Optimization of Machine Learning Models for Prediction of Personal Loan Default Rate

Yanghuai He[1], Yuzhe Jian[2], Tianyuan Liu[3], and Huaijin Xue[4(✉)]

[1] International Department of Guangzhou Foreign Language School, Guangzhou, China
[2] Malvern College, Worcestershire, UK
[3] Sacred Heart School of Halifax, Halifax, Canada
[4] School of Mechanical Engineering, Donghua University, Shanghai, China
guanghua.ren@gecacademy.cn

Abstract. The credit industry's continuing expansion depends on the application of modern information technology to lower the risk of credit default. Traditional credit default prediction model research places too much emphasis on the model's accuracy while ignoring some of its most important characteristics. Simultaneously, the parameter characteristics must be manually removed to reduce the model's complexity, which lessens the high-dimensional correlation between the analyzed data and lowers the model's prediction performance. Therefore, this paper constructs two personal credit loan default risk assessment models based on Random Forest (RF) and Light Gradient Boosting Machine (LightGBM), using Accuracy Rate (ACC) and Area Under the ROC Curve (AUC) as performance evaluation metrics. According to empirical studies, the most important determinants affecting loan defaults are 'debt_loan_ratio' and 'known_outstanding_loan'. The AUC of the LightGBM model is above 86%, while RF's AUC is just about 55%, indicating the better performance for the former one. Overall, these results shed light on the prediction of load default rate, which will be a guideline for further policy implementation.

Keywords: Machine Learning Models · LightGBM · Random Forest · Credit Default Prediction

1 Introduction

Loan consumption practices have gradually been accepted by a wider spectrum of social groupings as the quality of living has improved and lifestyles have changed. Meanwhile, the advantages of quick approval times and flexible loan terms have led to an increase in the number of people opting for personal loans over bank loans. Furthermore, several industrialized nations (e.g., Germany, Switzerland, and Japan) have entered the age of negative interest rates to boost economic development. In addition, bank interest rates have been dropping on a regular basis. Personal loans have become a new sort of

Y. He, Y. Jian, T. Liu and H. Xue—Contributed equally.

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 270–282, 2023.
https://doi.org/10.2991/978-94-6463-030-5_29

financial management for many financial investors on this basis. Rather than borrowing in the classic sense, this is a new way of thinking about borrowing.

Banks, as an industry with strong risk control needs, frequently face challenges in expanding their company due to the lack of awareness of new consumer groups and risk management of new customer segments. When commercial banks conducted credit risk evaluations of loan users in the past, they frequently depended on risk control professionals to make subjective judgements using the 5C categorization approach. They assessed loan applicants based on five key factors: personal character, credit history, solvency, and market economy. It is the guideline for judging and assessing whether or not to lend to the user. This system, which is heavily reliant on subjective judgement, is clearly inefficient, especially during the full evaluation process when the risk control personnel's personal judgement capacity will be severely disrupted. Furthermore, there is the possibility of internal risk control employees cheating from the standpoint of bank internal control. This highly human intervention risk assessment approach is clearly incapable of adapting to the rapid expansion of the market economy and meeting the demands of a large number of loan users, as well as meeting the risk management and control needs of banks. Banks must utilize various machine learning approaches to decrease the manual element of the monitoring and detection process and use automated ways to increase the accuracy and efficiency of loan assessment when confronted with tens of thousands of users who request loans.

Based on big data technologies, the lending sector has achieved some achievements in risk control and management. The analysis model developed by ZestFinance based on data mining and machine learning theories, as well as the FICO credit score, the most widely used personal credit scoring system in the United States and an important reference standard for lending decision-making in the US lending industry, are more mature products [11]. Contemporarily, the rapid growth of machine learning has enabled lending platforms to construct sophisticated risk control models using multi-dimensional big data, allowing them to more correctly assess personal credit status and effectively reduce the risk of default. Most researchers opt to pre-extract data characteristics before modelling to minimize model latitude to some level in the early stages. However, this feature extraction strategy will subject the model to the subjective impact of human variables once again, decreasing the model's complete capacity to interpret data relevance and perhaps raising control uncertainty. In this tutorial, we'll create models with random decision trees and LightGBM and compare them to discover the most efficient approach.

