A Solution to the Problem of Retail Credit Risk Pricing Problem Based on the Machine Learning XGBoost Algorithm

Jingxuan Ma[1], Xin Li[1], Jiajie Guo[2], and Qiuyue Li[1(✉)]

[1] College of Science, China Agricultural University, Beijing, China
lqyue@cau.edu.cn

[2] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China
23S005086@stu.hit.edu.cn

Abstract. Machine learning algorithms, represented by logistic regression and decision tree algorithms, have a wide range of applications in retail credit risk management, such as anti fraud models, application scoring models, behavioural scoring models and overdue collection scoring models, which are used to assess the credit risk level of customers at different stages of the retail credit process. The integrated decision tree model, represented by the XGBoost algorithm, constructs the model using more variables and although it lacks the necessary interpretability, its predictions are more accurate. Traditional retail risk pricing typically uses a preapplication scoring model to enforce different product prices for customers with different risk ratings. Based on the XGBoost algorithm and personal credit information from the People's Bank of China and other tripartite credit bureaus, we attempt to develop a retail risk pricing model based on the actual loan disbursement of the customers, which is used to apply different product prices to customers with different risk ratings, while at the same time remarketing to customers who have not disbursed historically. The data for the construction of the retail risk pricing model are taken from a retail personal loan business of a financial institution, and customers who are approved but not actually withdraw money are marked as 1, while those who are approved and actually withdraw money are marked as 0.

Keywords: Retail credit risk pricing · Machine learning XGBoost algorithm · Personal credit data

1 Introduction

As a multidisciplinary field that encompasses probability theory, statistics, approximation theory and complex algorithms, machine learning uses computers

J. Ma and X. Li—Co-First Author.

H.-H. Chen and W. Meng (Eds.): WiSATS 2024, LNICST 605, pp. 172–186, 2025.
https://doi.org/10.1007/978-3-031-86196-3_15

as tools to realistically simulate the way humans learn and to structure existing content in a way that effectively improves learning efficiency. Machine learning has a very wide range of applications in daily life, such as in intelligent finance [1,2], intelligent healthcare [3,4], intelligent transport [5,6], image recognition [7,8] and many other fields [9,10], and it has brought very many changes to our lives. In the field of machine learning, machine learning methods usually include supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning, where supervised learning includes algorithms such as Linear Regression, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Naive Bayes, and other algorithms, while unsupervised learning includes K-Means Clustering, Hierarchical Clustering, Gaussian Mixture Models, Principal Component Analysis, Association Rule Learning, etc. Decision tree algorithms include ID3 algorithm, C4.5 algorithm, CART (Classification and Regression Trees) algorithm, C5.0 algorithm and also integrated algorithms such as GBDT (Gradient Boosting Decision Tree) [11,12], XGBoost (Extreme Gradient Boosting) [13,14], LightGBM (Light Gradient Boosting Machine) [15,16] through are also considered as decision tree algorithms.

In the field of retail credit, the most widely used algorithms are logistic regression algorithm and XGBoost algorithm [13], for example, personal credit application scorecard constructed by using logistic regression algorithm can use fewer variables to evaluate the credit risk level of retail customers, For example, the personal credit application scorecard constructed by using logistic regression algorithm can use fewer variables to evaluate the credit risk level of retail customers, and it can be very intuitive to understand the factors that affect the evaluation results, while the personal credit application scorecard constructed by using XGboost algorithm [13] can autonomously find the factors that affect the evaluation results of customers from thousands of variables, which avoids a lot of human intervention, thus it can achieve the rapid iteration of the model. Retail credit risk pricing refers to the application of different pricing strategies to different risk segments in the consumer credit market in order to improve profitability. Typical risk pricing strategies are based on a retail credit application scorecard, whereby customers with high scores, i.e. those with a low risk of default, are offered lower prices and customers with low scores, i.e. those with a high risk of default, are offered higher prices.

In this paper, based on a financial institution's customer credit application data, including credit data from the People's Bank of China (PBOC) and third party credit bureaus A, B and C, we use the XGBoost algorithm [13] to build a retail credit risk pricing model, which not only allows us to implement different risk pricing strategies for different customers, but also allows us to remarket to customers who have passed credit approval but have not withdrawn their money. The modelling sample includes all approved customers, and the risk pricing model predictions are targeted at customers who have no actual credit transactions among the approved customers, and it should be noted that customers who have been rejected during the credit approval process are excluded from the sample for this modelling. This paper is organised as follows:

- In Sect. 2, we provide a brief overview of common machine learning algorithms.
- In Sect. 3, we provide a brief description of the methodology of the retail credit risk pricing model.
- In Sect. 4, we give the validation metrics of the retail credit risk pricing model and the effect of the applications.
- In Sect. 5, we present the conclusions of the retail credit risk pricing model and its shortcoming.

