ECONOMY AND FINANCE • VOL. 11. ISS. 3 • SEPT 2024 • DOI: 10.33908/EF.2024.3.1

ARTIFICIAL INTELLIGENCE FOR CREDIT RISK MODELS, OR HOW DO MACHINE LEARNING ALGORITHMS COMPARE TO TRADITIONAL MODELS?

László Rajka – Zoltán Pollák[1]

ABSTRACT

A new generation of credit risk management models has surfaced as a result of the technology revolution marked with artificial intelligence, which in short is a term for models based on machine learning. Expert systems represented the past in the development of credit risk models over some decades, while traditional statistical models, e.g., logistic regression are the present and machine learning methods are expected to be the future. The objective of this study is to describe and empirically analyse the classification algorithm XGBoost, one of the most promising examples of the latter machine learning models to reveal the degree of increase in efficiency machine learning algorithms can achieve compared to the traditional modelling methods currently regarded to be industrial best practice. In our study, both Artificial Neural Network (ANN) and XGBoost, models relying on artificial intelligence, have surpassed logistic regression in terms of efficiency of classification. Although machine learning methods have an excellent capability of prediction, the interpretation of decision-making models they offer is quite cumbersome compared to their traditional peers, which is a disadvantage. Because of the "black box nature" of machine learning methods based on artificial intelligence, banks are currently limited regarding their application. Therefore, the authors propose the current rules and guidelines corresponding to the traditional models should be reviewed so as to give way to banks for the application of machine learning models and, as a result, to improve the efficiency of their credit risk management.

JEL codes: C13, C25, C51, C53, C58, G21

Keywords: artificial intelligence, machine learning, application scoring, XGBoost, logistic regression, probability of default

1 *Rajka, László,* credit risk analyst, OTP Bank. E-mail: laszlo.rajka@gmail.com.
Pollák, Zoltán, corresponding author, associate professor, Budapest Business University, Faculty of Finance and Accounting, Department of Finance. E-mail: pollak.zoltan@uni-bge.hu.

1 INTRODUCTION

Applying scoring models having as far as possible the best prediction abilities is key for banks, since the management of credit risk by detecting potential non-performing borrowers is the number one risk for them.

The factor of profitability has been one of the drivers of ongoing development of credit risk systems over the last decades. Credit risk assessment in the second half of the 20th century was still dominated by expert systems including subjective components. Their hegemony was broken by traditional statistical models as information technology improved and increasingly efficient algorithms were offered with flagship logistic regression-based scoring systems becoming industrial best practice.

As artificial intelligence has shot forward, other revolutionary changes have become visible. Attention is directed to a new generation of credit rating scoring systems relying on machine learning. As technological development reaches new milestones, new algorithms relying on machine learning appear. One of them, XGBoost seems most promising, so it has been in the focus of this research.

Side by side with profitability, the increasingly sophisticated regulatory requirements of prudent banking operations have been driving the explosive development of credit risk management. Some of the crises of the past decades could be associated with credit risk, so international regulatory bodies have started to focus on the models mentioned above (MNB, 2002).

Internal Capital Adequacy Assessment Process (ICAAP) is part of the Basel capital adequacy calculations and banking risk management. According to it, banks apply internal procedures to assess their risks and quantify the capital required for coverage. Assessing the *probability of default* (hereinafter: PD) is the foundation of the procedure. Credit institutions are required not to apply case-by-case solutions but assess credit risks based on uniform modelling (MNB, 2018).

Reviewing the literature, one finds banks still mostly rely on traditional techniques, since their interpretation is easier for the business (and also for the regulators). Meanwhile machine learning methods have appeared and gained momentum in credit risk modelling, but their decision-making mechanism is much less transparent.

In simple terms one can say expert systems represent the past in the history of credit risk models, traditional statistical models are the present and machine learning methodology is expected to be the future.

The objective of this study is to introduce the classification algorithm XGBoost belonging to the last group and to analyse it on a real credit card database to learn

how predictions can be more efficient using machine learning algorithms compared to the currently dominant traditional modelling methodology.

Accordingly, our hypothesis is the following:

Machine learning methods including particularly XGBoost are able to surpass, in terms of classification, the performance of the traditional statistical models currently most widely used in industrial practice.

It should be noted this study is not aimed at generating the most effective model on the database studies. Modelling in our case is simply for illustration and, in addition to introducing the model, the authors focus on the difference of efficiency if the model is used on the same sample under the same circumstances.

The authors also touch upon the fact that while models relying on machine learning may be more efficient, their application raises further problems. Thus, side by side with their efficiency, the aspects of use and application of the models are also discussed.

Following the introduction, the relevant literature is briefly summed up including the findings of recent research studies as well as the requirements of international organisations regarding the models. In the second part of the study, each modelling methodology is discussed in detail including the metrics needed for modelling and their assessment, and finally, the findings received using different methods are compared.

2 LITERATURE REVIEW

In this chapter the currently available directions and modelling methodologies in the literature of PD modelling are presented. The development of the models, industrial best practices and the most effective methods of estimation are discussed. Also, the current guidelines and requirements regarding modelling procedures by EBA, the National Bank of Hungary (MNB) and the Basel standards are presented. So, the purpose of the literature review is to provide side by side the current views of the academia, practice and regulators on the topic.

2.1 Development of modelling methodologies

In the 1980s most financial institutions carried out credit risk assessment relying on subjective analysis or, in other words, expert opinion. Decision-making was totally subjective based on internal judgement analysing aspects such as reputation, capital, the debt portfolio or coverage (Altman–Saunders, 1997).

Similarly, Sommerville–Taffler (1995) found that bankers' subjective judgement was characteristic in risk assessment although the institutions using multiple-variable credit risk scoring systems were more effective in their operations. By the end of the century, financial institutions had more and more distanced themselves from subjective expert systems, so methodologies based on historical data gained momentum.

At the beginning of the 2000s, studies started to appear trying to improve the efficiency of the estimated probability of non-performance in traditional models. In his paper, West (2000) compared several neural network-based models using logistic regression. Although the model based on logistic regression provided better estimation than linear analysis, he found in his study that the neural network-based modelling had the best classification capabilities.

After the early 2000s, as information technology developed, more and more studies were published, which tried to further improve the accuracy of modelling. Huang et al. (2007) approached the methods of PD estimation related to credit card placements from the aspect of data mining tools. They used a hybrid version of classification algorithm SVM *(support vector machine)* with promising results compared to the methods of estimation applied at the time.

A paper by Yu (2020) also analysed the efficiency of credit card defaults. He also included the score figures of the American FICO in his modelling. He found methods of machine learning are more efficient or can provide more accurate estimations. In his analysis the random forest algorithm has proved to be the most accurate surpassing Adaboost, decision trees or logistic regression. The paper also underlined that the quantity of available data is key for modelling credit risk.

