Advanced Engineering Informatics 45 (2020) 101130

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers

Jasmina Nalić, Goran Martinović*, Drago Žagar

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, J.J. Strossmayer University of Osijek, KnezaTrpimira 2b, 31000 Osijek, Croatia

ARTICLE INFO

Keywords:
Credit scoring
Data mining
Ensemble classifier
Feature selection
Hybrid model

ABSTRACT

The aim of this paper is to propose a new hybrid data mining model based on combination of various feature selection and ensemble learning classification algorithms, in order to support decision making process. The model is built through several stages. In the first stage, initial dataset is preprocessed and apart of applying different preprocessing techniques, we paid a great attention to the feature selection. Five different feature selection algorithms were applied and their results, based on ROC and accuracy measures of logistic regression algorithm, were combined based on different voting types. We also proposed a new voting method, called *if_any*, that outperformed all other voting methods, as well as a single feature selection algorithm's results. In the next stage, a four different classification algorithms, including generalized linear model, support vector machine, naive Bayes and decision tree, were performed based on dataset obtained in the feature selection process. These classifiers were combined in eight different ensemble models using soft voting method. Using the real dataset, the experimental results show that hybrid model that is based on features selected by *if_any* voting method and ensemble GLM + DT model performs the highest performance and outperforms all other ensemble and single classifier models.

1. Introduction

One of the most critical risks that financial institutions are exposed to is certainly credit risk. In order to prevent losses caused by bad credit decisions, banks are widely adopting credit scoring models that support decision making process by applying different data mining algorithms. Credit scoring is a method used to predict whether the client belongs to either legitimate or suspicious client group [1]. The predictions based on manual estimation and assessment of clients became time and resource consuming [2], so automatization of this process by using machine learning made great savings for financial institutions. Accuracy of credit scoring is crucial, and therefore many researches lately, as well as this one, are focused on finding methods that will increase credit scoring model accuracy. Apart from that, reliable credit scoring model is beneficial in terms of improving cash flows, reducing possible credit losses, reducing cost of credit analysis, analyzing customers purchasing behavior etc. [10]. The hybrid and ensemble credit scoring models built by many researchers outperformed performance of the single classifier and statistical approach of credit scoring [3].

Ensemble learning is approach that combines results of two or more individually trained classifiers in order to make more accurate predictions. The multiple classifiers systems are based on aggregating of a pool of classifiers such that their fusion achieves higher performance than the single classifiers [4]. As mentioned, classifiers are trained individually to produce their own decisions, which are then subsequently combined by voting to produce one decision [5]. There are several types of voting, including soft voting, hard (majority) voting and weighted voting. Based on previous literature, most commonly used voting strategies are majority, as in [4] and [7] and weighted voting, as in [1] and [7]. In this research, we used adjusted version of unanimous majority voting in combining results of FS algorithms, and soft voting for ensemble classifiers.

Hybrid models combine power of improving performance by reducing dimensionality of dataset and using single classifier or ensemble approach. As the part of data preprocessing, feature selection algorithms have been proven by many researchers to improve all performance measures of data mining model.

The main objective of this paper is to combine techniques and methods that are shown by previous researches to be efficient in improving the model's performance, in order to produce the most accurate hybrid model. The paper also presents a new approach in feature selection stage since it combines five different feature selection algorithms in order to improve results of using a single one.

This paper is organized as follows. Second section represents related

* Corresponding author.
E-mail address: goran.martinovic@ferit.hr (G. Martinović).

https://doi.org/10.1016/j.aei.2020.101130
Received 11 February 2019; Received in revised form 17 March 2020; Accepted 29 May 2020
Available online 12 June 2020

Table 1
Comparison of related researches experimental results of feature selection.

Related research	Applied classifier(s)	FS algorithm	Accuracy (baseline classifier) %	Accuracy (preprocessed dataset) %
[9]	Decision tree	Gain Ratio	63.20	74.80
[15]	Decision tree	Minimum description length (MDL)	77.76	91.30
[20]	Random forest	Novel proposed	73.40	76.20
[21]	Random forest	Novel proposed	86.67	87.04

research work focused on hybrid and ensemble data mining models and analyzes their achievements and conclusions. Section 3 presents proposed methodology used in building a new hybrid ensemble credit scoring model. In this section, used algorithms are explained in detail, as well as the way of finding out the most accurate one. In Section 4, experimental results of proposed model applied on real dataset are presented and compared in order to find the most accurate hybrid model. In order to prove that proposed model is generally applicable for various cases, we applied another, much bigger real dataset. Results of the validation of the proposed model are presented in Section 5.

2. Related work

Since it is considered as one of the most challenging risks in banking sector, credit scoring, as the classification problem, has become very popular research field. Accordingly, there is a growing interest among researchers to build a new model that would outperform performances of existing ones. Recent researches are mainly focused on taking advantages of several data mining algorithms by combining them in order to build more efficient and accurate model. As this research has the same aim, in this section we will present a brief overview of previous related researches.

In [1], authors propose a hybrid ensemble credit scoring model based on neighborhood rough set (NRS) for feature selection and multilayer ensemble approach based on weighted voting. Study is performed based on six different feature selection algorithms, including stepwise regression (STEP), regression tree (CART), correlations (CORR), multivariate adaptive regression splines (MARS), T-test and NRS, and each of them, as well as dataset without dimensionality reduction, are combined with ensemble models based on five classifiers aggregated by four different voting types (majority voting, weighted voting, layered majority voting and layered weighted voting). Experimental results based on Australian and German credit dataset show that in terms of accuracy, sensitivity and G-measure, proposed model outperforms 27 other trained hybrid ensemble models. Accuracy of proposed model for Australian and German dataset is 95.39% and 86.47%, respectively, and sensitivity 94.46% and 92.66%.

