Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 2, 2024

Enhancing Algorithm Interpretability and Accuracy with Borderline-SMOTE and Bayesian Optimization

David Davis , Lisa Rodriguez

University of California,Santa Barbara,University of California,Santa Barbara

SDavidof1@gmail.com,Lisa77rod@gmail.com

Abstract: In recent years, machine learning technologies have made significant advancements across various domains. However, in sectors such as credit scoring and healthcare, the limited interpretability of algorithms has led to concerns, especially for tasks that require high security, occasionally resulting in suboptimal decisions by organizations. Enhancing both the accuracy and interpretability of algorithmic models is crucial for optimal decision-making. To address this, the Borderline-SMOTE method is proposed for data balancing, incorporating a control factor, posFac, to finely adjust the randomness in generating new samples. Additionally, a Bayesian optimization approach is utilized to refine the performance of the XGBoost model. SHAP values are then employed to interpret and analyze the predictive results of the optimized XGBoost model, identifying the most impactful features and the characteristics of input features. This approach not only improves the predictive accuracy of the XGBoost model but also its interpretability, paving the way for broader research and application in various fields.

Keywords: Credit score; Machine learning; Borderline-SMOTE; Bayesian optimization; XGBoost; SHAP tree interpreter.

1. Introduction

With the continuous maturity of machine learning technology, it has made great achievements in many fields such as medical treatment, face recognition, speech recognition, and has been successfully applied in people's life and work. At the same time, machine learning has also made great contributions to our human fight against the epidemic.

When machine learning completes many tasks that human beings cannot do for us, people continue to question and dispute its decision because of the lack of model interpretability. In the application process of machine learning, by inputting variables and then feeding back a decision result, not only ordinary users and decision makers cannot understand the decision basis behind it, but even developers cannot accurately explain it. The lack of interpretable algorithm model will lead people to question its reliability, especially in the face of life safety and property safety, a more reliable algorithm model is needed to solve the problems faced by people.

The application of algorithmic model in credit scoring will also be unable to eliminate security risks in time and further develop due to the lack of interpretability.

Improving the interpretability of the algorithm model not only increases its transparency, but also builds a bridge of trust between human beings and decision models. In recent years, the academic community has also put forward a variety of methods of model interpretability. Rodr í guez-P é rez Raquel and others used shapley value to explain machine learning model and applied it to compound potency and multi-objective activity prediction. Zhang Hong and others used the LIME algorithm to explain and analyze the stochastic forest algorithm model, providing a basis for decision-making in the practical application of the algorithm model. Ramprasaath R. Selvaraju and others combined Grad-CAM with existing fine-grained visualization to create a high-resolution guided Grad-CAM for classification of images, image description and visual Q&A, providing "visual interpretation" for a large class of model decisions of convolutional neural network (CNN), making it more transparent and interpretable. Kyunghyun Cho and others use the method of Attention Mechanism to interpret and analyze RNN coder-decoder (a new neural network architecture), which improves the performance and interpretability of statistical machine translation system. By improving the attention mechanism, Zhou Yong and others constructed a method that can extract component features using the attention mask, generate interpretable weights for feature components, and propose the ternary loss of significant components, which improves the recognition accuracy and interpretability of pedestrian recognition methods. However, different researchers have different ideas, ideas, and focuses on solving problems, which leads to different results of the model's interpretability. Therefore, it is necessary to further study the interpretation, summary and induction of the algorithm. Wen Chunhui and others have successfully applied machine learning to credit assessment and early warning in the financial field. Xu Liang and others introduced semi-supervised multi-layer convolution kernel learning algorithm into the credit scoring model, which enhanced the distinction between customer value and consumption preference, and further improved the prediction performance of the credit scoring model. It can be seen that the application of AI in credit scoring is developing rapidly at present. However, China's research on model interpretation is still at an early stage, unable to meet the user's understanding of the system, as well as the overall advantages and disadvantages of the system. Model interpretability evaluation needs to be seen from the perspective of interpretability, application scenarios, interpretation models, users, etc. This paper will start with the model to improve the prediction performance of the model and enhance the interpretability of the model.

The model is mainly divided into white box model and black box model . The white box model includes: decision tree model, rule model, linear model, etc. Black box models include: neural network, support vector machine, etc. Because credit scoring attaches great importance to accuracy and security, the algorithm mechanism of black-box model cannot provide a strong explanation basis for users, thus limiting the wide application of black-box algorithm model in the field of credit scoring. At present, the interpretation of the algorithm model is mainly focused on the white box model to strengthen the decision-making basis of the algorithm model in practical applications.

