Comparative analysis of data mining methods for predicting credit default probabilities in a retail bank portfolio

Adela Ioana Tudor, Adela Bâra, Simona Vasilica Oprea
Department of Economic Informatics and Cybernetics
Academy of Economic Studies, Bucharest, Romania
adela_sw@yahoo.com; bara.adela@ie.ase.ro; simona.oprea@transelectrica.ro

Abstract: Predictions about the borrowers’ ability to repay is the essence of credit risk management. A number of approaches and methodologies have been developed and implemented to date, but research continues to improve existing techniques in order to obtain optimal solutions, and design new models. In this context, we will examine different methods of data extraction and analysis in terms of their accuracy, in order to assess their comparative effectiveness for determining the probability of default. The data mining techniques used in this paper are Logistic Regression, Naive Bayes algorithm and Support Vector Machines. We consider very interesting and useful the fact that the probability of default is calculated for each debtor individually, and not for groups of customers with specific features, as practiced in most banks today.

Key words: data mining, classification, regression, support vector machine, attributes importance, business intelligence.

1. Introduction

In terms of lending, there are two major approaches to model credit risk at the transaction level: analytical approach based on accounting and statistical predictions. Because there is a large amount of information about the client in all financial business, a properly way of implementing an effective model is to use their own database together with data mining techniques. Business Intelligence (BI) solutions provides all the necessary functions to identify, integrate and analyze data, offering decision support at strategic management level. Data Mining (DM), as part of BI systems, has enjoyed great popularity in recent years, with advances in both research and commercialization. Data mining is focused on assessing the predictive power of models and performs analysis that would be too hard-working and time-consuming by using traditional statistical methods. Data mining technology can solve business problems in banking and finance areas by finding attributes, causalities and correlations that are not immediately obvious to managers, because the volume of data is too large or too quickly generated on the screen.

The main idea of the article is to build a model based on data mining techniques that can predict customer behaviour over time. Thus, we performed a comparative analysis of data mining techniques to assess the probability of default for credit card holders and determine the best statistical methods of prediction, in our view, for analyzing this indicator. The methods implemented are: Logistic Regression, Naive Bayes algorithm and Support Vector Machines.

The paper is structures as follows: Section 2 describes the model for data analysis based on a specific scoring model that we proposed. In section 3, we apply the attribute importance algorithm and improve the initial model in order to make a comparative analysis. Section 4 shows the conclusions.

For our study we combine methods with supervised and unsupervised learning mechanisms. For testing different methods we used Oracle Data Mining (ODM) that is organized around several generic operations, providing an unified interface for extraction and discovery functions. These operations include functions for construction, implementation, testing and manipulation of data to create models. ODM implements a series of algorithms for classification, prediction, regression, clustering, association, selection, and data analysis. Oracle Data Miner provides the following options for each stage: for transforming the data and build models (build), for testing the results (test) and for the evaluation and application on new data sets (apply).

2. The model for data analysis
Classification means grouping the observations based on a predictable attribute. Each case contains a set of attributes, of which one is the classification attribute (the predictable attribute). In our application, this attribute is called ‘RESTANTIER’. The operation consists in finding a model that describes the predictable attribute as a function of other attributes taken as input values. Basically, the objective of the classification is to first analyze the training data and develop a model for each class using the attributes available in the data. Such class descriptions are then used to classify future independent test data or to develop a better description for each class.

In order to train a classification model, the class values of each case in the data set must be known, values that are usually found in historical data [7]. We review the most popular classification algorithms used in our study: logistic regression (LR), Bayesian networks (BN) and support vector machines (SVM).

A Logistic Regression model is a special case of linear regression models and it specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. [1]

In logistic regression, there is no definition of the coefficient of determination ($R^2$) which is frequently used in the general linear model. $R^2$ has the desired interpretability as the proportion of variation of the dependent variable, which can be explained by the predictor variables of a given regression model. [6]

Bayesian networks show in a graphical manner the dependencies between variables. Based on the network structure, it can realize various kinds of inferences. There are two key optimization-related issues when using Bayesian networks. First, when some of the nodes in the network are not observable, that is, there is no data for the values of the attributes corresponding to those nodes, finding the most likely values of the conditional probabilities can be formulated as a non-linear mathematical program. In practice this is usually solved using a simple steepest descent approach. The second optimization problem occurs when the structure of the network is unknown, which can be formulated as a combinatorial optimization problem [5].

