FACTA UNIVERSITATIS
Series: **Economics and Organization** Vol. 18, N° 1, 2021, pp. 29 - 43
https://doi.org/10.22190/FUEO201028001R

Original Scientific Paper

CREDIT SCORING WITH AN ENSEMBLE DEEP LEARNING CLASSIFICATION METHODS – COMPARISON WITH TRADITIONAL METHODS

UDC 336.77:519.2

Ognjen Radović, Srđan Marinković, Jelena Radojičić

University of Niš, Faculty of Economics, Serbia

Abstract. *Credit scoring attracts special attention of financial institutions. In recent years, deep learning methods have been particularly interesting. In this paper, we compare the performance of ensemble deep learning methods based on decision trees with the best traditional method, logistic regression, and the machine learning method benchmark, support vector machines. Each method tests several different algorithms. We use different performance indicators. The research focuses on standard datasets relevant for this type of classification, the Australian and German datasets. The best method, according to the MCC indicator, proves to be the ensemble method with boosted decision trees. Also, on average, ensemble methods prove to be more successful than SVM.*

Key words: *credit scoring; classifier ensemble, deep learning, support vector machine*

JEL Classification: C38, C45, C55, G17, G24

1. INTRODUCTION

Credit scoring is a quantitative method for assessing the credit risk involved in granting a loan to a borrower. Credit scoring is one of the key stages in credit analysis. It is a method of assessing the creditworthiness of a client applying for a loan. The goal of creditworthiness assessment is to classify credit applications into acceptable and unacceptable, but also to provide the necessary inputs for the next phases of credit analysis, such as determining the credit volume, interest rates, collateral, restrictive clauses and the like. Traditional credit analysis relies on historical and verifiable information or accounting data. The most general credit analysis framework many traditional lenders use to assess credit users is the 5C approach. Credit analysts make decisions based on the following criteria: debtors' character, capacity (ability to repay), capital, collateral and market conditions. The main advantage of this method is that it can be

Received October 28, 2020 / Accepted December 02, 2020
Corresponding author: Jelena Radojičić
University of Niš, Faculty of Economics, Trg kralja Aleksandra 11, 18000 Niš, Serbia
E-mail: jelena.radojicic@eknfak.ni.ac.rs

used to make credit decisions in various types of business and consumer loans without significant adjustments. Traditionally, creditworthiness assessment of a corporate borrower is based on financial indicators that indicate possible problems in loan repayment and play an important role in calculating credit risk levels. These indicators are characterized by objectivity because they are calculated on the basis of data in borrowers' financial statements. The analyst's experience and impressions can play a significant role in creditworthiness assessment of the entity applying for a loan. Prior to the credit scoring model, the decision to grant a loan was made based on the credit analyst's assessment. One of the disadvantages of such an approach is the inability to process a large number of applications per day, which has given rise to various credit scoring models to quantify the credit risk (Dastile, Celik, & Potsane, 2020). Banks used information received within a loan application (e.g., number of dependents, period in the current job, etc.) to calculate the borrowers' numerical score (Lewis, 1992). Credit scoring, as a precise and automatic creditworthiness assessment tool, is a particularly important factor in the expansion of consumer lending (Thomas, Crook, & Edelman, 2002). Credit scoring is mostly applied in consumer loans, credit cards and mortgage loans (Einav, Jenkins, & Levin, 2013). Advances in information technology facilitate further development of credit scoring models for making objective and quick decisions (Thomas, Crook, & Edelman, 2002). Credit scoring is a technique that financial organizations use when making a decision to approve a loan or reject their clients' loan application (application scoring). Credit scoring models can be applied to analyze the behavior of already existing clients, and then the score represents a numerical summary of the bank's experience with the client (behavior scoring) (Hui, Li, & Zongfang, 2017). The Basel II agreement of 2004 expanded the field for developing more sophisticated credit scoring models, thus allowing banks to assess the probability of default on their own under the internal ratings-based (IRB) approach (Goh & Lee, 2019).

2. Traditional Statistical Methods Used in Credit Scoring

Credit scoring is a multi-stage process (Bequé & Lessmann, 2017), and it basically compares the borrower's characteristics and the characteristics of other clients from the previous period. The statistical model relies on historical data. Its goal is to predict future behavior in loan repayment based on previous experience with loan users with similar characteristics. If the borrower is similar to "bad" clients (who did not repay the approved loan properly), the application is rejected, and if the borrower is similar to "good" clients (who returned the approved loan properly), the loan application is approved. The loan applicant's score is compared with the established cut-off score. If the obtained score is higher than the cut-off score, credit is approved, and if the score is lower than the cut-off, application is rejected. The cut-off score is crucial for the usefulness of the credit scoring model and mainly relies on the credit decision makers' attitudes towards risk. So, there is no optimal cut-off value. "It varies from one environment to another and from one bank to another inside the same country" (Abdou & Pointon, 2011). The final step is to measure the accuracy of the credit scoring model and monitor business performance indicators. "The choice of a statistical model is crucial because it affects all subsequent activities and credit scoring performance" (Bequé & Lessmann, 2017).

The models that financial institutions use help them decide whether or not to grant a loan. As the final decision on approval is binary, there is a problem of binary classification. Credit scoring involves "formal statistical methods to classify loan applicants into "good" and "bad" risk classes" (Hand & Henley, 1997). Credit scoring algorithms are basically statistical in nature: they use empirical evidence to formulate predictions about the future. Prediction models assess a continuous variable while classification models predict class membership. In credit scoring, the dependent variable is actually binary, so most algorithms can be considered classification algorithms (Abdou & Pointon, 2011).

