Annals of Emerging Technologies in Computing (AETiC)
Vol. 7, No. 2, 2023

Research Article

A General Architecture for a Trustworthy Creditworthiness-Assessment Platform in the Financial Domain

Giandomenico Cornacchia[1,*], Vito Walter Anelli[1], Fedelucio Narducci[1], Azzurra Ragone[2] and Eugenio Di Sciascio[1]

[1]Politecnico di Bari, Italy
giandomenico.cornacchia@poliba.it; vitowalter.anelli@poliba.it; fedelucio.narducci@poliba.it; eugenio.disciascio@poliba.it
[2]Università degli Studi di Bari Aldo Moro, Italy
azzurra.ragone@uniba.it
*Correspondence: giandomenico.cornacchia@poliba.it

Received: 04th April 2022; Accepted: 15th March 2023; Published: 1st April 2023

Abstract: The financial domain is making huge advancements thanks to the exploitation of artificial intelligence. As an example, the credit-worthiness-assessment task is now strongly based on Machine Learning algorithms that make decisions independently from humans. Several studies showed remarkable improvement in reliability, customer care, and return on investment. Nonetheless, many users remain sceptical since they perceive the whole as only partially transparent. The trust in the system decision, the guarantee of fairness in the decision-making process, the explanation of the reasons behind the decision are just some of the open challenges for this task. Moreover, from the financial institution's perspective, another compelling problem is credit-repayment monitoring. Even here, traditional models (e.g., credit scorecards) and machine learning models can help the financial institution in identifying, at an early stage, customers that will fall into default on payments. The monitoring task is critical for the debt-repayment success of identifying bad debtors or simply users who are momentarily in difficulty. The financial institution can thus prevent possible defaults and, if possible, meet the debtor's needs. In this work, the authors propose an architecture for a Creditworthiness-Assessment duty that can meet the transparency needs of the customers while monitoring the credit-repayment risk. This preliminary study carried out an experimental evaluation of the component devoted to the credit-score computation and monitoring credit repayments. The study shows that the authors' architecture can be an effective tool to improve current Credit-scoring systems. Combining a static and a subsequent dynamic approach can correct mistakes made in the first phase and foil possible false positives for good creditors.

Keywords: *Credit Scoring; Early Warning; Explainability; Fairness*

---

1. Introduction

Artificial Intelligence (AI) is turning the banking and fintech industry into a smarter, more advanced, and versatile environment. Among the services banks offer, granting loans is one of the most frequent, and impactful, financial activities. For this task, in order to reduce the risk as much as possible for the bank to lose the money lent, a crucial step is represented by the creditworthiness assessment, namely the evaluation of the customer's capability to repay the loan.

1.1. Related Works

The traditional loan acceptance process needs to be faster, more efficient, and more effective. In particular, the traditional process had great difficulty in providing credit to a segment of the population

Giandomenico Cornacchia, Vito W. Anelli, Fedelucio Narducci, Azzurra Ragone and Eugenio Di Sciascio, "A General Architecture for a trustworthy Creditworthiness-Assessment Platform in the Financial domain", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 56-64, Vol. 7, No. 2, 1st April 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.02.005, Available: http://aetic.theiaer.org/archive/v7/v7n2/p5.html.

with no credit history [1] or disadvantaged economic conditions, resulting in unfair actions and/or inaccurate decisions. In that context, AI systems, in particular Machine Learning (ML) models, are helping to decrease the time required for making decisions and increase the process's accuracy. ML models can be key enabling tools also for unprivileged classes.

