9 (1) January 2023 18-28
Register: *Jurnal Ilmiah Teknologi Sistem Informasi*

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)

Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

Credit Risk Assessment in P2P Lending Using LightGBM and Particle Swarm Optimization

Yosza Dasril [a,], Much Aziz Muslim [b], M. Faris Al Hakim [c], Jumanto [d], Budi Prasetiyo [e]*

[a,b] *Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia*
[b,c,d,e] *Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia*
email: [a,*] *yosza@uthm.edu.my*, [b] *a212muslim@yahoo.com*, [c] *farishakim@mail.unnes.ac.id*, [d] *jumanto@mail.unnes.ac.id*, [e] *bprasetiyo@mail.unnes.ac.id*
* *Correspondence*

ARTICLE INFO

Article history:
Received 2 November 2022
Revised 2 January 2023
Accepted 18 February 2023
Available online 22 February 2023

Keywords:
LightGBM,
PSO,
Credit Risk Assessment,
P2P Lending
Machine Learning

Please cite this article in IEEE style as:
Y. Dasril, M. A. Muslim, M. F. Al Hakim, J. Jumanto and B. Prasetiyo, "Credit Risk Assessment in P2P Lending Using LightGBM and Particle Swarm Optimization," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 18-28, 2023.

ABSTRACT

Credit risk evaluation is a vital task in the P2P Lending platform. An effective credit risk assessment method in a P2P lending platform can significantly influence investors' decisions. Machine learning algorithm such as LightGBM can be used to evaluate credit risk. However, the results in evaluating P2P lending need to be improved. This research aims to improve the accuracy of the LightGBM algorithm by combining it with the Particle Swarm Optimization (PSO) algorithm. This research is novel as it combines LightGBM with PSO for large data from the Lending Club Dataset, which can be accessed on Kaggle.com. The highest accuracy also presented satisfactory results with 98.094% accuracy, 90.514% Recall, and 97.754% NPV, respectively. The combination of LightGBM and PSO has resulted in better outcome.

1. Introduction

Finance technology (fintech) has grown rapidly along with internet technology's improvement and big data's emergence. The integration of internet technology and the financial sector as an effective and efficient practical subject has been developed out of the widespread interest shown by world-class official and unofficial institutions, organizations or governments, and finance departments in various countries. In several conference sessions on world finance, the World Bank has examined and provided evaluations on the development of fintech. In fact, China is aggressively promoting policies related to fintech. According to china.org.cn, the term "Financial Inclusion" was in the resolution of the 18th central committee of the Chinese Communist Party's third plenary session on November 12th, 2013 (China.Org, 2015). Based on the China Digital Financial Inclusion Development report on June 5th, 2017, the Chinese government officially launched the "national Fintech sunshine project" [1]. The primary key of fintech is ensuring healthy financial development without experiencing any loss.

Peer to Peer Lending (P2P Lending) is a typical fintech representative with the primary concept of "inclusion." P2P Lending is a money lending method used by individuals or businesses [2]. The advantages of P2P lending are that it is more convenient, fast, and transparent in transactions than conventional financing methods. P2P lending directly connects investors with borrowers and ensures that both parties do not experience any loss. In addition, P2P lending platforms can result in lower borrowing costs, safer transactions, and higher returns at a fixed rate than conventional methods used by investors. In the investment market of America, P2P Lending is growing rapidly with a growth rate of more than 100% year to year [3]. By 2019, LendingClub's "Fourth Quarter and Full Year 2019 Results" showed US $12,290.1 billion in loans. (LendingClub Corporation, 2020). Meanwhile, in several Asian countries such as Korea, China, and Indonesia, P2P Lending is a fintech sector with more positive growth compared to other sectors [4], [5]. As of March 2017, the government of China asserted that China has the highest number of peer-to-peer loan platforms based on the investment industry, with approximately 2,300 platforms and CNY 9,208 loan volume [5].

There are two role categories in P2P lending: the borrower as a money borrower and the lender as a lender of money to borrowers. Borrowers who want to borrow money make a loan list on the P2P Lending platform and make an appointment. As shown in Figure 1, each loan has a limit on the amount of the loan and the duration of repayment (often less than a month), and all borrowers must meet the requirements and rules set by the lenders [6], [7]. However, the rapid development of the P2P lending platforms also comes with many problems. As described by [8], an incident in the city of Shuozheng Xu, China in 2016, showed that the P2P lending platform plays an essential role in the lives of the surrounding community, so there are various perceptions or opinions regarding P2P lending platforms. Some think that the P2P lending platform is a "financial innovation," and there are also those who call the P2P lending platform an "illegal fundraising" and a "Ponzi scheme." Based on the development of P2P lending platforms in China, there are a huge number of cases (including default and foreclosure platforms) that have caused large losses to lenders or investors. This case can also be found in developing countries, including Indonesia. Therefore, risk management control is the key to decreasing deficiencies on P2P lending platforms.

