A credit scoring analysis using data mining algorithms

Abstract: *Credit scoring is very important nowdays as it helps lenders to evaluate new credit applicants, it is an analysis through which banks can decide beforehand if a customer will be able to repay his debt, among with the interest, based on the historic data of former and present debtors. The purpose of this paper is to conduct a comparative study on the accuracy of classification models, the data base is formed of 150 individuals from a local bank in Romania, and there are used 18 dependent variables. The data mining algorithms selected for this comparison of credit scoring models are the following decision trees: CART, CHAID, Boosted Tree and Random Forest. The reason why the above mentioned models were selected in this study and not others is because in recent years decision trees were increasingly used to build credit scoring models, also their results can be easily interpreted and they can be applied on both categorical and continuous data. The results obtained in Statistica software indicate that Random Forest has higher accuracy rates and therefore outperformes the other proposed classification methods when it comes to distinguishing between good payers and bad payers.*

Introduction

With the continuous development of the financial and banking institutions, credit products play an increasingly important role in the economy. Economic globalization and the emergence of new channels through which services are provided, such as the internet, enables customers to search for and to select creditors without regional limitations. Because of this trend creditors must be willing and able to expand their business in other markets. Therefore, financial institutions are facing drastic global competition, prompting them to give more importance to the technology through which credit applications are analyzed.

The motivation of this study was the desire to conduct a credit scoring analysis for individuals, but not by using classical statistical methods, but by using data mining algorithms such as decision trees. The study was conducted by using the Statistica software, on a data base of 150 individual from a local bank in Romania, and its purpose is to encourage future studies to test the predictive performance of a new set of data mining algorithms.

The paper aims to answer the following research question:

- Which of the selected decision trees: CART, CHAID, Boosted Tree and Random Forest, can best distinguish between good payers and bad payers?

The content of this paper is as follows: the first section provides a formal definition of credit risk and the methods used to estimate it, section 2 describes the literature focused on credit scoring, section 3 presents a short description of the data base and of the data mining algorithms used in this study, section 4 presents the experimental results and section 5 gives remarks and provides a conclusion.

1. Credit Risk

Credit risk should be seen as the volatility of earnings of a bank. This volatility is the result of financial losses recorded by a financial institution, due to the borrower's failure to repay a loan or otherwise meet a contractual obligation.

Twenty years ago most financial institutions relied almost exclusively on subjective analysis or so-called banker expert systems to assess the credit risk on corporate loans. Essentially, bankers used information on various borrower characteristics, such as borrower character (reputation), capital (leverage), capacity (volatility of earnings) and collateral, to reach a largely subjective judgement as to whether or not to grant the credit.

Due to the fact that bankers tend to be overly pessimistic financial institutions have moved towards systems that are more objectively based. Among these are: univariate accounting based credit-scoring systems, where the decision-maker from a financial institution compares various key accounting ratios of potential borrowers with industry or group norms, and multivariate models, where key accounting variables are combined and weighted to produce either a credit risk score or a probability of default measure. There are at least four methodological approaches to developing multivariate credit-scoring systems: the linear probability model, the logit model, the probit model, and the discriminant analysis model.

2. Literature review

Credit scoring can be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent (Mester, 1997). The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively.

Credit scores help to reduce discrimination because credit scoring models provide an objective analysis of a consumer’s creditworthiness. This enables credit providers to focus on only information that relates to credit risk and avoid the personal subjectivity of a credit analyst (Fensterstock, 2005). They also help to increase the speed and consistency of the loan application process and allows the automation of the lending process (Rimmer, 2005). As such, it greatly reduces the need for human intervention on credit evaluation and the cost of delivering credit (Wendel, C., and Harvey, M., 2003). With the help of the credit scores, financial institutions are able to quantify the risks associated with granting credit to a particular applicant in a shorter time. Leonard’s (1995) study of a Canadian bank found that the time for processing a consumer loan application was shortened from nine days to three days after credit scoring was used. The time saved in processing the loans can be used to address more complex issues.

Most credit scoring articles focused on enterprise credit score: using audited financial accounts variables and other internal or external, industrial or credit bureau variables, the enterprise score is extracted, rather than individual (consumer) credit score: the individual credit score uses variables like applicant age, marital status, income and some other variables and can include credit bureau variables.

