Data Science and Management 6 (2023) 123–133

Contents lists available at ScienceDirect

Data Science and Management

journal homepage: www.keaipublishing.com/en/journals/data-science-and-management

Research article

Explainable prediction of loan default based on machine learning models

Xu Zhu [a], Qingyong Chu [a], Xinchang Song [a], Ping Hu [a], Lu Peng [a,b,*]

[a] School of Management, Wuhan University of Technology, Wuhan, 430070, China
[b] Research Institute of Digital Governance and Management Decision Innovation, Wuhan University of Technology, Wuhan, 430070, China

ARTICLE INFO

Keywords:
Explainable prediction
Machine learning
Loan default
Local interpretable model-agnostic explanations

ABSTRACT

Owing to the convenience of online loans, an increasing number of people are borrowing money on online platforms. With the emergence of machine learning technology, predicting loan defaults has become a popular topic. However, machine learning models have a black-box problem that cannot be disregarded. To make the prediction model rules more understandable and thereby increase the user's faith in the model, an explanatory model must be used. Logistic regression, decision tree, XGBoost, and LightGBM models are employed to predict a loan default. The prediction results show that LightGBM and XGBoost outperform logistic regression and decision tree models in terms of the predictive ability. The area under curve for LightGBM is 0.7213. The accuracies of LightGBM and XGBoost exceed 0.8. The precisions of LightGBM and XGBoost exceed 0.55. Simultaneously, we employed the local interpretable model-agnostic explanations approach to undertake an explainable analysis of the prediction findings. The results show that factors such as the loan term, loan grade, credit rating, and loan amount affect the predicted outcomes.

1. Introduction

Online personal loans have grown in popularity. Several personal credit systems are available on the Internet, including bank platforms and Ant Credit Pays. Some platforms have relatively simple lending requirements, making it easy for users to borrow money. However, the interest rates on postponed repayments and the handling charges have increased, raising the risk of non-payment. As a result, to assist decision makers in avoiding financial risk, it is important to identify the factors that can affect a loan repayment.

Various artificial intelligence algorithms have been applied to loan predictions (Li et al., 2021). For example, Emekter et al. (2015) used a logistic regression model to predict the default probability of borrowers and found that the revolving line utilization, Fair Isaac Corporation (FICO) score, debt-to-income ratio, and credit grade are important factors. Using consumer tradelines, credit bureaus, and macroeconomic data, Butaru et al. (2016) found that decision trees and random forests outperform logistic regression in forecasting credit card delinquencies. Fitzpatrick and Mues (2016) verified that boosted regression trees outperform penalized logistic regression in predicting mortgage defaults. Xia et al. (2017a) demonstrated the modeling process of a personal credit risk assessment based on the XGBoost model. They processed the missing values, normalized the original data, used an XGBoost model to rank the importance of the features, and finally applied the trained and optimized XGBoost model for predictive purposes. Xia et al. (2017b) proposed a CSXGBoost model for an individual risk assessment in the P2P field by combining cost-sensitive learning with a tree model. Deng (2019) conducted research using data provided by the LendingClub platform and identified the 20 variables with the highest influence. The author conducted a regression analysis using the logit model and generated qualitative analysis results by displaying the association between several variables. Kim and Cho (2019) introduced a semi-supervised learning strategy that considers the features of social lending data in accurately predicting social lending. They combined label propagation and the transformation of support vector machines. Zhou et al. (2019) linearly weighted the predictions of gradient-boosting decision tree, XGBoost, and LightGBM models. They experimented with a model on a P2P dataset in China and concluded that it can effectively cope with sparse unbalanced high-dimensional samples. Sadhwani et al. (2021) developed a nonlinear deep learning model for analyzing the behavior of mortgage borrowers using origination and monthly performance records for over 120 million mortgages originating across the U.S. Fuster et al. (2022) evaluated the change from traditional logit technology to machine learning technologies using a large dataset from the U.S. mortgage

Peer review under responsibility of Xi'an Jiaotong University.
* Corresponding author. School of Management, Wuhan University of Technology, Wuhan, 430070, China.
E-mail address: pengluhust@whut.edu.cn (L. Peng).

https://doi.org/10.1016/j.dsm.2023.04.003
Received 10 January 2023; Received in revised form 24 April 2023; Accepted 25 April 2023
Available online 5 May 2023

market. They also found that black and white Hispanic borrowers are predicted to lose relative to white and Asian borrowers.

The use of machine learning approaches has produced excellent outcomes in the studies mentioned above. Nonetheless, a prediction analysis does not provide decision makers with sufficient information. Even if the prediction accuracy of the model is excellent, the entire decision-making process for the prediction models may be irrational. It is critical for decision makers to understand the model and prediction rules. An increasing number of scholars have studied explanatory models to overcome the black box problem occurring with machine learning methods. For example, Chen et al. (2021) designed a deep matrix decomposition method with non-negative constraints that increases the robust interpretability of deep learning representations by defining interpretable losses. Dalmau et al. (2021) used the shapley additive explanations method to explain the importance of different features. Onchis and Gillich (2021) used local interpretable model-agnostic explanations (LIME) to train a simple interpretable model. Peng et al. (2022) added an attention mechanism to a long short-term memory network to explain the importance of the input variables. Wu et al. (2022) constructed an improved interpretable predictive model based on a multi-head attention mechanism, which can be used to evaluate the importance of the input variables. Zhou et al. (2022) designed an interpretable temporal attention network for the forecasting of COVID-19. However, few studies have conducted interpretability analyses of financial risk controls.

This study applies machine learning approaches and an interpretable model to the prediction and analysis of loan defaults. We compared the prediction performance of logistic regression, decision tree, XGBoost, and LightGBM models using a large-scale example. This study identifies the important characteristics affecting the probability of a default from the perspective of explainable machine learning. The contributions of this study are as follows:

(a) To reduce the number of feature dimensions, techniques such as deletion, a principal component analysis, feature interaction, and the population stability index are utilized.
(b) Personal loan default prediction models based on XGBoost, LightGBM, decision tree, and logistic regression are compared.
(c) The LightGBM model with the best prediction performance is analyzed using LIME.

2. Methodology

2.1. Logistic regression model

A logistic regression model is a machine-learning approach (Song et al., 2021; Sun et al., 2021). This model is an excellent classification algorithm that adds sigmoid function mapping to a linear regression, as shown in Eq. (1). This type of computation is inexpensive, its speed is high, and the required storage resources are low.

$$f(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Here, $z = \beta_0 + \sum_{k=1}^{m} \beta_k x_k$ refers to a weighted linear combination model, β_0 indicates the intercept of a function, $\beta_k (k = 1, \cdots, m)$ denotes the correlation coefficient of the function, m is the number of features, and x_k is the feature. The logistic regression value $f(z) \in (0, 1)$ indicates whether a loan is in default.

2.2. Decision tree model

A decision tree is an essential classification and regression technique (Shehadeh et al., 2021) based on the estimated probability of occurrence of distinct events. The model applies a decision tree to determine the likelihood that the anticipated value of the net present value will be greater than or equal to zero. It can then assess the project risk and viability of the decision analysis approach. A decision tree is a graphical method used for a probability analysis. It has a tree structure with the core nodes reflecting tests on the attributes, branches indicating the test outputs, and leaf nodes reflecting the categories. Decision trees have the advantages of low computing costs, insensitivity to a lack of intermediate values, and the ability to deal with irrelevant feature data. There are also various disadvantages to decision tree models, such as a difficulty in dealing with missing data, overfitting issues, and a neglection of the relationships between variables.