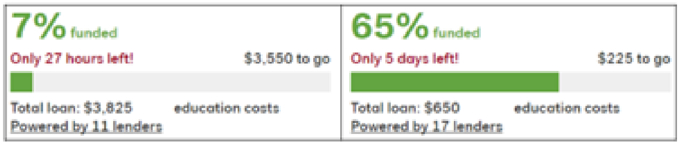

Fig. 1. Loans in P2P Lending [9]

Nowadays, big data is growing rapidly in the world investment market, including big data from the P2P lending industry. Implementing machine learning techniques is one of the most powerful and efficient approaches in data mining studies, which helps provide information based on big data analysis. P2P lending risk analysis using machine learning techniques can be carried out to determine the risks of P2P lending implementation [10],[11]. According to [12], they stated that the regulations for P2P lending are still unclear, which puts consumers at risk. Therefore, we can use a learning machine approach to control risk management in the P2P lending industry.

LightGBM uses a machine learning-based classifier approach. LightGBM is an enhanced decision tree-based gradient learning system with "weak" learner notions. Since the publication of LightGBM in 2016, several academicians have employed the big data machine learning algorithm to make predictions with extremely high accuracy, quick calculation, and outstanding performance in reducing relative over-fitting. Web search, Breast cancer to discover miRNAs [13], P2P lending platform default accuracy [8], [14], music recommendation [15], acoustic scene classification [16], smart grid load forecasting [17], estimation of reference evapotranspiration of agricultural or hydrological [18], prediction of construction cost [19], prediction of customer loyalty for fintech [20], and prediction of stream processing [21] are examples of the application of LightGBM. Based on publicly available experimental data, LightGBM is widely acknowledged as an algorithm that processes massive data quickly, learns data quickly, has high accuracy and strong model precision, and consumes little data memory. This advantage increases its effectiveness and efficiency over other machine-learning techniques [8]. LightGBM's benefits and flexibility can potentially encourage the development of credit evaluation in P2P lending platforms.

In the past, Particle Swarm Optimization (PSO) was introduced by Kennedy and Eberhart in 1995. PSO is one of the meta-heuristic algorithms for optimizing some problems. The concept of PSO was inspired by the motion of birds' flocks that fly together toward a place. In optimization, PSO works better than genetic algorithms [22]. PSO is started by initializing a swarm of particles that formed from a group of random possible solutions. Each particle is given a starting position and speed. When a

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

particle discovers a direction to the food source, other particles will take after it [23]. PSO generated better elapsed time than Ant Colony Optimization (ACO) when doing optimization tasks [24]. PSO has a simpler concept and source implementation compared to other heuristic techniques. PSO strategies can produce high-quality arrangements inside shorter calculation time and steady meeting characteristics than other stochastic strategies [24].

Based on the development of the P2P lending platform, credit scorecards were introduced for credit risk evaluation, and the researchers used machine learning techniques as a credit risk evaluation tool. Several researchers have proven that LightGBM performance is superior to other machine learning techniques [14],[25]. However, based on its application, this method focuses more on maximizing the level of accuracy and minimizing error rate and does not consider benefits for lenders or investors for the credit risk evaluation model. Meanwhile, maximizing profits is the investors' goal. Therefore, to achieve this objective, a profit score is introduced and used for the credit risk evaluation model [26]. According to [27] it is necessary to provide an overview of the difference in profit score and the real profit. The returns and losses from the remaining cases are not taken into account by the actual profit, which only counts profit and losses from cases classified as non-defaulters. If a non-defaulter becomes a defaulter, lenders or investors may lose their money. Profit scores can assist investors to avoid losses by identifying non-defaulters and defaulters. Thus, the profit score can be used to evaluate real profit and loss using potential profit and loss. Thus, the profit scores better measure and analyze credit risk evaluation than the actual profit.

Therefore, the LightGBM algorithm was used to improve the predictions of defaults of P2P lending platforms. PSO was used to improve decision tree comparison in a parameter-optimized LightGBM. The parameter-optimized LightGBM was used to minimize credit risk and maximize the profit score in credit risk evaluation in P2P lending platforms.

This research was structured as follows. After this introduction, section 2 will provide a review of relevant work. The review consists of credit risk evaluation in P2P lending platforms, enhancement of LightGBM, and PSO implementation. In section 3, the method used in this research is described comprehensively. All the data processed in this research is presented in section 4. The implementation of PSO as an optimizer for LightGBM algorithm to generate expectable results is also presented in section 4 . Finally, the last section summarizes the findings and provides the conclusion.

2. Materials and Methods

2.1. The Evaluation of Credit Risk in Peer-to-Peer Lending Platforms

P2P lending platforms employ credit risk evaluation in two primary ways. First, credit risk evaluations are binary-classified. Second, credit scores are used to estimate loan credit risk. A conventional credit evaluation approach is by using a credit scorecard. For example, LendingClub and Fair Isaac Corporation (FICO) self-launch a scorecard for the company [28]. This card delivers a credit score to each borrower quickly and effortlessly. However, according to [29], credit scorecards cannot differentiate between defaulters and non-defaulters.