In this regard, this paper proposes to use Bayesian algorithm to optimize the XGBoost (eXtreme Gradient Boosting) algorithm model, improve the performance of the algorithm model, and use SHAP to interpret and analyze the output results of XGBoost algorithm model, enhance the interpretability of the algorithm model, with a view to further promoting the reliability of machine learning in the field of credit scoring.

2. Data processing and feature analysis

2.1. Data preprocessing

In the process of data processing, if you want to further explore the null value and outlier value of the feature data set, and explore the distribution of each feature column, you will find that there are huge differences in the size of the feature columns displayed in the data distribution. To solve this problem, the StandardScaler function in sklearn library will be used to normalize and standardize the feature data set to ensure that all features have the same standard scale. For the missing characteristic data in the data set part, the experimental comparison of mean interpolation, random forest interpolation, deletion of missing values and other methods shows that the method of random forest interpolation proposed by Qian Chao and others has the best effect on the processing of missing data in this experiment. During data analysis, it is found that the data set has serious imbalance, which will lead to low recall rate and poor prediction effect for minority samples. For data imbalance processing, several oversampling and undersampling algorithms are studied, such as synthetic minority oversampling technology (SMOTE) , error undersampling algorithm , adaptive synthetic sampling (ADASYN) , etc. On the basis of the experimental results, Chen Yu and others used the Borderline-SMOTE algorithm improved based on SMOTE to deal with the data imbalance, making the classification effect of the convolution neural network model superior to that without Borderline-SMOTE algorithm. Unlike simple duplicate sample sampling, Borderline-SMOTE introduces the influence factor posFac to fine control the random number when synthesizing new samples, which can alleviate the limitations of the SMOTE algorithm when synthesizing a small number of samples. During the validation, only the training set is processed by Borderline-SMOTE, so as to avoid the data over-fitting problem.

2.2. Data feature analysis

In order to better understand the 11 characteristic variables, we carried out the correlation analysis experiment of the characteristic variables. All the paired correlation coefficients are represented by color coding diagram, as shown in the figure below. Black represents positive correlation, gray represents negative correlation, and the darker the color is, the greater the correlation coefficient is. X0 to X10 represent the characteristic values SeriousDlqin2yrs to NumberOfDependents respectively. From the thermodynamic diagram, we can see that the correlation coefficient of each eigenvalue is relatively low. The experiment uses correlation analysis as an exploratory measure, but does not use correlation as the standard for feature selection. Regardless of the size of the correlation coefficient of the two features, data classification can still be affected.

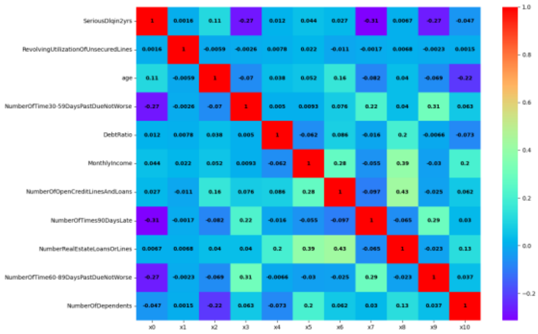

Figure 1. Thermal diagram of characteristic correlation coefficient

At the same time, this paper analyzes the IV diagram of data characteristic variables. It can be seen that the IV value of DebtRatio (x4), MonthlyIncome (x5), Number-Of OpenCreditLinesAndLoans (x6), NumberRealEstateLoasOr Lines (x8) and NumberOfDependencies (x10) variables is significantly lower. The IV value is less than 0.02, which means that the characteristic variables have no value. Ref er to the IV value and eliminate the DebtRatio (x4) varia ble. The characteristic variable IV diagram is shown in th e figure below.

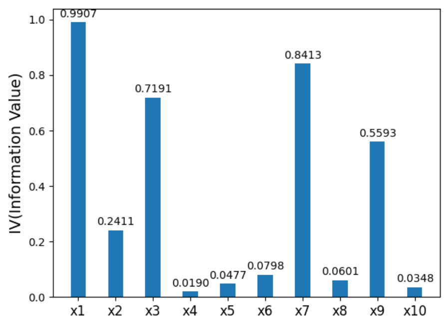

Figure 2. Characteristic variable IV diagram

3. Algorithm model and interpreter

In the previous study, four different algorithm models were used for comparative experiments, such as Decision Tree, Logical Regression, Back Propagation Neural Network (BPNN) and XGBoost. The results of this study show that XGBoost is superior to other algorithm models. In this study, considering the prediction performance of XGBoost algorithm and the calculation cost of model development, this paper focuses on using standard procedures of Bayesian algorithm to improve XGBoost algorithm.

