Decision Making in Manufacturing and Services
Vol. 14 • 2020 • No. 2 • pp. 193–208

Credit Risk Management Using Automatic Machine Learning

Bartłomiej Gaweł*, Andrzej Paliński*

Abstract. The article presents the basic techniques of data mining implemented in typical commercial software. They were used to assess the risk of credit card debt repayment. The article assesses the quality of classification models derived from data mining techniques and compares their results with the traditional approach using a logit model to assess credit risk. It turns out that data mining models provide similar accuracy of classification compared to the logit model, but they require much less work and facilitate the automation of the process of building scoring models.

Keywords: data mining, scoring, credit, loan

Mathematics Subject Classification: 91-08, 62P20

JEL Classification: C25, C53, C55, G21

Submitted: December 20, 2020

Revised: December 30, 2020

1. INTRODUCTION

After 2000, there was an avalanche of data collected in databases all over the world and the intensive development of techniques and tools for analyzing large data sets. Classic methods of building econometric models have been automated and built into commercial software. Parallel research in the field of artificial intelligence and machine learning has led to the development of advanced tools for automatic data analysis. These are, among others: classification and regression trees, neural networks, genetic algorithms, image recognition techniques (patterns recognition), association rules, fuzzy logic and rough sets, and others (e.g. Larose, 2006; Han *et al.*, 2012).

The contemporary approach to building empirical models goes beyond the traditional statistical and econometric models. Classic models require, on the one hand,

* AGH University of Science and Technology in Krakow, Faculty of Management, Department of Business Informatics and Management Engineering, Poland, e-mail: bgawel@zarz.agh.edu.pl; palinski@zarz.agh.edu.pl

DOI: https://doi.org/10.7494/dmms.2020.14.2.4379

the fulfillment of strict requirements for statistical inference, on the other hand, they impose the necessity to define a set of explanatory variables in advance. This dataset may be narrowed in the course of inference, but it is difficult to extend it later.

The construction of empirical models can currently be based on any wide set of potential explanatory variables, not necessarily directly related to the dependent variable, which is allowed by the contemporary so-called data-driven models that use research results from the following areas (Holdaway, 2014):

- artificial intelligence;
- computational intelligence – neural networks, fuzzy systems, evolution algorithms;
- soft computing – reasoning based on fuzzy terms – fuzzy rule systems;
- machine learning;
- data mining and knowledge discovery in databases.

The advantage of this approach is not only that it leaves the selection of variables to the algorithms embedded in the software, but also that it facilitates the possibility of any frequent update of the model form (e.g. quarterly, monthly or more frequently). Moreover, models built in this way can operate automatically and inform about the detection of deviations.

The aim of this article is to investigate whether widely available commercial software with embedded machine learning and data analysis tools is able to provide an effective instrument for building scoring models in any credit institution without significant workloads and the costs of professional analysis of loan portfolio. The research hypotheses are:

- commercial machine learning software provides ready-made tools for building scoring models with good prognostic quality;
- commercial machine learning software significantly speeds up and facilitates the creation of scoring models.

The remainder of the paper is organized as follows: in Section 2, a short overview of research in the area of credit scoring was conducted. Section 3 presents the characteristics of the data used in the research. Section 4 contains results of building classification models derived from data mining techniques for forecasting defaults of credit card holders. In Section 5, a logistic regression model was built for the same data. In Section 6, all previously used exploration and regression techniques were used on a balanced dataset with equal proportions of repaid and default cases. The paper is concluded with a short summary.

2. LITERATURE REVIEW

Managing individual credit risk for a single transaction using empirical models dates back to the 1960s. The model of Altman (1968) is considered to be the fundamental model from that period. In classical models, the risk of insolvency or bankruptcy was predicted on the basis of the linear discriminant function (e.g. Wilcox, 1973; Laitinen, 1991). Simultaneously, studies were carried out using the generalized regression method