Other articles on credit scoring compare data mining algorithms in order to identify the best performing model in terms of predictive capacity. Therefore, Yobas, Crook and Ross (2000) concluded that the linear discriminant analysis is superior to the genetic algorithm, and neural networks have a lower predictive ability than the linear discriminant analysis.

Lee (2006) demonstrated the effectiveness of credit scoring models using two algorithms: CART (Clasiffication and Regression Tree) and MARS (Multivariate Adaptive Regression Splines). The results of this study show that CART and MARS models are more efficient in terms of predictive ability than other traditional methods such as discriminant analysis, neural networks, logistic regression and SVM (Support Vector Machine).

Such as the articles mentioned above this paper aims to compare four data mining algorithms: CART, CHAID, Boosted Tree and Random Forest, in order to determine which one can best discriminate between good payers and bad payers.

3. Methodology

3.1 Data Preparation

Data is the main resource for data mining, therefore it should be prepared properly before applying any data-mining tool. Otherwise, it would be just a case of Garbage-In Garbage-Out (GIGO). Since major strategic decisions are impacted by these results, any error might give rise to huge losses. Thus, it is important to preprocess the data and improve the accuracy of the model so that one can make the best possible decision.

The following aspects of the data were noted during this stage:

- There are no outliers in the data;
- There are no missing values in the data;
- There are no redundant predictors;
- There are no duplicate values;
- There are no invariant data.

The next step is to split the original data set into two subsets: 30% of cases were retained for testing and 70% of cases were used for model building. Next, by using Stratified Random Sampling method an equal number of observations for both good and bad risk customers was extracted.

In order to reduce the complexity of the problem, the data set can be transformed into a data set of lower dimension. The Feature Selection and Variable Screening tool available in Statistica automatically found important predictors that clearly discriminate between good and bad customers, therefore the number of predictive variables is reduced from 18 to 10.

3.2 Data mining algorithms

These predictors determined above will be further examined using a wide array of data mining and machine learning algorithms such as:

- The CART tree is a non-parametric approach and consists of several layers of nodes: the first layer consists of a root node and the last layer consists of leaf nodes. Because it is a binary tree, each node (except the leaf nodes) is connected to two nodes in the next layer. In addition, CART is often able to uncover complex interactions between predictors which may be difficult or impossible using traditional multivariate techniques. It can be helpful if there are a lot of variables, as they can be used to identify important variables and interactions.

- CHAID: The acronym CHAID stands for Chi-squared Automatic Interaction Detector. This name derives from the basic algorithm that is used to construct (non-binary) trees, which for classification problems (when the dependent variable is categorical in nature) relies on the Chi-square test to determine the best next split at each step. CHAID output is visual and easy to interpret.
- Boosting Trees, the general idea is to compute a sequence of very simple trees, where each successive tree is built for the prediction residuals of the preceding tree, this method will build binary trees.
- Random Forest is a trademark term for an ensemble of non-binary decision trees. Unlike single decision trees which are likely to suffer from high variance (depending on how they are tuned) Random Forests use averaging to find a natural balance between the two extremes.

3.3 Model comparison

It is good practice to experiment with a number of different methods when modeling or mining data rather than relying on a single model for final deployment. Different techniques may shed new light on a problem or confirm previous conclusions. Therefore, the four model mentioned in the previous chapter will be deployed on the validation set.

3.4 Data base presentation

The example data set used in this case contains 150 cases and 18 variables (or predictors) with information pertaining to past and current customers who borrowed from a Romanian bank. The data set has a distribution of 70% credit worthy (good) customers and 30% not credit worthy (bad) customers. Customers who have missed a payment can be thought of as bad risks, and customers who have missed no payment can be thought of as good risks.

Following is the complete list of variables used in this data set:

- Basic Personal Information: age, sex;
- Family Information: marital status, number of dependents;
- Residential Information: years at current address, type of apartment;

- Employment Status: years in current occupation, years in work experience, occupation;
- Others: number of credit cards, car owner or not, studies, net income, number of guarantors, number of active credit facilities, bank exposure.

