Finance Research Letters 60 (2024) 104867

Contents lists available at ScienceDirect

Finance Research Letters

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

Loan default predictability with explainable machine learning ☆

Huan Li [a], Weixing Wu [b,*]

[a] School of Banking and Finance, University of International Business and Economics, Beijing, 100029, People's Republic of China
[b] School of Finance, Capital University of Economics and Business, Beijing, 100070, People's Republic of China

ARTICLE INFO

Keywords:
Loan default
Machine learning
SHapley additive exPlanations

ABSTRACT

This paper studies loan defaults with data disclosed by a lending institution. We comprehensively compare the prediction performance of nine commonly used machine learning models and find that the random forest model has an efficient and stable prediction ability. Then, we apply an explainable machine learning method, i.e., SHapley Additive exPlanations (SHAP), to analyze the important factors affecting loan defaults. Moreover, we conduct an empirical study and find that the significant influencing factors are clearly consistent with those suggested by SHAP: the older the lender and the longer their working experience, the lower the risk of loan default.

1. Introduction

With the evolution of society and changes in consumption concepts, credit loans have emerged as a popular method among consumers. Therefore, it has become more important for lending institutions to enhance credit risk management capabilities based on machine learning tools and extensive user data. As observed by Butaru et al. (2016), heterogeneity exists in risk prediction models across institutions, which motivates us to evaluate multiple (nine) machine learning tools for credit risk prediction, providing insights into the risk system construction for lending institutions. Moreover, we focus the theoretical analysis and empirical test on the relationship between the information of lenders and loan default probability, to our knowledge, which is less studied in previous works.

In the literature, since the pioneering research of Durand (1941), there has been substantial attention focused on analyzing credit/loan risk with linear statistical models, e.g., linear discriminant analysis (Myers and Forgy, 1963; Lane, 1972). Subsequently, Wiginton (1980) applied a logit model to assess credit applications. Later research gradually turned to nonlinear methods, such as tree-based models, ensemble methods, and neural networks (Bauer and Agarwal, 2014; Zieba et al., 2016). Kvamme et al. (2018) combined a convolutional neural network (CNN) with a random forest (RF) classifier and applied it to consumer transaction data to predict mortgage defaults; similarly, Albanesi and Vamossy (2019) combined a CNN with an XGBoost through arithmetic averaging to predict consumer default; Sigrist and Hirnschall (2019) introduced the Grabit model, which uses the gradient tree boosting method for the Tobit model (Tobin, 1958; Rosett and Nelson, 1975); and Barbaglia et al. (2021) investigated loan defaults in several European countries by comparing the performance of penalized logistic regression model, XGBoost, and random forest. However, these existing methods analyze and compare the capabilities of a limited number of machine learning models. In contrast, we comprehensively compare nine commonly used machine learning models and find that the RF model has the best overall performance due to its anti-overfitting and automatic feature selection. To explain our prediction model, we apply the SHAP method

☆ This work was supported by the National Natural Science Foundation of China [grant number 72373024].
* Corresponding author.
E-mail address: wxwu@cueb.edu.cn (W. Wu).

https://doi.org/10.1016/j.frl.2023.104867
Received 19 October 2023; Received in revised form 2 December 2023; Accepted 13 December 2023
Available online 15 December 2023

Table 1
Variable description.

Variables	Interpretation & Quantitative description
id	Loan user number
Lender characteristics:	
Income	Lender's annual income
Age	Lender's age when applying for loan
is_married	Single (0); Married (1)
City	City A (0); City B (1); City C (2); ...
Region	Region A (0); Region B (1); Region C (2); ...
current_job_years	Actual number of years expressed
current_house_years	Actual number of years expressed
house_ownership	Rented (0); Owned (1); Norend_noown (2)
car_ownership	Yes; No
Profession	Agency staff (0); Engineers (1); Doctors (2); ...
working_years	Actual number of years expressed
Dependent variable:	
Default	Yes (1); No (0)

This table illustrates all variables of loan default and the corresponding quantification methods.

(SHapley Additive exPlanations) to show that age and working years are important factors in predicting a lender's default risk, and these factors are also less discussed in the literature.

