The International Scientific Conference INPROFORUM 2015, *November 5 - 6, 2015, České Budějovice*, 166-170, ISBN 978-80-7394-536-7.

Predicting Bonity of Clients through Two Recursive Partitioning Methods

Michael Rost, Renata Klufová

Abstract: *The aim of the paper is to show and compare some classical classification approach with not so classical and typical classification method for identification of the factors influencing the credit scoring of some bank customers. For this purpose, we used statistical methodology like classification and regression trees (CART) and recursive partitioning method called PARTy. These classification methods are both not parametric methods and their application is not restricted by a strong normality assumption like LDA. Our approach is demonstrated on segment of the data coming from bank institution. Data pre-processing and the numerical computation were carried out in the programming language R.*

Key words: classification· credit scoring · CART · PARTY

JEL Classification: C38

1 Introduction

In these days every bank institution evaluates their customers – applicant – for bank loan. For this purpose they gathered and sore gigabytes of information about their customers. Consequently they construct predictive models with aim to uncover a potential risk linked with particular loan applicant. For this purpose bank institution usually use data mining methods like CART (Breiman, Friedman, Olshen, & Stone, 1998), ANN (Hastie, Tibshirani, & Friedman, 2001), and random forest (Hastie, Tibshirani, & Friedman, 2001) or some form of SVM (Vapnik, 2000) or their combinations see commercial software K-XEN.

Some of these methods provide insight into "structure" of such scoring process. From this point of view is very convenient methodology of classification trees which provide „simple" rules and decision tree providing deeper insight into the own decision process. In this paper we implement two classification methods, more concretely classification and regression trees methodology (CART) and more novel approach called recursive partitioning (PARTy).

2 Data and methods

All data, used in this paper, are collected from one bank institution in Czech Republic during the 2013 - 2014. Data could be characterized as random sample from bank internal database. This data set contains information from 2455 credit applicants. Initially there are nine different variables in this data set: Age (A), Income (I), Gender (G), Marital (Ma), Number of kids (Nk), Number of cards (Nc), Mortgage (Mo), Number of loans (L) and Risk for bank (R). In the table 1 there are provided basic descriptive statistics for these variables.

At this point we have to say that we apply discretization during the pre-processing of data set of following variables: age, income, number of kids, number of cards and number of loans. The detailed categories of particular variables is as follows: age with categories: a = [0,26), b = [26,51) measured in years; income with categories: a = (0,20], b = c = (20,40], d = (40,60] measured in thousand CZK; gender with categories: F – female, M –male; marital status with categories: divorced-separated; widowed; married and single; number of kids in family with categories: a = [0,1); b = [1,3); c = [3,5); number of debt or credit cards with categories: a= [0,1), b = [1,3) and c = [3,7); mortgage with categories: no, yes; loans with categories: a = [0,1), b = [1,4) and finally risk for bank with categories: bad loss; bad profit; good risk given by bank risk managers of bank institution.

Ing. Michael Rost, Ph.D., University of South Bohemia in České Budějovice, Faculty of Economics, Department of Applied Mathematics and Informatics, Studentská 13, České Budějovice, e-mail: rost@ef.jcu.cz

RNDr. Renata Klufová, Ph.D., University of South Bohemia in České Budějovice, Faculty of Economics, Department of Applied Mathematics and Informatics, Studentská 13, České Budějovice, e-mail: klufova@ef.jcu.cz

Table 1 Descriptive characteristic for used explanatory variables – before transformation

Numerical variables	Descriptive characteristics						
Name	Min.	The first quartile	Median	Mean	The third quartile	Max.	
Age (A)	18	23	31	31.89	41	50	
Income (I)*	15.018	20.559	23.620	25. 653	27. 618	59. 944	
Categorical variables	Categories and numbers of observation						
Gender	Female: 1228			Male: 1227			
Marital	Married: 1214		Single:711		Divorced or separated, widowed: 530		
Number of kids	0: 587	1: 789	2: 934	3: 150	4: 31		
Number of cards	0:325	1:651	2:661	3:367	4:121	5: 172	6:158
Mortgage	No: 546			Yes: 1909			
Risk for bank	Bad loss: 559		Bad profit: 1475		Good risk: 421		

Source: Own processing *Income is provided in thousands of CZK

2.2 Methods

With aim to uncover classification rules we used two methods based on "recursive partitioning" of the space. These methods could be briefly described as follows. During the building the set of classification rules by CART methodology, e.g. tree growing, we usually employ the following phases (Breiman, Friedman, Olshen, & Stone, 1998) or (Hastie, Tibshirani, & Friedman, 2001):