Angelini et al. (2008) described the options of using neural networks from the aspect of credit risk assessment. In addition to artificial neural networks, the authors also studied programmed learning mechanisms. They reported high efficiency from the aspect of accuracy of estimation in the case of both methods. They also found major faults in risk assessment in the model unless noisy data are filtered out. So, they suggest some kind of normalisation should be performed before modelling procedures are used. The study also states estimations can be further refined by optimising the parameters. The paper also suggests one should retain the traditional regression procedures and apply neural networks as their supplement and to refine the estimations.

Medema et al. (2009) supplemented the validation guidelines linked to the Basel Capital Conventions with an empirical study. Accordingly, a good PD model must be valid both in theory and in terms of data and statistics. In the study, the authors proposed a parameter vector applicable to validate PDs. They also underlined that missing data must be given special attention, further, the out-of-

sample efficiency of the models must also be considered. They proposed external databases to be involved or a bootstrap procedure to be applied.

Gurný–Gurný (2013) compared the performance of traditional logit, probit and LDA models in their paper. The study is unique. The models had been trained on the data of 298 US banks and then their efficiency was measured on the data of 100 different commercial banks. So, they tried to set up a model for the whole banking system, which would predict banking defaults in general. The model provided predictions for short-term (1 to 2 years) non-performance, and the logit model offered the most efficient predictions with an impressive ROC curve. The information used in the model consisted mainly of financial data, their ratios or logarithmic values. They also analysed in the study what delay should be applied in the risk models from input data to non-performance.

Butaru et al. (2016) studied the application of machine learning algorithms measuring the performance of credit risk models on a large credit card database containing the data of six banks. They used data from 2009 to 2013 to compare the results of decisions based on a decision tree or logistic regression. They found decision tree-based models explained delay better than the logistic regression procedure. In the study they compared the risk management practices of different banks and found there was a high degree of heterogeneity among the banks in terms of risk factors and risk sensitivity as a result of their businesses. Thus, they stated one cannot easily identify a model describing the whole banking system. They underlined the characteristics of a loan portfolio are not always sufficient to identify delays, since banks actively manage their portfolios all the time.

In their study, Sirignano et al. (2016) explored the data of mortgage loans in the USA from 1995 to 2014. The database used for the analysis was exceptionally large with 120 million observations including both loan and borrower specific as well as macroeconomic variables. They found the relationship between borrowers' behaviour and risk factors cannot be considered linear. They observed interaction of the variables studied in several cases. Using an out-of-sample analysis they proved the management of non-linearity of the data can significantly improve the accuracy of loan and pool level risk predictions, the investment performance of mortgage trading strategies and the assessment and coverage of mortgage loans. As for modelling results, they found neural networks surpassed the results provided by logistic regression.

Venkatesh–Jacob (2016*)* looked into the prediction power of algorithms BayesNet, Meta-Stacking, Naive Bayes, Random Forest, SMO and ZeroR from the aspect of credit risk. They used the database of the University of California for the study, which included the credit card data of a Taiwan bank. In addition to financial information in the database, they also used some information linked to borrowers for their modelling. They established in the study that both informational value

and correlational analysis can be effectively used to select explanatory variables in modelling credit risk. Their finding was that random forest, random tree and IBK classification algorithms provided the best prediction results, but the other methods also work at nearly 80% accuracy.

Addo et al. (2018) analysed the performance of deep learning algorithms for credit risk modelling. They compared the dominant models in the industry including logistic regression, random forest, gradient boosted models, and neural networks in the study. They found models based on decision tree provided more stable estimates than multi-level neural networks also surpassing traditional regression procedures. The authors explained it is important to check the quality of the training data, since it may cause distortions in terms of the setup of the training classes. They underlined the comparison of models should be looked at from several viewpoints during validation, since the comparison of different models is difficult. For instance, AIC, BIC and R^2 indicators cannot be interpreted for all models. So, they used the ROC curve and AUC and GINI values derived from it for evaluation. The authors emphasised in the paper the relevant regulations related to machine learning algorithms should be established as soon as possible to avoid the mistakes of using black box solutions.

The paper by Moradi–Mokhatab (2019) proposed a Fuzzy procedure to evaluate credit risk as opposed to the previous models. They believe the models do not usually use external data changing dynamically from one month to the next properly, such as political or economic sanctions. In their study, they analysed bad/dubious clients every month dynamically and tried to detect those that had met with difficulties because of the adverse effects of economic cycles. The risk model proposed by the authors can be used in credit rating as it includes information in case of bankruptcy or if hyperinflation appears.

Wang et al. (2023) presented an in-depth analysis of procedures aimed at selecting the characteristic features applied in credit risk models from the aspect of machine learning methods, in addition, they also presented risk assessment models in their study. Their research on the selection of the most efficient model has come up with a different result compared to the papers presented above, since the conventional procedure of logistic regression surpassed the predictions by XGBoost or the ones based on decision trees.

The paper by Rozo et al. (2023) is particularly interesting. The authors studied the impact of the world in lockdown because of the Covid-19 pandemic in terms of modelling. In the model, they used search data, the number of website visits and visits to bank branches as variables in the methodology of PD appraisals. They found they could reduce estimation errors in the models set up using traditional PD methodology if they used the data of web behaviour. Another finding they described was there was no significant difference in age among active web users

in the sample, so the data could be used properly in the estimation performed on the whole sample.

Reviewing the literature one can say modelling procedures have evolved significantly as the relevant technology developed. There have been major improvements and diversity both in terms of efficiency and modelling techniques over the past two decades. The quantity of available data and their informational power is also expanding, so models can start out of a wider range of data points.

With respect to the methods of machine learning, the research of the algorithm XGBoost and the areas of its application are more and more in the limelight abroad, as seen from our literature review, however, it is completely missing from the literature in this country. This paper aims to fill the gap and to open up a new path for researchers of the topic in Hungary.

2.2 Industrial practice and regulatory guidelines

As the volume of retail lending grows and competition is getting fierce, banks need to develop more accurate models to minimise their credit losses. High volumes of data allow banks to apply efficient methods relying on internal credit risk rating.

In addition to defining capital requirements, PD models play an important part in supporting lending decisions.

Existing and new customers must be differentiated from the aspect of credit risk. There are a lot of data about existing customers that are updated regularly, while a credit institution has little information about new customers. Banks can mainly rely on sociodemographic data for new loan applications, or – provided they have historic interbank information – they can use them too. Credit rating systems for new customers are termed application scoring, while credit institutions apply behavioural scoring for existing customers where they can also use the repayment data of earlier loans.

Logit-based models are quite popular for PD estimation in industrial practice, as logistic coefficients can be transformed into probability values. Models estimated using logistic regression can easily be changed into scorecards, where estimated probability values are converted into scores along a pre-defined scale.

