

An improved light gradient boosting machine algorithm based on swarm algorithms for predicting loan default of peer-to-peer lending

Much Aziz Muslim^{1,2}, Yosza Dasril¹, Muhammad Sam'an¹, Yahya Nur Ifriza³

¹Department of Technology Management, Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

²Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Apr 18, 2022

Revised Aug 26, 2022

Accepted Sep 5, 2022

Keywords:

Ant colony optimization

Bee colony optimization

Features selection

LightGBM

Loan default

P2P lending

ABSTRACT

Internet finance and big data technology are booming in the world. The launch of peer to peer (P2P) lending platforms is a sign and a great opportunity for entrepreneurs to easily increase their capital injection. However, this great opportunity has a high risk of impacting the sustainability and security development of the platform. One way to minimize loan risk is to predict the possibility of loan default. Hence, this study aims to find the best predictive model for predicting loan default of P2P Lending Club dataset. An improved light gradient boosting machine (LightGBM) via features selection by using swarm algorithms i.e. Ant colony optimization (ACO) and bee colony optimization (BCO) to the prediction analysis process. The best feature selection process is selected 6 out of 18 features. The synthetic minority oversampling technique (SMOTE) method is also provided to solve the unbalance class problem in the dataset, then a series of operations such as data cleaning and dimension reduction are performed. The experimental results prove that the LightGBM algorithm has been successfully improved. This success is shown by the prediction accuracy of LightGBM+ACO is 95.64%, LighGBM+BCO is 94.70% and LightGBM is 94.38%. This success also demonstrates outstanding performance in predicting loan default and strong generalizations.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Much Aziz Muslim

Postgraduate Student, Faculty of Technology Management and Business

Universiti Tun Hussein Onn Malaysia (UTHM)

Batu Pahat, Johor, 86400, Malaysia

Email: a212muslim@mail.unnes.ac.id

1. INTRODUCTION

Big data and Internet finance (Fintech) are currently trending and are often discussed in the world as internet lending industry is known as peer to peer (P2P) Lending. P2P lending as a Fintech platform has a unique characteristic in transactions, namely connecting individual loan borrowers to individual lenders or investors to make credit agreements and complete transaction procedures directly through the online platform, without commercial bank intermediaries. Gradually, the existence of P2P lending has become a solution for small and medium businesses to get loan capital so easily that every year the loan amount is very large. As reported by LendingClub Corporation [1] about "Fourth Quarter and Full Year 2019 Results" show that the loan amount had achieved US \$ 12,290.1 billion at the end of 2019. While Stern *et al.* [2]'s data showed that China's government noted that China became the most P2P loan platforms predicate of the investment market with quantitating to around 2.300 as of March 2017 and CNY 9.208 loan volume.

P2P lending presents an opportunity as well as a challenge, including China as a developed economy. In order to a large extent, P2P lending meets China's current economic needs as well as its risks. Financial risk can be seen from liquidity risk caused by insufficient liquidity funds, unbalanced information as a cause of credit risk and legal risk caused by unclear laws governing Fintech. In brief, the risk characteristics of Fintech are more complex than conventional finance. In addition, technical and virtual-based Fintech also triggers special risks that arise such as conventional financial risks as financial risks are sudden and spreading, besides that the increase in destructive risks is very serious and uncontrollable Challa *et al.* [3] said risk aversion is one of the hot topics, interesting and very important to be discussed among investors, policymakers, financial practitioners and made studies by researchers.

Generally, research and application of loan evaluation in P2P lending platforms are given two main directions. First, the use of credit scores to evaluate the credit risk of loans and second, transforming loan evaluations into a binary classification. A credit scorecard is a conventional loan evaluation method. Usually, Chen and Han [4] explained these scorecards are self-launched by P2P lending platforms for business needs, for example, Fair Isaac Corporation (FICO) score and LendingClub score. However, according to Malekipirbazari and Aksakalli [5], credit scorecards cannot distinguish between defaulter and non-defaulter. As big data technology matures many researchers use machine learning techniques to predict whether a loan can be returned or a loan repayment is due in P2P lending platform. Light gradient boosting machine (LightGBM) is a machine learning algorithm that is used as a classification. LightGBM is an improved version of the gradient learning framework based on decision trees and "weak" learner ideas. Since being developed by Microsoft in 2017 [6]. Since LightGBM was introduced in 2016, several researchers have applied the big data machine learning Algorithm in various fields and produce predictions with very high accuracy, fast-computationally and well-performance in minimizing relative over-fitting. Such as, web search, Breast cancer to identify miRNAs [7], the default accuracy prediction of P2P lending platform [8]-[11], music recommendation [12], the classification of acoustic scene [13], smart grid load forecasting [14], estimation of reference evapotranspiration of agricultural or hydrological [15], construction cost prediction [16], predict customer loyalty Fintech [17] and stream processing prediction [18].