In [2], authors present hybrid ensemble credit scoring model, based on two feature selection algorithms, Gabriel neighborhood graph editing (GNG) and MARS, and five base classifiers, including neural networks (NN), decision tree (DT), support vector machine (SVM), random forest (RF) and naive Bayes (NB). Apart from using traditional combining methods (MIN, MAX, AVG, majority voting, weighted voting), this study also proposes novel consensus classifier combination rule (consA). Seven datasets, including German, Australian and Japanese datasets, were used and based on accuracy, AUC, H-measure and brief score. A hybrid ensemble model based on consA combination rule outperforms all other models. Accuracy of models using consA is in range of 79% and 95%.

Authors in [3] propose a hybrid ensemble credit scoring model based on Principal Component Analysis (PCA), as feature selection algorithm, and ANN-adaptive boosting (AdaBoost) algorithm as most accurate ensemble. Apart from PCA, another tree feature selectors are tested, including genetic algorithm (GA), information gain radio and attribute evaluation as feature selectors and based of accuracy of SVM classifier, PCA shows the best results. Ensemble algorithms NB-AdaBoost, CART-AdaBoost, ANN-AdaBoost, SVM-AdaBoost, NB-bagging, CART-bagging, ANN-bagging and SVM-bagging are compared with single classifiers in terms of accuracy and AUC measure. Based on real life dataset, classification results show the proposed model's accuracy of 91%.

In [4], authors engage fuzzy clustering of dataset as pre-processing method and then build ensemble models based on three classifiers SVM, neural network (NNs) and DT combined by majority vote and novel proposed voting method named membership degree fusion (MDF) method. Using German dataset, hybrid model using MDF showed the highest performance in terms of accuracy. Authors in [8] also make similar research on the idea of clustering plus ensemble model that compares several versions of proposed hybrid model based on different combinations of clustering and classification algorithms.

There are also some researches, like [9], [15], [18], [20] and [21] that focus on feature selection as the way of improving model's performances. They compare combination of one or more feature selectors with one or more single classifiers and analyze results. Experimental results in all these studies show that algorithms perform better based on dataset with reduced dimensionality. Table 1 represents comparison of mentioned researches experimental results, based on methods that show the greatest performance improvement.

Regarding feature selection, some researches as [3] and [22], present models based on nature inspired algorithms that are very popular lately. As mentioned earlier, research [3] employs four feature selection algorithms and one of these is genetic algorithm. Still, based on algorithm's performance, authors in [3] propose model based on PCA for feature selection. As it is shown in Section 3.1, feature selection algorithm proposed by this research outperforms PCA algorithm. In [22], a new hybrid model consisting of particle swarm optimization (PSO) as feature selector and SVM as main classifier is proposed. Authors present model's results based on German and Australian UCI datasets and make comparison with the other model's in literature based on SVM classifier. Results show that accuracy of the PSO + SVM model based on the German and Australian datasets is 78.70% and 87.10%, respectively. Comparison of these results with results of other similar models shows that proposed models outperformed other models for both datasets.

Authors in [23] provide review and detailed discussions on credit scoring models proposed by researchers since 1997 to 2018 and based on SVM and metaheuristic algorithms (MA). Authors investigate usage of these algorithms as the main classifiers, as well as feature selectors. They report that both algorithms have been actively researched throughout the years and showed their great potential in the domain of credit scoring. Based on experimental results presented in observed studies that are based on generic datasets (German and Australian UCI datasets), authors conclude that hybrid modeling is the state-of-the-art approach for both methods. They also state that, according to the literature, GA is the most popular method to be hybridized with data mining classifiers, while SVM is current trend main classifier.

In this research, we propose a hybrid ensemble model based on cognition of previous related research, but we also propose a novel model building approach in order to outperform standard models. In terms of feature selection, we investigated the five most commonly used algorithms in related research. Instead of applying single feature selector, the contribution of this paper is reinforced by adopting a new feature selection approach that aggregated results of five feature

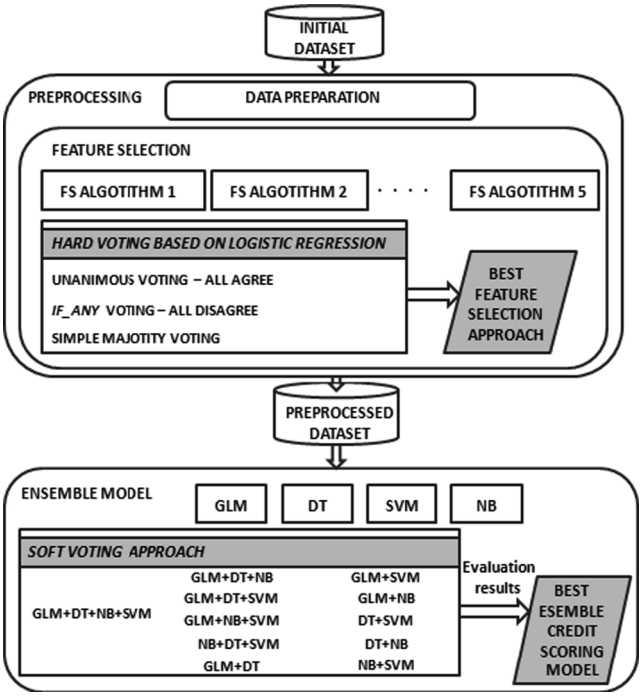

Fig. 1. A diagram of proposed hybrid credit scoring model.

selection algorithms. This approach shows performance improvement of the model based on a real dataset. Further, we engage four base classifiers that have been proven as the most efficient classification algorithms for credit scoring. With the respect to the similar related papers, we involve different combination of classification algorithms in order to investigate their impact on ensemble credit scoring models. In order to adopt the most efficient ensemble model, we analyze five different performance measures of eight ensemble learners.

