DECISION TREE OR LOGISTIC REGRESSION - WHICH BASIC MODEL IS BETTER?

Kitti Fodor
Department of Business Statistics and Economic Forecasting, Faculty of Economics, University of Miskolc, Miskolc, Hungary
kitti.fodor@uni-miskolc.hu

Abstract: *In this paper, my aim is to show which of the data in the Central Credit Information System are the ones that influence the factors that are then used to perform the analysis using a decision tree and logistic regression, and I would like to know, which of the two basic model is the better one. For the analyses, I used a random sample of 500 items, reflecting the proportions of performing and non-performing loans in the population. For both methods, one variable was found to be significant, which was the ratio of the repayment to the contract amount, so this is the most significant of the data recorded by the Central Credit Information System in terms of loan defaults. If I compare the two methods, I can conclude that both methods have a high level of accuracy, but logistic regression is the one that produced better results, as it was able to identify a higher proportion of defaulted loans. Unfortunately, the decision tree could not identify any defaulting loans despite its higher classification accuracy. The reason can be the unfavourable sample composition. Finally, the logistic regression was able to categorize the transactions with 81,1% accuracy and has better AUC value and better value for Gini coefficients.*

Keywords: *loan default; decision tree; logistic regression, random sample; classification; ROC curve*

JEL classification: B16; C38; C44; C53

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 and logistic regression can categorize defaulted loans, which variables are significant among the data recorded by the KHR or calculated based on KHR data and which method is better using different evaluation methods.

1.1 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).

2. Bankcrupty Methods

For bankcrupty forecasting discriminant analysis, logistic regression, decision tree and neural network are widely used. For this study I used decision tree and logistic regression.

2.1. Decision tree

This analysis is 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, 2018)

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 and I used the CHAID.

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.

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)

2.2 Logistic regression

In logistic regression, the goal is to classify observation units into predefined groups of dependent variables. In this case, the dependent variable has two categories, so I applied binomial logistic regression. In logistic regression, the analysis is based on the "odds", which determine the probability of the default. The odds can be expressed by the following formula:

$$odds_x = \frac{P_x}{1 - P_x}$$

In the logistic regression, we assume that the logarithm of the odds can be defined as a linear function of the independent variables, which can be written as follows:

$$\ln(odds_x) = logit(P_x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

The other central element of the analysis is the so-called cut point value. This value can be chosen arbitrarily, but it is important to keep in mind that the losses resulting from a false classification are kept to a minimum. (Hajdu 2003; Malhotra 2008; Sajtos & Mitev 2007; Varga & Szilágyi 2011)

3. 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 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 a database with 500 cases. For the sampling I used a random numbers generator and simple random sampling. This is a type of representative sampling.

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

4. Empirical research

Recent methods used for bankruptcy prediction include decision tree, logistic regression, and neural networks. I used two of these methods and I assume that these methods can be used to predict with high accuracy which customers or loan transactions will default, also in the case of retail lending.

To support this statement, I constructed classification models using decision tree and logistic regression. To perform the analysis, I used the database provided by the KHR and to validate the results, I divided the sample into a training and a test part. The training sample included 70% of the cases.

4.1 Logistic regression I. model

First, I performed a logistic regression analysis. Of the available explanatory variables, only the ratio of the repayment to the contract amount was found to be significant. The Omnibus test (p<0.001) and the Hosmer and Lemeshow goodness-of-fit test (p=0.212) showed a reliable model with a good fit. The generated model has medium explanatory power (Nagelkerke R^2=38.8%).

Table 1: Significant variables in the Logistic Regression I model

Sample		B	S.E.	Wald	Sig.	Exp(B)
Test	repayment ratio	,024	,007	13,249	<,001	1,025
	Constant	-3,594	,499	51,836	<,001	,027
Training	repayment ratio	,029	,004	49,181	<,001	1,030
	Constant	-3,448	,308	125,008	<,001	,032

Source: Own editing, SPSS output

The model equation can be written in the following form:

$$P_{(default)} = \frac{e^{0,032+1,030x_1}}{1+e^{0,032+1,030x_1}}$$

Where:
x_1: the ratio of the repayment to the contract amount.
For the evaluation of the model first I used the classification table.

