Multidiszciplináris tudományok, 13. kötet. 3. sz. (2023), pp. 212-220 **https://doi.org/10.35925/j.multi.2023.3.21**

CLASSIFICATION OF RETAIL LOANS USING DECISION TREE

Kitti Fodor
assistant lecturer, University of Miskolc, Institute of Economic Theory and Methodology
3515 Miskolc-Egyetemváros, e-mail: kitti.fodor@uni-miskolc.hu

Abstract
While there is extensive literature on the prediction of corporate bankruptcies, there is little literature on the classification of retail borrowers. There are several ways to analyse the data, which may yield different results. In this paper, my aim is to predict the default of household loans using decision tree. I found one significant explanatory variable, which was the ratio of the repayment to the contract amount. For my analysis I used two samples with different compositions. Both have high classification accuracy. Overall, the second model is better, with a classification accuracy of 84.4%.

Keywords: *loan default, decision tree, classification, different sample types*

1. Introduction

It is important for financial institutions to lend to customers with a low risk of non-repayment. However, it is difficult to identify which customers become defaulters. This is evidenced by the fact that banks have a credit assessment method, but there are still many non-performing loans registered in our country.

There are a lot of research on predicting corporate bankruptcies over the last 100 years, and I have based my own research on this. In this research, my goal is to examine the confidence with decision tree can categorize defaulted loans, which variables are significant among the data recorded by the KHR, and to examine the effect caused by the change in the sample composition.

2. The theoretical basis of the research

The first step of my research was to study the Hungarian and international literature. This is briefly summarised in this chapter. I believe this will contribute to the understanding of my research and its results.

2.1. A short history of lending

Lending has been part of human culture for thousands of years. The first rules date back to Hammurabi, who laid down rules for lenders and borrowers on stone tablets (Fekete, Tatay, 2012) In ancient times, non-payment of loans was punishable by serious consequences, including death. In the Middle Ages, the Council of Nicea imposed an interest ban, and this resistance persisted until the Reformation. In the 15th century, the idea was born that money should be circulated for the benefit of the economy. In the 1600s and 1800s, loans were mainly granted by wealthy landowners and citizens. In the continental countries, we can speak of lending from the 19th century. (Vértesy, 2008)

By 2004, millions of Hungarians had credit, banks were competing for lenders, which led to a steady easing of conditions. Even before the crisis, there were signs that financial awareness and financial

literacy among the Hungarian population was not favourable, which is a dangerous factor in terms of lending, and the lack of this knowledge can even lead to a debt trap. This was confirmed by the 50% increase in the number of non-performing loans in 2004.

By the last quarter of 2022, there was a significant decline in retail lending, with an 18% drop in personal loans and a 54% drop in housing loans compared to the same period a year earlier (MNB, 2023).

2.2. Bankruptcy models

Bankruptcy forecasting research does not yet have a 100-year history. The first attempts were made in the 1930s. The first real model was created by Altman, who built his model on 5 financial indicators that could predict insolvency with 95% confidence. A few years later, an extended seven-variable model was developed based on this model (Altman, 1968; Virág, 2004). Altman's models were not representative, and the sample included roughly equal proportions of surviving and failing firms. (Ohlson, 1980) The next novelty was the emergence of recursive partitioning algorithms, which dates back to the mid-1980s. Among the first adopters of this method were Altman, Frydman and Kao. The classification accuracy of the model was 94%, but there was a significant difference in the correct categorisation between surviving and failed firms (Frydman et al., 1985).

In the 2000s, McKee-Greenstein also attempted to carry out analyses using this method, but in the end the use of recursive partitioning algorithms did not spread in the literature (McKee, Greenstein, 2000).

3. Methods of predicting bankruptcy

For bankruptcy forecasting, the following methods are widely used:

- discriminant analysis
- logistic regression
- decision tree
- neural network

For this study I used only the decision tree.

3.1. Recursive partitioning algorithms

The methodology goes by several names, most commonly referred to simply as the decision tree. This analysis is also one of the classification methods. The resulting subgroups are called nodes. The basis for the prediction is the leaves, which are the part of the tree that is not further divided (Hajdú, 2008).

Initially, the method was only applied to categorical dependent and independent variables, but over time it was extended to metric variables. The aim is to minimize the within-group variance so that the between-group variance is as large as possible.