The sequel of this paper is organized as follows: Sect. 2 reviews the previous progress researchers have employed to forecast personal loan defaults. The proposed methods and the data utilized are described in depth in Sect. 3. The experimental findings and performance analysis on the database utilized in the work are presented in Sect. 4. Sections 5 and 6 explain and compare the results in further depth, respectively. Finally, Sect. 7 concludes by summarizing the contributions made in this paper.

2 Literature Review

Scholars have explored the application of a range of machine learning methods such as decision trees, neural networks, support vector machines, and integrated algorithms

in credit evaluation so far to assure the credit industry's continued, steady, and healthy development. According to Ref. [4], researchers constructed various models based on the Lending Club dataset to identify which factors are significant for forecasting loan defaults and which sorts of borrowers can pay interest on time. Furthermore, the paper finds that when the ACC is used solely as a model performance index, the RF model is the best classifier for identifying which borrowers are likely to default on a loan contract, and the decision tree (DT) model is the best choice for discriminating which customers have good credit. Scholars use restricted Boltzmann machines (RBMs) to extract the characteristics of the credit dataset before building a linear discriminant analysis (LDA) prediction model [8]. The ACC of the LDA model on the German credit dataset is better than that of several models such as LR (logistic regression), artificial neural network (ANN), SVM, and RF, but it is only 76.5%, which falls short of the actual application criteria. On the Lending Club loan dataset, a previous study builds an RF model for borrower status prediction, and the ACC of RF is 87% [5].

Apart from these, Ref. [2] discovered that when numerous characteristic variables have complicated non-linear correlations, classical Logistic regression does not perform as well as before. While the Logistic model may not be as accurate as the Machine Learning model in terms of prediction accuracy, it offers significant benefits in terms of variable explanatory and stability. As a result, several researchers have enhanced Logistic regression and used it to the prediction of borrowers' default behaviour. To create credit, the consumers are used to categorize into four groups: "cleared in advance", "currently normal", "suspicious", and "loss" from the two categories of "default" and "non-default" [12]. To perform an empirical investigation of online loan customer default prediction, state the transition equation, reverse-speculate the probable default probability of each type of client, and utilize the ordered multi-category Logistic model and ROC test. In terms of customer default prediction, the results reveal that the model is primarily geared at "currently normal" customers. Customers who are "suspicious" and "untrustworthy" have a greater accuracy rate. In previous literature, the L1 penalty Logit model is applied to empirically test the key influencing factors of P2P online loan credit default, as well as classification evaluation methods like confusion matrix and ROC curve to test the model's default prediction effect, and the research concludes that the L1 penalty Logit model is effective at predicting default [6]. The Logit model provides a good variable selection function that may efficiently identify the major elements impacting credit default and decrease managers' supervision costs; the L1 penalty is an excellent example of this. Ordinary Logit models, support vector machines, and other models can't match the accuracy of a Logit model. It can accurately anticipate the credit default status as a whole, as well as examine the impact of major influencing factors on the chance of default, which is useful for predicting and controlling credit risks.

Besides, scholars advocated the use of decision trees and other supervised learning methods to predict fraud, and the 10 factors with the highest explanatory character were picked among all variables using various screening methods for actual online business data [7]. To build a more stable and accurate decision tree model than one that uses all variables, an evaluation standard is proposed to reduce the error rate as much as possible under the premise of ensuring the overall error rate [10]. Specifically, it is realized by establishing a risk monitoring model with the decision tree algorithm as the core and

adjusting the parameters of the model to propose an evaluation standard to reduce the error rate as much as possible under the premise of ensuring the overall error rate by establishing a risk monitoring model with the decision tree algorithm as the core [10]. According to Ref. [9], the decision tree model has significant applicability, accuracy, and interpretability in understanding the reasons for loan default, splitting credit ratings, and lowering default rates.

Furthermore, most of these researches ignored two significant issues: the first is that, in addition to ACC, the predictive model's generalization capacity is also critical, as it may keep the trained model stable on unknown data; the second is the data set's manual feature. The high-dimensional correlation between the analytic data will be reduced by extraction, which may result in a decline in model performance. This article tries to develop a personal credit default prediction model using LightGBM, which can increase ACC and stability while also reducing human impact.