During the modelling process, weight of evidence (WoE) values generated from default cases are grouped (Siddiqi, 2006). Those WoE values represent the rate of default within a sub-population compared to the whole sample.

The evaluation of the scoring models thus established is particularly important. Procedures often used include Kolmogorov-Smirnov (KS) statistics, the ROC

curve and the GINI value generated from it (Kovács–Marsi, 2018). They will be used later for comparing the efficiency/effectiveness of the models.

Banking scoring systems typically work in a hybrid manner, i.e., in addition to application and behavioural models, credit risk models can be differentiated based on their PD character. Banks often apply both *Point in Time* (PiT) and *Through-the-Cycle* (TTC) models at the same time to assess risks, therefore, the authors intend to present the difference of the two types of PD.

Estimation in a PiT system depends on the actual phase of an economic cycle, so its variables may be cyclical. As a result of using cycle-sensitive indicators in a PiT system, some transactions are mass migrated into lower rating categories during an economic recession as the values of the indicators worsen. The capital requirement linked to PiT systems fluctuates in time due to its „Point in Time" nature.

In TTC systems the approach to estimation (through the cycle as in the name) is much more expressed. Using TTC systems, institutions try to cleanse non-performance risk of the volatility caused by economic cycles and to measure customer risk during the cycle. A TTC rating will not respond to the changes of the economic cycle, changes in fundamental features only can cause migration. Applying TTC systems can lead to capital requirements to be more stable in time.

According to current market practice relating to CRR-based capital calculations, credit institutions are advised to design their scoring systems so that they guarantee the stability of estimation of their rating categories through time and economic cycles. Estimation of TTC PD based on internal rating is typically easy in the retail segment, since corporate portfolios can hardly be considered without cleansing them from economic cycles (Bíró–Nagy, 2018).

According to the recommendation of the European Banking Authority (2017) (EBA), PD modelling has many components to be considered during development. If you use statistical models, you must consider all circumstances that can be relevant for assessment. According to the recommendation by EBA, using ratings from a third party can also carry risks, so they must be used with care. You should clarify, before building your model, the relevant data requirements and the timeframe of modelling. During your analysis, you should also pay attention to how often the data are updated. The guidelines state business experts should also be involved side by side with a statistical approach, so that the information used is properly implemented in the different estimation models, in addition, a person's identical loans must also be managed.

Naturally, the Bank for International Settlements (2023) (BIS) also has minimum requirements if a credit institution wants to apply a methodology relying on internal rating. According to the Basel Guidelines, credit rating models often use mechanical classification procedures to provide an estimation on the probability

of default. The models also support the business, since the lending process can even be automated using them. However, it should be noted the models are not perfect, so you must be prepared for inaccuracies of estimation in the classification. Therefore, the models used must be constantly monitored and any out-of-scope information must be taken into account.

BIS (2023) provides no detailed guidelines in terms of methodological expectations regarding models based internal rating. Its guidelines are general statements. According to it, banks must present their own definitions for non-performance and loss used internally and must prove their consistency with the definitions identified in the Basel standards. If a bank applies statistical models for rating, it must provide detailed documentation on its methodology. The documents must comply with the following standards:

1. A model must exactly define the mathematical and/or empirical foundations for the estimation for categories, the estimations for single obligors and of the theory or assumptions of how they are assigned to exposures or pools; they must also present the datasource(s) used for the estimation of the model.
2. Strict statistical procedures must be established to validate a model (including performance tests relating to other periods as well as out-of-the-sample i.e., independent sample performance tests).
3. The documents on statistical models must specify the circumstances that could prevent the model from operating effectively.

The above bullet point list of guidelines are further detailed in a Manual published by the National Bank of Hungary (MNB) (2018). To sum up, one can say they emphasise very similar principles in compliance with the documents of BIS (2023). They require full scale documentation, exact identification of concepts and procedures, assurance of proper data quality, annual validation and regular reports on the performance and stability of the models. With respect to models, an updated inventory must be established, and suitability must be verified with continuous backtesting. Next, the models must ensure relevant risk distinction, and changes in portfolio quality need to be reported.

The guidelines emphasise the categories identified by rating must reflect monotony, i.e., customers with higher non-performance rate have to be assigned to the lower rating categories, while higher rating categories have to be defined for lower risks. In addition, credit institutions must measure the migration from one rating category to another from time to time including monitoring if there is real deterioration in the background of the changes or if the process can be explained by reasons originating from the errors of the PiT system. The capital requirements Manual stipulates the application of TTC is a must if PiT rating systems are used.

The MNB states with regard to its guidelines all models must comply with the guidelines of the ECB and of BIS related to internal rating systems.

One should mention the forward-looking activity of international regulatory authorities too. The EBA (2020) published a report on the necessity of the institutional application of Environmental, Social, and Governance (ESG) risks. The report underlines using ESG risks is an important point for defining portfolio quality, but its use can be cumbersome in the case of models based on historical data such as PD or LGD estimations. The application of ESG is still in its infancy in credit risk analysis, however, some studies have already been published to prove there is a link between ESG events and non-performance (Henisz–McGlinch, 2019).

3 MODEL BUILDING AND EVALUATION OF FINDINGS

In this chapter the modelling procedures used to estimate PD are introduced. In the first part, the database used is presented with a short statistical description. Next, the evaluation techniques are presented to make the results of the models comparable. After that, in accordance with industrial best practice, the results of a logistic regression model using grouped variables are described. Finally, the efficiency of the so termed "black box" solutions is analysed from the aspect of estimation. The analyses were run in Python3 environment (Rossum–Drake, 2009), relying mostly on the Scikit-learn library (Pedregosa et al., 2011). Our goal with the research is to find the procedure with the minimum error among the models predicting credit card defaults. For that purpose, an application PiT PD model is specified, then its accuracy is compared with a ROC curve, the AUC value and the GINI index on a test sample.

3.1 Introduction and descriptive analysis of the database

The modelling procedures and comparisons in the study were established on a detailed credit card database with numerous variables *Kaggle (n.d.).* It contained 122 variables with non-performance as its target variable. Credit default was given as a binary variable where the default rate was 8% rounded on the whole sample.

As described above, banks use both application and behavioural models to estimate probability of default. The database mostly included application information (i.e., it consisted of application data), but it also contained data from earlier applications. Application data mostly meant the data of loan applicants and their property elements supplemented with some other information from external sources. Since earlier applications were only available for less than half of the loan

transactions, they were excluded from our analysis. Further, any other variables were excluded where the rate of missing data was higher than 30%. 26 variables were excluded using the first filter while the 30% missing data reduced the database to 50 variables. Next in the course of building our model, the observations with missing data for the selected variables were deleted. The initial sample included 307,511 records, which was reduced to 263,947 following the exclusions.