2.3. XGBoost model

The main concept of the XGBoost model is to fit the prediction residuals using a tree after each feature is split, and the predicted value of the sample is equal to the sum of the sample features (Abedi et al., 2022; Nguyen et al., 2022).

At completion of the process, there are K trained trees, and T is the fundamental tree model. The forecasting results are then derived using Eq. (2).

$$y_i^p = \sum_{k=1}^{K} f_k(x_i), f_k \in T \tag{2}$$

where $f_k(x_i)$ represents the score for every leaf node, and y_i^p is the predicted value.

The objective function is thus represented by Eq. (3).

$$Obj = \sum_i L(y_i^p, y_i) + \sum_k \Omega(f_k) \tag{3}$$

Here, L is the loss function representing the difference between the projected value y_i^p and the actual value y_i. As stated in Eq. (4), Ω is a regularization function used to reduce any overfitting.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{4}$$

where w is the weight of the leaves of each tree, and γ and λ are factors preventing an overfitting.

Eq. (5) yields the optimal weights when the derivative of the objective function is zero.

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{5}$$

where g_i denotes the first-order derivative of the loss function, and h_i represents the second-order derivative. In addition, I_j is the total number of leaf nodes.

The objective function is then obtained using Eq. (6).

$$Obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{6}$$

2.4. LightGBM model

LightGBM is an efficient gradient-boosting decision-tree algorithm. LightGBM uses two methods to improve the training speed (Hao et al., 2022; Liu et al., 2021). The first is gradient-based one-sided sampling (GOSS). GOSS selects data with large gradients from the samples to improve the accuracy and sends most samples with small gradients. The second method involves exclusive-feature bundling (EFB), which can bind exclusive features in such a way that it takes a nonzero value at the same time. In practical applications, because the feature space is sparse, EFB can merge some features of the data to reduce the number of data dimensions.

2.5. LIME model

Ribeiro et al. (2016) developed the LIME algorithm. LIME explains a black box model in such a way that can be better understood. The core principle is to shift the focus to certain parts that can be fitted with a linear model when the model is difficult to explain at the global level. The local sample is sampled and classified using a classification model, the results of which are then obtained. The weight is assigned according to the distance between the sample and the characteristics of the original data. Sampling refers to the perturbation of the characteristics of the original sample.

The core concept of LIME is described below. The explanatory model is defined as $g \in G$. Here, $\pi_x(z)$ is the proximity between z and x. In addition, ξ is an objective function. The L function describes how g approximates an f complex model under a local definition. When $\Omega(g)$ is sufficiently low, the L function obtains the optimal solution. The explanation provided by LIME is shown in Eq. (7).

$$\xi(x) = \text{argmin}_{g \in G} L(f, g, \pi_x(z)) + \Omega(g) \tag{7}$$

Ribeiro et al. (2016) defined the distance of the sample similarity before and after a perturbation. The similarity is calculated using Eq. (8).

$$\pi_x(z) = exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \tag{8}$$

In Eq. (8), $\pi_x(z)$ indicates a Gaussian kernel function used for reflection and measurement. In addition, σ is the kernel width, which is usually 0.75-times the square root of the number of columns in the training data used by the LIME algorithm. Moreover, similar instances can affect the model when the kernel width is small, and examples of large cosine distances can also affect the model when the kernel width is large. The distance function uses the cosine similarity distance, as shown in Eq. (9).

$$D(x,z) = 1 - \frac{x \cdot z}{|x||z|} \tag{9}$$

Thus, the objective function can be rewritten as Eq. (10).

$$\xi(x) = \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \tag{10}$$

where $f(z)$ is the predicted value of the perturbed sample for the original feature (D-dimensional space), and g (z') is the forecasting value for the interpretable feature (D′-dimensional space). z represents the dataset comprising of perturbed samples with the associated labels. The objective function is optimized through a linear regression using the similarity as a weight to produce the outcome.

The principle of LIME is illustrated in Fig. 1. Different colored regions represent sample points with different labels, which are divided into two categories using the model, where the sample point represents the sample to be interpreted. The perturbed sample is selected according to the cosine similarity rule. The linear model is optimized using a linear regression. Thus, a local surrogate model of the complex model of the sample is used for the explanation.

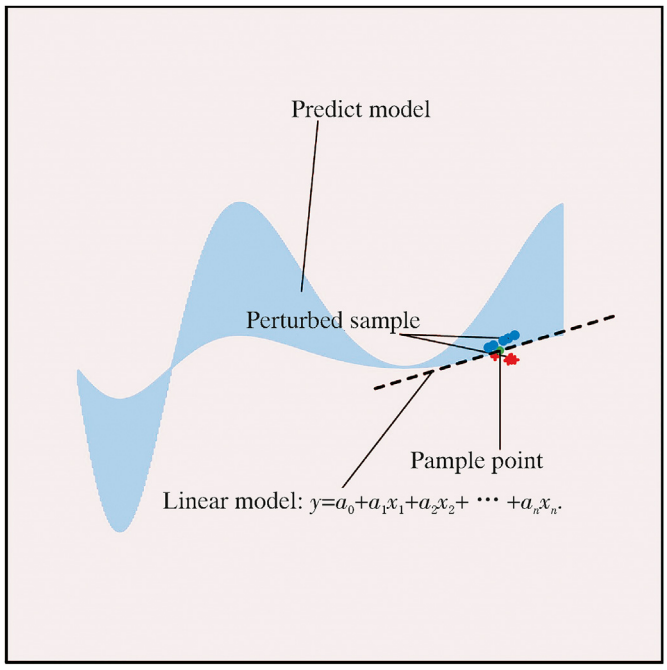

Fig. 1. Example intuition of local interpretable model–agnostic explanations (LIME).

3. Data description and preparation

3.1. Data sources

The experiments were conducted using data provided by Datawhale. The dataset was obtained from the website (https://tianchi.aliyun.com/competition/entrance/531830/information). These variables are listed in Table 1. Table 2 provides partial raw data.

This study utilized a data set consisting of 800,000 entries. In the dataset, "isDefault" is listed as a description of the default state, where 1 is a default and 0 is an on-time repayment. A total of 46 variables, including 15 anonymous variables, were selected. The 15 anonymous variables range from n0 to n14. The statistical results for the partial raw data are presented in Table 3. The number of data points, minimum, maximum, average, standard deviation, skewness, and kurtosis values are listed in Table 3.

3.2. Data preprocessing

First, data pre-processing includes the handling of missing values and

Table 1
Description of some variables used.

