International Journal of Computers and Applications

ISSN: 1206-212X (Print) 1925-7074 (Online) Journal homepage: www.tandfonline.com/journals/tjca20

Feature selection in P2P lending based on hybrid genetic algorithm with machine learning

Muhammad Sam'an, Muhammad Munsarif, Safuan & Yahya Nur Ifriza

To cite this article: Muhammad Sam'an, Muhammad Munsarif, Safuan & Yahya Nur Ifriza (2023) Feature selection in P2P lending based on hybrid genetic algorithm with machine learning, International Journal of Computers and Applications, 45:12, 764-775, DOI: 10.1080/1206212X.2023.2276553

To link to this article: https://doi.org/10.1080/1206212X.2023.2276553

INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS
2023, VOL. 45, NO. 12, 764–775
https://doi.org/10.1080/1206212X.2023.2276553

Feature selection in P2P lending based on hybrid genetic algorithm with machine learning

Muhammad Sam'an[a], Muhammad Munsarif[a], Safuan [a] and Yahya Nur Ifriza[b]

[a]Departement of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia; [b]Departement of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia

ABSTRACT

The emergence of online lending services such as peer-to-peer (P2P) lending has simplified transactions for lenders, eliminating the need for traditional bank intermediaries. However, accurately predicting potential lender defaults is a critical challenge in preventing financial distress, as lenders bear the burden of default risk. This challenge becomes even more complicate with P2P lending datasets that feature a multitude of complex attributes. To improve predictive accuracy, the study focuses on optimizing feature selection in this data-rich environment. Its main objective is to identify the most influential features for predicting default risk among P2P lenders through feature optimization techniques, leveraging the Genetic Algorithm (GA) in conjunction with ten different Machine Learning models. The study employed a hybrid approach of GA with three mutation rate levels, basic (MR = 0), moderate (MR = 0.5), and extreme (MR = 1), to provide insights into model responsiveness and performance under various mutation scenarios. Notably, the results highlight GA+XGBoost as the best-performing model with a stable fitness score of 86.132% compared to others. This research holds significant potential for improving lender risk management in the P2P lending landscape by effectively identifying higher risk lenders. Ultimately, the findings contribute positively to the mitigation of financial risks for lenders within the P2P lending ecosystem.

ARTICLE HISTORY
Received 21 May 2023
Accepted 24 October 2023

KEYWORDS
P2P lending; feature selection; genetic algorithms; machine learning

1. Introduction

Peer-to-Peer (P2P) lending has significant potential to reshape the future of conventional banking. As the world's largest online lending marketplace, P2P lending offers a wide range of lending services, including personal loans, business loans, medical loans, and more. Initially, the primary goal of P2P lending in 2005 was to democratize more efficient consumer financial services. In this model, individuals can make loan offers, and investors or lenders can approve loans without involving formal financial institutions. In 2005, P2P lending was first launched in the UK by Zopa. A year later, platforms such as LendingClub and Prosper were established in the U.S., while China also developed similar loan facilities.

Many researchers have studied P2P lending data analysis by implementing and proposing many models. In Korea, the herding behavior and the lack of marginal effects as bid advances in P2P lending have been studied [1]. Jin and Zhu [2] compared five machine learning (ML) models: two decision trees, two neural networks and one support vector machine for predicting credit risk in P2P lending. Byanjankar et al. [3] conducted a study on ANN-based model credit score for default and non-default classification in P2P lending. Malekipirbazari and Aksakalli [4] applied Random Forest (RF) to identify the best borrowers and compare them with FICO credit scores based on LC scores. Outlier detection based on bad credit scores was utilized to find abnormal investors and predict potential investors [5]. Internal rate of return for predicting profits expected by investors [6]. An imbalance between investors and borrowers resulted from unstructured data mining related to text data on the motivation of investors and borrowers [7]. Hybrid pessimistic contrastive probability-based rejection inference model with advanced LightGBM as a classifier (CPLE-LightGBM) in semi-unsupervised P2P lending related to inference rejection problems [8]. Default prediction in P2P lending using machine learning [9–13]. Embedded technique and stacking ensemble learning for predicting credit risk of P2P lending [14].

The data set obtained from P2P lending platforms often contains irrelevant and excessive features. As a result, the accuracy of the model in the data classification process is diturbed [15]. The problem becomes more complex in the case of data with very large sizes and dimensions, as this has an impact on the efficiency and effectiveness of the model's performance [16]. For example, the processing time becomes longer because of the need to handle numerous features. A solution to this problem is the technique of feature selection, which is a part of feature engineering in the preclassification phase [17–19]. This technique focuses on selecting features that are relevant to the data and have an impact on the classification results [20–22]. Feature selection also plays a role in reducing data dimensions and improving the accuracy performance of classification models by eliminating irrelevant features [23–25].

Feature selection for predicting loan defaults in P2P lending, which is a crucial step in optimizing the accuracy of predictive model, has been conducted. Xu et al. [26] utilized RF and Support Vector Machine (SVM), which proved to be excellent in detecting potential fraud in loan applications. The results of this research revealed that the use of RF and SVM could overcome the limitations of basic features, thereby improving the accuracy of default prediction and increasing the reliability of P2P lending systems in identifying risks. On the other hand, the Restricted Boltzmann Machine has been proven to be effective in eliminating irrelevant features and reducing

CONTACT Muhammad Sam'an muhammad92sam@unimus.ac.id

errors in credit score prediction based on deep learning models [27]. This method aids in identifying features that significantly influence credit prediction outcomes, allowing for a reduction in model complexity and ultimately increasing prediction efficiency and accuracy. In other words, feature selection not only assists in removing irrelevant features, but also allows models to focus on the most relevant information. The feature selection process in this research involved reducing the dimensionality of the filters from 30 features to 19 features using the recursive feature elimination method, a technique known to effectively reduce data complexity without sacrificing accuracy. After feature selection, the classification process was performed using RF, which is known as one of the robust algorithms for classification tasks. The final results of this study achieved significantly higher accuracy compared to cases without feature selection [28].

Yang et al. [29] developed the SSA-CatBoost model, which combines the Sparrow Search Algorithm (SSA) and CatBoost to improve the accuracy of credit scoring. This approach also leverages the Recursive Feature Elimination technique as a feature selection method to mitigate the impact of risky data and improve computational efficiency. As a result, this research significantly improved the accuracy of credit prediction while maintaining data processing efficiency. Yin et al. [30] proposed an ensemble stacking model for predicting credit default risk. In their method, they used the Max-Relevance and Min-Redundancy (MRMR) method to select relevant features and applied k-means to eliminate irrelevant features. The results of their research showed that the ensemble stacking model outperformed the approach using a single model in terms of accuracy, precision, and recall.

The research conducted by Xu et al. [26], Ha et al. [27], Li et al. [28], Yang et al. [29], and Yin et al. [30] consistently emphasized the importance of feature selection as a crucial step in improving the quality of default prediction in the context of P2P lending. The primary motivation behind this research is to improve the efficiency and accuracy of credit risk management in the increasingly vital P2P lending industry. By employing meticulous feature selection methods, this research aims to provide a more robust tool for identifying potential high-risk borrowers. As a result, it is expected that P2P lending platforms will be able to make more effective lending decisions and reduce the risk of default. Therefore, the results of this research have a significant implications for improving the quality of financial services provided by P2P lending platforms to borrowers and investors, while also helping to mitigate potential risks within the industry.

Recently, genetic algorithms (GAs), a subset of the evolutionary algorithms (EAs), have been applied to feature selection. In GAs, chromosomes are represented as features to be selected [31]. Fitness values are obtained from the model training process in machine learning. Features in the context of P2P lending, which serves as chromosomes, are processed through the search space in GAs. High-dimensional P2P lending datasets require a large search space. However, the limited search space in GAs can lead to premature or local optima, meaning that not all machine learning models hybridized with GAs can significantly improve classification accuracy. For example, [32] hybridized GAs with three machine learning models, namely SVM (GA+SVM), RF (GA+RF), and LR (GA+LR). The result showed an average increase in classification accuracy of 0.1% in compared to the baseline model. This limitation is also influenced by one of the operators used in GAs, namely mutation. This operator plays a role in generating a new generation that is more fit. The smaller the mutation rate is, the higher the chance of obtaining a fit generation becomes [33]. However, it is essential to remember that these statements only guarantee a fit generation or an optimal solution.

This research proposes feature selection in P2P lending based on hybrid genetic algorithm with machine learning. The main contributions of this research are outlined below.

