# Credit Risk Prediction Model Using Feature Engineering and Machine Learning Techniques

**Chonlada Muangthanang[1], Surasak Mungsing[2], and Nivet Chirawichitchai[3*]**

[1, 2]School of Information Technology, Sripatum University, Bangkok, Thailand
[3]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: chonladamuangthanang@gmail.com, surasak.mu@spu.ac.th, nivetchi@pim.ac.th

Received: June 19, 2023/ Revised: July 18, 2023/ Accepted: July 19, 2023

*Abstract*— **Credit scoring is a crucial step in the risk management process of the financial industry and commercial banks. The objective of this research is the development of a credit risk prediction model using feature engineering and machine learning techniques. This research was used to test the algorithm with a peer-to-peer (P2P) lending dataset and measure performance with classification accuracy. The experiment in this research found the XGB algorithm provided the most effective classification accuracy of 88.94%, which is better than other classifiers. Therefore, the proposed research framework of this research, working with feature engineering, feature selection, and machine learning techniques, is suitable and effective for credit scoring problem analysis.**

*Index Term*—**Credit Scoring, Feature Engineering, Machine Learning**

## I. Introduction

Peer-to-peer (P2P) lending is a novel financial system that leverages an internet-based platform to enable direct lending between borrowers and lenders, bypassing traditional financial intermediaries such as banks. The P2P network lending sector has experienced substantial expansion in recent times, primarily attributed to the progress made in big data and Internet finance. The proliferation of peer-to-peer lending platforms on the Internet is indicative of the industry's advancement. Peer-to-peer (P2P) lending is an online financial service that facilitates direct connections between individual investors and loan borrowers, thereby bypassing the involvement of commercial banks as intermediaries. Small and medium-sized businesses, as well as individuals in need of loans, now have this type of lending as a significant option. Lending Club, the world's largest online financial platform for borrowers and investors, has processed loans for over 3 million borrowers and attracted investments totaling more than $50 billion. By utilizing the power of the Internet, Lending Club has developed a marketplace that offers cheaper costs and higher investment returns than traditional commercial banks. The innovative methodology employed has facilitated the attainment of accessibility and simplicity in the borrowing and investing processes for all individuals. Peer-to-peer lending is a suitable match for the present economic progress of the nation, presenting noteworthy prospects. Nonetheless, it poses specific obstacles and potential hazards. The primary financial risks associated with peer-to-peer (P2P) lending are attributed to inadequate liquidity of funds, credit risks that arise due to information asymmetry, operational risks, and legal risks that result from incomplete laws and regulations in the domain of Internet finance. In contrast to conventional finance, online financial risks exhibit more intricate features. The virtual and technical aspects of Internet technology give rise to supplementary risks that surpass those encountered in traditional finance. Internet-based financial risks have a tendency to manifest abruptly and spread rapidly, with a higher potential for causing significant harm while being less manageable. As a result, the concept of risk aversion has emerged as a significant and crucial subject of discourse among investors, policymakers, scholars, and financial professionals. The presence of these risks significantly increases the probability of borrowers defaulting, thereby exposing P2P lending to credit risks that may arise from such defaults. The expansion of P2P lending platforms as well as investor profitability are both negatively impacted by loan defaults. As a result, loan evaluation has been extensively researched by scholars from both domestic and international backgrounds. The present assessment constitutes a valuable instrument for peer-to-peer (P2P) platforms to evaluate and manage credit risks. Peer-to-peer lending platforms commonly employ a credit scorecard as a basis for constructing their loan assessment framework, which is customized to meet their particular business needs. The utilization of a credit scorecard has the potential to expeditiously allocate a credit score to individual loans.

However, its efficacy in accurately distinguishing between borrowers who are prone to default and those who are not is limited [1]-[4].

As big data technology has grown and matured, risk management systems in the financial sector increasingly rely on machine learning and artificial neural networks. The utilization of Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Deep Neural Networks (DNNs) is prevalent in the prediction of stock prices and their movements, demonstrating their efficacy in forecasting financial time series. Additionally, researchers have found that the Random Forest technique exhibits superior performance compared to alternative methods in the context of loan assessment for peer-to-peer lending [3]. Building on previous research, this study uses the Random Forest algorithm to create a loan default prediction model utilizing data from Lending Club loans in the first quarter of 2019. Furthermore, four distinct methodologies are utilized and contrasted with Random Forest in subsequent evaluations. The results of our study hold great importance, as they aid in improving the loan evaluation process and promoting the sustainable growth of peer-to-peer lending [4]-[6]. The next sections of this document are categorized into five distinct parts. Section 2 provides a concise overview of the existing literature pertaining to loan evaluation and credit risk assessment. Section 3 goes through the specifics of the Lending Club and machine learning. Section 4 then discusses the experiments and their outcomes. In conclusion, Section 5 provides a summary of the preceding sections.