Beyond using only machine-learning techniques to predict the loan default probability, some researchers have analyzed other deeper aspects of credit risk. For instance, Baesens et al. (2003) utilized neural network rule extraction techniques to model the explanatory relationship between the characteristics of loan application and final decisions. Butaru et al. (2016) compared risk management actions and the reasons for delinquency across six commercial banks. Fuster et al. (2022) studied the distribution of less risky lenders and deemed riskier across societally important categories (e.g., race). As a complement to previous studies, we introduce a *theoretical analysis* of the lender's age and the years of employment for loan default risk, inspired by the information provided by the prediction model. Then, we conduct empirical research and a robustness test on these two factors and find that the older the lender and the longer the working years, the lower the default risk.

More recently, Ma et al. (2023) applied five machine learning models to predict the credit default of Chinese real estate listed companies. Gao et al. (2023) predicted a framer's loan default with three machine learning algorithms and found relevance between climate change and loan default risk. Similarly, Zhu et al. (2023) employed four kinds of models to predict a loan default and studied how the information of the loan itself affected default. Nevertheless, our method has several apparent differences from these existing works: (1) we evaluate the prediction performance of more models (i.e., nine types); (2) we conduct a theoretical analysis of the two factors of lenders on loan default risk; and (3) we apply empirical and robustness tests to validate our hypotheses, which are also missed by the above methods.

2. Theoretical analysis and research hypotheses

The lender's default will bring certain default benefits; however, it will also generate costs of social shame (Gross and Souleles, 2002). Social stigma costs include two categories: nonmonetary costs (e.g., personal reputation and psychological pressure) and monetary costs (e.g., restrictions on future loans due to damaged credit records). These social stigma costs are positively related to the age of the lender, because the older the age, the wider the personal social network, and the wider the spread of information about the lender's default. In addition, a large number of studies have shown that negative information spreads faster and more widely in social networks (Berger and Milkman, 2012; Fang and Ben-Miled, 2017). Therefore, we propose a hypothesis that according to opportunity cost theory, the older the lender is, the higher the social shame cost caused by loan default; therefore, the lender is less likely to make a loan default decision.

Generally, an applicant's working years are positively related to career stability. If the individual's job is stable, the lender will have a positive and optimistic attitude toward the future economic situation. According to reward theories in psychology and behavioral economics, people tend to take actions to achieve their desired positive outcomes and avoid negative punishments (Kahneman and Tversky, 1979; Thaler, 2016). If a lender feels optimistic about future economic prospects, they may be more motivated to make prudent financial decisions to avoid possible risks. Based on the above theoretical analysis, we propose a hypothesis that the longer the lender's working experience and the higher their career stability, the more active they are in avoiding loan defaults.

3. Data and processing

In this work, we choose publicly available data from Baidu PaddlePaddle (Baidu, 2021) for loan default research. The data come from a lending institution and have been desensitized, containing a total of 84,000 loan records. The variables and quantification methods selected for this study are reported in Table 1. There are 11 characteristics that describe a lender. To depict the data more

Table 2
Summary statistics.

Variables	N	Mean	Std. dev.	Minimum	Median	Maximum
Income	84,000	50,014.61	28,762.38	103	50,104	99,992
Age	84,000	49.94	17.09	21	50	79
current_job_years	84,000	6.32	3.65	0	6	14
current_house_years	84,000	12	1.40	10	12	14
working_years	84,000	10.08	6.01	0	10	20

This table reports the number of observations (N), the mean, standard deviation (Std. dev.), minimum, median, and maximum of loan default data.

clearly, we measured the degree of dispersion and central tendency of continuous variables such as income, age, current_job_years, current_house_years, and working_years. The statistical results are shown in Table 2.

Considering that the order of magnitude difference between income and other variables is relatively large, we take the natural logarithm of lender income. For the discrete variables, including is_married (single; married), house_ownership (rented; norent_noown; owned), and car_ownership (yes; no), we add four dummy variables: Nomarried, Rented, Norentown, NoCar.

4. Prediction model of loan default risk

In this section, we comprehensively compare nine commonly used machine learning algorithms and choose the model with the best performance to assess loan default risk. These methods are k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), logistic regression (Logistic), Gaussian naive Bayes (GauNB), multilayer perceptron (MLP), random forest (RF), extreme gradient boosting (XGBoost), and stacking. We briefly review these methods below.