and the logit or probit models (e.g. Zmijewski, 1984; Li and Miu, 2010), or game theory (Palinski, 2018). More extensive research on the use of discriminant analysis and logistic regression in the risk assessment of Polish enterprises was presented in the work of Jagiełło (2013). In recent years, new insight was also brought by generalized partial linear models (Weber *et al.*, 2012). In most of the mentioned models, the borrower is assigned a point value, based on which the probability of default on the debt is determined. These are the so-called scoring models. In credit scoring, attempts have been also made to use non-classical techniques like artificial intelligence methods including: neural networks, genetic algorithms or expert systems (Jonc and Kraska, 2001; Matuszyk, 2013). The advantage of data mining techniques over classic statistical models is the easy with which they deal with missing and dirty data. The typical machine learning methods used in credit scoring include: K-nearest neighbors, naive Bayes, artificial neural networks, classification trees (Yeh and Lien, 2009), credit scorecard and decision tree (Yap *et al.*, 2011). In recent years, research in the field of credit risk has increasingly used team and enhanced classifiers, SVM, and fuzzy sets (Rębiasz *et al.*, 2017; Moradi and Mokhatab, 2019). A comprehensive overview of the methods used in credit scoring can be found in the work of Sadatrasoul *et al.* (2013), Keramati and Yousefi (2011) and Moradi and Mokhatab (2019). Nevertheless, the current approach to credit scoring assumed an individual and dedicated procedure for building a scoring model. Research efforts on the quality of models obtained with various machine learning methods have been carried out in research departments using an individual approach to each method. Applying an individual approach to building scoring models is time-consuming, requires professional knowledge, and is expensive. It is interesting and practical to examine whether the out of the box tools offered by software producers are able to build scoring models quickly and automatically, and without excessive loss of quality.

3. DATA

This paper uses a publicly available dataset available in the Machine Learning Repository (2020), previously used in a modified form in the article by Yeh and Lien (2009). The data concerns the repayment of credit card debt in a large Taiwanese bank in 2005. The aim of the study is to build a model forecasting the failure to repay the debt (default) from the credit card account in the next month, which is October 2005 for the dataset. The variable explained is a binary variable informing that the card holder did not repay the debt on the credit card in October 2005. The dataset with 30,000 observations includes 23 explanatory variables in addition to the dependent variable. All variables are as follows:

DEFAULT (depended variable) – whether a customer failed to pay off his credit card debt? (Yes = 1, No = 0);

LIMIT – the amount of credit granted (in Taiwanese dollars), this variable includes both individual consumer credits and credits of the indebted person's immediate family (supplementary credit);

SEX – gender (1 = male, 2 = female);
EDUCATION – education (1 = primary school graduate, 2 = university, 3 = high school, 4 = other);
MARRIAGE – marital status (1 = married, 2 = single, 3 = other);
AGE – age;
PAY_1...PAY_6 – explanatory variables relating to the history of recent payments (monthly payments from April to September 2005, coded as follows: PAY_1 = repayment status in September 2005; PAY_2 = repayment status in August; PAY_3 = repayment status in July; PAY_4 = repayment status in June; PAY_5 = repayment status in May; PAY_6 = repayment status in April 2005. The repayment measurement scale is as follows: –2 means "customer pays duly", –1 – "customer's payment is delayed by one month", 0 – "payment is delayed by two months", 1 – "payment is delayed by 3 months", 2 – "payment delayed by 4 months", 3 – "payment delayed by 5 months", 4 – "payment delayed by 6 months", 5 – "payment delayed by 7 months", 6 – "payment delayed by 8 months", 7 – "payment delayed by 9 months" and 8 – "more than 9 months delay in payment";
BILL_AMT1... BILL_AMT6 – respectively, history of amounts on credit card account in a given month (BILL_AMT1 = amount on the account in September 2005,..., BILL_AMT6 = amount on the account in April 2005);
PAY_AMT1... PAY_AMT6 – history of previous payments amounts (in Taiwanese dollars) (PAY_AMT1 = amount paid in September 2005,..., PAY_AMT6 = amount paid in April 2005).

Table 1 contains the basic descriptive statistics of quantitative variables.

Table 1. *Descriptive statistics of quantitative variables*

Variable	Minimum	1 quartile	Median	Mean	3 quartile	Maximum
LIMIT	10,000	50,000	140,000	167,484	240,000	1,000,000
AGE	21	28	34	35.49	41	79
PAY_1	–2	–1	0	–0.0167	0	8
PAY_2	–2	–1	0	–0.1338	0	8
PAY_3	–2	–1	0	–0.1662	0	8
PAY_4	–2	–1	0	–0.2207	0	8
PAY_5	–2	–1	0	–0.2662	0	8
PAY_6	–2	–1	0	–0.2911	0	8
BILL_AMT1	–165,580	3,559	22,382	51,223	67,091	964,511
BILL_AMT2	–69,777	2,985	21,200	49,179	64,006	983,931

Table 1 cont.