In the early stage of fintech innovation, automatic credit-scoring systems were software exploiting statistical inference to transform data into numerical measures to guide credit decisions [1]. They helped humans to make faster and more accurate decisions. With the increase of available information and computation power, these statistical models have been gradually substituted with Machine Learning (ML) models [2, 3]. In this respect, vast literature focused on exploiting machine learning models in the financial credit market [4, 5]. Baesens [6] compared several algorithms (i.e., Support Vector Machines, Logistic Regression, Decision Trees, and K-Nearest Neighbors) with artificial neural networks to find the best technique for a fraud detection task. Due to the strong imbalance of data in the financial credit domain by defaulter categories, several authors have proposed hybrid models and two-stage ensemble models to achieve more accurate predictions [7-9]. Chang *et al.* [10] were one of the first researchers who exploited the extreme gradient boosting model (XGBoost) for financial risk assessment. XGBoost is now considered one of the more effective models for loan default predictions. However, after a credit application, a critical phase occurs, when the customer slowly repays the loan, and it will only end with the debit extinction. At this stage, behaviour prediction scores are precious. In practice, behaviour scores consist mainly of user segmentation techniques that exploit multiple information sources (e.g., recent and frequent transactions). Such techniques mainly served marketing purposes (e.g., customer targeting, profiling, and pricing) or to predict user behaviour (e.g., late payment or insolvency) [1]. Behaviour scores turn out to be valuable tools used to direct the decision-making process when dealing with over-limit management, assess-portfolio risk, and other tasks. Even in this context, the first approaches were based on the statistical analysis of user behaviour with traditional clustering and data mining techniques applied to databases [11].

In the ML domain, these tasks can also be addressed by Early Warning Systems (EWS). Several authors have presented Early Warning models based on temporal data analysis. Ayvaz *et al.* [12] presented an EWS that leveraged Long-Short Term Memory (LSTM) neural architectures. They exploited macroeconomic variables to predict economic crises and to develop proper strategies. Similarly, Du [13] proposed to exploit AutoRegressive Integrated Moving Average (ARIMA) models to predict financial indicators (i.e., stock price, industrial value-added growth rate, domestic and foreign real deposit interest rate differential, and foreign direct investment as a percentage of Gross Domestic Product, GDP) and two Logit models to predict possible Chinese financial crises. Cheng *et al.* [14] proposed a graph attention architecture with a recursive self-attention mechanism to predict the default probability of Small and medium-sized enterprises. Geng *et al.* [15] used various ML models (e.g., SVM, NN, and Decision Trees) to predict the financial distress of Chinese companies. Liang et Cai [16] compared different architectures such as ANN, ARIMA, and SVM with LSTM for a Peer2Peer lending platform (i.e., Lending Club). The LSTM model outperforms other models in default prediction accuracy thanks to its ability to extract dynamic information. However, any of these works do pose a particular accent on the analysis of the customer's financial history.

1.2. Motivations

In this work, the authors design a platform that combines the creditworthiness-assessment task with the repaying monitoring of the loan. The contributions of this work are twofold:

i. Propose a general architecture for a Creditworthiness-Assessment Platform.
ii. An extensive evaluation of the two critical components: the credit scoring model and the early warning detector.

The remainder of the paper is organized as follows: Section 2 depicts the platform’s architecture, while Section 3 introduces and discusses the experimental evaluation. The conclusion and future work are drawn in Section 4.

2. A Trustworthy Credit Assessment Platform

In the last few years, the research has been focused more on applying ML models to predict customer creditworthiness rather than predicting whether the customer will effectively repay the loan. In this work, the authors propose a Creditworthiness-Assessment Platform that deals with both tasks.