(i) To achieve more accurate and diverse results, we employed a combination of genetic algorithms (GA) with ten different machine learning models: Logistic Regression (GA+LR), Gaussian Naive Bayes (GA+GNB), K-Nearest Neighbors (GA+KNN), Decision Tree (GA+DT), Gradient Boosting Decision Trees (GA+GBDT), Random Forest (GA+RF), XGBoost (GA+XGBoost), LightGBM (GA+LightGBM), Adaboost (GA+Adaboost), and CatBoost (GA+CatBoost). This combination allowed us to explore different machine learning approaches and techniques in our efforts to improve prediction quality and model accuracy in the context of P2P lending. By utilizing these diverse models, we identified the most appropriate ones for specific problems and optimize the overall prediction outcomes.

(ii) This research focused on the development of feature selection methods in the context of P2P lending, with an emphasis on determining the mutation rate. The mutation rate is a key parameter in the genetic algorithm (GA) used for feature selection. This study aims to investigate the impact of different mutation rates, both at regular and extreme levels, on the performance of feature selection in the context of P2P lending. A regular mutation rate may be more conservative, while an extreme mutation rate may introduce genetic changes more aggressively. The results of this research are expected to provide insights into the optimization of GA parameters for selecting the most relevant features for predicting loan defaults in the P2P lending industry.

This research paper is structured in the following way The proposed system is described in the 'Methods' section. Experimental results and discussion are presented in the 'Results and Discussion' section. Conclusions and future research directions are discussed in the 'Conclusions and Future Work' section.

2. Methodology

Figure 1 visualizes the workflow of the proposed model for feature selection in P2P lending default prediction.

2.1. P2P lending dataset

The P2P lending dataset was collected from the Lending Club, which is available on Kaggle.com [38]. This dataset does not include the name of the borrower. The P2P lending dataset included 42.538 observations and 111 features in 2019.

2.2. Data preparation

2.2.1. Data pre-processing

The goal of data preprocessing was data integration, data cleaning, and data reduction. The output of this step was a good quality dataset so that it can improve the performance of the prediction model accuracy. In addition, dataset analysis is also important to understand the various features relevant to the proposed model. For example, employment status is helpful for data exploration to understand the demographics of the loan base. Loan status provides information about the condition of the loan when the dataset was created. This feature helps to serve as the dependent variable.

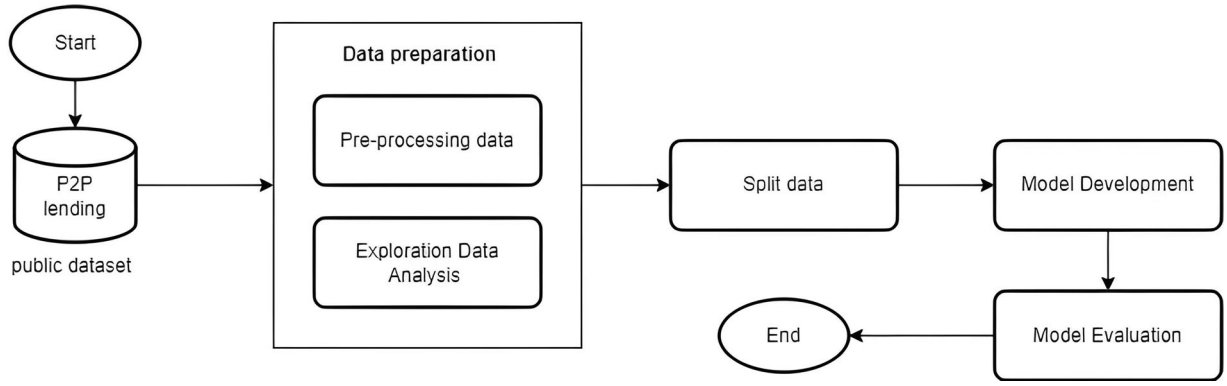

Figure 1. The workflow of the proposed model.

2.2.2. Exploratory data analysis

The EDA (Exploratory Data Analysis) identified missing values, patterns and outliers in the P2P lending datasets. The EDA aims to understand the relevant features that influence the dependent variable. EDA can also improve the predictive ability of the model. This technique process coincides with data preprocessing to ensure that data irregularities can be identified and resolved appropriately.

2.3. Split data

The comparison of the dataset ratio was 80:20, where 80% was for training data and 20% was for test data. A random status of 42 was added to the training process so that the data separation was consistent and repeatable.

2.4. Model development

The essence of this paper is the exploration of EA algorithms, especially GA, for feature selection optimization. Hybrid GA with 10 ML was proposed for feature selection in P2P lending datasets.

2.4.1. Genetic algorithm (GA)

Genetic algorithm (GA) can efficiently solve NP-hard problems by simulating biological and genetic evolution processes. Each chromosome represents an optimal solution. Chromosome evaluation uses the fitness function to determine the next parent. The two operators are crossover and mutation to produce the next new generation. A chromosome with the best fitness value is used as the optimal solution or output of the GA. Figure 2 visualizes the workflow of feature selection based GA in this paper. Chromosome coding, fitness function and genetic operations (crossover, mutation and selection) are shown in detail.

(1) Chromosome encoding
Chromosomes represent possible combinations of optimized features. The features that most influence the prediction model are denoted by (F or features). Specifically, an F-length string was used to represent the selected features, and the nth digit of the chromosome represented the nth feature in the feature selection. The value of each digit was 0 and 1. This means the corresponding feature was selected or omitted. For example, the possible combinations of features in feature selection that have n features are shown in Figure 3.

(2) Fitness function
The fitness function determines the convergence speed and the resulting optimal solution. This function is important for GA because it is directly related to the performance of the solution. This study aims to select the most influential features; therefore, the fitness function was designed as the best accuracy score of each ML model: LR, Gaussian NB, KNN, DT, RF, GBDT, XGBoost, LightGBM, AdaBoost, and CatBoost. The larger the chromosome fitness value, the higher the accuracy score of ML models suitable for feature selection in predicting payments or defaults in P2P lending datasets

(3) Crossover operation
In order to produce a new generation of offspring, GA can be done through the crossover operator. Two-point crossover was used in this paper, as shown in Figure 4. For example, from the parents and crosses 1 and 2, a pair of chromosomes Cr_1 and Cr_2 were randomly selected. The two crossing points were divided into 3 parts, where the middle gene Cr_1 was crossed with the middle gene Cr_2. After the operation, two new generations were produced. Each parental chromosome was operated using the crossover to increase the possibility of individual diversity.

(4) Mutation operation
Mutation operations can avoid early or local optimality in GA by increasing the randomness of the solution. This operator determines the probability of selecting an individual. In this paper, the mutation rate (MR) was configured in 3 levels: basic (MR = 0), medium (MR = 0.5), and extreme (MR = 1). Figure 5 visualizes the multipoint used as the mutation operator. In the mutation process, each digit of the chromosome was changed with a pre-configured probability, which means that several features were dropped or selected. This operator introduces slight randomness into the search process and maintains the convergence of the model.

(5) Selection operation
The selection operation is used to select individuals with the highest fitness value or accuracy score from their offspring and parents to produce the next generation. In this paper, the tournament selection technique was adopted for individual selection. Two individuals were randomly selected to be compared, and the best fitness value was selected to become the next generation. This technique allows individuals with the best fitness value to have a more remarkable survival.

2.5. Model evaluation

The performance evaluation of the proposed method uses the confusion matrix for a binary classification problem. The confusion matrix includes the true positive (TP), true negative (TN), false negative (FN), and false positive (FP). The confusion matrix was used to calculate accuracy, recall, precision, and F1-score, which are formulated