Many researchers deploy machine learning techniques to forecast whether a P2P lending platform loan can be repaid or when repayment is due. The machine learning technique is applied to increase the accuracy of spotting defaults, such as Logistic Regression [30]–[32], Decision Tree [33], Neural Networks [34]–[36], Support Vector Machine [37], Random Forest [38]–[40], Gradient Boosting Decision Trees [14], [25], [41], and Convolutional Neural Networks [42],[43]. Other algorithms, such as naïve bayes [44][45], C4.5 [46],[47], and K-Nearest Neighbor [48],[49], can also be used as they have the potential to do classification tasks.

Some researchers have also conducted a comparative study of different machine-learning techniques to evaluate loans in P2P lending platforms. Support Vector Machine has better performance than neural networks [50]. Random Forest outperforms the other machine learning techniques (SVM, K-Nearest Neighbor, and LR) with actual P2P lending platform data at LendingClub in [29]. Using 60,000 records of data in 2015 from LendingClub, [51] indicated that Random Forest has better prediction accuracy than Decision Tree and Bagging. [52] presented Binary PSO and SVM with two different classification algorithms: Extremely Randomized Tree and Random Forest or BPSOSVM-ERT and BPSOSVM-RF. They showed that BPSOSVM-ERT is superior to BPSOSVM-RF, SVM, and RF. The

http://doi.org/10.26594/register.v9i1.3060

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

combination of Transductive Support Vector Machine and label propagation with the Dempster Shafer theory outperforms the existing machine learning techniques such as Decision Tree, Co-training, Simple Averaging, TSVM, and Label Propagation in [42]. A study by [14] found that the three decision tree-based single models (GBDT, XGBoost and LightGBM) performed more superior prediction than the other six benchmark single models (NN, LR, RF, SVM, AdaBoost, and KNN).

The main objective of evaluating loans in P2P lending platforms is to help lenders or investors to get big profits. According to [53], traditional credit risk assessment methods are more concerned with cutting or minimizing default rates. Therefore, P2P lending platforms must develop new credit risk rating criteria to raise or maximize revenues rather than reduce or minimize default rates [54]. For example, a study by [55] proposed innovative credit and profit scoring systems. In addition, some researchers argue that classification accuracy is not always precise or profitable. According to [56], they proposed a method for estimating prospective candidates' profitability that is more accurate than existing methods for predicting default. For the credit risk evaluation model, [57] presented a new profit concept based on the classification of performance criteria. In this concept, all performance is measured using the Expected Maximum Profit (EMP) metric, which is particularly effective in selecting a high-profit credit risk evaluation method. Furthermore, [58] integrated credit scoring into a profit-based return annualized rate. Meanwhile, [51] used cost-sensitive learning and Extreme Gradient Boosted (XGBoost) to look at loans using a cost-sensitive boosted tree. This procedure makes it easier to distinguish between defaulters and non-defaulters.

Based on the literature review described above, it can be concluded that decision algorithms, such as LightGBM and random forest, work better and have a strong potential to be improved than other machine learning techniques. As a result, profit score can perform better in choosing a credit risk evaluation model that brings great profit potential than conventional models of criteria measurement such as AUC and accuracy. Based on the review of the previous work above, the researchers would like to address two problems in this present study. First, the researchers put more focus on the actual returns and actual losses categorized as non-defaulters and did not consider the impact of the potential returns and losses of such cases. Second, the researchers also focused on comparative studies but paid less attention to the possible significant impact of improved credit risk evaluation algorithms in P2P lending platforms.

2.2. The Improvement of Light Gradient Boosting Machine

Several researchers had proven that LightGBM was an algorithm that can achieve good classification accuracy compared to other machine learning techniques [14][19][21][25]. LightGBM parameters also have the potential to be optimized to obtain a greater level of classification accuracy while maintaining a reasonable level of operational efficiency. In loan evaluation in P2P lending, [29] examined the impact of maximum tree depth, split features number, and forest scale on Random Forest works. In computer science, [59] investigated the implementation of RF on several datasets to investigate the impact of parameter selection on classification tasks and found that parameters are significantly related to their level of accuracy and how parameter optimization also affects the performance of RF. Although the performance of the RF can be improved by optimizing the parameters, this increase does not have a significant effect because Random Forest is less sensitive to parameter selection [60].

Based on the findings above, [41] optimized the parameters of XGBoost with a PSO, and it was proven to improve the performance of XGBoost. Such a step can also be applied to the LightGBM algorithm because, based on our observations, no one has optimized the parameter of LightGBM for credit risk evaluation in the P2P lending platforms.