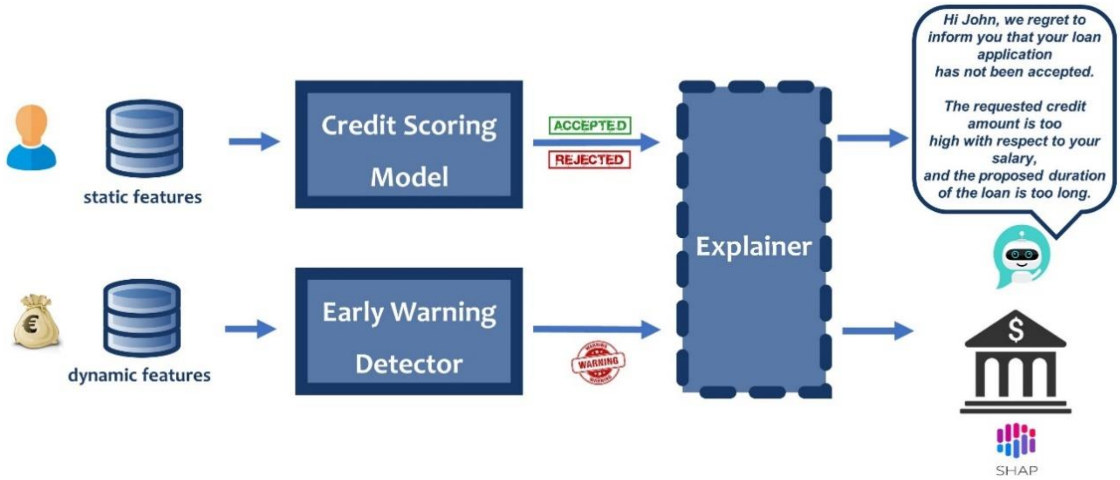

Figure 1. Creditworthiness Assessment Platform prediction

Figure 1. Creditworthiness Assessment Platform predictiondepicts the general architecture of the platform. First, the system can exploit two kinds of user data: static and dynamic. The former does not change over time (or change slowly). Data such as demographics, income, and gender belong to this first group. It is worth noticing that some of those features are considered *sensitive* by the legislation. The second group belong to features that change very frequently over time. In this scenario, these features are user transactions. Static and dynamic features are the input to two distinct modules: the Credit Scoring Model (CSM) and the Early Warning Detector (EWD).

The CSM is a binary classifier. Given a set of static user characteristics, it can decide whether that user will be able to repay its debt. The decision of the bank to grant or not the loan to a specific customer depends on this module. The task addressed by this module is particularly crucial because, as stated earlier, some user features are considered sensitive. Indeed, the last EU Commission regulation (April 21, 2021)[1] considered the financial domain one of the most regulated. The law proposal presents a pyramidal division of risk-based application of AI systems from minimal risk to unacceptable risk. Financial applications of AI systems (e.g., credit scoring) are considered *high-risk applications*. AI systems should comply with key ethical and trustworthy requirements since they need to pass different assessment steps. Therefore, it is really important that the algorithm implemented by the CSM does not put in place any kind of discrimination.

The second key component is the EWD. In this work, the authors preliminary evaluated and compared, in terms of accuracy performance, these two modules. The Early Warning Monitoring System will rely only on accepted credit requests. Customer card transactions periodically feed the system (e.g., daily, weekly, or at predefined intervals) that models customer trends in terms of expenses and the available balance ratio for each transaction. Given the transaction, the model will predict a potential future bankruptcy. Once the EWS has triggered the Business Intelligence team, they will analyze the model's output and decide if it corresponds to a False Positive or a True Positive situation. More in detail, the input of this module consists of the user transactions. When a customer makes a payment, purchase, or whatever financial transaction, this module checks whether that action could in some way jeopardize her ability to repay the debt. In contrast to the output of the Credit Scoring Model, which generally does not change over time, the prediction of the EWD is extremely fickle. In this case, EWD considers all the customer's history. Thus, the decision to trigger or not the warning depends on all the actions the user has done so far. The more the warning is true and early, the more the component is effective.

To comply with current regulations, the Business Intelligence team can use explanatory tools (e.g., Shapley values) to understand better which transactions have been responsible for this warning. Indeed,

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

regulations impose the bank to be totally aware of potential poor decisions and perform human-controlled actions in critical and life-changing situations.

The last component is the Explainer. Once the model performs the prediction task, the customer will be provided with an explanation, especially in case of rejection. In previous work [17-19], the authors provide different pipelines for generating natural language-based explanations, using both Shapley values and Counterfactual reasoning. As a game theory approach, the Shapley values give ranked feature importance of the most discriminating features for the decision task. It corresponds to the first stage of a user-friendly explanation. Shapley values theory is recognized as an effective tool for unveiling complex model decisions and a useful business intelligence analysis tool.

In contrast, counterfactual reasoning is used to discover a polarity between attribute and feature values to generate a natural-language-based explanation. Furthermore, the counterfactual exploration could provide plausible actions to receive the required credit. Again, in this case, the legislation plays a crucial role. More specifically, in the EU, the GDPR sets off the *right to explanation*: users have the right to ask for an explanation about an algorithmic decision made about them. In the UK, the Financial Conduct Authority (FCA) requires firms to explain why a more expensive mortgage has been chosen if a cheaper option is available. The G20 has adopted the OECD AI Principles[2] for a trustworthy AI, which underlined that users should understand AI outcomes and be able to challenge them.