Statistical methods such as linear regression and discriminant analysis require the assumption of a linear relationship between variables. In credit scoring, linear regression is used as a binary classification problem. Discriminant analysis is a simple parametric statistical technique for classifying loans into good and bad. Fischer (1936) suggests the application of discriminant analysis as a classification technique, and in 1941 Durand used discriminant analysis to classify "good" and "bad" car loans, thus beginning the trend of applying statistical models in credit scoring. Later, Altman (1968) developed a Z-score model with financial indicators, using variables from corporate financial statements as input variables for a discriminant analysis model to predict company bankruptcies. "Discriminant analysis is one of the frequently used techniques in credit scoring" (Abdou & Pointon, 2011). Ogrler (1971) applies regression analysis for credit scoring of consumer loans in banks, after using this method somewhat earlier to evaluate the already existing commercial loans (Orgler, 1971). He concludes that "information not included in the loan application form has a greater predictive ability" to assess loan quality in the future than the information included in the form.

In 1980, Ohlson proposed the use of logistic regression (LOGIT) as a creditworthiness assessment method in companies (Ohlson, 1980). The outcome variable in logistic regression is dichotomous (outcome 0/1), so this method is suitable for modeling binary outcomes and is widely used in creditworthiness assessment due to its simplicity and transparency (Dastile, Celik, & Potsane, 2020; Abdou & Pointon, 2011). Within traditional methods, logistic regression has become the standard credit scoring model due to its compliance with the Basel II standard (Goh & Lee, 2019).

3. Application of Machine Learning in Credit Scoring

More sophisticated methods of credit scoring that literature has offered in recent years are machine learning algorithms and data mining methods. The point of using sophisticated techniques is their ability to model extremely complex functions (Abdou & Pointon, 2011). Machine learning models learn from available data, thus allowing the calculation of predictive value. A machine learning algorithm learns a set of rules based on the information available in a training set of examples. Machine learning models have the potential to replace the logistic regression model in credit scoring because they show great prediction accuracy. However, the impossibility of certain models to explain the predictions, i.e. lack of transparency, limits their application in regulated financial institutions. The application of machine learning algorithms may include (Dastile, Celik, & Potsane, 2020): „k-*Nearest Neighbor* (k-NN), *Decision Trees* (DTs), *Support Vector Machines* (SVMs), *Artificial Neural Networks* (ANNs), *Random Forests* (RFs), *Boosting*, *Extreme Gradient Boost* (XGBoost), *Bagging*, *Restricted Boltzmann Machines* (RBMs), *Deep Multi-Layer Perceptron*, (DMLP), *Convolutional Neural Networks* (CNNs) and *Deep Belief Neural Networks* (DBNs)".

The Decision Tree creates a discriminant function in the form of a tree, from the root to the leaves. Each node represents a logical test of the attribute value, and the leaves denote classes. Input observations are recursively split into subbranches, i.e. until the final result (class designation). Mathematical formulas such as the Gini index (CART) or entropy (in ID3, C4.5, C5, J4.8 decision tree algorithms) are used to determine the splitting threshold (Patil, Aghav, & Sareen, 2016; Bequé & Lessmann, 2017). The tree learns by asking questions that will solve the problem in the fastest and most accurate way. Each time the algorithm is repeated, the attribute value is compared to the threshold. Thresholds determine which attribute should be tested and when tree growth should be stopped (Bequé & Lessmann, 2017). After the training, the tree can predict the outcome if applied to data of the same type and format. The low decision tree accuracy may be affected by low depth and presence of noise (overfitting). This classification technique is widely used in credit scoring models (Dastile, Celik, & Potsane, 2020).

Since its introduction into the theory of statistical learning in 1998 (Vapnik, 1998), the support vector machines method (Support Vector Machines – SVM) has been used as a binary classifier of machine learning. SVM algorithms in the field of credit scoring were introduced by Baesens et al. (2003) and through comparison with other classification algorithms (logistic regression, discriminant analysis, k-nearest neighbor, neural networks, decision tree) indicated its good performance. SVM, as a discriminant model, is based on the margin for classification between classes, i.e. it focuses on finding the boundary that separates two classes with the smallest error. In the SVM model, data is viewed as vectors in n-dimensional space. The focus is on the maximum margin between classes, i.e. the space between two hyperplanes that separate data from different classes, i.e. which belong to one or more support vectors. The separating hyperplane is located farthest from the data and is determined by the position of the data of both classes closest to it. If the classes are denoted by y=+1 and y=-1, in the version of the linear SVM classifier, the margin $\|\alpha\| = \sqrt{\sum_{i=1}^{m} \alpha_i^2}$ between the negative and positive hyperplanes is maximized. The following equation is used to assign a class (Dastile, Celik, & Potsane, 2020):

$$y = \begin{cases} +1, if\ b + \alpha^T x \geq +1 \\ -1, if\ b + \alpha^T x \leq -1 \end{cases} \tag{1}$$

where b is bias.

The SVM method can also be used for data classes that are not linearly separable. “For nonlinear classification, a kernel trick is used to modify the SVM formulation” (Dastile, Celik, & Potsane, 2020). Using the appropriate kernel function, the example is mapped to a space with a larger number of dimensions and the nonlinear problem is transformed into a linear one.

