Advances in Intelligent Systems and Computing 1353

Marcin Paprzycki · Sabu M. Thampi · Sushmita Mitra · Ljiljana Trajkovic · El-Sayed M. El-Alfy *Editors*

Intelligent Systems, Technologies and Applications

Proceedings of Sixth ISTA 2020, India

Springer

Advances in Intelligent Systems and Computing

Volume 1353

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within "Advances in Intelligent Systems and Computing" are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Indexed by DBLP, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST).
All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/11156

Marcin Paprzycki · Sabu M. Thampi ·
Sushmita Mitra · Ljiljana Trajkovic ·
El-Sayed M. El-Alfy
Editors

Intelligent Systems, Technologies and Applications

Proceedings of Sixth ISTA 2020, India

Springer

Editors
Marcin Paprzycki
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland

Sushmita Mitra
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, West Bengal, India

El-Sayed M. El-Alfy
College of Computer Sciences
and Engineering
King Fahd University of Petroleum
and Minerals
Dhahran, Saudi Arabia

Sabu M. Thampi
Indian Institute of Information Technology
and Management Kerala
Trivandrum, Kerala, India

Ljiljana Trajkovic
Simon Fraser University
Burnaby
BC, Canada

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-16-0729-5 ISBN 978-981-16-0730-1 (eBook)
https://doi.org/10.1007/978-981-16-0730-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Conference Organization

Chief Patron

G. Viswanathan, Chancellor, Vellore Institute of Technology

Patrons

Sankar Viswanathan, Vice-President, Vellore Institute of Technology
Sekar Viswanathan, Vice-President, Vellore Institute of Technology
G. V. Selvam, Vice-President, Vellore Institute of Technology
Sandhya Pentareddy, Executive Director, Vellore Institute of Technology
Kadhambari S. Viswanathan, Assistant Vice-President, Vellore Institute of Technology
Rambabu Kodali, Vice Chancellor, Vellore Institute of Technology
S. Narayanan, Pro-Vice Chancellor, Vellore Institute of Technology, Vellore
V. S. Kanchana Bhaaskaran, Pro-Vice Chancellor, Vellore Institute of Technology
P. K. Manoharan, Additional Registrar, Vellore Institute of Technology

General Chairs

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Poland
Marcin Paprzycki, Systems Research Institute, Polish Academy of Sciences, Poland
Sushmita Mitra, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

General Executive Chair

Sabu M. Thampi, Indian Institute of Information Technology and Management-Kerala (IIITM-K), Trivandrum

Program Chairs

Juan Manuel Corchado Rodriguez, University of Salamanca, Spain
El-Sayed M. El-Alfy, King Fahd University of Petroleum and Minerals, Saudi Arabia
Ljiljana Trajkovic, Simon Fraser University, Canada

Organizing Chairs

S. Geetha, VIT, Chennai
R. Jagadeesh Kannan, VIT, Chennai

Organizing Secretaries

C. Sweetlin Hemalatha, VIT, Chennai
G. Suganya, VIT, Chennai
R. Kumar, VIT, Chennai

Organizing Co-chairs

S. Asha, VIT, Chennai
R. Pattabiraman, VIT, Chennai
V. Viswanathan, VIT, Chennai

TPC Members

http://www.acn-conference.org/2020/ista2020/committee.html

Organized by

Vellore Institute of Technology (VIT), Chennai, India

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Preface

This volume of proceedings offers to readers a selection of refereed papers that were presented at the Sixth International Symposium on Intelligent Systems Technologies and Applications (ISTA'20). ISTA'20 brought together researchers in related fields to explore and discuss various aspects of intelligent systems technologies and their applications. The symposium was organized by Vellore Institute of Technology (VIT), Chennai, India, during October 14–17, 2020. Due to the recent pandemic situation, it was conducted as a virtual event. ISTA'20 was co-located with the International Conference on Applied Soft Computing and Communication Networks (ACN'20).

All submissions were evaluated on the basis of their significance, novelty, and technical quality. A double-blind review process was conducted to ensure that the names and affiliations of authors were unknown to TPC. This volume consists of 28 papers (19 regular and 9 short papers) that were virtually presented at the symposium. The papers cover different areas such as big data analytics, security and privacy, Internet of Things, machine and deep learning, health informatics, visual computing, signal processing, and natural language processing.

We are very grateful to the many people who contributed to the organization of the symposium. Our heartfelt thanks go to all the authors for submitting their research results to the symposium and to the members of the program committee for their careful and insightful reviews. We are grateful to Vellore Institute of Technology (VIT), Chennai, for organizing the symposium. It could not have taken place without the commitment of the local organizing committee, which in many ways helped to effectively coordinate the event. We are grateful to General Chairs and Program Chairs for their great support. We express our most sincere thanks to all keynote speakers who have shared their expertise and knowledge with us. Finally, we would

like to thank our publisher Springer, in particular Senior Editor Aninda Bose, for their collaboration.