3. Proposed methodology and the model

As mentioned above, this research presents a new hybrid model that combines both novel feature selection and ensemble approach, with the same goal to build a model that will outperform performance of single approach models.

A block diagram of proposed model is presented in Fig. 1.

In the first stage, initial data set is preprocessed. Data pre-processing is of crucial importance for the efficiency and accuracy of data mining models [15]. That implies usage of various tasks and methods to prepare dataset in order to achieve better performance of the classification model. In our experimental work, various data preparation techniques were applied, and so prepared dataset is then used in a feature selection process. Instead of using a single feature selection algorithm, we engage five different feature selection algorithms and aggregate their results in order to find a set of parameters that achieve higher accuracy of the model. As the aggregation rule, we use three different types of hard voting and adapt a voting model with the highest accuracy based on logistic regression.

In the next stage, we apply four different classifiers, including generalized linear model (GLM), decision tree (DT), support vector machine (SVM) and naive Bayes (NB), to reduced dataset. Based on soft voting approach, we adapt an ensemble credit scoring model that has the highest performance and also outperforms performance of single classifiers. All single classifiers involved in model building are trained on 60% of dataset and the classification results are based on the rest 40%.

3.1. Feature selection

Feature selection is a data pre-processing technique that is used to produce a dataset with the most relevant features in order to achieve the highest model performances. According to [1], [15] and [16], it improves both model accuracy and running time.

In this research, we trained individually five different FS algorithms:

- *Classifier feature evaluation (ClassFE)*-uses decision tree classifier to evaluate the worth of a feature.
- *Correlation feature evaluator (CorrelationFE)* - evaluates the worth of a feature by measuring the Pearson's correlation between that feature and the class.
- *Gain ratio feature evaluator (GainRFE)* - evaluates the worth of a feature by measuring the gain ratio with respect to the class. GainR (Class, Attribute) = (H(Class) - H(Class | Attribute))/H(Attribute).H (Class) and H(Class | Attribute) signifies respectively the degrees of uncertainty over the choice before and after the attribute is selected [24].
- *Information gain feature evaluator (InfoGainFE)* - evaluates the worth of a feature by measuring the information gain with respect to the class. InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute). Based on explanation of previous formula, information gain represents a reduction of uncertainty over the choice Attribute. The attribute with the highest information gain is the one that reduces the degree of uncertainty the most [24].
- *Relief feature evaluator (RefilefFE)* - evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. Relief algorithm is successful attribute estimator able to detect conditional dependencies between attributes and provide a unified view on the attribute estimation in classification and regression [9].

All FS algorithms were trained on 26 features dataset and each of them selected different set of features. Table 2 represents outcomes of FS algorithms, where features are marked as F1 to F26, FS algorithms are marked as FSA1 to FSA5 and "X" denotes that algorithm recognized

Table 2
Results of five feature selection algorithms.

Feature	FS 1	FS 2	FS 3	FS 4	FS 5
F1					
F2					
F3					X
F4					
F5	X				X
F6					X
F7					X
F8					X
F9		X			
F10				X	X
F11					X
F12					X
F13					
F14					X
F15					
F16				X	
F17	X	X	X	X	
F18		X	X	X	X
F19	X				
F20	X	X	X	X	X
F21	X		X		X
F22					
F23					X
F24		X			
F25		X			X
F26		X			X

Table 3
Comparison of LR performances for different voting types.

Voting type	ROC	Accuracy
Unanimous	0.799	82.0539
Majority	0.748	79.6406
if_any	0.8	82.1309

that feature as the worth one.

Since every FS algorithm suggested different set of features as the most optimal one, we applied different types of voting in order to make fusion of these five algorithms and outperform performance of single FS algorithm. Results, presented in the Section 3.1.1, are compared on accuracy of logistic regression (LR), as the one of the most adopted algorithms for credit scoring models ([4,10,11,12,15;18]).

3.1.1. Hard (majority) voting

Hard or majority voting is based on binary decision rule that selects features with major of votes. There are two types of majority voting: simple and unanimous majority voting. Simple majority voting represents type of voting where feature is selected if it has half or more votes and unanimous voting denotes that feature is selected if it has all the votes. In addition to these types, we proposed *if_any* voting type that selects those features that are selected by any algorithm. In other words, *if_any* voting does not select those features that are recognized as unworthy by all algorithms (all algorithms disagree). Results of all three voting types are compared in Table 3, based on ROC and accuracy of logistic regression. It shows that combination of feature selection algorithms based on *if_any* voting type outperformed combinations based on unanimous and majority voting types.

Since a simple majority voting, out of 26 features, selects only four, it was expected that LR performs with poor accuracy. On the other side, unanimous all agree algorithm showed high performance, but it selected only one feature, so these results are irrelevant. In that way, *if_any* voting type outperformed the other one, so we adopted this voting type for the building pre-processed dataset.

In order to validate results of proposed feature selection algorithm *if_any*, we engaged Principal Component Analysis (PCA) algorithm, as one of the most commonly used feature selection algorithms for credit scoring. PCA is statistical technique that reduce dataset dimensionality by extracting new independent features [3]. It selects covariance or correlation matrix based on characteristics of the initial features and then computes the characteristics roots of the covariance or correlation matrix [18].

PCA is applied on the same 26 features dataset and it created 55 new features. Table 4 presents comparison of the feature selection algorithm based on *if_any* voting type with PCA. Results based on LR classifier show that proposed *if_any* feature selection algorithm outperforms PCA feature selector in terms of accuracy and ROC.