Neural network consists of several neurons that work in parallel, without centralized control. An artificial neural network mimics the way a biological neural network processes information. Neurons are usually complex in layers. The neural network usually consists of three layers: input, hidden, and output parameters. There are several types of neural networks, but their common components are a set of nodes and connections between nodes. Nodes represent computer units whose task is to receive inputs, process inputs and produce and output (Bequé & Lessmann, 2017). First, the input characteristics are processed to the hidden parameters, and then the hidden parameters calculate the adequate weight before forwarding the information to the output parameters. Each layer of the neural network “consists of several elements, i.e. neurons. The number of input neurons depends on the number of predictors, the number of hidden neurons is a setting parameter determined by the analyst, and the number of output neurons is determined by the modeling task itself, e.g. for binary classification it is one" (Bequé & Lessmann, 2017). Artificial neural networks were first applied by Odom and Sharda (1990) in credit scoring.

For a given vector of the input attribute x, the three-layer neural network calculates the output value$\hat{y}$ as follows (Dastile, Celik, & Potsane, 2020):

$$\hat{y} = a_2\big(a_1\left(\alpha^{(1)}x + \alpha_0^{(1)}\right)\alpha^{(2)}x + \alpha_0^{(2)}\big) \tag{2}$$

where $(\alpha_0^{(1)}, \alpha^{(1)})$, $(\alpha_0^{(2)}, \alpha^{(2)})$ are weights, and a_2 and a_2 are activation functions between the input and hidden layer.

Neural networks are trained through a training set, and the final decision is made by applying the decision function to $\hat{y}$.

4. Ensemble of Classifiers – Literature Review

Elementary classifiers represent a unique set of statistical relationships. Ensemble of classifiers consist of a set of individually trained base classifiers whose decisions are combined in a certain way (weighted voting or unweighted voting) when new examples are classified. Ensemble of classifiers is a "combination of classifiers so that their fusion achieves better performance than stand-alone classifiers" (Nanni & Lumini, 2009). Combining different machine learning algorithms can improve the accuracy of results (Dastile, Celik, & Potsane, 2020). Ensemble algorithm techniques are used to aggregate the results of "unstable" algorithms in which small changes in the training set lead to large changes in the learned set of rules (e.g. neural networks, decision trees). The application of ensemble learning techniques requires the simultaneous fulfillment of the following assumptions (Pławiak, Abdar, & Acharya, 2019): "a) quality, b) statistical independence (diversity) and c) efficiency (speed)".

Ensemble of classifiers is used to achieve better performance in various research areas such as computer intelligence, statistics, and machine learning (Ren, Zhang, & Suganthan, 2016). Different ensembles of classifiers are used in literature (Pławiak, Abdar, & Acharya, 2019): a) Boosting (AdaBoost), b) Bagging (Bootstrap aggregation), c) Random Forest, d) Stacking (Stacked Generalization) and e) Mixtures of Experts.

Boosting is the most commonly used method in the ensemble of classifiers (Dastile, Celik, & Potsane, 2020). It starts with a weak model (for example, a shallow decision tree), and then the models are iteratively evaluated and amplified (Freund & Schapire, 1997). Boosting produces a series of classifiers (Bequé & Lessmann, 2017). Each subsequent classifier focuses on examples that have been misclassified by the previous classifier. Each example in the training set is assigned a weight in accordance with the significance of the example in the set. Examples misclassified by the previous model are assigned a higher weight. After individual classifier learning, the weights are updated on the test set. The accuracy of individual classifiers on a test set is determined by the weight of that classifier in the classification of new examples by applying an ensemble of classifiers.

One of the most well-known boosting techniques is Adaptive Boosting (AdaBoost) (Freund & Schapire, 1997). AdaBoost algorithm has one parameter T – the number of generated classifiers (iteration) and is characterized by simplicity and efficiency. AdaBoost assigns a class to an input attribute vector that is classified (x) as follows:

$$\hat{y} = sign\left(\sum_{t=1}^{T} \alpha_t \phi_t\,(x)\right) \tag{3}$$

where α_t is the weight of the classifier $\phi_t(x)$.

Bagging generates multiple versions of classifiers that are used as an aggregate predictor through a voting mechanism (Breiman, 1996). Bootstrap Aggregation is used to generate classifiers, with no iterative division into a training set and a test set, but random selection with return. The training set is formed by successive sampling (with repetition) of data from the initial set. Data never selected form a test set, while the rest is used for training. The process is repeated several times, and the overall score is obtained as the average score on all thus formed sets for verification (Breiman, 1996). Bootstrapping generates K training sets, and then one basic classifier is trained on each of them. The class rating is awarded by a majority vote of Kclassifiers, as follows (Dastile, Celik, & Potsane, 2020):

$$y = \underset{y\in\{+1,-1\}}{argmax} \sum_{i=1}^{K} 1(y = \phi_i(x)) \tag{4}$$

where

$$1(y = \phi_i(x)) = \begin{cases} 1, \ if\ y = \phi_i(x); \\ 0, if\ y \neq \phi_i(x). \end{cases} \tag{5}$$

The Random Forest algorithm consists of several decision trees. The new examples are classified by the voting method based on the decisions of individual trees. Not all samples and attributes are taken for training individual trees, but a certain number of randomly selected attributes and samples from the training dataset. Each decision tree develops on a subset of randomly selected attributes. The best attribute is chosen for the decision tree node. Selecting the right attributes (questions) to be tested in a particular node reduces the entropy, or provides additional information about the sample. In a random forest algorithm, multiple decision trees learn from randomly selected data leading to greater tree diversity and depth. The greater depth of the trees makes the random forest algorithm more resistant to underfitting (insufficiently good interpretation of the relationships between variables within the dataset) and overfitting (noise in the data) compared to individual decision trees.