Variable	Minimum	1 quartile	Median	Mean	3 quartile	Maximum
BILL_AMT3	–157,264	2,666	20,089	47,013	60,165	1,664,089
BILL_AMT4	–170,000	2,327	19,052	43,263	54,506	891,586
BILL_AMT5	–81,334	1,763	18,105	40,311	50,191	927,171
BILL_AMT6	–339,603	1,256	17,071	38,872	49,198	961,664
PAY_AMT1	0	1,000	2,100	5,664	5,006	873,552
PAY_AMT2	0	833	2,009	5,921	5,000	1,684,259
PAY_AMT3	0	390	1,800	5,226	4,505	896,04
PAY_AMT4	0	296	1,500	4,826	4,013	621,000
PAY_AMT5	0	252.5	1,500	4,799.4	4,031.5	426,529
PAY_AMT6	0	117.8	1,500	5,215.5	4,000	528,666

4. DEFAULT PREDICTION USING CLASSIFICATION ALGORITHMS

In the first stage, models covering all of the data were built. In the basic dataset of 30,000 observations, 22.1% are defaults. It is a typical structure of an imbalanced dataset containing a relatively small positive class and a much larger negative class. The initial imbalanced dataset was analyzed. The balanced dataset will be utilized in Section 5 of this article.

The original dataset was loaded into a database created in Microsoft SQL Server Enterprise 2017 (Microsoft, 2017a). It is one of the world's most important products for managing relational databases, data warehouses and advanced data analytics (data mining). It is basically a group of products that also includes Business Intelligence tools and data analytics tools that are a component of Visual Studio – we used Visual Studio 2017 (Microsoft, 2017b).

These IT tools are treated in this paper as only a typical example of tools for collecting data and advanced data analysis, presenting the capabilities of this type of software. The products of other leading manufacturers of database software and data analysis (e.g. IBM, SAP, Oracle, Teradata, SAS, Rapid Miner, KNIME) have similar capabilities (see Edjlali *et al.*, 2017; Linden *et al.*, 2017).

Using easy-to-use wizards for building data mining models, 3 models were built for different classification algorithms: a classification tree, an artificial neural network and logistic regression. The program automatically and randomly divided the data into a training set and a test set, on which the accuracy of predictions of models built on the basis of the training set is checked.

In the case of the analyzed dataset, which is an imbalanced dataset, apart from the assessment of overall accuracy, it is necessary to use detailed measures of classification accuracy. The assessment of the accuracy of the classification is based on the so-called confusion matrix (Tab. 2).

Table 2. *Confusion matrix*

		Actual class	
		negative (0)	positive (1)
Predicted class	negative (0)	True negative (TN)	False negative (FN)
	positive (1)	False positive (FP)	True positive (TP)

Source: own study based on (Lantz, 2013)

Based on the confusion matrix, the following measures of classification accuracy are introduced (Lantz, 2013).

Accuracy – the ratio of correctly qualified observations to the total sample. The measure, which evaluates the overall accuracy of classification.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

Precision – the proportion of positive observations that were correctly qualified. The measure determines how often a model makes the right choice when predicting a positive class.

$$precision = \frac{TP}{(TP + FP)} \tag{2}$$

Sensitivity (true positive rate, recall) – the relation of correctly classified positive classifications to the total number of positive observations entered into the model. A high value of this measure indicates successful detection of positive cases.

$$sensitivity = \frac{TP}{(TP + FN)} \tag{3}$$

Specificity (true negative rate) – the ratio of correctly classified negative class observations to all negative predictions. This measure shows how the model copes with prediction of negative cases.

$$specificity = \frac{TN}{(TN + FP)} \tag{4}$$

In the further part of the article two indicators will be used: accuracy – measuring the overall accuracy of the classification and sensitivity – measuring the correctness of the positive class classification – default.

Figure 1 shows the results of building a classification tree, which was one of the three compared classification models. The results of the models' performance on the test sample of 1,000 elements are summarized in Table 3. The process of building, processing, and testing the models took only a few minutes.

The data on the diagonals from 0-0 to 1-1 in Table 3, which is the confusion matrix for the three analyzed models, inform about the number of cases of correct classification and is used to calculate the accuracy ratio (ACC).