2 Related Work

In recent years, machine learning algorithms have been used more and more widely, both in traditional financial institutions such as banks [23–25], and in the retail credit segment of some Internet. Commonly used machine learning algorithms include logistic regression algorithms and integrated decision tree algorithms, of which integrated decision tree algorithms include GBDT, XGBoost and LightGBM.

2.1 Logistic Regression Algorithms

Logistic regression is a commonly used binary classification model [17,18], which is usually used to predict binary classification problems, such as whether a customer will click on a product link, whether a credit customer will default on a loan, whether a certain evaluation is positive or negative, and so on. In the financial field, logistic regression is widely used in credit assessment, default prediction, and customer value analysis. Logistic regression is used to predict the value of an output variable by learning the relationship between input features and output labels, and its loss function is usually the cross entropy loss function. The difference between logistic regression and linear regression is that linear regression is used to predict continuous variables where the output variable is a continuous value, such as the price of a commodity, wage income, while logistic regression is used to solve binary problems where the output variable is a binary value, such as whether or not to buy a product, whether or not to default on a contract, and so on. The core principle of the logistic regression algorithm can be expressed as

$$P(y = 1|x; w) = \frac{1}{1 + e^{-(w_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n)}} \tag{1}$$

where $P(y = 1|x; w)$ denotes the probability that the output variable is 1 when the input feature is x, and $x_1, x_2, ..., x_n$ denote the input features, while $w_0, w_1, w_2, ..., w_n$ denote the weight vectors. We use stochastic gradient descent to optimise the loss function $L = -(ylog(p) + (1 - y)log(1 - p))$ until the optimal weight parameters are found. The advantages of using logistic regression to construct a retail credit risk scorecard are that the modelling process is simple

and efficient, it is easy to understand and implement, and it directly predicts classification probabilities without making prior assumptions about the data distribution, thus avoiding problems associated with inaccurate assumptions about the distribution, and the parameters it fits represent the impact of each feature on the results. Similarly, its disadvantages are more obvious, such as the model is easily underfitting, in most cases manual feature engineering is required to construct the combination of features, the classification accuracy may not be high, it is essentially a linear classifier which does not deal well with the correlation between the features, and it is very sensitive to covariance, but there may be multiple covariance which needs to be tested by the VIF (Variance Inflation Factor) test.

2.2 Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is an iterative decision tree algorithm that effectively combines decision trees with integration ideas by constructing a set of weak learners and accumulating the results of multiple decision learners as the final prediction output. The decision tree used by GBDT is a CART regression tree [28,29], whether the problem is regression, binary classification or multiple classification. The core idea is that each tree learns the residuals of all previous decision trees, and the residuals are essentially the difference between the true value and the predicted value. In the learning process, it first learns a regression tree and then subtracts the true value from the predicted value to get the residuals, and then takes the residuals as a learning target to learn the next regression tree, and so on, until the residuals are less than a certain threshold close to 0, or the number of regression trees reaches a certain threshold. The core idea is to reduce the loss function by fitting the residuals in each round. It uses a gradient descent algorithm to optimise the loss function. It approximates the loss function with a firstorder Taylor expansion of the loss function during model optimisation, which can reduce the amount of computation in the model optimisation process. The GBDT model has high accuracy, it performs well on both training and test sets, while it can handle complex problems such as high dimensionality [35,36], sparse features [37,38] and nonlinear relationships, and it also has strong generalisation ability as it reduces the risk of overfitting by combining several weak classifiers into a single strong classifier. However, as GBDT is a serial algorithm, it needs to construct each decision tree sequentially, so the training time is long, while it is sensitive to outliers, and it is easily affected by outliers during the training process, which may lead to the degradation of the model's performance. In addition, although the generalisation ability of GBDT model is better, when the model has less sample data or the number of model features is small, it is easy to have overfitting problem.

2.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an efficient gradient boosting decision tree algorithm [26,27] that uses the second order Taylor expansion of the

loss function to approximate the loss function, which greatly improves the model performance compared to GBDT. As a forward additive model [33,34], it adopts the same integration idea as GBDT, which works by integrating multiple weak learners into one strong learner by certain methods, and uses multiple trees for joint decision making. In the XGBoost model, the prediction result of each decision tree is the difference between the target value and the prediction results of all previous decision trees, so the final prediction result of the model is obtained by adding the prediction results of all decision trees. Unlike other models, which typically require a separate preprocessing step, XGBoost can handle the missing value problem internally, with the algorithm finding the best imputation values for the missing values and then storing them for future prediction. Although the augmentation algorithm is prone to overfitting problems, XGBoost improves the generalisation of the model by incorporating the L1 (lasso) [19,20] and L2 (ridge) [21,22] regularisations directly into the objective function during training. For sparse data problems, XGBoost uses compressed, memory efficient data structures and its algorithm is designed to efficiently traverse sparse matrices. Most integrated methods provide feature importance metrics, but XGBoost provides a more comprehensive set of feature importance metrics, including gain, frequency and coverage, which allows for a more detailed interpretation of the model. For categorical variable problems, XGBoost's treatment of categorical variables is more nuanced than simple binary partitioning, allowing complex relationships to be captured without additional preprocessing.XGBoost's unique capabilities make it not only a state of art machine learning algorithm in terms of predictive accuracy, but also efficient and customisable. Its ability to handle real world data complexities such as missing values, sparsity and multicollinearity, while being computationally efficient and providing detailed interpretability, makes it an invaluable tool for a wide range of data science tasks.