Its use in bankruptcy prediction dates back to the 1980s. The method combines univariate and multivariate analyses, as 1-1 splitting is done by one variable, but overall it includes more variables in the analysis. At each step, the algorithm tries to reduce misclassifications. The algorithm is an iterative process designed specifically for computers.

There are several types of decision trees, of which the following have a clear methodology:

- CHAID: Chi-squared Automatic Trees
- Exhaustive CHAID
- C&RT: Classification and Regression Trees

- QUEST: Quick, Unbaised, Efficient Statistical Trees

The great advantage of the analysis is that there is no restriction on the variables included, both metric and non-metric variables can be included.

The process consists of three main steps: merging, splitting and stopping.

Merging means "For each explanatory variable, for the dependent variable, it means combining statistically independent, or more precisely, the least statistically related categories" (Hámori, 2001: 704).

For a given explanatory variable, it examines all possible pairings and uses Pearson's $\chi 2$ test to examine the probability that the outcome variable categories and the pairs of categories of the explanatory variable are independent for different pairings, and then the algorithm finds the case with the highest p value and compares it to the threshold for merging. It does this until the highest p value is less than the merge threshold, at which point the loop stops. The process is repeated for each explanatory variable.

Splitting is "the partitioning of observations into categories of explanatory variables that are considered to be the least independent with respect to the dependent variable" (Hámori, 2001: 704).

The explanatory variable with the smallest p-value is selected, and this value is compared to the allocation threshold. If smaller, the split is created and repeated until the sub-dataset cannot be split any further.

Stopping means "The algorithm continues to recursively merge categories and split cases until it reaches a predefined stopping criterion". (Hámori, 2001: 704)

The split-divide cycle is repeated until some stopping criterion is reached:

- p exceeds the allocation threshold
- no difference between cases for explanatory or outcome variables
- the number of elements in the sub-database is less than the predefined number of cases
- the maximum depth of the tree is reached.

The process described above results in a tree, an example of which is shown in the figure below:

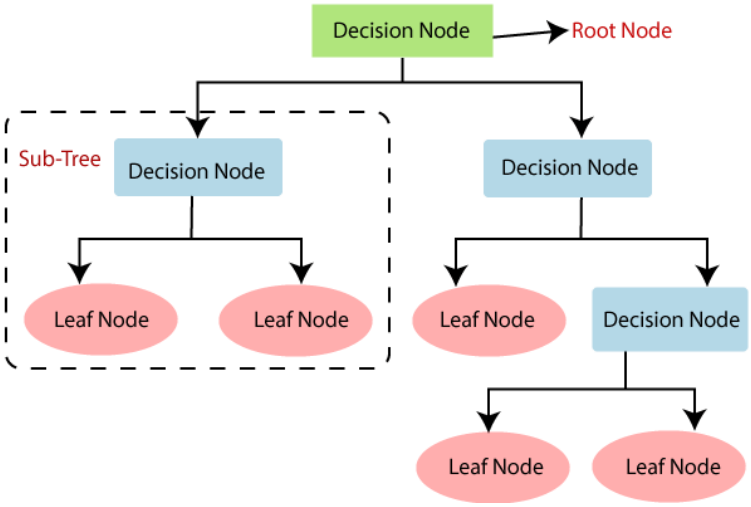

Figure 1. *Typical decision tree*

Source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

An argument in favour of this method is that the conditions do not include a normal distribution of variables. It is easiest to apply when there are binary separations. As a result, a high proportion of the population is assigned the appropriate solvency classification, the exact classification data can be found in the classification matrix.

The disadvantage of this methodology is that it cannot be used for forecasting purposes, as it is mostly specialized for the training database. However, the problem can be solved by using the method developed to control over-learning in artificial intelligence models, i.e. by dividing the data into a training and a testing part and examining whether similar results are obtained in both cases (Hámori, 2001).

4. Database

In Hungary, information on household creditors is kept by the Central Credit Information System, or KHR, which helps banks to share information on creditors, assist in credit assessment and reduce the risk of over-indebtedness. The KHR maintains a so-called complete list, i.e. customers who meet their obligations on time are also included in the register.