3.1. XGBoost

XGBoost (its full name is eXtreme Gradient Boosting) algorithm model is a boosting algorithm based on CART (its full name is Classification And Regression Tree) proposed by Dr. Chen Tianqi in 2016. Since the launch of the model, it has made remarkable achievements in various model competitions, and has been widely concerned with its excellent learning ability and efficient training speed.

CART is a regression tree, which is the most basic component module of XGBoost. CART constructs the classification tree according to the training data and features, and then judges the prediction results of each data, and uses Gini index to calculate the gain of the construction tree to realize the feature selection of the construction tree.

D is the data set, p_k k is the classification category, K is the k probability in D, is the number of categories.

Formula (1) is Gini index formula:

$$Gini(D) = \sum_{k=1}^{K} p_k(1-p_k) \tag{1}$$

, D_2 respectively $Gini(D_1)$ represent the A data set A of the D middle feature and the A data set of the non-feature, and represent the Gini index of the feature data set.

Formula (2):

$$Gini(D,A) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \tag{2}$$

However, only one CART can not effectively predict the results. Therefore, based on CART, boosting tree is used to combine multiple trees for prediction.

By calculating the residual, the formula is:

$$r_{mi} = y_i - f_{(m-1)}(x_i), i = 1,2,...,n \tag{3}$$

Fitting residual r_{mi} learning regression tree, get $T(x{:}\theta_m)$.

Update formula:

$$f_m(x) = f_{(m-1)}(x) \qquad + T(x{:}\theta_m) \tag{4}$$

Finally, the regression lifting tree is obtained:

$$f_M(x) = \sum\nolimits_{m=1}^{M} T(x{:}\theta_m) \tag{5}$$

For XGBoost, first, define an objective function:

$$Obj(t) = \sum_{i=1}^{n} L(y_i, \hat{y}^{t-1} + f_t(x_i)) + \Omega(f_t) + constant \tag{6}$$

T $constant$ Is a constant, w_j representing the j weight of the second leaf node, and is the number of leaf nodes. The regular term $\Omega(f_t)$ formula is as follows:

$$\Omega(f_t) \quad = \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T} w_j^2 \tag{7}$$

Formula (6) is expanded by Taylor expansion:

$$Obj(t) \approx \sum\nolimits_{i=1}^{n}\left[L(y_i, \hat{y}^{t-1}) + g_i f_i(x_i) + \frac{1}{2}h_i f_i^2\right] + \Omega(f_t) + constant \tag{8}$$

The objective function is simplified as follows:

$$Obj(t) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T + C \tag{9}$$

w_j When the simplified objective function is introduced, it can be seen that the objective function of XGBoost can be customized, and only its first and second derivatives are needed in the calculation process. After the formula is simplified, the gain brought by the selected feature is calculated, and then the appropriate split feature is selected.

Gain's formula is as follows:

$$\begin{aligned} Gain(\phi) &= Gain(before) - Gain(after) \\ &= \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R^2)^2}{H_l+H_R+\lambda}\right] - \gamma \end{aligned} \tag{10}$$

The definition of Gain is the Obj of two sub-nodes after the Obj of a single node minus the sharding. When the value of Gain is positive, it is worth sharding. When the value of the left half of Gain is greater than g, it is also worth sharding.

XGBoost's traversal of sample data can be transformed into traversal on leaf nodes, which greatly improves the efficiency of the algorithm, and the feature selection of data and node segmentation can also be performed in parallel, which can greatly reduce the calculation time. Therefore, XGBoost algorithm is superior to traditional GBDT algorithm in terms of performance and prediction accuracy.

3.2. Bayesian algorithm

Bayesian optimization, for any machine learning algorithm, it is necessary to undergo initialization of hyperparameters before learning. Few algorithms are independent of hyperparameters. In addition, the prediction accuracy of XGBoost machine learning algorithm is deeply affected by several super parameters, including the number and depth of trees. Therefore, proper adjustment of these parameters is crucial to improve the accuracy of the learning model. However, hyperparametric optimization is a process that needs to select a group of optimal hyperparameters; Therefore, the gradient descent algorithm used to optimize general parameters cannot be directly applied to this process. As a Bayesian method, its optimization aims to search the global extremum of functions (especially high-dimensional nonlinear non-convex functions).