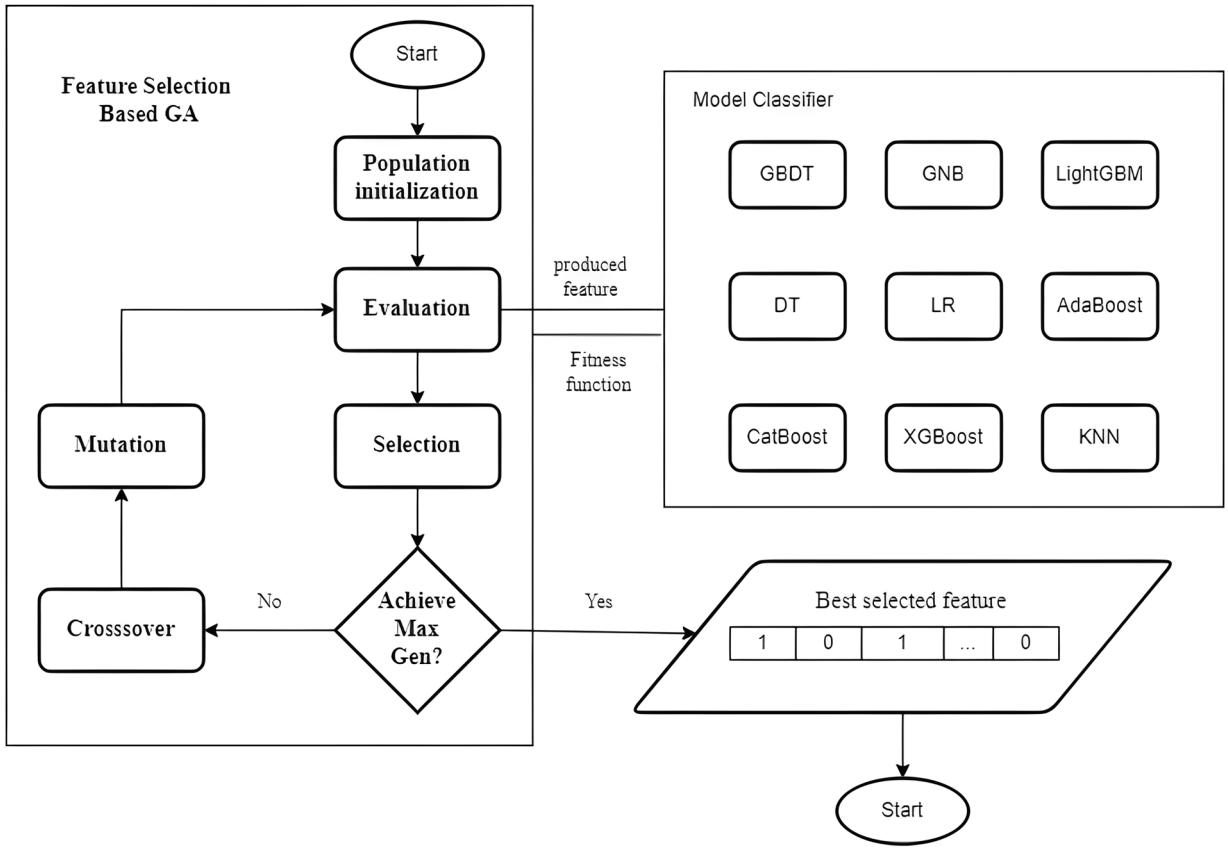

Figure 2. The workflow of feature selection based GA.

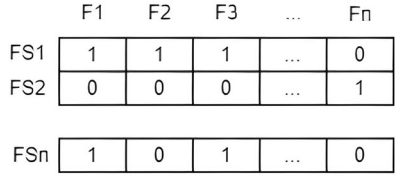

Figure 3. Chromosome encoding.

as follows:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{TN + FP + TP + FN} \\
Recall &= \frac{TP}{TP + FN} \\
Precison &= \frac{TP}{TP + FP} \\
F1 - score &= 2 \times \frac{Precison \times Recall}{Precison + Recall}
\end{aligned}
\tag{1}
$$

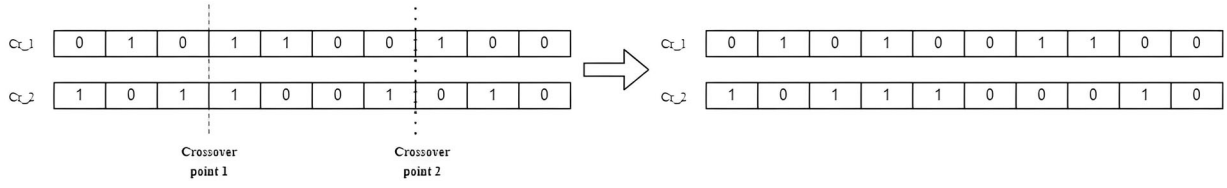

Figure 4. Crossover operation.

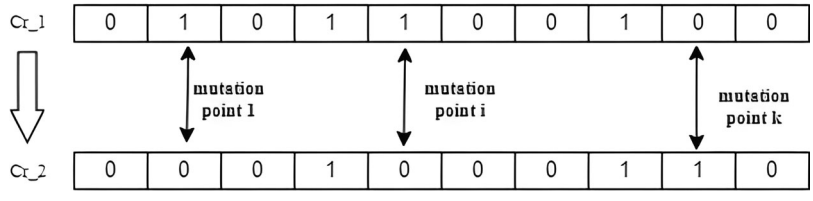

Figure 5. Mutation operation.

3. Data analysis

3.1. Data management and description

Based on the P2P lending dataset in Section 3.1, the exploration eliminated features with more than 68.51% missing value. This process reduced 111 features to 37 features. With so many features, the performance of the prediction model was not optimal because the features were too complex. Based on previous research, the solution is to further restrict the feature space as suggested [4,34,35]. The results of the features are presented in Table 1.

The tenure of the feature was converted to a numeric variable with ' < 1-year' tenure set to 0 and ' > 10 years' tenure set to 10. The homeownership feature containing the 'ANY' level was not found in the Loan Club data dictionary. The 'purpose' feature contained 14 levels, including 'Marriage' and 'Education,' which represent multiple observations. The levels were then moved to the 'Other' level. A new feature called earliest credit from issue date was created by subtracting the date from the earliest credit limit and the issue date to determine how farg the credit limit is from the date of the loan. This new feature will be explained in a matter of months.

Since the focus of the study is to provide a prediction for probable defaulters, the levels 'Late (31–120 days),' 'In Grace Period,' and 'Late (16–30 days)' were removed from the 'loan status' feature. The 'Default' level was then renamed 'Charged Off' as both levels had similar definitions. For the numeric features, the debt-to-income ratio should only consist of positive values, but observations with negative values were removed from the dataset. Similarly, a right censor similar to [4] was applied to the delinquency in the last 2 years, inquiries in the last 6 months, and derogatory public records features. The rest of the numerical features contained outliers, and their distribution was significantly left-skewed. The top 1% of outliers were removed from each numeric feature, which improved the skewness and kurtosis values

Table 1. Lending Club.

Feature name	label	Description
loan_amnt	F0	The listed amount of the loan is requested by the borrower
int_rate	F1	Interest Rate on the loan
installment	F2	The monthly payment the borrower will owe if the loan originates
annual_inc	F3	Lending Club assigned loan grade
loan_status	F4	The current status of the loan
dti	F5	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
delinq_2yrs	F6	The number of delinquency 30+ days past-due in the borrower's credit file in the last 2 years
inq_last_6mths	F7	The number of inquiries in the last 6 months
open_acc	F8	The number of open credit lines in the borrower's credit file
pub_rec	F9	Number of derogatory public records
revol_bal	F10	Total revolving credit balance
revol_util	F11	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
total_acc	F12	the borrower is using relative to all available revolving credit
home_ownership	F13–F17	The total number of credit lines currently in the borrower's credit file Homeownership status, as provided by the borrower during enrollment or obtained from the credit report.. Our values are: RENT (17), OWN (16), MORTGAGE (13), NONE (14), OTHER (15).
verification_status	F18–F20	Indicating whether the income was verified (18) by LC, not verified (19), whether the income source was verified (20)
purpose	F21–F33	A category provided by the borrower for the loan request. Our values are: car (21), credit card (22), debt consolidation (23), educational (24), home improvement (25), house (26), major purchase (27), medical (28), moving (29), other (30), renewable energy (31), small business (32), vacation (33), wedding (34)
term	F35, F36	The number of payments on the loan. Values are in months and can be either 36 (35) or 60 (36).

3.2. Data pre-processing

Missing values are handled using imputation mode for categorical variables and mean imputation for numeric variables. Categorical variables are encoded using one-hot coding to get them in numeric format. We coded the loan status 'Current,' 'Paid' in full, 'Issued' as usual = 0, encoded 'Default,' 'Billed,' 'Within grace period,' 'Lately (16–30 days)' and 'Lately (31–120days)' as default = 1. Then we can visualize the loan status as shown in Figure 6. The sample with 'normal' loan status accounts for 79.92%, but the 'default' sample is only 20.08%, indicating the category imbalance in the data set. Meanwhile, we abstracted features 'emp_length' and 'grade' and other features encoded once.

Missing values were handled using imputation mode for categorical variables and mean imputation for numerical variables. Categorical variables were encoded using one-hot coding to get them into numerical format. We coded the loan status 'Current,' 'Paid' in full, 'Issued' as usual = 0, coded 'Default,' 'Billed,' 'Within grace period,' 'Recent (16–30 days)' and 'Recent (31–120days)' as default = 1. Then loan status is visualized in Figure 6. The sample with 'normal' loan status accounted for 79.92%, but the 'default' sample was only 20.08%, indicating the category imbalance in the data set. Meanwhile, we abstracted the features 'emp_length' and 'grade' and other features that were encoded once.