The trend in defaults for 2021 shows that the number of defaults has steadily decreased over the year, with the number of outstanding debts falling by 13.9% from January to December. The trend in outstanding debts has also been influenced by legislative changes, such as the gradual reduction of the moratorium on repayments.

In terms of duration of defaults, 12.21% of the outstanding defaults have been outstanding for up to one year, 6.4% for less than 720 days and a significant proportion, 81.39%, for more than almost 2 years (KHR Annual Information, 2021).

The necessary database for the analyses was provided by BISZ Zrt. The data were extracted on 30 September 2021, so the database contains the persons registered on that date. A unit in the database represents one loan transaction, so there may be persons in the database who are listed more than once with different loan transactions. Overall, on that date, the register contained 10,767,452 credit transactions and 21 variables. In addition to the original variables, I added more variables to the database. For the analysis the relevant variables are:

- default
- age
- gender
- loan maturity
- repayment amount as a percentage of contract amount

Before starting the analyses, the first step was to clean the database and narrow it down to the research objectives; after that I had 2,887,470 cases in the database. For the analysis I used 2 database with 500 cases. For the sampling I used a random numbers generator. For the first sample, I used simple random sampling. This is a type of representative sampling. For the second sample, I also used random sampling, but in this case the proportion of performing and non-performing loans is the same. The second sample type is a good and applied practice in this area.

I classified as default the loan transaction that had a default amount.

5. Empirical research

5.1. Decision tree I.

Before starting the analysis, it is important to note that one of the disadvantages of the decision tree is its tendency to over-learn, which is also a risk in this case, as the sample is predominantly composed of good performing loans (93%).

In the case of the decision tree, the algorithm had four explanatory variables, of which the ratio of the repayment to the contract amount proved to be a good discriminating variable based on the algorithm. The decision tree run on the training and test sample is shown in *Figure 2*.

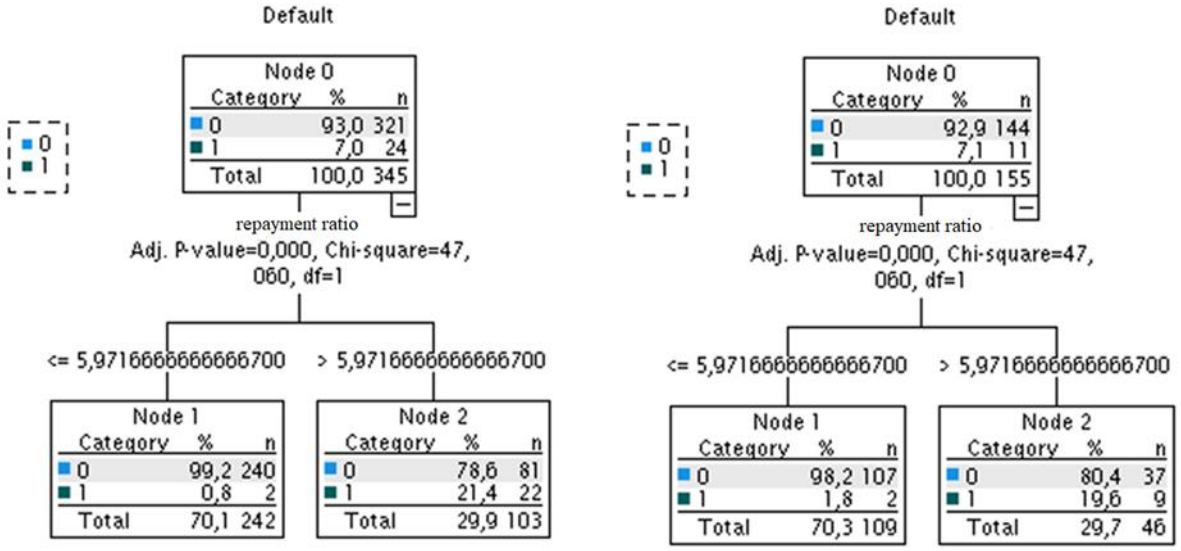

Figure 2. *Decision tree for the training and test sample*
Source: SPSS output, own editing

This decision tree consisted of a level 0 and a level 1. Level 0 shows the entire database in one view and the distribution and item number of each category of the dependent variable. This is followed by an iterative process, the algorithm performs the analysis for each explanatory variable and then selects the one that has the greatest influence. In this case, this variable is the ratio of the repayment to the contract amount. If the algorithm then finds more significant variables, the tree is extended by additional levels, if not, the tree ends at that level.