4.1. Machine learning algorithms

The KNN algorithm makes predictions based on similarities among samples. For a test sample, KNN first finds the K training samples closest to the test sample and then obtains the class or value of the test sample by combining those of these K samples. SVM performs classification or regression tasks by finding an optimal hyperplane in the feature space. Specifically, the data are first mapped into a high-dimensional feature space; then, the maximum margin hyperplane is found in the feature space. DT is a tree model that builds a predictive model through a series of selections and splits on input data. Logistic regression is a linear model that converts linear combinations of input variables into probability values through the sigmoid function. GNB assumes that the features of each category are independent and obey a Gaussian distribution under a given category, and then calculates the posterior probability through Bayes' theorem. An MLP is composed of multiple neuron layers, where each neuron layer is connected to the next layer through weights and nonlinear activations. The main idea is to learn the complex nonlinear relationship between the input features and output. The following three methods belong to ensemble learning. The RF is a machine learning algorithm based on ensemble learning that performs classification and regression tasks by building multiple decision trees and combining their predictions. XGBoost is a gradient-boosting tree algorithm that gradually improves the model's predictive power by training multiple weak classifiers (usually decision trees) in series. Stacking aims to improve the model's performance by combining the prediction results of multiple basic classifiers (also called "weak learners").

4.2. Model prediction performance

To assess loan default risk, we construct nine prediction models based on machine learning, where the stacking model uses RF and XGBoost as the base estimator and logistic regression as the classifier. To reduce the impact of the imbalance in the number of default samples and nondefault samples, we adopt a random undersampling method to match the number of default samples via sampling without replacement. The performance evaluation results of these models are shown in Table 3. RF has the highest accuracy, AUC, and F1 score, and DT has the highest recall. The recall and F1 score of SVM, Logistic, GauNB, and MLP are relatively low. Overall, the RF model has the best predictive power because RF can reduce the risk of overfitting due to its random nature and has good robustness for potential noise and outliers in the data.

To further analyze the direction and degree of influence of each independent variable on default risk, we introduce the SHAP method to rank the feature importance of the random forest model. Feature importance refers to the contribution of the feature to the prediction result. SHAP is an additive explanation model based on game theory, and all features are considered "contributors". For a single sample, each feature has a SHAP value. To obtain the importance of each feature, the absolute values of the SHAP values of all samples are averaged. The feature importance results of the RF model are shown in Fig. 1. Age, income, working_years, city, and profession are the top five most important factors for evaluating default risk. In addition, Fig. 2 shows the SHAP values of all samples, where the abscissa represents the SHAP value, the ordinate lists all features, and each point corresponds to a sample. We can see that the older the lender is, the lower the probability of default; longer working experience reduces the probability of default. For features of age and working_years, many red points have negative SHAP values.

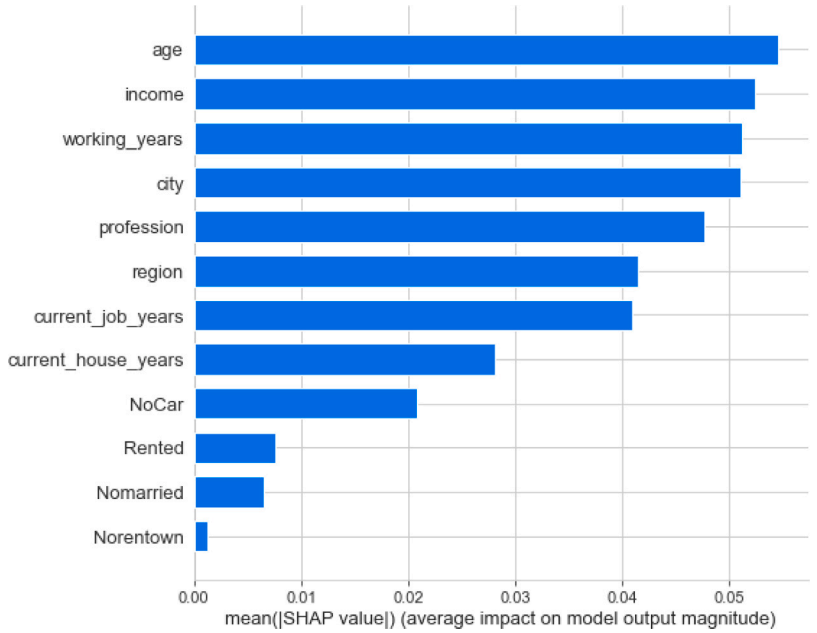

Fig. 1. Feature importance distribution of random forest model. The abscissa represents the average SHAP value, and the ordinate lists all features.

Table 3
Model prediction performance.