The second indicator – *sensitivity* or *true positive rate* (TPR) is calculated based on the values in the "Actual 1" column. TPR is a more important indicator in the

case of imbalanced datasets, in which the correct classification of positive observations (default) is more important than the negative ones.

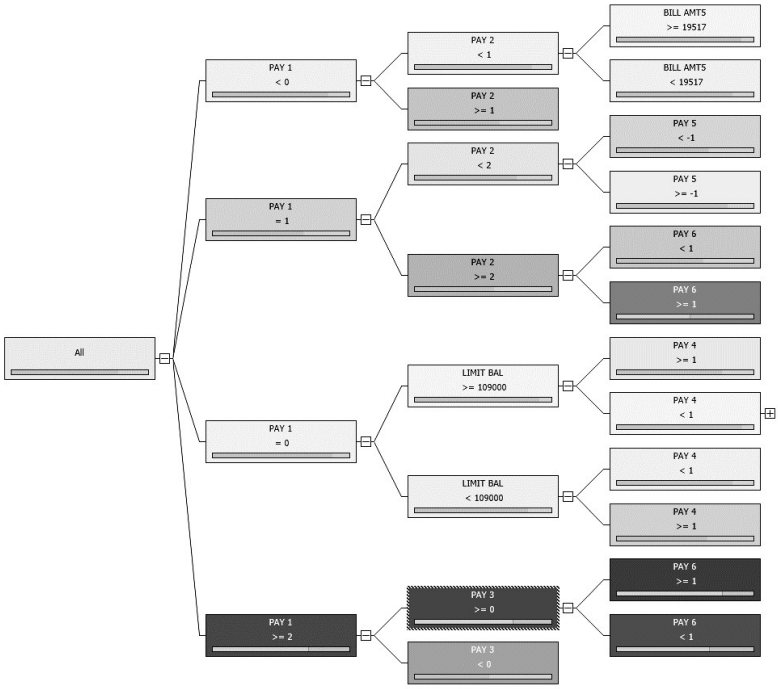

Fig. 1. *Classification tree created in Microsoft SQL Server. The darker the color indicates a higher number of defaults, represented by the right side of the bottom bars inside the rectangles*

Table 3. *Summary of data mining results (test set of 1,000 observations)*

Predicted	**Actual**	
	0	1
Classification tree		
0	768	135
1	41	56
Artificial neural network		
0	771	147
1	38	44
Logistic regression		
0	781	151
1	28	40

The accuracy of the forecasts of all models is high and similar: classification tree – ACC = 82.4%, neural network – ACC = 81.5% and logistic regression – ACC = 82.1%. However, the ability to predict defaults, which should be the primary goal of this type of models, is not very good: classification tree – TPR = 29.3 %%, neural network – TPR = 23.0% and logistic regression – TPR = 20.9%. This may be due to the improper structure of the training dataset, which contains a low number of positive cases, i.e. debt default.

5. LOGISTIC REGRESSION IN DEFAULT PREDICTION

The same dataset was used in the next step to build a logistic regression model following a typical econometric approach. However, it was necessary to transform the qualitative variables into a binary form. The gender has a value of 0 for a man, 1 – for a woman, education – higher education is assumed as a value of 1. The remaining forms of education constitute a reference group and assume 0. "Married" was taken as the reference for marital status, while other states received the value 1. Additionally, the variables PAY_1 – PAY_6 were scaled so that the value of the variables determined the delay in repayment in months from 0 to 10.

The logistic regression model is a generalized linear model that uses the logit, or the natural logarithm of the odds ratio, as a linking function (e.g. Maddala, 2001; Górecki, 2010; Hosmer and Lemeshow, 2000). In a situation where the dependent variable takes only two values, in our case – will pay off credit (0) or not pay off (1), the classical linear regression model cannot be used. For such a dichotomous dependent variable, the regression model is as follows:

$$y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{\mathrm{ij}} + u_i \tag{5}$$

where coefficients β_0 and β_j are parameters of the model and y_i^* is an unobservable or "latent" variable. What we observe is a binary variable y_i, which is defined by:

$$y_i = \begin{cases} 1, \text{ if } y_i^* > 0 \\ \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

The transformation of not observed variable is as follows:

$$Z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{7}$$

where Z_i is an estimated model. From Equation (6) we see that multiplying y_i^* by any positive constant does not change the value of the empirical variable y_i. Therefore, the β parameters can only be estimated up to the positive multiplier. The usual assumption is that $var(ui) = 1$, which fixes the scale of y_i^*. From the above we get:

$$P_i = P(y_i = 1) = P(u_i > Z_i) = 1 - F(-Z_i) \tag{8}$$

where P_i is the probability that the event will occur and F is the cumulative distribution function of the error term u.