It can be seen that in the case where the value of the variable is less than 5.9717, the number of non-performing loans is negligible.

Information on the accuracy of the classifications is provided by the classification matrix.

For the training database, the model achieved a classification accuracy of 93.0%, but did not correctly categorise any of the non-performing loans. This is because the number of non-performing loans was too low in the sample, so the algorithm overestimated the classification of performing loans. A solution to this problem could be to design a sample with (approximately) equal proportions of performing and non-performing loans.

Table 1. *Classification matrix*

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	312	0	100.0%
	1	24	0	0.0%
	Overall Percentage	100.0%	0.0%	93.0%
Test	0	144	0	100.0%
	1	11	0	0.0%
	Overall Percentage	100.0%	0.0%	92.9%

Source: Own editing

5.2. Decision tree II.

Again, the algorithm was based on the same four explanatory variables, and in this case the same variable was found to be significant as in the first case, shown in *Figure 3*. This variable is the ratio of the repayment to the contract amount.

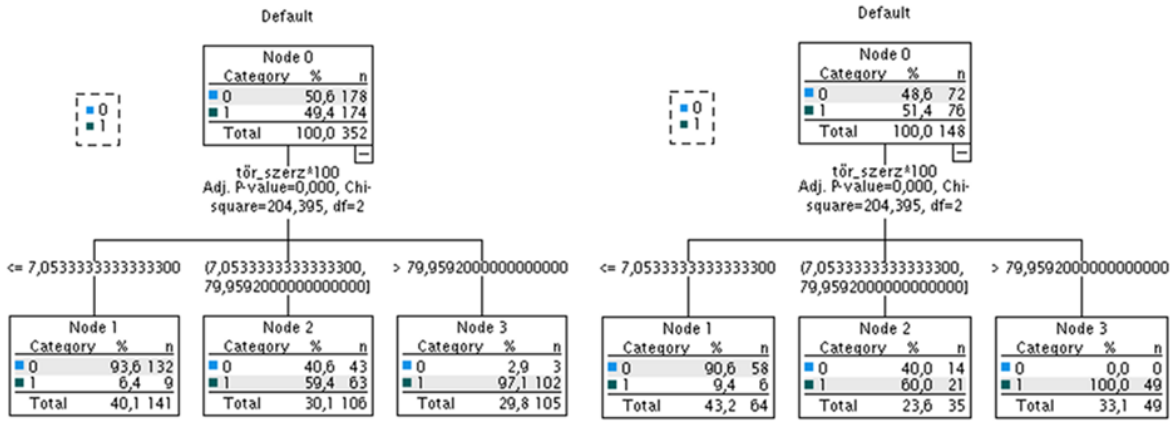

Figure 3. *Decision tree for the training and test sample*
Source: SPSS output, own editing

This decision tree also has a level, but in this case, instead of the previous 2 leaves, there are now 3.

The resulting tree shows that if the value of the variable does not exceed 7.053, then the percentage of non-performing loans is low, but if it exceeds 79.959, then it is almost certain that the loan is a non-performing loan.

Information on the accuracy of the classifications is provided by the classification matrix.

Although the accuracy of the first model (93.0%) decreased in this case, the fact that the first model could not categorise any non-performing loan transaction was not good for the purpose of the analysis. In this case, however, a significant proportion of non-performing loans were categorised in the correct group, and the results obtained on the training and test samples do not differ significantly, so I consider the results obtained to be valid.

Table 2. *Classification matrix*

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	132	46	74.2%
	1	9	165	94.8%
	Overall Percentage	40.1%	59.9%	84.4%
Test	0	58	14	80.6%
	1	6	70	92.1%
	Overall Percentage	43.2%	56.8%	86.5%

Source: Own editing

6. Comparison of results and conclusion

In the analyses, I found that when using the decision tree, one explanatory variable was significant. It can be concluded that the most significant variable of the data recorded by the KHR in terms of loan defaults is the ratio of the repayment to the contract amount.

There is also a significant difference in the classification accuracy of the different sample-based methods, as summarised in the table below.