3 Data and Method

3.1 Data

The data used in this paper are from Datafountain's official website. The data of the Grand Prix published by China Computer Society and Central Plains Bank are downloaded from the official website, which contains 750,000 lines of observation data and 42 variables. The target in this paper is to predict the default rate of borrowers that asked for loans online or offline.

The 42 variables used in this paper can be roughly divided into three types of data. Generally, for the loan information applied by the borrower, the amount of loan from the bank is between 500 yuan and 40,000 yuan. In fact, the amount of loans in the data set of this paper is all in this range, with the maximum value of 40,000 yuan and the minimum value of 500 yuan. The current loan status can be simply divided into repaid, current and default, which is divided into several periods according to the overdue time. The loan interest rate is the annual interest rate set when the loan project transaction is reached. In this paper, the annual interest rate ranges between 5.31% and 30.99%, i.e., the maximum annual interest rate is 30.99% and the minimum annual interest rate is 5.31%. Similar data types include the amount payable each month, the purpose of the loan and the month in which the loan was made. Basic information about the borrower is also given, e.g., the debt-to-income ratio provided by the borrower during the registration process ranges from −1% to 999%. The length of service of the borrower ranges from 0 to 10, where 0 means less than 1 year and 10 means 10 years or more. The ownership status of the home provided by the borrower at the time of registration or obtained from a credit report is denoted as property status, which can be classified as rented, owned, mortgaged or other. Similar data types include the borrower's state, job title and number of inquiries in the last six months. Credit information of the borrower is given simultaneously. A very important data set is to give credit ratings of A to borrowers, which has A total of seven grades of A, B, C, D, E, F and G, and each grade is divided into five sub-grades of 1, 2, 3, 4 and 5, with A total of 35 grades. There is also data such as the number of accounts the borrower is currently defaulting on, the number of transactions in the last 24 months, the number of bad public records, and publicly recorded bankruptcies.

The data used in this paper is directly obtained from official websites. The original data still has problems (e.g., missing values and label leakage), hence it cannot be directly used for modelling. A series of data cleaning processes are required to establish the model for data mining including deleting the category variables as well as variables with a high missing ratio.

3.2 Random Forest

Random forest is a model based on decision trees, which was first introduced by Bell Laboratory Researchers [3]. It is mentioned that a single decision tree was famous for its high efficiency in execution. However, it becomes incompatible while dealing with arbitrary complexity. As a result, Random Forest theory was introduced to improve the inaccuracy. Then, Professor L Breiman from UC Berkeley combined Ho's theories with his theory of Bagging and then invented the exact algorithm which is available in the machine learning aspect [1].

3.3 LightGBM

LightGBM is a gradient elevation decision tree framework proposed by Microsoft Research Asia in 2016. The model has the advantages of faster and more efficient training, lower memory utilization, better accuracy, support for parallel and GPU learning, and the ability to process large-scale data, according to the official document on the model. Before introducing this algorithm, we need to introduce the gradient lifting method first.

Boosting is a logical, sequential generation learner that learns from mistakes made in the previous model and adjusts the weight of data according to the learning results of the previous learner. In addition, Boosting is a logical, sequential generation learner that learns from mistakes made in the previous model. Gradient Boosting is a new decision tree that is generated based on the negative Gradient of the loss function, i.e., a new tree is generated based on the direction of residual reduction. Therefore, the base learner is a regression tree in the CART decision tree in both regression and classification problems solved by the gradient lifting method.

The gradient lifting decision tree framework before LightGBM consumes a lot of computation and memory during the search for optimal feature divisions in the decision trees. Even XGBoost uses the pre-sorting method to split features in parallel to accelerate the training of the model, it brings about twice the memory consumption.

LightGBM uses a histogram algorithm, which first discretizes continuous floating-point data into K discrete values, and constructs a histogram of width K to calculate the cumulative statistics of each discrete value in the histogram. In feature splitting, we only need to find the optimal splitting point by traversing the histogram. The discretized feature can reduce memory consumption, not only does not need additional storage space but also can reduce the storage space required by the original data. The histogram construction is faster than the pre-sorting of XGBoost, and the histogram of the leaf node can be directly obtained by the difference between the histogram of the parent node and the histogram of the brother node. The training speed of the model is further accelerated. Data discretization will lead to certain information loss, so the split point found is not exactly the optimal segmentation point. However, according to the experiment, the split

point after discretization has little influence on the final accuracy, and the data after discretization can further prevent overfitting.