Deep Learning, as one of the machine learning fields, is based on a hierarchical architecture that includes multiple layers of nonlinear operations and steps in information processing. Some of the deep machine learning techniques used in credit scoring are (Pławiak, Abdar, & Acharya, 2019): „(a) deep discriminant models such as: Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), as well as Convolutional Neural Networks (CNNs), (b) unsupervised learning (generative models) such as: Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), Deep Boltzmann Machines (DBMs) as well as Regularized Autoencoders." Deep learning classifiers are not widely used in credit scoring (Dastile, Celik, & Potsane, 2020). Disadvantages of deep learning are (Pławiak, Abdar, & Acharya, 2019): "(1) computationally complex training, (2) long and inefficient training and (3) the overfitting effect, which prevents its effective practical use".

Predictive accuracy is a basic measure of classification success. Prediction accuracy is the percentage of success in classifying new examples using learned rules. Even small errors in creditworthiness assessment can lead to large losses, so increasing the accuracy of credit scoring is of great importance for the profitability of banks (Pławiak, Abdar, & Acharya, 2019). In this regard, more sophisticated credit scoring models have significant potential (Dastile, Celik, & Potsane, 2020). In one of the earliest reviews of statistical methods and data mining methods applied in credit scoring, Hand and Henley (1997) conclude that further development should go towards more complex DM models, which later literature reviews confirmed (Sadatrasoul, et. al. 2013). A comparison of different approaches to credit scoring shows that advanced machine learning-based techniques may have better predictive ability than conventional techniques such as logistic regression and discriminant analysis (Abdou & Pointon, 2011). Nanni and Lumini (2009) investigate several ensemble of classifier systems used in credit scoring and bankruptcy prediction and improve the performance obtained using the stand-alone classifiers. Wang et al. (2011) carry out a comparative performance evaluation of “three ensemble methods (Bagging, Boosting, Stacking) of basic classifiers: logistic regression, decision tree, artificial neural network, and support vector machine (SVM)”. The results show that ensemble method improves performance and that bagging gives better results than boosting. Lou et al. (2017) compare classification success of deep learning algorithms in credit scoring and widely used models such as logistic regression and SVM, to find that deep learning models have better performance. Li et al. (2017) develop a “model for credit risk assessment using deep neural networks” and show that the proposed algorithm has greater accuracy in credit risk assessment. The results of the simulation by Zhou et al. (2012) show that Extreme learning machines (ELM) is a more suitable approach for credit risk assessment than SVM. Bequé and Lessmann (2017) investigate the potential application of ELM for credit scoring and compare it with other classifiers (neural networks, k-nearest neighbor, SVM, classification and regression decision tree, logistic regression) within three dimensions: “ease of use, computational complexity and predictive performance”. They conclude that “ELM shows competitive or better results in each dimension of comparison and especially proves high discriminant power, both in isolation and within the ensemble and, therefore, represents a competitive alternative to already established classifiers in the field of credit scoring”. Neagoe et al. (2018) design a credit scoring model using a neural network classifier, namely: The Multilayer Perceptron (MLP) approach and the thirteen-layer DCNN variant. The obtained results confirm the efficiency of the proposed approach, indicating a significant advantage of DCNN over MLP. Proceeding from the idea to imitate the work of the human brain in terms of fusion and information flow, Plawiak et al. (2019) create the “Deep genetic cascade ensembles of classifiers (DGCEC) based on the fusion of stratified 10-fold CV method, ensemble learning, deep learning, layered learning and supervised training. The applied model combines three machine learning techniques: evolutionary, ensemble and deep learning.” The solution the authors propose provides a fast and efficient approach to training, which increases the accuracy of creditworthiness assessment. In the DGCEC method, each first-layer classifier is trained to increase the recognition performance of accepted or rejected borrowers based on the pre-processed data on borrowers. In other layers, based on the pre-processed user data and the classifier response from the first and previous layers using deep learning techniques and selection of genetic characteristics, a knowledge extraction process takes place that leads to the final result. The results show better performance of this approach compared to previously applied approaches in terms of the accuracy of creditworthiness assessment of borrowers in Australia. The highest accuracy of creditworthiness assessment in previously conducted studies is 91.97%, while the method proposed by the authors allows higher prediction accuracy, i.e. of 97.39%. (Pławiak, Abdar, & Acharya, 2019). Recent literature research shows that ensemble models have better performance than individual classifiers and that deep learning models give better results compared to statistical and traditional machine learning models (Dastile, Celik, & Potsane, 2020).

5. Research Methodology

Empirical research includes checking the performance of deep learning algorithms over known credit scoring datasets. In this paper, we use two datasets, the Australian Credit and the German Credit (UCI Machine Learning repository, Asuncion & Newman, 2010).