3.3. Block-Based GA for feature selection

Application of GA with the ML model was a fitness function to determine the best chromosome or feature set. Operator configuration on GA, such as Crossover Rate (CR), was 0.8, and Three MR levels were 0.1, 0.5, and 1.0. These two operators were used to develop a series of

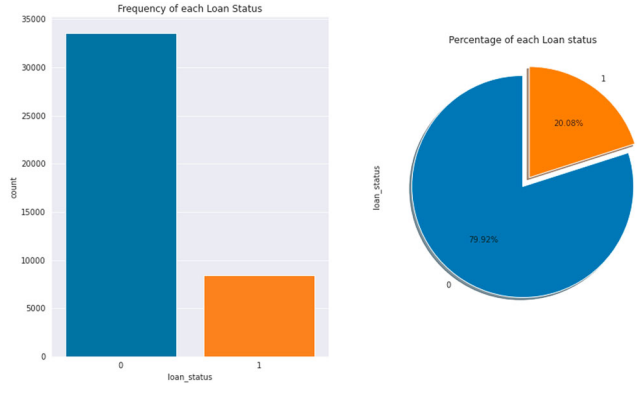

Figure 6. Percentage of loan status.

matching individuals. Furthermore, tournament selection was used to select the parents to be married. A number of strings were selected from the population by replacement, and the most suitable pair was obtained for mating. In this paper, three individuals were selected for the tournament selection process. Apart from being selected for mutation, each offspring produced has a slight chance of inverting its features to pass on population diversity. The population size was 20, and the generations were 100 for a total of 2000 iterations. At the end of the generation, which was the 100^{th} generation, GA may produce a chromosome or feature set with the best fitness value. Nevertheless, this event is not guaranteed because more than one set of features may produce the best fitness value. This event is known as the multiple optimal solutions

4. Result and discussion

The list of the most influential features was the output of feature selection using hybrid GA with 10 ML models with the P2P lending dataset as input. Computationally, GA is very time-consuming due to the large number of observation data and features in the dataset. In this paper, 20 populations multiplied by 100 generations resulted in 20.00 iterations, 0.8 CR, and 3 MR levels (0, 0.5, and 1), which are the configurations used by GA. The difference in MR values aims to determine the effect of MR on the fitness value of each offspring or new generation produced.

The fitness level of each model based on the basic mutation rate or MR = 0 is displayed in Figure 7. The results of the fitness values indicate that four models experience performance improvements over time, signifying that these models are becoming more proficient in predicting lender default risk over generations. This improvement in fitness values reflects a positive response to the basic mutation rate in the Genetic Algorithm (GA). In addition, certain models performed best accross generations. For instance, GA+KNN achieved 84.501% accuracy in Generation-16, GA+RF attained 85.938% accuracy in Generation-18, GA+XGBoost reached 86.119% accuracy in Generation-4, and GA+LightGBM achieved 85.213% in Generation-7. These results indicate that certain models tend to excel at certain points in the experiment, and selecting the appropriate model for a specific generation can have a significant impact on the results. Additionally, five models exhibited stable fitness levels, maintaining consistent performance across all generations. These models were GA+LR at 86.016%, GA+GNB at

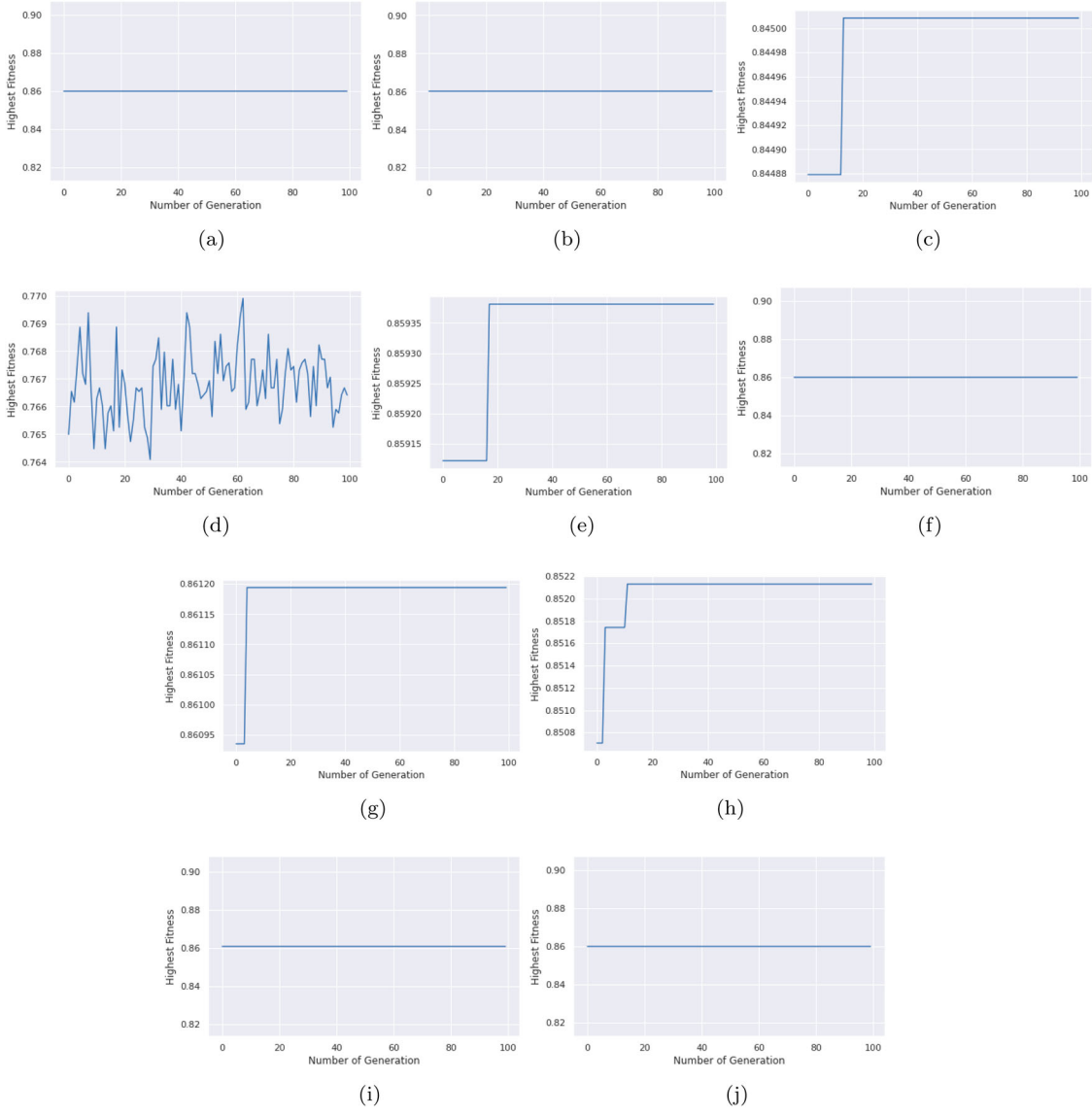

Figure 7. The fitness value of each generation based on basic mutation rate (MR = 0). (a) GA+LR (b) GA+GNB (c) GA+K-NN (d) GA+DT (e) GA+RF (f) GA+GBDT (g) GA+XGBoost (h) GA+LightGBM (i) GA+AdaBoost (j) GA+CatBoost.

86.029%, GA+GBDT at 86.016%, GA+AdaBoost at 86.093%, and GA+CatBoost at 86.016%. The stability in the fitness values of these models can be taken as an indication that they can provide consistent performance in predicting lender default risk. However, it should be noted that the GA+DT model exhibited unstable fitness levels, with significant variations in performance across generations. This suggests that this model may be less consistent in predicting lender default risk. Meanwhile, the GA+XGBoost model remained the best, with stable fitness levels and high accuracy of approximately 86.132%. Thus, the GA+XGBoost model is considered the most effective in predicting lender default risk in the context of P2P services

Based on the medium mutation rate or MR = 0.5, as observed in Figure 8, five models experienced an increase in fitness values across generations. These models achieved their best fitness values, with GA+LR reaching 86.068% in Generation-25, GA+KNN attaining 84.540% in Generation-35, GA+XGBoost achieving 86.119% in Generation- 22, GA+LightGBM reaching 85.174% in Generation-17, and GA+AdaBoost scoring 86.119% in Generation-17. Three other models, namely GA+GNB, GA+GBDT, and GA+CatBoost, demonstrated stable fitness values across all generations, achieving 86.029%, 86.016%, and 86.016%, respectively. However, the GA+DT and GA+RF models exhibited irregular behavior, with their best fitness values achieved in different generations, specifically 77.198% in Generation-37 and 86.003% in Generation-80, respectively. The research results indicate that the GA+XGBoost and GA+AdaBoost models displayed the best performance in terms of fitness values compared to the other models, highlighting the stability and potential of these models in optimal feature selection for lender default risk assessment in P2P services.

Figure 9 illustrates the fitness values in each generation based on the extreme mutation rate or MR = 1. In this experiment, six models exhibited increasing fitness values as the generations progressed, with each achieving its best fitness value. GA+LR reached 86.068% in Generation-40, GA+KNN achieved 84.902% in Generation-100, GA+RF attained 86.106% in Generation-18, GA+XGBoost reached 86.132% in Generation- 75, GA+LightGBM achieved 85.342% in Generation-5, and GA+AdaBoost reached 86.106% in Generation-43. However, similar to MR = 0.5, three other models remained constant across all generations with the same best fitness value.

