Enhancing Default Prediction in P2P Lending using Random Forest and Grey Wolf Optimization-based Feature Selection

Bagus Winarko Nugroho¹, Purwanto²
¹,²Universitas Dian Nuswantoro, Universitas Dian Nuswantoro
E-mail: P31202002371@mhs.dinus.ac.id, purwanto@dsn.dinus.ac.id

Heribertus Himawan³
³Universitas Dian Nuswantoro
E-mail: himawan26@dsn.dinus.ac.id

Abstract - Online lending services such as Peer to Peer (P2P) loans provide convenience for lenders to transact directly without involving banks as intermediaries. Identifying potential loan recipients who are at risk of default is a crucial step in preventing financial losses, as lenders are responsible for default risk. However, predicting default risk becomes a challenge when P2P lending datasets have various complex features. Some features in P2P lending are redundant, while others do not significantly contribute to an effective solution. Therefore, feature selection is an important process to choose a relevant subset of features from input or target data. Traditional feature selection methods often fail to provide optimal results. A better approach is to use heuristic search algorithms capable of finding suboptimal feature subsets. We employ the Grey Wolf Optimization (GWO) technique, inspired by the hierarchy of leadership and grey wolf hunting mechanisms. Combined with Random Forest (RF), which has limitations in classifying data with very high dimensions, our GWO+RF combination has proven to enhance classification performance better than previous research. It achieves an accuracy score of 97.31%, compared to previous research with scores of only 67.72% for RBM+RF, 64% for Binary PSO+ERT, and 92% for GA+RF.

Keywords – P2P lending; Feature selection; Grey Wolf Optimization; Random Forest; Accuracy

1. INTRODUCTION

P2P lending has the potential to bring about significant changes to the trajectory of traditional banks in the future. Being the world’s largest digital credit marketplace, P2P lending offers various types of loans, including personal, business, and medical loans. The primary objective of P2P lending when it emerged in 2005 was to democratize access to more efficient consumer financial services. This approach involves individuals presenting loan proposals, which are then approved by investors or lenders, bypassing the need for a formal financial institution’s involvement. The inception of P2P lending occurred in the UK with Zopa in 2005. Subsequently, in 2006, the United States introduced LendingClub and Prosper, while China established its own credit platforms.

Numerous researchers have examined P2P lending through the introduction of a diverse array of models. This includes the utilization of a stacking ensemble model for the assessment
of personal credit risk[1]. Logistic regression, hinged on credit scores, and linear regression, based on profit scores, have been employed to forecast the likelihood of default and potential profits within a novel loan recommendation framework[2]. In terms of default prediction, three statistical models (Logistic Regression (LR), Bayesian Classifier, and Linear Discriminant Analysis (LDA)) along with five AI models (Decision Tree, Random Forest, Light-GBM, Artificial Neural Network (ANN), and Convolutional Neural Network (CNN)) have been utilized[3]. A multi-view deep neural network has been crafted to address default prediction challenges inherent in imbalanced and intricate datasets[4]. Within the realm of credit risk scoring, there exists a benchmarking model grounded in machine learning techniques[5]. The forecasting of default risk on unbalanced datasets has been approached via three models: Random Forest, Neural Network, and Logistic Regression[6]. Additionally, an innovative resampling ensemble model, rooted in data distribution, has been proposed for the assessment of credit risk within imbalanced datasets[7]. The exploration of multi-view ensemble learning, incorporating the distance-to-model concept and adaptive clustering, has been carried out for credit risk assessment within imbalanced datasets[8]. Predictive models, encompassing Naive Bayes, Decision Tree, and Boosted Decision Tree methodologies, have been employed for default prediction[9][10]. Credit risk assessment has been facilitated through Regression models[11]. A multi-round ensemble learning model, leveraging heterogeneous ensemble frameworks, has been developed for the prediction of defaults[12]. ANN-based models have been harnessed to categorize credit scores for both default and non-default classifications in the context of P2P lending[13]. Malekipirbazari and Aksakalli have implemented Random Forest (RF) for the identification of top-tier borrowers, contrasting them with FICO credit scores derived from LC scores[14]. Abnormal investor identification and the prediction of potential investors have been rooted in outlier detection utilizing poor credit scores[15]. Lastly, the utilization of the internal rate of return has been employed to anticipate the anticipated profits for investors[16].