4. Results

The obtained results are summarized in the following two tables:

Table 1

Classification rates

Modele	Good payers	Bad payers	Overall predictive accuracy
CART	96.55%	100%	98.27%
CHAID	96.55%	65.51%	81.03%
Boosted Tree	82.14%	93.33%	87.93%
Random Forest	89.65%	96.55%	93.10%

Table 2

Type I and type II errors

Modele	Type I error	Type II error
CART	3.45%	0%
CHAID	3.45%	34%
Boosted Tree	17.86%	7%
Random Forest	10.35%	3%

Deciding which algorithm is the most efficient in terms of predictive ability based on the information summarized above is not indicated, the comparison must be made on a new set of data, the validation sample put aside at the beginning.

For a financial institution it is particularly important to correct classify the bad payers, as granting them credit facilities my cause financial loss. Thus, the following graphs have been generated in order to select the model with the highest predictive accuracy.

- Gains Chart

The gains chart, **ANNEX 1**, provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting categorical dependent variable. Specifically, the chart summarizes the utility that one can expect by using the respective predictive models, as compared to using baseline information only.

This chart depicts that the Random Forest with Deployment model is the best among the available models for prediction purposes. For this model, if we consider the top three deciles (after sorting based on the confidence of prediction), we would correctly classify approximately 85 percent of the cases in the population belonging to category "bad." The baseline indicates the expected result if no model were used at all.

Corresponding values of Gains can be computed for each percentile of the population (in this case loan applicants sorted based on the confidence level of prediction) to determine the percentile of cases that should be targeted to achieve a certain percentage of predictive accuracy. You can see from the above graph that the gains values for different percentiles can be connected by a line and it will typically ascend slowly and merge with the baseline if all customers (100%) were selected.

- Lift Chart

This chart is visible in **ANNEX 2**, if we consider the top three deciles, we would end up with a sample that has almost 2.83 times the number of 'bad' customers when compared to the baseline model. In other words, the relative gain or lift value by using Random Forest model is approximately 2.83.

- ROC curve

The receiver operating characteristic (ROC) chart, **ANNEX 3**, graphically displays sensitivity (percentage of the defaulters predicted correctly as defaulters) versus 1-specificity (percentage of the non-defaulters wrongly classified as defaulters), or the ratio of the true positive rate versus the false positive rate. It can be seen that there are four lines in the ROC graph. The straight line is the baseline and every decision tree is represented by a separate line. The baseline indicates that at each point of the line, the percentage of the defaulters predicted correctly as defaulters (true positive rate) is equivalent with the percentage of the non-defaulters wrongly classified as defaulters (false positive rate). The ROC charts show that the Random Forest scorecard model outperforms the other decision tree models, as the area under the curve is higher in the case of this algorithm (0,97).

5. Conclusions

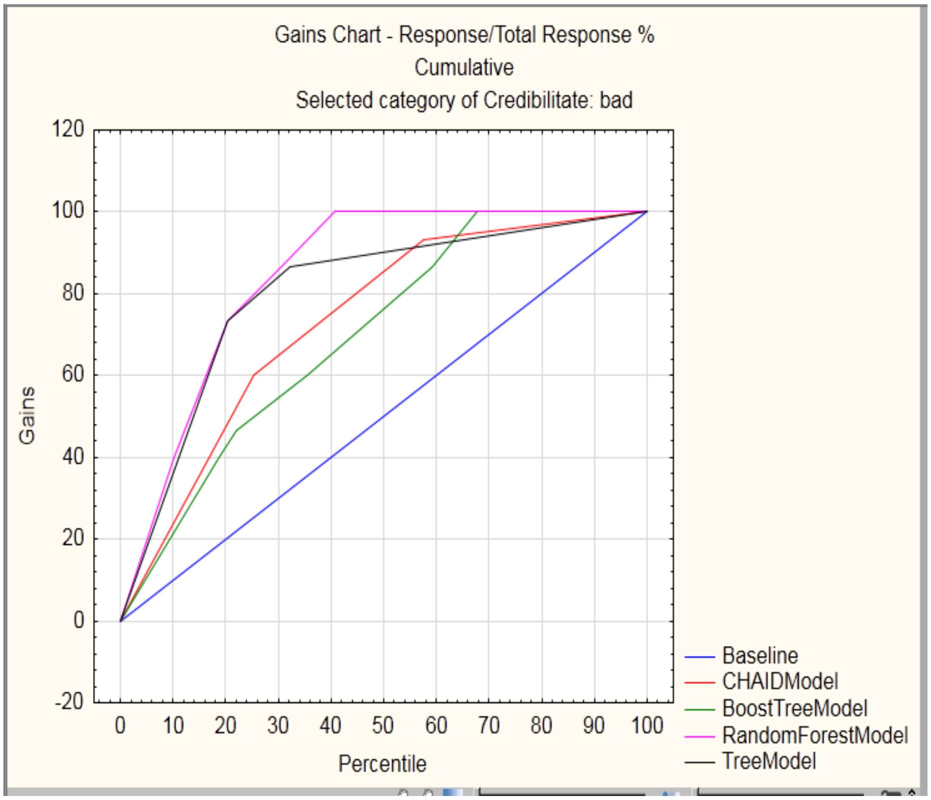
In this section there is formulated an answer to the questions that were raised at the beginning of the paper.