3.2. Ensemble model

Ensemble modeling focuses on gathering the insight of several classifiers trained in order to solve the same problem, using their opinions to reach an effective and accurate decision [2]. The error and deviation of one classifier in ensemble classification models are compensated by the other members of ensemble, therefore ability of ensemble model is usually much stronger than that of a single classifier [4].The ensemble model in this research combines four classification algorithms - GLM, DT, SVM and NB.

Table 4
Comparison of PCA and IF_ANY feature selection algorithms performances based on LR classifier.

Feature selection	ROC	Accuracy
PCA	0.775	80.8974
if_any	0.800	82.1309

Generalized linear model is a parametric modeling technique with ability to predict bounds of confidence (probability interval) and assume the distribution of the data [25]. It produces significant amount of statistic indicators that makes the results of this model very understandable and convenient for interpretation. In this research, a binary logistic regression with the binomial variance and the logit link function has been used. Logistic regression is used to identify relationship between the dependent binary variable or response, and the explanatory variables called predictors [27]. Logistic regression is presented by following equation.

$$Logit\,(p) = \log[\frac{p}{1-p}] = \alpha + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn \quad (1)$$

In presented equation, p represents the probability of outcome, $\beta 1 \cdots \beta n$ represent regression coefficients in the linear combination of descriptive predictors $X1 \cdots Xn$, and α is intercept.

According to many authors [25,27,28], due its simplicity and understandability, logistic regression models are the most popular and the most practical ones for credit scoring problem. Another advantage of using these models is their ability to provide information about the dependence between variables and outcome, since credit regulators today mostly insist on reason of loan application rejection [29].

Decision tree is a very popular classification algorithm, due to its ability to be interpreted by simple rules with minimum user intervention. A tree is started by root node that represents an attribute and then branches according to attribute value until all branches on node reach certain criteria or until certain value is achieved [9]. Credit scoring model developed on decision trees is very efficient since the model built is used for sequence of simple tests based on a single predictor. At any given level, the question asked is an outcome of previous answers, thus supporting the ultimate goal of this algorithm where answers taken together uniquely identify a specific target value. Briefly, decision trees generalize better for unobserved instances, they are computationally efficient, and the tree architecture provides high level of understanding regarding classification of instances [10]. This makes decision tree model suitable for credit scoring problems in terms of providing understandable information about approval or rejection of loan application.

Support Vector Machine is an artificial intelligence technique or an optimization technique used to find an optimal hyper plane that maximizes the margin between itself and the nearest training samples in the high-dimensional space and minimizes the expected generalization error [4]. Basically, in the process of prediction, this technique does not require any information regarding correlation among variables. For the non-linear data, the original model is upgraded with a function to map the data into a higher dimensional space and thus producing an optimal line (hyperplane) that can almost perfectly separate the two classes in space [14]. SVM is one of the best algorithms to classify small samples and very well suited for consumer credit-risk analytics with possible complex relationships between consumers' characteristics and their transactions [19].

Naive Bayes is a simple and effective classification algorithm based on Bayes' theorem that provides high level of prediction and scalability across many different fields of research where the input feature space is very high. It enables usage of probabilistic knowledge for the clear and unambiguous expression of its statistical components and projected results [13]. When using this algorithm, a classifier is calculated by multiplying previous probability of a class before encountering any data with the likelihood of the data given its class. For example, in a context of credit scoring, let the training set D = {X1,X2,..Xn} where each × consists of n characteristics of attributes {X11, X12, …X1n} and supported with a class label c that represents either good or bad loan. The main goal of NB classifier is to determine mapping function f: (x11,

…x1n}- > (c) that can predict a label of an unknown example × = (x1,..xn) bad on analysis of the training set [14].

In this research, classifiers were trained individually and then combined into eight different combinations (each-to-each) based on soft voting approach applied on classifiers outcomes and probability of outcomes for each dataset instance.

3.2.1. Soft voting

The soft voting corresponds to the classifiers that have probabilistic outcome. Weighted soft computing method can be defined at classifier level, class level or instance level [14]. Soft voting is able to achieve better performance than hard voting, because it gives more weight to highly confident vote [17] and takes into account each classifier's certainty in the final decision.

In this research, we combined output probabilities of a single instance for every combination of single classifiers or for every of eight ensemble models. Decision of each instance is made by average probability for target class. In combination of all four classifiers, if average probability of class *bad* (bad client) for an instance is higher than the average probability of class *good*, then common decision is class *bad*.

The algorithm used to combine classifiers by soft voting approach is presented by following pseudo-code.

Pseudo - code for combining algorithms into ensemble classifier by soft voting

```
Input: T - table consisting of i records and predicted class probabilities for each re-
    cord.
  T (i) [predicted_class (n), probability (n)]
x - number of combined algorithms [2.0.4]
Output: C_output (i) predicted class for record i
Step 1: Initialize
P_good (i) represents probability of outcome good for record i
  P_bad (i) represents probability of outcome bad for record i
Step2: for each i in T do
  for n = 1 to x do
    if predicted_class(n) is good
P_good (i) = P_good (i) + probability (n)
    else
    P_bad (i) = P_bad (i) + probability (n)
    end if
  end for
    if P_good (i) > P_bad (i) do
    C_output (i) = good
    else
    C_outpu t (i) = bad
    end if
end for
```

After calculating outcomes for every ensemble model, results of each are analyzed and compared in terms of ensemble model accuracy, type I error, type II error, sensitivity and F-measure. These results are presented and discussed in Section 4.3.