LightGBM is known as an algorithm that is fast data learning, faster when handling big data, high accuracy, good model precision, low data memory consumption so that this algorithm is considered more effective and efficient than other machine learning techniques [8], [19]. According to Rao *et al.* [20], feature selection in a big data set as a significant phase performs several tasks such as image classification, cluster analysis, data mining, pattern recognition, and image capture [21], [22]. Many methods have been proposed, improved and discussed for feature selection. Alickovic and Subasi [23]-[24] improved whale optimization algorithm (WOA) to optimize features in the dataset. Zhu *et al.* [25] presented a method of uncontrolled spectral feature selection to maintain local and global features of the feature during the redundant feature removal process. Wan and Freitas [26] evaluated the hierarchy method in optimizing the feature selection of aging related gene data sets. Rao *et al.* [20] used artificial bee colony and gradient boosting decision tree to select features of eight UCI data sets and produced and the experimental results proven that Rao's method is able to reduce the dimensions of the data set and achieve superior classification accuracy. Ghosh *et al.* [27] improved the wrapper-filter feature selection method based on ant colony optimization to reduce computational complexity.

Based on the previous research described above, increasing prediction accuracy via feature selection techniques is focus of this study. Therefore, we use two swarm algorithms, i.e. ant colony optimization (ACO) algorithm and Bee Colony Optimization (BCO) algorithm as a feature selection and LightGBM as a tool to evaluate P2P lending data sets. This study aims to determine the two swarm algorithms performance in the feature selection process, then the prediction performance of the LightGBM algorithm. In addition, we also use the synthetic minority oversampling technique (SMOTE) to address data class imbalances. This technique is believed to also be able to improve the accuracy of predictions as has been proven by Faris *et al.* [28] to predict the bankruptcy of companies with highly imbalanced data classes.

2. RELATED KNOWLEDGE AND THEORY

2.1. Lending club

The lending club has lanced an impact on risk management. Loan applications can be approved are very small, around 10% of all applications. In addition, there are lending club levels i.e. A to G to classify loans based on risk. The main role of the Lending club is to make it easier for borrowers and lenders or investors to transact and provide information related to However. In fact, there are many problems in this transaction model, such as loan money is not returned by the borrower according to the agreement so that investors experience losses. Determining loan interest rates according to loan credit and loan term. The Lending club business pattern is shown in Figure 1.

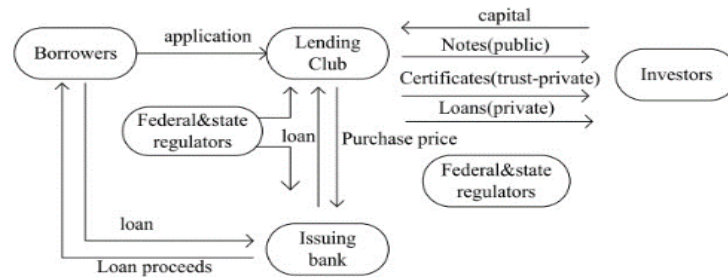


Figure 1. Lending club business pattern [29]

2.2. Bee colony optimization

Bee colony optimization (BCO) is one of bio-inspired methods by bees gathering nectar behavior [30]. Akay and Karaboga [31] said that the value of global optimum is determined by neighborhood search optimization of each bee. Wen *et al.* [32] said that artificial bee colony method able to locate the global optimum solution for the global optimization problems. When compared to other bio-inspired heuristic algorithms, BCO has many strengths i.e. a simple structure, requires few control parameters and is easy for implementing [33]. Because of this strength, BCO has attracted the attention of researchers to study and apply it in various fields.