Number	Variables	Description
1	loanAmnt	Loan amount
2	term	Loan year
3	interestRate	Loan rate
4	installment	Installment amount
5	grade	Loan grade
6	subGrade	Sub level of loan grade
7	employmentTitle	Title of employment
8	employmentLength	Years of employment
9	homeOwnership	Home ownership status provided by the borrower at the time of registration
10	annualIncome	Annual income
11	verificationStatus	Verify status
12	issueDate	The month in which the loan was made available
13	purpose	The borrower's loan purpose at the time of loan application
14	postCode	The first three digits of the borrower's zip code provided in the loan application
15	regionCode	Area code
16	dti	Debt-to-income ratio
17	delinquency_2 years	Number of default occurrences in the borrower's credit file that has been late for more than 30 days in the last two years
18	ficoRangeLow	The borrower's minimum Fair Isaac Corporation (FICO) at the time of loan issuance
19	ficoRangeHigh	The borrower's maximum FICO at the time of loan issuance
20	openAcc	The amount of unavailable credit in the borrower's credit file
21	pubRec	The number of derogatory public records

Table 2
Partial raw data.

ID	loanAmnt	term	interestRate	installment	grade	n9	n10	n11	n12	n13	n14
0	35,000	5	19.52	917.97	5	2	7	0	0	0	2
1	18,000	5	18.49	461.90	4	4	13	0	0	0	2
2	12,000	5	16.99	298.17	4	3	11	0	0	0	4
3	11,000	3	7.26	340.96	1	6	9	0	0	0	1
4	3,000	3	12.99	101.07	3	7	12	0	0	0	4
5	11,000	3	7.99	344.65	1	2	19	0	0	0	0

Table 3
Statistical results of the partial raw data.

Variable	Number of data	Minimum value	Maximum value	Average value	Standard deviation value	Skewness value	Kurtosis value
term	800,000	3	5	3.48	0.86	1.21	−0.54
employmentTitle	799,999	0	378,351	72,005.35	106,585.60	1.38	0.55
homeOwnership	800,000	0	5	0.61	0.68	0.68	−0.47
annualIncome	800,000	0	10,999,200	76,133.91	68,947.51	46.19	4,902.03
purpose	800,000	0	13	1.75	2.37	1.25	1.25
dti	799,761	−1	999	18.29	11.15	27.30	2,153.80
ficoRangeLow	800,000	630	845	696.20	31.87	1.28	1.65
ficoRangeHigh	800,000	634	850	700.20	31.87	1.28	1.65
pubRec	800,000	0	86	0.21	0.61	13.52	999.90
pubRecBankruptcies	799,595	0	12	0.13	0.38	3.44	20.02
revolUtil	799,469	0	892.30	51.80	24.52	−0.02	0.93
initialListStatus	800,000	0	1	0.42	0.49	0.34	−1.89
n14	759,730	0	30	2.18	1.84	1.49	4.83

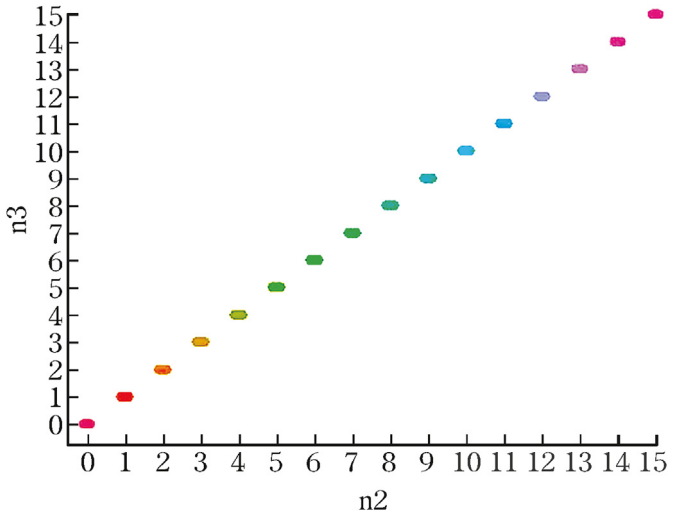

Fig. 2. Distribution diagram of n2 and n3.

outliers. Missing values are handled first. In this study, the median was used to fill in all the missing values. For the time variable, we converted the time format into numeric values. The second method involves treating outliers. We use standard deviation to detect outliers, 3σ (σ represents standard deviation) principle is used to mark and remove the outliers. If the data follow a normal distribution, outliers are the values which are not in the interval of (μ–3σ, μ+3σ). μ represents mean value. The possibility of the value which is not in the interval of (μ–3σ, μ+3σ) is less than 0.3%. After removing the outliers, the final sample size was 612,742. The inputs of the predictive model were based on 612,742 samples.

Invalid features that lack practical meaning increase the operational complexity. Correlations between features increase the difficulty of the analysis. Some variables have poor stability, which significantly interferes with prediction results. Therefore, this study used methods such as deletion, principal component analysis (PCA), feature interaction, and population stability index (PSI) to process the variables in the dataset. To minimize the loss of information in the original dataset, we reduced the number of variables that needed to be analyzed. The datasets were then comprehensively analyzed. The validity of the variables and accuracy of the prediction results can also be improved.

3.2.1. Deletion

As shown in Fig. 2, n2 and n3 are the anonymous variables with the same distribution. Therefore, n2 was retained as a feature. Additionally, since some variables in "grade" are represented by "E, D, D, A, C" and their corresponding variables in "subGrade" are "E2, D2, D3, A4, C2". Thus, the "subGrade" contains all information in the "grade". The "subGrade" was then retained as a feature.

Because the correlation between "interestRate" and "subGrade" is 97.83%, "subGrade" represents "interestRate". Therefore, the "interestRate" variable is removed.

To improve the training speed of the model, the redundant variables must be removed. The indicator used to determine the amount of feature information is the information value (IV), which is calculated using Eq. (11).

$$\mathrm{IV} = \sum_{i=1}^{10}\left(\frac{bad_i}{badtotal} - \frac{good_i}{goodtotal}\right) \times \ln\left(\frac{bad_i}{badtotal} \Big/ \frac{good_i}{goodtotal}\right) \tag{11}$$

When calculating IV, the variables are first divided into bins and then into 10 boxes for training. Here, bad_i represents the number of customers defaulting in any box, $good_i$ represents the number of non-defaulting individuals in a box, *badtotal* represents the total number of defaults, and *goodtotal* is the total number of individuals who did not default. The larger the IV value, the greater the value of the information corresponding to the variables, and the greater the contribution in determining whether the customer is in breach of contract. It is generally believed that when IV is less than 0.02, variables have almost no predictive ability. Based on the IV calculation, the values for employmentLength, postCode, regionCode, delinquency_2years, revolBal, n0, n6, n11, n12, and n13 are all less than 0.02 and are therefore deleted.

3.2.2. PCA

As shown in Table 4, the correlations between some of the variables were high. A principal component analysis (PCA) was used to reduce the

dimensionality of the features (Wang et al., 2021). It was also used for the loanAmnt and installment variables used to construct a new feature, debtpca1-1. Similarly, a PCA was used for the totalAcc and openAcc variables to construct a new feature, accpca1-1. By testing the correlations between the variables, the correlations between n14 and the other variables were found to be less than 0.5. Thus, n14 is not included in the reduced dimensionality. For the anonymous variables n1, n2, n4, n5, n7, n8, n9, and n10, the degree of variance contribution corresponding to each principal component was obtained through a PCA, as shown in Fig. 3. The variance contribution of each principal component is uneven. When the first three principal components were retained, the sum of their variance contribution reached 92% of the sum of the variance contribution of all principal components, as shown in Fig. 4. Therefore, only three principal components were retained to summarize the information. The new features npca1-1, npca1-2, and npca1-3 were constructed to replace the anonymous variables n1, n2, n4, n5, n7, n8, n9, and n10.