4. Experimental results

4.1. Dataset

This research is based on real-life dataset of a microfinance institution in Bosnia and Herzegovina, consisting of client personal, demographical, social, financial and credit history data. Initial dataset represented loan applications gathered over period of 12 months and described by 32 features that microfinance institution, based on their experience, considered as significant ones for decision making process. Structure of the used dataset is similar to the most commonly used German and Australian credit datasets from UCI Machine learning Repository, with the difference that used dataset very likely contains more outliers and data noise.

Before accessing a feature selection, we applied various pre-processing techniques on initial dataset, including:

(1) Removing records with missing values of most attributes. Since microcredit institution started to collect some of information recently (duration on address, duration of employment, number of adults and number of non-adults), records without these values were removed.
(2) Removing redundant attributes. For example, feature date of birth is removed, since we use feature age.
(3) removing attributes with the same value of all records. For example, we removed feature Type of client, since all clients of microfinance institution are individuals.
(4) Data aggregation. We aggregated several features (number of client's incomes, number of dependent member's income, number of independent member's income) into one feature - number of incomes in household
(5) Equal interval binning. For instance, we divided values of feature age into four age groups. The same we did for the attribute Amount of loan.

After applying all these techniques, initial dataset consisting of 32 features and 13,600 records was modified into 26 features and 12,983 records. Target feature is binary variable with value *bad* denotes clients that were late with repayment of their obligations more than 30 days, and *good* if they paid their obligations on time. 60% of dataset was used for building the model and the rest 40% for training the model.

As mentioned previously, we aggregated the results of five different FS algorithms, using *if_any* voting. Out of 26 features, proposed methodology selected 20 features as important ones and we used that dataset to build ensemble model. Table 5 represents selected features.

4.2. Performance Measures

All results gained in this research are evaluated by certain measures of performance. As presented earlier, results of combining feature selection algorithms are analyzed in terms of accuracy and ROC and comparison of eight different ensemble models is based on model accuracy, type I error, type II error, G-measure and Sensitivity, as the most standard measurements for evaluation of classification performance [26]. All these measures are calculated based on confusion matrix presented in the Table 6. Columns of matrix present predicted

Table 5
Description of selected features used for building ensemble model.

Feature	Feature description and values
Qualification	High school education, University degree, low education etc.
Own business	Yes/No
Social status	Domestic/Returnee
Nationality	Bosnian/Croatian/Serbian/Other
Urban/Rural	Urban/Rural
Marital status	Single/Married/Divorced/Widow
Age	4 Age groups
Duration on address	Short/Medium/Long/Very long
Source of income	Public sector/Private company/War veteran/ Retiree/Small house business/Other
Duration of employment	Short/Medium/Long/Very long
No of incomes in household	Number of household's members with income
No of non-adults	Number of children in household
No of adults	Number of adults in household
No of pets	Number of pets in household
No of dependent members	Number of dependent adult members in household (unemployed etc.)
Amount of loan	Low/Medium/High/Very high
Type of loan	Agriculture/Multipurpose/Trade/Services/ Residential/Production
Client cycle	Number of previous loans in the institution
Collaterals	Yes/No
Client segment	Employed/Unemployed/Family member employed

Table 6
Confusion matrix.

	Good	Bad
Good	True negative (TN)	False negative (FN)
Bad	False positive (FP)	True positive (TP)

values and rows present actual values. Since it is of the much greater importance to identify bad client for the given problem of credit scoring, our target class is bad.

In terms of class prediction, outcome of the binary classification model can be rated as follows:

- TP denotes actually bad clients that are classified as bad
- TN denotes actually good clients that are classified as good
- FP denotes actually bad clients that are classified as good
- FN denotes actually good clients that are classified as bad

Calculation of measures of performance are presented in the following equations.

$$Accuracy = \frac{TP + TN}{POS + NEG} \tag{2}$$

$$Type\ I\ error = \frac{FN}{POS + NEG} \tag{3}$$

$$Type\ II\ error = \frac{FP}{POS + NEG} \tag{4}$$

$$F - measure = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

Accuracy is considered to be the main measure of models performance. From the aspect of credit scoring, accuracy denotes a ratio of model's correct predictions to total number of instances.

Type I error denotes model's misclassifications in terms of predicting good clients and type II error represents wrong prediction of actually bad clients.

Sensitivity or recall represents the accuracy of positive class prediction. It is the measure that represents ratio of correctly classified bad clients to total number of clients predicted as bad. The ratio of correctly classified bad clients to the total number of actually bad clients is presented by precision of the model. F-measure represents harmonic mean of recall and precision [8].

4.3. Results and analysis

Results of FS algorithms are already discussed earlier while in this section an analysis and comparison of eight ensembles will be presented. Every ensemble is built based on dataset consisting of features selected by *if_any* voting on five different FS algorithms. Confusion matrix is created for every ensemble model and based on those, performance measures are presented in the Table 7. The best result for each performance measure is presented as bolded in a table

As presented in Table 7, DT classifiers showed the best performance out of all other single classifiers, in terms of all measures, except in terms of Type II error where the GLM single classifier showed the best results. It is also evident that all ensemble classifiers that involve DT showed the better performances than those without DT algorithm. This was expected, since almost all features have discrete values and DT performs very well when dealing with discrete values.

Ensemble model GLM + DT outperformed other ensemble models and single classifiers in terms of accuracy, type I error, F-measure and Sensitivity. As a single classifier, DT showed the best results in terms of same measures, and it took advantage of GLM algorithm, as the algorithm with the lowest type II error. Out of 2930 actually bad clients, single GLM classifier made 877 wrong decisions, that is the best result in terms of type II error. Combining great GLM's power of recognizing bad clients with high accuracy and low Type I error of DT classifier, ensemble model GLM + DT showed the best performances, although it did not show the best results in terms of Type II error rate. Since these two error type should not be considered separately [6], it is useful to involve error rate measure, because for the given problem of credit scoring, no matter how important it is to recognize a bad client, it is also as important not to decline a good one. Error rate represents the percentage of misclassifications in testing dataset and for GLM + DT model it is 17.985%. Other ensemble models, especially those that do not include DT classifier, have significantly higher error rate. Higher Type II error of GLM + DT model is compensated with twice lower Type I error rate, so in terms of error rate, this model still outperforms all other ensemble single classifiers. According to [28], logistic regression is superior to other algorithms in terms of sensitivity or prediction of bad clients in credit scoring. The presented results in Table 5 show sensitivity improvement of proposed GLM + DT model compared to GLM model based on logistic regression, as well as compared to other single classifiers.