2.3. Ant colony optimization

Ant colony optimization (ACO) is one of bio-inspired algorithms by ant colony behavior [34]. Ants cannot see. However, through indirect communication, the ants can find the shortest route from nest to the food source [35]. Ants modify their environment (by disguising pheromone) to influence another ant behavior is named Stigmergy. The concept of ACO algorithms for foraging ant behavior. Algorithms often discussed and applied are ant system (AS), ant colony system (ACS), max-min ant system (MMAS). In solving the optimization problem using the ACO algorithm, several artificial ants are used to model the solution iteratively. For each iteration, the ants will store a certain amount of pheromone which is proportional to solution quality. In each rarity, Tabakhi and Moradi [36] explained that the ant calculates a series of feasible solutions to the current partial solution and one of the choices depends on two factors i.e. local heuristics and prior knowledge, three phases need to be addressed i.e. Graph representation, Heuristic desirability and Pheromone update rule.

2.4. Light gradient boosting machine

Light gradient boosting machine (LightGBM) is a fast and efficient gradient boosted decision tree (GBDT) algorithm with an open-source promotion work objective that was created by Microsoft MSRA in 2016. This algorithm is used for sorting, classification, regression, and many other machine learning techniques assignments and supports efficient parallel training. In contrast to XGBoost, LightGBM algorithm uses a histogram to speed up the training process, reduce memory space, and implement a wise growth strategy with depth constraints. The basic idea of LightGBM using a histogram is to discrete the continuity of floating-point eigenvalues to k bins and create a histogram with a width of k . LightGBM does not require large storage of pre-sorted results, can store 8-bit integers and can also reduce memory consumption to 1/8 of the original. This rough partition does not reduce the mode of LightGBM accuracy. The LightGBM is a boosting type that has three steps. For simplicity, X is given as a pre-processed streaming data set.

Step 1. Initialize the weak learner by (1).

$$f_0(x) = \operatorname{argmin}_c \sum_{i=1}^n L(y_i, c) \quad (1)$$

where: $f_0(x)$ as the weak learner basis function, $L(y_i, c) = L(y, f(x)) = (y - f(x))^2$ as the function of loss, n as the amount of samples.

Step 2. Calculate weak learners M times, Iteratively.

- a. For the sample $x_i \in X \forall i = 1, 2, \dots, n$ calculate the negative gradient of loss function evaluated in the existing model in (2).

$$r_{mi} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{m-1}(x)} \quad (2)$$

where r_{mi} as negative gradient of the loss function.

b. The residual r_{mi} resulted is taken as sample new real value. Fit a regression tree for;

$\{(x_1, r_{m1}), \dots, (x_n, r_{mn})\}$ and make a new regression tree $f_m(x)$.

c. Calculate the best-fit value of the leaf area $j = 1, 2, \dots, J$. By using c_{mj} in (3) as linear search to predict leaf node region value for minimizing the loss function.

$$c_{mj} = \operatorname{argmin}_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i + c)), i = 1, \dots, M. \tag{3}$$

d. Update the robust learner by using (4).

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^J c_{mj} I(x \in R_{mj}) \tag{4}$$

where $f_m(x)$ as the the existing weak leaner, $f_{m-1}(x)$ as pre-weak leaner, I as the indicator function.

Step 3. Determine the final regression tree by using (5).

$$F(x) = \sum_{m=1}^M \sum_{j=1}^J c_{mj} I(x \in R_{mj}) \tag{5}$$

The significance of a feature is calculated as the normalized total reduction of criterion brought by that feature. It is also known as the Gini significance Gini is denoted by Gini (p) in (10).

$$Gini(p) = \sum_{l=1}^L P_l(1 - p_l) = 1 - \sum_{k=1}^L p_k^2 \tag{6}$$

where: L as the number of labels p_k as the weight of l-label.

3. RESEARCH METHOD

The research method of loan default of P2P lending prediction analysis uses several phases, i.e. Dataset pre-processing, data oversampling, ensemble classification and performance evaluation. Generally, the research framework can be shown in Figure 2.