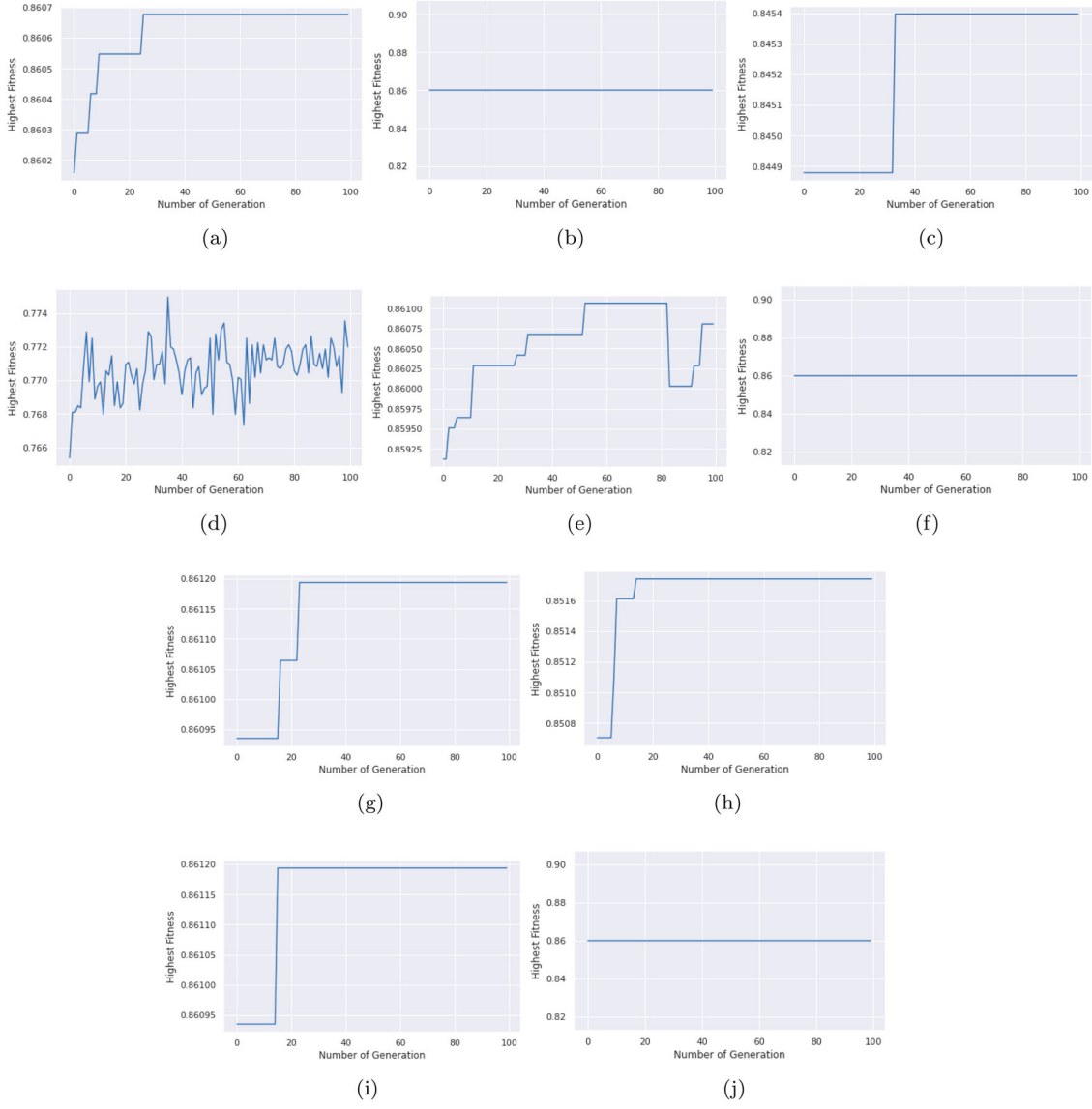

Figure 8. The fitness value of each generation based on medium mutation rate (MR = 0.5). (a) GA+LR (b) GA+GNB (c) GA+K-NN (d) GA+DT (e) GA+RF (f) GA+GBDT (g) GA+XGBoost (h) GA+LightGBM (i) GA+AdaBoost (j) GA+CatBoost.

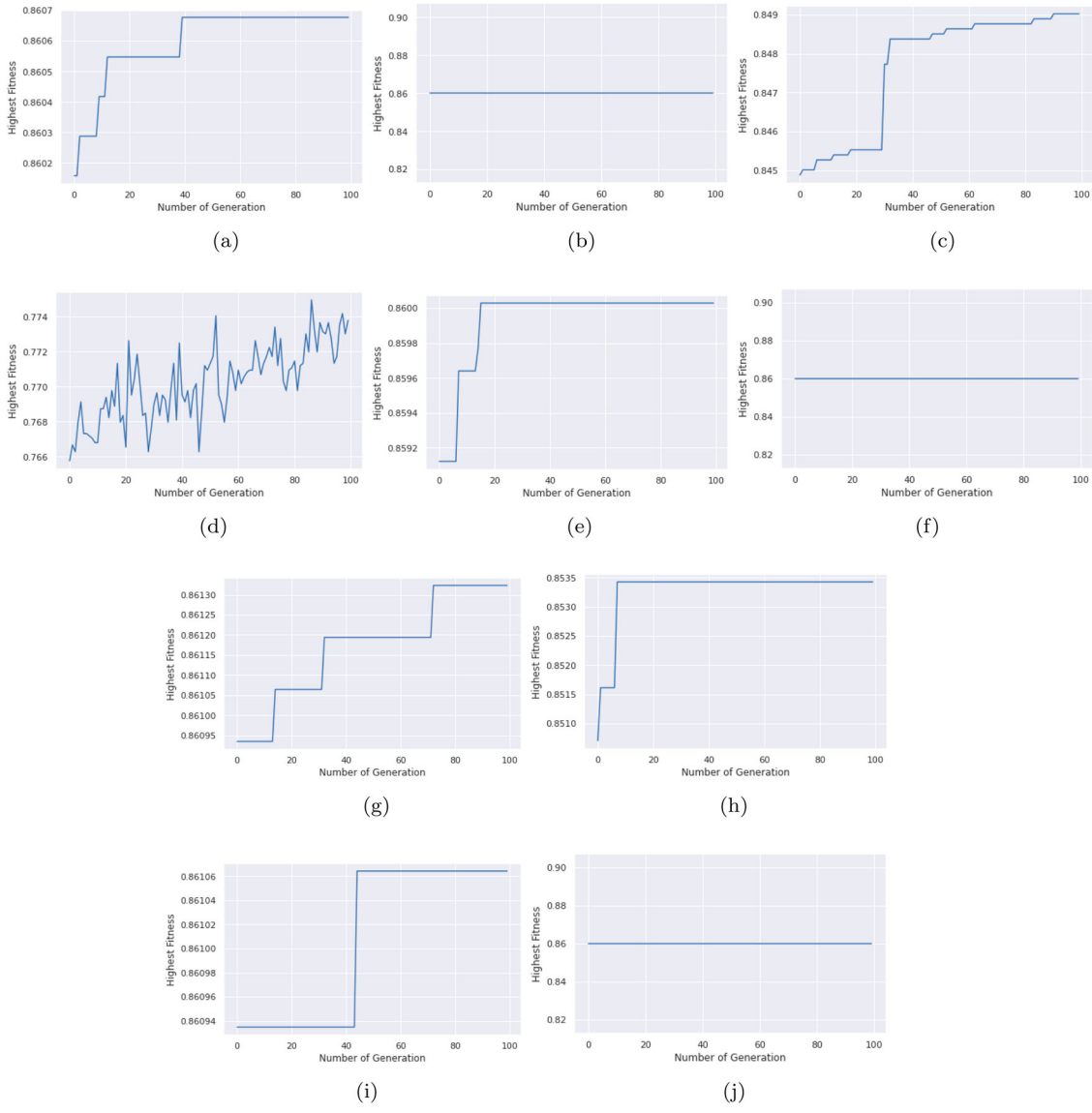

Figure 9. The fitness value of each generation based on extreme mutation rate (MR = 1). (a) GA+LR (b) GA+GNB (c) GA+K-NN (d) GA+DT (e) GA+RF (f) GA+GBDT (g) GA+XGBoost (h) GA+LightGBM (i) GA+AdaBoost (j) GA+CatBoost.

The random model GA+DT had the best fitness value of 77.496% in Generation-85. Changing the mutation rate to MR = 1 in this experiment influenced the improved performance of certain models, with GA+XGBoost and GA+AdaBoost consistently outperforming others, indicating a positive response to mutation variation in this study.