Table 3. *Performance of the models developed using different evaluation techniques*

		Accuracy			AUC (%)	Gini (%)
		0	1	∑		
Decision tree	I.	100	0	93	81,6	63,2
	II.	74.2	94.8	84.4	91.7	83.4

Source: Own editing

For the AUC value, a value between 80-90% is considered to be outstanding. Both models have AUC values above 80%. A similar conclusion can be drawn for the Gini coefficient, where a value above 70% indicates a very strong model. The value of the second model is higher than this criterion.

Based on the above, it can be concluded that the classification accuracy of the initial model was higher, but it could not properly categorise non-performing loans. For the re-sampled model, although the predictive ability was reduced, the classification accuracy of non-performing loans was significantly improved, and higher AUC and Gini coefficient values were obtained for the second model, so that this model can be considered as better.

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23 (4), p. 589. **https://doi.org/10.2307/2978933**

[2] Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, 1 (1), pp. 29–54. **https://doi.org/10.1016/0378-4266(77)90017-6**

[3] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, p. 71. **https://doi.org/10.2307/2490171**

[4] Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, 12 (1), p. 1. **https://doi.org/10.2307/2490525**

[5] Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10 (1), 167. **https://doi.org/10.2307/2490225**

[6] Fekete O., Tatay T. (2012). *Hitelezők és adósok kapcsolatának szabályozási kérdései*. https://kgk.sze.hu/images/dokumentumok/kautzkiadvany2012/penzugy/fekete_tatay.pdf.

[7] Frydman, H., Altman, E. I., & Kao, D.-L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, 40 (1), pp. 269–291. **https://doi.org/10.1111/j.1540-6261.1985.tb04949.x**

[8] Hajdu O. (2003). *Többváltozós statisztikai számítások* [Multivariate statistical calculations]. Budapest: Központi Statisztikai Hivatal. **https://doi.org/10.20311/stat2018.10.hu1021**

[9] Hámori G. (2001). A CHAID alapú döntési fák jellemzői. *Statisztikai Szemle*, 79. évf., 8. sz., 703–710. http://www.ksh.hu/statszemle_archive/2001/2001_08/2001_08_703.pdf.

[10] KHR Annul Information (2021). https://www.bisz.hu/dokumentumtar (May 2023).

[11] Ketskeméty L., Izsó L., Könyves Tóth E. (2011). *Bevezetés az IBM SPSS Statistics programrendszerbe* [Introduction to IBM SPSS Statistics]. Budapest: Artéria Stúdió Kft.

[12] Malhotra, N. K. (2008). *Marketingkutatás* [Marketing research]. Budapest: Akadémiai Kiadó.

[13] McKee, T. E., Greenstein M. (2000). Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. *Journal of Forecasting*, 19, pp. 219–230. **https://doi.org/10.1002/(SICI)1099-131X(200004)19:3<219::AID-FOR752>3.0.CO;2-J**

[14] MNB (2023). *Hitelezési folyamatok. Magyar Nemzeti Bank. 2023. március*.

[15] Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *1990 IJCNN International Joint Conference on Neural Networks*, pp. 163–168. **https://doi.org/10.1109/IJCNN.1990.137710**

[16] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18 (1), p. 109. **https://doi.org/10.2307/2490395**

[17] Olmeda, I., & Fernández, E. (1997). Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction. *Computational Economics*, 10 (4), pp. 317–335. **https://doi.org/10.1023/a:1008668718837**

[18] Sajtos L., Mitev A. (2007*). SPSS kutatási és adatelemzési kézikönyv* [SPSS research and data analysis handbook]. Budapest: Alinea Kiadó.

[19] Vértesy L. (2008). *A pénzügyi intézmények finanszírozási tevékenységének jogi szabályozása Magyarországon*. PhD-értekezés.

[20] Virág M. (2004). A csődmodellek jellegzetességei és története [Characteristics and history of bankruptcy models]. *Vezetéstudomány*, 35 (10), pp. 24–32.

[21] Virág M., Kristóf T. (2005). Az első hazai csődmodell újraszámítása neurális hálók segítségével. [Recalculation of the first domestic bankruptcy model using neural networks]. *Közgazdasági Szemle*, 52 (2), pp. 144–162.

[22] Zhang, G., Hu, M. & Patuwo, B. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*, 116, pp. 16–32. **https://doi.org/10.1016/S0377-2217(98)00051-4**