Models	Accuracy (%)	Recall (%)	AUC	F1 score
KNN	83.19	74.38	0.85	0.52
SVM	55.83	51.22	0.56	0.22
Logistic	51.35	57.04	0.55	0.22
DT	84.77	87.41	0.85	0.59
RF	88.05	79.99	0.93	0.62
XGBoost	82.45	81.50	0.87	0.53
Stacking	86.76	82.80	0.93	0.61
GauNB	31.84	80.63	0.55	0.23
MLP	69.18	44.44	0.63	0.26

This table presents the performance of nine commonly used machine learning models on loan default prediction. We choose four evaluation metrics: accuracy, recall, the area under the curve (AUC), and F1 score. Accuracy represents the proportion of correct predictions by the model among all samples. Recall is the percentage of correct positive predictions among all actual positive samples. AUC measures the area underneath the ROC curve, which plots the true positive rate versus the false positive rate at different thresholds. The F1 score presents the harmonic mean of precision and recall, where precision is the percentage of correct positive predictions among all positive predictions. The best two scores are marked with red and blue font.

5. Empirical test

As mentioned previously, the machine learning model predicts loan default by analyzing multiple variables. Meanwhile, it can also provide the variable's importance. Based on this information, we conduct further empirical tests. Because default is a 0–1 variable, we choose the Probit model for regression, and its formula is expressed as follows:

$$LD_i = \alpha + \beta_1 Age_i + \beta_2 WorkYear_i + \gamma X_i + \mathcal{V}_r + \mathcal{V}_p + \epsilon_i, \tag{1}$$

where loan default (LD) is the response variable, age (Age) and years of employment ($WorkYear$) are the key explanatory variables, X is the control variable, ϵ is unobservable errors, and $\mathcal{V}_r$ and $\mathcal{V}_p$ represent the fixed effects of region and profession, respectively. The main regression results are shown in Table 4. Both age and working years are negatively related to loan default at the 1% significance level. Table 5 shows the marginal effects of age and working years on loan defaults. From the table, it can be found that the marginal effects of the age and working years of lenders are also significant at the 1% level. The regression results demonstrate that age and working years can reduce loan default probability.

6. Robustness test

The previous sections concluded that the lender's age and years of employment can reduce the probability of loan default. In this section, we apply the matching sample method to prove the robustness of the previous empirical test. Sample matching can

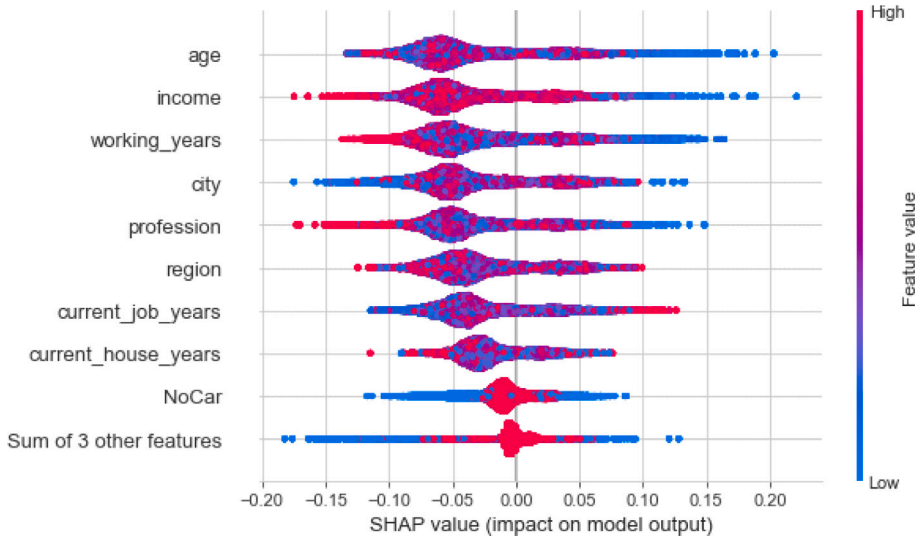

Fig. 2. Top nine features that have an impact on default prediction. Red (blue) represents high (low) feature values, that is, the numerical value of the predictor is big (small).

Table 4
Main regression.

Default	Coef.	Std. Err.	t− value	p− value
Age	−0.001856***	0.000329	−5.64	0
working_years	−0.011537***	0.00126	−9.15	0
income	−0.020484***	0.005701	−3.59	0.000327
current_job_years	0.005088**	0.002058	2.47	0.013423
current_house_years	−0.000265	0.004036	−0.07	0.947689
City	0.000086	0.000064	1.34	0.180777
Nomarried	0.127377***	0.019497	6.53	0
NoCar	0.072584***	0.012443	5.83	0
Norentown	0.048134	0.044704	1.08	0.281605
Rented	0.204274***	0.027984	7.30	0
Constant	−1.043359***	0.115923	−9.00	0
Pseudo R^2	0.0122	Number of obs	84,000	

This table presents the regression results of loan default. *, **, and *** refer to significance at 10%, 5%, and 1%, respectively.