- **Which of the selected decision trees: CART, CHAID, Boosted Tree and Random Forest, can best distinguish between good payers and bad payers?**

Based on the Gains, Lift charts and ROC curve we can conclude that that the Random Forest algorithm can be further applied to a new set of data, in order to make a prediction regarding their status of good or bad payers, depending on the dependent variables introduced.

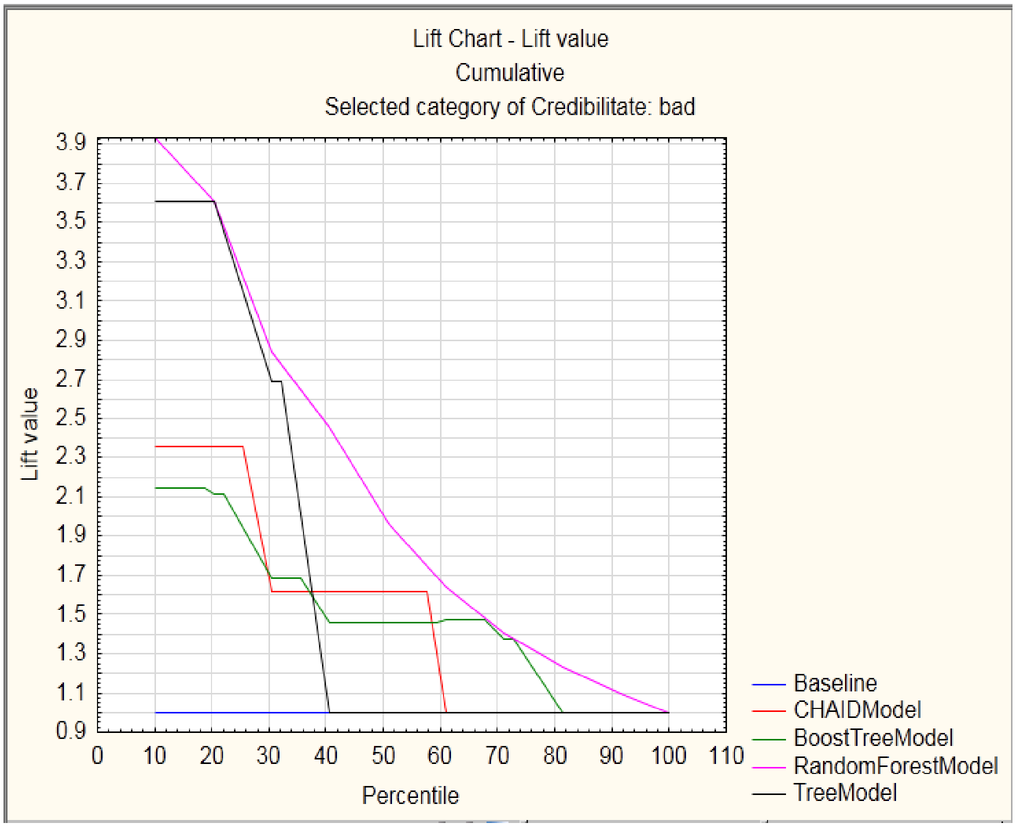
Annex 1

Gains Chart



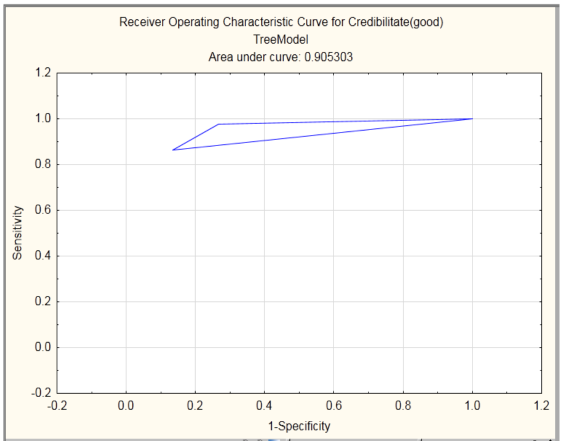
Annex 2

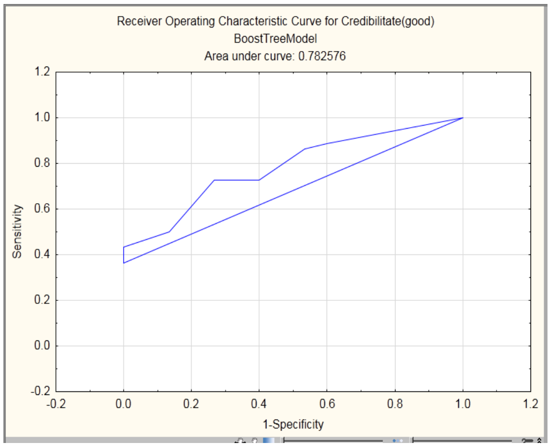
Lift chart

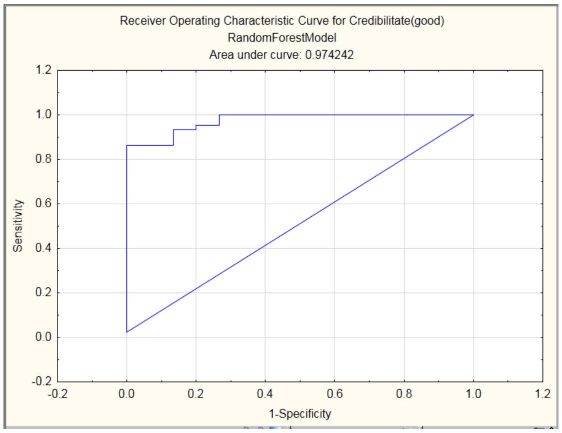


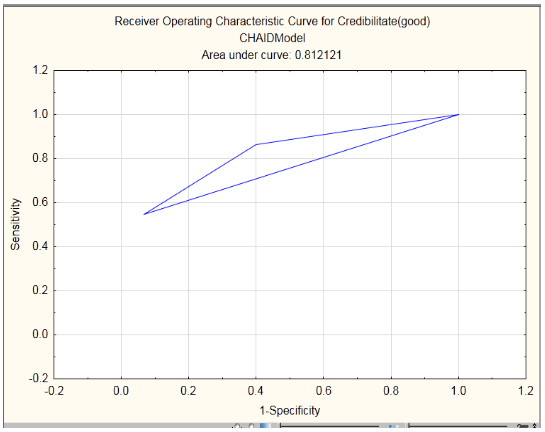
Annex 3

ROC curve









REFERENCES

1) **Fensterstock, A. (2005)**, *Credit scoring and the next step*, Business Credit 107(3) 46-49.
2) **Lee, T. S., Chiu, C.C., Chou, Y.C., Lu, C.J., (2006)**, *Mining the customer credit using classification and regression tree and multivariate adaptive regression splines*, Computational Statistics & Data Analysis 50 1113 – 1130
3) **Leonard, K.J. (1995)**, *The development of credit scoring quality measures for consumer credit applications*, International Journal of Quality and Reliability Management 12(4) 79-85.
4) **Mester, L.J. (1997)**. *What's the point of credit scoring?*, Business Review (September) 3-16.
5) **Rimmer, J. (2005),** *Contemporary changes in credit scoring*, Credit Control 26(4) 56-60.
6) **Wendel, C., și M. Harvey. (2003),** *Credit scoring: Best practices and approaches*, Commercial Lending Review 18(3) 4-7
7) **Yobas, M.B, Crook, J.N., and Ross, P. (2000*)***, *Credit scoring using neural and evolutionary techniques*, IMA Journal of Mathematics applied in business and industry 11, 111-125.

***http://www.statsoft.com/Textbook/Basic-Statistics