If the distribution F is symmetric, then $1 - F(-Z_i) = F(Z_i)$, which gives:

$$P_i = F\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right) \tag{9}$$

The observations of the dependent variable are realizations of the binomial process, the probability of which is given by Equation (9). The form of the function F depends on the assumption about the distribution of the error term u_i. If we assume that F is a logistic distribution, we get the logit transformation, which is as follows:

$$F\left(Z_i\right) = \frac{exp(Z_i)}{1 + exp(Z_i)} \tag{10}$$

hence:

$$\log \frac{F(Z_i)}{1 - F(Z_i)} = Z_i \tag{11}$$

Ultimately, the *logit* model is:

$$\log \frac{P_i}{1 - P_i} = \mathbf{x}_i^{'}\beta = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{12}$$

The left side of the above equation is the logarithm of the *odds ratio*, which is the ratio of the probability that an event will occur to the probability of its non-occurrence.

In the first step, it was necessary to divide the dataset into a training and test sets, which required additional work. Then the training set was used to build the logit model using Gretl software (Kufel, 2020). Initially, some of coefficients of the structural parameters of the model were statistically insignificant at the level of 0.05, hence were removed from the model using the backward elimination method. The final structure of the model is presented in Table 4.

Table 4. *Logit model results for all 30,000 observations. Dependent variable – logit*

Variable	**Coefficient**	**Standard error**	***p*-value**	**Significance**
const	–2.64557	0.093076	<0.0001	*
LIMIT_BAL	–6.24068e-07	1.50726e–07	<0.0001	*
SEX	–0.117495	0.0306529	0.0001	*
MARRIAGE	–0.175321	0.0336843	<0.0001	*
AGE	0.00499464	0.00178313	0.0051	*

Table 4 cont.

Variable	Coefficient	Standard error	*p*-value	Significance
PAY_1	0.578097	0.0176654	<0.0001	*
PAY_2	0.0805973	0.0201638	<0.0001	*
PAY_3	0.0819264	0.0203352	<0.0001	*
PAY_5	0.0521533	0.0178973	0.0036	*
BILL_AMT1	–5.75839e–06	1.13352e–06	< 0.0001	*
BILL_AMT2	3.20824e-06	1.28963e-06	0.0129	**
BILL_AMT5	1.5308e-06	6.61443e-07	0.0206	**
PAY_AMT1	–1.39868e–05	2.29982e–06	< 0.0001	*
PAY_AMT2	–8.6983e–06	1.85019e–06	< 0.0001	*
PAY_AMT3	–3.67426e–06	1.53076e–06	0.0164	**
PAY_AMT4	–4.49954e–06	1.61708e–06	0.0054	*
PAY_AMT5	–3.11652e–06	1.50576e–06	0.0385	**

* – significance at 0.01 level, ** – significance at 0.05 level

The resulting model must still be transformed from the logarithm of the odds ratio to a probability according to the formula (Górecki, 2010):

$$P_i = \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}} \tag{13}$$

which requires additional transformations and workload. In the next step, it was necessary to build a forecasting model for the test data, which was performed in Microsoft Excel.

The accuracy of forecasts achieved using the logit model were similar to the automatic models built into Microsoft SQL Server and amounted to ACC = 81.0%, and in relation to the default predictions, it was only TPR = 23.6%, which gives similar result to the results of automatic models.

6. CLASSIFICATION AND LOGISTIC REGRESSION FOR IMBALANCED DATASET

An imbalanced dataset is one in which the minority class contains much fewer examples than the other classes. Usually, the main goal is to identify examples from the minority class, e.g. borrower's insolvency, bankruptcy, insurance and tax fraud, etc. Classification and machine learning on imbalanced dataset is a considerable problem, because the algorithms optimizing the objective function improve the classification accuracy, and do not take into account the class to which the individual examples belong. Thus, the minority class loses its importance.