These are the motivations behind putting this component in the architecture. For this component, the shape is dashed since the authors propose only a possible implementation inspired by [17, 18]. The module will be able to provide two kinds of explanation: a technical explanation for bank professionals, and a user-friendly explanation for the user. The former is based on SHAP that is inspired by the cooperative game theory based on the Shapley Values [20]. Each feature is considered a player that contributes differently to the outcome (i.e., the algorithm decision). SHAP provides a ranked list of the features that contributed the most to the less to the outcome. In this case, the bank analyst can understand what are the features that impacted the most algorithm decision.

The second form of explanation is in natural language. An effective solution could be to transform the output produced by SHAP into a natural language sentence. The natural language generation might be based on a set of rules that transform the Shapley values into natural language sentences. As an example, if the credit amount, the salary, and the loan duration have been the most important features for the algorithm decision, thus the natural language explanation could be:

"The credit amount is too high based on the salary and the duration is too long."

The platform provides different steps that cope with Fairness and Explainability requirements. Considering the Creditworthiness Assessment step, the model should provide evidence of fair decisions based on a specific metric of fairness before being placed on the market. Several metrics can be used to evaluate the algorithm's fairness [21]. However, choosing which one to optimize is a complex task since each metric can belong to different statistical criteria (i.e., Independence, Separation, and Sufficiency) and to different fairness concepts (e.g., group fairness, individual fairness, sub-group fairness). Choosing the right fairness metrics remains a challenging task [22]. In the study at hand, it has been chosen to operate in the context of *fairness under unawareness*, where sensitive features are not used as predictor variables. However, the analysis of the model's fairness is out of the scope of this study, and it will be carried out in future works.

3. Experimental Evaluation

As reported in the previous Section, in this work, the authors propose an implementation for the Credit Scoring Model and the Early Warning Detector modules. More specifically, they evaluated the capability of different machine learning algorithms to predict user defaults. The authors compared five different state-of-the-art models, i.e., Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Light Gradient Boosting (LGB), and eXtreme Gradient Boosting (XGB). The same models were used for both the credit scoring and early warning tasks to make the analysis comparable.

[2] https://oecd.ai/ai-principles

3.1. Datasets

The study has been carried out on an open-source dataset, the "1999 Czech Financial Dataset"[3], a real anonymized Czech bank database containing customer transactions, account information, and loan records released for the Financial PKDD'99 Discovery Challenge.

The dataset consists of the following different views: *account*, *customer*, *disposition*, *standing order*, *transaction*, *loan*, *credit card*, and *demographic data*. Because it is taken from a real bank database and contains both static and dynamic data, it perfectly fits the proposed architecture. The dataset contains customers who have had a loan and have paid it off, customers who are current on their payments, customers who are overdue on their payments, and customers who had failed to pay off the loan. In this experiment, the first two types of customers belong to the class of *good customers* and the last two to the *defaulting customers*. These are the two classes on which the binary classifiers have been learned. Other statistical information can be found in Table 1.

Table 1. Dataset information

View	Number of features	type	Number of records
Account	4	static	4500
Customer	3	static	5369
Disposition	4	static	5369
Permanent Order	6	static	6471
Transaction	10	dynamic	1056320
Credit Card	4	static	892
Demographic Data	16	static	77

3.2. Preprocessing

A static dataset has been created using all the views reported in Table 1. It does not contain *demographic data* that consists of demographic information of the users and *transactions* that have been used for the EWS. Another sensitive category, "*birth number*", which also includes gender information, has been omitted to comply with the regulations and avoid possible direct discrimination. The *permanent orders* relation has been added to the static data. Specifically, the different permanent order types have been One Hot Encoded and multiplied by the order amount for each user, thus condensing the number of permanent orders into a single feature.

