Techniques for Customer Behaviour Prediction: A Case Study for Credit Risk Assessment

Laura Maria Badea Stroie¹
¹Bucharest Academy of Economic Studies – Doctoral School, Department of Cybernetics and Statistics, e-mail: laura.maria.stroie@gmail.com

Abstract

The continuous increase in the amount of information that needs to be processed and the necessity to reduce the time spent on data classification have directed researchers’ interest towards new modelling techniques. Artificial neural networks are non-parametric estimation methods which have proven high flexibility in modelling the data from all types of areas. Considering the current economic context which has determined financial institutions to raise protection against default event that might incur in case of companies’ loans, this paper tests artificial neural networks against traditional techniques in a credit risk assessment process. Even if logistic regression and decision trees proved good results, neural networks outperformed the other two methods, providing the highest overall detection rate and still keeping its generalization capacity on the out-of-sample set at a high level.

Keywords: neural networks, decision trees, logistic regression

1. Introduction

Data classification has become an important issue, as the amount of information that is generated and needs to be handled has significantly increased. Modellers have to choose between the easiness of using traditional techniques and the complexity of newer methods which in the end might, nevertheless, generate better results.

Using the loan information provided by a Romanian credit institution on small-medium enterprises (henceforth also SMEs), this paper focuses on identifying the best model in terms of default prediction, but also in terms of flexibility in operating with the data.

In Section 2, a brief description of SMEs credit quality in Romania will be brought to discussion. Related studies regarding the evaluation of companies’ credit quality using decision trees, logistic regression and neural networks will be presented in Section 3. Afterwards, variable selection and sampling information regarding the data used in the analysis will be provided. In Section 5 of the paper, the estimation credit risk models will be developed and their results will be detailed. The sixth part focuses on making a comparative assessment of the three models used, while the last one concludes the main aspects in respect with the best model emerged from the performed analysis.
2. SMEs in Romania – Access to financing and creditworthiness

SMEs play an important role within national economy, especially through the high contribution of working places from the labour market, but also through the increased potential of adapting their supplies to consumer demands and needs. However, over the past three years, in the context of the financial and economic crisis, SMEs have been very affected by limited access to financing. This is mainly because of higher risks they are exposed to and also because of increased restrictions imposed by the National Regulator on the lending activity. Publications from the National Bank of Romania (2012, p. 12) have defined rather “concerning” SMEs’ access to financing, with an increase of only 4.9% in loan amounts over the time period December 2010 – June 2012, compared with large companies where the rise has been of 8.2%. During December 2008 – December 2011, the number of SMEs with loans granted by resident financial institutions has decreased by 16.5%, down to 76,000, the total rate of SMEs with loans reaching only 21% out of the overall Romanian active small-medium enterprises. At the same time, SMEs credit risk has faced a rise, the rate of nonperforming loans (defined as loans and interests with more than 90 days past dues and/or those for which legal enforcement acts have been initiated on the activity or on the debtor, related to all registered credits and interests) reaching 23.2% in July 2012, compared with 15.1% in December 2010. This is the result of subunit interest coverage ratio, decreasing cash flows generated by the main activity, and also decreasing gross profit margin that SMEs have faced during the economic crisis.

3. Credit risk models and related studies

The accelerated increase in the rate of nonperforming loans, along with additional prudential requirements imposed on the lending activity by the National Bank of Romania, have raised the need for financial institutions to develop advanced credit risk models, capable of estimating the creditworthiness of loan applicants. This is an important process, not only because it impacts the quality of the lending activity, but it also influences the value of capital requirements that the banks must hold in order to keep the risk within bearable parameters. According to Dietsch and Petey (2004), SMEs induce a higher credit risk compared with large companies, but they also can bring more benefits by representing a large number of potential clients. Originally introduced by Altman (1968), multiple discriminant analysis became a common and easy to understand method in estimating the credit risk through the z-score, which consisted of five financial variables (out of 22 analysed). These, combined in a model and applied on a set of 66 manufacturing firms, were capable of predicting the bankruptcy event. Later, Altman et al. (1977) gave further evidence on the performance of the z-score model through the improvements brought to the variables used (“Zeta model”). Since then, various methods have been used for default prediction, each having its upsides and limitations. Next, a particular interest will be paid to three default estimation techniques, namely decision trees, logistic regression and neural networks.
3.1. Decision Trees