Based on the analysis of Figures 7, 8, and 9, this experiment has yielded significant results. First, three models, namely GA+GNB, GA+GBDT, and GA+CatBoost, demonstrated consistency in their performance across different mutation rates, providing reliability in feature selection. Second, most models, including GA+LR, GA+KNN, GA+DT, GA+RF, and GA+XGBoost, showed improved performance with increasing mutation rates, indicating a positive response to mutation rate variations. This highlights the potential of mutation rate elevation strategies in feature selection optimization. Third, only the GA+DT model exhibited a performance decline at mutation rate MR = 0.5, but its performance recovered at MR = 1, highlighting the sensitivity of this model to mutation rate changes. Fourth, GA+XGBoost consistently proved to be the best-performing model with stable performance across different mutation rates, demonstrating its potential in feature selection and risk management in P2P services. The comparison of fitness values of each model based on the MR level is shown in Figure 10. Mutation rates play a crucial role in model performance for optimal feature selection, and this research has shown that variations in mutation rate can be leveraged to improve model performance. GA+XGBoost has proven to be a preferred model with stable performance, while there is still potential for further research into factors that influence model responsiveness to mutation rates. This research has the potential to make a valuable contribution to financial risk management in the P2P lending industry. (Table 2)

The features selected based on the hybrid GA model with 10 machine learning (ML) models are detailed in Figure 2. Each feature is represented as F (e.g. F1: loan_amnt; F2: int_rate; F3: installment; F4: annual_inc, etc), where 1 indicates a selected feature, and 0 indicates an unselected feature. On average, 50% of the total 37 features were selected, with breakdowns of 41% for GA+LR, 51% for GA+GNB, 54% for GA+KNN, 38% for GA+DT, 35% for GA+RF, 54% for GA+GBDT, 54% for GA+XGBoost,

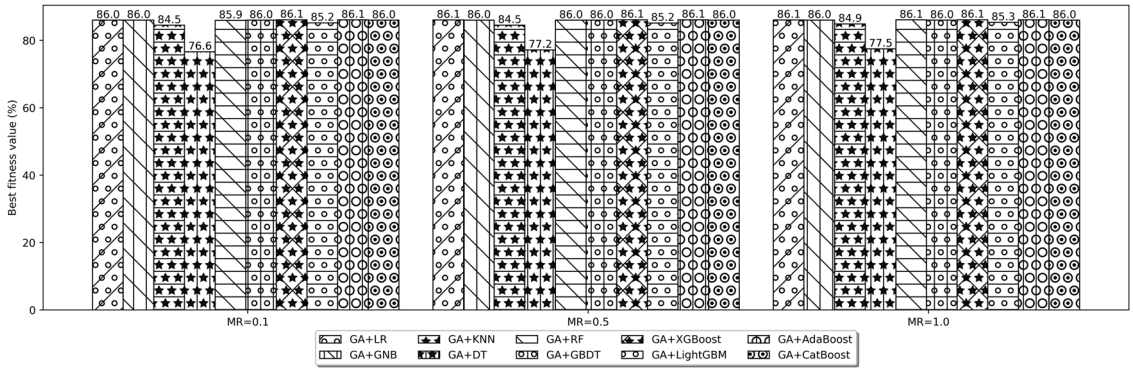

Figure 10. The fitness value comparison of each model based on the MR level.

Table 2. The selected feature on P2P lending dataset.

	GA+ Hybrid Model									
F	LR	GNB	KNN	DT	RF	GBDT	XGBoost	LightGBM	AdaBoost	CatBoost
F0	0	0	0	1	0	1	0	0	0	1
F1	1	0	1	0	1	1	0	1	1	0
F2	0	1	0	1	1	1	1	1	0	0
F3	0	1	0	1	1	1	1	1	1	0
F4	0	0	0	0	0	0	0	0	0	0
F5	0	1	1	0	1	0	1	1	0	0
F6	0	1	0	0	0	1	1	0	1	1
F7	0	0	0	1	1	0	0	1	1	1
F8	0	1	0	0	1	1	1	1	0	1
F9	0	0	0	1	0	1	0	1	0	1
F10	0	0	1	1	0	0	0	1	1	1
F11	0	0	1	1	0	0	0	1	1	1
F12	0	1	1	0	1	1	1	0	1	1
F13	1	1	1	0	1	0	1	1	1	0
F14	1	1	0	1	0	1	1	1	1	1
F15	1	0	1	0	0	0	0	0	1	1
F16	0	1	0	0	0	1	1	1	0	1
F17	0	1	0	0	0	1	1	1	1	1
F18	1	1	1	1	0	1	1	0	1	1
F19	0	0	1	1	1	0	0	1	0	1
F20	0	0	1	0	0	1	0	1	0	0
F21	0	0	0	1	0	1	0	0	0	1
F22	1	0	0	0	0	0	0	0	1	1
F23	1	1	1	0	0	0	1	0	0	0
F24	0	1	1	0	0	0	1	0	1	0
F25	0	1	0	0	1	1	1	1	1	1
F26	0	0	0	0	1	0	0	1	1	1
F27	1	0	1	1	1	1	0	0	0	1
F28	1	1	0	0	0	0	1	0	0	0
F29	0	0	0	1	1	0	0	1	0	0
F30	0	0	1	0	0	1	0	1	1	0
F31	1	1	1	0	0	1	1	0	0	0
F32	1	0	1	1	0	0	0	1	1	1
F33	1	1	1	0	0	1	1	1	0	0
F34	1	0	1	0	0	1	1	1	0	0
F35	1	1	1	0	0	0	1	1	0	0
F36	1	1	1	0	0	0	1	1	1	0
T	15	19	20	14	13	20	20	24	19	20

Note: 1 is selected; 0 is not selected

65% for GA+LightGBM, 51% for GA+AdaBoost, and 54% for GA+CatBoost. These results indicate a relatively balanced distribution of feature selection among the various ML models used in the experiment, with GA+LightGBM showing the highest feature selection rate at 65%. This suggests that GA+LightGBM places more emphasis on certain features compared to other models, possibly recognizing the significance of these features in making accurate predictions. On the other hand, models such as GA+RF and GA+DT had lower feature selection rates, indicating a more conservative approach to feature selection. The diversity in feature selection among the models reflects their unique strategies and preferences, which can be valuable in understanding the contribution of specific features in assessing lender default risk. Overall, a comprehensive analysis of feature selection contributes to a deeper understanding of how different ML models approach feature importance, which can potentially support more informed feature engineering decisions for risk assessment in P2P lending services.

The Hybrid GA+10 ML model for feature selection in the P2P lending service dataset has significantly improved various model evaluation parameters. Figure 11(a) shows a comparison of accuracy scores between the original ML model and the proposed model. Hybrid GA+10 managed to improve the accuracy scores of the original model with detailed improvements of 0.155% for LR, 1.217% for GNB, 0.764% for K-NN, 1.619% for DT, 0.285% for RF, 2% for GBDT, 0.117% for RF, 0.414% for AdaBoost, and 0.078% for CatBoost. Figure 11(b) illustrates the comparison of the recall scores between the original ML models and the proposed one. There was a significant increase in recall scores, with LR increasing by 5.49%, Gaussian NBC by 7.86%, K-NN by 6.77%, DT by a remarkable 22.02%, RF by 6.8%, GBDT by 7.57%, XGBoost by 2.48%, LightGBM by 4.81%, and both AdaBoost and CatBoost by 3%.

Similarly, Figure 11(c) displays the comparison of the precision scores between the original ML models and the proposed one. The results show an increase in precision for several models, such as LR increasing by 1.9%, Gaussian NBC improving by 3.14%, K-NN increasing by 2.44%, DT showing a significant increase of 7.79%, RF increasing by 2.65%, GBDT with an increase of 2.78%, XGBoost improving by 0.79%, LightGBM increasing by 6.56%, and both AdaBoost and CatBoost increasing by 2.86%. Figure 11(d) depicts the comparison of f1-score between the original ML models and the proposed one. In terms of f1-score, there was a significant increase in several models, with the most notable improvement observed in the DT model with an increase of 15.59%, followed by Gaussian NBC with an increase of 5.61%, K-NN with an increase of 4.70%, RF with an improvement of 4.80%, XGBoost with an increase of 5.26%, LightGBM with an increase of 5.64%, and both AdaBoost and CatBoost with an increase of 2.93%.