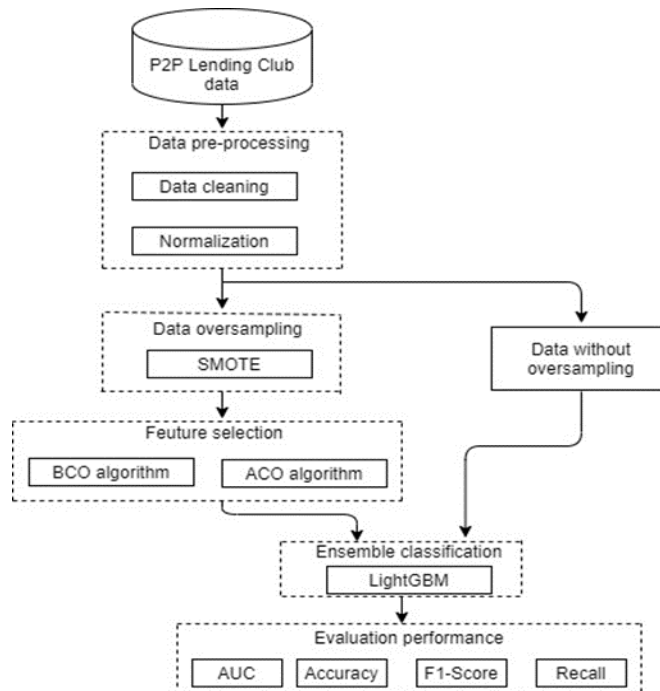


Figure 2. Framework of study

3.1. Dataset pre-processing

Data pre-processing is a sequence of process parts that are practiced to prepare the dataset for analysis and modeling. Therefore, this phase is believed to be an important step in the data mining process [37]. In this study, data preprocessing includes data cleaning, data normalization, and data retrieval. In data cleaning, missing values, inconsistencies and noise (e.g., incorrect data input) are eliminated [38]-[41]. We use the Lending club data set for the 2019 quarter downloaded from Kaggle.com containing 20,875,146 original user loans with 18 attributes. Furthermore, after data pre-processing, the missing value is filled via interpolation mode and multiple or not effect attributes are removed so that we get six attributes and Table 1 shows the attributes used in the experiment.

Table 1. The selected attributes and pre-processing

Feature name	Description and pre-processing	Type	Algorithm
amount_borrowed	the principal amount of the loan upon which interest will accrue	numeric	A,B,C
borrower_rate	the interest rate at which money may be borrowed	numeric	A,B,C
installment	the monthly payment owed by the borrower if the loan originates	numeric	A,B,C
principal_paid	a payment toward the original amount of a loan that is owed	numeric	A, B, C
interest_paid	a payment of interest on a loan or mortgage	numeric	A, B, C
grade	lending club assigned loan grade	nominal	B
term	the loan repayment amount the Value represented 36 months from binary number to discretization	numeric	A
loan_status	the source of our answer to the core question if people are paying the loans they take out	nominal	C

Note: A = LighGBM without swarm algorithms, B = LighGBM with BCO algorithm, C = LighGBM with ACO algorithm

3.2. Synthetic minority oversampling technique

Based on data pre-processing 3.1, significant differences in the number categories of normal and default on target variable 'loan_status' can complicate learning modeling. SMOTE is an oversampling method to overcome imbalanced data sets, the SMOTE rationale as follows [29],

- To calculate the K-nearest neighbor of each minority sample with the Euclidean Distance as the standard, the neighbor algorithm is used.
- Adjusting a sampling proportion with the unbalance sample proportion and each sample x minority class, a few samples are randomly selected from its K-neighbors.
- Suppose x_n is the selected neighbor. For each randomly selected neighbor x_n , a new sample can be generated using (11) with the respective original samples.

$$x_{new} = x_i + rand(0,1) * |x - x_n| \quad (7)$$

By iteratively, for each sample x_i , the original sample size of minority class can be widened to an ideal ratio.

3.3. Feature selection

First, we define the "installment" feature to represent the user's monthly fee payment as a percentage of their monthly revenue. The greater the "installment" value, the more loans provided by investors will be more burdened and tend to default. Second, feature abstraction. We encoded the loan status 'Current', 'Completed' as usual=0, encoding 'Default', 'Charge off' and 'Canceled' as default=1. Next, we plot loan_status. That 89% of loan_status is "Default" and the rest is only 11% for "Normal". Based on these results, it indicates a serious imbalance of datasets. After scaling the features, third is feature selection. The selected feature attributes have high relevance or correlation value and remove irrelevant features or low correlation. This elimination can reduce difficulties in the training process. We use swarm algorithms i.e. ACO and BCO to select 6 features with the strongest correlation with the target variable and remove features step by step to achieve the reduction of the first dimension with variables 18 to 6. We illustrate a Pearson correlation graph of 18 features, as shown in Figure 3.