The dataset was subsequently preprocessed by removing features affected by collinearity. The final static dataset consists of 682 users and 15 features. For each transaction, the authors calculated as a feature to give as input to the classifier the ratio between the account balance and the transaction itself. For example, if the user makes a \$100 transaction and its balance is \$1,000, the feature will be \$0.1. The formula for creating the dynamic user vector is available in Equation 1.

$$x^i(t) = \frac{transaction_t}{balance_{t-1}}, \qquad \text{with } x^i =< x_1^i, x_2^i, \dots, x_{max(t)}^i > \tag{1}$$

The customer profile at a given time is composed of a set of ratios (e.g., \$0.1, \$0.18, \$0.083) that show the trend of her financial profile (a ratio close to 1 means that the user is spending all his financial resources). For each new transaction, the whole list of transactions up to that time is used by the classifier to predict the customer default and trigger the warning.

3.3. Model Evaluation and results

Once both datasets have been preprocessed, the static dataset has been split with a train-test hold-out 70/30 methodology, and the train set has furthermore been divided with a k-fold cross-validation methodology with the number of folds equal to five. The dynamic dataset has been split with a different procedure. The test set comprises the transactions of 15 good and 15 bad creditors randomly chosen. The model validation has been performed on a balanced Validation set, with six good and six bad creditors. The rest of the transactions as part of the training set. Due to the imbalance between the two classes, all models have been fine-tuned on the training set using weighted class prediction and maximizing the F1 score. Then,

[3] https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm

the models have been evaluated on the test set. Splitting information is available in Table 2. Dataset Splitting information

Table 2. Dataset Splitting information

Dataset	**Train records**	**Validation records**	**Test records**	**Splitting ratio**	**Predicted Labels**
Static Dataset	382	95	205	Stratified	'status'
Dynamic Dataset	640	12	30	Balanced	'status'

The confusion matrix has been used for the evaluation procedure as a summary of the correct and incorrect predictions (i.e., TP, FP, TN, FN). Based on these results, the authors calculated, for each classifier, the Accuracy (ACC) metric. Furthermore, they enhanced the analysis, including the Area under the ROC Curve (AUC). ACC is the number of correct predictions divided by the total number of predictions, and AUC indicates how well the classifier can separate the classes of good debtors and defaulters, varying a probability cutoff threshold from 0 to 1.

Table 3. Static Dataset Results

	AUC	**Accuracy**	**TP**	**FP**	**TN**	**FN**
Logistic Regression	0,798	0,732	20	52	130	3
Random Forest	0,701	0,571	20	85	97	3
XGB	0,827	0,781	14	36	146	9
SVM	0,730	**0,790**	14	34	148	9
LGB	**0,830**	0,756	16	43	139	7

Table 4. Dynamic Dataset Results

	AUC	**Accuracy**	**TP**	**FP**	**TN**	**FN**
Logistic Regression	0,991	**0,867**	11	0	5	4
Random Forest	0,500	0,500	0	0	15	15
XGB	0,902	0,700	6	0	15	9
SVM	**0,996**	**0,867**	13	2	13	2
LGB	0,920	0,700	6	0	15	9

Table 3. Static Dataset Resultsshows the results of the different models for the CSM. Looking into the results, all the models seem to perform well for the creditworthiness task except for Random Forest (RF). RF does not separate the good creditors from the defaulters correctly. The best model in Accuracy is the SVM, but its ability to correctly separate the two classes is less distinctive.

In Table 4. Dynamic Dataset Results, it can be found the results of the several classifiers exploited for the EWD. Looking into the results, two models show the best performance, i.e., Logistic Regression and Support Vector Machines. The Random Forest classifier is the worst also in this case. Indeed, RF is unable to predict a warning situation as its TP is equal to 0 and FN is equal to 15. This tells us that the classifier has overfitted the negative class. As proof of this, its AUC is equal to 0.5, indicating a poor ability to separate the two classes. The SVM with the highest ACC, AUC, and TP is the classifier that best identifies a possible defaulter.