Based on the literature review above, we proposed LightGBM-PSO for credit risk evaluation on P2P lending platforms. However, in several studies with the same objective, high accuracy is used as a parameter to see that the optimization of the algorithm is successful. For the evaluation of loans on the P2P lending platform, the parameters cannot represent a high potential profit or a small risk of loss. It is necessary to define a new objective function based on the evaluation characteristics of P2P lending platform to optimize LightGBM through PSO.

2.3. Method

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

The P2P lending prediction analysis procedure started with collecting the dataset. The dataset used in this study was obtained from LendingClub. Then, the next step was data pre-processing. The next step was classifying the data using LightGBM and LightGBM optimized by PSO. Then, the final step was the measurement evaluation to evaluate the model performance. Figure 2 illustrates the research framework in general.

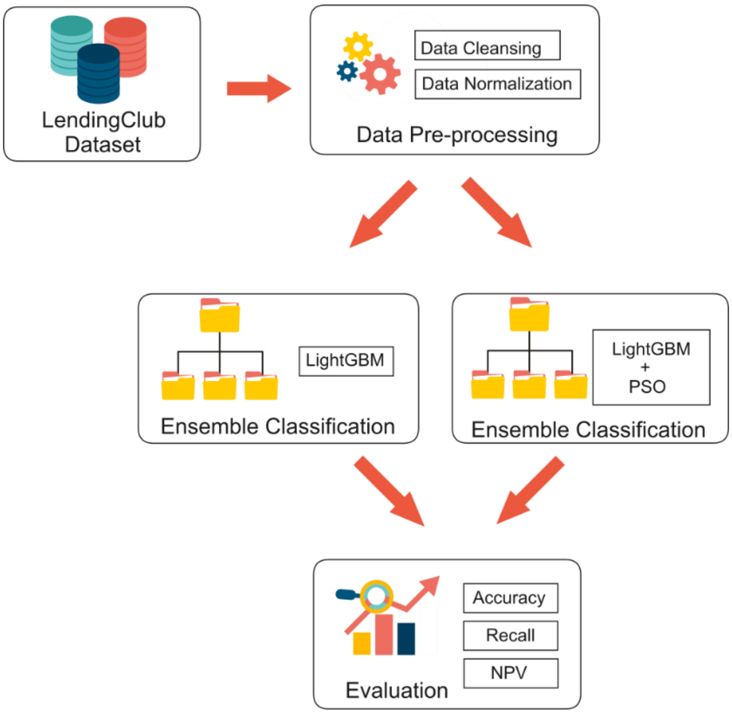

Fig. 2. Framework of research

2.3.1 Data Description

This research used the Lending Club dataset during the 2019 quarter that was accessed from Kaggle.com. Initially, the dataset consisted of 20,875,146 customer loans, containing eighteen attributes. In addition, after the data were pre-processed, the multiple or non-effect attributes were eliminated, leaving six attributes, as shown in Table 1 below.

Table 1. Attributes Selection and Pre-processing

Attributes Name	Description and Preprocessing	Type
amount_borrowed	the loan's principal amount on which interest will accrue	Numeric
borrower_rate	the amount of money that can be borrowed at a given interest rate	Numeric
Installment	The amount of the borrower's monthly bill if the loan is approved	Numeric
principal_paid	a payment made toward the original loan balance	Numeric
interest_paid	a mortgage or credit interest payment	Numeric
Grade	The lending club issued a loan rating	Numeric
Term	From binary number to discretization, 36 months' loan payback.	Numeric
loan_status	the origin of our response to the core issue of whether or not people pay back the loans they take out	Numeric

2.3.2 Data Pre-processing

The data pre-processing was done to obtain better model performance. Several invalid data, such as empty, incomplete, or null data, were cleaned. Noise and inconsistent data were also cleaned. There were several methods of doing data cleaning, for example: removing tuples, filling the value manually using global constants, mean or median, and the closest value were alternatives to handling missing value. Accordingly, in this study removing tuples was chosen as a method to deal with missing values.

2.3.3 Light Gradient Boosting Machine

http://doi.org/10.26594/register.v9i1.3060

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

In 2016, Microsoft MSRA created LightGBM as a fast and powerful Gradient Boosted Decision Tree (GBDT) algorithm for open-source promotion [61]. This parallel training method works for regression, classification, sorting, and other machine-learning tasks. Unlike XGBoost, LightGBM uses a histogram to expedite training, reduce memory, and apply a realistic growth plan with depth constraints. LightGBM discretizes floating-point eigenvalues into k bins, creates a histogram with a width of k and stores 8-bit integers, and pre-sorted results in 1/8 the memory. The accuracy of LightGBM is unaffected by this rough partition.