## II. Literature Review

Presently, the primary emphasis of research on P2P platforms, both domestically and internationally, centers on loan defaults and the evaluation of credit risk, utilizing machine learning techniques. Serrano-Cinca and colleagues investigated default variables using Lending Club loan sample data, a single-factor mean test, and survival analysis [7]. Yao et al. utilized advanced-support vector regression (SVR) methodologies to forecast default loss for corporate bonds, exhibiting superior performance of SVR variants in comparison to alternative techniques [8]. Malekipirbazari and Aksakalli have proposed a method of classification based on Random Forest to identify P2P borrowing customers who possess high-quality attributes. Upon comparison with several machine learning techniques, the findings revealed that the Random Forest approach exhibited a significantly superior performance in distinguishing creditworthy borrowers when compared to FICO credit scores and LC grades. Emekter and colleagues used a logistic regression (LR) model to predict Lending Club borrowers' default likelihood, finding that credit grade, debt-to-income ratio, FICO score, and revolving line utilization were key [9]. Bagherpour used a huge dataset to forecast loan defaults using KNN, SVM, Random Forest, and Sand Factorization Machines (FM) algorithms [10]. The authors Byanjankar et al. [11] introduced a credit-scoring framework that employs artificial neural networks (ANNs) to categorise peer-to-peer (P2P) loans into two groups, namely default and non-default, and demonstrated its efficacy in identifying default applications. In their study, Cao et al. [12] conducted a comparative analysis of the efficacy of eight classification techniques, namely LDA, LR, DT, SVM, RF, GBDT, MLP, and XGBoost, using datasets sourced from Kaggle. Using accuracy, area under the curve of ROC, and logistic loss, the XGBoost model performed better. Kvamme et al. employed convolutional neural networks (CNNs) to forecast loan defaults, utilizing time series data pertaining to customer transactions across diverse accounts and cards. The study conducted by the researchers indicated that the CNN model exhibited superior performance compared to the Random Forest classifier [13]. In their study, Kim et al. introduced a novel approach that integrated label propagation, transduction support vector machine (TSVM), and Dempster-Shafer theory to effectively forecast defaults in social lending through the utilization of unlabeled data [14].

The field of credit risk assessment has witnessed the emergence of diverse methodologies and models. The trust spiral model was first proposed by Tang et al. and was utilized to investigate credit risk in the lending association between small businesses and banks [15]. Moradi and Mokhatab Rafiei devised an adaptive network-based fuzzy inference system by subjecting it to training with monthly data extracted from a customer profile dataset. Subsequently, a follow-up evaluation was carried out utilizing recently established variables and their corresponding regulations within a fuzzy inference mechanism. The outcome of this process led to the development of a month roster of unfavorable clients and a flexible framework for evaluating credit hazard [16]. Brown et al. conducted actual investigations and discovered that Random Forest and Gradient Boosting classifiers performed remarkably well in credit scoring, particularly when dealing with severe class imbalances within the datasets [17]. Li conducted a qualitative analysis to determine the likelihood of loan defaults among lending club borrowers. Loan purpose, income, residence address, and work seniority were considered in this analysis. A logistic regression model was then used to construct credit scores and forecast borrower default [18]. Zhang et al. used Multiple Instance Learning (MIL) to create a novel credit scoring model that included sociodemographic and loan application information, as well as the applicants' transaction history [19]. In order to estimate credit card survival models, Djeundje and Crook employed Generalised Additive Models (GAMs) with cubic B-splines, showing that GAMs performed better than other techniques in terms of increasing prediction accuracy [20]. Chen et al. Predicting default risk on peer-to-peer lending imbalanced datasets. The objective of this research is to utilize not only several machine learning schemes for predicting the default risk of P2P lending but also re-sampling and cost-sensitive mechanisms to process imbalanced datasets [21]. Masmoudi and colleagues utilized a discrete Bayesian network that incorporated a latent variable to construct a model for loan subscribers exhibiting default payment behavior. The objective of this model is to assess the credit risk associated with loan subscribers and group them into clusters [22]. Papouskova and Hajek created a two-stage consumer credit risk model utilizing heterogeneous ensemble learning. This model predicted credit scoring and default exposure using class-imbalanced ensemble learning and regression ensemble approaches [23]. Ma et al. and Coser et al. used LightGBM, XGBoost, Logistic Regression, and Random Forest to create a set of prediction models for estimating the likelihood of loan default among clients [24]-[25]. Finally, Cho et al. suggested an instance-based entropy fuzzy support vector machine categorization investment choice model for the P2P lending market [26]. Although there has been a large amount of research focused on forecasting loan defaults in the Lending Club, our study attempts to add to the existing body of work that employs the Random Forest approach [27].