Support Vector Machines have been implemented in many financial applications recently, mainly in the area of time series prediction and classification. Support Vector Machine is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. [2]

The algorithm is considering the following steps:

a) Class separation: we must find the optimal separating hyper plane between the two classes. Linear programming can be used to obtain both linear and non-linear discrimination models between separable data points or instances. If this hyper plane exists, then there are many such planes. The optimal separating hyper plane is the one that maximizes the sum of the distances from the plane to the closest positive example and the closest negative example.

b) Overlapping classes: data points on the wrong side of the discriminated margin are weighted down;

c) Nonlinearity: when we cannot find a linear separator, data points are projected into a higher-dimensional space where the data points effectively become linearly separable (this projection is realised via kernel techniques);

d) Problem solution: the whole task can be formulated as a quadratic optimization problem which can be solved by specific techniques.

### 2.1. Data

The study is based on financial data in 2009 from an important bank in Romania and the target customers are credit card holders. Among the 18239 instances, 1489 record arrears. The research involves normalization of attribute values and a binary variable as response variable: 1 for the default situation and 0 for non-default. Explanatory variables or the attributes are found in the scoring model proposed by us that includes: Credit amount (the amount limit by which the debtor may have multiple withdrawals and repayments from the credit card), Credit balance, Opening date of the credit account, credit product identification, Category, Currency, Client’s name, Gender, Age, Marital status, Profession, Client with history (including other banking products and payment history), Deposit state, Amounts of deposits opened in the bank, if applicable, Scoring rate (scoring points from 1 to 6, 1 is for the best and 6 for the weakest debtor).

Data was divided into two tables, one used for model construction and one for testing and validating the model.

### 2.2. Applying the algorithm on all attributes

Oracle Data Mining enables the application of the algorithm on the recorded observations and the
possibility to configure the models. For results interpretation, there are made available a series of suggestive views. Thus, based on each client attributes found in the table above, we determine the customer profile in terms of solvency (good payer, bad payer). Let’s interpret the results on cost matrix and confusion matrix obtained from the construction and testing of each model. These are generated from the Accuracy module. Accuracy refers to the percentage of correct predictions made by the model compared to the real classifications from the test data. In the following figure, the top screen shows the cost matrix and the second screen shows the confusion matrix.

As we can see, for LG model, there are 16,750 instances with 0 value (debtors who repay) from which 98,15% were correct predicted, and 1,489 default cases, with 1 value, from which the model predicted correctly 93,22%. The total cost is 411 records.

The confusion matrix displays the model errors, as follows: the figure 1 from left cell bottom shows false negative predictions, predictions of 0 when the real value is 1 (the client is a good payer but he records default) and 310 from the right cell above indicates false positive predictions, predictions of 1 when the actual value is 0 (the customer is a bad payer but he repays the loan).

LG model made 17.828 (16.440 + 1.388) correct predictions and 411 (101 + 310) wrong predictions from a total of 18.239 instances. The error rate is 0,022. The false positive rate is 0,067 and the false negative rate is 0,018.

In case of Naive Bayes model, there are 16,750 instances with 0 value from which 97,37% were correct predicted, and 1,489 default cases, with 1 value, from which the model predicted correctly 98,52%. The total cost is 781,28.

The confusion matrix displays the model errors, as follows: NB model made 17.542 (16.055 + 1.487) correct predictions and 697 (695 + 2) wrong predictions from a total of 18.239 instances. The error rate is 0,038. The false positive rate is 0,001 and the false negative rate is 0,041.

In case of Support Vector Machines, there are 16,750 instances with 0 value from which 97,37% were correct predicted, and 1,489 default cases, with 1 value, from which the model predicted correctly 98,52%. The total cost is 462 u.m.

The confusion matrix displays the model errors, as follows: SVM model made 17.777 (16.310 + 1.467) correct predictions and 462 (440 + 22) wrong predictions from a total of 18.239 instances. The error rate is 0,025. The false positive rate is 0,014 and the false negative rate is 0,026.

In conclusion, we can say that in terms of cost and predictions, logistic regression showed both the lowest cost and the lowest error rate followed by a very small distance of Support Vector Machines, the error rates being almost similar.

3. Improving the model for default analysis

3.1. Attribute importance algorithm

Oracle Data Mining owns a specific characteristic named Algorithm Importance (AI) which is based on Minimum Description Length (MDL) algorithm that orders the attributes by significance in making predictions. Any system, method, or program using
AI to reduce time and computing resources necessary to build data mining models, goes through a process of selecting a subset of original attributes by eliminating redundant, irrelevant or informal ones and identify those attributes that can be extremely useful in making predictions.