Loan size was a key variable, which contained 5,603 individual observations. The lowest loan amount was 45,000, the highest 4,050,000 units (the data base did not include the currency). Loan size distribution took a shape with a slope to the left and stretching to the right, i.e., the loans issued were typically of low value and their frequency gradually declined as loan sizes increased. The average of the loans disbursed was close to 600,000 units with 404,312 unit spreads. The findings are presented in Table 1.

The next key variable of the analysis was age at the time loan disbursement. In terms of the distribution of the variables, all age groups evenly occur in the sample; the oldest credit card user was about 70 years old while the youngest was 21.

Table 1
Descriptive analysis of key variables following selection

	No of days since change of previous telephone	No of days since change of previous ID card	Loan size	No of days since registration data changed	Income size	No of days since birth
Average	–988	–3 049	606,587	–4,990	171,124	–16,120
Spread	833	1,491	404,312	3,523	249,021	4,308
Minimum	–4,185	–7,197	45,000	–24,672	26,100	–25,201
25% percentile	–1,603	–4,319	272,579	–7,479	112,500	–19,716
50% percentile	–798	–3,335	521,280	–4,517	157,500	–15,816
75% percentile	–286	–1,814	813,195	–2,007	202,500	–12,574
Maximum	0	0	4,050,000	0	117,000, 000	–7,489

Source: own design

In terms of total income, the minimum was 26,100 while the highest income was 10 to the eighth magnitude ($10^{8)}$. Naturally, the reality of such an outstanding value can hardly be decided, but the problem of outlier values was managed in the course of modelling using WoE grouping. As for income, 171,124 units were the average with 249,021 unit spread. Like with the loan amounts granted, the sample was quite heterogeneous.

One could observe close positive relationship in the sample between the loan amounts granted and income. In other words, customers having really high income used the credit card limits granted to a higher extent. It is not really surprising, as banks can usually provide high-income customers with bigger credit card limits, so the amounts to be drawn and used were also higher in their cases.

A piece of interest might be the number of days since the last cell phone replacement, which had become a significant variable in the model. The two extremes were 0 and 4,185 days. It is interesting to imagine a customer using a credit card who did not change his cell phone for 10 years, but it was, in fact, an outlier value. In the sample, customers purchased their cell phones 988 days earlier on average, which means approximately 2.5 years. The spread of the number of days since buying the last phone was 833 days, i.e., a spread of 2.26 years belongs to the average 2.5 years with respect to phone swap.

In the sample, the next interesting variable was the number of days spent in the current job. The number days worked was more or less evenly distributed, however, there were quite many data errors there, since values with an opposite sign appeared in the database. The problem was managed via selection for the model.

The database also said how many days passed since the debtor changed their ID cards. The distribution was more or less even with respect to ID card replacement, but older ID cards had a higher ratio in the sample. The customers in the sample replaced their ID cards 8 years ago on average with a spread of 4 years.

The value of the goods purchased with the loan was another variable. Like income and loan amount, it took a shape slanting to the left and spreading to the right. The average of the variable was 606,587 units with a spread of 404,312 units. It should be noted that the loan amount was a subset in the value of goods purchased with the credit card and the correlation between the two variables was high. This indicates customers withdrew higher loan amounts from their card limits to buy higher-value goods. Filtering had to be performed for regression because of the high correlation, which will be explained in detail discussing the steps of modelling.

3.2 Methodology of comparing the models

Before going into details about modelling results, the metrics used to compare their performances are described. As mentioned in the Introduction, the objective of the study is to present how much efficiency can be improved by applying machine learning methods compared to the industrial best practice. For the models to be comparable, a well described uniform indicator should be identified, which is suitable to compare estimation performance.

The Kolmogorov-Smirnov (KS) statistics in the validation of credit risk models is a tool assisting evaluators to identify how a model performs in predicting distribution variables (for instance, loan or credit risk indicators). KS statistics is a non-parameter test to find out whether or not two datasets are significantly different. KS statistics analyse a portfolio cut in two distinct parts along a cutoff (lying between the defaulted and non-defaulted transactions estimated by the model). The two cumulative distribution functions are compared and the maximum value between the two curves provides the KS statistics (Madar, 2015).

To evaluate credit risk, the *Receiver Operating Characteristic* (ROC) curve and the area under the curve (AUC, *Area Under Curve*) are often used, as well as the GINI coefficient/factor generated from them. You need the PD, or the score established from it as well as the real default figures to create the ROC curve. Next, you must establish the exact hit rate for each cutoff value for defaults and non-defaults using the PD values estimated by the model. The so termed *hit rate* and *false alarm rate* provide the two axes for the ROC curve. The gradient of the ROC curve will also illustrate how good a given model is, since it is the one separating real defaults and non-real ones.

It should be noted you need to identify a tangible coefficient, which can measure discrimination power well, to compare the models properly. AUC is perfect for the purpose, showing the area under the ROC curve, thus, the larger the area under the curve is, the better the given model is. For comparison the value of AUC may be adjusted according to the following formula: $GINI = 2 \times AUC - 1$. In the case of this metrics, the highest value will be linked to the model having the best separation power (Madar, 2015).

3.3 Results of traditional modelling

Before going into details regarding the results of traditional modelling, the relevant methodology needs to be described. The variables must be reviewed, or, in the case of regression, they must be categorised. The categorisation allows making more accurate risk predictions by using the default rate.

Weight of Evidence (WoE) is one of the most popular procedures. The optimum is searched in the logistic space and the variables are provided with categorised values. WoE grouping was performed before logistic regression so as to observe non-linear effects. To identify bins, a decision tree-based segmentation was selected further refined by modifying the threshold values if necessary. So, using a statistical approach, the essence of WoE cuts is to construe subpopulations from a large population applying certain information, relying on the information on non-performance.

Using WoE has its advantages, since the substituted values received will retain the non-linearities of the variables. Additionally, using WoE values is appropriate because the next aggregation of the groups construed will identify an information value, which can be used to compare the power of different variables.

The information value should always be assessed separately for any given modelling problem, where a higher value indicates a variable with more explanatory power. The WoE values received can be used to generate a new dataset holding WoE values rather than the original value set. In practice it means to use WoE values linked to the band value of the group they belong to for further modelling (Kovács and Marsi, 2018).

Logistic regression was selected out of traditional industrial practice, and it was used to make an estimation on the probability of non-performance (default). Logistic regression is a statistical method often used to solve problems with categorical or binary outputs. The method is usually applied for categorisation when the goal is to identify the category of a given observation on the basis of one or more than one input variable. It models probability of output by means of the logistic function according to the input variables. The "S" shaped logistic function using binary classification returns probability values from 0 to 1, which can be used to interpret binary problems. The logistic function can be written as follows:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_n x_n)}} \quad (1)$$

where $p(x)$ is the probability of an event, while β_0 provides logit/log odds value if input variables are 0. β_n represents weights belonging to the input variables (coefficient) indicating how much the values of input variables affect output probability. x_n represents the input variables used by the logistic regression model to estimate the probability of event occurrence.