Datasets include a different number of independent variables. Australian credit data consists of 307 cases of creditworthy candidates and 383 cases of candidates to whom credit should not be granted. The German dataset is somewhat more asymmetric, with many more creditworthy examples (700) than those that should not be granted credit (300). The Australian dataset has 14 attributes, while the German has 24 attributes. Both sets have two classes {approved, rejected} and are a good mix of different types of attributes: continuous and nominal. Variables can be grouped into several categories (Beque and Lessmann, 2017): financial (assets, monthly income, etc.), socio-demographic (age, place of residence, etc.), others (possession of a credit card or a mobile phone).

K-fold cross-validation is used to assess the classification model, which works as follows (Dietterich, 1998):

1. Dividing a training dataset into k randomly selected non-overlapping data subsets of approximately equal size;
2. One subset is used to validate the model of trained over the remaining data subsets;
3. This procedure is repeated k-times so that each subset is used exactly once for model validation;
4. Performance is assessed for each partition and the average error on all k-partitions is reported.

This is one of the most popular techniques for cross-validation and is good at assessing the predictive accuracy of the classification model. In our study, 5-fold cross-validation is used for both credit rating datasets. Also, in order to reduce the dimensionality of the predictor space, principal component analysis (PCA) is used. PCA linearly transforms predictors to remove redundant dimensions and prevent overfitting.

To assess the performance of classification models, several standard indicators are used, for the calculation of which the values of correct and incorrect predictions are used: the number of borrowers correctly classified as {approved} (defaults) (True Positives -TP), the number of borrowers incorrectly classified as {approved (defaults) (False Positives - FP), the number of borrowers correctly classified as {rejected} (non-defaults) (True Negatives - TN), and the number of borrowers incorrectly classified as {rejected} (non-defaults) (False Negatives - FN). The total number of examples is N=TP+FP+FN+TN. The false positive rate (FP) is defined as the share of misclassified loan approval cases. In contrast, the rate of false negative results (FN) is defined as the share of misclassified cases refused to be given a loan (qualifying for a loan).

The set of indicators consists of: PCC (Percentage Correctly Classified), AUC (area under the curve), sensitivity, specificity, precision, G-mean, F-measure and Matthews correlation coefficient - MCC. The selection of indicators is based on previous research (Oztekin, Al-Ebbini, Sevkli, & Delen, 2018; Kim et al, 2020).

Percentage Correctly Classified (PCC) is the ratio of correct predictions of a case classification model in two categories {approved, rejected}. It is calculated as PCC = (TP+TN)/(TP+TN+FP+FN). PCC is the average percentage of correctly classified cases and is a measure of correct classification over sets unused for learning (subset for validation in 5-fold cross-validation). Sensitivity/Recall is the ratio of correctly classified cases in the class {approved} to the total number of examples in the class {approved} and is calculated as SEN = TP/(TP+FN). Specificity is the ratio of correctly classified cases in the class {rejected} to the total number of examples in the class {rejected} and is obtained as TN/(TN+FP). Sensitivity and specificity show the accuracy of class-level classifiers.

Geometric-Mean (G-mean) is obtained as follows:

$$G\text{–}mean = \sqrt{\frac{TN}{(FP+TN)} x \frac{TP}{(TP+FN)}} \tag{6}$$

F-measure is calculated as:

$$F\text{–}measure = \frac{2\times Sensitivity \times Precision}{Precision + Sensitivity} \tag{7}$$

where Precision = TP/(TP+FP). G-mean and F-measure indicate imbalance between classes.

AUC or Area Under the Receiver Operating Characteristic Curve is one of the indicators that illustrates the performance of the classification model. The larger this area, the better the model. AUC can be seen as the ability to distinguish positive from negative classification.

Matthews correlation coefficient (MCC) is another performance indicator. The MCC value is -1 to 1. Perfect prediction has a value of 1, completely incorrect prediction is -1, while random prediction has a value of 0. The MCC generates a high score only if the model can correctly predict most positive credits and most correctly rejected credits. It is considered one of the best indicators of accuracy in the evaluation of machine learning algorithms. The formula for calculating the MCC is as follows (Matthews, 1975; Jurman et al, 2012):

$$MCC = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \tag{8}$$

The following classifiers are used to compare the performance of ensemble models: Logistic regression and Support vectors. The SVM models used to compare classification performance use different learning algorithms (Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, Coarse Gaussian). The tested ensemble models use three learning algorithms (Boosted Trees (AdaBoost), Bagged Trees and RUSBoosted Trees) over decision trees (Random Forest techniques). The research is conducted in the Classification Learner application module using the Matlab 2019b software package. All calculations are performed in a Windows 10 environment (AMD Ryzen 7 3700U with 12GB RAM).

6. Results and Discussion

Tables 1 and 2 show the performance of predictive models for the Australian and German datasets, respectively. The tables show assessment of categorical prediction accuracy and the discriminant ability of the included classifiers. Performance assessment is average accuracy on test sets. For each model, AUC (Area Under the Receiver Operating Characteristic Curve), True Positive Classification Rate (TP), False Positive Classification Rate (FP), True Negative Classification Rate (TN), False Negative Classification Rate (FN), and PCC are shown (Percent Correctly Classified). These results are the average values determined for each of the 5 independent non-overlapping partitions of the dataset used in the 5-fold cross-validation.