3.2.3. Feature interaction

We constructed a new variable "netprofit" based on the actual meaning of the income and income-debt ratio. We set netprofit = annualincome × (1–dti). The "netprofit" variable is primarily used to reflect the income of the borrower under normal repayment conditions. In the feature construction process, the integration of anonymous variables with loan grades can improve the nonlinear modeling ability of the model. The n4 and subGrade variables are discrete numerical features, and the original data contain positive 33-bit integers of 0–32. A new variable "subGradetomeann4" is generated by interacting with the n4 and subGrade variables using Eq. (12).

$$Z = \frac{Y_{X=i}}{\overline{Y}_{X=i}} (i = 1, 2, \cdots, 32) \tag{12}$$

In Eq. (12), $\overline{Y}_{X=i}$ represents the mean of the loan subgrade under the same number of behaviors, Y represents the loan grades of any sample, and Z reflects the different levels of the sample loan grades relative to the same behavioral conditions. This nonlinear interaction mode effectively increases the amount of information. The numbers of behaviors of anonymous variables are often closely related to the loan level. The interplay between mutually affecting factors can significantly improve the learning ability of the model.

Table 4
Correlations between some variables.

Variable	Variable	Correlation
loanAmnt	installment	0.9519
purpose	title	0.8063
subGrade	interestRate	0.9783
grade	interestRate	0.9518
grade	subGrade	0.9747
totalAcc	openAcc	0.6675
n5	n8	0.7949
n2	n7	0.7769
n5	n7	0.5282
n8	n10	0.5873
n2	n4	0.5713
n4	n10	0.5034
n1	n2	0.7610
n1	n9	0.7581
n1	n7	0.5735
n2	n8	0.5298
n4	n9	0.5655
n9	n10	0.5989
n7	n8	0.7229
n7	n10	0.7739
n2	n3	1.0000
n4	n5	0.6735
n8	n9	0.5300
n7	n9	0.7796
n1	n5	0.5320
n2	n9	0.9939
n4	n7	0.6607
n2	n10	0.5964
n1	n4	0.7962

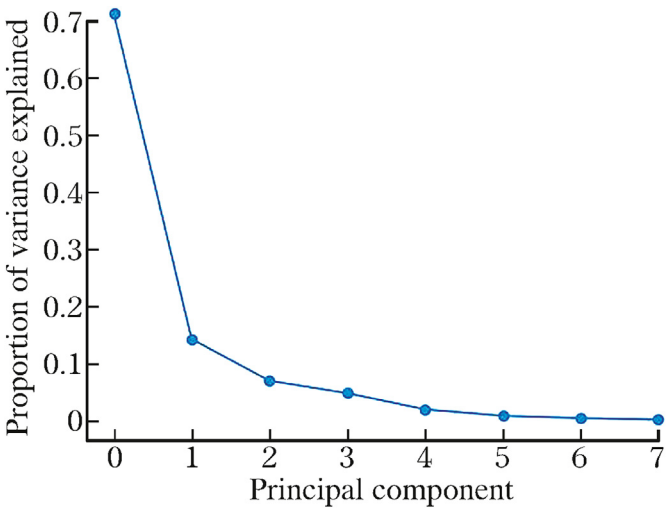

Fig. 3. Degree of variance contribution obtained by principal component analysis.

3.2.4. PSI

Similar to Huang et al. (2022), the population stability index (PSI) measures the deviation between the predicted value of the model and the actual value, as shown in Eq. (13). The data are divided into five parts, of which 80% are the training set and 20% are the test set. Here, A_i is the proportion of variable distribution of the training set, and E_i is the proportion of the variable distribution of the test set.

$$PSI = \sum_{i=1}^{n} (A_i - E_i) \times \ln \frac{A_i}{E_i} \tag{13}$$

The PSI values for each variable are obtained as shown in Table 5.

The stability of the model is extremely high when the PSI is less than 0.1. When the PSI value exceeds 0.25, the model stability is poor. If the PSI value of the net profit is greater than 0.25, the variable net profit is deleted. However, the features obtained by PCA are retained to improve the stability of the feature data. Therefore, npca1-1, npca1-2, and npca1-3 are retained.

Finally, the inputs of the forecasting methods are shown in Table 6.

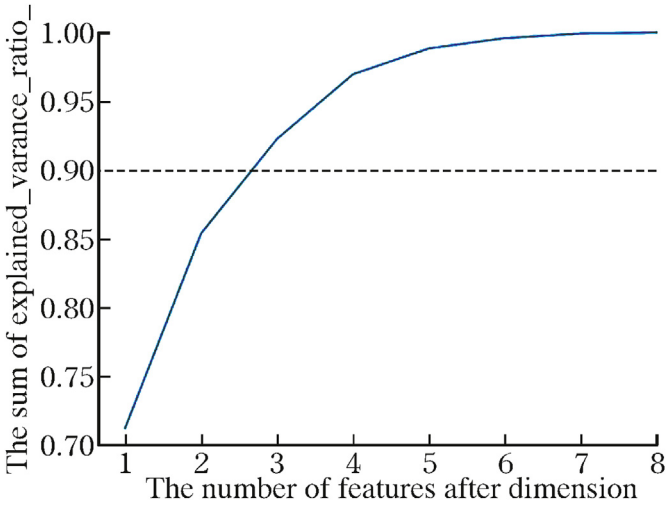

Fig. 4. Sum of the variance contribution.

Table 5
Population stability index (PSI) value of different variables.

Variable	PSI
term	0.000012
subGrade	0.000366
employmentTitle	0.158981
homeOwnership	0.000017
annualIncome	0.040213
verificationStatus	0.000013
issueDate	0.001376
Purpose	0.000026
dti	0.042133
ficoRangeLow	0.000182
ficoRangeHigh	0.000182
pubRec	0.000008
pubRecBankruptcies	0.000016
revolUtil	0.010612
initialListStatus	0.000012
applicationType	0.000000
earliesCreditLine	0.000462
title	0.017561
n14	0.000042
subGradetomeann4	0.005101
npca1-1	0.291531
npca1-2	0.291531
npca1-3	0.291542
debtpca1-1	0.217407
accpca1-1	0.011534
netprofit	0.392611

4. Evaluation indicators

The area under curve (AUC), precision, accuracy, and Kolmogorov-Smirnov (KS) values are used as evaluation indicators. Before introducing the evaluation indicators, a basic confusion matrix is introduced,

Table 6
Inputs for the forecasting method.

Number	Input
1	term
2	subGrade
3	annualIncome
4	homeOwnership
5	employmentTitle
6	verificationStatu
7	issueDate
8	dti
9	purpose
10	ficoRangeHigh
11	ficoRangeLow
12	pubRec
13	pubRecBankruptcie
14	revolUtil
15	title
16	applicationType
17	initialListStatus
18	earliesCreditLine
19	n14
20	subGradetomeann4
21	npca1-1
22	npca1-2
23	npca1-3
24	debtpca1-1
25	accpca1-1
-	-

Table 7
Confusion matrix.