Warsaw, Poland — Marcin Paprzycki
Trivandrum, India — Sabu M. Thampi
Kolkata, India — Sushmita Mitra
Burnaby, Canada — Ljiljana Trajkovic
Dhahran, Saudi Arabia — El-Sayed M. El-Alfy
October 2020

Contents

About the Editors

Marcin Paprzycki is Associate Professor in the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. He has an M.S. from Adam Mickiewicz University in Poznań, Poland, a Ph.D. from Southern Methodist University in Dallas, Texas, and a Doctor of Science from the Bulgarian Academy of Sciences. He is a senior member of IEEE, a senior member of ACM, Senior Fulbright Lecturer, and IEEE CS Distinguished Visitor. He has contributed to more than 500 publications and was invited to the program committees of over 800 international conferences. He is on the editorial boards of 12 journals and a book series.

Sabu M. Thampi is Professor at the Indian Institute of Information Technology and Management-Kerala (IIITM-K), Technopark Campus, Trivandrum, Kerala, India. His current research interests include cognitive computing, Internet of things (IoT), authorship analysis, trust management, biometrics, social networks, nature-inspired computing, and video surveillance. He has published papers in book chapters, journals, and conference proceedings. He has authored and edited a few books. Sabu has served as Guest Editor for special issues in few journals and a program committee member for many international conferences and workshops. He has co-chaired several international workshops and conferences. He has initiated and is also involved in the organization of several annual conferences/symposiums. Sabu is currently serving as Editor for *Journal of Network and Computer Applications* (JNCA), Elsevier; *Connection Science*, Taylor & Francis; Associate Editor for *IEEE Access* and *International Journal of Embedded Systems*, Inderscience, UK, and Reviewer for several reputed international journals. Sabu is a senior member of IEEE and ACM. He is Founding Chair of ACM Trivandrum Professional Chapter.

Sushmita Mitra is full Professor at the Machine Intelligence Unit (MIU), Indian Statistical Institute, Kolkata. From 1992 to 1994 she was in the RWTH, Aachen, Germany as a DAAD Fellow. She was a Visiting Professor in the Computer Science Departments of the University of Alberta, Edmonton, Canada in 2004, 2007; Meiji University, Japan in 1999, 2004, 2005, 2007; and Aalborg University Esbjerg, Denmark in 2002, 2003. Dr. Mitra received the National Talent Search Scholarship (1978–1983) from NCERT, India, the University Gold Medal in 1988, the IEEE

TNN Outstanding Paper Award in 1994 for her pioneering work in neuro-fuzzy computing, the CIMPA-INRIA-UNESCO Fellowship in 1996, and Fulbright-Nehru Senior Research Fellowship in 2018–2020. She was the INAE Chair Professor during 2018–2020. Dr. Mitra is the author of the books *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing* and *Data Mining: Multimedia, Soft Computing, and Bioinformatics* published by John Wiley, and *Introduction to Machine Learning and Bioinformatics*, Chapman & Hall/CRC Press, beside a host of other edited books. Dr. Mitra has guestedited special issues of several journals, is an Associate Editor of *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *Information Sciences*, *Fundamenta Informatica*, and is a Founding Associate Editor of *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (WIRE DMKD). She has more than 150 research publications in refereed international journals. Dr. Mitra is a Fellow of the IEEE, Indian National Science Academy (INSA), International Association for Pattern Recognition (IAPR), and Fellow of the Indian National Academy of Engineering (INAE) and The National Academy of Sciences, India (NASI). She is an IEEE CIS Distinguished Lecturer and the current Chair-Elect, IEEE Kolkata Section. She has visited more than 30 countries as a Plenary/Invited Speaker or an academic visitor, and served as General Chair and/or Program Chair of several international conferences. Her current research interests include data mining, pattern recognition, soft computing, medical image processing, and Bioinformatics.

Ljiljana Trajkovic received the Dipl. Ing. degree from University of Pristina, Yugoslavia, in 1974, the M.Sc. degrees in electrical engineering and computer engineering from Syracuse University, Syracuse, NY, in 1979 and 1981, respectively, and the Ph.D. degree in electrical engineering from University of California at Los Angeles, in 1986.

She is currently a Professor in the School of Engineering Science at Simon Fraser University, Burnaby, British Columbia, Canada. From 1995 to 1997, she was a National Science Foundation (NSF) Visiting Professor in the Electrical Engineering and Computer Sciences Department, University of California, Berkeley. She was a Research Scientist at Bell Communications Research, Morristown, NJ, from 1990 to 1997, and a Member of the Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ, from 1988 to 1990. Her research interests include high-performance communication networks, control of communication systems, computer-aided circuit analysis and design, and theory of nonlinear circuits and dynamical systems.