For both datasets, the approved case rate is higher than the rejection rate for all models tested, meaning that false negative results are less common than false positive results relative to the overall prediction (with the exception of the Fine Gaussian SVM). It is debatable whether false negative predictions are more serious than false positive ones from the point of view of credit risk. If someone is predicted to be able to repay the loan and it turns out that they are not able to, that can lead to certain losses. Conversely, if someone is rejected, and they are able to repay the loan, it leads to a loss of earnings on the loan.

At the Australian dataset, according to the AUC criterion, the best classifier is the Ensemble Bagged Trees model (Table 1) with a value of 0.92 (87.41%). The Gaussian SVM model has the lowest false positive case rate (FP), while the linear Gaussian SVM models have the lowest false negative case rate (FN). According to the criterion of false positive cases (FP), ensemble models show the best results. This means that these models prove to be less risky in terms of incorrectly granted credit. According to the PCC criteria, the Ensemble Bagged Trees model proves to be the best classifier (Table 1). However, some SVM methods (Gaussian SVM) show better results than other ensemble methods. However, in general, ensemble methods show better results than SVM methods. On average, according to the PCC indicator, ensemble models are better than other tested models. Our results confirm the Beque and Lessmann (2017), but it should be noted that our research relies on the Matlab package while Beque and Lessmann (2017) use the R programming environment. Our research on the Australian dataset shows that logistic regression is inferior to SVM and ensemble methods.

Table 1 Summary of results of individual classifiers obtained in predicting the Australian credit scoring dataset

Classification Learner	AUC	TP	FP	TN	FN	PCC
Logistic regression						
Logistic regression	0.90 (0.13,0.82)	266 (87%)	69 (18%)	314 (82%)	41 (13%)	84.06%
Support Vector Machines (Box constraint level, Manual kernel scale)						
Linear (1,-)	0.91 (0.07,0.80)	**285** **(93%)**	78 (20%)	305 (80%)	**22** **(7%)**	85.51%
Quadratic (1,-)	0.90 (0.16,0.84)	257 (84%)	62 (16%)	321 (84%)	50 (16%)	83.77%
Cubic (1,-)	0.90 (0.18,0.84)	253 (82%)	60 (16%)	323 (84%)	54 (18%)	83.48%
Fine Gaussian (1,0.94)	0.89 (0.24,0.89)	233 (76%)	**43** **(11%)**	**340** **(89%)**	74 (24%)	83.04%
Medium Gaussian (1,3.7)	0.92 (0.07,0.80)	**285** **(93%)**	77 (20%)	306 (80%)	**22** **(7%)**	85.65%
Coarse Gaussian (1,15)	0.92 (0.07,0.80)	**285** **(93%)**	77 (20%)	306 (80%)	**22** **(7%)**	85.65%
Ensemble (Method, Maximum number of splits, Number of learners, Learning rate)						
Boosted Trees (AdaBoost, 20,30,0.1))	0.92 (0.20,0.87)	246 (80%)	49 (13%)	334 (87%)	61 (20%)	84.06%
Bagged Trees (Bag, 689,30,-)	**0.92** **(0.12,0.87)**	269 (88%)	49 (13%)	334 (87%)	38 (12%)	87.39%
RUSBoosted Trees (RUSBoost, 20,30,0.1)	0.91 (0.16,0.86)	259 (84%)	52 (14%)	331 (86%)	48 (16%)	85.51%

Source: Data processed by the author

In the German dataset, according to the AUC criterion, the best classifier is the ensemble RUSBoosted Trees with a value of 0.76 (70.38%), and logistic regression, linear and medium Gaussian SVM are close to it with 0.77 (about 65%) (Table 2). The RUSBoosted Trees model has the lowest rate of false positive cases (FP), while the Gaussian SVM models have the lowest rate of false negative cases (FN). According to the criterion of false positive cases (FP), ensemble models give the best results. According to the PCC criteria, the Medium Gaussian SVM model (Table 2) proves to be the best classifier with accuracy of 74.10%. On average, according to the PCC indicator, ensemble models are equal to SVM models. In the German dataset, logistic regression yields results on a par with SVM and ensemble methods.

Table 2 Summary of results of individual classifiers obtained in predicting the German credit scoring dataset

Classification Learner	AUC	TP	FP	TN	FN	PCC
Logistic regression						
Logistic regression	**0.77** **(0.54,0.86)**	602 (86.0%)	163 (54.3%)	137 (45.7%)	98 (14.0%)	73.90%
Support Vector Machines (Box constraint level, Manual kernel scale)						
Linear (1,-)	**0.77** **(0.60,088)**	616 (88.0%)	179 (59.7%)	121 (40.3%)	84 (12.0%)	73.70%
Quadratic (1,-)	0.75 (0.57,0.85)	592 (84.6%)	171 (57.0%)	108 (43.0%)	129 (15.4%)	70.00%
Cubic (1,-)	0.70 (0.53,0.78)	547 (78.1%)	158 (52.7%)	142 (47.3%)	153 (21.9%)	68.90%
Fine Gaussian (1,0.94)	0.71 (0.99,1.0)	699 (99.9%)	296 (98.7)	4 (1.3%)	**1** **(0.1%)**	70.30%
Medium Gaussian (1,3.7)	**0.77** **(0.68,0.92)**	**645** **(92.1%)**	204 (68.0%)	96 (32.0%)	55 (7.9%)	74.10%
Ensemble (Method, Maximum number of splits, Number of learners, Learning rate)						
Boosted Trees (AdaBoost, 20,30,0.1))	0.76 (0.61,0.87)	608 (86.9%)	182 (60.7%)	118 (39.3%)	92 (13.1%)	72.60%
Bagged Trees (Bag, 689,30,-)	0.75 (0.58,0.86)	605 (86.4%)	173 (57.7%)	127 (42.3%)	95 (13.6%)	73.20%
RUSBoosted Trees (RUSBoost, 20,30,0.1)	0.76 (0.25,0.65)	458 (65.4%)	**74** **(24.7%)**	**226** **(75.3%)**	242 (34.6%)	68.40%