LightGBM also accelerates in two ways based on the characteristics of the gradient lifting decision tree algorithm. On the one hand, the gradient-based one-side Sampling method (GOSS) is adopted in the data. In Gradient lifting, according to the output results of the previous base learner, each data will have different Gradient values. Data with a small Gradient means that the model has been well learned. Although computations on small gradient data ensure the total data contribution, it is time-consuming for the process on the other hand. Thus, GOSS to random sampling method for small gradient data, calculate the information gain constant multiplier is introduced to offset the data distribution caused by the scene, this algorithm not only model will be more focused on not learn good data also are not subject to the change of the impact of data distribution. EFB (Exclusive Feature Bundling) is used to reduce data dimensions by Bundling Exclusive Feature Bundling to reduce data dimensions without losing information. Obviously, it is generally difficult to achieve complete Exclusive Feature Bundling. In this case, the conflict ratio is used to measure the degree of non-exclusive features. In other words, the EBF method first sorts features according to the number of non-zero values, then calculates the conflict ratio between different features, and tries to combine different features to minimize the conflict ratio.

3.4 Metrics

In this section, we measure the accuracy of these two models listed above based on the ROC curve and using calculations of AUC as criteria. ROC curves are curves that illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. To be short, the curve plots two parameters, with the vertical axis representing True Positive Rate (TPR) and with the horizontal axis representing False Positive Rate (FPR) with the following formulae:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR \tag{1}$$

$$FPR = \frac{FP}{P} = \frac{FP}{FP + FN} = 1 - TNR \tag{2}$$

After measuring all the data using the formulae, the ROC curve can be generated. Several typical ROC is represented in Fig. 1.

After the ROC was drawn, we can measure the AUC by integration. AUC measures the entire two-dimensional area under the ROC in the range and domain of (0,1) using integral calculus (see an example shown in Fig. 1). The best possible prediction method (shown as blue curve) would yield a point in the upper left corner or coordinate (0,1) of the ROC curve, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). In this curve, the AUC = 1. A random guess, for example, the dice tosses, would give a point along a diagonal line (shown as the black curve) from the left bottom to the top right corners. As the size of the sample increases, a random classifier's ROC point tends towards the diagonal line corresponding to the AUC = 0.5.

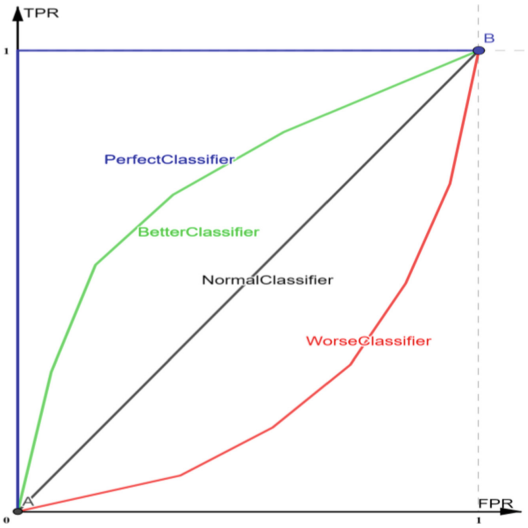

Fig. 1. Typical performance of classifiers.

Thus, it became two thresholds for measuring the accuracy of the classifier. Specifically, if the AUC of one classifier ranged from 0.5 to 1 (shown as green curve), then the classifier acts better than random guess which shows credibility in risk predictions. Else, if the AUC of the classifier ranged from 0 to 1 (shown as red curve), the classifier acts worse than random guess which is useless.

4 Results

4.1 Random Forest Model-Projections with Reliability Analysis

We measure the Pearson product-moment correlation coefficient (PPMCC) between each feature selected and the dependent variable. The absolute value of the PPMCC coefficient ranged from 0 to 1, representing the significance of the correlations between the independent variable and dependent variable.