Meanwhile, the results of the reduction of the first dimension, the redundant features are selected and removed using the Pearson correlation graph based on the swarm algorithm used. The feature dimensions reduced from 18 to 6 are shown in Figure 4, Figure 4(a) shown that features selection of BCO algorithm is amount_borrowed, borrower_rate, installment, principal_paid, grade and on Figure 4(b) shown that features selection of ACO algorithm is amount_borrowed, borrower_rate, installment, principal_paid and loan_status.

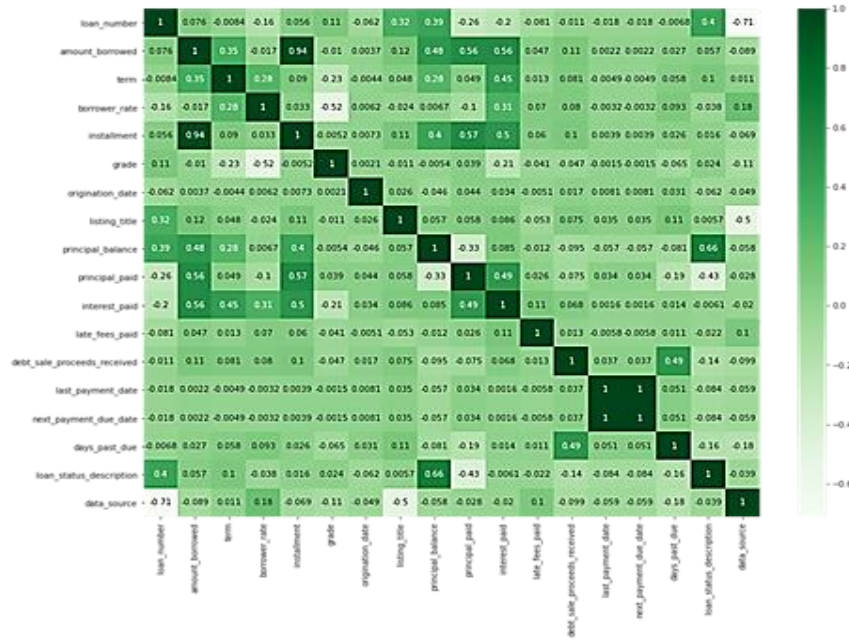
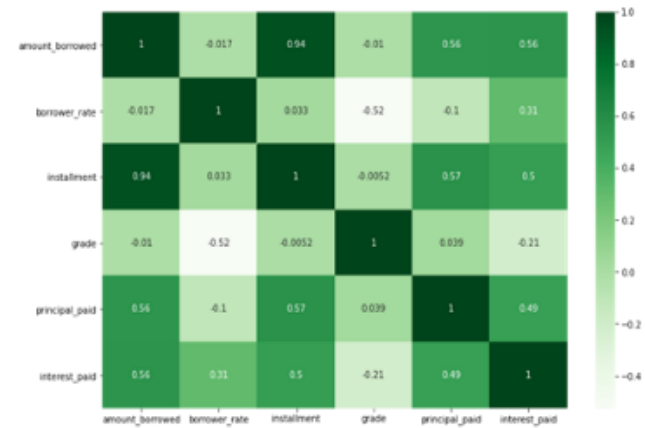


Figure 3. Person correlation of 18 features



(a)



(b)

Figure 4. Person correlation of 6 features, (a) The features selection of BCO algorithm and (b) The features selection of ACO algorithm

Population correlation coefficient is formulated as the covariance and standard deviation between two variables. Predict covariance and standard deviation of sample to determine the Pearson correlation coefficient of sample. Finally, we use the swarm algorithm i.e. ACO and BCO to select the importance of the feature and reduce the learning difficulty to optimize the model calculation.

3.4. The evaluation performance model

In this study, we use three parameters i.e. accuracy, AUC and ROC to evaluate and assess the performance of our proposed model. Accuracy is the ratio of the number of correct sample classifications to the total number of samples for a particular test data set as shown in (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

where: TP=True Positives, TN=True Negatives, FP=False Positives and FN=False Negatives.

Recall is called the fraction of all positive instances (default) where the classifier categorizes true as positive or known as the TP ratio. A balanced F score or F1-score is called the balanced average of Precision and Recall.