An important step of the application of logistic regression is to transform a probability value into a binary variable. It is done by identifying a threshold value and regard an event to have occurred if it is above the threshold or to not have occurred if it is below it. A probability value can be used well because the measure of the threshold also has financial relevance if the problem is linked to an economic event (Peng et al., 2002).

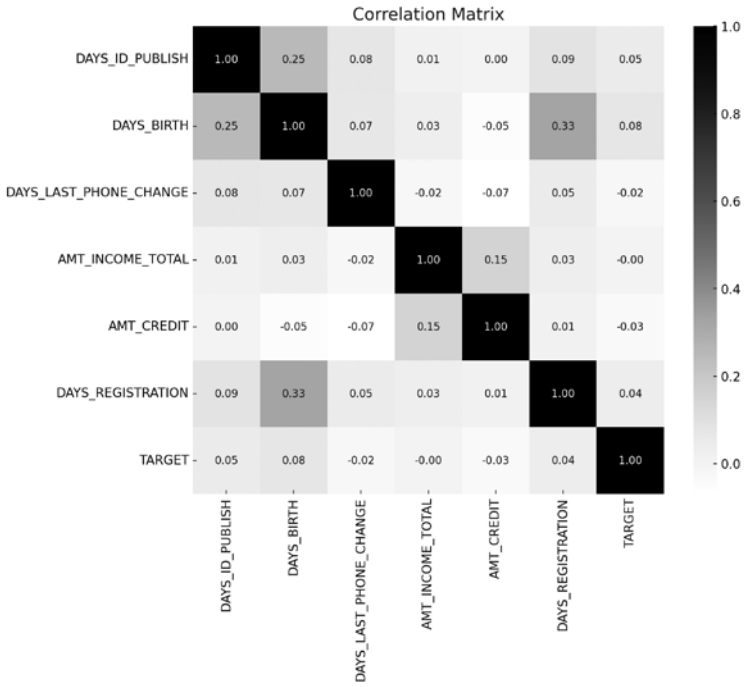
To start model building, a training and a test sample had to be construed. The training sample was defined at 70% while the test sample at 30% by random selection. Next, the variables that will be suitable for estimation had to be selected from the many variables. The scorecardpy python library was used for that (Shichen, 2023). The library automatically mapped the groups for all the variables, and after cutoffs, selection could be performed using information value. Following expert

advice, it was decided to leave in the analysis all variables with information value higher than 0.08. In the end, 10 variables were selected.

There is high correlation between the size of instalment and the price of the product bought with the credit card and the loan amount disbursed; in addition, there was strong correction between the number of days since birth and the number of days spent at current job. Retaining variables with high correlation in the model may result in over representation in a regression procedure, so those variables were omitted (Peng et al., 2002).

Using the remaining variables, the thermal image in *Figure 1* was construed for correlation. You can see the correlation between the variables retained is not so big as to result in distortions of estimation. It should be noted the relationship between the variables retained and the target variable can be said to be quite low, as even the highest value has a correlation coefficient of 0.08, so the explanatory power of the model is expected to be low. In this study our goal is to compare the efficiency of different models, so even if prediction power is weak, the performance of more efficient estimation models will be visible.

Figure 1
Correlation of variables in logistic regression



Source: own design

One can also see that elderly people have held their ID cards for longer, which is not a surprise as – in line with local regulations - young ones may have to renew their cards more often. The closeness between the other variables can be regarded to be negligible.

The data were regrouped before modelling as far as it was justified. With respect to information value, the number of days since birth provided the highest explanatory power with IV at 0.08, but the information on loan size and the number of days since last change of cell phone was not lagging behind much either.

Thus, a logistic regression model was built on the training sample using grouped variables. Next, p values were used to check if different variables explained the default significantly. With 5% significance, all selected variables proved to be significant. Several procedures can be applied for evaluation, as shown above, in our case AUC and the GINI value construed from it were selected. We are going to use them later on as well, since they are suitable for the comparison of the models from the aspect of efficiency of estimation, which will answer the research question too.

Table 2
Findings of estimation using logistic regression

Procedure	GINI	AUC
Logistic regression	0.26	0.63

Source: own design

As you can see in *Figure 2*, 0.26 percent GINI could be achieved using logistic regression, which means 0.63 AUC. It cannot be said to be a strong estimation model, but it should be emphasised that the database mostly includes low value credit card loans with a low rate of default. In addition, the model only covers application information, so track history or behavioural data could not be analysed in the sample, which narrows down the options of estimation.

3.4 Models based on artificial intelligence

In the next part, the findings of models built with the help of the data processing capability of artificial intelligence are presented. Several models were built for analysing the problem, but only the ones will be detailed that proved to be more efficient than the traditional approach. Following a short introduction on methodology, the results relating to efficiency of estimation are introduced in detail.

3.4.1 Artificial neural network (ANN)

Artificial neural networks are based on biology. The nervous system is basically a complex system of connections. It is built from neurons arranged into a network. Signals are transmitted via synapses when the electric potential of a cell in the dendrite of a neuron reaches a threshold value. It is also important how different the strength of the connections between the synapses is (Pruves et al., 2019).

The operation of artificial neural networks is also based on weighting according to certain activation functions. During a simulation, a network is trying to learn the data pattern relying on the combinations of a weight matrix. In the process, all the above variables are input data, so the first so termed input level consists of sixty-four neurons. It turned out during modelling that a certain increase of the number of neurons in the second level (the first hidden level) was an important impacting factor. In the end, ten neurons were assigned there, since a higher number did not significantly improve the quality of learning. The second hidden level received five neurons, finally, one neuron represented the output level. Neuron levels form a fully connected linear combination in neural networks.