P2P lending datasets, extracted from P2P lending platforms, frequently contain irrelevant or redundant features. Consequently, the predictive performance of models often falls short of optimal levels, yielding inaccurate outcomes[17]. The extensive size and complexity of P2P lending datasets lead to suboptimal and inefficient model performance[18]. For instance, the processing duration tends to elongate due to the extensive feature processing required[19]. To mitigate these issues, feature selection emerges as a solution. This method discerns pertinent features, those exerting significant influence on the prediction process. Moreover, it serves to curtail data dimensions and eliminate irrelevant features, thereby enhancing classification model accuracy[20].

Feature selection methods tailored for P2P loan default prediction have been extensively proposed. The Max-Relevance and Min-Redundancy (MRMR) technique is employed for feature selection, and k-means clustering aids in discarding irrelevant attributes[1]. Various models, including LightGBM (Light Gradient Boosting Machine)[21], Random Forest[22][23], Logistic Regression[24], Random Forest and XGBoost[24], Binary Particle Swarm Optimization (PSO) with Support Vector Machines (SVM)[25], Adaptive Feature Selection based on Most Informative Graph and Most Relative Graph[26], and Grey Relational Clustering, have demonstrated the superior accuracy of results achieved through feature selection compared to unselected features[27].

However, traditional feature selection methods face limitations in addressing the challenge. The transition from a set of different N-sized features to a 2N-sized vector exacerbates the already vast feature space[28]. The evaluation of 2N subsets falls under the category of np-hard problems[20][22]. As an alternative solution for such complexities, Evolutionary Algorithms (EA) step in. The Gray Wolf Optimization (GWO), introduced by Mirjalili, Mirjalili, and Lewis, is a type of EA well-suited for optimizing feature selection within extensive feature spaces[29]. GWO
incorporates a hierarchical leadership concept and the hunting mechanism of gray wolves, making it effective for addressing optimization problems. Therefore, this paper introduces a novel feature selection approach founded on the GWO technique and Machine Learning models. The objective is to achieve highly accurate loan default prediction for P2P lending, utilizing data collected from P2P lending platforms. What sets it apart from previous research is that prior studies employed less effective feature selection algorithms when dealing with high-dimensional feature datasets. Therefore, the GWO algorithm combined with Random Forest classification serves as the optimal solution for enhancing the evaluation performance, as we will discuss in the results section.

2. RESEARCH METHOD

Figure 1 illustrates the process flow of the research methodology, specifically the employment of Gray Wolf Optimization for Feature Selection in the prediction of defaults within the context of P2P lending.

2.1. P2P lending dataset

The research dataset is sourced from the Kaggle repository, a dataset of online P2P loans available at the following link: https://www.kaggle.com/competitions/ec-dsm-1-spring-2023-lendingclub/data in the year 2023. The P2P loan dataset comprises 16,000 records with 29 features, as depicted in Figure 2.

2.2. Data Preparation

2.2.1. Exploratory Data Analysis
Exploratory Data Analysis (EDA) is employed to detect absent data, trends, and anomalies within P2P lending datasets. The objective of EDA is to grasp the pertinent attributes influencing the dependent variable. EDA also contributes to enhancing the model's predictive capability. This procedural approach aligns with data preprocessing to detect and rectify data inconsistencies effectively. Figure 3 illustrates the findings from the exploration and analysis of data, unveiling the correlations among numerical attributes within the P2P Lending dataset.

2.2.2. Pre-Processing Data

The objective of preprocessing the data involves amalgamating, refining, and minimizing the dataset. The Sklearn library is utilized to detect any missing values, and all features within the P2P lending dataset exhibit no missing values. Features indicating outliers, specifically 'grade' and 'sub grade,' are excluded. Subsequently, One-Hot Encoding is applied to object data types such as term, home ownership, purpose, verification status, and payment plan. The conclusive form of the P2P lending dataset, prepared for analysis, is illustrated in Figure 4. It comprises 41 independent features, with 'default' serving as the target variable.

```
RangeIndex: 15996 entries, 0 to 15999
Data columns (total 29 columns):
#   Column          Non-Null Count   Dtype
0   Loan_amnt      15996 non-null   int64
1   term           15996 non-null   object
2   grade          15996 non-null   int64
3   sub_grade      15996 non-null   int64
4   emp_length     15996 non-null   int64
5   home_ownership 15996 non-null   object
6   purpose        15996 non-null   object
7   annual_inc     15996 non-null   float64
8   dti             15996 non-null   float64
9   verification_status 15996 non-null object
10  delinq_2yrs    15996 non-null   int64
11  fico_range_low 15996 non-null   int64
12  fico_range_high 15996 non-null  int64
13  revol_bal      15996 non-null   int64
14  total_acc      15996 non-null   int64
15  installment    15996 non-null   float64
16  int_rate       15996 non-null   float64
17  out_prncp      15996 non-null   float64
18  last_pymnt_amt 15996 non-null   float64
19  total_rec_int  15996 non-null   float64
20  total_rec_late_fee 15996 non-null float64
21  pymnt_plan     15996 non-null   object
22  recoveries     15996 non-null   float64
23  collection_recovery_fee 15996 non-null float64
24  default        15996 non-null   int64
```