Dr. Trajkovic serves as IEEE Division X Delegate/Director (2019–2020) and served as IEEE Division X Delegate-Elect/Director-Elect (2018). She served as Senior Past President (2018–2019), Junior Past President (2016–2017), President (2014–2015), President-Elect (2013), Vice President Publications (2012–2013, 2010–2011), Vice President Long-Range Planning and Finance (2008–2009), and a Member at Large of the Board of Governors (2004–2006) of the IEEE Systems, Man, and Cybernetics Society. She served as 2007 President of the IEEE Circuits and Systems Society and a member of its Board of Governors (2004–2005, 2001–2003). She is Chair of the IEEE Circuits and Systems Society joint Chapter of the Vancouver/Victoria Sections. She was Chair of the IEEE Technical Committee on

Nonlinear Circuits and Systems (1998). She is General Co-chair of SMC 2020 and SMC 2020 Workshop on BMI Systems and served as General Co-chair of SMC 2019 and SMC 2018 Workshops on BMI Systems, SMC 2016, and HPSR 2014, Special Sessions Co-chair of SMC 2017, Technical Program Chair of SMC 2017 and SMC 2016 Workshops on BMI Systems, Technical Program Co-chair of ISCAS 2005, and Technical Program Chair and Vice General Co-Chair of ISCAS 2004. She served as an Associate Editor of the *IEEE Transactions on Circuits and Systems (Part I)* (2004–2005, 1993–1995), the *IEEE Transactions on Circuits and Systems (Part II)* (2018, 2002–2003, 1999–2001), and the *IEEE Circuits and Systems Magazine* (2001–2003). She is a Distinguished Lecturer of the *IEEE Systems, Man, and Cybernetics Society* (2020–2021) and the *IEEE Circuits and Systems Society* (2010–2011, 2002–2003). She is a Professional Member of IEEE-HKN and a Life Fellow of the IEEE.

El-Sayed M. El-Alfy is a Professor in Information and Computer Science Department, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He has more than 25 years of experience in industry and academia involving research, teaching, supervision, curriculum design, program assessment and quality assurance in higher education. He is approved ABET/CSAB Program Evaluator (PEV), IEEE Senior Member, and reviewer and consultant for NCAAA and several universities and research agents in various countries. He is an active researcher with interests in fields related to machine learning and nature-inspired computing and applications to data science and cybersecurity analytics, pattern recognition, multimedia forensics and security systems. He has published numerously in peer-reviewed international journals and conferences, edited a number of books published by reputable international publishers, attended and contributed in the organization of many world-class international conferences, supervised master and Ph.D. students, served as a guest editor for a number of special issues in international journals, and been in the editorial board of a number of premium international journals including *IEEE/CAA Journal of Automatica Sinica*, *IEEE Transactions on Neural Networks and Learning Systems*, *International Journal on Trust Management in Computing and Communications*, and *Journal of Emerging Technologies in Web Intelligence* (JETWI). He is also a member of *IEEE Computational Intelligence Society*, and ex-member of *ACM and IEEE Computer Society*. His work has been internationally recognized and received a number of awards.

Probability of Loan Default—Applying Data Analytics to Financial Credit Risk Prediction

Aleksandra Łuczak, Maria Ganzha, and Marcin Paprzycki

Abstract In the banking industry, one of the important issues is how to establish credit worthiness of potential clients. With the possibility of collecting digital records of results of past credit applications (of all clients), it can be stipulated that machine learning techniques can be used in "credit decision support" systems. There exists a substantial body of literature devoted to this subject. Moreover, benchmark datasets have been proposed, to establish effectiveness of proposed credit risk assessment approaches. The aim of this work is to compare performance of *seven* different classifiers, applied to *two* different benchmark datasets. Moreover, capabilities of, recently introduced, methods for combining results from multiple classifiers, into a meta-classifier, will be evaluated.

Keywords Meta-classification · Neural network · Random forest · XGBoost · Adaboost · K-nearest neighbours · Support vector machine · Parameter optimization · Credit risk · Credit risk prediction · Loan default · Data analytics · Machine learning · Performance comparison

1 Introduction

Banking industry of today collects "all possible" data concerning all of its customers. Moreover, for all practical purposes, such data is never discarded (even if clients close their accounts [1]). One of the areas, where collected data is expected to be of great value, is to provide support for the so-called credit risk assessment. Specifically, it is assumed that application of data analytics methods, to the collected data, can help correctly assess probability of loan defaults. This belief can be observed, among others,

A. Łuczak · M. Ganzha
Warsaw University of Technology, Warsaw, Poland
e-mail: M.Ganzha@mini.pw.edu.pl

M. Paprzycki (✉)
Systems Research Institute Polish Academy of Sciences, Warsaw, Poland
e-mail: marcin.paprzycki@ibspan.waw.pl

M. Paprzycki et al. (eds.), *Intelligent Systems, Technologies and Applications*,
Advances in Intelligent Systems and Computing 1353,
https://doi.org/10.1007/978-981-16-0730-1_1

in a large body of literature devoted to the subject (see, Sect. 3). However, existing work is, mostly, focused on specific classifiers, with the goal of improving their *individual* performance. The aim of our work is to *compare* performance of different credit risk assessment methods, when applied to two standard "benchmark datasets". Moreover, we will investigate usefulness of a, recently introduced, meta-classifiers, which "merge" predictions of individual classifiers to provide a "combined score".