## III. Research Framework

The data were subjected to data cleaning and feature engineering in order to facilitate the extraction of characteristics and model training. The research framework comprises seven distinct steps, including data cleaning and elimination of features redundant, feature engineering, handling of missing data and scaling, oversampling, feature selection, splitting a dataset into training and testing, and machine learning techniques shown in Fig. 1.

### *A. Dataset*

The dataset used for this study comprises publicly available peer-to-peer lending data sourced from the lending club. This dataset includes all the data gathered by the platform during the lending procedure. The primary components comprise the personal particulars of the borrower, the intended use of the loan, the individual's credit background, their current debt status, and additional relevant factors. Therefore, we used the loan dataset period from January 2007 to December 2016, corresponding to a total of 396,031 loans with 151 features each. We have used loan status as the reference label for default, where fully paid means the applicant has fully paid the loan (the principal and the interest rate), and Charged-off means the applicant has not paid the installments in due time for a long period of time and has defaulted on the loan.

### *B. Data Cleaning*

The P2P lending datasets generally have many features, many of which are empty for most records, to help extract characteristics and train algorithms. We cleaned the data using feature engineering. This was a five-step process that comprised deleting superfluous features, converting features, dealing with missing data and scaling, and performing feature selection. First, we deleted irrelevant details like the borrower's lending club membership ID. We removed non-analyzable descriptive elements like loan purpose paragraphs. Furthermore, we eliminated characteristics that were excessively unvarying, exhibiting a homogeneity of over 99% in the data, for instance, application classifications that were primarily composed of personal loans. In addition, features that were acquired after loan approval, such as the repayment date of the previous loan, were excluded. Credit features identified by lending club and those with an excessive number of missing values, where more than 99 percent of the data was lacking, were also eliminated.

### *C. Feature Engineering*

Because the majority of the data consisted of categorical variables, which are unsuitable for model training, the data had to be converted into numerical representations. The initial reference label is denoted as the default loan status. The category labeled charged off was assigned a value of 0, while the category labeled fully paid was assigned a value of 1. The variable employee length denotes the duration of an individual's employment in terms of years. We converted this sequential property into ordered integers using ordinal encoding. We awarded a numerical value of 10 to those who had worked for more than ten years, a value of 0 to those who had worked for less than one year, and the corresponding numerical values to those who had worked for one to ten years. The third variable pertains to the rate of revolving credit utilization, expressed as a percentage. The decimal form was obtained through conversion. It appears to be a historical time stamp for the earliest credit history. Utilize an apply function to extract the year from the given feature, followed by converting it into a numeric feature. Regarding the remaining category features, such as loan purpose and housing ownership, which lack a sequential relationship, we employed one-hot encoding to transform them. This entailed creating independent binary features for each category, with a binary value of either 0 or 1.

### *D. Feature Scaling and Handling Missing Values*

Because there were some missing values in the dataset, it was important to address this issue before proceeding with the model training. Given that the "N/A" value signifies the absence of a default record in the past, it is imperative that this information is not disregarded. The logical way to fill in the value was to set a second feature that shows if the missing value in the first feature is present; the missing value in the first feature was filled in with values that don't typically occur. Assuming the data were normal-distributed, we filled in the average of other attributes without missing values. Moreover, in the context of employing machine learning algorithms that use the mean square error as the loss function, it is noteworthy that the magnitude of the features can significantly impact the predictive efficacy of the model. This is due to the model's propensity to exhibit sensitivity towards features that possess substantial scales. As a result, before training the model, we standardized the data to ensure that each feature only had a proportional impact on the prediction outcome. Feature scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that vary in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling. In this research, we use MinMaxScaler for feature scaling.