In terms of performance measures, comparison with hybrid and ensemble models proposed by other authors in [1], [2], [3] and [4], is not relevant due to various reasons. In this research, we proposed a hybrid ensemble model based on combination of classifiers that, according to our knowledge, is not proposed in the literature. In addition to that, mentioned similar models are mostly based on smaller dataset from UCI Machine learning Repository. Our research is based on more than ten times larger real-life dataset. That makes our performance measure results not directly comparable with the results of other proposed models. Considering usage of hybrid ensemble models, our research confirms the cognitions of previous related researches:

- using hybrid models that include feature selection algorithm in combination with classification algorithm is much more efficient then using single classifier, as in [7], [8], [9], [18], [19] and [20]
- combination of two or more feature selection algorithms in the model is efficient in improving model's performances([8,6])
- performances of hybrid ensemble models outperforms single classifiers, as well as models consisting of feature selector and single classifier ([1–4]).

Based on presented research, we can conclude that due to simplicity and interpretability of both GLM and DT models, ensemble consisting of these two models, that outperformed single classifiers, is very suitable for the given problem of credit scoring.

5. Model application

In order to validate proposed credit scoring model, we involved the new, bigger dataset. Our aim was to test the model on a bigger dataset, so we applied the same structure dataset of other microfinance institution consisting of 70,348 records. Out of total number of records, 9.81% or 6898 records have target value *bad* and the rest 90.19% or 63,450 are clients classified as *good* ones. We performed testing for all eight ensembles, as well as for single classifiers. Results of model application are presented in Table 8 and those results releted to the ensemble model with highest performance are bolded.

Application of the proposed credit scoring model based on the new and bigger dataset confirms earlier results presented in Section 4. Ensemble model GLM + DT outperformed other ensembles and single classifiers in terms of all measures of performance, even in terms of Type II error. Confusion matrix of proposed ensemble GLM + DT is presented in Table 9.

As noticeable from the Table 9, out of 70,348 records, the proposed

Table 7
Measure of performances of base and ensemble classifiers.

Single classifier/Ensemble	Accuracy (%)	Type I error (%)	Type II error (%)	Sensitivity (%)	F measure (%)
GLM	73.42679	19.81822	**6.754987**	44.37959	54.34092
DT	80.92891	10.05161	9.019487	57.40862	58.69203
SVM	74.94416	17.93114	7.124702	46.27279	55.21135
NB	71.27782	21.19695	7.525225	41.50903	51.15914
GLM + DT + SVM + NB	76.81584	15.75137	7.432797	49.00249	56.62824
GLM + DT + SVM	75.25225	17.62305	7.124702	46.70394	55.5171
GLM + DT + NB	75.93776	16.86821	7.194023	47.68275	56.09893
DT + NB + SVM	76.08411	16.62174	7.294154	56.08825	47.88698
GLM + DT	**82.01494**	**8.618963**	9.366094	**60.50124**	**59.48291**
GLM + SVM	74.38188	18.66287	6.955249	45.55056	54.93225
GLM + NB	72.97235	19.93376	7.093892	43.70241	53.38116
DT + SVM	81.46037	9.574058	8.96557	58.69059	59.47129
DT + NB	81.5605	8.819225	9.620273	59.48337	58.40862
SVM + NB	73.7811	19.20203	7.016868	44.74734	54.25961

Table 8
Measure of performances of model application.

Single classifier/Ensemble	Accuracy (%)	Type I error (%)	Type II error (%)	Sensitivity (%)	F measure (%)
GLM	79.739296	17.547052	2.7136521	41.178655	79.739296
DT	85.061409	11.099107	3.8394837	44.40565	85.061409
SVM	79.342696	17.549895	3.1074089	39.338788	79.342696
NB	76.431455	20.212373	3.3561722	35.370702	76.431455
GLM + DT + SVM + NB	78.758458	18.375789	2.8657531	39.519165	78.758458
GLM + DT + SVM	75.25225	17.62305	7.124702	46.70394	55.5171
GLM + DT + NB	78.022119	19.27418	2.7037016	39.256669	78.022119
DT + NB + SVM	77.989424	19.197419	2.8131574	38.851592	77.989424
GLM + DT	**87.686928**	**10.740888**	**1.572184**	**57.21624**	**87.686928**
GLM + SVM	76.111617	20.887587	3.000796	36.294022	76.111617
GLM + NB	74.016319	23.443453	2.5402286	35.865408	74.016319
DT + SVM	83.44658	13.549781	3.0036391	45.109592	83.44658
DT + NB	83.081253	13.244158	3.6745892	42.020655	83.081253
SVM + NB	75.344004	21.720589	2.9354068	35.785421	75.344004

credit scoring model made 61,686 correct classifications. That makes model's accuracy of 87.69%. In comparison with the results of the first dataset presented in Section 4.3., apart from improvement in terms of type I error rate, the proposed credit scoring shows great improvement also in terms of type II error rate. This is because the second dataset belongs to the institution that exists for a very long time and gathers data more carefully and more precisely, so this dataset is poor with outliers. In contrast to that, the first dataset belongs to the financial institution that uses existing software and data source for only a few years. Some of the used features are considered as important lately, so these are not collected for all records. Apart from that, loan officers are less familiar with the purpose of some information and that probably makes more outliers and noisy data in this dataset.