To this effect, we proceed as follows. We start from a brief summary of necessary background knowledge related to the banking industry (Sect. 2). Next, in Sect. 3, we present an overview of related literature. We follow, in Sects. 4 and 5, with an overview of the two datasets used in our work and the experimental setup. This allows us, in Sect. 6, to summarize results of performed experiments. We conclude our contribution in Sect. 7.

2 How Banks Assess Customers—Short Introduction

In all banks, there exist complicated procedures, used to comprehensively evaluate credit worthiness of their (potential) customers. Moreover, each bank has it own method to do so, and such methods are a closely guarded intellectual property. While one of the authors worked in a credit department of a large bank, for obvious reasons, we will be able to share only "common open knowledge/information" about the process of assessment of probability of loan default. Keeping this in mind, let us start from basics of credit risk assessment.

2.1 Need of Credit Risk Assessment

One of the most sensitive stages, of a bank processing a credit application, is the decision-making process. Following the digital transformation, banks introduced online lending and borrowing mechanisms. Here, the key element is a semi-automatic credit decision process, which is based, among others, on the demographic and the financial data. Here, the *semi-automatic process* means, usually that automatic positive decisions are accepted, while possible "problems" are remanded to human employees for final decisions.

The possibility of applying for loans online is one of the reasons that banks are granting more and more credits. Here, note that, at the same time, the total number of physical bank branches is systematically decreasing (as predicted, for instance, in [2]). The fact that the increase in number of credits is a global phenomenon is confirmed, among others, by the Polish Financial Supervision Authority [3]. According to the available data, at the end of December 2019, household debt, to the banks, amounted to 671 billion PLN. In addition, as much as 15.44 million, or 40.6% of Poles were actively repaying a loan, or a credit.

With the increasing prevalence of loans, and the decreasing requirements of banks (which earn money on the credits), there is also a recognized risk of approving loans to persons who cannot cope with their repayment, hence the need to automatically (as the total number of bank employees is decreasing) evaluate credit risk of an increasing number of loan applicants.

In this context, the *Probability of Default* (PD) indicates the probability of a loss, associated with a given credit. It is, usually, considered within a certain time-horizon (usually one year). Simply said, high value of PD indicates that customer may stop paying back the loan (temporarily, or permanently). The probability of customer's load default is the primary parameter in calculating creditworthiness. This factor plays a major role in loss control and income maximization [4]. PD can be estimated with the use of a number of analytical tools, and majority of such tools are based on some form of, broadly understood, "data analytics". In practice, in majority of banks, linear discriminate analysis [5] or logistic regression [6] is applied, because these algorithms combine simplicity and efficiency and are easy to explain (high interpretability) [7]. The latter makes then also slightly more desirable than neural network-based approaches, which are not explainable and thus raise variety of ethical issues.

Separately, in scientific literature, a number of approaches have been studied. Moreover, to support this research, at least two open, tagged, datasets have been created, which can be used to compare performance of considered methods.

3 Related Work

This being the case, let us now briefly present state of the art of analytic credit risk assessment, as represented in scientific publications. Since, as mentioned above, each bank has its own credit risk assessment procedures, which are a closely guarded intellectual property, only published literature can be discussed. We split the material into two parts. First, we summarize the proposed approaches, dataset(s) that they were applied to, and the reported quality of the loan default prediction. The latter will be used to compare results of our experiments to. Second, for the standard methods of data analytics, described in the reviewed literature, we provide short descriptions and basic references.

3.1 Data Analytics for Credit Risk Analysis

In the credit risk assessment, the most common method used to predict the PD is a scoring card [8]. It consists of assessing (weighting) selected characteristics of the potential borrower. After assigning a score to each one of them, these ratings are combined (summed up). Here, if the total score is higher than the (financial institution dependent) cut-off point, client receives a credit [9]. This is a very simple method

and is very easy to operationalize, but has many disadvantages. For example, it has a relatively weak predictive power and cannot handle large datasets and features with complicated relationships [10].

Nowadays, more focus is devoted to the predictive power of classification models, used to control credit risk, because even a small increase in performance results in a large increase in profits [11, 12]. Therefore, with the development of machine learning methods, the improvement of the predictive power of scoring models is within reach of financial institutions, and they are taking up this approach [13–15]. In most cases, algorithms from the supervised ML family are used in the credit risk assessment, to find a correlation between the clients attributes, and their expected loan repayment. Here, note that in many articles, proposed approaches are confirmed by high accuracy [16–18].

Recently, many researchers have focused their attention on ensemble methods, and use of different base classifiers, to make the accuracy of prediction "as high as possible". Such heterogeneous classifiers can, for instance, ensemble three state-of-the art classifiers: logistic regression (LR), artificial neural network (ANN), and support vector machines (SVM) [19]. Hybrid ensemble methods were also based on Random Forest, Support Vector Machines, Decision Trees, Artificial Neural Network, Multidimensional Neural Network [20]. The method based on a hybrid associative memory with translation was reported in [21]. There exists a study, from 2015, that compares 41 different classification algorithms [22] and a study that compares Bagging DT, Random Subspace DT, Random Forest and Rotation Forest [23]. All this research shows good model performance, as measured in terms of "standard" metrics such as accuracy (ACC) and area under the ROC curve (AUC), which are above 0.7.