Figure 2. Feature of P2P Lending dataset original
Figure 3. Lending dataset correlation matrix

Figure 4. Dataset P2P Lending after pre-processing
2.3. Proposed Model

Mirjalili, Mirjalili, and Lewis introduced an approach to Evolutionary Algorithms (EA) that draws inspiration from social hierarchies. This strategy, known as Gray Wolf Optimization (GWO), is influenced by the hunting behavior of gray wolves. Gray wolves are recognized as apex predators within the food chain, typically residing in packs with an average size of 5-12 individuals.

• Alpha (α): The leader of the wolf pack, either male or female, responsible for crucial decisions like hunting, sleeping, and setting the pack’s schedule.
• Beta (β): Subordinates who assist the α in making decisions and other tasks. β could be male or female, and often emerges as a potential successor to α.
• Delta (δ): These individuals submit to α and β but have authority over ϖ. Roles like scouts, sentinels, elders, hunters, and caretakers fall within this category.
• Omega (ϖ): Acts as a scapegoat and must obey the commands of other pack members. ϖ represents the lowest-ranking and weakest wolf.

Beyond the hierarchical structure exhibited by the Gray Wolf, the phenomenon of group hunting constitutes a captivating social behavior within this species. The primary stages of the Gray Wolf’s hunting process encompass tracking, pursuit, and assault of prey. Both the hunting techniques of Gray Wolves and their social hierarchy are subjected to mathematical modeling to construct the GWO and carry out optimization procedures. Within the mathematical model of GWO, α is denoted as the most potent solution, β as the second-best solution, and ϖ as the third-best solution. The remaining solution candidates are treated as ϖ. During the hunt, guidance is provided by α, β, and δ, while ϖ follows these three contenders. Consequently, the herd surrounds the prey before initiating the hunt. To express this circular behavior mathematically, the following equations (1)-(2) are employed.

\[ \ddot{X}(t + 1) = \ddot{X}_p(t) + \ddot{A} \cdot \ddot{D} \]  
\[ \ddot{D} = \left| \ddot{C} \cdot \ddot{X}_p(t) - \ddot{X}(t) \right| \]  

Within this context, \( \ddot{D} \) is established according to equation (2), where ‘t’ signifies the total iterations. Coefficient vectors \( \ddot{A} \) and \( \ddot{C} \) come into play, while \( \ddot{X}_p \) denotes the prey’s location and \( \ddot{X} \) represents the Grey wolf’s position. The specific values of \( \ddot{A} \) and \( \ddot{C} \) are ascertained through the application of equations; (3) and (4).

\[ \ddot{A} = 2a \cdot \ddot{r}_1 - a \quad \text{(3)} \]
\[ \ddot{C} = 2\ddot{r}_2 \quad \text{(4)} \]

Here, the value of ‘a’ gradually diminishes in a linear fashion from 2 to 0 throughout the iteration process. Vectors \( \ddot{r}_1 \) and \( \ddot{r}_2 \) are randomly generated with values within the range of [0, 1], thereby facilitating the update of the Grey wolf’s position as outlined in Equation (5).

\( \ddot{X}^\alpha, \ddot{X}^\beta, \) and \( \ddot{X}^\delta \) representing the initial three most optimal solutions within the group during a specific iteration. \( \ddot{A}^1, \ddot{A}^2, \) and \( \ddot{A}^3 \) are defined following the equation in (3). \( \ddot{D}^\alpha, \ddot{D}^\beta, \) and \( \ddot{D}^\delta \) are established based on the equation in (7).
where \( \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3 \) are defined in Eq. (4)

Hence, the Gray Wolf Optimization (GWO) involves a revision of the 'a' parameter, which governs the balance between exploration and exploitation. This parameter is subject to linear updating in each iteration, ranging from 2 to 0, as outlined in Equation (8).

\[
\alpha = 2 - t \frac{2}{\text{max iter}}
\]

\( \text{max iter} \) denotes the maximum permissible number of iterations for the optimization process.