The activation functions used by neurons in the hidden levels for the learning mechanism also need to be presented. Firstly, the statistical process behind weighting needs to be detailed. Neurons are connected with directed links. Those links are provided with so termed associated $w_{j,i}$ weights. The weights will make up a weight matrix. Each neuron (unit i) will first perform weighting of the input connection as follows:

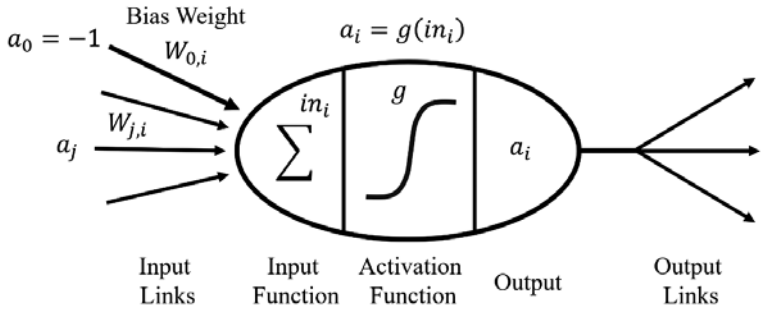
$$in_i = \sum_{j=0}^{n} W_{j,i} a_j \tag{2}$$

Value a_j generates the connection of the activation value from unit j to unit i. The activation function is responsible for weighting in the neuron. The model will set a_o as zero to an input of –1, and assigns a shift weight $w_{o,i}$ to it. So, the new activation is made according to the following formula:

$$a_i = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i} a_j\right) \tag{3}$$

where g is the activation function. The activation function should not be linear, because the model could easily be transformed into a simple linear function. *Figure 2* sums up the operation of a neuron:

Figure 2
How a neuron in an artificial neural network works



Source: Russal–Norvig (2005)

The levels were defined in different ways according to the activations. The model was weighted according to a sigmoid activation function in the last two output levels while the so termed rectified linear unit (Relu) was used in the first two levels for weighting. Relu provides 0 value for negative inputs and leaves positive ones uncut, while the sigmoid function can also be regarded as a kind of logistic function. The different outputs are defined using the functions and are set according to threshold points. The threshold point of a function is set by the actual point of the shift weights. Therefore, a unit will be activated if $\sum_{j=0}^{n} W_{j,i} a_j$ surpasses $w_{o,i}$-t (Russel–Norvig, 2003).

Table 3
Findings of the artificial neural network (ANN) procedure

Procedure	GINI	AUC
Artificial neural network	0.33	0.67

Source: own design

The estimation performance of the above model was assessed on the test sample. GINI produced 0.33 with AUC of 0.67. Thus, the artificial neural network model surpassed the results of the logistic regression model. Table 3 illustrates the results. It should, however, be noted that the interpretation of the so termed "black box" solutions or the explanation of the results is difficult, which might be a hindrance in areas such as healthcare or finance (Maheshwari, 2018).

3.4.2 XGBoost

XGBoost is short for *Extreme Gradient Boosting*, where the procedure *Gradient Boosting* was first described by Friedman (2001). The procedure is based on supervised learning. The starting point is a classical problem of modelling where construed explanatory variables x_i are available, which are used to define a target (y). So, one must identify a target function during the process with XGBoost serving optimisation. Those target functions consist of two components: estimation loss measured on the training data (*training loss*) and a regularising factor:

$$(\theta) = L(\theta) + \Omega(\theta) \tag{4}$$

where L indicates estimation loss and Ω is the regularising factor. In the case discussed, the estimation loss will show how accurate the prediction of a model is while the regularising factor helps to avoid overfitting and regulates the complexity of the model. The RMSE value is often used to write L, but the target function can be modified during the process. Using the methodology described, the performance of decision trees, random forest and boosted forest procedures are compared to formally approach the modification of the parameters.

In addition to supervised learning, XGBoost is based on the aggregation of decision trees (CART) that can be used to build a model via optimisation whether it is classification or regression. For instance, you can find out if somebody likes a computer game based on the sample of their family members considering age, gender, occupation, etc. In a CART model you can place family members into different levels where the levels have real scores, which is a strong tool for optimisation. Since a single tree cannot provide sufficient information, more than one trees should be assessed. To gain a better understanding of how trees are used, let us take a unit of two trees. The prediction scores of each tree are added to get the final score while the two trees supplement each other. This can be written as follows:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{5}$$

where K is the number of trees, f_k a function in the F function space, and F contains all possible outputs of the given CART. As detailed above, you must find the most accurate output on a combination set. To do so, the following optimising target function can be used:

$$(\theta) = \sum_{i}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \omega(f_k) \tag{6}$$

where $\omega(f_k)$ denotes the complexity of a tree while f_k is the function defined earlier. The generation of tree units is similar to random forest, as both are based on

using tree units. The difference comes from the way training is done on the training data. So, the prediction function can be applied in both cases.

The question may arise what parameters are used to differentiate the construction of the tree units. In other words, what strategy is applied when an algorithm, which contains the structure of the tree and the scores of the level, is used to assess the learning and assessment of f_i functions. Learning the structure of trees is more complex than a simple optimisation problem, where the gradients are considered only, so an additive strategy must be applied. We must check in practice what is it we have learnt so far and how it can be improved to generate a new tree. Thus, a prediction value, such as $\widehat{y_i^{(t)}}$ can be written down step by step. So, a modified prediction value will be written as follows:

$$\begin{aligned}
&\widehat{y_i^{(0)}} = 0 \\
&\widehat{y_i^{(1)}} = f_1(x_i) = \widehat{y_i^{(0)}} + f_1(x_i) \\
&\widehat{y_i^{(2)}} = f_1(x_i) + f_2(x_i) = \widehat{y_i^{(1)}} + f_2(x_i) \\
&\ldots \\
&\widehat{y_i^{(t)}} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y_i^{(t-1)}} + f_t(x_i)
\end{aligned} \tag{7}$$

If using an additive process, the selection of trees is important for each step. Use the tree that optimises your target (it is to be added):

$$\begin{aligned}
obj^t = \sum_{\{i=1\}}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{\{i=1\}}^{t} \omega(f_i) = \\
\sum_{\{i=1\}}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \omega(f_t) + constant\Big)
\end{aligned} \tag{8}$$

As detailed above, an identified target function must be selected you wish to optimise. It will be used to modify the function detailed above. Side by side with optimisation, a regularising factor must be identified that will represent the complexity of the tree. It can be written as follows:

$$f_{t(x)} = w_{q(x)}, w \in R^T, q: R^d \to \{1,2,\ldots,T\}, \tag{9}$$

where w is the vector of the score of the leaves, while q is a function assigning each data point to a suitable leaf. T is the number of leaves. Thus, complexity can be given as follows:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{10}$$

Reorganising the tree structure provides a score termed structure score in the literature. The structural value is given as follows:

$$w_j^* = -\frac{G_j}{H_j+\lambda} \tag{11}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j+\lambda} + \gamma T \tag{12}$$

In the equation w_j values are independent of each other, the best value will provide the best $q(x)$ structure, which will tell you how well a given tree performs. In effect, in the structure of a particular tree, g_i and h_i statistics are assigned to the appropriate leaves, then they are added, and the performance of the tree is calculated by means of the formula. The score will be the purity indicator of a decision tree, which also considers the complexity of the model.