### *E. Feature Selection*

In the past, reducing the dimensionality of the data involved using an extraction method for features. The new feature is a projection of the previous one when using this type of procedure, such as principal component analysis. However, a feature extraction method removes the original features and may not have empirical meanings, making it unsuitable for business applications [26]. As a result, we must do feature selection, giving precedence to features that are highly related to the aim and deleting irrelevant features to lower the complexity of learning. In this research, we used Pearson's correlation to analyze the significance of the features. Pearson's correlation coefficient is the test statistic that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. We used the Feature Elimination approach to identify features that had the strongest association with the target variable and then removed them one by one to achieve the initial dimensionality reduction, with the independent variable decreasing from 151 to 26, as shown in Fig. 1to Fig. 2. Here is the information on this particular data set, shown in Table I.

### *F. Oversampling*

The task of imbalanced classification pertains to the construction of predictive models for classification datasets that exhibit a significant disparity in the number of instances between the classes. The issue of working with imbalanced datasets is that most machine learning algorithms will disregard, and so perform poorly on, the minority class, despite the fact that performance on the minority class is generally the most essential. Oversample the minority class to correct skewed datasets. The simplest method is to duplicate minority class examples, which don't provide any new information to the model. Alternatively, novel instances can be generated through the amalgamation of pre-existing exemplars. The Synthetic Minority Oversampling Technique (SMOTE) is a form of data augmentation that is used to address imbalanced class distribution, specifically for the minority class [21].