Confusion matrix	Positive (actual)	Negative (actual)
Positive (predicted)	True positive (TP)	False positive (FP)
Negative (predicted)	False negative (FN)	True negative (TN)

Table 8
Evaluation indicators of different models.

Evaluation indicators	Logistic regression	Decision tree	XGBoost	LightGBM
Area under curve	0.7080	0.7034	0.7206	0.7213
Accuracy	0.6555	0.6317	0.8098	0.8104
Precision	0.3110	0.3015	0.5583	0.5751
Kolmogorov-Smirnov	0.3033	0.2985	0.3225	0.3215

as shown in Table 7. True positive (TP) refers to positive data correctly classified by the model. False positive (FP) refers to negative data that the model incorrectly classifies as positive. True negative (TN) indicates that the model properly identifies negative sample data. False negative (FN) indicates positive data that the model incorrectly classifies as negative.

The ordinate of the receiver operating characteristic (ROC) curve is the true positive rate (TPR), whereas the abscissa is the false positive rate (FPR). TPR and FPR are calculated using Eqs. (14) and (15). The area contained by the axis below the ROC curve is the curve value, which is less than or equal to 1. Because the ROC curve is often above the $y = x$ line, the AUC ranges from 0.5 to 1. The closer the AUC is to 1.0, the more reliable the detection algorithm. When it equals 0.5, the authenticity is the lowest, and no application value exists. There are two approaches to compare the performances of the two models. If the ROC curve of Model A completely encapsulates the ROC curve of Model B, we consider Model A to be better than Model B. If the two curves cross, we can make a judgment by comparing the area of the curve enclosed by the ROC and x- and y-axes. The larger the area, the better the model performance.

$$\mathrm{TPR} = \frac{TP}{TP + FN} \tag{14}$$

$$\mathrm{FPR} = \frac{FP}{FP + TN} \tag{15}$$

The precision is only for correct positive-case data, which manifests as the extent to which the predicted positive data are true positive data. The precision is expressed through Eq. (16).

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

The accuracy is the most commonly used classification performance metric, which calculates the number of samples correctly predicted by the classifier. The model is used to classify the test set when predicting, and the accuracy of the model prediction is calculated as the number of

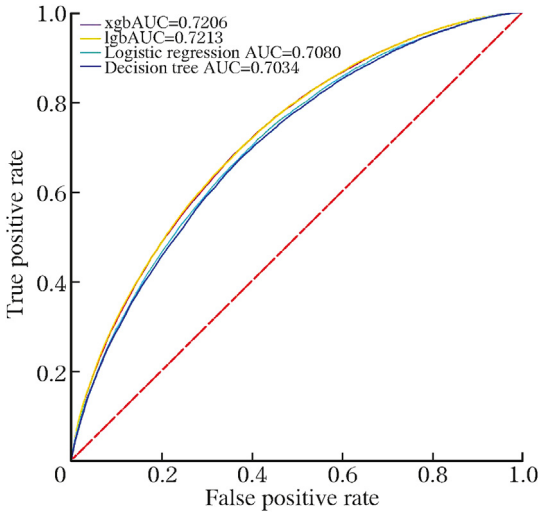

Fig. 5. Area under curve of different prediction models.

correctly predicted samples as a percentage of all samples. The accuracy is expressed through Eq. (17).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{17}$$

The KS value is the highest absolute value of the difference between the TPR and FPR, as shown in Eq. (18), which is used to assess the risk discrimination abilities of the model. The difference in the cumulative distribution between excellent and poor samples is measured using an indicator. The wider the cumulative difference is between the good and poor samples, the higher the KS index and risk discrimination ability of the model. In general, for the model to be considered useful, the KS value must be more than 0.2, and the higher the KS score, the better the model.

$$KS = \max(TPR - FPR) \tag{18}$$

5. Empirical analysis and discussion of the results

In this study, 80% of the data randomly selected from the total sample were used as the training set, and the rest was used as the test set. Logistic regression, decision tree, XGBoost, and LightGBM models were used to obtain the prediction results. The corresponding probability values obtained by the prediction model on the test set are classified, and samples with a probability of greater than 0.5 are judged to be default samples. The corresponding AUC, accuracy, precision, and KS are used to measure the accuracy of the different models (Table 8). The AUC values of the different prediction models are shown in Fig. 5.

As indicated in Table 8, the AUC values of all model indicators exceed 0.7. The AUC of LightGBM is 0.7213, which is better than that of the other models. In addition, the accuracies of LightGBM and XGBoost exceed 0.8. The precisions of LightGBM and XGBoost exceed 0.55. The predictive abilities of LightGBM and XGBoost are significantly better than those of the logistic regression and decision tree models. Moreover, the accuracy and precision of LightGBM are slightly higher than those of XGBoost. Finally, the KS values of XGBoost and LightGBM are slightly higher than those of the logistic regression and decision tree models. Therefore, we conclude that (a) more complex models, such as LightGBM and XGBoost, have a higher predictive ability than traditional logistic regression and decision tree models; (b) the predictive ability of LightGBM is slightly higher than that of the XGBoost model.

The predictive performance of the LightGBM model used in this study is better than that of the other models. In addition, we tested the prediction performance of the LightGBM model through a 5-fold cross-validation. The method randomly divides the original data into five parts, i.e., a test set and four training sets. The accuracy and AUC values of the LightGBM are listed in Table 9.

According to the cross-validation results, the performance of the LightGBM model in the test and training sets was stable. The variance in the AUC and accuracy was small, indicating that the prediction effect of the LightGBM model was relatively stable. Therefore, no obvious overfitting occurred.

6. Interpretability analysis using LIME

The prediction performance of the LightGBM model is stable with high quality for different test sets. However, as an integrated black box model, we cannot intuitively understand the influence of different variables on the prediction results. In this study, the LIME method was used to conduct an explainable analysis of the ensemble learning model.

Several samples were selected from each of the two categories, i.e., default and non-default. An explainable LIME was used to explain single samples. The degree of effect of each variable on the projection results was obtained for a single sample. Furthermore, the average effect of each variable on the predicted outcomes was assessed. Consequently, the degree of the effect of each variable on the total sample projection outcomes was calculated. It is then straightforward to identify the most relevant characteristics of the LightGBM model in generating the prediction results.

6.1. Interpretation of individual samples

First, we choose the default sample as an example. Assuming that the probability of a default is y and the value of each variable is x, we obtain a straight line using LIME, as shown in Eq. (19).

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{19}$$

where a_0 indicates the intercept of the LIME model. That is, a_0 is the intercept of a simple linear model. In addition, a_nx_n represents the degree of effect of variable n on a sample. If the value is positive, variable n promotes the prediction result of the sample to move closer to the default. If the value is negative, variable n promotes the prediction result of the sample to move closer to no default. The larger the absolute value of a_nx_n, the greater the effect of variable n on the forecasting results of the sample.

Table 9
The accuracy and area under curve (AUC) values obtained through a 5-fold cross-validation.