The result is contained in Table 1.

Thus, the feature "early_return" shows the strongest correlation to the dependent feature "isDefault", indicating that the significance of the subject having early returns on the final results of default or not. After cross-validations to optimize parameters, several random forest classifier parameters are displayed. Thus, the cross-validations then measure the R-square Score of 0.85. Therefore, it proves that the prediction of the default rate fits the actual work without under-fitting and overfitting (Table 2).

The predicted several classifications results are given in Table 3 and 4. In these predictions, the majority of the objectives don't default while asking for loans. According to the results, only 71 of objects are predicted to default on loan repayment.

Then the ROC is computed and the curve is illustrated in Fig. 2. The AUC is calculated as 0.55, which indicates that the Random Forest method is only slightly better than

Table 1. PPMCC OF EACH SELECTED FEATURE

Features	PPMCC
work_year	−0.002919
house_exist	0.046535
debt_loan_ratio	0.087238
scoring_low	−0.057055
scoring_high	−0.049325
known_outstanding_loan	0.065271
recircle_b	−0.015423
recircle_u	0.038105
f0	0.092384
f1	0.004833
f2	−0.007458
f3	0.025876
f4	0.011866
early_return	−0.351473
early_return_amount	−0.293792
early_return_amount_3mon	−0.216375
isDefault	1.000000

Table 2. OPTIMIZED PARAMETERS AFTER CROSS- VALIDATIONS, (CV = 50)

Parameters	Optimized Value
max_depth	5
min_samples_leaf	3
min_samples_split	2
n_estimators	40

Table 3. COUNTS OF DIFFERENT CLASSIFICATIONS (I)

Labels Classification	Value Counts of Predictions	Value Counts of True
1 = Default	454	1274
0 = Not Default	7046	6226
Total Observations in Both Dataset	7500	7500

Table 4. COUNTS OF DIFFERENT CLASSIFICATIONS (II)

Labels Classification	Value Counts of Predictions	Value Counts of True
1 = Default	71	409
0 = Not Default	2429	2091
Total Observations in Both Dataset	2500	2500

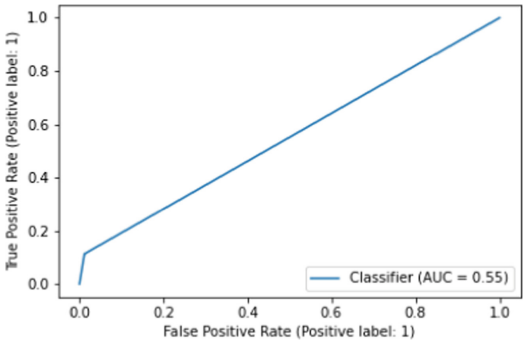

Fig. 2. ROC curve of the Random Forest Model (Figure 2 photo credit: Original)

random prediction with mere advantages. On this occasion, random forest is not suitable to be utilized as a default classifier while only the financial activities of objects are given as parameters.

4.2 LightGBM

The experimental environment in this paper is based on Windows 10 system, 16 GB memory, python3.7 program environment, code writing environment is Google Colab. The library is used for processing data in Pandas, Numpy, Matplotlib, Scikit-Learn, and LightGBM. Matplotlib and Seaborn are mainly used in the process of data visualization. Scikit-learn library is used to model random forests. LightGBM framework is modelled by Microsoft official open-source LightGBM library. The main parameter Settings of light GBM in this paper are summarized in Table 5.

The characteristic importance of LightGBM model is also displayed in Fig. 3. Thereinto, the top five are the number of outstanding credit lines in the borrower's file, the total credit turnover balance, debt-to-income ratio, the accumulated amount of payer's prepayment, and the number of months since the opening of the earliest revolving account.