Table 5
Margin effects.

	dy/dx	Std. Err.	z	P>z
Age	−0.000373***	0.000066	−5.64	0
working_years	−0.002319***	0.000253	−9.16	0

This table reports the margin effects of age and working_years on loan default. *** indicates significance at the 1% level.

make control variables relatively similar such that the differences in default risk are more closely related to the differences in the explanatory variables. To perform sample matching, we use propensity score matching (PSM) (Caliendo and Kopeinig, 2008). The explanatory variables in Eq. (1) contain age and years of employment. Therefore, it is necessary to construct experimental groups and control groups for these two variables. Specifically, for each variable, a quarter of the samples above the 75th quantile are selected and defined as the experimental group. Then, PSM is used to match the remaining three-quarters of the samples one-to-one to obtain the control group with the same number as the experimental group. For the explanatory variables of age and working_years, we introduce indicator variables treat_age and treat_work, respectively. Specifically, the value of the indicator available in the experimental group is 1 and that in the control group is 0. We again conduct Probit regression and compute the margin effects, and the corresponding results are reported in Tables 6–9. These tables indicate that there is still a significant negative relationship between loan default and age and working_years. In addition, we conduct the regression with the Logit model, and the significance is consistent with that of the Probit model.

Table 6
Robustness test on age.

Default	Coef.	Std. Err.	t− value	p− value
treat_age	−0.108838***	0.015728	−6.92	0
working_years	−0.01434***	0.001768	−8.11	0
Income	−0.012498	0.007929	−1.58	0.114959
current_job_years	0.012258***	0.002873	4.27	0.00002
current_house_years	0.013742**	0.005632	2.44	0.014689
City	0.000347***	0.00009	3.87	0.00011
Nomarried	0.119251***	0.026136	4.56	0.000005
NoCar	0.032863*	0.017100	1.92	0.054633
Norentown	−0.051961	0.061419	−0.85	0.397544
Rented	0.165955***	0.037673	4.41	0.000011
Constant	−1.356457***	0.159585	−8.50	0
Pseudo R^2	0.018222	Number of obs	42,808	

This table presents the regression results of loan default when performing PSM on age. *, **, and *** refer to significance at 10%, 5%, and 1%, respectively.

Table 7
Margin effects.

	dy/dx	Std. Err.	z	P>z
treat_age	−0.022484***	0.003248	−6.92	0
working_years	−0.002962***	0.000365	−8.11	0

This table shows the margin effects of treat_age and working_years on loan default. *** indicates significance at the 1% level.

Table 8
Robustness test on working_years.

Default	Coef.	Std. Err.	t− value	p− value
Age	−0.002049***	0.000442	−4.64	0.000004
treat_work	−0.050462***	0.015076	−3.35	0.000816
Income	−0.024555***	0.007511	−3.27	0.001079
current_job_years	0.004766**	0.002306	2.07	0.038797
current_house_years	−0.003679	0.005381	−0.68	0.494151
City	−0.000021	0.000086	−0.25	0.806035
Nomarried	0.139580***	0.025757	5.42	0
NoCar	0.089112***	0.016685	5.34	0
Norentown	0.043528	0.058289	0.75	0.455201
Rented	0.208224***	0.036649	5.68	0
Constant	−1.139372***	0.156033	−7.30	0
Pseudo R^2	0.013624	Number of obs	48,260	

This table presents the regression results of loan default when performing PSM on working_years. *, **, and *** refer to significance at 10%, 5%, and 1%, respectively.

Table 9
Margin effects.

	dy/dx	Std. Err.	z	P>z
Age	−0.000403***	0.000087	−4.64	0
treat_work	−0.009930***	0.002967	−3.35	0

This table shows the margin effects of age and treat_work on loan default. *** indicates significance at the 1% level.

7. Conclusions and future work

In this study, we employ nine commonly used machine learning algorithms, including support vector machine, decision tree, multilayer perceptron, random forest, etc., to evaluate the default risk on the loan data shared by a lending institution. We find that the random forest model outperforms other models, with an accuracy of 88.05%. Then, we introduce the SHAP method to analyze the important factors that affect loan defaults and the direction of influence. Inspired by this information, we theoretically analyze the relationship between loan default probability and the age and working years of a lender based on the opportunity cost theory and reward theories. With the empirical and robustness tests, we find that the older the lender and the longer the working experience, the lower the risk of loan default.