Decision trees are classification techniques, organized in tree-like graphs, with decision branches, and with possible results as leaves. These methods are based on maximizing the performance measure chosen as related to the target variable. Even though it is fairly easy to understand its output ("white-box" character), decision trees approach has a rather restrictive capacity in generalizing the results and in handling large number of variables. Frydman et al. (1985) showed that results obtained through Recursive Partitioning Algorithm outperform discriminant analysis when using financial variables for credit risk evaluation. Also, Simha and Satchidananda (2006) used the agricultural loan information provided by two banks from India to predict default risk using both decision tree learning scheme and logistic regression method. They found that decision trees technique generated good results (the overall default precision using decision trees was 90%, while using logistic regression was 83%) and have the benefit of being easy to understand and implement.

3.2. Logistic regression

Binary logistic regression is a common technique used for estimating the probability of a certain event happening. This approach does not make any assumptions regarding the normal distribution of variables, it is based on the maximum likelihood method and it is easy to interpret in terms of probability because the output varies between 0 and 1. The logistic regression (logit) estimates the probabilities that each event happens as follows:

\[
P(y = 1 \mid X) = \frac{1}{1 + e^{-(w_0 + W^T X)}} \quad \text{and} \quad P(y = 0 \mid X) = \frac{e^{-(w_0 + W^T X)}}{1 + e^{-(w_0 + W^T X)}}
\]

(1)

Where, \( y_i \in \{0,1\} \) is the binary target variable, \( X \) a n-dimensional input vector, \( W \) is the vector of parameters and \( w_0 \) is the intercept. For parameter estimation, the maximum likelihood method is used. The probability of each event happening is:

\[
P(y \mid X) = P(y = 1 \mid X)^y (1 - P(y = 1 \mid X))^{1-y}
\]

(2)

The likelihood and the log-likelihood become then:

\[
\prod_{i=1}^{N} P(y_i = 1 \mid x_i)^{y_i} (1 - P(y_i = 1 \mid x_i))^{1-y_i}
\]

(3)

\[
LogL = \sum_{i=1}^{N} y_i \log(P(y_i = 1 \mid x_i)) + (1 - y_i) \log(1 - P(y_i = 1 \mid x_i))
\]

(4)

Where, \( N \) is the number of observations from the development sample. Using Newton-Raphson algorithm, the \( LogL \) function will be maximized. The logistic regression assumes a linear decision boundary, therefore it is necessary that \( P(y = 1 \mid X) \), or a monotone transformation of the \( P(y = 1 \mid X) \) is linear in \( X \):

\[
\log \left( \frac{P(y = 1 \mid X)}{1 - P(y = 1 \mid X)} \right) = w_0 + W^T X
\]

(5)
Where, \( \frac{P(y = 1 | X)}{1 - P(y = 1 | X)} \) represents the odds that \( y = 1 \).

Logistic regression has been widely used for credit risk purposes. In order to avoid the limitations imposed by the multiple discriminant analysis, Ohlson (1980) performed a logistic regression to predict company bankruptcy using nine predictors. Also, Behr et al. (2004) applied a binary logistic regression on a set of quantitative and qualitative indicators in order to predict the default event at SMEs with loans from a big bank in Germany. Altman and Sabato (2007) performed the logistic regression on a sample of 2,000 US SMEs in order to develop a credit risk model. One year later, Yazdanfar and Nilsson (2008) identified the main predictors of SME’s bankruptcy in Sweden. Using multiple discriminant analysis and logistic regression, the accuracy ratios were higher in case of the latter approach (77.8% up to 83.5% in case of logistic regression, compared with 75.4% up to 82.8% in case of discriminant analysis).