**Algorithm 1:** Pseudo code of the classical grey wolf optimization algorithm

Initialize the prey wolf population \( \mathbf{X}_i \) \( (i = 1, 2, \ldots, n) \)

Initialize \( a, \mathbf{A} \) and \( \mathbf{C} \)

Calculate the fitness of each search agent

\( \mathbf{x}_\alpha \) is the best search agent

\( \mathbf{x}_\beta \) is the second best search agent

\( \mathbf{x}_\delta \) is third best agent

while \( (t < \text{max iter}) \)

for each search agent

update the position of the current search agent by Eq. (15)

end for

Update \( a, \mathbf{A}, \mathbf{C} \)

\( a, \mathbf{A} \) and \( \mathbf{C} \)

Update \( \mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\delta \)

\( t = t + 1 \)

end while

return \( \mathbf{x}_\alpha \)

This paper proposes an effective feature selection method, namely GWO-RF (Grey Wolf - Random Forest), to predict default in P2P lending. There are two main phases in using GWO-RF. First, redundant and irrelevant features based on Figure 2 are eliminated by seeking the best features using GWO. GWO generates population initialization, and then the population's
positions are updated in the discrete search space. Second, the GWO-RF models are executed for the classification process based on the optimal feature set obtained in the first phase. Figure 5 illustrates the workflow of the GWO-RF model.

GWO efficiently exploits the feature space to identify the optimal features within the P2P lending dataset. The optimal feature is the solution that yields the highest classification accuracy with the chosen attributes. Typically, the number of features is reduced compared to the original dataset. The fitness function defined in Equation 9 is employed to maximize the accuracy performance of the ML model and leverages GWO for assessing the selected features.

\[
\text{Fitness} (t) = \alpha \rho + \beta \frac{N - L}{L}
\]  

(9)

Where \( \rho \) represents the classification accuracy of the Random Forest Model. \( L \) denotes the length of the chosen feature set, \( N \) stands for the total count of features within the P2P lending dataset. Meanwhile, \( \alpha \) and \( \beta \) are parameters associated with the weight of classification accuracy and the feature selection quality. Here, \( \alpha \) ranges within \([0, 1]\) and \( \beta = 1 - \alpha \). 't' signifies the iteration.

Figure 6 visually demonstrates a subset of features that correspond to potential solutions. In this context, we adopt the binary chromosome model for configuring feature subsets. 'd' denotes the total count of features, equivalent to the chromosome's length. The values '1' or '0' are assigned based on the chromosome's position. If the value of the \( i^{th} \) bit is 1, the corresponding feature is included; otherwise, if the \( i^{th} \) bit is 0, the feature is excluded.

2.3.1. Performance Evaluation

The assessment of the proposed approach’s effectiveness involves employing a confusion matrix for a binary classification scenario. This matrix encompasses metrics like true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP). Utilizing the confusion matrix facilitates the computation of accuracy, recall, precision, and F1-score, which are defined as:

\[
\text{Accuracy} = \frac{TP + TN}{TN + FP + TP + FN}
\]  

(10)

\[
\text{Recall} = \frac{TP}{TP + FN}
\]  

(11)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  

(12)

\[
F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

(13)
Figure 5. The workflow of GWO-RF

Figure 6. Solution representation of feature selection

3. RESULTS AND DISCUSSION

The outcome of the feature selection process employing GWO along with 3-ML models using the P2P lending dataset as input yields a compilation of the most influential attributes. The configuration of parameter values encompasses elements such as the iteration count, the quantity of wolves, the number of features or dimensions, the scope of the search domain, as well as alpha and beta parameters for the fitness function. These specifics are elaborated in Table 1. Executing the proposed GWO-RF model follows the outlined procedure in Figure 5, producing the fitness function output computed according to Equation (9).

As depicted in Figure 7, the fitness value consistently rises with each iteration. GWO exhibits the capability to yield the optimal solution or the most fitting attributes for the GWO-RF model classification procedure. The count of selected features (L), representing the optimal features based on the GWO-RF model, is visualized in Table 1. This particular feature wields significant influence over default prediction within P2P lending.