2.4 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) [30,32] is an efficient and scalable machine learning algorithm based on Gradient Boosted Decision Trees, which combines the advantages of the GBDT algorithm and the XGBoost algorithm with a series of optimisations to the model. In order to solve the problem of the GBDT algorithm framework's computational inefficiency [39] when dealing with large amounts of data, LightGBM greatly improves the computational efficiency of GBDT at the expense of a very small computational accuracy. Despite the sacrifice in computational accuracy, the actual modelling effect of LightGBM is almost at the same level as XGBoost. From another perspective, although LightGBM sacrifices computational accuracy to a certain extent, it suppresses the overfitting problem [40,41] of the model. Therefore, in many scenarios, the algorithmic effect of LightGBM will even be better than XGBoost. In terms of algorithmic principles, LightGBM also adopts a gradient based boosting algorithm, but it uses the optimisation technique of Gradient based One Side Sampling (GOSS) to speed up the training process by retaining samples with larger gradients. In terms of data processing, when the amount of data is very large,

XGBoost may face the problem of insufficient memory, but LightGBM significantly improves the data processing capability by adopting the GOSS technique, and is able to process larger data sets efficiently, while outperforming XGBoost in terms of computational speed and memory consumption. In addition, LightGBM uses the histogram algorithm for node splitting, which can take full advantage of the parallel computing capabilities of multi core CPUs and significantly accelerate the training speed. LightGBM uses a leaf wise growth strategy to find a leaf node with the largest splitting gain from all the current leaves, then splits it, and then loops to execute this strategy. However, this strategy grows deeper decision trees, which in turn creates an overfitting problem, so LightGBM adds a maximum depth limit on top of leaf wise to prevent overfitting while maintaining high efficiency.

3 Methodology

Retail credit risk pricing is an assessment of the price of a risky asset. In order to maximise operating profit, financial institutions usually implement different pricing strategies for customers with different levels of credit risk, whereby customers with lower credit risk usually receive financing services at lower prices, while those with higher credit risk usually need to receive financing services at higher prices, i.e., users with poorer levels of risk need to be supplemented with a risk premium. In this paper, based on the actual business data of financial institutions, including customers' application information data, approval result data, credit data from the People's Bank of China and credit data from other three party financial institutions, we use the XGBoost machine learning algorithm to construct a retail credit risk pricing model that evaluates the customer's withdrawal propensity by predicting whether the customer who passed the credit approval will have an actual withdrawal behaviour in the next three months. Based on the results of the model's prediction, a higher pricing strategy is implemented for customers with a high propensity to withdraw, while a relatively lower pricing strategy is implemented for customers with a relatively low propensity to withdraw.

3.1 Model Sample Segmentation Strategies

We randomly selected 10,000 customers from all retail credit approvals for use in building the retail credit risk pricing model. 77% of the 10,000 customers modelled made a withdrawal within three months of being approved, but the remaining 23% did not actually make a withdrawal within three months of being approved for credit. In order to test the actual predictive power of the model and whether there is an overfitting or underfitting problem in the model, we divided all the model samples into training and test sets according to a 2:1 ratio, where the number of training sets totals 6667, while the numbers of samples in the test set totals 3333. As a total of 23% of the 10,000 customers modelled did not make an actual withdrawal, the percentages of target customers in both the training and test sets are about 23%.

The model feature variables used in this modelling are mainly from four parts, which are the credit data of the People's Bank of China (PBOC), the data of the three party credit agency A, the data of the three party credit agency B and the data of the three party credit agency C. The data of the PBOC contains about 15,000 feature variables, the data of the three party credit agency A contains about 1,250 model variables, the data of the three party credit agency B contains about 410 model variables, while the data of the three party credit agency C contains about 60 model variables. We categorise the PBOC's credit data according to their actual business meanings, with a total of 15 different categories, where the business meanings of the feature variables within each category were similar or the same, and the businmess meanings of the feature variables differe significantly of different categories. With 15 categories of PBOC credit data and 3 categories of credit data from tripartite credit agencies, we have a total of 18 categories of credit feature variables to construct the retail credit risk pricing model.