Nevertheless, in this contribution, we will compare obtained results to those reported in [24, 25]. This is related to the fact that they use similar datasets. Moreover, they follow the same general methodology (see, Sect. 5). Nevertheless, we recognize the fact that a more broad comparative study (similar to that reported in [22], but aligned with our investigation) could be of value.

3.2 Classifiers Used in Our Work

Let us now summarize the specific machine learning approaches that have been used in the context of credit risk assessment.

3.2.1 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a classifier, which is very often used in pattern recognition [26], and to build credit scoring models [27]. It is a simple algorithm, based on adding new elements to the class that they are "best fitting", on the basis of a process of comparing it to the nearest set of observations (k-nearest neighbours).

The typical distance function is an Euclidean distance. In our work, we used the KNN modules available in *sklearn.neighbors.KNeighborsClassifier*.[1]

3.2.2 Support Vector Machines

Support vector machines (SVMs) is an ML method proposed by Vapnik [28]. It is based on the search for a hyperplanes separating classes. Established hyperplanes should be such that the closest observations from each class are as far away from each other as possible. In our work, we have experimented with the SVM in two versions: with (a) linear and (b) radial kernel functions. Here, we used *sklearn.svm.LinearSVC*[2] and *sklearn.svm.SVC*[3] modules.

3.2.3 Random Forest (RF)

The Random Forest (RF) algorithm was proposed by Breiman [29, 30]. It is based on the decision tree, and the classification committee. This means that the final response of the algorithm is selected as the average response of each decision tree (seen as an "independent classifier"). For the random forest, we used *sklearn.ensemble.RandomForestClassifier*[4] modules.

3.2.4 Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) is a generalization of boosting [31] to an arbitrary differential loss functions [32]. GBDT is an accurate and effective off-the-shelf procedure that can be used for both the regression and the classification problems. It can be applied in a variety of areas. To experiment with the GBDT, we used *sklearn.ensemble.GradientBoostingClassifier*[5] modules.

3.2.5 AdaBoost

An AdaBoost is an adaptive boosting algorithm, proposed by Freund and Schapire [33]. This is a meta-estimator that begins by fitting a classifier into the original

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html?highlight=svm#sklearn.svm.LinearSVC.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=random%20forest#sklearn.ensemble.RandomForestClassifier.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier.

dataset and then fits additional copies of the classifier on the same dataset. However, in this case, the weights of incorrectly classified instances are adjusted, such that the subsequent classifiers focus more on difficult cases. In case of AdaBoost, we used *sklearn.ensemble.AdaBoostClassifier*[6] modules.

3.2.6 Extreme Gradient Boosting

The Extreme Gradient Boosting (XGBoost) algorithm is one of the most popular, and most effectively implemented algorithms, in the family of algorithms based on the gradient boosting method [34]. For the XGBoost, we used *xgboost.XGBClassifier*[7] modules.

3.2.7 Meta-Classifiers

While the above listed approaches can be (and have been) applied individually to credit risk assessment (see, Sect. 3.1), recently it has been realized that each one of them captures "separate aspects" of the data. The question has thus been posed: Is it possible to combine results from multiple classifiers to develop a meta-classifier, which will improve the overall quality of prediction? Such meta-classifiers have been tried, with moderate success, in non-financial domains (see, for instance, [35, 36]). However, they involve relatively complex methods that require a lot of attention to be correctly implemented.

Searching in the same direction, recently, a different class of simple meta-classifiers has been proposed. Specifically, Weighted Average Recall Error (WARE) and Weighted Average Type-1 Error (WA-T1E) algorithms are based on an easy to implement consensus models. These classifiers are similar to the Weighted Average Prediction Error (WA-PE) method used in [37]. Here, the data sample is divided into two parts, one for training the classifiers and one to calculate the Recall Recall_m, and the Type I error $T1err_m$ (where m is the index of the classifier). After normalizing the weights, for the Recall_m and the $T1err_m$, posteriori probabilistic classes of observations are computed as follows:

$$\text{WARE}_1(x) = \sum_{m=1}^{M} \text{Recall}_m \, p_1^{h_m}(x) \tag{1}$$

$$\text{WA} - T1E_1(x) = \sum_{m=1}^{M} T1err_m \, p_1^{h_m}(x) \tag{2}$$

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html?highlight=adaboost#sklearn.ensemble.AdaBoostClassifier.

[7] https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn.

where p_1 is the posteriori probability of belonging to class 1 and h_m is a classifier. Here, observation x is assigned to class 1 if $\text{WARE}_1(x) > \text{WARE}_0(x)$, or to class 0 otherwise. The same applies to the WA-T1E method, where the observation x is assigned to class 1 if $WA - T1E_1(x) > WA - T1E_0(x)$, or to class 0 otherwise.

Therefore, taking into account their simplicity, we have decided to experiment also with these two meta-classifiers, to see if their application can improve the overall performance of the credit risk estimators.

