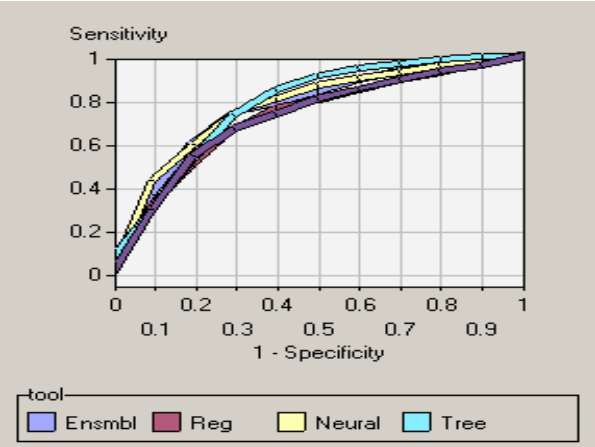


**Figure 4. The ROC curves for estimated models**

The probabilistic interpretation of the ROC curve is based on an article of Bamber (1975) [3]. The area under an ROC curve is equal to the probability that a randomly selected positive case (defaulted loan) will receive a higher score than a randomly selected negative case (non-defaulted loan). Better curves are closer to the upper-left corner. When the curve of one model is completely above the curve of another model, it is clear that the model will perform better regardless of what threshold is used [11]. As figure 4 shows, the best model is CART and Combined is the second one. The significance of differences in error rates between each pair of algorithms is not statistically tested. However, it can be concluded that the difference between CART and combined is significant.

# 7. Conclusion

Modeling techniques like traditional statistical analyses and Data Mining (artificial intelligence) techniques have been developed in order to successfully attack the credit scoring problems. Probit and logistic regressions are the most commonly used statistical credit scoring techniques, but are often criticized due to their model assumptions. On the other hand, the Data Mining approach is becoming a real alternative in credit scoring tasks due to its associated memory characteristic, generalization capability, and outstanding credit scoring capability. However, it is also being criticized for its long training process, inability to identify the relative importance of potential input variables, and certain interpretative difficulties.

This paper has investigated the accuracy of the quantitative models under consideration by the Iranian private bank for credit scoring applications. The results of this research suggest that in training test among statistical models, logit regression and among the kinds of Data Mining models, which we tested, Bagging credit scoring model can achieve fractional

improvements in credit scoring accuracy, therefore, as regards in-sample (or training) data, Bagging model is much better in terms of accuracy. As out of sample test is concerned, you can see that Bagging and the combined model are good alternatives to the logistic regression. At the same time, Logistic regression is slightly more accurate than other Data Mining models (CART, Neural Network, and MARS) except Bagging.

In theory, there are many potential benefits of variable and feature selection: facilitating data visualization, data understanding, and reducing the dimension. In addition to examining various credit scoring models, we analyzed the possible advantages of features selection. Although the conclusions in the paper need to be validated with other data sets in future researches, a general conclusion can be drawn from this study: the automated data mining feature selection technique provides an effective method for selecting the most predictable features and Data Mining credit scoring models. As the cost of misclassification for out of sample with Data Mining feature selection is considered, you see that Bagging model gets better results. To sum up, this paper recommends that Bagging and Logit are good methods for fitting models when there is a limited sample of data.

## 8. Future works

As our studies mainly use financial variables as independent variables, future studies may aim at collecting more variables that are important, e.g. business environment, in improving the credit scoring accuracies. Integrating fuzzy discriminate analysis, genetic algorithms, with neural networks and/or support vector machines are possible research directions in further improving the credit scoring accuracies.

To reach a more general conclusion, further experiments need to be conducted on other larger data sets. In addition, it would be meaningful to show the impact of the studied feature selection methods on other classification criteria, for example, the area under the receiver operating characteristic curve.

## 9. Acknowledgements

The authors would like to express their acknowledgement to Karafarin Bank, our executive managing-director Dr. Parviz Aghili Kermani, managing-director adviser Mr.Massoud Rad, Mr. Mehdi Parsa, and his kind colleagues in Credit Department, for their valuable supports, comments and cooperations.

## 10. Reference

[1] Breiman, L., Bagging predictors, Machine Learning, 26123-140, 1996a.

[2] Chakrabarti, Baijayanta, and Varadachari, Ravi, Quantitative methods for default probability estimation, 2004, i-flex solution.

[3] Engelmann, Bernd, Hayden, Evelyn, and Dirk Tasche, Measuring the Discriminative Power of Rating Systems, Discussion paper Series 2 Banking and Financial Supervision, No.01, 2003.

[4] Guyon, Isabelle, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 1157-1182, 2003.

[5] H¨ardle ,Wolfgang, On Credit Scoring Estimation, Submitted to: Institute for Statistics and Econometrics Humboldt University Spandauer Str. 1, D–10178 Berlin.

[6] Fogarty, James, S. Baker, Ryan, and Scott E. Hudson, Case Studies in the use of ROC Curve Analysis Sensor-Based Estimates in Human Computer Interaction, Human Computer Interaction Institute, Carnegie Mellon University.

[7] Hayden, Evelyn, Are credit scoring models sensitive with respect to default definition?, Evidence from the Austrian market, 2003.

[8] J. Lewis, Roger, An Introduction to Classification and Regression Tree (CART) Analysis, Department of Emergency Medicine Harbor-UCLA, Medical Center Torrance, California.

[9] Kiss, Ferenc, Credit scoring process from a knowledge management perspective. Periodical polytechnica ser.soc.man.scl. VOL. 11, NO. 1, PP. 95–110, 2003.

[10] Liu, Yang, "The evaluation of classification models for credit scoring". Arbeitsbericht. Nr. 02, 2002.

[11] Mester, Loretta, What's the point of the credit scoring?, Business Review-Federal Reserve Bank of Philadelphia, 1997.

[12] Leea, Tian-Shyug, Chiub, Chih-Chou, Chouc, Yu-Chao, and Chi-Jie Lud, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines.

[13] Y, Liu, and M. Schumann, Data mining feature selection for credit scoring Models, Journal of the Operational Research Society 56, 1099–1108, 2005.