Table 2: Classification table for Logistic regression I model

Sample	Observed		Predicted		
			Default		Percentage Correct
			0	1	
Test	Default	0	138	4	97,2
		1	4	4	50,0
	Overall Percentage				94,7
Training	Default	0	261	62	80,8
		1	4	23	85,2
	Overall Percentage				81,1
a. The cut value is ,039					

Source: Own editing, SPSS output

The selected cut-off value is different from the default value of 0.5. There are several recommendations to determine it, of which I used the Youden rule. Youden's rule considered 0.039 to be the optimum value, so I used it. Although the cut off value chosen in this way reduced the accuracy of the classification from 93.7% to 81.1%, it increased the correct categorisation rate for non-performing loans from 40.7% to 85.2% and can therefore be considered as more favourable.

Overall, the model created correctly categorised loan transactions with an accuracy of 81.1%, with 66 items incorrectly categorised. As the cut off value used is low, the random classification classified all transactions as non-performing and compared to the random categorisation (7.7%), the 81.1% value can be considered as a significant increase.

A significant difference between the training and test sample is observed for classification accuracy, sensitivity and specificity. The sensitivity of the test sample is significantly lower than that of the training sample. This may be due to the predominance of performing loans in the sample, i.e. the sample composition is unfavourable for analysis. To improve this, I will perform the analysis on a new sample in the future, so for the time being I consider this model as final.

Logistic regression imposes several conditions on the analysis, so these conditions need to be checked:

level of measurement of the variables: the dependent variable is a dichotomous variable, and the independent variables can be measured at any scale, so this condition is met.

independence of data: an item represents a loan transaction that is independent of other loan transactions, so this condition is also met.

sample size: sample of 500 items.

multicollinearity: only one explanatory variable was found to be significant in the analysis.

Thus, it can be concluded that the model created meets all the criteria and the validation was successful.

4.2 Decision tree I. model

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%).

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. For the analysis I used a training and a test sample. The decision tree run on the training and test sample is shown in Figure 1.

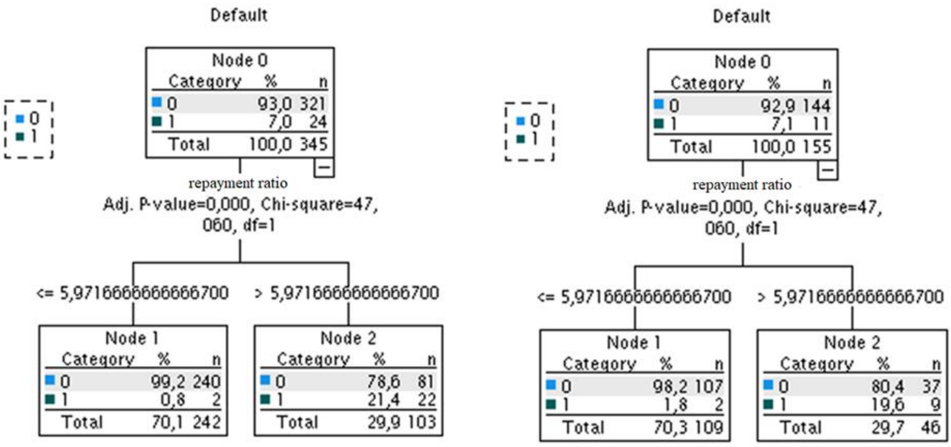

Figure 1: Decision tree for the training and test sample

Source: Own editing, SPSS output

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.
Based on the analysis, it can be found, 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.

Table 3: Classification matrix for Decision tree I model

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	312	0	100,0
	1	24	0	0,0
	Overall Percentage			93,0
Test	0	144	0	100,0
	1	11	0	0,0
	Overall Percentage			92,9

Source: Own editing

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.

4.3 Comparison of the models

I based the models on four explanatory variables, and the table below summarises which explanatory variables were found to be significant by the different methods.

Table 4: Summary of variables used by classification models

Name of the variable	Log. regr.	Decision tree
ratio of repayment	X	X
loan maturity		
age		
gender		

Source: Own editing

Based on the above, it can be concluded that the most significant of the data recorded by the KHR in terms of loan defaults is the ratio of the repayment to the contractual amount.

In addition to the classification matrix, I also used the ROC curve, the AUC value, and the Gini coefficient to evaluate the models. On the Figure 2 can we see the ROC curve on the left side for the decision tree, on the right side for the logistic regression.