Bayesian optimization involves two core steps: prior function (PF) and acquisition function (AC). The former mainly adopts Gaussian process regression, while the latter combines multiple methods, such as EI, PI and UCB. In addition, AC can also achieve the balance between development and exploration. There are three types of acquisition functions: upper confidence bound (UCB), maximum improvement rate (PI) and expected improvement (EI). The PI acquisition function is used in this study, as follows:

$$PI(x) = P(f(x) \geq f(x^+) + \nu) = \Phi\left(\frac{\mu(x) - f(x^+ - \nu)}{\sigma(x)}\right) \tag{11}$$

Super parameters are μ used to adjust the balance before exploration and production. $\mu = 0$ Represents the $f(x^+)$ trend of convergence at, $\Phi(\cdot)$ represents the cumulative distribution function of standard normal, and $f(x^+)$ represents the current maximum. X is the observation point, σ but the standard deviation of all observation points. The Bayesian optimization process is shown in the figure below.

Bayesian optimization sets more favorable parameters for the model according to the previous results by continuously updating the probability model. By referring to the previous prediction, Bayesian method can save a lot of performance consumption and time when performing the next set of super parameters. The combination of Bayesian optimization and XGBoost can effectively reduce the over-fitting and calculation workload of the algorithm, and greatly improve the efficiency of the model algorithm by optimizing the parameters. The model can predict quickly and accurately to achieve the expected purpose.

In the experimental study, we carried out parameter optimization comparative experiments on four algorithms: Grid search , Random search , Particle Swarm Optimization (PSO) [18], and Bayesian Optimization (BO) . For the optimization performance, first of all, the effect of Grid search and Random search is average, second of all, the optimization efficiency of the Patch Swarm Optimization algorithm is not particularly high, and needs enough initial sample points. Finally, the Bayesian Optimization algorithm is selected for the final experiment. Compared with other algorithms, its optimization efficiency is higher, and the model performance improvement effect is the best. The experiment uses 1000 matches, adds the logarithm range with the best performance to Bayesian Optimization, finds the optimized logarithm set through the logarithm space, and then applies the logarithm with the best performance to the final model. After optimization, the maximum depth of the tree, the minimum weight of the leaf node and the proportion of the sub-samples used for training the model in the whole sample set are 13, 2 and 0.1 respectively. Although this tuning method takes a long time and requires multiple searches in the parameter range, the model can make more accurate predictions after tuning. Comparison of optimization parameters is shown in the following table:

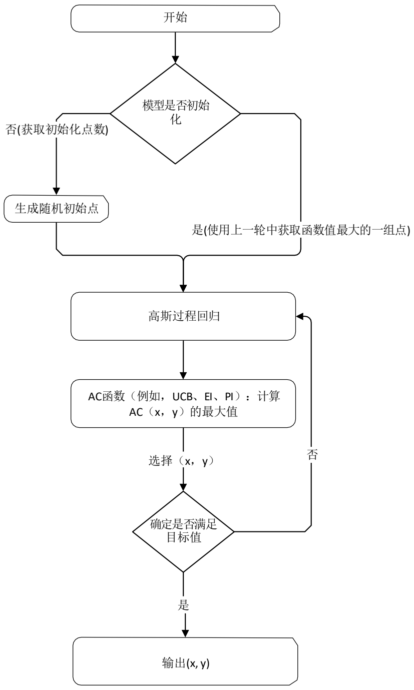

Figure 3. Bayesian optimization flow chart

Table 1. Comparison Table of Hyperparametric Optimization Parameters

attribute	Parameters before optimization	Parameters after optimization
colsample_by_tree	0.9	0.97
colsample_bylevel	1	0.73
gamma	0.01	3.35
learning_rate	0.3	0.18
max_delta_step	5.0	6.0
max_depth	10	13
min_child_weight	2	5.0
n_estimators	10	66.0
reg_alpha	0	0.48
reg_lambda	0	106.15
scale_pos_weight	1	1.02
subsample	0.1	0.48

3.3. Model performance evaluation index

The model performance is evaluated using a variety of standard evaluation indicators. Common performance indicators include Precision, Recall, Accuracy and receiver operating characteristic curve (ROC curve). In this study, precision is the ratio of predicted value to total actual value; Recall rate is the ratio of correct predicted value to all actual values; Accuracy is the most intuitive metric, which is described as the ratio of the correctly marked value to the entire value pool. ROC curve is a commonly used binary classifier tool. ROC curve can draw the curve between true positive rate (TPR) and false positive rate (FPR). In this study, the ROC curve, precision, recall and accuracy will be used to evaluate the performance of the model. The method for calculating the evaluation data is based on the second classification. The positive example is Positive, the negative example is Negative, the correct prediction is True, and the error prediction is False. The summarized results are FP, FN, TP, TN, which are represented by the confusion matrix, as shown in the following table.