Source: Data processed by the author

Tables 3 and 4 show the performance indicators of different classifiers for Australian and German datasets, respectively. The models with the best performance indicators are in bold. All classifiers in both tested datasets record a high value of sensitivity, with large differences observed in terms of specificity indicator in German dataset. In both datasets, the Fine Gaussian SVM has the highest sensitivity value. In the Australian dataset, SVM algorithms have the highest specificity. However, the best ensemble in the German dataset is the Ensemble RUSBoosted Trees. The sensitivity of SVM methods is on average higher than ensemble methods in both datasets. However, the specificity of ensemble methods is on average higher than SVM in both datasets. Similarly, in terms of G-average and F-measure, which show a balance between sensitivity and specificity, ensemble methods, on average, show slightly better results than SVM in terms of G-average but lower in F-measure. In terms of MCC indicators, as the most relevant for the assessment of binary classification techniques of machine learning, in both datasets, the best results are recorded with ensemble techniques. In the Australian dataset, the MCC correlation of 74.60% Ensemble Bagged Trees shows that the predicted class and the correct class are highly correlated. However, the MCC/Ensemble RUSBoosted Trees correlation of 37.44% shows that the predicted class and the correct class are not as highly correlated (as with the Australian dataset).

Taking into account all indicators, the analysis of classifiers and their predictive abilities do not identify an individual classifier with high predictive power for all, or at least most performance indicators.

Table 3 Performance comparison of individual classifiers for the Australian credit scoring dataset

Method	Sensitivity/Recall	Specificity	G-mean	F-measure	MCC
Logistic regression	81.98%	86.64%	84.28%	83.01%	68.24%
Linear SVM	79.63%	**92.83%**	85.98%	82.47%	72.13%
Quadratic SVM	83.81%	83.71%	83.76%	83.79%	67.31%
Cubic SVM	84.33%	82.41%	83.37%	83.90%	66.63%
Fine Gaussian SVM	**88.77%**	75.90%	82.08%	85.81%	65.60%
Medium Gaussian SVM	79.90%	**92.83%**	86.12%	82.67%	72.37%
Coarse Gaussian SVM	79.90%	**92.83%**	86.12%	82.67%	72.37%
Ensemble Boosted Trees	87.21%	80.13%	83.59%	85.60%	67.64%
Ensemble Bagged Trees	87.21%	87.62%	**87.41%**	**87.30%**	**74.60%**
Ensemble RUSBoosted Trees	86.42%	84.36%	85.39%	85.96%	70.70%

Source: Data processed by the author

Table 4 Performance comparison of individual classifiers for the German credit scoring dataset

Method	Sensitivity/ Recall	Specificity	G-mean	F-measure	MCC
Logistic regression	86.00%	45.67%	62.67%	79.49%	34.23%
Linear SVM	88.00%	40.33%	59.58%	80.22%	32.16%
Quadratic SVM	82.11%	38.71%	56.38%	75.57%	21.96%
Cubic SVM	78.14%	47.33%	60.82%	73.23%	25.60%
Fine Gaussian SVM	**99.86%**	1.33%	11.54%	**82.51%**	7.73%
Medium Gaussian SVM	92.14%	32.00%	54.30%	82.14%	30.90%
Ensemble Boosted Trees	86.86%	39.33%	58.45%	79.09%	29.47%
Ensemble Bagged Trees	86.43%	42.33%	60.49%	79.27%	31.71%
Ensemble RUSBoosted Trees	65.43%	**75.33%**	**70.21%**	66.88%	**37.44%**

Source: Data processed by the author

7. Conclusion

Credit scoring is a widely used technique that helps banks decide when granting loans to applicants. In addition to using standard statistical decision-making techniques, such as logistic regression or decision tree, credit scoring is a very interesting task for machine learning and artificial intelligence methods. In recent years, machine learning technologies have been developing rapidly and ensemble learning is being studied more and more.

Several papers have shown the advantages of deep learning over traditional credit scoring methods. In this paper, we investigated the predictive capabilities of ensemble algorithms over credit scoring decision trees and compared them with traditional methods – logistic regression and SVM support vectors.

Ensemble methods are promising classifier and predictive techniques and represent an alternative to classical artificial neural networks (ANN). A large number of studies have shown that in problems of credit scoring classification, ensemble techniques are better than SVM techniques, as well as than traditional techniques such as logistic regression and decision tree. According to each comparison criterion, ensemble methods show better results or at least competitive results with tested predictive techniques.

Based on Australian and German credit scoring data, performance of different classification models is compared. The performance of logistic regression (LR), support vector machine (SVM), and deep learning based on decision tree ensembles is analyzed. Several different performance indicators are used. According to the MCC indicator, which is considered to be the most adequate for classification problems, on average, deep learning methods prove to be the best. Individually, ensembles with boosted trees work best in the Australian dataset, and ensembles with RUSBoosted trees in the German dataset.