4 Datasets Used in Experiments

Let us now move our attention to the data used in our experiments. As mentioned above, there exist multiple tagged datasets that can be used to evaluate performance of approaches to credit risk prediction. In our work, we have used two of them: (a) German Credit,[8] and (b) Give Me Some Credit.[9] Both datasets contain information about the demographic characteristics and the financial situation, of potential borrowers. Let us now describe each one of them in more detail.

4.1 German Credit Dataset

The German redit data contains 1,000,000 observations, representing two classes: good creditor (699,774 observations) and bad creditor (300,226 observations). The dataset is described by 20 features, including 7 continuous features.

- *duration*—Duration in month
- *credit_amount*—Credit amount
- *installment_commitment*—Installment rate in percentage of disposable income
- *residence_since*—Present residence since
- *age*—Age in years
- *existing_credits*—Number of existing credits at this bank
- *num_dependents*—Number of people being liable to provide maintenance for

and 13 categorical features

- *checking_status*—Status of existing checking account
- *credit_history*—Credit history
- *purpose*—Purpose of loan
- *savings_status*—Status of savings account/bonds
- *employment*—Present employment since
- *personal_status*—Personal status and sex

[8] https://www.openml.org/d/260.

[9] https://www.kaggle.com/c/GiveMeSomeCredit/data?select=cs-training.csv.

Table 1 Information about missing values in categorical features of the German credit dataset

Feature name	Number of missing values	Representations of missing values	Ratio of missing values (%)
checking_status	394,051	'no checking'	39.40
credit_history	40,856	'no credits/all paid'	4.08
savings_status	184,016	'no known savings'	18.40
other_parties	905,898	none	90.58
property_magnitude	151,898	'no known property'	15.18
other_payment_plans	808,505	none	80.85
own_telephone	594,649	none	59.46

- *other_parties*—Other debtors/guarantors
- *property_magnitude*—Magnitude of personal property
- *other_payment_plans*—Other installment plans
- *housing*—Housing (own or rent)
- *job*—Status of present job
- *own_telephone*—Telephone
- *foreign_worker*—Foreign worker.

The data has been curated, resulting in "no data missing". However, some categorical variables contain information that they do not represent "real data". For example, in some cases, the variable *saving_status* contains the value *no known savings*, indicating that the actual data is missing. All such cases are summarized in Table 1.

To apply machine learning, all categorical variables have been manually converted, from string variables, to numerical variables. In addition, strings listed in Table 1 were transformed into label 0. Nevertheless, no data was removed from the dataset. This decision may be questioned, and we plan to investigate this aspect further in the future.

4.2 Give Me Some Credit Dataset

This dataset is smaller and uses 10 features, to describe 150,000 borrowers. On the basis of the data provided, it should be anticipated whether the person described there will experience financial difficulties, in the next two years. This feature is tagged as *SeriousDlqin2yrs*. This information allows one to determine the credibility of the borrower and can be used to make a decision about approving loan application. Below, we present features available in the dataset, with their description.

- *Age*—Age in years
- *NumberOfDependents*—Number of the borrower's dependants
- *MonthlyIncome*—Amount of monthly income

- *DebtRatio*—The ratio of the sum of monthly debt. alimony and maintenance costs to monthly income expressed as a percentage
- *RevolvingUtilizationOfUnsecuredLines*—The ratio of the total balance on cards and personal lines of credit to the sum of credit limits expressed as a percentage
- *NumberOfOpenCreditLinesAndLoans*—Number of loans and open credit lines
- *NumberRealEstateLoansOrLines*—Number of housing loans and credits and household credit lines
- *NumberOfTime30–59DaysPastDueNotWorse*—Number of cases where the borrower was 30–59 days late in making payments in the last 2 years
- *NumberOfTime60-89DaysPastDueNotWorse*—Number of cases where the borrower was 60–89 days late in making payments in the last 2 years
- *NumberOfTimes90DaysLate*—Number of cases where the borrower was above 90 days late in making payments in the last 2 years

Data available in this dataset is incomplete. Specifically, there are two features that have considerable data gaps. For the *MonthlyIncome* attribute, data is missing for 29,731 observations, while for the *NumberOfDependents* attribute, data is missing for 3924 observations.

However, most of the tested algorithms (mentioned in Sect. 3.1) require that the dataset should *not* contain any missing data. To avoid "simple imputing", which means: to fill in the missing values with 0 or -1, we used the KNNImputer algorithm from the sklearn package. This algorithm calculates the k-th nearest neighbours of a given observation, based on the Euclidean distance. Next, it selects the value of the feature to be completed from the k-neighbors. This data augmentation step was applied after the division of the dataset into the teaching and the testing parts.

5 Experimental Setup

Let us now summarize the key technical aspects of the performed experiments.

- We have separately experimented (including applied preprocessing), with the two datasets described in Sect. 4.
- In each case, we have applied all classifiers, described in Sect. 3.
- Only open-source classifiers were used (as specified in Sect. 3.1).
- In each case, available data was randomly divided into the standard 7:3 ratio, between the training and testing subsets.
- During the training, the models were fitted, and their hyperparameters were optimized using special software, described in what follows.
- During the testing, all metrics described in Sect. 5.1 were calculated.
- Meta-classifiers were used to combine "suggestions" of the individual classifiers.