The main idea behind the MDL principle is that "any regularity in the data can be used to compress the data, i.e. to describe it using fewer symbols than the number of symbols needed to describe the data literally." [3].

In this study, the objective is to build a predictive model in order to identify those borrowers with a high probability of default and to get a clear distinction between valuable customers of the bank and the less beneficial for the credit system. The first question we ask is: "what features should have a good payer customer? What represents him the best?". In order to receive a smart response, we apply the MDL algorithm on the following attributes:

![Figure 2 - Attributes on which the MDL algorithm is applied](image)

The result is summarized in the chart below and the table includes the attributes listed in order of their importance and significance:

![Figure 3 - The results after applying AI algorithm](image)
It is good to mention that the scoring points and the amount of liquidity held by the client in deposits (if any) are extremely important for credit decision making. We also believe that this information is very useful both individually and also for improving the predictive models built, especially Naive Bayes and Support Vector Machines.

3.2. Applying the algorithm after the attributes selection

After the selection of the most important attributes, we will continue applying data mining algorithms, the results being interpreted as follows:

For LG model, there were predicted correctly 98.88% out of 16.750 instances with 0 value (good payers) and 90.8% out of 1.489 default cases, afferent to 1 value. The total cost is 324 u.m. The confusion matrix displays the following errors: there are 137 false negative predictions and 187 false positive predictions. LG model made 17.915 (16.563 + 1.352) correct predictions and 324 (137 + 187) wrong predictions out of a total of 18.239 instances. The error rate is 0.017.

In case of NB, there were predicted correctly 95.81% out of 16.750 instances with 0 value and 99.93% out of 1.489 default cases, with 1 value. The total cost is 776,65 u.m. The confusion matrix displays the following errors: one false negative prediction and 702 false positive predictions. NB model made 17.536 (16.048 + 1.488) correct predictions and 703 (702 + 1) wrong predictions out of a total of 18.239 instances. The error rate is 0.038.

In case of SVM, there were predicted correctly 97.71% out of 16.750 instances with 0 value and 98.52% out of 1.489 default cases, with 1 value. The total cost is 406 u.m. The confusion matrix displays the following errors: there are 22 false negative predictions, and 384 false positive predictions. SVM model made 17.833 (16.366 + 1.467) correct predictions and 406 (22 + 384) wrong predictions out of a total of 18.239 instances. The error rate is 0.022.

3.3. Comparative analysis of the models after applying Attribute Importance algorithm

After applying the three models on the set of instances, we obtained a series of predictions for credits reimbursement. We have centralized these results in a virtual table that contains data about each customer (name, account, ID number), the real value of the RESTANŢIER attribute, predictions made by the three models and the likelihood of achieving these predictions (see fig. 4).

Figure 4 - Comparative results of predictions
Comparative results show that only 845 of 18 239 records are incorrect predictions, representing 4.63% of total. We also made a comparison of the incorrect predictions released by the three models. As we have noticed before, SVM registered 406 incorrect predictions (2.22%), NB recorded 703 incorrect predictions (3.85%), and LG registered only 324 wrong predictions (1.77%). From cost point of view, LG recorded the lowest cost, followed by SVM, and NB is detaching pretty much.

As a conclusion, on our observations, logistic regression showed the best results, but the other two models registered also small errors (less than 5%). We can consider that all three models can be successfully applied in banking practice.

4. Conclusions
Studies show that in recent years there have seen a dramatic explosion in the level of interest in data mining. The users wanted to take advantage of the tools offered by this technology in order to gain an intelligent competitive advantage. The management of financial risks has many dimensions and involves many types of decisions. The importance of this article comes from the complex issue of credit risk management in order to assure financial stability. Credit risk is considered the most dangerous category of banking risk and in order to prevent it, banks must meet a series of regulations and use different predictive models. Therefore, credit default is a major research area due to its effect. In this research various attribute selection and data mining techniques have been used to build a few predictive models for banking practice. It has been found that Logistic Regression performs very well in predicting the defaulters, followed closely by the Support Vector Machines. Further work is under progress to apply the results in terms of effects generated by the current international crisis.

Acknowledgements

This article is a result of the project POSDRU/88/1.5/S/55287 „Doctoral Programme in Economics at European Knowledge Standards (DOESEC)“. This project is co-funded by the European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies in partnership with West University of Timisoara. Also, this paper presents some results of the research project PN II, TE Program, Code 332: “Informatics Solutions for decision making support in the uncertain and unpredictable environments in order to integrate them within a Grid network”, financed within the framework of People research program.

References