Logistic regression performance is relatively poor compared to the ensemble method and SVM. Logistic regression, as the best of the traditional methods, fails to match the tools of machine learning.

The results of this paper confirm previous research on the advantages of machine learning methods over traditional models. Also, the division regarding the obvious advantage of deep learning over SVM methods is confirmed. According to certain criteria, SVM shows better characteristics compared to ensembles. Nevertheless, ensemble methods are promising tools and provide potential for future research.

REFERENCES

Abdou, H., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting Finance & Management*, *18*, 59–88.

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

Asuncion, A., & Newman, D. J. (2010). *UCI machine learning repository.* School of information and computer science, Retrieved from: http://archive.ics.uci.edu/ml/

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *The Journal of the Operational Research Society*, *54*(6), 627-635.

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems With Applications*, *86*, 42-53.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, *91*, 1-21. https://doi.org/10.1016/j.asoc.2020.106263

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning. *Neural Computation*, *10*(7), 1895–1923.

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing*. New York: National Bureau of Economy Research.

Einav, L., Jenkins, M., & Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, *44*(2), 249–274.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179-188.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

Goh, R., & Lee, L. (2019). Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. *Advances in Operations Research*, 1-30. https://doi.org/10.1155/2019/1974794

Hand, D. J., & Henley, W. E. (1997). Statistical classifcation methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*, *160*(3), 523-541.

Hui, L., Li, S., & Zongfang, Z. (2017). The Model and Empirical Research of Application Scoring Based on Data Mining Methods. *Procedia Computer Science*, *17*, 911-918.

Jurman, G. R. S. (2012). A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE*, *7*(8), 41882.

Kim, A. Y.-C. (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, *283*(1), 217-234.

Lewis, E. (1992). *An Introduction to Credit Scoring*. San Rafael: Fair, Isaac and Co., Inc.

Li, Y., Lin, X., Wang, X., Shen, F., & Gong, Z. (2017). Credit Risk Assessment Algorithm Using Deep Neural Networks with Clustering and Merging. 13th International Conference on Computational Intelligence and Security (CIS) (pp. 173-176). Hong Kong: IEEE.

Lou, C., Wu, D., & Wu., D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, *65*, 465-470.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451.

Nanni, L., & Lumini, A. (2009). AAn experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, *36*, 3028–3033.

Neagoe, V., Ciotec, A., & Cucu, G. (2018). Deep Convolutional Neural Networks Versus Multilayer Perceptron for Financial Prediction. *2018 International Conference on Communications (COMM)* (pp. 201-206). Bucharest: IEEE

Odom, M., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *1990 IJCNN International Joint Conference on Neural Networks*, 2, pp. 163-168.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109-131.

Orgler, Y. (1971). Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research*, *2*(1), 31-37.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. *European Journal of Operational Research*, *266*(2), 639–665.

Patil, P. S., Aghav, J. V., & Sareen, V. (2016). An Overview of Classification Algorithms and Ensemble Methods in Personal Credit Scoring. *International Journal of Computer Science and Technology*, *7*(2), 183-188.

Pławiak, P., Abdar, U., & Acharya, R. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing Journal*, *84*, 105740.

Ren, Y., Zhang, L., & Suganthan, P. (2016). Ensemble Classification and Regression-Recent Developments, Applications and Future Directions. *IEEE Computational Intelligence Magazine*, *11*(1), 41-53.

Sadatrasoul, S. M., Gholamian, M. R., Siami, M., & Hajimohammadi, Z. (2013). Credit scoring in banks and fnancial institutions via data mining techniques: a literature review. *Journal of AI and Data Mining*, *1*(2), 119-129.

Thomas, L., Crook, J., & Edelman, D. (2002). *Credit Scoring and Its Applications, Second Edition*. Philadelphia: Society for Industrial and. and Applied Mathematics.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley- Interscience.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223-230.

Zhou, H., Lan, Y., Soh, Y., Huang, G., & Zhang, R. (2012). Credit risk evaluation with extreme learning machine. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (pp. 1064-1069). Seoul.

KREDITNO BODOVANJE POMOĆU ANSAMBLERSKIH METODA DUBOKOG UČENJA ZA KLASIFIKACIJU – POREĐENJE SA TRADICIONALNIM METODAMA

Kreditni skoring (kreditno bodovanje) privlači posebnu pažnju finansijskih institucija. Poslednjih godina, posebno su interesantni metodi zasnovani na dubokom učenju. U ovom radu, upoređujemo performanse ansamblerskim metodama dubokog učenja zasnovanih na stablima odlučivanja sa najboljom klasičnom metodom, logističkom regresijom i već postavljenim reperom za metode mašinskog učenja, mašinama sa vektorskom podrškom. Za svaku metodu testirano je više različitih algoritama. Takođe, korišćeni su različiti indikatori performansi. Istraživanje je izvršeno nad standardnim bazama za ovu vrstu klasifikacije, Australijskim i Nemačkim skupom podataka. Kao najbolja metoda, prema MCC indikatoru, pokazala se ansemblerska metoda sa boosted stablima odlučivnja. Takođe, u proseku, ansemblerske metode su se pokazele uspešnijim od SVM.

Ključne reči: *kreditno bodovanje, ansambli za klasifikaciju, duboko učenje, mašine za vektorsku podršku*