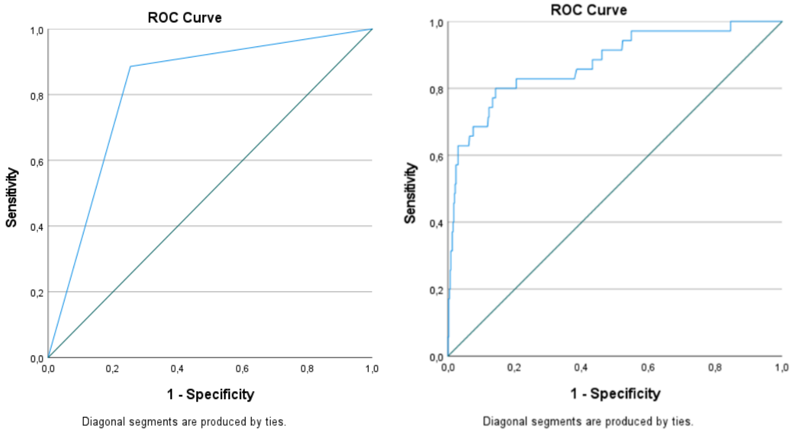

Figure 2: ROC curve for Decision tree I. and for Logistic regression I.
Source: Own editing, SPSS output

Based on the ROC curve can be the AUC (Area Under the Curve) value calculated. The AUC ranges from 0 to 100%, where 100% is considered perfect. If the AUC is around 80-90%, it is considered outstanding.

One other method for the evaluation is the Gini coefficient, which can be calculated in several ways. The simplest way to calculate it is:

$$Gini = 2\ (AUC - 0{,}5)$$

The maximum value of the indicator is 1. If the value is between 60% and 70%, the model is considered to be correct. If the value is above 70%, the model is considered to be good. (Engelman et al, 2003; Olawale, 2020)

Table 5: Evaluation of the models developed using different methods

	Accuracy			AUC (%)	Gini (%)
	0	1	∑		
Logistic regression	80,8	85,2	81,1	87,7	75,4
Decision tree	100	0	93	81,6	63,2

Source: Own editing

In both cases I achieved a value over 80% for AUC, so the models can be considered as outstanding.
The value of the Gini coefficient is 63,2% in the case of decision tree, so the model can be considered as correct, and in the case of the logistic regression the value is 75,4%, so the model can be considered as good.

5. Summary

In the analyses, I found that when using both methods, 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 are significant differences in the evaluation systems. The reason is that in the sample the proportion of performing loans was higher and it is unfavourable for the selected methods. The classification matrix of decision tree showed that none of the cases could the model correctly categorise of the non-performing loans and the Gini coefficient value also indicated that the model is not the best. The accuracy of the logistic regression was lower, but it could categorise the non-performing loans with higher proportion, and the value of the other evaluation methods was higher, but there was a significant difference between the test and the training sample.
There are several possible solutions to the problem. One option is to use a new sample that is more favourable to the analytical methods. Another possible solution is to add new variables to the current sample that are not recorded by the KHR and that may be important in the borrowing process. But answering these questions will be the subject of another study.

References

Database: provided by Bisz Zrt.

1. Altman, E. I. (1968): Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, The Journal of Finance, Vol. 23. No. 4. 589-609. old.
2. Engelmann, B., Hayden, E., Tasche, D. (2003): Testing rating accuracy, 2003 január, Credit Risk, www.risk.net.
3. Frydman, H., Altman, E. I., Kao, D. L. (1985): Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress, The Journal of Finance, Vol. 40. No. 1. 303-320. old.
4. Hajdu, O. (2003): Többváltozós statisztikai számítások, Központi Statisztikai Hivatal, Budapest
5. Hajdú, O. (2018): Többváltozós statisztikai R Open alkalmazások (2018. szept 28.) Statisztikai szemle
6. Hámori, G. (2001): A CHAID alapú döntési fák jellemzői, Statisztikai Szemle, 79. évf. 8. sz. 703-710. old. http://www.ksh.hu/statszemle_archive/2001/2001_08/2001_08_703.pdf
7. KHR Annul Information (2021). https://www.bisz.hu/dokumentumtar (May 2023)
8. Malhotra, N. K. (2008): Marketingkutatás (Akadémiai Kiadó, Budapest)

9. McKee, T.E., Greenstein M. (2000): Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. Journal of Forecasting, 2000, no.19. pp. 219-230.
10. Ohlson, J. (1980): Financial ratios and the probabilistic prediction of bankruptcy, Journal of Accounting Research, Vol. 18. No. 1. 109-131. old.
11. Ojo Olawale: The CAP curves (https://waleblaq.medium.com/the-cap-curves-the-cumulative-accuracy-profile-58a141e01fae
12. Sajtos L., Mitev A. (2007): SPSS kutatási és adatelemzési kézikönyv (Alinea Kiadó, Budapest)
13. Virág, M. (2004): A csődmodellek jellegzetességei és története, Vezetéstudomány, 35. évf. 10. sz. 24-32. old.