- The split criterion for each node of growing tree is chosen. This problem is usually solved by impurity measure. As an impurity measure, we chose the Gini index. Other possibilities are for example Misclassification error, Cross-entropy or deviance. For more technical details see (Breiman, Friedman, Olshen, & Stone, 1998) or (Therneau, &Atkinson, 2011)).
- To decide which node becomes a leaf (terminal node of tree) is an essential problem solved in the second stage. Usually pruning of tree is used. After building T_{max} tree (each leave contains the objects only from one class, or the number of classifying objects in each leave is smaller than the prescribed value) is this tree pruned to tree T_{optim}. The new tree T_{optim} is the subset of $_{Tmax}$. This pruning of tree T_{max} to tree T_{optim} minimize estimation of relative error of classification.
- The third phase is the simplest part of the tree growing process. Each of the classes is assigned to one of the leaves. The idea is following: to assign correctly the specific class to leaves is to assign the value, that minimizing the estimate of the misclassification error. More information about CART methodology can be found in (Breiman, Friedman, Olshen, & Stone, 1998) or in (Hastie, Tibshirani, & Friedman, 2001). As we can see, the major advantage of the recursive binary tree is its nice interpretability. The whole feature space partition is fully described by one tree.

The algorithm of Hothorn et al. (2006b) for binary recursive partitioning can be described in three steps:

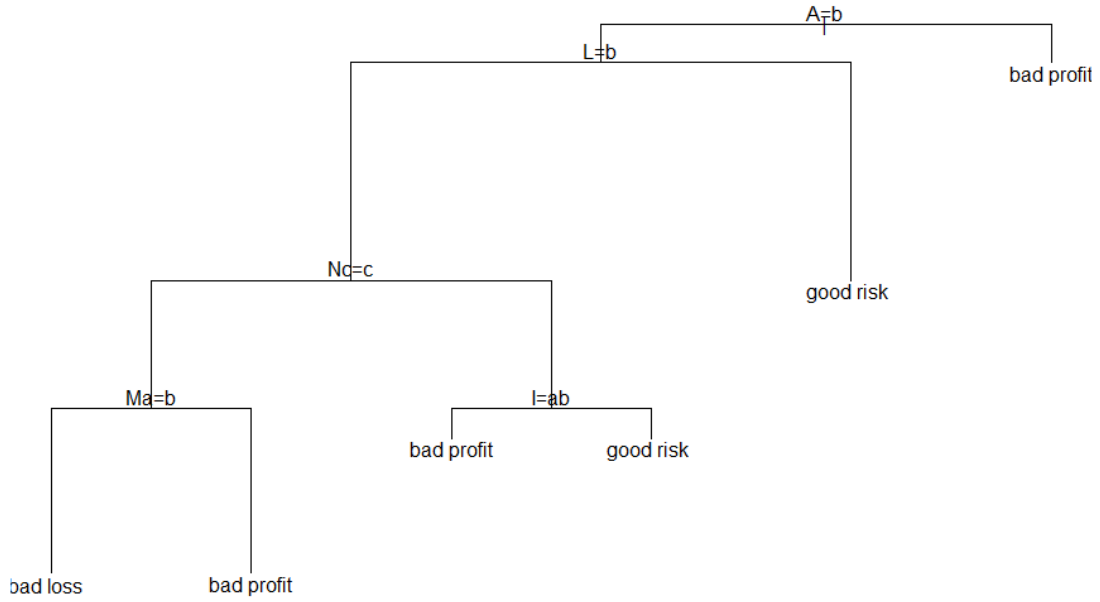
- Firstly, beginning with the whole sample, the global null hypothesis that there is no relationship between any of the covariates and the response variable is evaluated. If no violation of the null hypothesis is detected, the procedure stops. If, however, a significant association is discovered, the variable with the largest association is chosen for the split.
- Secondly, the best cutpoint in this variable is determined and used to split the sample into two groups according to values of the selected covariate.
- Then the algorithm recursively repeats the first two steps in the subsamples until there is no further violation of the null hypothesis, or a minimum number of observations per node is reached.

For more technical details about algorithm see (Hothorn, Torsten, Hornik, & Zeileis, 2006b). Data pre-processing and the numerical computation were carried out in the programming language R (R package, 2011).