Based on results presented in Table 8, proposed model shows great performance when applied on much bigger real-life dataset, so proposed model can be considered as generally applicable toward different cases. Also, considering that in real life, for the given problem of credit scoring, it is common to work with big datasets, it can be concluded that proposed model would be very efficient in practice.

Table 10 presents comparison of the other research's results with results of hybrid ensemble credit scoring model proposed in this research. Comparison is made in terms of accuracy, since this is the only measure of model's performance presented in all observed researches.

As mentioned earlier, results presented in Table 10 are not directly comparable with each other, since the models proposed in observed researches are based on different approaches and different datasets.

Table 9
Confusion matrix of DT + GLM ensemble model.

Predicted values				
		GOOD	BAD	TOTAL
Actual Values	GOOD	55,894	7556	63,450
	BAD	1106	5792	6898
		57,000	13,348	70,348

Table 10
Comparison of proposed ensemble model with the similar models in the related researches.

Research	Dataset	Number of records	Number of feature	Accuracy (%)
[1]	German	1000	20	86.47
	Australian	690	14	95.39
[2]	German	100	20	79.00
	Australian	690	14	88.10
	Japanese	690	15	88.70
	Iranian	1000	27	95.80
	Polish	240	30	81.30
	Jordanian	500	12	87.40
	UCSD	2435	38	87.50
[3]	Real life	777	30	91.00
[4]	German	1000	20	81.42
This research	Real life dataset 1	12,983	20	82.01
	Real life dataset 2	70,348	20	87.69

However, considering that model proposed in this research is based and validated on much bigger datasets than those used in other researches, we can conclude that results achieved in this research are quite impressive. Apart from that, unlike researches [1], [2] and [4], this research is based on real-life dataset of financial institution gathered during certain time period. All record from that time period are taken into account except the small number of records without values for the most of the features. It is to be expected that such dataset contains more outliers and noisy data than generic dataset. Considering that fact, accuracy of the model proposed by this research is much more imposing then the higher accuracies of the models based on generic and much smaller datasets. Even though it is based on a real life dataset, research [3] is based on much smaller dataset consisting of only 777 records. This makes the accuracy improvement of this model irrelevant. Apart from that, most of the similar hybrid models proposed by related researches, are based on only one feature selector chosen among several tested feature selection algorithms, as the best one. This research proposes the novel feature selection algorithm based on the results of five different feature selectors. This algorithm outperformed each of these single feature selectors, as well as PCA, as one of the most commonly used feature selection algorithms.

6. Conclusion

The aim of this paper is to present the importance of hybridization and ensemble learning approach in order to improve performance of machine learning algorithms. We proposed a new hybrid ensemble credit scoring model based on fusion of five feature selection algorithms combined by three different types of voting and eight different ensemble models combined by soft voting approach. We also proposed a new fusion method, *if_any,* that outperformed performance of other voting methods in terms of feature selection.

Experimental results showed that the best algorithm in terms of accuracy, type I error rate, sensitivity and F-measure is the one based on five feature selectors combined by *if_any* voting and ensemble GLM + DT. Accuracy of this model is 82.015%, while the single DT classifier's accuracy is 80%. In terms of error type II, single classifier GLM showed the lowest type II error, but accuracy of 73.427%. Since GLM + DT ensemble performed higher accuracy than the single DT classifier, and error type II rate is still pretty high, it is obvious that ensemble GLM + DT is very efficient in recognizing and classifying bad clients correctly. Since, for the given problem of credit scoring, both error type I and error type II rates have significant impact on model performance, then it should be considered within same measure - error rate. Among all other ensembles, model GLM + DT has the lowest error rate, so we can conclude that this ensemble hybrid model is the most accurate and efficient one in comparison with other models being trained in this research. The proposed model is validated by applying much bigger dataset and it showed even higher accuracy. GLM + DT ensemble is confirmed as the ensemble with the highest performance with accuracy of 87.69%. Based on these results and comparison with similar models proposed in the literature, proposed model can be considered as generally applicable among various cases and very efficient in practice.

The only limitation of the model proposed in this research is that it requires certain performance resources in terms of high level of computing power, preferably hardware configuration based on the multicore processor and large volume of RAM memory, to perform in acceptable time period. Since it uses various techniques and methods in order to achieve the highest performance of hybrid ensemble model, running time of this model is driven by resource performance.

Although the approaches used in this research resulted in development of high performance model that is generally applicable for various cases, our future work will be focused on proposing credit scoring model based on lately popular approaches. That implies engagement of some other classifiers, such as neural networks, as well as other voting types, weighted voting for instance. Another recommendation is to involve nature-inspired feature selection algorithms, for example genetic algorithms and particle swarm optimization algorithm, but also algorithms such as ant colony optimization algorithm, bat algorithm, firefly algorithm etc., that are not used for the given problem of credit scoring, according to our knowledge. In addition, since proposed *if_any* voting type outperformed others the most popular voting types, we will try to test *if_any* voting type based on results of nature-inspired FS algorithms.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been supported by the European Regional Development Fund under the Grant KK.01.1.1.01.0009 (DATACROSS).