Table 2. Confusion matrix

Actual class \ Forecast	Positive example	Counterexample
Positive example	True Positive (TP)	False Positive (FP)
Counterexample	False Negative (FN)	True Negative (TN)

Precision calculation formula is:

$$\frac{TP}{TP+FP} \tag{12}$$

The calculation formula of Recall is:

$$\frac{TP}{TP+FN} \tag{13}$$

Accuracy calculation formula is:

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

3.4. The interpretability of SHAP model

SHAP is a widely used method to explain the machine learning model based on cooperative game theory. The SHAP interpretation method is to calculate the Shapley value from the cooperative game theory. The characteristic value of the data instance is the participant in the cooperation. However, the contribution of each participant to the cooperation is different. The Shapley value ensures that each participant will obtain a fair share according to the contribution size. The Shapley value of the characteristic value can be calculated by the following equation:

$$\phi_j = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(f(S\cup\{x_j\}) - f(S)\right) \tag{15}$$

Including:

S is the feature subset used in the model,

P is the number of features,

$x_j (j=1,\ldots,p)$ Characteristic value,

Prediction of training model $f(S\cup\{x_j\})$ with characteristic j,

Prediction of training model $f(S)$ without feature j.

This study will use SHAP to analyze the XGBoost model and predict the kaggle open source data set to evaluate the prediction contribution of 11 features. Specifically, SHAP's TreeExplainer will be used to explain tree-based machine learning models, such as decision trees, random forests, gradient enhancement trees, etc. The reason why SHAP is selected to provide interpretability is that it can provide global and local interpretability. Each observation can obtain its SHAP value, so SHAP can help the model to interpret globally and locally. In addition, contrary to the existing machine learning model for finding important features, SHAP can identify the positive and negative values of input feature contributions, which is of great benefit to this experiment. SHAP's analysis of XGBoost model helps to explore the logic principle of its "black box" to enhance the interpretability of the model and the possibility of its practical application in credit scoring. Specifically, the research focused on using SHAP tools to solve the following problems:

What are the most influential characteristics of the model output?

What are the characteristics of the eigenvalues of the model prediction results?

4. Experimental results and analysis

4.1. Model results and analysis

This experiment is based on Windows 10 64-bit Microsoft operating system, and the experimental data is the kaggle open source data set. The experiment divides the kaggle data set into training set (70%) and test set (30%), which are used for model training and model validation respectively. In the process of data segmentation, the method of stratified sampling of the data set is used for data screening to ensure that the samples extracted from each mobile subgroup are correct, so as to ensure that the test set represents the overall data and ensure the authenticity of the accuracy rate of experimental verification. The training data set is further divided into training set and verification set, which are used in cross validation and super parameter tuning. As shown in the test data in the following table, the evaluation indicators of XGBoost model are given, with an accuracy rate of 0.87 and a recall rate of 0.90, indicating that 90% of the minority cases are correctly predicted.

Table 3. Prediction accuracy of test data (%)

Model algorithm	Precision	Recall	Accuracy	AUC
XGBoost	86	90	87	94
BO-XGBoost	86	90	87	95
LR	70	75	70	80
GBDT	84	82	83	92
BPNN	82	75	80	86

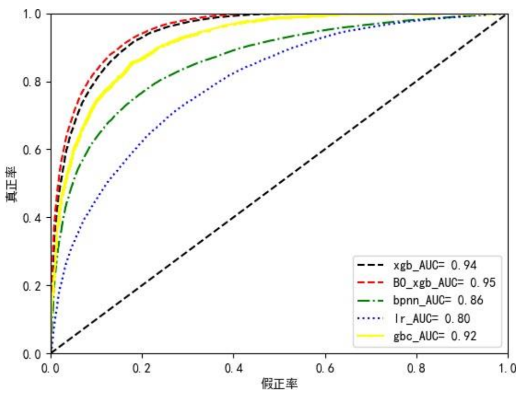

Figure 4. ROC curve of each model

The figure above shows the ROC curves and corresponding AUC values of multiple algorithm models. The most ideal point in ROC space is in the upper left corner of the graph, that is, the AUC value is 1. AUC is a very important evaluation index to evaluate the performance of the data model. The base reference is higher than Accuracy. The smaller the difference between AUC and 1, the better the overall performance of the built model. In this experiment, as shown in the figure above, the AUC value of XGBoost model is 0.95, which is significantly higher than other machine learning models, proving that XGBoost has the best performance compared with other algorithm models.