Table 1. The setting of parameter for the proposed method.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of iteration</td>
<td>10</td>
</tr>
<tr>
<td>Number of wolves</td>
<td>95</td>
</tr>
<tr>
<td>Search Domain</td>
<td>[0 1]</td>
</tr>
<tr>
<td>Number of Dimensions</td>
<td>41</td>
</tr>
</tbody>
</table>
Figure 7. The fitness value of the proposed model.

Table 2. The selected features in P2P Lending.

<table>
<thead>
<tr>
<th>Models</th>
<th>Order of features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
</tr>
<tr>
<td>RF</td>
<td>21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40</td>
</tr>
<tr>
<td>RF</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

From Table 2, it is known that the features selected by the Grey Wolf Optimizer algorithm are Loan amount, term, last_pymnt_amnt, total_rec_int, and recoveries. Out of a total of 41 existing features, the most influential ones, after going through the GWO algorithm processing, are narrowed down to just 5 features. These 5 features are then processed using the Random Forest algorithm for classification.

Table 3. Performance evaluation comparison between GWO-RF and original RF.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>F1-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>GWO</td>
<td>Original</td>
<td>GWO</td>
</tr>
<tr>
<td>RF</td>
<td>96.53</td>
<td>97.31</td>
<td>76.83</td>
<td>85.39</td>
</tr>
</tbody>
</table>

Based on the results from Table 3, a performance comparison between GWO-RF and the original RF is conducted. GWO-RF outperforms the Original RF based on performance evaluations for Accuracy, recall, and F1-score. Feature selection using GWO effectively improves predictive performance against the Random Forest model. For Accuracy, it increases by 0.78%, Recall by 8.56%, and F1-score by 3.6%, while Precision experiences a decrease of 3.76%.

Furthermore, Table 4 presents a comparison of the proposed model’s accuracy performance with previous related studies. This table confirms that the proposed model outperforms the research conducted by Nguyen et al. [30], which used the Restricted Boltzmann Machine (RBM) feature selection algorithm in conjunction with six Machine Learning models: LDA, LR, ANN, KNN, SVM, RF, with the highest score being 81.20% for the RBM+LDA mode. Subsequently, Setiawan, Suharjito, and Diana[25] employed the Hybrid Binary PSO+ERT in predicting P2P Lending defaults, resulting in a score of 64%. Finally, Victor and Raheem [31] used GA as a feature selection method along with three Machine Learning models: LR, RF, and SVM,
with the highest score achieved by GA+RF at 92% accuracy. In contrast, the model proposed in this study, GWO+RF, obtained a score of 97.31%, which represents the highest accuracy score in predicting P2P Lending defaults. Therefore, GWO has proven to be a suitable feature selection optimization algorithm. However, there is a need to expand it through an extended search space to accommodate high-dimensional datasets.

Table 4. Comparison between the model proposed in this study and previous related research.

<table>
<thead>
<tr>
<th>Study</th>
<th>Feature selection</th>
<th>Models</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nguyen et al. [30]</td>
<td>Restricted Boltzmann Machine</td>
<td>LDA</td>
<td>81.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>81.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>66.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KNN</td>
<td>72.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVM</td>
<td>76.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RF</td>
<td>67.72</td>
</tr>
<tr>
<td>Setiawan, Suharjito and Diana [25]</td>
<td>Binary Particle Swarm Optimisation</td>
<td>ERT</td>
<td>64</td>
</tr>
<tr>
<td>Victor and Raheem [31]</td>
<td>Genetic Algorithm</td>
<td>LR</td>
<td>86</td>
</tr>
<tr>
<td>Proposed model</td>
<td>GWO</td>
<td>RF</td>
<td>97.31</td>
</tr>
</tbody>
</table>

4. CONCLUSION

The conclusion that can be drawn from the evaluation of the Grey Wolf Optimizer-Random Forest method is the feature selection approach proposed in this research. The GWO-RF model can select relevant features and disregard irrelevant ones within the P2P Lending dataset. Comparative studies of three performance evaluations (accuracy, recall, and F1-score) indicate that the GWO-RF model outperforms the original Random Forest method in predicting defaults in P2P Lending. The proposed method is also superior to three previous related studies based on accuracy. Furthermore, there is a need to enhance GWO by expanding the search space to handle high-dimensional datasets.

REFERENCES