Because the decision tree algorithm still has some weaknesses, the segmentation points do not have to be accurate. Coarser segmentation also affects regularization, which can effectively reduce overfitting. The concept of level-wise as a growth strategy for a decision tree algorithm is deemed inefficient since it treats leaves of the same layer, resulting in excessive memory waste. In contrast, the leaf-wise notion of the decision tree algorithm is regarded as a more effective technique, as it can locate the leaves with the maximum branching yield at any time among all leaves and traverse the branching cycle. As a result, the blade can reduce mistakes and attain better precision in the same length of segmentation time when compared to the horizontal direction. The weakness of leaf orientation is that it can produce a deeper decision tree, resulting in overfitting. LightGBM imposes a maximum depth limit to the top of the leaf to minimize overfitting while maintaining high efficiency. The schematic diagram of decision tree growth strategies of leaf-wise and level-wise can be seen in Figure 3 below.

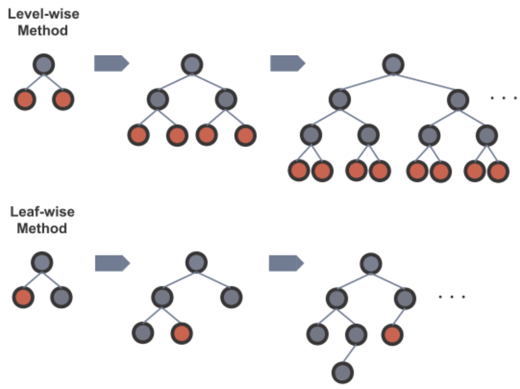

Fig. 3. The decision tree growth strategies of level wise and leaf-wise [17]

2.3.4 Particle Swarm Optimization

The particle swarm optimization algorithm is inspired by bird flock movement and is usually used to optimize solving a problem [23]. Those particle swarms encompass volume-less particles with random velocities, every of which represents a possible solution. The best solution is decided by shifting the particles in the solution area. PSO begins by initializing the particle population using random values. Particles move in the seek space to find the optimal solution by changing the position of each particle, depending on its experience and that of its neighbors. During the process, xi = (xi1, xi2, …, xiD), where D is the number of dimensions of the search space. The particle velocity i is represented using vi = (vi1, vi2, …, viD). The particle velocity is limited by vmax (maximum velocity) and vtid e (-vmax, vmax). PSO has pbest (Personal Best) and gbest (Global Best). Pbest represents the optimal position of the personal particle, while Gbest represents the optimal position of all particles. Particle position and velocity were updated by PSO to obtain the best solution. The process of updating is governed by the following equations:

$$v_{id}(t+1) = w * v_{id}(t) + c_1 * r_1 * \left(p_{id} - x_{id}(t)\right) + c_2 * r_2 * \left(p_{gd} - x_{id}(t)\right) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

http://doi.org/10.26594/register.v9i1.3060

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

where t is the t-th iteration in the algorithm process, d ∈ D represents the d-th dimension of the search space, and w denotes inertia weight. Acceleration constants are represented by c1 and c2. Meanwhile, r1i and r2i has random values distributed in [0, 1] uniformly. pid and pgd are used to represent the elements of Pbest and Gbest in the d-th dimension. The overall process of PSO is described in Figure 4 below.

Algorithm 1 Optimizing LGBM using PSO

```
Begin
Initialize particles using random values for LGBM attributes
while maximum iteration is not met do
    Calculate fitness (trained LGBM function)
    for i = 1 → number of particles do
        Update pbest of particle i
        Update gbest of particle i
        for d = 1 → Dimensionality do
            Update the velocity of particles
            Update the position of particles
        end for
    end for
end while
Return the position of gbest
Evaluate using testing data
End
```

Fig. 4. Pseudocode of Optimizing LightGBM using PSO

2.3.5 Measurement Evaluation

The last procedure of this research method was measurement evaluation. The confusion matrix was used to evaluate the performance of a classification algorithm. The confusion matrix is composed of four categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The confusion matrix is shown in Table 2.

Table 2. The Confusion Matrix

	Actual "Fully-Paid"	Actual "Charged-Off"
Predicted "Fully-Paid"	True Positive (TP)	False Positive (FP)
Predicted "Charged-Off"	False Negative (FN)	True Negative (TN)

2.3.5.1 Accuracy

Accuracy is the proportion of data points successfully predicting out of the total number of data points.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

2.3.5.2 Recall

The recall rate indicates the proportion of true positive cases that are accurately identified.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

2.3.5.3 Negative Predictive Value (NPV)

The negative predictive value (NPV) quantifies the proportion of correctly detected expected negative cases.

$$\boldsymbol{NPV = \frac{TN}{TN+FN}} \quad (7)$$

3. Results and Discussion

3.1. Experimental Results

Several parameters used by LightGBM were tuned using PSO to generate optimal results. Some of them were learning rate, num_leaves, min_child_samples, and max_depth. PSO tuned all of them to an optimal value. The optimal value in each parameter produced optimal results of accuracy. Accuracy, Recall, and Negative Predictive Value (NPV) were also used as the fitness value in optimization. PSO parameter was set on w 0.9, C1 0.5. C2 0.5, and 20 iterations. The results of parameter tuning experiment using PSO are presented in Table 3 as follows.