3.2 Model Indicator Screening Strategies

When faced with a large number of features, relying on all of them for modelling can lead to dimensional disaster and overly complex models, while also introducing redundant or irrelevant features, reducing the accuracy and interpretability of the model [42,43]. The goal of multi feature screening [46,47] is to select the subset of features from all available features that are most important for credit risk model building and prediction. By properly selecting and carefully screening features, we can improve the predictive power of the model, simplify the model complexity, reduce the risk of overfitting, and better explain the model predictions. These optimisations will help to improve the stability, reliability and practicality of the credit risk model and provide effective support for risk management and decision. In total, there are about 17,000 different feature variables in the 18 different categories of credit data mentioned above, and there is a high degree of multicollinearity among feature variables in the same category. Although the XGBoost algorithm can theoretically handle an infinite number of different feature variables and does not need to preprocess the multicollinearity problem, from a practical application point of view, we need to filter the feature variables to some extent to ultimately achieve lightweight deployment.

For each set of feature variables of different categories, we use the XGBoost algorithm to model the target customers on the training set and perform parameter tuning by adjusting the different model hyperparameters, in particular, we can adjust the weight coefficients of the different variables by adjusting the L1 regularity coefficients and the L2 regularity coefficients. The L1 regularity is the sum of the absolute values of the elements in the pointers, also known as Lasso regularisation, in the high dimensional case, if a feature is unimportant, even if the weights of that feature are large, it has a small impact on the loss function but a large impact on the regularity term, which is then filtered out in the presence of the regularity term. L2 regularisation, also known as ridge regression, is the process of summing the squares of the elements of a quantity

and then finding the square root, the effect of the L2 regular term is to keep all the parameters close to 0 so that the model isn't particularly sensitive to any particular feature, i.e. when running on a test set, even if there is unusually strong noise on one of the features, that noise won't have a big effect on the result for the output of the overall model. The process of feature screening using the XGBoost algorithm is as follows:

- Train the initial XGBoost model based on the feature variables to be screened.u
- Rank the feature variables according to the importance of the feature variable [44,45], eliminate the variables with lower importance rankings.
- Continue to perform feature variables and screening, and so on, until the number of retained variables reaches a certain threshold or a significant degradation in model performance occurs after modelling the excluded feature variables.

3.3 Model Training Strategies

Machine learning models are trained using specific algorithms and data to create a model that can solve a specific problem. Throughout the model training process, the model learns the associative relationships hidden behind the data from the training data set, with the goal of being able to accurately predict or classify new data. The model training process typically involves selecting model training data, dividing the model samples into training and test sets, tuning the model parameters, and testing the model. Through model training, the model is able to learn the laws of the feature variables and use these laws to make predictions or decisions on new data. Ultimately, the model obtained through training determines a set of parameter values at which the model is better able to handle new data.

In the process of model training, the model training data we use includes 10,000 training samples, in which the proportion of the prediction target in the whole sample is about 23%, and these 10,000 data are divided into two parts, the training set and the test set, according to the ratio of 2:1, in which the training set is used to adjust the parameters of the model, and the test set is used to evaluate the generalisation ability of the model. In the process of model training, we always consider factors such as the model's KS value (Kolmogorov Smirnov value), AUC (Area Under Curve) value the accuracy rate and the actual percentage of customer withdrawals in different probability intervals, and at the same time we evaluate the model's generalisation ability by comparing the model's performance in the training set and the test set. Overall, model training is a process of constantly adjusting model parameters, and this process requires making full use of the data set to evaluate model performance in order to obtain a well performing model with good generalisation ability.

3.4 Model Testing Strategies

The testing of machine learning models consists of two aspects: on the one hand, it refers to whether the model's performance on the training set is stable and the prediction results are up to the expected standard; on the other hand, it refers to whether the model trained on the training set has a good generalisation on the test data set. The performance of the model on the training set can indicate whether the model has sufficiently learned the rules hidden behind the training data, while the performance of the model on the test set can indicate whether the performance of the model can achieve stable performance on new data. Through several rounds of tuning the model parameters, we finally retained 34 feature variables as the final model features, and the KS values of the model on the training and test sets are 0.28 and 0.26, while the AUC values are 0.69 and 0.67, respectively.

The KS curves of the model on the training and test sets are shown in Fig. 1:

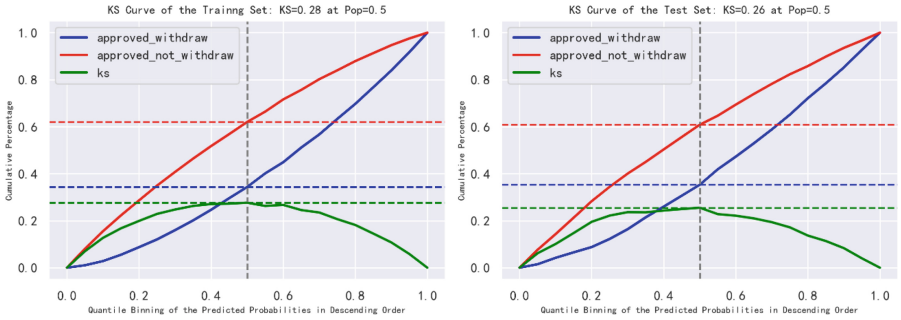

Fig. 1. KS curves of the training set and test set

The AUC curves of the model on the training and test sets are shown in Fig. 2:

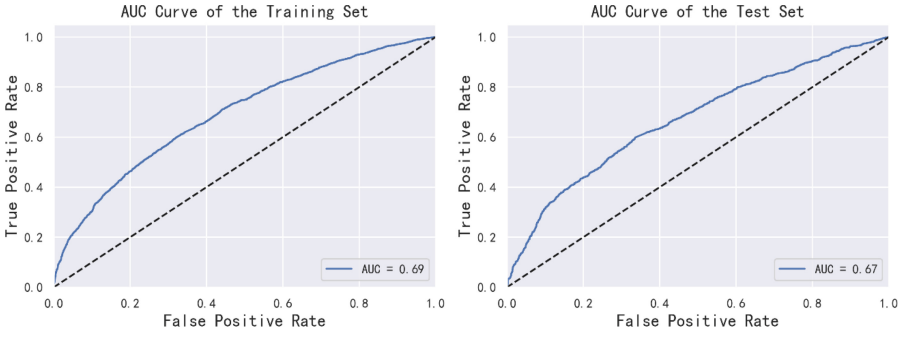

Fig. 2. AUC curves of the training set and test set

The probability of withdrawals in different probability intervals for training set customers and test set customers as shown in Fig. 3:

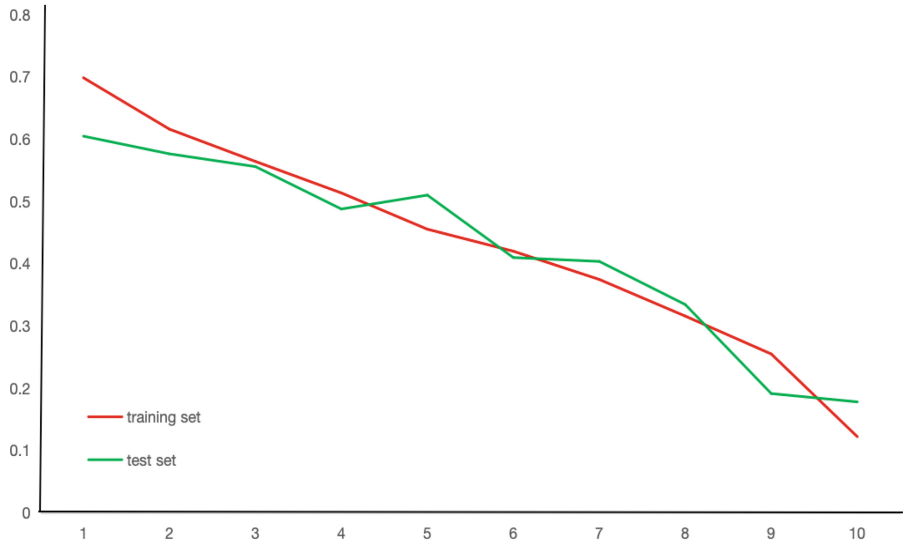

Fig. 3. Withdrawal probabilities for different probability intervals

From the perspective of model generalisation, the metrics of the training set and the test set are relatively close to each other, i.e. there is no obvious overfitting phenomenon. The results are in line with expectations, and the model results can be practically applied. From the perspective of model metrics, the KS of the training set is about 0.28, with an accuracy of about 69%, and the KS of the test set is about 0.26, with an accuracy of about 67%, which is a little lower than expected, but can be worth trying to apply. From the perspective of the future extension space, we only use the credit data of the People's Bank of China (PBOC) and the credit data of three third party credit agencies, although the number of feature variables of the PBOC credit data is large, but the meanings of most of the feature variables are overlapping. There are more feature variables related to customers' credit risk, but fewer characteristic variables related to customers' willingness to withdraw money, so there is some room for improvement.

4 Applications

In this paper, the application of the retail credit risk pricing model has two aspects: firstly, we develop a differentiated pricing strategy for new customers based on their propensity to withdraw, and secondly, we remarket historically approved but underserved customers by lowering the credit price.

4.1 Differentiated Pricing Strategies

For credit customers with low withdrawal propensity, the price of credit products can be lowered to increase the turnover rate under the premise of protecting the credit risk of the customers, while for customers with high withdrawal propensity, the price of credit products can be raised under the premise of protecting the turnover rate and credit risk of the customers to increase the return of the customers, so as to achieve the effective distribution of funds and the optimal allocation of resources. In addition, for customers with poor credit qualifications, differentiated pricing can be used to allow more people to receive financial services, thus achieving the purpose of true financial inclusion.

From the Fig. 4, we can see that, overall, for retail credit customers in different willingness bands, customers' willingness to withdraw decreases and then slowly increases as the interest rate on the loan increases. However, customers' willingness to withdraw decreases rapidly when the interest rate increases from 16% to 17% and does not decrease with the increase in the interest rate when the interest rate is above 17%. That is to say, the lower the interest rate, the higher the willingness of customers to withdraw money, but once the interest rate level reaches a certain level, the willingness of customers to withdraw money no longer decreases as the interest rate rises, and even increases to a certain extent.