The main difficulties in the learning phase for imbalanced datasets result from the fact that:

- training algorithms assume balanced data;
- classification strategies favor majority classes;
- there is a difficulty in distinguishing incorrect (dirty) data from examples from the minority class.

The groups of methods for solving the problem of imbalanced classification found in the literature are (He and Garcia, 2009; Galar *et al.*, 2012; Mahani and Ali, 2020):

- methods of data modification, the so-called external approach where data is processed prior to the use of classifiers; the data is independent of the selected classifier learning algorithm;
- methods of algorithm modification, the so-called internal approach, in which classic algorithms are enriched with mechanisms that take into account class disproportion; this approach uses inductive bias and learning in which only examples from the minority class are taken into account, and examples from other classes are omitted;
- transformations to the cost-sensitive learning task, being a combination of the two previous methods; the input data are modified by assigning them different weights (costs) and the learning algorithms are enriched with mechanisms taking into account different weights assigned to the observations; this method is used in cases where there are significant differences in costs related to wrong decisions;

In the further part of the paper, an external approach is applied consisting of preprocessing (modifying) imbalanced data. Several methods for modifying imbalanced datasets are possible (Maalouf and Trafalis, 2011; Mahani and Ali, 2020), the most common of which are:

- random-undersampling of objects from the majority class. The downside of this method is the risk of discarding potentially important data;
- conscious elimination (Neighbor Cleaning Rule) using the K-NN algorithm of the nearest neighbors; for each example in the dataset, the three closest neighbors (3NN) are found. If an example belongs to a majority class and 3NN points to a minority class, then such example is deleted; if the example belongs to the minority class and the 3NN algorithm misclassifies it, then 3 contiguous cases are deleted;
- random-oversampling, which consists in replicating observations from the minority class through randomized sampling; the downside of this method is that it is more likely to overfit a model, as it makes exact copies of existing examples;
- intelligent sampling by generating synthetic observations based on examples from the dominated class; one of the most popular methods is the SMOTE (Synthetic Minority Oversampling Technique) algorithm, in which for each observation from the minority class a synthetic example is generated using the two closest neighbors (2NN) from the minority class.

In the basic imbalanced dataset of 30,000 observations, only 22.1% are defaults. Therefore, in order to obtain a higher accuracy of forecasting the risk of default, the simplest method of balancing the data set was used – random undersampling. A sample was selected from the entire data set, containing all cases of insolvency and an equal group of randomly selected cases of correct debt repayment. This resulted in a total of 13,272 observations constituting the balanced sample. Table 5 summarizes the results of the predictive quality of particular models for the balanced dataset.

Table 5. *Summary of data mining results for a balanced dataset (test set of 1,000 observations)*

Predicted	**Actual**	
	0	1
Classification tree		
0	401	185
1	99	315
Artificial neural network		
0	422	224
1	78	276
Logistic regression		
0	430	240
1	70	260

The accuracy of default predictions in the case of the automatic models for the balanced dataset is similar for all three exploration techniques used and amounts to: classification tree – TPR = 63.0%, neural network – TPR = 55.2%, logistic regression – TPR = 52, 0%. The total correctness of the model prediction measured by the ACC is equal to 71.6%, 69.8% and 69.0%, respectively.

The accuracy of default predictions in the case of the "manually made" logit model is similar, only slightly worse than the classification tree, and for the balanced dataset is TPR = 61.3%. The overall accuracy of the model classification is equal to ACC = 66.7%. The detailed specification of the logit model is presented in Table 6. A summary for all models and imbalanced and balanced datasets is gathered in Table 7.

Table 6. *Logit model results for balanced dataset containing 13,272 observations. Dependent variable – logit*

Variable	**Coefficient**	**Standard error**	***p*-value**	**Significance**
const	–1.28334	0.109193	< 0.0001	*
LIMIT_BAL	–6.03006e–07	1.80295e–07	0.0008	*
MARRIAGE	–0.237689	0.0423378	< 0.0001	*
AGE	0.00567909	0.00225075	0.0116	**

Table 6 cont.