The Hybrid GA+10 ML model for feature selection in the P2P lending service dataset has successfully made a significant positive impact on various aspects of model evaluation. The experiment results indicate that this approach consistently improves the performance of the original ML models in predicting the risk of borrower default in P2P lending services. There was a substantial improvement in the accuracy, recall, and precision scores of various models, such as LR, GNB, K-NN, DT, RF, GBDT, XGBoost, LightGBM, AdaBoost, and CatBoost. These results reflect the significant

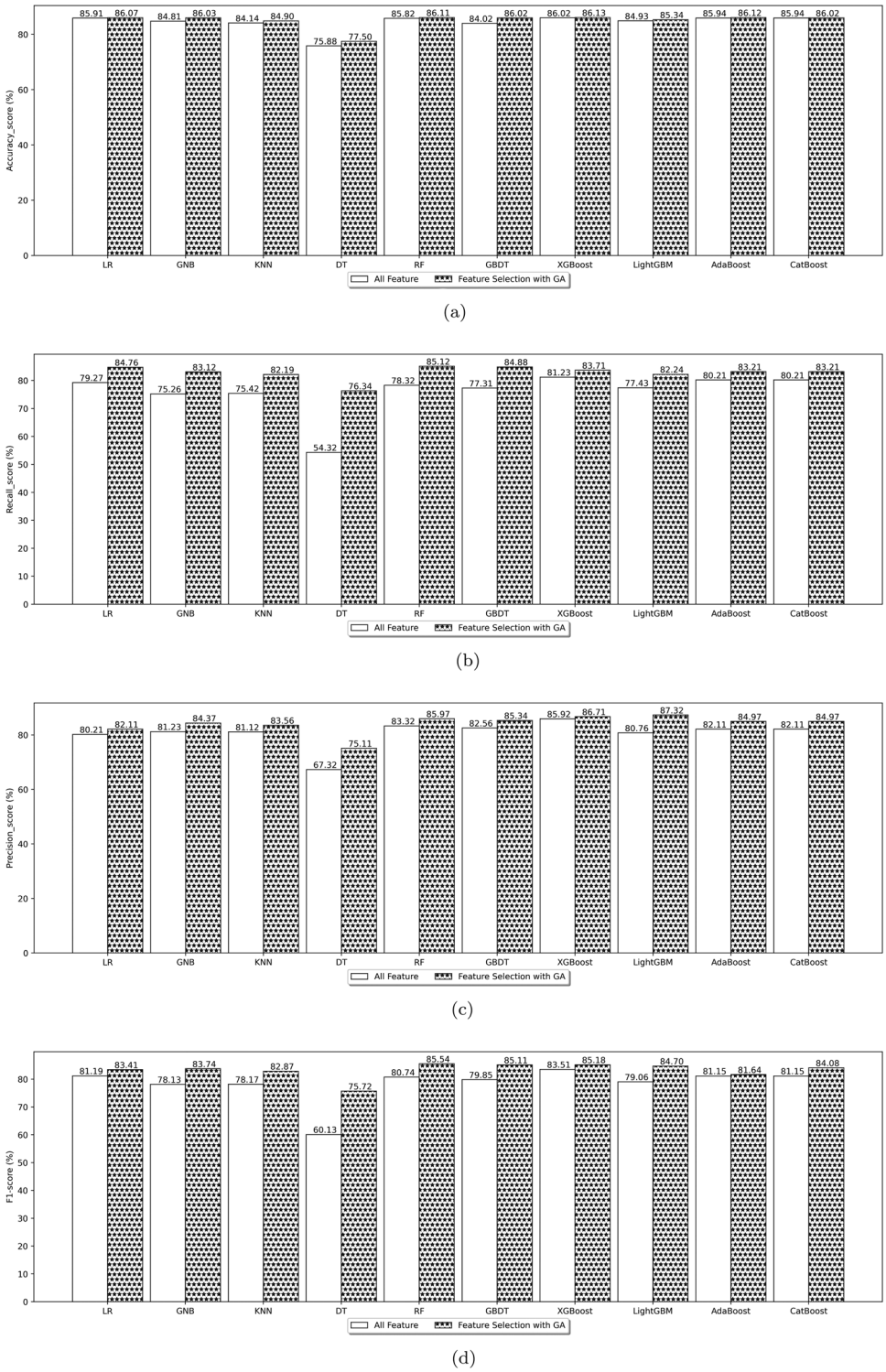

Figure 11. Comparison of original ML model with the GA+ML hybrid model. (a) Accuracy (b) Recall (c) Precision (d) F1-Score.

potential of the Hybrid GA+10 ML approach to support risk management and better decision making in the P2P lending industry. With improved accuracy, these models can more accurately identify potential defaulting borrowers, reducing risk for lenders. Additionally, the increased recall and precision also suggest that this approach can help minimize prediction errors, allowing for smarter credit allocation decisions. Thus, this experiment provides valuable insights into the effectiveness and potential applications of the Hybrid GA+10 ML approach in improving the quality of risk assessment in P2P lending services.

The proposed model (GA+XGBoost) showed the highest performance, achieving an accuracy of 86.132%. This high accuracy indicates that this model is very good at predicting the risk of borrower default. The implication is that P2P lenders can effectively utilize this model as a tool to identify potential borrowers who are highly likely to repay their loans on time. With such a high level of

Table 3. Comparison of the proposed model with those of the previous related works.

Method	Feature selection	#of Feature (%)	Accuracy (%)
LDA	RBM	80 (49.65%)	81.2
LR			81.05
ANN			66.08
K-ANN			72.05
SVM			72.6
RF			67.72
ERT	BPSO-SVM	19 (55.88%)	64
Proposed Model	GA+ML	20 (54%)	86.132

accuracy, lenders can better minimize their risks. They can confidently approve loans to borrowers who have been identified by this model as having a high likelihood of repayment, thereby reducing the risk of default and improving the overall quality of their P2P loan portfolios. Furthermore, the high level of accuracy can assist in identifying high-risk borrowers, allowing lenders to take additional steps, such as charging higher interest rates or denying loans to borrowers with higher risk profiles. This, in turn, improves risk management and profitability for lenders. Overall, achieving a high level of accuracy with the GA+XGBoost model is a positive step toward more effective risk management in the P2P lending industry, ultimately increasing lender confidence and reducing potential losses.

We compared the performance of the proposed model with the results of previous studies that have used the same Lending Club dataset. Table 3 summarizes the comparative study of the proposed model with previous related works. The proposed model has generally outperformed all the models in previous studies applied to the same dataset, including the Restricted Boltzmann Machine (RBM) with several classifier models such as LDA, LR, ANN, k-NN, SVM, and RF [36], and Binary Particle Swarm Optimization-SVM (BPSO-SVM)+Extremely Randomized Trees (ERT) [37].

This research provides an important contribution to the existing research on the prediction of the risk of default in P2P lending services. The research results show that the proposed model, GA+XGBoost, has an excellent performance, with an accuracy rate of 86.132%. Comparing the performance of this model with that of previous related studies, GA+XGBoost significantly outperforms the models in previous studies. For example, this model outperforms RBM with several classification models, such as LDA, LR, ANN, k-NN, SVM, and RF, in previous studies. This shows that the approach used in this research, namely the combination of GA and XGBoost, has brought significant improvements in predicting default risk in P2P lending. These results provide a solid foundation for further research in this domain.

This research makes a significant contribution to existing studies in predicting default risk in P2P lending services. The proposed model, GA+XGBoost, has demonstrated excellent performance, outforming models in previous related research. This means that P2P lenders can effectively use this model to identify borrowers with a high likelihood of repaying loans on time, thereby reducing the risk of default and improving the quality of their loan portfolios. Additionally, P2P lending platform operators can utilize this model as an additional evaluation tool to assess the quality of borrowers registering on their platforms, while investors can use it as a guide to select lower-risk investment projects. The research also provides a strong foundation for further studies in this domain, with the potential for the data used in this research to become a standard dataset for future research in predicting default risk. Overall, this research has significant practical implications for the P2P lending industry in terms of risk management and investment decision making. Furthermore, it highlights the potential of feature selection in improving the performance of default risk prediction models, providing valuable guidance to lenders or P2P lending platforms in selecting the most influential features in credit decision making. This research provides a solid foundation for further exploration of feature selection methods that can improve the performance of default risk prediction models in the P2P lending industry.