4.2. Results and analysis of the interpretability and feature importance of the model

The ultimate goal of this experiment is to explore the variables that affect the performance of the model, so as to make objective data-driven decisions on the credit scoring model. The previous section explained that XGBoost model has the best performance, and its prediction results can prove that the selected variables can better predict the credit score. This section focuses on exploring the most important features of the model, which are used to verify the correctness of the model and strengthen the interpretability of the model. The experiment will use the SHAP tool to further explore the impact of the input variables of the model on the prediction of XGBoost model. The global importance factor of the input variable is shown in the figure below. The global importance estimation is to use the average value of the absolute Shapley value of each feature in the data to rank the input variables in an important degree, that is, the higher the SHAP mean value of the variable, the more important the feature variable is. In the figure below, "RevolvingUtilisationOfUnsecured Lines", "NumberOfDependents", "age", "DebtRatio", "NumberRealEstateLoasOrLines" and other characteristics seem to be the five most influential characteristics on the performance of credit score prediction. Among them, "the recycling of unsecured loans" is the most important feature in the credit scoring model, providing about 31% of the model's explanatory power, followed by "the number of family members", providing 18% of the explanatory power, and the first seven features provide 90% of the model's explanatory power. Through interpretive analysis, this is also an important seven feature selection in user information collection.

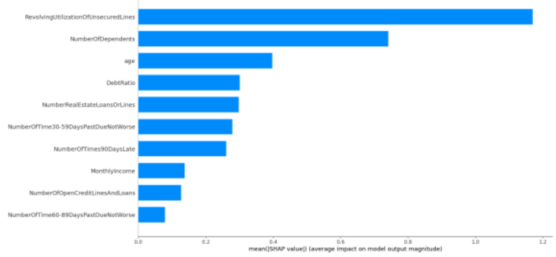

Figure 5. SHAP waterfall diagram explaining the importance of features

This paper shows the influence range and distribution of input characteristics on model prediction by using the summary chart of SHAP tool, as shown in the figure below. Each point shown in the figure is the Shapley value of input variables and instances, and the y-axis displays the input variables in the order of "importance" from top to bottom. Each point in the figure is filled with color by entering the value of the variable, from low to high. The higher the density of points in the figure, the more concentrated the distribution of points in the data.

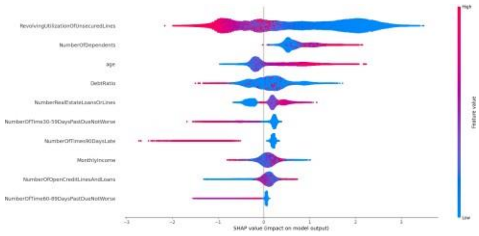

Figure 6. Range and distribution of influence of input characteristics

It can be seen from the above figure that the smaller the value of the features "recycling of unsecured loans" and "debt ratio", the greater their SHAP value, thus pushing the forecast downward (negative). It can be understood that the lower the characteristic value of "the recycling of unsecured loans", the greater the contribution to the prediction results of the model. In addition, the larger the values of "age" and "number of family members" are, the larger their SHAP values are, which will push the prediction upward (positive). It can be understood that the higher the characteristic values of "age" and "number of family members", the greater the contribution to the prediction results of the model. These results can be explained as follows: the less the number of users' loans and liabilities, the greater the probability of users' compliance; The greater the user's age and the number of family members, the greater the user's probability of compliance. Through the interpretation of the model, the characteristics of the influence of input characteristics on the prediction results of the model are obtained.

5. Interpretability

5.1. Interpretative analysis

The results in Section 4.2 can answer the two questions (1) and (2) in Section 3.4. The results show that the two characteristics of "recycling of unsecured loans" and "number of family members" have the greatest impact on the model used in the experiment (answer question (1)). Through the analysis of SHAP value, it is concluded that the smaller the number of loans and liabilities of users, the greater the age and the number of family members, the higher the probability of compliance (answer question (2)). The SHAP value can not only answer the two questions raised in this paper, but also describe the impact of the interaction of the two variables on the target value, which makes the model and optimization proposed in this paper more interpretable. As shown in the figure below.

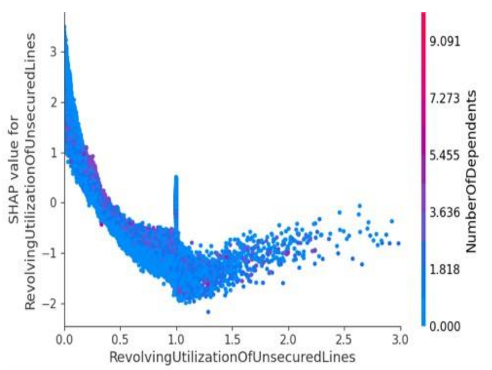

Figure 7. Interaction diagram of SHAP value attribute (recycling of unsecured loans)

The figure above shows that there is a certain interaction between "the recycling of unsecured loans" and "the number of family members". The smaller the number of family members of loan users, the more times unsecured loans will occur, showing a negative correlation trend.