Evaluation indicators	Part 1	Part 2	Part 3	Part 4	Part 5	Mean	Variance
AUC	0.7199	0.7213	0.7216	0.7184	0.7218	0.7206	0.0013
Accuracy	0.8068	0.8080	0.8100	0.8087	0.8084	0.8085	0.0010

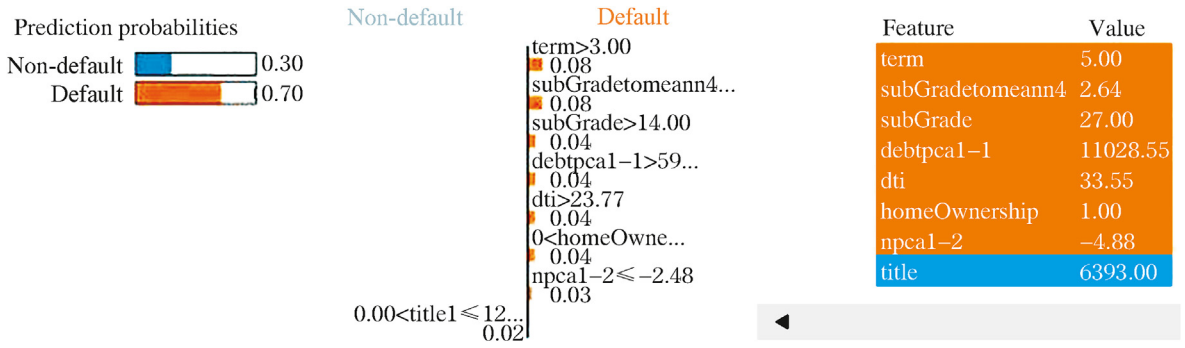

Fig. 6. Results of a sample with a high default rate obtained using local interpretable model-agnostic explanations (LIME).

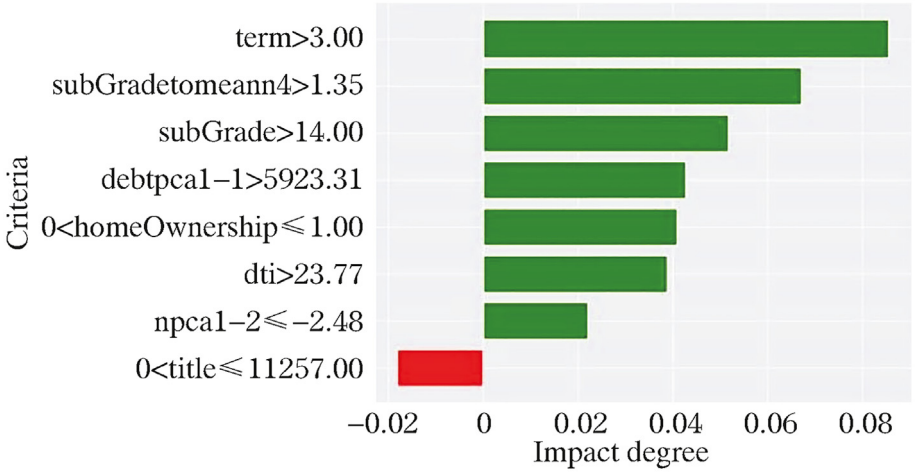

Fig. 7. Local explanation for default.

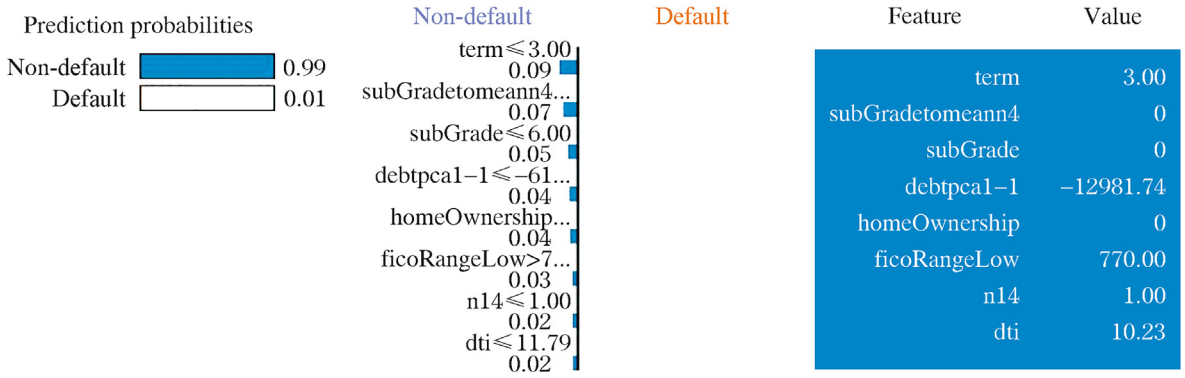

Fig. 8. Results of a sample with a low default rate obtained using model-lagnostic explanations (LIME).

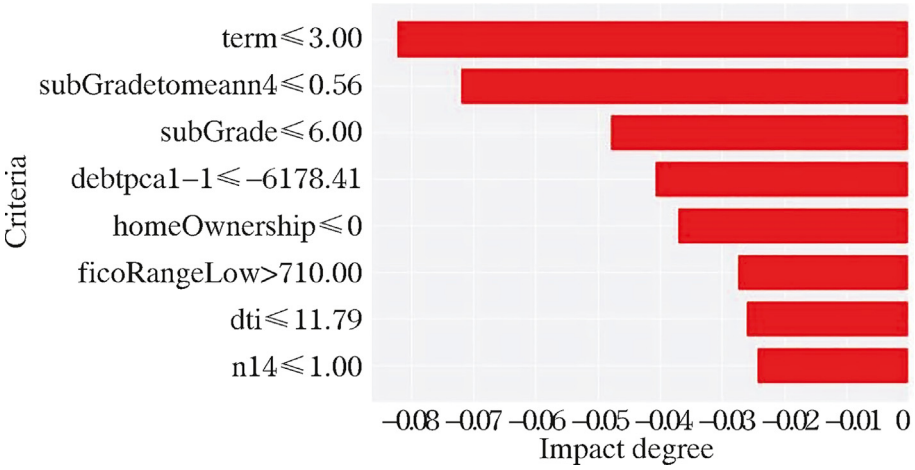

Fig. 9. Local explanation for non-default.

6.1.1. Sample with a high default rate

For a sample with a high default rate, the 117,949th sample in the test set is considered as an example. The sample is interpreted based on the LIME interpretation method, the results of which are shown in Fig. 6.

We can obtain the following conclusions from Fig. 6:

(1) The upper-left corner of Fig. 6 shows the probability of the classification results obtained using the prediction model for the sample. The probability of the prediction model dividing the sample into a default sample is 0.70, and the probability of no default is 0.30.

(2) The data frame on the right side of Fig. 6 displays the features and their input data. The features in the orange background lead to an increase in the probability of a default for the sample. The features in the blue background lead to a decrease in the probability of default for the sample. They are sorted sequentially according to the degree of influence of the characteristic variables.