3.3. Artificial Neural Networks

Artificial Neural Networks (henceforth also ANNs) are learning techniques developed after the way in which the human brain behaves. These are capable of modifying their internal parameters and adapting themselves in order to reach certain results. ANNs are composed of neurons (nodes) which are interconnected in a pre-specified way, depending on the type of neural network. These connections represent weights associated with the intensity and the amount of information used in the learning process. Given their lack of transparency in model building, ANNs are considered “black-boxes”, but after the proper architecture is chosen they nevertheless provide good results and have a high generalization power, if the learning process stops in time.

Multilayer perceptron (MLP) using the “back-propagation” learning algorithm is one of the most common types of neural networks used in credit risk modelling. A MLP contains an input layer, one or more hidden layers and an output layer, and operates in a feed-forward manner (each neuron from a layer is connected to each neuron from the next layer). The outputs from hidden and output layers are generated as a weighted sum of the bias and the weighted inputs using transfer functions. The most common types of transfer functions are sigmoid, hyperbolic tangent and linear. The weights from the hidden and output layers are initialized in a random manner and afterwards are iteratively adjusted in order to minimize the error function.

Wu and Wang (2000) used small business financial information provided by an important New York bank to develop a credit decision tool based on neural networks. Compared with other traditional methods, ANNs proved to have high predictive power, generating good results with reduced evaluation time and costs. Pang et al. (2002) used MLP to generate a credit-scoring model for companies applying for loans in China. Results indicated high capacity of the model in discriminating between “good” and “bad” (risky) companies. Also, Vallini et al. (2009) used neural networks, multiple discriminant analysis (MDA) and logistic regression to predict the default event at SMEs on a sample of 6,113 Italian firms. Compared with the traditional MDA and logistic regression techniques which reached overall detection rates of 65.9% and 67.2% respectively, ANNs generated better results, with a correct classification rate of 68.4%.
4. Dataset and variable selection

For this study, small-medium enterprises loan information from a Romanian financial institution was taken into consideration. SMEs were picked based on the size of the annual net turnover; following the bank’s internal segmentation, were categorized as SMEs those companies having less than 10 MEUR net turnover at the end of 2010. Afterwards, all selected SMEs were observed over the time period 1st of January 2011 – 31st of December 2011 and were assigned with the following status in terms of default event: i) “non-default” = no default event in 2011, ii) “default” = default event in 2011. Default was defined as one of the following events: more than 90 days in overdue, specific provisions required, insolvency case, or restructuring case. The final database consisted of 1,015 non-default enterprises and 264 default SMEs, generating an overall default rate of 20.64%.

Altman and Narayanan (1997) have pointed out the need of including financial information related to the following categories: profitability, debt coverage and leverage, liquidity, capitalization ratios, and earnings variability. Going deeper, Hayden (2003) analyzed the following extended categories of financial ratios as possible predictors of the default event: leverage, debt coverage, liquidity, activity, productivity, turnover, profitability, size, and growth. Some of these were also considered in the present paper as candidates for a default prediction model development.

However, as sometimes financial variables are not the only triggers of a default, qualitative data was also taken into consideration. Lehmann (2003) proved that the default estimation model significantly improved when soft facts information was included in the model. For the current analysis, the individual accuracy ratios of 50 hard facts indicators developed from the available financial information were calculated. Considering the financial standings from 31st of December 2010, the accuracy ratios, and an optimal combination of financial categories, ten hard facts variables were chosen for the final model (Table 1).