Generally speaking, the willingness of customers to withdraw has a correlation with their credit qualification, i.e. the better the credit qualification, the lower the willingness to withdraw, while the worse the credit qualification, the higher the willingness to withdraw. For customers with higher credit qualification, they have more channels to obtain credit products, so they are more sensitive to the interest rate of credit products, while for customers with lower credit qualification, they have fewer channels to obtain credit products, so they are not sensitive to the interest rate of credit products. In the past, during the stage of rough development, financial institutions usually offered low interest rate products to customers with better credit qualifications and high interest rate products to customers with poorer credit qualifications. Therefore, in the low interest rate band, the overall customer qualification is relatively good, and they are more sensitive to interest rates, so when the interest rate is raised, the probability of withdrawal is significantly reduced. However, when the interest rate is higher than a certain level, the customers' credit qualifications are relatively poor and they are not sensitive to interest rates, so when the interest rate is raised, the withdrawal rate of the customers does not decrease, but rather, it is raised to a certain extent. Thus, for customers in different willingness zones, the actual probability of withdrawal first decreases significantly when the interest rate level of the credit product goes from low to high, but then increases slightly when the interest rate rises to a certain level.

We know from detailed calculations that implementing a differentiated pricing strategy can increase the profitability of a single transactional customer by 15%.

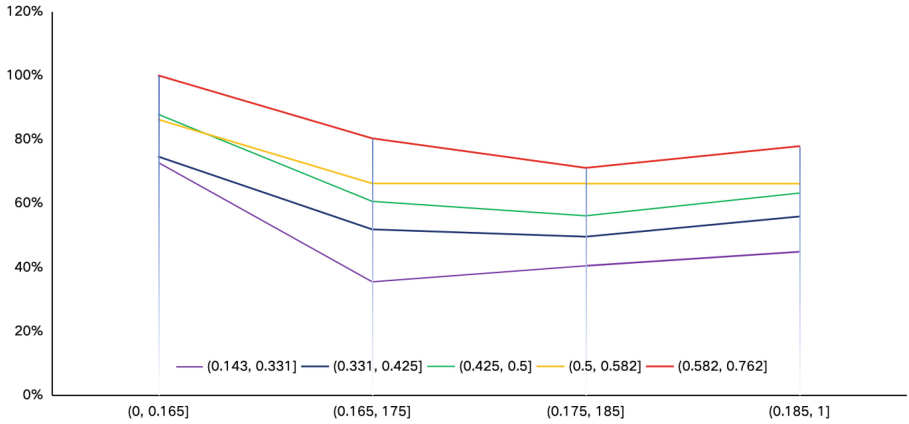

Fig. 4. Willingness to withdraw under different interest rate bands

4.2 Remarketing to Unwithdrawn Customers

For customers whose credit approvals have been passed, whether or not the customer actually takes out the loan is closely related to the price of the credit product. In the pass through case, the lower the price of the credit product, the higher the probability that the customer will complete the transaction and, conversely, the higher the price of the credit product, the lower the probability that the customer will draw down the loan.

For all 10,000 samples used to construct the retail credit risk pricing model, we divide them into five parts according to their withdrawal propensity, from low to high, and label them A1, A2, A3, A4 and A5. If we conduct secondary marketing for customers in groups A1, A2, and A3 who do not make withdrawals, and implement differentiated interest rates of 14%, 15%, and 16% for customers in groups A1, A2, and A3 who do not make withdrawals, and assume that they have response probabilities of 5%, 6%, and 7%, respectively, we can add ¥410,000, ¥350,000, and ¥320,000, respectively, to our additional profits. The overall withdrawal probability of customers in A4 and A5 is relatively high, and based on the results in Fig. 4, it can be seen that if we increase the effective interest rate for the interest rate insensitive customers in this segment, the withdrawal probability of customers does not decrease . For this group of customers, if we increase the effective interest rate by 1 per cent, it is calculated that this group of customers can increase their income by about ¥ 4,500,000.

5 Conclusions

In this paper, we describe the construction of a retail credit pricing model using the XGBoost machine learning algorithm. The entire model construction process includes identifying the model data, defining the prediction target of the

model, dividing the model samples into training and test sets, performing feature variable screening, performing model training, performing model testing, and applying the model. However, this method of pricing retail credit risk only models the customer's propensity to withdraw and does not take into account corporate social responsibility and national policies. In fact, the pricing of credit products by financial institutions should not only consider the business cost, risk and necessary profit from their own actual situation, but also make the price acceptable to customers from the customers' point of view, and also consider the competitors' pricing strategy and the price level of the market.