(3) The middle part of Fig. 6 shows the degree and direction of influence of the corresponding feature variables on the predicted probability of the model. The variables are arranged in descending order based

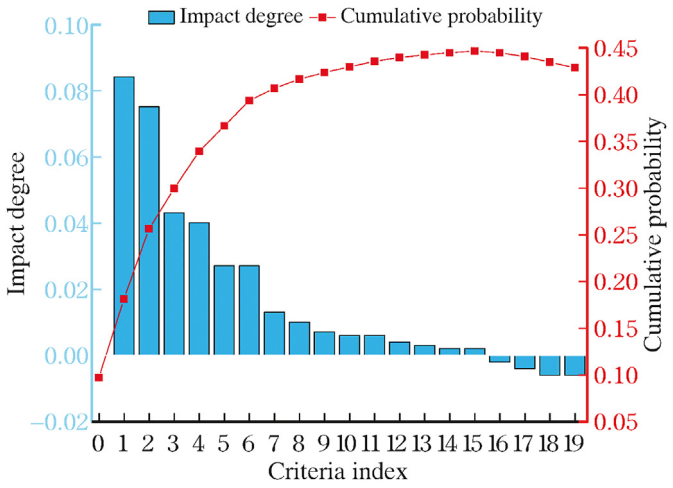

Fig. 10. Degree of impact and cumulative probability for default samples.

on the absolute value of their degree of influence. The orange bar represents the feature variable increasing the probability of a default. The blue bar represents a feature variable reducing the probability of a default. The bar length indicates the degree of influence.

By further enlarging the weight graph in the middle part of Fig. 6, the degree of influence of the different variables is intuitively shown in Fig. 7.

In Fig. 7, the ordinates represent the characteristic variables and their value ranges. The abscissa represents the corresponding weights. Green bars indicate that the feature variables increase the probability of a default. The red bars indicate that the feature variable reduces the probability of a default. In the sample, eight variables exhibited higher absolute values. The term, subGradetomeann4, subGrade, debtpca1-1, homeOwnership, dti, and npca1-2 variables increase the probability of a default, whereas the variable title decreases it.

6.1.2. *Sample with a low default rate*

For a sample with a low default rate, the 85,312th sample in the test set is considered as an example. The sample is interpreted based on the LIME interpretation method, the results of which are shown in Fig. 8. The probability of the prediction model categorizing a sample as a default is 0.01, whereas the probability of a non-default categorization is 0.99.

By further enlarging the weight graph in the middle part of Fig. 9, the degrees of influence of different variables can be intuitively shown. The term, subGradetomeann4, subGrade, debtpca1-1, homeOwnership,

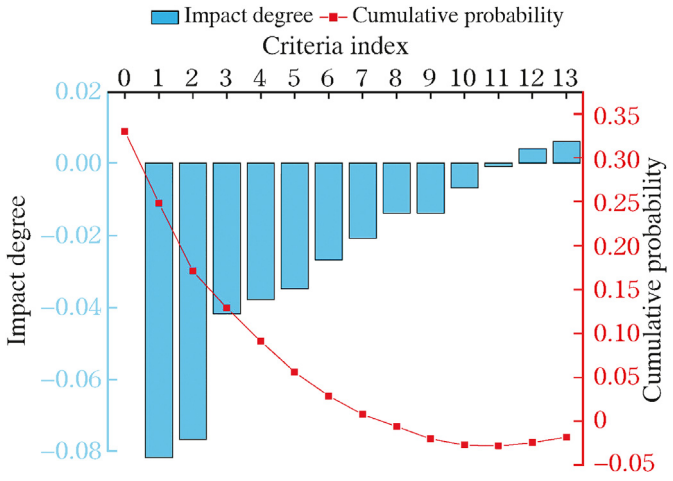

Fig. 11. Degree of impact and cumulative probability for the non-default samples.

Table 10
Notations used in Fig. 10.

Numeric ordinal number	The meaning of the numeric ordinal number
0	intercept
1	term > 3.00
2	subGradetomeann4 > 1.35
3	subGrade > 14.00
4	0.00 < homeOwnership ≤ 1.00
5	debtpca1-1 > 5,923.31
6	dti > 23.77
7	n14 > 3.00
8	title ≤ 0.00
9	npca1-2 ≤ −2.48
10	npca1-3 ≤ −1.77
11	annualIncome ≤ 45,000.00
12	ficoRangeLow ≤ 670.00
13	purpose > 4.00
14	ficoRangeHigh ≤ 674.00
15	2.00 < n14 ≤ 3.00
16	ficoRangeHigh > 714.00
17	ficoRangeLow > 710.00
18	pubRec ≤ 0.00
19	0.00 < title ≤ 11,257.00

Table 11
Notations used in Fig. 11.

Numeric ordinal number	The meaning of the numeric ordinal number
0	intercept
1	term ≤ 3.00
2	subGradetomeann4 ≤ 0.56
3	subGrade ≤ 6.00
4	homeOwnership ≤ 0.00
5	debtpca1-1 ≤ −6,178.41
6	ficoRangeLow > 710.00
7	dti ≤ 17.46
8	n14 ≤ 1.00
9	npca1-2 > 2.18
10	0.00 < title ≤ 11,257.00
11	62,000.00 < annualIncome ≤ 87,500.00
12	purpose > 4.00
13	title ≤ 0.00

ficoRangeLow, dti, and n14 variables reduce the probability of a default. These eight variables prompt the model to classify a sample as a non-default.

6.2. *Interpretation of all samples*

To further explore the impact of different feature variables on default and non-default samples, we selected samples with high and low default probabilities. We analyzed and judged the impact of different feature variables according to the integration weights of the feature variables associated with different samples and degrees of impact. Figs. 10 and 11

Table 12
Features and their importance of non-default.

Criteria	Impact degree
term ≤ 3.00	−0.082
subGradetomeann4 ≤ 0.56	−0.077
subGrade ≤ 6.00	−0.042
homeOwnership ≤ 0.00	−0.038
debtpca1-1 ≤ −6178.41	−0.035
ficoRangeLow > 710.00	−0.027
dti ≤ 17.46	−0.021
n14 ≤ 1.00	−0.014
npca1-2 > 2.18	−0.014
0.00 < title ≤ 11,257.00	−0.007
62000.00 < annualIncome ≤ 87,500.00	−0.001
purpose > 4.00	0.004
title ≤ 0.00	0.006

Table 13
Features and their importance of default.

Criteria	Impact degree
term >3.00	0.084
subGradetomeann4 > 1.35	0.075
subGrade >14.00	0.043
0.00 < homeOwnership ≤ 1.00	0.040
debtpca1-1 > 5,923.31	0.027
dti > 23.77	0.027
n14 > 3.00	0.013
title ≤ 0.00	0.010
npca1-2 ≤ −2.48	0.007
npca1-3 ≤ −1.77	0.006
annualIncome ≤ 45,000.00	0.006
ficoRangeLow ≤ 670.00	0.004
purpose > 4.00	0.003
ficoRangeHigh ≤ 674.00	0.002
2.00 < n14 ≤ 3.00	0.002
ficoRangeHigh > 714.00	−0.002
ficoRangeLow > 710.00	−0.004
pubRec ≤ 0.00	−0.006
0.00 < title ≤ 11,257.00	−0.006

show the impacts of the different variables on the prediction results. Tables 10 and 11 list the notations used in Figs. 10 and 11.

According to Figs. 10 and 11 and Tables 12 and 13, among the important influencing variables of the default and non-default samples, the term, subGradetomeann4, subGrade, dti, homeOwnership, and ficoRange have an influence. Thus, term, subGradetomeann4, subGrade, dti, homeOwnership, and ficoRange play important roles in determining whether the sample loan defaults.