3.4.1. Receiver operating characteristic curve

In statistics, receiver operating characteristics or ROC known as a two-dimensional graphical plot illustrates the performance of a binary classifier. The curve of ROC is made in various threshold settings by plotting true positive ratio (TPR) to the false positive ratio (FPR) by using (9). Intuitively, this curve represents the performance of the classifier.

$$TPR = \frac{TP}{TP+FN} \quad FPR = \frac{FP}{FP+TN} \quad (9)$$

3.4.2. AUC value

The AUC represents the area under the curve of ROC in the test data-set. Suppose that the curve of ROC is formed by a sequential relationship of points with coordinates of $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$. Thus, the value of AUC can be formulated by using (10).

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1}) \quad (10)$$

where the AUC value range is [0.5,1.0] and if the AUC value is almost 1.0 then the classifier has a good performance.

4. RESULTS AND DISCUSSION

In research, the improved LightGBM algorithm as a classifier via features engineering or feature selection using a swarm algorithm i.e. ACO and BCO are evaluated and assessed their performance using several parameters i.e. accuracy, AUC, F1-Score, recall and ROC curves. The results obtained are shown in Table 2.

Table 2. The evaluation metrics comparison of the proposed model

Classifier model	Accuracy	AUC	F1-score		Recall		Rank
			0	1	0	1	
LighGBM+ACO	95.64 %	0.956	0.97	0.97	0.96	0.97	1
LighGBM+BCO	94.70 %	0.947	0.93	0.93	0.94	0.93	2
LighGBM	94.38 %	0.943	0.90	0.90	0.90	0.92	3

Table 2 shows that the performance of the LightGBM algorithm increases after the application of feature selection using the swarm algorithm. The performance of LightGBM+ACO algorithm is superior to LightGBM+BCO algorithm and LightGBM without swarm algorithm. Precision and Recall prediction models based on LightGBM using either the evaluation algorithm or not, all above 0.90. This value indicates that the model has strong generalizability. Meanwhile, the ROC curve graph is illustrated in Figure 5. This table shows that the closer the ROC curve is to upper left corner, the higher the prediction rate of model. The point of the ROC curve closest to upper left corner is best classification with lowest error based on the maximum threshold and the least total number of FPR and TPR. So from the curve, we can conclude that the LightGBM+swarm is superior to LightGBM.

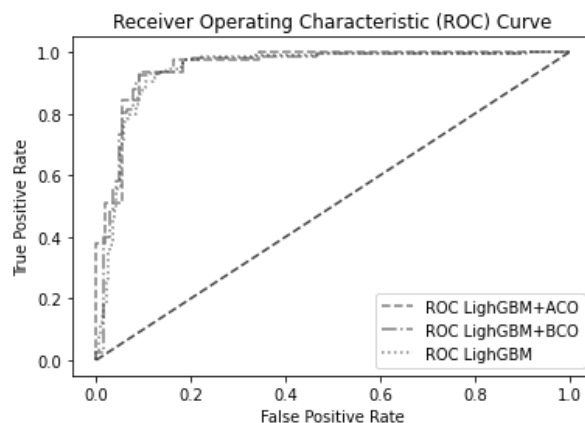


Figure 5. The ROC curve performance comparison LightGBM+ACO, LightGBM+BCO and LightGBM

5. CONCLUSION

In this study, the LightGBM algorithm is improved through feature engineering or feature selection using the BCO algorithm and the ACO algorithm to create a P2P loan evaluation model, especially the prediction of credit defaults. The experiment uses data sets from kaggle.com to show that improved LightGBM is successful. The best feature selection process is selected 6 out of 18 features. The SMOTE method is also provided to solve the unbalance class problem in the dataset, then a series of operations such as data cleaning and dimension reduction are performed. The experimental results prove that the LightGBM Algorithm has been successfully improved. This success is shown by the prediction accuracy of LightGBM + ACO is 95.64%, LighGBM + BCO is 94.70% and LightGBM is 94.38%. This success also demonstrates outstanding performance in predicting loan default and strong generalizations.

ACKNOWLEDGEMENTS




The authors would like to express their gratitude and appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM) through the research grant TIER 1 (H777).