Our study illustrates the potential insights and benefits that machine learning models with the SHAP method can bring to banks, lending institutions, and regulatory bodies, leading to more informed risk management systems. The explainable machine learning

method facilitates the understanding of the model's decision-making process, which can provide banks and lending institutions with a clearer basis for the loan application process while improving customer satisfaction. In addition, for regulatory bodies, our findings can provide more suggestions about regulatory evaluation standards, helping to enhance financial stability.

Due to the limitations of public loan data, for example, the exact regions are unknown (only symbols), we cannot further analyze the relationship between loan default risk and the region's economy. In the future, we would like to collect large-scale data with more features, e.g., time dimension, different institutions, and exact regions, and then conduct a more comprehensive study on loan default.

CRediT authorship contribution statement

Huan Li: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Weixing Wu:** Conceptualization, Methodology, Supervision, Validation, Visualization, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

Acknowledgment

This work was supported by the National Natural Science Foundation of China [grant number 72373024].

References

Albanesi, S., Vamossy, D.F., 2019. Predicting Consumer Default: A Deep Learning Approach. Working Paper 26165, National Bureau of Economic Research.
Baesens, B., Setiono, R., Mues, C., Vanthienen, J., 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. Manage. Sci. 49 (3), 312–329.
Baidu, 2021. Loan default data. https://aistudio.baidu.com/aistudio/datasetdetail/112664. (visited on 2022-10-11).
Barbaglia, L., Manzan, S., Tosetti, E., 2021. Forecasting loan default in europe with machine learning*. J. Financ. Econom. 21 (2), 569–596.
Bauer, J., Agarwal, V., 2014. Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. J. Bank. Financ. 40, 432–442.
Berger, J., Milkman, K.L., 2012. What makes online content viral? J. Mar. Res. 49 (2), 192–205.
Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A.W., Siddique, A., 2016. Risk and risk management in the credit card industry. J. Bank. Financ. 72, 218–239.
Caliendo, M., Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. J. Econ. Surv. 22 (1), 31–72.
Durand, D., 1941. Risk Elements in Consumer Instalment Financing. NBER.
Fang, A., Ben-Miled, Z., 2017. Does bad news spread faster? In: International Conference on Computing, Networking and Communications. pp. 793–797.
Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., 2022. Predictably unequal? The effects of machine learning on credit markets. J. Finance 77 (1), 5–47.
Gao, W., Ju, M., Yang, T., 2023. Severe weather and peer-to-peer farmers' loan default predictions: Evidence from machine learning analysis. Finance Res. Lett. 58, 104287.
Gross, D.B., Souleles, N.S., 2002. An empirical analysis of personal bankruptcy and delinquency. Rev. Financ. Stud. 15 (1), 319–347.
Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47 (2), 263–291.
Kvamme, H., Sellereite, N., Aas, K., Sjursen, S., 2018. Predicting mortgage default using convolutional neural networks. Expert Syst. Appl. 102, 207–217.
Lane, S., 1972. Submarginal credit risk classification. J. Financ. Quant. Anal. 7 (1), 1379–1385.
Ma, Y., Zhang, P., Duan, S., Zhang, T., 2023. Credit default prediction of Chinese real estate listed companies based on explainable machine learning. Finance Res. Lett. 58, 104305.
Myers, J.H., Forgy, E.W., 1963. The development of numerical credit evaluation systems. J. Amer. Statist. Assoc. 58 (303), 799–806.
Rosett, R.N., Nelson, F.D., 1975. Estimation of the two-limit probit regression model. Econometrica 43 (1), 141–146.
Sigrist, F., Hirnschall, C., 2019. Grabit: Gradient tree-boosted tobit models for default prediction. J. Bank. Financ. 102, 177–192.
Thaler, R.H., 2016. Behavioral economics: Past, present, and future. Amer. Econ. Rev. 106 (7), 1577–1600.
Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26 (1), 24–36.
Wiginton, J.C., 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. J. Financ. Quant. Anal. 15 (3), 757–770.
Zhu, X., Chu, Q., Song, X., Hu, P., Peng, L., 2023. Explainable prediction of loan default based on machine learning models. Data Sci. Manag. 6 (3), 123–133.
Zieba, M., Tomczak, S.K., Tomczak, J.M., 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Syst. Appl. 58, 93–101.