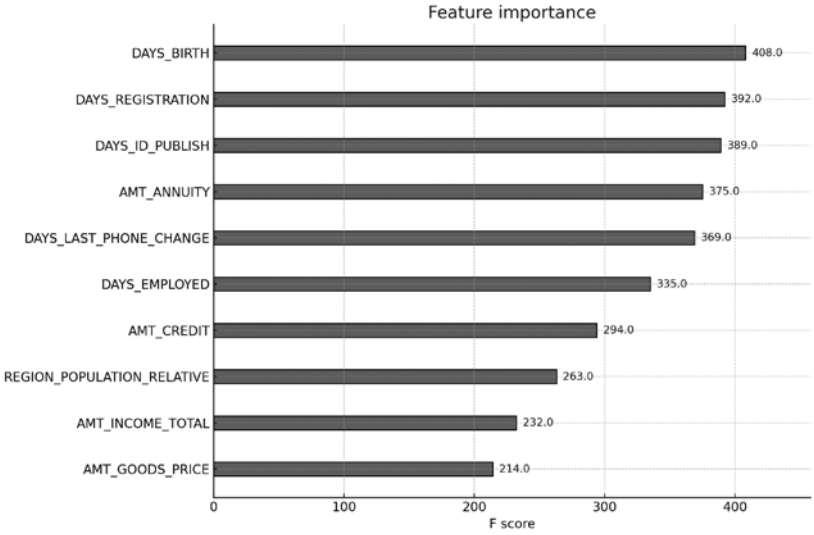
In an ideal world, if you can already measure the efficiency of a tree, you should list all possible trees and select the best of them. In practice, however, only one or another level of a tree is optimised, i.e., a branch is divided into two leaves to identify the pollution level of a decision tree:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G_L^2+G_R^2}{H_L+H_R+\lambda}\right] - \gamma \tag{13}$$

The interpretation of the formula is the following: the first component is the score of the new leaf on the left, the second is that of the new leaf on the right, the third is the score of the original leaf, finally, the regularisation of the leaf is defined. Please note, if the growth value is lower than γ, the branch in question should not be added to the model. For real value data, that is how an optimal division is approached, i.e., optimisation is performed with cutting from left to right (xgboost developers, 2022a). The procedure described can efficiently structure input information through optimisation, so not surprisingly it has become one of the most popular methods at data science competitions organised online (Reinstein, 2017).

Using XGBoost, estimation was made with the help of setting basic parameters. It can also be found on the website xgboost developers (2022a). During model building, the importance of the variables can also be defined (*feature importance*), which provides feedback on the importance of the information used via value *F*. *Figure 3* illustrates the most important variables in the XGBoost model included age, the number of days since change of registration, the length of time the account has been held with the bank and the size of the instalment. Less important variables included total income, or the value of the goods purchased with the credit card, but they also provided useful information in XGBoost since their value was quite high regarding *F*.

Figure 3
Values of importance of variables used in XGBoost model (*F* score)



Source: own design

Using the procedure, a total of 0.366 GINI values could be connected with 0.68 AUC. With respect to comparison, the model XGBoost provided the best result surpassing all earlier models. *Table 4* presents the results of the models studied.

Table 4
Comparison of results of different modelling procedures

Procedure	GINI	AUC
Logistic regression	0.26	0.63
Artificial neural network	0.33	0.67
XGBoost	0.37	0.68

Source: own design

Like in the case of artificial neural networks (ANN), you can say for XGBoost too that variable selection need not be performed separately during model building, XGBoost will apply importance estimation based on its informative ability during training the model, which allows the relevant variables to be identified (Goodarzi et al., 2009). Experience has shown an expert revision of the selection of variables has not improved the efficiency of the model (xgboost developers, 2022b).

An XGBoost model was also built using the variables selected for a logistic regression model for comparison. The GINI value received by logistic approach improved from 0.26 to 0.27. Obviously, those lag behind the efficiency of a model built using a full set of variables.

Both ANN and XGBoost optimise better for modelling than Information Value relying on WoE estimation or logistic regression based on correlation selection. Models relying on machine learning select information better, however, their interpretation is more difficult, which is a disadvantage the discussion of which goes beyond the scope of this paper.

4 SUMMARY

The authors have presented that modelling credit risk has been a continuously and intensively developing specific area. As information technology has been improving over the past decades, expert approach has been more and more frequently replaced by traditional statistical models based on masses of empirical data.

The current technological revolution driven by artificial intelligence has paved the way for a new generation of models in the field of credit risk summarily termed models based on machine learning.

In the study, a credit card database was used to compare application (PiT) PD modelling with the traditional models, which are dominant in banking practice, and machine learning-based models. In an empirical study of logistic regression, 7 variables were selected following the selection of information values and variables for correlation. The variables were grouped in line with the WoE method, finally, a logistic regression was run on the data construed from the values.

Of methods based on artificial intelligence, the model using an artificial neural network (ANN) provided significantly better GINI coefficient values compared to the results achieved using traditional logistic regression. The algorithm XGBoost provided the model considered to be the most efficient.

As described, you need not perform preliminary selection of the variables for XGBoost and ANN, since during its training the model will automatically overrepresent the important information. It might be the reason why the efficiency of machine learning methods is stronger than the traditional approach of logistic regression if further useful information is also used.

The prediction capabilities of machine learning methods are excellent. However, the interpretation of decision models is more difficult – due to their black box nature – compared to traditional approaches regarded to be industrial best practice. To manage the issue, many software packages keep appearing to make

the interpretation of decision mechanisms simpler in the case of more complex models. SHAP is such a package, which can simplify the decision mechanism of XGBoost with the help of interaction values (Lundberg, 2018). The authors believe as algorithms and model interpretation software packages improve, and as more and more information becomes available, machine learning methods will take the lead in the field of credit risk modelling.

Finally, the authors wish to express a policy recommendation addressed to the regulatory and supervisory authorities. Because of the black box nature of machine learning based on artificial intelligence, banks currently cannot apply them. The difficulties of interpretation of the results and of the decision-making mechanism cause problems in several points of the regulatory and supervisory set of requirements. Our recommendation is to review the regulations and recommendations tailored to fit the traditional models so as to allow room for the banks to apply machine learning models. The increase in efficiency presented in this paper can contribute to increasingly prudent banking operations, which is a core interest for all parties involved.

REFERENCES

Addo, P. M. – Guegan, D. – Hassani, B. (2018): Credit risk analysis using machine and deep learning models. *Risks,* 6(2), 38, https://doi.org/10.3390/risks6020038.

Angelini, E. – Di Tollo, G. – Roli, A. (2008): A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance,* 48(4), 733–755, https://doi.org/10.1016/j.qref.2007.04.001.

Altman, E. I. – Saunders, A. (1997): Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, 21(11–12), 1721–1742, https://doi.org/10.1016/s0378-4266(97)00036-8.

Basel Committee on Banking Supervision (2023): Calculation of RWA for credit risk – CRE36: IRB approach: minimum requirements to use IRB approach.