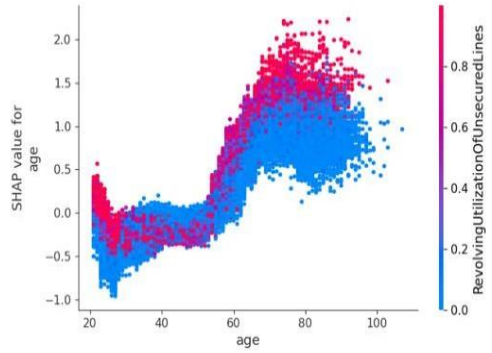

Figure 8. SHAP value attribute (age) interaction diagram

The figure above shows that the age of borrowers is mostly between 50 and 60 years old, indicating that middle-aged people have a large demand for loans. With the increase of age, the number of loans shows a positive correlation trend.

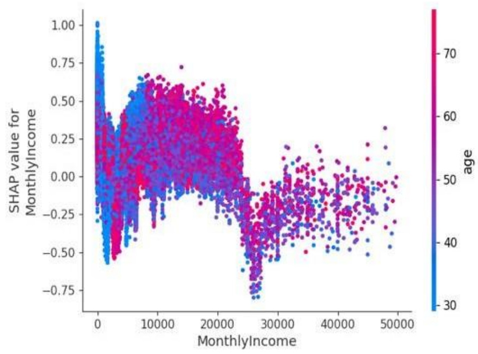

Figure 9. Interaction diagram of SHAP value attribute (monthly income)

The figure above shows that the impact between "monthly income" and "age" is relatively stable, and loan users are concentrated in the population with a monthly income of less than 20000. The characteristics of credit user population can be explained by obtaining the interaction diagram of SHAP values with positive or negative impact characteristics.

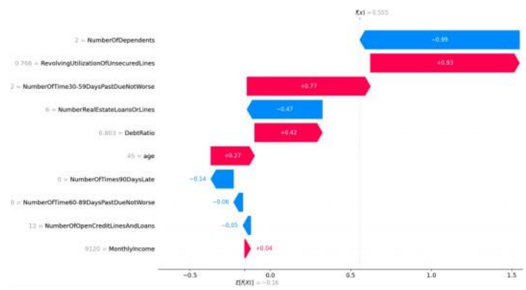

Figure 10. Single sample SHAP interpretation waterfall diagram

The figure above shows the waterfall diagram of a single sample of the model. The abscissa axis is the SHAP value, and the ordinate axis represents the value of each feature of the input sample. The gray indicates that the impact of the feature on the model prediction results is negative, while the gray arrow pointing to the left indicates that the SHAP value of the feature decreases, the black indicates that the impact of the feature value on the model prediction results is positive, and the black arrow pointing to the right indicates that the SHAP value of the feature increases. The displayed value of E [f (x)] below the abscissa axis in the figure is the reference value interpreted by SHAP, that is, the mean value calculated by the current prediction model, and E represents the expected sign.

According to the above figure, from the bottom up, there are three least important characteristic values, which have a positive impact on the model prediction of 0.04, a negative impact of 0.05, and a negative impact of 0.06. From the top up, the most important characteristic of the model is "the recycling of unsecured loans", which has a positive impact on the prediction model of 0.96. Finally, the SHAP value at the top right in the figure is shown as f (x)=0.555.

Figure 11. Single sample SHAP interpretation

The figure above shows the figure calculated by the single sample SHAP interpretation. In fact, the waterfall diagram is arranged horizontally, which can more intuitively see the positive and negative impact and proportion of each characteristic value of the model on the prediction results.

5.2. Conclusion

This paper establishes an interpretable prediction model for the credit scoring model. This model combines XGBoost algorithm and SHAP to predict credit scoring performance. This paper uses the SHAP value to determine the contribution of each input characteristic parameter to the performance of the model, while XGBoost can accurately predict customer credit, and the results can accurately reflect the impact of various characteristic parameters on the prediction performance of the credit model. The model also provides a basis for the formulation of credit scoring strategies. And successfully solved the problem of poor interpretability of credit scoring model in the actual application process. According to the explanatory analysis of the model, the loan limit of customers at different levels can be adjusted. Therefore, the application of the explanation model helps to accurately analyze the causes of bad customers, so that credit service personnel can accurately control the strategy, avoid blind decision-making, and improve the quality and efficiency of work.