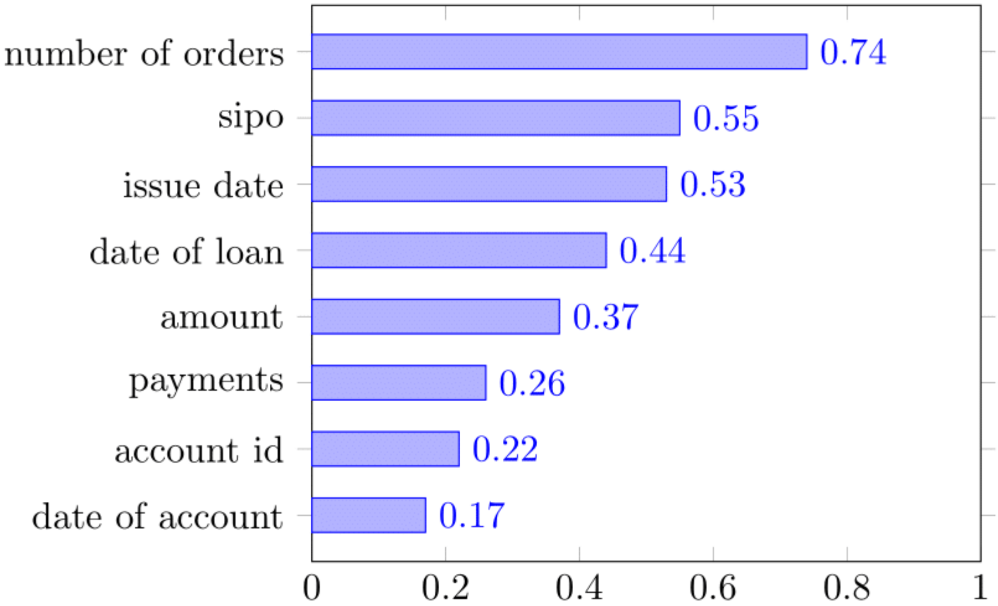

Figure 2. Shapley value for the LGB algorithm

By comparing the results on the static dataset with the dynamic one, it is possible to observe that the SVM and LR achieve the best overall accuracy with dynamic features. This is an interesting result since EWD can identify the defaulting customers with very high accuracy. However, when dynamic features are unavailable, static user characteristics can be really useful for deciding to approve a loan application.

The authors used one of the best models in the previous evaluation (i.e., SVM) for extracting the Shapley values by SHAP [20].

The most relevant features that emerged in Figure 2. Shapley value for the LGB algorithm are the number of permanent orders, type of permanent orders, amount, other features relative to the loan, card and account data. This example confirms the authors' intuition that providing an explanation based on Shapley values is not so useful, mainly when it is intended for customers who are usually unaware of how an ML algorithm works.

4. Conclusion

This work proposes a general architecture for the Credit Worthiness-Assessment platform. The platform aims to tackle two well-known and crucial tasks: credit and behavior scoring. The architecture consists of three Artificial Intelligence-powered modules: Credit Score, Early Warning Detection, and Explainer.

Inspired by recent literature, the authors investigated several Machine Learning algorithms to implement the Credit Score and the Early Warning Detection modules. Extensive experiments for Credit Score prediction showed that financial institutions could solely rely on static features to decide whether to grant or not a loan. The best-performing models for Creditworthiness-Assessment are SVM and LGB when dealing with static data, whereas SVM and logistic regression show to be the best models for dynamic data. Even though the number of datasets does not allow generalization and states that mentioned models are always the best, the performance shows that this set of state-of-the-art models can generally lead to very competitive results. For Early Warning Detection, the experiments show that a time window-based approach achieves state-of-the-art performance in monitoring user transactions and catching anomalous behaviors that can lead to a default.

According to the recent advances in explanation generation techniques, the Explainer module provides Shapley values that should help the user interpret the prediction outcome. Notwithstanding, the Shapley values do not show wide exploitation since ordinary users do not straightforwardly interpret their meaning. This further motivates authors' idea of taking a step further and developing an explainer which can generate natural language explanations.

In another future work, the authors will investigate the combination of static and dynamic features to improve the system accuracy further. Finally, they will assess which algorithms promptly detect warning signs before the others since an earlier warning could be a competitive advantage for financial institutions.

Acknowledgement

The authors acknowledge partial support of the projects: Smart Rights Management Platform Fincons III, PID2019-108965GB-I00, Fincons CdP3, Servizi Locali 2.0, ERP4.0, Secure Safe Apulia, PON ARS01_00876 BIO-D, Casa delle Tecnologie Emergenti della Città di Matera, PON ARS01_00821 FLET4.0, PIA Servizi Locali 2.0, H2020 Passapartout - Grant n. 101016956, PIA ERP4.0, IPZS-PRJ4_IA_NORMATIVO.