REFERENCES




- [1] G. Attigeri, M. M. Manohara Pai, and R. M. Pai, "Framework to predict NPA/Willful defaults in corporate loans: A big data approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 3786–3797, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3786-3797.
- [2] C. Stern, M. Makinen, and Z. Qian, "FinTechs in China – with a special focus on peer to peer lending," *Journal of Chinese Economic and Foreign Trade Studies*, vol. 10, no. 3, pp. 215–228, Oct. 2017, doi: 10.1108/JCEFTS-06-2017-0015.
- [3] M. L. Challa, V. Malepati, and S. N. R. Kolusu, "Forecasting risk using auto regressive integrated moving average approach: an evidence from S&P BSE Sensex," *Financial Innovation*, vol. 4, no. 1, p. 24, Dec. 2018, doi: 10.1186/s40854-018-0107-z.
- [4] D. Chen and C. Han, "Comparative Study of online P2P Lending in the USA and China," *Journal of Internet Banking and Commerce*, 2012.
- [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, Jun. 2015, doi: 10.1016/j.eswa.2015.02.001.
- [6] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 3147–3155, 2017.
- [7] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," in *ACM International Conference Proceeding Series*, 2017, pp. 7–11, doi: 10.1145/3155077.3155079.
- [8] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications*, vol. 31, pp. 24–39, Sep. 2018, doi: 10.1016/j.elerap.2018.08.002.
- [9] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122370, Nov. 2019, doi: 10.1016/j.physa.2019.122370.
- [10] Y. Wang and X. S. Ni, "Improving investment suggestions for peer-to-peer lending via integrating credit scoring into profit scoring," in *ACMSE 2020 - Proceedings of the 2020 ACM Southeast Conference*, Apr. 2020, pp. 141–148, doi: 10.1145/3374135.3385272.
- [11] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," *Information Sciences*, vol. 525, pp. 182–204, Jul. 2020, doi: 10.1016/j.ins.2020.03.027.
- [12] W. Zhang, H. Quan, and D. Srinivasan, "Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination," *Energy*, vol. 160, pp. 810–819, Oct. 2018, doi: 10.1016/j.energy.2018.07.019.