http://doi.org/10.26594/register.v9i1.3060

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

Table 3. The Experiment Result of Parameter Tuning

	Accuracy	Recall	NPV
LightGBM+PSO	98,094%	90,514%	97,754%
LightGBM	97,986%	89,935%	97,620%

Figure 5 below shows the comparison between LightGBM with normal parameters and LightGBM optimized using PSO. It can be seen that the results are different in each iteration. LightGBM-PSO always produced better accuracy than LightGBM in each iteration. The accuracy produced by LightGBM is 97.986%. Meanwhile, in LightGBM+PSO, the best accuracy is 98.094%. These results are proof that LightGBM combined with PSO is better than without PSO.

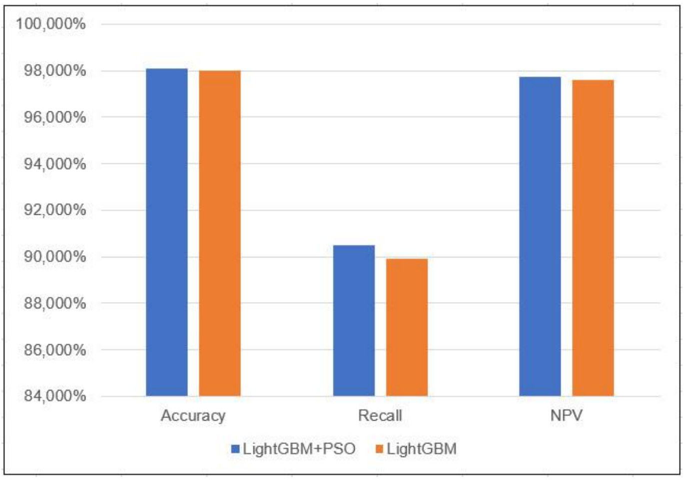

Fig. 5. The comparison of accuracy between LightGBM+PSO and LightGBM

The optimization was done by PSO to yield the best result for the LightGBM classification task. The value of LightGBM parameters generated by PSO reaches 98.094% of accuracy with 0.46 learning rate, 47 num_leaves, 14 max_depth, and 33 num_child_samples. The combination value generated by PSO for several parameters used in LightGBM could improve the result of the classification tasks. These results provide evidence that PSO can significantly help improve LightGBM algorithms.

3.2. Discussion

Even though the higher the data, the higher the accuracy generated rule applies, it can not be concluded that the accuracy will be perfect if more data is added. It needs a special study in LightGBM classification data to support the conclusion. The iteration is too. It can not be deduced that the higher the iteration, the higher the accuracy reached.

From the experiment results, the accuracy produced by LightGBM never exceeds that produced by LightGBM-PSO. It was one of the discoveries in this approach. This finding strengthens that LightGBM still needs more improvement to reach better results. Compared to previous studies, the findings in this study provide more superior accuracy than those obtained by [52][62]. This study also found that the combination of LightGBM and PSO works better to solve the problem.

4. Conclusions

A combination between LightGBM and PSO improved the model's accuracy in predicting the credit risk evaluation. LightGBM+PSO managed to outperform the normal LightGBM. LightGBM, when coupled with PSO, also generated satisfactory results. The highest accuracy also presented satisfactory results with 98.094% accuracy, 90.514% Recall, and 97.754% NPV, respectively. This finding can be used as a solid stepping stone for the following research. However, several things still need to taken into consideration. The time of experiments needs to be analyzed specifically. In addition, the number of data can be more varied because of imbalanced data. Imbalanced data can be another major problem in credit risk evaluation prediction. Finally, other optimization algorithms can be tried as an alternative approach.

Acknowledgment

This research was supported by University Tun Hussein Onn Malaysia through TIER1, H777.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

References

[1] Practices and Explorations in Digital Financial Inclusion, “China Digital Financial Inclusion Development Report,” 2017.

[2] Z. Fang, J. Zhang, and F. Zhiyuan, “Study on P2P E-finance platform system: A case in China,” *Proc. - 11th IEEE Int. Conf. E-bus. Eng. ICEBE 2014 - Incl. 10th Work. Serv. Appl. Integr. Collab. SOAIC 2014 1st Work. E-Commerce Eng. ECE 2014*, pp. 331–337, 2014.

[3] P. Renton, “Peer To Peer Lending Crosses $1 Billion In Loans Issued,” 2012. [Online]. Available: https://techcrunch.com/2012/05/29/peer-to-peer-lending-crosses-1-billion-in-loans-issued/.

[4] S. Lee, “Evaluation of mobile application in user’s perspective: Case of P2P lending apps in FinTech industry,” *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 2, pp. 1105–1115, 2017.

[5] C. Stern, “Fintechs and their emergence in banking services in CESEE,” *Focus Eur. Econ. Integr.*, no. Q3/17, pp. 42–58, 2017.