References

[1] Richardson Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, 1st ed. Oxford, UK: Oxford University Press, 2007, Print ISBN: 978-0199226405, Available: https://global.oup.com/academic/product/the-credit-scoring-toolkit-9780199226405?cc=it&lang=en&.

[2] Eliana Costa e Silva, Isabel Cristina Lopes, Aldina Correia and Susana Faria, "A logistic regression model for consumer default risk", *Journal of Applied Statistics*, pp. 2879-2894, Vol 47, No. 13-15, 2020, DOI: 10.1080/02664763.2020.1759030, Available: https://www.tandfonline.com/doi/abs/10.1080/02664763.2020.1759030.

[3] Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying and Kuiyi Liu, "A study on predicting loan default based on the random forest algorithm", *Procedia Computer Science*, Print ISSN: 1877-0509, pp. 503-513, Vol. 162, 2019, DOI: 10.1016/j.procs.2019.12.017, Available: https://www.sciencedirect.com/science/article/pii/S1877050919320277.

[4] Selçuk Bayraci and Orkun Susuz, "A Deep Neural Network (DNN) based classification model in application to loan default prediction", *Theoretical and Applied Economics*, Vol. XXVI, No. 4, pp. 75-84, 2019, Available: https://ideas.repec.org/a/agr/journl/vxxviy2019i4(621)p75-84.html.

[5] I O. Eweoya, Ayodele A. Adebiyi, Ambrose Azeta and Olufunmilola Amosu, "Fraud prediction in loan default using support vector machine", *Journal of Physics: Conference Series*, pp. 12-39, vol. 1299, No. 1, August 2019, Published by IOP Publishing, DOI: 10.1088/1742-6596/1299/1/012039, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1299/1/012039/pdf.

[6] Baesens Bar, T V. Gestel, S Viaene, M Stephanova, J Suykens *et al.* "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring." The Journal of the Operational Research Society, pp. 627–635, Vol. 54, No. 6, 2003, DOI: 10. 1057/palgrave.jors.2601545, Available: https://www.jstor.org/stable/4101754.

[7] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen, "Credit scoring using the hybrid neural discriminant technique", *Expert Systems with Applications*, Print ISSN: 0957-4174, pp. 245-254, Vol. 23, No. 3, 2002. DOI: 10.1016/S0957-4174(02)00044-1, Available: https://www.sciencedirect.com/science/article/pii/S0957417402000441.

[8] Jih-Jeng Huang, Gwo-Hshiung Tzeng and Chorng-Shyong, "Two-stage genetic programming (2SGP) for the credit scoring model", *Applied Mathematics and Computation*, Print ISSN: 0096-3003, pp. 1039-1053, Vol. 174, No. 2, 2006, DOI: 10.1016/j.amc.2005.05.027, Availabe: https://www.sciencedirect.com/science/article/pii/S009630030500531X.

[9] Shu L. Lin, "A new two-stage hybrid approach of credit risk in banking industry", *Expert Systems with Applications*, Print ISSN: 0957-4174, pp. 8333-8341, Vol. 36, No. 4, 2009, DOI: 10.1016/j.eswa.2008.10.015, Available: https://www.sciencedirect.com/science/article/pii/S0957417408007094.

[10] Yung-Chia Chang, Kuei-Hu Chang and Guan-Jhih Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions", *Applied Soft Computing*, Print ISSN: 1568-4946, pp. 914-920, Vol. 73, 2018, DOI: 10.1016/j.asoc.2018.09.029, Available: https://www.sciencedirect.com/science/article/pii/S1568494618305465.

[11] Guoqiang Zhang, Michael Y. Hu, B E. Patuwo and Daniel C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *European Journal of Operational Research*, Print ISSN: 0377-2217, pp. 16-32, Vol. 116, No. 1, 1999, DOI: 10.1016/S0377-2217(98)00051-4, Available: https://www.sciencedirect.com/science/article/pii/S0377221798000514.