It should be noted that all algorithms were automatically optimized. For the scientific context of tuning of hyperparameters of classifiers, see [38]. Overall, the main goal of optimization was to minimize the function:

$$\text{loss}(f(X)) = 1 - \text{AUC}(f(X), y). \tag{3}$$

where X is the feature, y is the target, and f is the decision function. To achieve the needed optimization, the *hyperopt*[10] package has been used. Additionally, all algorithms had a parameter set to unbalanced classes (matching the characteristics of the datasets), to further improve the performance.

5.1 Performance Evaluation—Methodological Considerations

In the works summarized in Sect. 3, to assess performance of the proposed approaches, the following standard performance metrics have been used: accuracy (ACC), Matthews correlation coefficient (MCC), the area under the ROC curve (AUC), Recall, Precision, Type I error, and Type II error. The formulae, representing these metrics, have been summarized in the following equations:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Type I error} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{8}$$

$$\text{Type II error} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{9}$$

All variables, found in equations above, are calculated on the basis of the error matrix, represented in Table 2.

Table 2 Confusion matrix used in this study

Predicted values			
		Positive[a]	Negative[b]
True values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

[a]Positive (True) means bad debtors
[b]Negative (False) means good debtors

[10]https://github.com/hyperopt/hyperopt.

It is important to make the following meta-level observation. Regardless of the academic discussions concerning value of each performance metrics (e.g., the famous "tug-of-war" between the Precision and the Recall), our approach is rooted in the real-world practices, followed by the banking industry. As noted, one of authors has worked in a large bank and, therefore, we can reasonably claim that from the bank perspective it are the Recall and the Type I error that are the most important. In other words, bank's focus is on making sure that bad loans will not happen. Hence, these two metrics are crucial from the point of view of credit risk, as they indicate how much the financial institution could potentially lose, by giving funding to a person who will stop paying the debt. Type I error indicates the ratio of how many people the model predicted as good[11] and they turned out to be bad[12] to the total number of all "good people". Recall, on the other hand, indicates the fraction of how many people the model correct predicted as bad, to all bad creditors. In an ideal world, the financial institution would like the Recall to be close to 1. This fact provided us with the guideline, which results should be considered "the best". Therefore, in our work, *we have focused on obtaining the best Recall score*.

This fact has two important consequences. (1) Since our goal was to achieve the best Recall score, it could have happened that the fact that our results are better than that reported in the literature/obtained using other classifiers can be seen as slightly unfair (if they were focused on improving other performance metrics). (2) At the same time, our results are *are real-world grounded* as the performance metrics of our choice are the ones that are actually pursued by the banks (the Recall, in particular). Obviously, in this way our approach differs from the academic research found in the literature.

6 Experimental Results

Let us now summarize the obtained results (keeping in mind the methodological considerations described in Sect. 5). For performance comparison, the results from [24, 25] were used as a "baseline". Let us now describe individual results for the two datasets.

6.1 Results for the German Credit Data

In Table 3, we summarize the results obtained for the German Credit dataset. Moreover, as a comparison, first, we show results from [24, 25], where its author used the same classifiers, but did not optimize the hyperparameters. Second, we report results from [25], where authors, firstly, used unsupervised algorithms to cluster

[11] good—debtors not delayed in repayment.

[12] bad—debtors delayed in repayment.

Table 3 Results for the German credit data

	AUC	ACC	MCC	Precision	Type I error	Type II error	Recall	Recall [24]	Recall [25]
XGBoost	**0.890**	**0.832**	**0.587**	**0.761**	0.143	0.239	0.645	0.393	–
Adaboost	0.779	0.831	0.584	0.754	0.091	0.351	0.649	0.44	–
GBDT	0.779	0.830	0.583	0.752	0.092	0.350	0.650	0.393	0.418
Random Forest	0.648	0.750	0.350	0.639	0.095	0.610	0.390	0.343	0.754
SVM_rbf	0.530	0.672	0.082	0.398	0.112	0.828	0.172	0.097	–
LinearSVM	0.516	0.328	0.081	0.308	0.957	**0.011**	**0.989**	0.477	0.459
KNN	0.548	0.695	0.139	0.481	**0.083**	0.821	0.179	0.373	0.443

observations, and then they build supervised individual models. We believe that both comparison fairly illustrate the comparative quality of our results.

The first conclusion is that there is no single model that has the highest score across all results. However, the model that "leads the way" is the XGBoost (highest AUC, ACC, MCC and Precision). However, as mentioned above, we focus on the Recall value. Hence, the "winner" seems to be the LinearSVM algorithm that has the highest score (0.989). Unfortunately, it also has the highest Type I error (0.957). This means that, in the banking practice, this approach should be avoided (as discussed in Sect. 5). Therefore, we cannot say that this is the best algorithm.

This being the case, the AdaBoost and the GBDT are the "best overall algorithms", to apply to our problem, from the bank's point of view. The Recall score is high (0.65), and they have relatively low Type I error (0.09; although, again, not the lowest), so these are the algorithms that banks may want to be most interested in.