<table>
<thead>
<tr>
<th>Financial Indicator</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary business income / Operating income</td>
<td>Profitability</td>
</tr>
<tr>
<td>Net Income / Net sales</td>
<td>Profitability</td>
</tr>
<tr>
<td>Cash / Current liabilities</td>
<td>Liquidity</td>
</tr>
<tr>
<td>Cash / Assets</td>
<td>Liquidity</td>
</tr>
<tr>
<td>Bank debt / Assets</td>
<td>Leverage</td>
</tr>
<tr>
<td>(Net result + Interest expenses + Amortization) / Interest expenses</td>
<td>Debt coverage</td>
</tr>
<tr>
<td>Accounts payable / Net sales</td>
<td>Activity</td>
</tr>
<tr>
<td>Accounts payable / Material costs</td>
<td>Activity</td>
</tr>
<tr>
<td>(Net sales - Material costs) / Personnel costs</td>
<td>Productivity</td>
</tr>
<tr>
<td>Material costs / Operating income</td>
<td>Productivity</td>
</tr>
</tbody>
</table>

A particular attention was paid in avoiding high correlations between variables, by removing indicators when correlations coefficients exceeded 0.35 in absolute value.
Qualitative factors included in the model are: “Management qualification and experience”, “Business strategy”, “Industry” and “Maximum delay so far” (if new loan applicants didn’t have any Credit Bureau information due to no past loans, they were assigned with the category “no delays”). These were also tested for their predictive power using the statistics weights-of-evidence and information-value. Since all four variables have an Information-value above 0.1, indicating good discriminatory power (Table 2), they were all kept in the model. However, results indicate that “Quality of the management” is an important aspect in SMEs business development, most of their activities being driven by the managers’ capacity to conduct efficient operations and behave operatively also within difficult conditions.

Table 2 – Information value for qualitative variables

<table>
<thead>
<tr>
<th>Qualitative indicator</th>
<th>Information-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management qualification and experience</td>
<td>0.5809</td>
</tr>
<tr>
<td>Business strategy</td>
<td>0.3809</td>
</tr>
<tr>
<td>Industry</td>
<td>0.1159</td>
</tr>
<tr>
<td>Maximum delay so far</td>
<td>0.4977</td>
</tr>
</tbody>
</table>

For evaluation purposes, the initial dataset was partitioned into training set (60%), validation set (25%), used for model assessment, and test set (15%), representing an out-of-sample database. For each set, the same default rate was kept as in the initial database, approximately 21%.

5. Model development and results

The main focus in this paper is on identifying the model with best results in predicting the risky SMEs who apply for new loans. In this respect three models were configured and compared in terms of detection rates: decision trees, logistic regression and neural networks. All models were built and tested against the same datasets defined above.

5.1. Decision trees

Considering all 14 variables as possible candidates for the model, a decision tree was built using the following characteristics: i) the splitting rule for interval variables was F-statistic and for nominal variables the Chi-Square statistic (both having a threshold of 0.2 for p-value); ii) maximum branch number was set at 2; iii) maximum depth (maximum number of generations of nodes) and maximum categorical size (maximum number of observations for a categorical value) were set at 5; iv) input variables were used only once; v) the leaf size, which specifies the smallest number of observations from a training leaf, was set at 5; vi) the assessment measure was the misclassification rate on the validation dataset; vii) the building method was based on choosing the smallest sub-tree with the best results in terms of misclassification rate.
Based on these rules, the decision tree follows the splits from Figure 1, assigning high importance to qualitative information, such as “Maximum delay so far” and “Management qualification and experience”.

Results available in Table 3 indicate a good overall detection rate (above 80%) on all three datasets at a 0.5 cut-off point. On the training dataset, as well as on the test sample, the decision tree model reaches a correct classification rate of 82.9%, marking a good profit rate.

Table 3 – Decision trees results on all three datasets

<table>
<thead>
<tr>
<th>Statistics Label</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Squared Error</td>
<td>0.1290</td>
<td>0.1329</td>
<td>0.1431</td>
</tr>
<tr>
<td>Overall detection rate</td>
<td>82.92%</td>
<td>83.70%</td>
<td>82.90%</td>
</tr>
</tbody>
</table>

5.2. Logistic regression

The logistic regression was performed using the following particularities: i) the link function is \( \text{logit} \); ii) nominal values were encoded using deviation method; iii) the model included all 14 variables.
Results at a cut-off value of 0.5 are available in Table 4 and show a detection rate of 85.4% on the training database. Thus, logistic regression technique has reached better detection rates on the training and validation samples compared with decision trees. However, on the test set, the model loses its generalization capacity, with an overall classification rate of 80.83%.