[6] E. M. Gerber, J. S. Hui, and P.-Y. Kuo, “Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms,” *Proc. Int. Work. …*, no. April 2014, p. 10, 2012.

[7] Z. Wei and M. Lin, “Market mechanisms in online peer-to-peer lending,” *Manage. Sci.*, vol. 63, no. 12, pp. 4236–4257, 2017.

[8] L. Ma, X. Zhao, Z. Zhou, and Y. Liu, “A new aspect on P2P online lending default prediction using meta-level phone usage data in China,” *Decis. Support Syst.*, vol. 111, pp. 60–71, 2018.

[9] H. Zhang, H. Zhao, Q. Liu, T. Xu, E. Chen, and X. Huang, “Finding potential lenders in P2P lending: A Hybrid Random Walk Approach,” *Inf. Sci. (Ny).*, vol. 432, pp. 376–391, 2018.

[10] J. Hegde and B. Rokseth, “Applications of machine learning methods for engineering risk assessment – A review,” *Saf. Sci.*, vol. 122, no. September 2019, p. 104492, 2020.

[11] M. A. Muslim *et al.*, “New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning,” *Intell. Syst. with Appl.*, vol. 18, p. 200204, May 2023.

[12] S. C. Hsueh and C. H. Kuo, “Effective matching for P2P lending by mining strong association rules,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F1309, pp. 30–33, 2017.

[13] D. Wang, Y. Zhang, and Y. Zhao, “LightGBM: An effective miRNA classification method in breast cancer patients,” *ACM Int. Conf. Proceeding Ser.*, pp. 7–11, 2017.

[14] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, “Default prediction in P2P lending from high-dimensional data based on machine learning,” *Phys. A Stat. Mech. its Appl.*, vol. 534, p. 122370, 2019.

[15] Z. Liu, W. Xu, W. Zhang, and Q. Jiang, “An emotion-based personalized music recommendation framework for emotion improvement,” *Inf. Process. Manag.*, vol. 60, no. 3, p. 103256, May 2023.

[16] 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,” *Detect. Classif. Acoust. Scenes Events*, no. November, pp. 1–5, 2017.

[17] N. Jha, D. Prashar, M. Rashid, S. K. Gupta, and R. K. Saket, “Electricity load forecasting and feature extraction in smart grid using neural networks,” *Comput. Electr. Eng.*, vol. 96, p. 107479, Dec. 2021.

[18] 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,” *Agric. Water Manag.*, vol. 225, no. August, p. 105758, 2019.

[19] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, “A novel construction cost prediction model using hybrid natural and light gradient boosting,” *Adv. Eng. Informatics*, vol. 46, no. September, p. 101201, 2020.

[20] 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,” *14th Int. Conf. Comput. Sci. Educ. ICCSE 2019*, no. Iccse, pp. 1111–1116, 2019.

[21] 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,” *Inf. Sci. (Ny).*, vol. 535, pp. 107–129, 2020.

[22] T. Sousa, A. Silva, and A. Neves, “Particle Swarm based Data Mining Algorithms for classification tasks,” *Parallel Comput.*, vol. 30, no. 5–6, pp. 767–783, 2004.

 http://doi.org/10.26594/register.v9i1.3060

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

[23] J. Kennedy, "Particle swarm: Social adaptation of knowledge," *Proc. IEEE Conf. Evol. Comput. ICEC*, pp. 303–308, 1997.

[24] T. Herlambang, D. Rahmalia, and T. Yulianto, "Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) for optimizing PID parameters on Autonomous Underwater Vehicle (AUV) control system," *J. Phys. Conf. Ser.*, vol. 1211, no. 1, 2019.

[25] 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," *Electron. Commer. Res. Appl.*, vol. 31, pp. 24–39, 2018.

[26] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decis. Support Syst.*, vol. 89, no. June, pp. 113–122, 2016.

[27] X. Ye, L. an Dong, and D. Ma, "Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score," *Electron. Commer. Res. Appl.*, vol. 32, no. July, pp. 23–36, 2018.

[28] A. C. B. Ortega and F. Bell, "Online social lending: Borrower-generated content," *14th Am. Conf. Inf. Syst. AMCIS 2008*, vol. 3, no. January 2008, pp. 1957–1964, 2008.

[29] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, 2015.

[30] M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M. J. Segovia-Vargas, "Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending," *IEEE Access*, vol. 8, pp. 64873–64890, 2020.

[31] Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong, "Instance-based credit risk assessment for investment decisions in P2P lending," *Eur. J. Oper. Res.*, vol. 249, no. 2, pp. 417–426, 2016.

[32] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of default in P2P lending," *PLoS One*, vol. 10, no. 10, pp. 1–22, 2015.

[33] Y. Zhang, H. Jia, Y. Diao, M. Hai, and H. Li, "Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending," *Procedia Comput. Sci.*, vol. 91, no. Itqm, pp. 168–174, 2016.