3 Research results

At the beginning of the building classification rules, the sufficiently branched tree T_{max} was created. To manage the growing process, the complexity parameter *cp* was specified by zero value because the low value of the complexity parameter made the tree sufficiently branched. Visualization of this tree is proposed on picture 1.

Figure 1 Unpruned classification tree obtained by CART algorithm with setting complexity parameter *cp* = 0.



Source: Own processing

Age (A), Income (I), Marital (Ma), Number of cards (Nc), Mortgage (Mo), Number of loans (L).

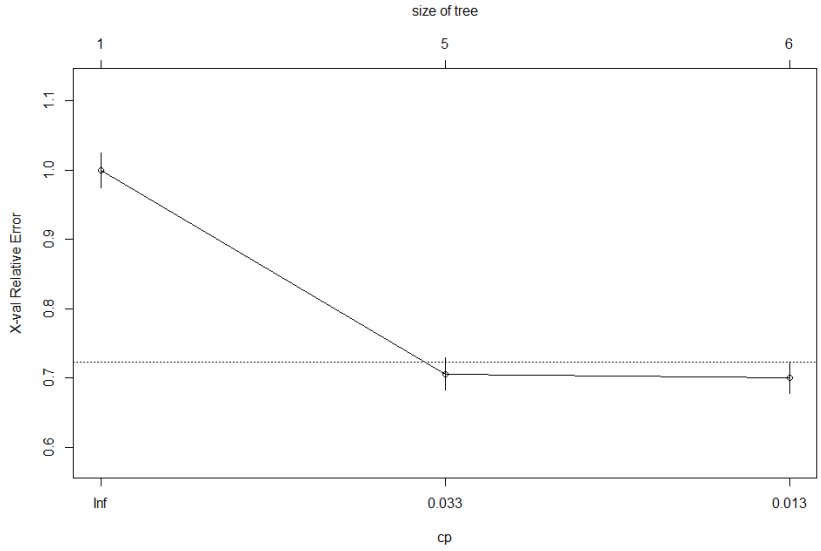
For decision where to stop the growing process and to prune the tree, we used 1-*SE* rule, see table 2 for more details and characteristics. These characteristics are also provided in graphical form on graph 2.

Table 2 Complexity table

Root node error: 980/2455 =0,39919					
	cp	Number of splits	Relative error	X-error	X-std
1	0.062755	0	1.00000	1.0000	0.024760
2	0.017347	4	0.70612	0.70612	0.022747
3	0.010000	5	0.68878	0.68878	0.022574

Source: Own processing

Figure 2 Graphical visualization of the relative error vs. size of tree and complexity parameter