References

1. Liu, X., Salem, S., Bian, L., Seong, J.T., Alshanbari, H.M.: Application of machine learning algorithms in the domain of financial engineering. Alexandria Eng. J. **95**, 94–100 (2024)
2. Nazareth, N., Reddy, Y.V.R.: Financial applications of machine learning: a literature review. Expert Syst. Appl. **219**, 119640 (2023)
3. Ponsiglione, A.M., et al.: Combining simulation models and machine learning in healthcare management: strategies and applications. Prog. Biomed. Eng. **6**(2), 022001 (2024)
4. Yaghoobpoor, S., et al.: Machine learning approaches in the prediction of positive axillary lymph nodes post neoadjuvant chemotherapy using MRI, CT, or ultrasound: a systematic review. Eur. J. Radiol. Open **12**, 100561 (2024)
5. Agarwal, S., Gupta, S., Kachroo, P., Dhingra, N.: A machine learning based approach for smart and automated data collection: applications in transportation. Transport. Dev. Econ. **10**(1), 15 (2024)
6. Durlik, I., Miller, T., Dorobczyński, L., Kozlovska, P., Kostecki, T.: Revolutionizing marine traffic management: a comprehensive review of machine learning applications in complex maritime systems. Appl. Sci. **13**, 8099 (2023)
7. Wu, C.W., et al.: Recognition of glaucomatous fundus images using machine learning methods based on optic nerve head topographic features. J. Glaucoma, 10-1097 (2024)
8. Li, H., Li, R.W., Shu, P., Li, Y.Q.: Machine learning-based identification of contaminated images in light curve data preprocessing. Res. Astron. Astrophys. **24**(4), 045025 (2024)
9. Singh, K.N., Mantri, J.K.: An intelligent recommender system using machine learning association rules and rough set for disease prediction from incomplete symptom set. Decis. Anal. J. **11**, 100468 (2024)
10. Qian, R., et al.: Predictive value of machine learning for the severity of acute pancreatitis: a systematic review and meta-analysis. Heliyon **10**, e29603 (2024)
11. Xie, F., Wang, H., Ni, S., An, C.: Efficiency optimization control of permanent magnet synchronous motors for pure electric vehicles based on GBDT. J. Power Electron. **24**(2), 215–226 (2024)
12. Zhang, G.: Application of project-based learning model based on GBDT model in higher vocational civics classes at the time of innovation. Appl. Math. Nonlinear Sci. **9** (2024)
13. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. CoRR arxiv:1603.02754 (2016)

14. Wu, Z., Zhao, J., Li, Y., Wang, Z., He, B., Chen, L.: A GAN-BO-XGBoost model for high-quality patents identification. Sci. Rep. **14**(1), 9560 (2024)
15. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. Neural Inf. Process. Syst. (2017)
16. Lu, Y., Wang, J., Wang, D., Yoo, C., Liu, H.: Incorporating temporal multi-head self-attention convolutional networks and LightGBM for indoor air quality prediction. Appl. Soft Comput. **157**, 111569 (2024)
17. Zhou, B., Yuan, Y., Song, Q.: A proximal forward-backward splitting based algorithmic framework for Wasserstein logistic regression using heavy ball strategy. Int. J. Syst. Sci. **55**(4), 644–657 (2024)
18. Huang, Y.: Analysis of the impact of ADDIE education model based on logistic regression model on teaching contemporary cultural and creative product design. Appl. Math. Nonlinear Sci. **9** (2024)
19. Binns, M., Usai, A., Theodoropoulos, C.: Identifiability methods for biological systems: determining subsets of parameters through sensitivity analysis, penalty-based optimisation, profile likelihood and LASSO model reduction. Comput. Chem. Eng. **186**, 108683 (2024)
20. Han, D., Modisette, V., Forthofer, M., Paul, R.: Hierarchical Bayesian adaptive lasso methods on exponential random graph models. Appl. Netw. Sci. **9**(1), 9 (2024)
21. Han, F., Cheng, C.: The innovation path of virtual practice teaching in college Civics class based on the Ridge regression model. Appl. Math. Nonlinear Sci. **9** (2024)
22. Selman, M., Özge, A., Atila, G., Necla, G.: A new robust ridge parameter estimator having no outlier and ensuring normality for linear regression modele. J. Radiat. Res. Appl. Sci. **17**, 100788 (2024)
23. Coskun, T., Murat, D.: Constructing early warning indicators for banks using machine learning models. North Am. J. Econ. Finan. **69**, 102018 (2024)
24. Gaurav, K., Ramizur, R.M., Abhinav, R., Kumar, M.A.: Predicting systemic risk of banks: a machine learning approach. J. Model. Manag. **19**, 441–469 (2024)
25. Nguyen, L.Q.T., Matousek, R., Muradoglu, G.: Bank capital, liquidity creation and the moderating role of bank culture: an investigation using a machine learning approach. J. Finan. Stabil. **72**, 101265 (2024)
26. Li, X., et al.: An improved method for broiler weight estimation integrating multi-feature with gradient boosting decision tree. Animals **13**(23), 3721 (2023)
27. Ait Naceur, H., Abdo, H.G., Igmoullan, B., Namous, M., Alshehri, F., Albanai, J.A.: Implementation of random forest, adaptive boosting, and gradient boosting decision trees algorithms for gully erosion susceptibility mapping using remote sensing and GIS. Environ. Earth Sci. **83**(3), 121 (2024)
28. Dacko, M., Oleksy, A., Synowiec, A., Klimek-Kopyra, A., Kulig, B., Zajac, T.: Plant-architectural and environmental predictors of seed mass of winter oilseed rape in southern Poland based on the CART trees regression model. Ind. Crops Prod. **192**, 116109 (2023)
29. Sharma, D.N., Iqbal, S.I.M.: Applying decision tree algorithm classification and regression tree (CART) algorithm to gini techniques binary splits. Int. J. Eng. Adv. Technol. **12**(5), 77–81 (2023)
30. Zhang, S., Hu, Y., Tan, Z.: Research on borrower's credit classification of P2P network loan based on LightGBM algorithm. IJES **11**, 602–612 (2019)
31. Guo, Q., et al.: Mobile user credit prediction based on LightGBM. In: Proceedings of 2019 International Conference on Big Data, Electronics and Communication Engineering (BDECE 2019) (2019)