[34] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach," *Rev. Dev. Financ.*, vol. 4, no. 1, pp. 20–28, 2014.

[35] P. Giudici, B. Hadji-Misheva, and A. Spelta, "Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms," *Front. Artif. Intell.*, vol. 2, no. May, pp. 1–8, 2019.

[36] L. Liang and X. Cai, "Forecasting peer-to-peer platform default rate with LSTM neural network," *Electron. Commer. Res. Appl.*, vol. 43, p. 100997, Sep. 2020.

[37] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014.

[38] Y. Jin and Y. Zhu, "A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending," *Proc. - 2015 5th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2015*, pp. 609–613, 2015.

[39] C. Rao, M. Liu, M. Goh, and J. Wen, "2-stage modified random forest model for credit risk assessment of P2P network lending to 'Three Rurals' borrowers," *Appl. Soft Comput. J.*, vol. 95, p. 106570, 2020.

[40] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput. Sci.*, vol. 162, no. Itqm 2019, pp. 503–513, 2019.

[41] 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," *Inf. Sci. (Ny).*, vol. 525, pp. 182–204, 2020.

[42] J. Y. Kim and S. B. Cho, *Deep Dense Convolutional Networks for Repayment Prediction in Peer-to-Peer Lending*, vol. 771. Springer International Publishing, 2019.

[43] C. Wang, Y. Zhang, W. Zhang, and X. Gong, "Textual sentiment of comments and collapse of P2P platforms: Evidence from China's P2P market," *Res. Int. Bus. Financ.*, vol. 58, p. 101448, Dec. 2021.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 18-28

[44] R. Hamonangan, M. B. Saputro, C. Bagus, S. Dinata, and K. Atmaja, "Accuracy of classification poisonous or edible of mushroom using naïve bayes and k-nearest neighbors," *J. Soft Comput. Explor.*, vol. 2, no. 1, 2021.

[45] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, May 2018.

[46] M. A. Muslim, A. J. Herowati, E. Sugiharti, and B. Prasetiyo, "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease," *J. Phys. Conf. Ser.*, vol. 983, no. 1, 2018.

[47] I. G. A. Suciningsih, M. A. Hidayat, and R. A. Hapsari, "Comparation analysis of naïve bayes and decision tree C4.5 for caesarean section prediction," *J. Soft Comput. Explor.*, vol. 2, no. 1, pp. 46–52, 2021.

[48] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, p. 032024, Jun. 2020.

[49] N. Hidayat, M. F. Al Hakim, and ..., "Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning," *Sci. J. …*, vol. 8, no. 2, pp. 314–319, 2021.

[50] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study," *Decis. Support Syst.*, vol. 37, no. 4, pp. 543–558, 2004.

[51] Y. Xia, C. Liu, and N. Liu, "Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending," *Electron. Commer. Res. Appl.*, vol. 24, pp. 30–49, 2017.

[52] N. Setiawan, Suharjito, and Diana, "A comparison of prediction methods for credit default on peer to peer lending using machine learning," *Procedia Comput. Sci.*, vol. 157, pp. 38–45, 2019.

[53] R. A. Eisenbeis, "Pitfalls in the application of discriminant analysis in business, finance, and economics," *J. Finance*, vol. 32, no. 3, 1977.

[54] T. LC, "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers," *Int. J. Forecast.*, vol. 16, p. 149, 2000.

[55] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015.

[56] M. C. So, L. C. Thomas, H. V. Seow, and C. Mues, "Using a transactor/revolver scorecard to make credit and pricing decisions," *Decis. Support Syst.*, vol. 59, no. 1, pp. 143–151, 2014.

[57] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," *Eur. J. Oper. Res.*, vol. 238, no. 2, pp. 505–513, 2014.

[58] Y. Wang and X. S. Ni, "Improving investment suggestions for peer-to-peer lending via integrating credit scoring into profit scoring," *ACMSE 2020 - Proc. 2020 ACM Southeast Conf.*, pp. 141–148, 2020.

[59] B. F. F. Huang and P. C. Boutros, "The parameter sensitivity of random forests," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–13, 2016.

[60] W. Yuyan, D. Wang, Y Wang, and Y. Jin, "Predicting Survivability of Colorectal Cancer by an Ensemble Classification Method Improved on Random Forest," *J. Manag. Sci.*, vol. 30, no. 1, pp. 95–106, 2017.

[61] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[62] V. K. L, S. Natarajan, S. Keerthana, K. M. Chinmayi, and N. Lakshmi, "Credit Risk Analysis in Peer-to-Peer Lending System Vinod Kumar L Keerthana S , Chinmayi K M , Lakshmi," *2016 IEEE Int. Conf. Knowl. Eng. Appl.*, pp. 193–196, 2016.

http://doi.org/10.26594/register.v9i1.3060