Variable	Coefficient	Standard error	*p*-value	Significance
PAY_1	0.469484	0.0211087	< 0.0001	*
PAY_2	0.134265	0.0219151	< 0.0001	*
PAY_4	0.0856546	0.0208286	< 0.0001	*
BILL_AMT1	–4.8861e–06	8.01277e–07	< 0.0001	*
BILL_AMT3	3.87778e–06	9.03725e–07	< 0.0001	*
PAY_AMT1	–1.40891e–05	2.27701e–06	< 0.0001	*
PAY_AMT2	–1.11411e–05	1.98576e–06	< 0.0001	*

* – significance at 0.01 level, ** – significance at 0.05 level

Table 7. *Summary of predictive quality of particular models*

Measure	Imbalanced dataset	Balanced dataset
Classification tree		
Accuracy (ACC)	82.4	71.6
Sensitivity (TPR)	29.3	63.0
Artificial neural network		
Accuracy (ACC)	81.5	69.8
Sensitivity (TPR)	23.0	55.2
Logistic regression (automatic)		
Accuracy (ACC)	82.1	69.0
Sensitivity (TPR)	20.9	52.0
Logistic regression (manual)		
Accuracy (ACC)	81.0	66.7
Sensitivity (TPR)	23.6	61.3

Summarizing the results of testing default prediction models, two conclusions can be drawn.

1) Building a scoring model with the use of data mining software via classification algorithms was much faster and required much less analyst involvement than in the case of the manual construction of the logistic regression model. It did not even require specialized financial and econometric knowledge.
2) The prognostic capacity of the automatic classification models were slightly better than that of the logistic regression, but still not sufficiently high for defaults. The reason for this fact may be a small set of potential predictors, which lacked demographic data related to work, material situation, residence, family size or others.

7. CONCLUSION

The analysis of the capabilities of software for managing databases and data warehouses, as well as data mining in the assessment of credit risk showed a great usefulness of this type of tool. The automatic construction of scoring models was simple and fast, and the accuracy of the forecasts of these models ware even slightly higher than that of the logit model built by a analyst without the use of automatic tools. The classification tree showed the highest accuracy of predictions of all models. Moreover, the manual implementation of a scoring model required much more work in relation to the models created automatically. Based on the presented case study, the research hypotheses were confirmed.

We used a preprocessed dataset, but the software has built-in ETL tools (ETL – extract transform load), which enable the automation of the process of updating a dataset with new data, which further simplifies the entire process of credit risk assessment. It becomes possible to frequently update the scoring model with the inflow of new credit data.

The answer to the question, whether commercial machine learning tools can be an effective tool for credit risk management, seems to be in the positive. For retail loans for which credit institutions have a large number of credit histories, automating credit risk management is relatively straightforward. The prognostic effectiveness of IT tools, their ease of use and a high degree of automation allow for the creation and frequent updating of scoring models without the need to involve a large group of highly qualified staff. These tools can generate a "black box" model from any broad set of potential predictors, including macroeconomic, political and social, which may lead to the elimination of human involvement in much of the work related to credit risk assessment.

The automated attitude to credit scoring analyzed in the paper can be extremely convenient for enterprises which face problems of granting trade credit to customers. Without professional credit staff it is possible to manage credit risk arising from sales with deferred payment.

Further research on the effectiveness of machine learning methods in scoring should include models with better selected learning parameters – the depth of the decision tree and the number of layers and neurons in the neural network. Moreover, the econometric models used for comparison can be supplemented with more modern models such as generalized partial linear models (GPLM, CGPLM, RCGPLM – see Özmen and Weber, 2012) or the Granger panel bootstrap causality approach (Kawa *et al.*, 2020). The next research question is whether scoring models built with the use of commercial software tools are not too "sensitive" to short time series. However, the answer to this question requires further research.

REFERENCES

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), pp. 589–609.

Edjlali, R., Ronthal, A., Greenwald, R., Beyer, M., Feinberg, D., 2017. *Magic Quadrant for Data Management Solutions for Analytics. Gartner*, https://www.gartner.com/home.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Systems, Man, and Cybernetics Society*, 42(4), pp. 3358–3378.

Górecki, B.R., 2010. *Ekonometria. Podstawy teorii i praktyki*. Wydawnictwo Key Text, Warszawa.

Han, J., Kamber, M., Pei, J., 2012. *Data Mining: Concepts and Techniques*. Elsevier, San Francisco.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp.1263–1284.

Holdaway, K., 2014. *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models*. John Wiley & Sons, New Jersey.

Hosmer, D.W., Lemeshow, S., 2000. *Applied Logistic Regression*. John Wiley & Sons, New York.