[12] Ayvaz Ednan, Kaplan Kaplan and Melih Kuncan, "An Integrated LSTM Neural Networks Approach to Sustainable Balanced Scorecard-Based Early Warning System", *IEEE Access*, pp. 37958-37966, vol. 8, 2020, DOI: 10.1109/ACCESS.2020.2973514, Available: https://ieeexplore.ieee.org/document/8995520.

[13] Qingyu Du, "Financial Crisis Early Warning Based on Panel Data and Dynamic Dual Choice Model", *Complexity*, pp. 1-10, Vol. 2021, 2021, Published by Hindawi, DOI: 10.1155/2021/5596384, Available: https://www.hindawi.com/journals/complexity/2021/5596384/.

[14] Dawei Cheng, Zhibin Niu and Liqin Zhang, "Delinquent Events Prediction in Temporal Networked-Guarantee Loans", *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-13, 2020, DOI: 10.1109/TNNLS.2020.3027346, Available: https://ieeexplore.ieee.org/document/9222478.

[15] Ruibin Geng, Indranil Bose and Xi Chen, "Prediction of financial distress: An empirical study of listed Chinese companies using data mining", *European Journal of Operational Research*, Print ISSN: 0377-2217, pp. 236-247, Vol. 241, No. 1, 2015, Published by Elsevier B.V., DOI: 10.1016/j.ejor.2014.08.016, Available: https://www.sciencedirect.com/science/article/pii/S0377221714006511.

[16] Longyue Liang and Xuanye Cai, "Forecasting peer-to-peer platform default rate with LSTM neural network", *Electronic Commerce Research and Applications*, Print ISSN: 1567-4223, p. 100997, Vol. 43, 2020. DOI: 10.1016/j.elerap.2020.100997, Available: https://www.sciencedirect.com/science/article/abs/pii/S1567422320300740.

[17] Giandomenico Cornacchia, Fedelucio Narducci and Azzurra Ragone, "Improving the User Experience and the Trustworthiness of Financial Services", *18th International Conference promoted by the IFIP Technical Committee 13 on Human Computer Interaction - (INTERACT 2021)*, August 30 - 3 September 2021, Bari, Italy, Online ISBN: 978-3-030-85606-9, E-ISBN: 978-3-030-85607-6, DOI: 10.1007/978-3-030-85607-6_19, pp. 264-269, Published by Springer, Available: https://link.springer.com/chapter/10.1007/978-3-030-85607-6_19.

[18] Giandomenico Cornacchia, Fedelucio Narducci and Azzurra Ragone, "A general model for fair and explainable recommendation in the loan domain", *Joint Workshop Proceedings of the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM RecSys 2021, Amsterdam, 2021*, September 27–1 October 2021, Amsterdam, Netherlands, vol. 2960, 2021, Available: https://ceur-ws.org/Vol-2960/paper12.pdf.

[19] Giandomenico Cornacchia, Vito W. Anelli, Giovanni M. Biancofiore, Fedelucio Narducci, Claudio Pomo *et al*., "Auditing fairness under unawareness through counterfactual reasoning", *Information Processing & Management*, Print ISSN: 0306-4573, p. 103224, Vol. 60, No. 2, 2023, DOI: 10.1016/j.ipm.2022.103224, Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457322003259.

[20] Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions", *Advances in neural information processing systems*, Print ISBN: 9781510860964, pp. 4768-4777, Vol. 30, 2017, Published by Curran Associates Inc., DOI: 10.5555/3295222.3295230, Available: https://dl.acm.org/doi/10.5555/3295222.3295230.

[21] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt *et al*., "Fairness Measures for Machine Learning in Finance", *The Journal of Financial Data Science*, Print ISSN: 2640-3943, pp. 33-64, Vol. 3, No. 4, 2021, DOI: 10.3905/jfds.2021.1.075, Available: https://jfds.pm-research.com/content/3/4/33.

[22] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Giuseppina I. Penco *et al*., "A clarification of the nuances in the fairness metrics landscape", *Scientific Reports*, Print ISSN: 2045-2322, Vol. 12, No. 1, 2021, DOI: 10.1038/s41598-022-07939-1, Available: https://rdcu.be/c7t6L.