References

[1] D. Tripathi, D.R. Edla, R. Cheruku, Hybrid Credit Scoring Model Using Neighborhood Rough Set and Multi-Layer Ensemble Classification, J. Intell. Fuzzy Syst. 34 (3) (2018) 1543–1549.
[2] M. Ala'raj, M.F. Abbod, A New Hybrid Ensemble Credit Scoring Model Based on Classifiers Consensus System Approach, Expert Syst. Appl. 64 (2016) 36–55.
[3] F.N. Koutanaei, H. Sajedi, M. Khanbabaei, A Hybrid Data Mining Model of Feature Selection Algorithms and Ensemble Learning Classifiers for Credit Scoring, J. Retailing Consumer Services 27 (2015) 11–23.
[4] A. Ghodselahi, A Hybrid Support Vector Machine Ensemble Model for Credit Scoring, Int. J. Comput. Appl. 17 (5) (2011) 1–5.
[5] W. Zang, P. Zhang, C. Zhou, L. Guo, Comparative Study Between Incremental and Ensemble Learning on Data Streams: Case Study, J. Big Data 1 (1) (2014) 5.
[6] M. Abedini, F. Ahmadzadeh, R. Noorossana, Customer Credit Scoring Using a Hybrid Data Mining Approach, Kybernetes 45 (10) (2016) 1576–1588.
[7] S. Dahiya, S.S. Handa, N.P. Singh, Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set, Industrija 43 (4) (2015) 163–174.
[8] A.G. Armaki, M.F. Fallah, M. Alborzi, A. Mohammadzadeh, A Hybrid Meta-Learner Technique for Credit Scoring of Banks' Customers, Eng. Technol. Appl. Sci. Res. 7 (5) (2017) 2073–2082.
[9] Z. Davoodabadi, A. Moeini, Building Customers Credit Scoring Models with Combination of Feature Selection and Decision Tree Algorithms, Adv. Comput. Sci. Int. J. 4 (2) (2015) 97–103.
[10] M. Khashei, A. Mirahmadi, A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification, Int. J. Financial Studies 3 (3) (2015) 411–422.
[11] S.M. Sadatrasoul, Matrix Sequential Hybrid Credit Scorecard Based on Logistic Regression and Clustering, Iranian J. Manage. Studies 11 (1) (2018) 91–111.
[12] H. Pabuçcu, T.Y. Ayan, The Development of an Alternative Method for the Sovereign Credit Rating System Based on Adaptive Neuro-Fuzzy Inference System, American J. Operat. Res. 7 (01) (2016) 41.
[13] A. Onan, S. Korukoğlu, H. Bulut, A Multiobjective Weighted Voting Ensemble Classifier Based on Differential Evolution Algorithm for Text Sentiment Classification, Expert Syst. Appl. 62 (2016) 1–16.
[14] S. Shukla, R.N. Yadav, Unweighted Class Specific Soft Voting Based Ensemble of Extreme Learning Machine and Its Variant, Int. J. Comput. Sci. Information Security 13 (3) (2015) 59.
[15] J. Nalić, A. Švraka, May. Importance of Data Preprocessing in Credit Scoring Models Based on Data Mining Approaches, IEEE, 2018, pp. 1046–1051.
[16] J. Nalić, A. Švraka, March. Using Data Mining Approaches to Build Credit Scoring Model: Case Study - Implementation of Credit Scoring Model in Microfinance Institution, INFOTEH-JAHORINA (INFOTEH), 2018 17th International Symposium, 2018, pp. 1–5.
[17] F. Nelli, Machine Learning with scikit-learn, Python Data Analytics, Apress, Berkeley, CA, 2015, pp. 237–264.
[18] R. Rimiru, C. Otieno, A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression, Int. J. Comput. (IJC) 27 (1) (2017) 84–102.
[19] S.M. Anaei, M. Moradi, A New Method Based on Clustering and Feature Selection for Credit Scoring of Banking Customers, Int. J. Modern Trends Eng. Res. (SJIF) 3 (2) (2016) 2393–8161.
[20] H. Van Sang, N.H. Nam, N.D. Nhan, A Novel Credit Scoring Prediction Model Based on Feature Selection Approach and Parallel Random Forest, Indian J. Sci. Technol. 9 (20) (2016).
[21] H. Van Sang, N.H. Nam, N.D. Nhan, March. A Hybrid Feature Selection Method for Credit Scoring. EAI Endorsed, Transactions 4 (11) (2017).
[22] L. Yun, Q.Y. Cao, H. Zhang, December. Application of the PSO-SVM Model for Credit Scoring, 2011 IEEE Seventh International Conference on Computational Intelligence and Security, 2011, pp. 47–51.
[23] R.Y. Goh, L.S. Lee, Credit Scoring: A Review on Support Vector Machines and

Metaheuristic Approaches, Adv. Operations Res. (2019).
[24] H. Du, Data Mining Techniques and Applications: An Introduction, Cengage Learning, Boston, 2010, pp. 113–147.
[25] J. Xu, E. Chi, K. Lange, Generalized Linear Model Regression under Distance-to-set Penalties, Adv. Neural Info. Processing Syst. (2017) 1385–1395.
[26] J.J. McArthur, N. Shahbazi, R. Fok, C. Raghubar, B. Bortoluzzi, A. An, Machine Learning and BIM Visualization for Maintenance Issue Classification and Enhanced Data Collection, Adv. Eng. Inf. 38 (2018) 101–112.
[27] S.Y. Sohn, D.H. Kim, J.H. Yoon, Technology Credit Scoring Model with Fuzzy Logistic Regression, Appl. Soft Comput. 43 (2016) 150–158.
[28] Y.L. Eddy, E.M.N.E.A. Bakar, Akademia Baru, J. Adv. Res. Business Manage. Studies 7 (2) (2017) 29–41.
[29] A. Samreen, F.B. Zaidi, A. Sarwar, Design and Development of Credit Scoring Model for the Commercial Banks in Pakistan: Forecasting Creditworthiness of Corporate Borrowers, Int. J. Business Commerce 2 (5) (2013) 1–26.