Table 5. Main Parameter Setting Used in LightGBM

Parameters	Parameter meaning	Optimized Value
Boosting type	Training methods	Binary
Num_leaves	Evaluation function	20
Learning_rate	Number of leaves per tree	0.2
Feature_fraction	Learning rate feature sampling	0.7
Bagging_fraction	The proportion of sampled data without resampling	1
Bagging_freq	Resampling is performed every k iteration	3
Max_depth	The maximum depth of a tree	4

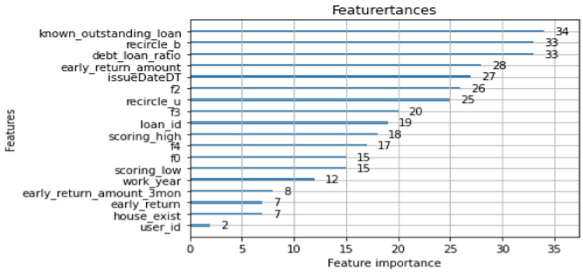

Fig. 3. LightGBM model feature importance diagram (Figure 3 photo credit: Original).

From the perspective of personal information, annual income, FICO loan and monthly debt-to-income ratio can reflect a person's loan repayment ability. Since banks are most willing to lend money to those who don't need it, loans to rich people have lower risks. Account opening records can reflect a person's awareness of instalment payment, starting instalment payment early has better repayment habits. Whereas, the more loan accounts, the greater the monthly repayment pressure.

From the perspective of loans, loan interest rate and loan amount also affect loan repayment to a large extent. The larger the loan amount, the higher the interest rate means the greater the repayment pressure. However, the larger the loan amount, the stronger the repayment ability during loan evaluation, and the lower the loan interest rate, the better the credit record.

Based on the analysis of the experimental results, after 30 iterations, the LightGBM model comes about that the best iteration is the 27th one, the AUC value of the training set is 0.909375, and the AUC value of the verification set is 0.861582. The ROC curves are illustrated in Fig. 4. It can be seen that the LightGBM algorithm is better than the random forest algorithm. The generalization capability of the model has improved a lot.

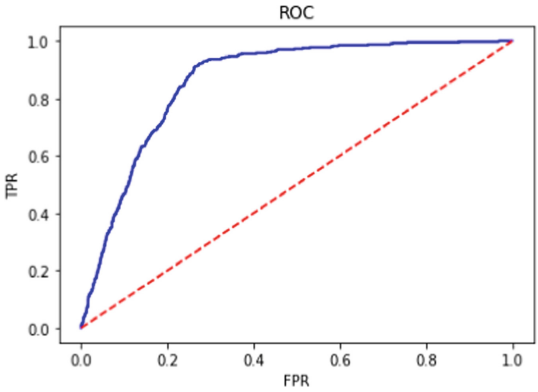

Fig. 4. ROC curve of the LightGBM Model

5 Discussion

5.1 Random Forest

The main limitation of Random Forest classifier models is the slow and inefficient real-prediction when using a large number of trees. Aside from this, Random Forest models cannot produce natural and continuous predictions when it comes to regression related problems as well. We can possibly infer from the results that users generally have very little control over what Random Forest Model does, i.e., the model feels almost like a black box. It is like a paradox to attempt to improve the prediction accuracy of this model. Improving the prediction accuracy of random forest classifiers requires us to increase the number of trees while increasing the number of trees utilized in the algorithm slow down the prediction time and can be quite ineffective in terms of real-time prediction.

While we discuss the disadvantages of random forest models, one shall keep in mind that each algorithm has its specific use cases. The Random Forest models are predictive models, rather than prescriptive models. We appreciate the ease to understand RF models, the simplicity to fine-tune RF models and the high prediction accuracy of RF models. Thanks to Random Forest models, we won't need to worry about overfitting the algorithm as long as there are enough trees in the model. All those brilliant features of Random Forest models combined together is what makes Random Forest models a better choice in terms of detecting customers who are likely to repay their debt on time and other similar tasks in the finance industry. However, that is not an excuse to stop us from looking for a better model to predict the default rate of personal loans.

5.2 LightGBM

Unlike Random Forest, which is essentially a bagging model, boosting models like LightGBM are even more prone to overfitting. They work by identifying the error at the end of a prediction cycle and continue to work on minimizing the error. They are more efficient and faster generally in terms of training the models by classifying continuous feature values into discrete bins. They also replace continuous values with discrete bins

which results in lower memory usage. Apart from higher prediction accuracy, the most important feature that LightGBM models typically have over Random Forest models is that they are able to handle larger datasets using a shorter training time.