Butaru, F. – Chen, Q. – Clark, B. – Das, S. – Lo, A. W. – Siddique, A. (2016): Risk and risk management in the credit card industry. *Journal of Banking & Finance,* 72, 218–239, https://doi.org/10.1016/j.jbankfin.2016.07.015.

European Banking Authority (2017): Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures.

European Banking Authority (2020): On management and supervision of ESG risks for credit institutions and investment firms – EBA/REP/2021/18.

Friedman, J. H. (2001): Greedy function approximation: a gradient boosting machine. *Annals of statistics,* 1189–1232, https://doi.org/10.1214/aos/1013203451.

Goodarzi, M. – Deshpande, S. – Murugesan, V. – Katti, S. B. – Prabhakar, Y. S. (2009): Is feature selection essential for ANN modeling? *QSAR & Combinatorial Science,* 28(11–12), 1487–1499, https://doi.org/10.1002/qsar.200960074.

Gurný, P. – Gurný, M. (2013): Comparison of credit scoring models on probability of default estimation for us banks. *Prague Economic Papers,* 22(2), 163–181, https://doi.org/10.18267/j.pep.446.

Henisz, W. J. – McGlinch, J. (2019): ESG, Material Credit Events, and Credit Risk. *Journal of Applied Corporate Finance*, 31(2), 105–117, https://doi.org/10.1111/jacf.12352.

Huang, C. L. – Chen, M. C. – Wang, C. J. (2007): Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4), 847–856, https://doi.org/10.1016/j.eswa.2006.07.007.

Hungarian Central Bank (MNB) (2002): Tanulmányok a bankszektor középtávú fejlődési irányairól [Studies on mid-term development directions of the banking sector]. *MNB Műhelytanulmányok*, 26.

Hungarian Central Bank (MNB) (2018): ICAAP-ILAAP-BMA kézikönyv [ICAAP-ILAAP-BMA Manual]. Budapest: Magyar Nemzeti Bank. https://www.mnb.hu/letoltes/icaap-ilaap-bma-kezikonyv-2018-januar.pdf.

Kaggle (n. a.): Loan Defaulter Dataset. https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter?select=application_data.csv (accessed at 01.04.2023).

Kovács, L. – Marsi, E. (szerk.) (2018): *Bankmenedzsment – banküzemtan* [Bank management, bank operations]. Budapest: Magyar Bankszövetség. ISBN: 978-963-89653-2-5.

Lundberg, S. (2018): Basic SHAP Interaction Value Example in XGBoost. https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Basic%20SHAP%20Interaction%20Value%20Example%20in%20XGBoost.html (accessed at 11.06.2023).

Madar, L. (2015): Scoring rendszerek hatásai a gazdasági tőkeszámítás során alkalmazott portfoliómodellek eredményeire [Impact of scoring systems on results of portfolio models used for economic capital calculations]. PhD Dissertation, Kaposvár University.

Maheshwari, S. (2018): The Explainable Neural Network. Elérési link: https://medium.com/@shagunm1210/the-explainable-neural-network-8f95256dcddb (accessed at 25.05.2023.).

Medema, L. – Koning, R. H. – Lensink, R. (2009): A practical approach to validating a PD model. *Journal of Banking & Finance,* 33(4), 701–708, https://doi.org/10.1016/j.jbankfin.2008.11.007.

Moradi, S. – Mokhatab Rafiei, F. (2019): A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financial Innovation,* 5(1), 1-27, https://doi.org/10.1186/s40854-019-0121-9.

Nagy G. – Biró G. (2018): PIT and TTC problems related to IRB PD parameter estimation in the light of supervisory reviews. *Economy and Finance,* 5(3), 250–278. https://bankszovetseg.hu/Public/gep/2018/250-278%20Nagy%20Gabor-Biro%20Gergelyuj.pdf.

Pedregosa, F. – Varoquaux, G. – Gramfort, A. – Michel, V. – Thirion, B. – Grisel, O. (2011): Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12 (Oct), 2825–2830.

Peng, C. Y. J. – Lee, K. L. – Ingersoll, G. M. (2002): An introduction to logistic regression analysis and reporting. *The Journal of Educational Research,* 96(1), 3–14, https://doi.org/10.1080/00220670209598786.

Purves, D. – Augustine, G. J. – Fitzpatrick, D. – Hall, W. – LaMantia, A. S. – White, L. (2019): *Neurosciences. De Boeck Supérieur.*

Reinstein, I. (2017): XGBoost: A Top Machine Learning Method Explained in Kaggle Competitions. KDnuggets. https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html (accessed at2023.05.21.).

Rossum, G. – Drake, F. L. (2009): *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Rozo, B. J. G. – Crook, J. – Andreeva, G. (2023): The role of web browsing in credit risk prediction. *Decision Support Systems*, 164, 113879, https://doi.org/10.1016/j.dss.2022.113879.

Russell, S. – Norvig, P. (2005): *Artifical intelligence. A modern approach* (hungarian translation). Budapest: Panem Könykiadó, http://project.mit.bme.hu/ (accessed at 25.05.2023.).

Shichen, X. (2023): Credit Risk Scorecard- scorecardpy 0.1.9.6 (Python könyvtár). https://github.com/ShichenXie/scorecardpy.

Siddiqi, N. (2006): Credit Risk Scorecards. NJ, Hoboken: John Wiley & Sons.

Sirignano, J. – Sadhwani, A. – Giesecke, K. (2016): Deep Learning for Mortgage Risk. https://arxiv.org/abs/1607.02470.

Somerville, R. A. – Taffler, R. J. (1995): Banker judgement versus formal forecasting models: The case of country risk assessment. *Journal of Banking & Finance*, *19*(2), 281–297, https://doi.org/10.1016/0378-4266(94)00051-4.

Venkatesh, A. – Jacob, S. G. (2016): Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. *International Journal of Computer Applications* (0975–8887), 145(7), https://doi.org/10.5120/ijca2016910702.

xgboost developers (2022a): Introduction to boosted trees. https://xgboost.readthedocs.io/en/stable/tutorials/model.html (accessed at 20.05.2023.).

xgboost developers (2022b): Understand your dataset with XGBoost. https://xgboost.readthedocs.io/en/stable/R-package/discoverYourData.html (accessed at 10.06.2023.).

Wang, K. – Li, M. – Cheng, J. – Zhou, X. –Li, G. (2022): Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science,* 199, 1128–1135, https://doi.org/10.1016/j.procs.2022.01.143.

West, D. (2000): Neural network credit scoring models. *Computers & Operations Research*, 27(11–12), 1131–1152, https://doi.org/10.1016/s0305-0548(99)00149-5.

Yu, Y. (2020): The Application of Machine Learning Algorithms in Credit Card Default Prediction. In 2020 International Conference on Computing and Data Science (CDS), 212–218. *IEEE,* https://doi.org/10.1109/cds49703.2020.00050.