Jagiełło, R., 2013. Analiza dyskryminacyjna i regresja logistyczna w procesie oceny zdolności kredytowej przedsiębiorstw. *Materiały i Studia NBP*, No. 286.

Jonc, A., Kraska, M., 2001. *Credit-scoring. Nowoczesna metoda oceny zdolności kredytowej*. Zarządzanie i Finanse, Warszawa.

Kawa, P., Wajda-Lichy, M., Fijorek, K., Denkowska, S., 2020. Do Finance and Trade Foster Economic Growth in the New EU Member States: Granger Panel Bootstrap Causality Approach. *Eastern European Economics*, 58(6), pp. 458–477.

Keramati, A., Yousefi, N., 2011. A Proposed Classification of Data Mining Techniques in Credit Scoring. *Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management. Kuala Lumpur, Malaysia, January 22–24*.

Kufel, T., 2020. *Gretl*. http://www.kufel.torun.pl/ [20.11.2020].

Laitinen, E.K., 1991. Financial ratios and different failure processes. *Journal of Business Finance & Accounting*, 18, pp. 649–673.

Lantz, B., 2013. *Machine Learning with R*. Packt Publishing, Birmingham.

Larose, D., 2006. *Odkrywanie wiedzy z danych*. Wydawnictwo Naukowe PWN, Warszawa.

Li, M.Y.L., Miu, P., 2010. A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. *Journal of Empirical Finance*, 17, pp. 818–833.

Linden, A., Krensky, P., Hare, J., Idoine, C., Sicular, S., Vashisth, S., 2017. *Magic Quadrant for Data Science Platforms*, Gartner, https://www.gartner.com/home.

Maalouf, M., Trafalis, T., 2011. Rare events and imbalanced datasets: an overview. *International Journal of Data Mining, Modelling and Management*, 3(4), pp. 375–388.

Machine Learning Repository, 2020, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients# [20.11.2012].

Maddala, G., 2001. *Introduction to Econometrics*. John Willey & Sons, Chichester.

Mahani, A., Ali, A., 2020. Classification Problem in Imbalanced Datasets. In: Sadollah, A. (ed.), *Recent Trends in Computational Intelligence*. IntechOpen, London.

Matuszyk, A., 2013. *Credit scoring*. CeDeWu, Warszawa.

Microsoft, 2017a. https://www.microsoft.com/pl-pl/sql-server/sql-server-2017-editions [20.10.2017].

Microsoft, 2017b. https://docs.microsoft.com/en-us/visualstudio/ide/visual-studio-ide [20.10.2017].

Moradi, S., Mokhatab, M., 2019. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financial Innovation*, 5(15), pp. 1–27.

Özmen, A., Weber, G.-W., 2012. Robust conic generalized partial linear models using RCMARS method – A robustification of CGPLM. *Global Conference on Power Control and Optimization*, Dubai, UAE, 1–3 June 2011.

Paliński, A., 2018. Loan Payment and Renegotiation: The Role of the Liquidation Value. *Argumenta Oeconomica*, 1(40), pp. 225–252. Doi: https://dx.doi.org/10.15611/aoe.2018.1.10.

Rębiasz, B., Gaweł, B., Skalna, I., 2017. Hybrid Framework for Investment Project Portfolio Selection. In: Pełech-Pilichowski, T., Mach-Król, M., Olszak, C. (eds.), Advances in Business ICT: New Ideas from Ongoing Research. *Studies in Computational Intelligence*, vol. 658. Springer, Cham. Doi: https://doi.org/10.1007/978-3-319-47208-9_6.

Sadatrasoul, S., Gholamian, M., Siami, M., Hajimohammadi, Z. 2013. Credit scoring in banks and financial institutions via data mining techniques: A literature review. *Journal of AI and Data Mining*, 1(2), pp. 119–129.

Weber, G.-W., Çavuşoğlu, Z., Özmen, A., 2012. Predicting default probabilities in emerging markets by new conic generalized partial linear models and their optimization. *Optimization*, 61(4), pp. 443–457.

Wilcox, J.W., 1973. A prediction of business failure using accounting data. *Journal of Accounting Research*, 11, pp. 163–179.

Yap, B., Ong, S., Husain, N., 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, 38, pp. 13274–13283.

Yeh, I.C., Lien, C.H., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), pp. 2473–2480.

Zmijewski, M.E., 1984. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 22, pp. 59–82.