It is also worth mentioning that, compared to [24], algorithms XGBoost, AdaBoost, GBDT, LinearSVM, SVM_rbf and RF have a higher Recall score. Finally, compared to [25], GBDT and LinearSVM also have a higher Recall score. However, the Random Forest and the KNN algorithm have worse scores. This may be due to differences in data preparation for modeling.

6.2 Results for the Give Me Some Credit Dataset

In order to verify the effectiveness and stability of our strategy, we have also applied the same methodology to the Give Me Some Credit dataset. Comparison between our classifiers and these from [24] is shown in Table 4. Recall, that there, authors build the same classifiers, but did not optimize hyperparameters.

Note that, here, the XGBoost algorithm did not have the highest values of the four performance metrics, as in above Section. In this case, only the MCC metric (0.348) had the highest value, but this value is much lower than the satisfactory one (above 0.5). The best algorithm in terms of the number of "best" values of metrics was

Table 4 Results for the give me some credit data

	AUC	ACC	MCC	Precision	Type I error	Type II error	Recall	Recall [24]
XGBoost	0.869	0.81	**0.348**	0.228	0.02	0.772	0.769	0.17
Adaboost	**0.873**	0.801	0.345	0.222	0.019	0.778	0.784	0.178
GBDT	0.872	0.792	0.341	0.215	**0.018**	0.785	**0.799**	0.17
Random Forest	0.867	0.778	0.326	0.204	**0.018**	0.796	0.796	0.152
SVM_rbf	0.656	0.643	0.12	0.106	0.044	0.894	0.584	0.0
LinearSVM	0.513	**0.934**	0.12	**0.621**	0.065	**0.379**	0.027	0.006
KNN	0.54	0.931	0.055	0.265	0.066	0.735	0.019	0.013

the SVM with the linear kernel, because it had the highest ACC (0.934), Precision (0.621) and the lowest Type II error (0.379).

However, note that the most important metric for us is Recall. Here, the algorithm that had the highest Recall score (0.799) was the GBDT. It also had the second highest AUC (0.872), and the lowest Type I error (0.018, including Random Forest), which (as noted above) is the second equally important metric in credit risk control. Note also that AdaBoost had the highest AUC (0.873). The worst results were achieved by the SVM_rbf and the KNN models.

Compared to the [24], all algorithms calculated with our methodology achieve higher Recall scores, especially in boosting and bagging models. For instance, for XGBoost, AdaBoost, GBDT and Random Forest, the difference is more than 50% percentage points. This is an improvement in the quality of the classifiers of over 300%. This may be due to the fact that the data was very unbalanced and our algorithms took this fact into account.

6.3 Results for the Meta-Classifiers

Let us now discuss the results for the two consensus models: the **Weighted Average Recall Error (WARE)** and the **Weighted Average Type-1 Error (WA-T1E)**, described in Sect. 3.1. Both were applied to the two datasets. They were tested because they had promising results reported in [37]. In Table 5, results for the German Credit dataset are presented. Next, in the Table 5, results for the Give Me Some Credit dataset are summarized.

In Table 5, we can notice that the meta-classifiers have not reached a single higher metric, over the best metric from the individual models, shown in Table 3. However, the MCC metric is high and is comparable to the results of the XGBoost model. The Recall score was lower than expected, and the Type I error was one of the highest.

Table 6 shows more promising results. The MCC metric achieved a better result (0.395) than the best result from individual models (0.348, for the XGBoost model).

Table 5 Results for the German credit data for consensus models

	WARE	WA-T1E
AUC	0.884	0.869
ACC	0.827	0.806
MCC	0.568	0.507
Recall	0.585	0.458
Precision	0.786	0.817
Type I error	0.161	0.196
Type II error	0.214	0.183

Table 6 Results for the give me some credit data for consensus models

	WARE	WA-T1E
AUC	0.873	0.867
ACC	0.876	0.924
MCC	0.395	0.391
Recall	0.667	0.432
Precision	0.305	0.432
Type I error	0.026	0.041
Type II error	0.695	0.568

The AUC achieved the same result (0.873), as the best result from the individual models (for the AdaBoost model). Moreover, ACC, Precision and Type II error have the second best result, compared to the individual models shows in Table 4. However, for our most important metrics, i.e., Recall and Type I error, the results are one of the worst.

7 Concluding Remarks

With the expanding and rapidly growing credit market, and the popularity of machine learning methods, ML-based credit scoring approaches are becoming an increasingly important aspect of differentiating between good and bad borrowers. Due to the area in which banks operate, even a small change in the predictive power of the models results in large income growth. In this work, we propose to change the popular (in scientific literature) approach to credit scoring models so that the main goal is the best possible Recall score and a methodology to achieve this.

In this context we have discussed: optimization of hyperparameters of the models, in terms of the best Recall score, and taking into account class imbalance. We have conducted experiments on two publicly available benchmarks (Sect. 4). In both cases, the best Recall was achieved by the SVM model with the linear kernel, but it had the