6.3. Management implications and discussions

When complex machine learning models are used to predict a loan default, LIME can be used to add an interpretability analysis to the model results. Managers can understand the predictive principles of complex models in depth. Managers no longer accept the conclusions of a model passively, but rather under the premise of ensuring the credibility of the model. Managers actively make decisions and choose whether to use model recommendations for achieving more accurate and effective decisions.

LIME offers significant advantages in terms of versatility, pertinence, and plasticity. It can be used to select representative samples for analysis according to different requirements. Taking the experiment conducted in this study as an example, we analyzed the variable importance of different probability intervals separately by stratifying the default probability. By contrast, the results of the LIME method are relatively easy to interpret.

However, LIME has certain limitations. First, owing to its local interpretability, not all samples can be interpreted. Moreover, the time cost in applying LIME is high. LIME is used for training and explaining a model locally. In the selection of the perturbation data range, the similarity is difficult to define, which also leads to poor stability. Different sample data of a disturbance frequently lead to different conclusions. In addition, owing to the diversity of the prediction models, the output results and formats are extremely different. Thus, the prediction model often does not directly build an explanatory model, which brings about greater challenges to the interpretation. In the experiment conducted in this study, the output format of the predictive model used by LightGBM was adjusted to match the need to build explanatory models, which greatly reduced the simplicity.

7. Conclusions and future research

In this study, logistic regression, decision tree, XGBoost, and LightGBM models were applied to the prediction of loan default. The prediction abilities of the XGBoost and LightGBM models were better than those of the logistic regression and decisiontree models.

For the LightGBM prediction model, LIME was used to explain the prediction results. It was concluded that the loan term, loan grade, home ownership status provided by the borrower at the time of registration, loan amount, installment amount, debt-to-income ratio and credit score of the borrower are significant factors influencing a personal loan default. Thus, explaining the model and making the prediction rules explicit using LIME can boost user faith in the model.

Prediction and explainable models were combined to predict the personal loan defaults in our study. The predictive ability of machine learning models will be further improved, and their complexity will also increase. Explainable models can effectively help with model recognition and judging whether the model is consistent with reality. Therefore, a combination of prediction and explainable models can be used in other forecasting problems (Wakjira et al., 2022). In the future, other explainable models can also be applied to estimate the overall influence of the features used (Lim et al., 2021).

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

During the process of writing this paper, we encountered many difficulties and received significant assistance. Thus, we thank the Tianchi platform in particular for access to the data used, and we would like to thank the Wuhan University of Technology for their financial support. This research was partially supported by Fundamental Research Funds for the Central Universities (WUT: 2022IVA067).

References

Abedi, R., Costache, R., Shafizadeh-Moghadam, H., et al., 2022. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto Int. 37 (19), 5479–5496.
Butaru, F., Chen, Q., Clark, B., et al., 2016. Risk and risk management in the credit card industry. J. Bank. Finance 72 (Nov.), 218–239.
Chen, Z., Jin, S., Liu, R., et al., 2021. A deep non-negative matrix factorization model for big data representation learning. Front. Neurorob. 15 (Jul.), 701194.
Dalmau, R., Ballerini, F., Naessens, H., et al., 2021. An explainable machine learning approach to improve take-off time predictions. J. Air Transport. Manag. 95 (Aug.), 102090.
Deng, T., 2019. Study of the prediction of micro-loan default based on Logit model. In: 2019 International Conference on Economic Management and Model Engineering (ICEMME). IEEE, pp. 260–264.
Emekter, R., Tu, Y., Jirasakuldech, B., et al., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 47 (1), 54–70.
Fitzpatrick, T., Mues, C., 2016. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. Eur. J. Oper. Res. 249 (2), 427–439.
Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., et al., 2022. Predictably unequal? The effects of machine learning on credit markets. J. Finance 77 (1), 5–47.
Hao, X., Zhang, Z., Xu, Q., et al., 2022. Prediction of f-CaO content in cement clinker: a novel prediction method based on LightGBM and Bayesian optimization. Chemometr. Intell. Lab. Syst. 220 (Jan.), 104461.
Huang, Y., Rameezdeen, R., Chow, C.W., et al., 2022. Monitoring the health status of water mains using a scorecard modelling approach. Water Supply 22 (3), 3114–3124.
Kim, A., Cho, S.B., 2019. An ensemble semi-supervised learning method for predicting defaults in social lending. Eng. Appl. Artif. Intell. 81 (May), 193–199.
Li, M., Yan, C., Liu, W., 2021. The network loan risk prediction model based on Convolutional neural network and Stacking fusion model. Appl. Soft Comput. 113 (Dec.), 107961.
Lim, B., Arık, S.Ö., Loeff, N., et al., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int. J. Forecast. 37 (4), 1748–1764.
Liu, J., Gao, Y., Hu, F., 2021. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. Comput. Secur. 106 (Jul.), 102289.
Nguyen, V.Q., Tran, V.L., Nguyen, D.D., et al., 2022. Novel hybrid MFO-XGBoost model for predicting the racking ratio of the rectangular tunnels subjected to seismic loading. Transp. Geotech. 37 (Nov.), 100878.
Onchis, D.M., Gillich, G.R., 2021. Stable and explainable deep learning damage prediction for prismatic cantilever steel beam. Comput. Ind. 125 (Feb.), 103359.
Peng, L., Wang, L., Xia, D., et al., 2022. Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. Energy 238 (Jan.), 121756.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1135–1144.
Sadhwani, A., Giesecke, K., Sirignano, J., 2021. Deep learning for mortgage risk. J. Finance Econom. 19 (2), 313–368.
Shehadeh, A., Alshboul, O., Al Mamlook, R.E., et al., 2021. Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression. Autom. ConStruct. 129 (Sep.), 103827.
Song, X., Liu, X., Liu, F., et al., 2021. Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. Int. J. Med. Inf. 151 (Jul.), 104484.
Sun, D., Xu, J., Wen, H., et al., 2021. Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. Eng. Geol. 281 (Feb.), 105972.
Wakjira, T.G., Ibrahim, M., Ebead, U., et al., 2022. Explainable machine learning model and reliability analysis for flexural capacity prediction of RC beams strengthened in flexure with FRCM. Eng. Struct. 255 (Mar.), 113903.
Wang, L., Wang, S., Yuan, Z., et al., 2021. Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu index: taking Beijing city as an example. Data Sci. Manag. 2 (Jun.), 12–19.
Wu, B., Wang, L., Zeng, Y.R., 2022. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. Energy 252 (Aug.), 123990.
Xia, Y., Liu, C., Li, Y., et al., 2017a. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst. Appl. 78 (Jul.), 225–241.
Xia, Y., Liu, C., Liu, N., 2017b. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron. Commer. Res. Appl. 24 (Jul.–Aug.), 30–49.
Zhou, B., Yang, G., Shi, Z., et al., 2022. Interpretable temporal attention network for COVID-19 forecasting. Appl. Soft Comput. 120 (May), 108691.
Zhou, J., Li, W., Wang, J., et al., 2019. Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A 534 (Nov.), 122370.