32. Mao, X., et al.: A variable weight combination prediction model for climate in a greenhouse based on BiGRU-Attention and LightGBM. Comput. Electron. Agric. **219**, 108818 (2024)
33. Lahiri, A., Paria, B., Biswas, P.K.: Forward stagewise additive model for collaborative multiview boosting. IEEE Trans. Neural Netw. Learn. Syst. **29**(2), 470–485 (2016)
34. Zhong, W., Duan, S., Zhu, L.: Forward additive regression for ultrahigh-dimensional nonparametric additive models. Stat. Sin. **30**(1), 175–192 (2020)
35. Telea, A., Machado, A., Wang, Y.: Seeing is learning in high dimensions: the synergy between dimensionality reduction and machine learning. SN Comput. Sci. **5**(3), 279 (2024)
36. Tan, W.G.Y., Xiao, M., Wu, Z.: Robust reduced-order machine learning modeling of high-dimensional nonlinear processes using noisy data. Dig. Chem. Eng. **11**, 100145 (2024)
37. Hiyama, K., Takeuchi, K., Omodaka, Y., Srisamranrungruang, T.: Operation strategy for engineered natural ventilation using machine learning under sparse data conditions. Jpn. Arch. Rev. **5**(1), 119–126 (2022)
38. Chen, X., Chen, H., Nan, S., Kong, X., Duan, H., Zhu, H.: Dealing with missing, imbalanced, and sparse features during the development of a prediction model for sudden death using emergency medicine data: machine learning approach. JMIR Med. Inf. **11**, e38590 (2023)
39. Zhang, L., Csányi, G., van der Giessen, E., Maresca, F.: Efficiency, accuracy, and transferability of machine learning potentials: application to dislocations and cracks in iron. Acta Mater. **270**, 119788 (2024)
40. Chao, B, Zhang, Z.: Research on overfitting problem and correction in machine learning. In: Journal of Physics: Conference Series, vol. 1693, p. 012100 (2020)
41. Salam, M.A., Azar, A.T., Elgendy, M.S., Fouad, K.M.: The effect of different dimensionality reduction techniques on machine learning overfitting problem. Int. J. Adv. Comput. Sci. Appl. **12**(4), 641–655 (2021)
42. Presciuttini, A., Cantini, A., Costa, F., Portioli-Staudacher, A.: Machine learning applications on IoT data in manufacturing operations and their interpretability implications: a systematic literature review. J. Manuf. Syst. **74**, 477–486 (2024)
43. Baur, L., et al.: Explainability and interpretability in electric load forecasting using machine learning techniques–a review. Energy AI **16**, 100358 (2024)
44. Wen, H.T., Wu, H.Y., Liao, K.C.: Using XGBoost regression to analyze the importance of input features applied to an artificial intelligence model for the biomass gasification system. Inventions **7**(4), 126 (2022)
45. Song, T., Yan, Q., Fan, C., Meng, J., Wu, Y., Zhang, J.: Significant wave height retrieval using XGBoost from polarimetric Gaofen-3 SAR and feature importance analysis. Remote Sens. **15**(1), 149 (2022)
46. Jovanovic, L., et al.: Improving phishing website detection using a hybrid two-level framework for feature selection and xgboost tuning. J. Web Eng. **22**(3), 543–574 (2023)
47. Abbas, Z., ur Rehman, M., Tayara, H., Zou, Q., Chong, K.T.: XGBoost framework with feature selection for the prediction of RNA N5-methylcytosine sites. Molec. Therapy J. Am. Soc. Gene Therapy **31**(8), 2543–2551 (2023)