- [13] E. Fonseca, R. Gong, D. Bogdanov, O. Slizovskaia, E. Gomez, and X. Serra, "Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, no. November, pp. 1–5, 2017.
- [14] Q. Zhang, N. Cui, Y. Feng, D. Gong, and X. Hu, "Improvement of Makkink model for reference evapotranspiration estimation using temperature data in Northwest China," *Journal of Hydrology*, vol. 566, pp. 264–273, Nov. 2018, doi: 10.1016/j.jhydrol.2018.09.021.
- [15] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, "Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data," *Agricultural Water Management*, vol. 225, p. 105758, Nov. 2019, doi: 10.1016/j.agwat.2019.105758.
- [16] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Advanced Engineering Informatics*, vol. 46, p. 101201, Oct. 2020, doi: 10.1016/j.aei.2020.101201.
- [17] M. R. Machado, S. Karray, and I. T. De Sousa, "LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry," in *14th International Conference on Computer Science and Education, ICCSE 2019*, Aug. 2019, pp. 1111–1116, doi: 10.1109/ICCSE.2019.8845529.
- [18] Z. Chu, J. Yu, and A. Hamdulla, "LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network," *Information Sciences*, vol. 535, pp. 107–129, Oct. 2020, doi: 10.1016/j.ins.2020.05.042.
- [19] M. A. Muslim, A. Nurzahputra, and B. Prasetyo, "Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction," in *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, Mar. 2018, vol. 2018-January, pp. 141–145, doi: 10.1109/ICOIACT.2018.8350753.
- [20] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing Journal*, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/j.asoc.2018.10.036.
- [21] B. Prasetyo, Alamsyah, and M. A. Muslim, "Analysis of building energy efficiency dataset using naive bayes classification classifier," *Journal of Physics: Conference Series*, vol. 1321, no. 3, p. 032016, Oct. 2019, doi: 10.1088/1742-6596/1321/3/032016.
- [22] A. Nurzahputra, M. A. Muslim, and B. Prasetyo, "Optimization of C4.5 algorithm using meta learning in diagnosing of chronic kidney diseases," *Journal of Physics: Conference Series*, vol. 1321, no. 3, p. 032022, Oct. 2019, doi: 10.1088/1742-6596/1321/3/032022.
- [23] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and rotation forest," *Neural Computing and Applications*, vol. 28, no. 4, pp. 753–763, Apr. 2017, doi: 10.1007/s00521-015-2103-9.
- [24] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Applied Soft Computing*, vol. 62, pp. 441–453, Jan. 2018, doi: 10.1016/j.asoc.2017.11.006.
- [25] X. Zhu, S. Zhang, R. Hu, Y. Zhu, and J. Song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 517–529, Mar. 2018, doi: 10.1109/TKDE.2017.2763618.
- [26] C. Wan and A. A. Freitas, "An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features," *Artificial Intelligence Review*, vol. 50, no. 2, pp. 201–240, Aug. 2018, doi: 10.1007/s10462-017-9541-y.
- [27] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7839–7857, Jun. 2020, doi: 10.1007/s00521-019-04171-3.
- [28] H. Faris *et al.*, "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market," *Progress in Artificial Intelligence*, vol. 9, no. 1, pp. 31–53, Mar. 2020, doi: 10.1007/s13748-019-00197-9.
- [29] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.
- [30] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied Soft Computing Journal*, vol. 8, no. 1, pp. 687–697, Jan. 2008, doi: 10.1016/j.asoc.2007.05.007.
- [31] B. Akay and D. Karaboga, "A modified Artificial Bee Colony algorithm for real-parameter optimization," *Information Sciences*, vol. 192, pp. 120–142, Jun. 2012, doi: 10.1016/j.ins.2010.07.015.
- [32] G. K. Wen, Y. Bin Dasril, N. Bujang, M. Mohamad, M. D. H. Gamal, and L. C. Soon, "Hybridization gradient descent search with artificial bees colony algorithm in general global optimization problems," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 4, pp. 999–1008, 2021.
- [33] M. Schiezzaro and H. Pedrini, "Data feature selection based on Artificial Bee Colony algorithm," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 47, Dec. 2013, doi: 10.1186/1687-5281-2013-47.
- [34] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Systems with Applications*, vol. 36, no. 3 PART 2, pp. 6843–6853, Apr. 2009, doi: 10.1016/j.eswa.2008.08.022.
- [35] Y. N. Ifriza and M. Sam'an, "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *Journal of Soft Computing Exploration*, 2021, [Online]. Available: <https://shmpublisher.com/index.php/joscecx/article/view/42>.
- [36] S. Tabakhi and P. Moradi, "Relevance-redundancy feature selection based on ant colony optimization," *Pattern Recognition*, vol. 48, no. 9, pp. 2798–2811, Sep. 2015, doi: 10.1016/j.patcog.2015.03.020.
- [37] C. F. Tsai and K. C. Cheng, "Simple instance selection for bankruptcy prediction," *Knowledge-Based Systems*, vol. 27, pp. 333–342, Mar. 2012, doi: 10.1016/j.knosys.2011.09.017.
- [38] G. Mogos and N. S. Mohd Jamail, "Study on security risks of e-banking system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 1065–1072, Feb. 2020, doi: 10.11591/ijeecs.v21.i2.pp1065-1072.
- [39] S. Kim and K. You, "Data analysis of financial burden index through KBO league FA pitcher's performance and contract amount size," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2525–2562, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2555-2562.
- [40] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5549–5557, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.
- [41] K. Budiman and Y. N. Ifriza, "Analysis of earthquake forecasting using random forest," *Journal of Soft Computing Exploration*, 2021, [Online]. Available: <https://www.shmpublisher.com/index.php/joscecx/article/view/51>.




BIOGRAPHIES OF AUTHORS

Much Aziz Muslim    PhD candidate in the faculty of management technology at Universiti Tun Hussein Onn Malaysia (UTHM). The scope of research he is currently working on is in the fields of Data Mining. besides that he is also a lecturer in the computer science department of the Universitas Negeri Semarang. He can be contacted at email: a212muslim@mail.unnes.ac.id.






Yosza Dasril    received PhD and master's degree degree in Applied Mathematics from Univeriti Putra Malaysia and Bachelor Degree in Mathematics from Universitas Riau, Indonesia. He is a Lecturer at Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia. His research interests are in Optimization, Engineering Mathematics. He can be contacted at email: yosza@utem.edu.my.



Muhammad Sam'an    graduated in Master of Mathematics from Universitas Diponegoro in 2018. Currently Lecturer at department of computer sciences, Universitas Muhammadiyah Semarang. He has interested in fuzzy optimization and operation research. He can be contacted at email: muhammad.92sam@gmail.com.



Yahya Nur Ifriza    graduated in Master of Informatic System from Universitas Diponegoro in 2017. Currently, he is a Lecturer at department of computer sciences, Universitas Negeri Semarang. He has interested research in data mining and wireless sensor network. He can be contacted at email: yahyanurifriza@mail.unnes.ac.id.