6. Conclusion

This paper provides an interpretable analysis of the credit scoring model through machine learning technology. The performance comparison and analysis results of the four algorithm models show that XGBoost algorithm model has the advantage of 87% prediction accuracy in predicting credit loan users based on feature selection. The parameters of XGBoost algorithm are optimized by Bayesian algorithm, and the prediction accuracy of credit score is improved according to the selected variables. By using SHAP exploration model to analyze the interaction between features, identify the importance of features and decode the complex potential relationship between input variables, explain the impact of input features on the data model, and explore the importance and contribution of features to specific prediction. SHAP is used to intuitively explain the complex nonlinear behavior of the basic model, improve the interpretability of the credit scoring model, and provide reference value for commercial credit analysis of user characteristics and product launch.

The single algorithm model proposed in this paper will have some shortcomings in model stability. In the future, model fusion or optimization algorithm can be considered to further improve the accuracy of model prediction, algorithm stability and performance, or the algorithm model and causal calculation can be combined to design a more robust and interpretable integrated algorithm model.

Acknowledgements

National Natural Science Foundation (No.51474095);

Henan University Science and Technology Innovation Team Support Program (No.17IRTSTHN010);

Henan University of Science and Technology Innovation Team Project (No.2015XTD011);

Major industry-university-research cooperation and cultivation fund project of Henan University of Science and Technology (No.2015ZDCXY03).

References

[1] Rodríguez -Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions[J]. Journal of Computer-Aided Molecular Design, 2020, 34(10): 1013- 1026.

[2] Study on landslide susceptibility zoning in hilly and gentle slope areas based on interpretable machine learning_ Zhang Hong. pdf [Z]

[3] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 618-626.

[4] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[M]. arXiv, 2014[2022-11-29]. http://arxiv.org/abs/1406.1078.

[5] Zhou Yong, Wang Hanzheng, Zhao Jiaqi, etc Pedestrian recognition method based on interpretable attention component model [J] Journal of Automation: 1-16

[6] Wen C, Yang J, Gan L, et al. Big data driven Internet of Things for credit evaluation and early warning in finance[J]. Future Generation Computer Systems, 2021, 124: 295-307.

[7] Xu L, Cui L, Weise T, et al. Semi-supervised multi-Layer convolution kernel learning in credit evaluation[J]. Pattern Recognition, 2021, 120: 108125.

[8] Guidotti R, Monreale A, Ruggieri S, et al. A Survey of Methods for Explaining Black Box Models[J]. ACM Computing Surveys, 2019, 51(5): 1-42.

[9] Rzycho ń M, Ż oga ł a A, Róg L. SHAP - based interpretation of an XGBoost model in the prediction of grindability of coals and their blends[J]. International Journal of Coal Preparation and Utilization, 2022, 42(11): 3348-3368.

[10] Qian Chao, Chen Jianxun, Luo Yanbin, et al The imputation method of highway tunnel operation missing data based on random forest [J] Transportation System Engineering and Information, 2016, 16 (03): 81-87

[11] Feng S, Keung J, Yu X, et al. Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction[J]. Information and Software Technology, 2021, 139: 106662.

[12] Bao L, Juan C, Li J, et al. Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets[J]. Neurocomputing, 2016, 172: 198-206.

[13] Ren J, Li B, Zhao M, et al. Optimization for Data-Driven Preventive Control Using Model Interpretation and Augmented Dataset[J]. Energies, 2021, 14(12): 3430.

[14] Chen Y, Chang R, Guo J. Effects of Data Augmentation Method Borderline-SMOTE on Emotion Recognition of EEG Signals Based on Convolutional Neural Network[J]. IEEE Access, 2021, 9: 47491-47502.

[15] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[16] Sumreddee P, Hay E H, Toghiani S, et al. Grid search approach to discriminate between old and recent inbreeding using phenotypic, pedigree and genomic information[J]. BMC Genomics, 2021, 22(1): 538.

[17] Do B, Ohsaki M. A random search for discrete robust design optimization of linear-elastic steel frames under interval parametric uncertainty[J]. Computers & Structures, 2021, 249: 106506.

[18] Deng C, Zhang X, Huang Y, et al. Equipping Seasonal Exponential Smoothing Models with Particle Swarm Optimization Algorithm for Electricity Consumption Forecasting[J]. Energies, 2021, 14(13): 4036.

[19] Xiao Y, Nazarian S, Bogdan P. A stochastic quantum program synthesis framework based on Bayesian optimization[J]. Scientific Reports, 2021, 11(1): 13138.

[20] Xue Y, Deng Y. A real Shapley value for evidential games with fuzzy characteristic function[J]. Engineering Applicationsof Artificial Intelligence, 2021, 104: 104350.