Overall, these intriguing and excellent features don't come at a free price. Generally speaking, Light GBM models are more complex to handle and not so intuitive to comprehend compared to Random Forest models. With so many hyperparameters, we are given much more control of what the model actually does in contrast to Random Forest models, which is an absolute advantage in terms of fine-tuning the model.

6 Comparison

It is already quite obvious that the LightGBM algorithm suits better for questions like predicting default rate of personal loans. With a larger AUC of 0.91, compared to the AUC of 0.55 obtained from the Random Forest model, it is suggested that a greatly improved generalization ability at binary classification problems e.g., predicting the default rate of personal loans. In other words, it is a no brainer that the LightGBM model is a winner as well in terms of power efficiency while performing the same task compared to the Random Forest models. It utilizes less memory, less processing power and performs real-time predictions with a large number of trees and significantly improved results. With a few manual tweaking and fine-tunings, we can easily boost the performance of the LightGBM algorithm and find the best possible threshold with the least amount of classification errors, while the Random Forest model barely gives users the control of what the model actually does, which results in fewer manual adjustment possibilities.

7 Conclusion

In summary, we investigate personal loan default rate prediction based on classification regression in terms of two machine learning approaches (Random Forest and LightGBM). According to the analysis, the performance of Random Forest is rather worse with an accuracy ratio of about 55%, which is slightly larger than the random ratio (50%). However, the performance of the LightGBM is much better than the Random Forest model with an accuracy ratio of about 86%, indicating the model is less overfitting and more suitable in this case. These results pave a path and offer a new sight for future model selection for the prediction of personal loan default.

References

1. Breiman L, "Random Forests," Machine Learning, 45 (1): 5–32, Bibcode: 2001MachL..45....5B. https://doi.org/10.1023/A:1010933404324, 2001.
2. Correa Bahnsen. A, "Feature Engineering Strategies for Credit Card Fraud Detection," Expert Systems with Applications, 51, 134–142, June 2016. Available at: https://doi.org/10.1016/j.eswa.2015.12.030
3. Ho, Tin Kam, "Random Decision Forests (PDF)," Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282, June 2016.

4. L. Vinod Kumar, S. Natarajan, S. Keerthana et al., "Credit risk analysis in peer-to-peer lending system," in Proceedings of the 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), pp. 193–196, Singapore, September 2016. Available at: https://doi.org/10.1109/ICKEA.2016.7803017
5. M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," Expert Systems with Applications, vol. 42, no. 10, pp. 4621–4631, June 2015. Available at: https://doi.org/10.1016/j.eswa.2015.02.001
6. Ruan Sumei and Zhou Zelin, "Identification and Prediction of P2P Network Lending Credit Default Based on L1 Penalty Logit Model," [J] Finance and Trade Research, 29(2): 54–63, 2018.
7. Tsang. S, "De-tecting Online Auction Shilling Frauds Using Supervised Learning", Expert Systems with Applications, 41, 3027–3040, May 2014. Available at: https://doi.org/10.1016/j.eswa.2013.10.033
8. V. Ha, D. Lu, G. S. Choi et al., "Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning," in Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), pp. 511–515, PyeongChang, Korea, December 2019. Available at: https://ieeexplore.ieee.org/abstract/document/8701943
9. Wang Chenglong and Chen Cheng, "Research on the Credit Rating System of P2P Online Loan Platform Based on Decision Tree," [J] Rural Finance Research, (12): 45–50, 2016.
10. Wang Maoguang, Ge Leilei, and Zhao Jiangping, "Research on risk monitoring of micro-online loan platform based on C5.0 algorithm," [J] China Management Science, 24(S1): 345–352, 2016.
11. Wang Xiangting, Zhao Zixuan, Wang Shutan, and Liu Ningning, "Research on Default Prediction of Online Lending Borrowers Based on Machine Learning," Service Science and Management, 8, 40–48. https://doi.org/10.12677/SSEM.2019.81006, January 2019. Avaliaeble at: https://m.hanspub.org/journal/paper/28465
12. Xiong Zhengde, Liu Zhenxuan, and Xiong Yipeng, "Research on Internet Financial Customer Default Risk Based on Orderly Logistic Model," [J] Systems Engineering, 35(8): 29–38, 2017.