5. Conclusion and future work

This study presented a feature selection approach in P2P lending datasets utilizing the GA+10ML hybrid model. The GA was configured with 20 populations, 100 generations, and 5000 iterations, along with a crossover rate (CR) of 0.8 and mutation rate (MR) levels, including basic (MR = 0), medium (MR = 0.5), and extreme (MR = 1). The results of the model runs demonstrated four models with consistently high fitness values: GA+GNB, GA+GBDT, GA+GBDT, and GA+CatBoost. GA+LR, GA+KNN, GA+DT, GA+RF, and GA+XGBoost exhibited an increase in fitness values from MR = 0 to MR = 1. Only one model experienced a decrease in fitness value from MR = 0 to MR = 0.5, followed by an increase after MR = 1. The best fitness value for each model varied with the MR level in each generation. MR significantly influenced the fitness values generated in each generation, with GA+XGBoost being the most stable model with consistent fitness values compared to others. Additionally, the feature selection process resulted in an average selection of 50% of the total 37 features. Future work in this area could explore various population sizes, generation counts, and iterations to further optimize the GA configuration. Additionally, exploring alternative mutation and crossover strategies may provide insight into improving feature selection performance. Evaluating the robustness and scalability of the model on larger P2P lending datasets is another avenue of research. Moreover, an in-depth analysis of the interpretability and relevance of the selected features for credit risk assessment may provide valuable insights for both lenders and borrowers in the P2P lending industry. Furthermore, extending this research to consider temporal dynamics and changes in borrower behavior over time may lead to more accurate credit risk prediction models, ultimately benefiting the P2P lending industry and its stakeholders.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on Contributors

Muhammad Sam'an received Bachelor Degree from Universitas Negeri Semarang and Master Degree from Universitas Diponegoro in Mathematics 2010 and 2016 respectively. His research interests are in optimization, fuzzy mathematics and computational mathematics. He can be contacted at email: muhammad92sam@unimus.ac.id.

Muhammad Munsarif received the Doctoral Degree in Computer science from Dian Nuswantoro University (UDINUS). Currently, he is a lecturer in informatics Engineering at Muhammadiyah University, Semarang (UNIMUS). His research interests include computer vision, data science and technopreneuership. He can be contacted at email: m.munsarif@unimus.ac.id.

Safuan received the Master Degree in Informatics Engineering from Dian Nuswantoro University (UDINUS) in 2015. Currently, he is a lecturer in informatics Engineering at Muhammadiyah University, Semarang (UNIMUS). His research interests include data mining, programming and web security. He can be contacted at email: safuan@unimus.ac.id.

Yahya Nur Ifriza received 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.

References

[1] Lee E, Lee B, Chae M. Herding behavior in online P2P lending: An empirical investigation. 7-11 July 2011. Brisbane, Australia. PACIS 2011 - 15th Pacific Asia Conference on Information Systems: Quality Research in Pacific. 2011.

[2] Jin Y, Zhu Y. A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending. 04-06 April 2015. Gwalior, India. 2015 Fifth International Conference on Communication Systems and Network Technologies. 2015.

[3] Byanjankar A, Heikkila M, Mezei J. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. 11 January 2016. Cape Town, South Africa. 2015 IEEE Symposium Series on Computational Intelligence. 2015.

[4] Malekipirbazari M, Aksakalli V. Risk assessment in social lending via random forests. Expert Syst Appl. 2015;42(10):4621–4631. doi: 10.1016/j.eswa.2015.02.001

[5] Li H, Zhang Y, Zhang N, et al. Detecting the abnormal lenders from P2P lending data. Procedia Comput Sci. 2016;91:357–361. doi: 10.1016/j.procs.2016.07.095

[6] Serrano-Cinca C, Gutierrez-Nieto B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decis Support Syst. 2016;89:113–122. doi: 10.1016/j.dss.2016.06.014

[7] Yan J, Wang K, Liu Y, et al. Mining social lending motivations for loan project recommendations. Expert Syst Appl. 2018;111:100–106. doi: 10.1016/j.eswa.2017.11.010

[8] Xia Y, Yang X, Zhang Y. A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. Electron Commer Res Appl. 2018;30:111–124. doi: 10.1016/j.elerap.2018.05.011

[9] Madaan M, Kumar A, Keshri C, et al. Loan default prediction using decision trees and random forest: a comparative study. IOP Conf Ser: Materials Science and Engineering. 2021;1022(1):012042. doi: 10.1088/1757-899X/1022/1/012042

[10] Sharma AK, Li LH, Ahmad R. Identifying and predicting default borrowers in P2P lending platform: A machine learning approach. 29-31 August 2021. Taichung, Taiwan. 2021 IEEE International Conference on Social Sciences and Intelligent Management (SSIM). 2021.

[11] Tumuluru P, Burra LR, Loukya M. Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. 23-25 February 2022. Coimbatore, India. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). 2022.

[12] Wang Y, Zhang Y, Lu Y, et al. A comparative assessment of credit risk model based on machine learning – a case study of bank loan data. Procedia Comput Sci. 2020;174:141–149. doi: 10.1016/j.procs.2020.06.069

[13] Yamparala R, Saranya JR, Anusha P, et al. Predictıng the loan using machine learning. In: Soft computing for security applications. Advances in intelligent systems and computing. 2023. p. 701–712.

[14] Munsarif M, Sam'an M, Safuan . Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning. Bull Electrical Eng Inf. 2022;11(6):3483–3489.

[15] Chen RC, Dewi C, Huang SW, et al. Selecting critical features for data classification based on machine learning methods. J Big Data. 2020;7(1):52. doi: 10.1186/s40537-020-00327-4

[16] Al-Zoubi AM, Faris H, Alqatawna J, et al. Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. Knowl Based Syst. 2018;153:91–104. doi: 10.1016/j.knosys.2018.04.025

[17] Kumar N, Singh AK, Srivastava S. Feature selection for interest flooding attack in named data networking. Int J Computers Appl. 2021;43(6):537–546.

[18] Thiyam B, Dey S. Statistical methods for feature selection: unlocking the key to improved accuracy. International Journal of Computers and Applications. 2023;45:433–443. doi: 10.1080/1206212X.2023.2223795

[19] Wang HD. Research on the features of car insurance data based on machine learning. Procedia Comput Sci. 2020;166:582–587. doi: 10.1016/j.procs.2020.02.016

[20] Jain D, Singh V. A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification. Int J Computers Appl. 2021;43(6):524–536.

[21] Papouskova M, Hajek P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decis Support Syst. 2019;118:33–45. doi: 10.1016/j.dss.2019.01.002

[22] Zheng X. Feature selection algorithm of network attack big data under the interference of fading noise. Int J Computers Appl. 2022;44(9):807–813.

[23] Angadi S, Reddy VS. Multimodal sentiment analysis using reliefF feature selection and random forest classifier. Int J Comput Appl. 2021;43(9):931–939.

[24] Fathima MD, Samuel SJ, Natchadalingam R, et al. Majority voting ensembled feature selection and customized deep neural network for the enhanced clinical decision support system. Int J Comput Appl. 2022;44(10):991–1001.

[25] Gu S, Cheng R, Jin Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Comput. 2018;22(3):811–822. doi: 10.1007/s00500-016-2385-6

[26] Xu J, Chen D, Chau M. Identifying features for detecting fraudulent loan requests on P2P platforms. 28-30 September 2016.Tucson, AZ, USA. 2016 IEEE Conference on Intelligence and Security Informatics (ISI). 2016.

[27] Ha V-S, Lu D-N, Choi GS. Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Feature selection with Deep learning. 17-20 February 2019. PyeongChang, Korea (South). 2019 21st International Conference on Advanced Communication Technology (ICACT). 2019.

[28] Li X, Ergu D, Zhang D. Prediction of loan default based on multi-model fusion. 9-11 july 2021. Chengdu, China. the 8th International Conference on Information Technology and Quantitative Management. 2021.

[29] Yang R, Wang P, Qi J. A novel SSA-CatBoost machine learning model for credit rating. J Intell Fuzzy Syst. 2023;44(2):2269–2284. doi: 10.3233/JIFS-221652

[30] Yin W, Kirkulak-Uludag B, Zhu D, et al. Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending. Appl Soft Comput. 2023;142:110302. doi: 10.1016/j.asoc.2023.110302

[31] Cao W, He Y, Wang W, et al. Ensemble methods for credit scoring of Chinese peer-to-peer loans. J Credit Risk. 2021;17(3):79–115.

[32] Victor L, Raheem M. Loan default prediction using genetic algorithm: a study within Peer-To-Peer lending communities. Int J Innovative Sci Res Technol. 2021;6(3):1195–1205.

[33] Lappas PZ, Yannacopoulos AN. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl Soft Comput. 2021;107:107391. doi: 10.1016/j.asoc.2021.107391

[34] Ye X, Dong L, Ma D. Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score. Electron Commer Res Appl. 2018;32:23–36. doi: 10.1016/j.elerap.2018.10.004

[35] Zhu L, Qiu D, Ergu D, et al. A study on predicting loan default based on the random forest algorithm. Procedia Comput Sci. 2019;162:503–513. doi: 10.1016/j.procs.2019.12.017

[36] Nguyen Truong T, Khuat Thanh S, Ngo Thi Thu T, et al. Improve risk prediction in online lending (P2P) using feature selection and deep learning. Int J Computer Sci Network Security. 2019;19(11):216–222.

[37] Setiawan N, Suharjito D.. A comparison of prediction methods for credit default on peer to peer lending using machine learning. Procedia Comput Sci. 2019;157:38–45. doi: 10.1016/j.procs.2019.08.139

[38] George N. All lending club loan data. Kaggle, 2019. Available at https://www.kaggle.com/datasets/wordsforthewise/lending-club/metadata.