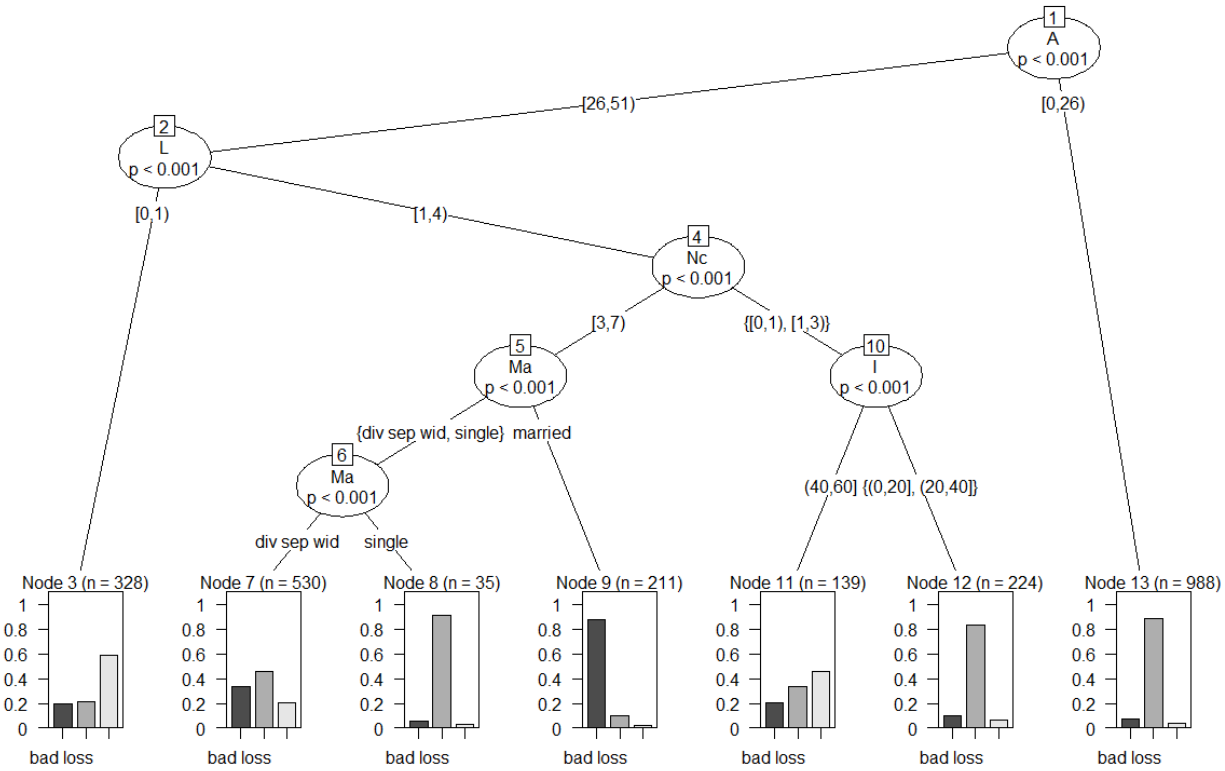


Source: Own processing

From this is clearly evident that if we apply 1-*SE* rule we cannot prune T_{max} classification tree. So the final classification tree obtained by CART methodology is T_{max} tree, e.g. tree with five splits and six terminal nodes.

It is evident that "driving" variables discovered by CART methodology are: age (A), number of loans (L), number of cards, marital (Ma) and income (I). On picture 3 is provided final classification tree obtained by PARTy methodology. We can see that this tree has seven terminal nodes. From this picture is obvious that classification rules are based only on following variables: age (A), loans (L), number of cards (Nc), marital (Ma) and income (I). So the same variables as identified in the previous tree. So we can say that partitioning of the sample space is stable from this point of view.

Figure 3 Resulting classification tree obtained by PARTy algorithm



Source: Own processing

The efficiency of the two mentioned classification approaches is evaluated on training data set. The results are presented in error table in table 3. In the case of classical approach (CART methodology) we reached classification efficiency of 72.50 % for training data. The classification efficiency for PARTy approach reached exactly the same value for test data. For more details see table 3.

Table 3 Error classification table for two methodologies applied on training and test data

Truth	**Prediction**			**Sum**
	Applicant is classified as: "bad loss"	Applicant is classified as: "bad profit"	Applicant is classified as: "good risk"	
Applicant is "bad loss"	Training data set: **186**[a] ; **186**[b]	Training data set: **280**[a]; **280**[b]	Training data set: **93**[a] ; **93**[b]	559; 559
Applicant is "bad profit"	Training data set: **21**[a] ; **21**[b]	Training data set: **1337**[a] ; **1337**[b]	Training data set: **117**[a] ; **117**[b]	1475; 1475
Applicant is "good risk"	Training data set: **4**[a]; **4**[b]	Training data set: **160**[a]; **160**[b]	Training data set: **257**[a] ; **257**[b]	421; 421
Sum	211; 211	1777; 1777	467; 467	2455

[a]Results for classical approach through CART methodology; [b]Results for PARTy methodology

4 Conclusions

Using the above mentioned methodology, we identified some essential factors influencing the classification of customers. From the both classification methodology we identified the same influencing variables. More concretely we identified following variables: Age, Income, Marital, Number of cards (Nc), Mortgage (Mo), Number of loans (L).

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, Ch. J. (1998). *Classification and Regression trees.* Chapman & Hall/CRC, Boca Raton, 359, ISBN: 0-412-04841-8

Hastie T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning, Data Mining, Inference and Precision*. Springer, New York, ISBN 0-387-95284-5

Vapnik, V. N., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 264–280.

R package, version 3.1-50 (online). Retrieved 2011-10-20, from http://CRAN.R-project.org/package=rpart

Sewell, M. (2008). *Structural Risk Minimization* (online), Department of Computer Science University College London, London 2008. Retrieved April 3, 2013, from http://www.svms.org/srm/srm.pdf

Therneau, T. M., & Atkinson, B, R port by Brian Ripley (2011). *rpart: Recursive Partitioning*.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory. Information Science and Statistics*. Springer-Verlag. ISBN 978-0-387-98780-4.

Hothorn, Torsten, Kurt Hornik, & Achim Zeileis. (2006b). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics,* 15(3), 651–674.