Table 4 – Logistic regression results on all three datasets

<table>
<thead>
<tr>
<th>Statistics Label</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Squared Error</td>
<td>0.1162</td>
<td>0.1268</td>
<td>0.1347</td>
</tr>
<tr>
<td>Overall detection rate</td>
<td><strong>85.40%</strong></td>
<td><strong>84.01%</strong></td>
<td><strong>80.83%</strong></td>
</tr>
</tbody>
</table>

5.3. Neural networks

Neural networks performances are highly dependent on the model architecture and training process. That is why, sometimes, they require more time spent on the modelling part, but in the end, performance is superior to those obtained from using traditional statistical models.

In this paper the neural networks model was designed using the following parameters and assumptions: i) the network was built in block layers (which means that hidden layers were added as additional layers with a uniform number of neurons); ii) the learning process will stop when over-fitting is detected; iii) the maximum number of iterations to perform during training is 15; iv) the number of hidden units is 7; v) the chosen activation function in the hidden and output layers is *tangential*; vi) the learning algorithm is “back-propagation”.

Results of the model built using neural networks are available in Table 5. These indicate better detection rates on all analysed datasets when compared with those obtained from using traditional statistical methods. Compared with logistic regression, the correct classification rate on the training sample is by 1.5pps higher, reaching 86.96%. The model performs good also on the test dataset, with an overall detection rate of 83.94%, the highest of all three analyzed techniques.

Table 5 – Neural Networks results on all three datasets

<table>
<thead>
<tr>
<th>Statistics Label</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Squared Error</td>
<td>0.0975</td>
<td>0.1238</td>
<td>0.1278</td>
</tr>
<tr>
<td>Overall detection rate</td>
<td><strong>86.96%</strong></td>
<td><strong>86.21%</strong></td>
<td><strong>83.94%</strong></td>
</tr>
</tbody>
</table>

6. Model comparison

Thus, results indicate neural networks approach as having the best performance in terms of predicting default applications. Even if past studies indicated neural networks as being very exposed to “over-fitting” phenomenon, the developed model manages to perform well even on the out-of-sample test set, with an 83% detection rate and, thus, a good generalization capacity.

In terms of detection rates per class (“default”/”non-default”), results from Table 6 reveal lower performances in case of “default” applications. These are nevertheless available for
a 0.5 cut-off point. In practice, financial institutions can modify this cut-off point in order to detect more default cases, but this way the precision on non-default observations will also decrease, forcing a trade-off. However, considering that in case of “default” events the banks would risk losing not only the gain from interests, but also the capital lent, it is wiser to choose a model that detects more “default” cases in the detriment of rejecting some potentially good customers. The highest detection rate for “default” customers, at a 0.5 cut-off point, is reached by neural networks model, with 49.2% precision on the validation set, significantly higher compared with logistic regression and decision trees.

<table>
<thead>
<tr>
<th>Category</th>
<th>Decision Trees</th>
<th>Logistic Regression</th>
<th>Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Validation</td>
<td>Training</td>
</tr>
<tr>
<td>“Default”</td>
<td>27.2%</td>
<td>29.2%</td>
<td>38.6%</td>
</tr>
<tr>
<td>“Non-default”</td>
<td>97.4%</td>
<td>97.6%</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

The ROC (“Receiver Operating Curve”) available in Figure 2, performed on the training and validation datasets, strengthens the evidence of good performance provided by neural networks model (the closer the curve gets to the point (0,1), the better the model discriminates).

7. Conclusions

Compared with traditional credit risk estimation techniques such as decision trees and logistic regression, neural networks provide the best results in terms of correctly classified firms, proving also flexibility when operating with large number of variables. Moreover, in this study, neural networks generated significantly higher detection rates for “default” cases, which represent in fact the main focus when developing credit risk models. Even if it takes more time building and configuring them, with the proper architecture and an optimal stopping rule, neural networks become highly performant techniques with a good generalization power.
References