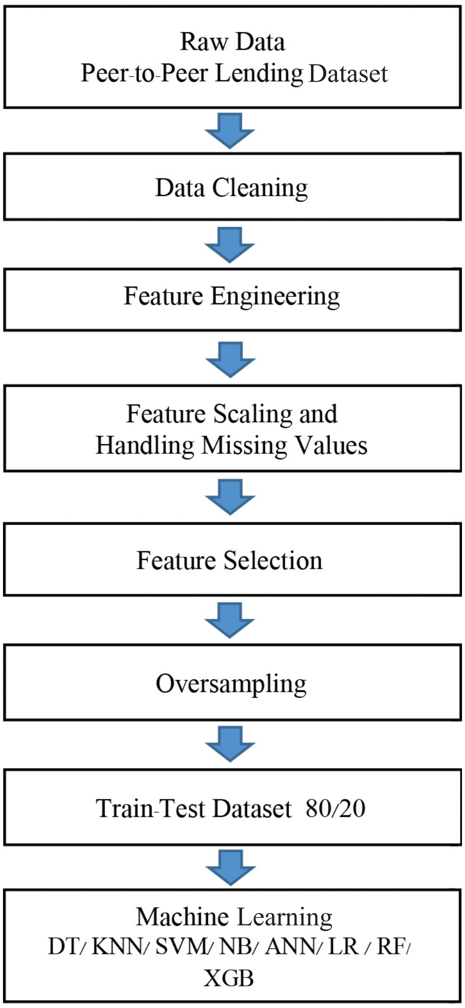

Fig. 1. Research framework

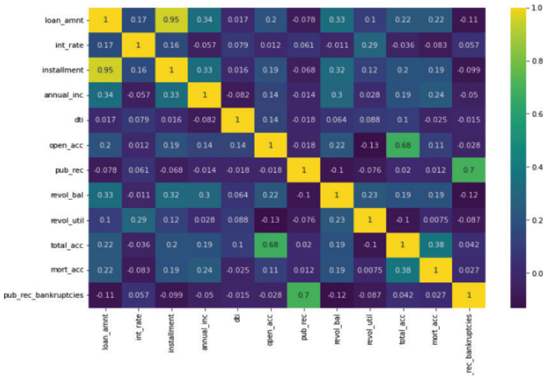

Fig. 2. Person correlation of features

TABLE I
RESULTS OF FEATURE SELECTION

| Loan Stat New | Description |
|---|---|
| Loan Amnt | The loan amount requested by the borrower is indicated in the listed amount. If the credit department decides to decrease the loan amount at any given time, this change will be reflected in this value. |
| Term | The loan duration is denoted by the number of payments, which is measured in months and can be either 36 or 60. |
| Int Rate | Interest Rate on the loan |
| Installment | The borrower is responsible for a monthly payment that is due when the loan is initiated. |
| Grade | Loan grade |
| Sub_Grade | Loan subgrade |
| emp_title | The occupation or job position provided by the borrower during the loan application process. |
| Emp_Length | The duration of employment is stated in years. The available options range from 0 to 10, with 0 representing less than one year and 10 indicating ten or more years of employment. |
| Home_Ownership | The borrower's residential property ownership status is indicated during registration or obtained from the credit report. |
| Annual_Inc | The annual income was disclosed by the borrower during the registration process. |
| Verification_Status | It indicates whether the borrower's income was verified by LC, not verified, or if the source of income was verified. |
| Issue_D | The month in which the loan was financed. |
| Loan_Status | Current status of the loan |
| Purpose | A classification is provided by the borrower regarding their loan request. |
| Title | The description or title of the loan provided by the borrower. |
| Zip_Code | The initial three digits of the zip code are given by the borrower in the loan application. |
| Addr_State | The state is mentioned by the borrower in the loan application. |
| Dti | The ratio is derived by dividing the borrower's total monthly debt payments, excluding mortgage and the requested LC loan, by the borrower's self-reported monthly income. |
| Earliest_Cr_Line | The month when the borrower initially established their reported credit line. |
| Open_Acc | The count of active credit lines in the borrower's credit file. |
| Pub_Rec | The count of negative public records associated with the borrower. |
| Revol_Bal | It refers to the overall balance of revolving credit accounts. |
| Revol_Util | The percentage of revolving credit utilized by the borrower, indicates the amount of credit they are currently using compared to the total available revolving credit. |
| Total_Acc | The overall count of credit lines currently present in the borrower's credit file. |
| Initial_List_Status | The initial status is assigned to the loan listing. It can have one of two values:<br>W or F. |
| Application_Type | It indicates whether the loan application is made by an individual or involves a joint application with two co-borrowers. |
| Mort_Acc | It represents the count of mortgage accounts. |
| Pub_Rec_Bankruptcies | It represents the count of bankruptcies listed in public records. |

## IV. Experiments and Results

We performed experiments using a collection of peer-to-peer lending datasets obtained from the lending club website. We implemented all algorithms in Python using the scikit-learn library. To evaluate the performance of our network, we calculate four metrics: accuracy, precision, recall, and F-measure, TP = the number of true positives, FP = the number of false positives, TN = the number of true negatives, FN = the number of false negatives, P = the number of positives in ground truth, N = the number of negatives in ground truth [28]-[30]. Classification effectiveness is usually measured by accuracy, precision, and recall. Precision is the proportion of true positive examples labeled positive by the system that was truly positive and recall is the proportion of true positive examples that were labeled positive by the system. The F-measure function which combines precision and recall is computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F-measure = \frac{2\times Precision\times Recall}{Precision+Recall}$$

We tested all algorithms using the validation test set of 20%. The results in terms of accuracy, precision, recall, and F-measure is the averaged values calculated across all cross-validation experiments reported in Table II to Table III and Fig. 3 to Fig. 4.

TABLE II
TRADITIONAL FRAMEWORK PERFORMANCE COMPARISONS %

| List | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|
| ANN | 80.60 | 76.5 | 80.60 | 78.49 |
| XGB | 80.60 | 76.24 | 80.60 | 78.36 |
| RF | 80.53 | 75.97 | 80.53 | 78.18 |
| DT | 70.88 | 71.76 | 70.88 | 71.32 |
| NB | 80.03 | 74.14 | 80.03 | 76.97 |
| KNN | 74.73 | 69.73 | 74.73 | 72.14 |
| SVM | 79.94 | 71.37 | 79.94 | 75.41 |
| LR | 80.31 | 64.49 | 80.31 | 71.54 |

TABLE III
PROPOSED FRAMEWORK PERFORMANCE COMPARISONS %

| List | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|
| ANN | 88.72 | 88.38 | 88.72 | 88.55 |
| XGB | **88.94** | 89.19 | **88.94** | 89.06 |
| RF | 88.89 | 89.52 | 88.89 | 89.20 |
| DT | 82.81 | 83.24 | 82.81 | 83.02 |
| NB | 86.41 | 85.68 | 86.41 | 86.04 |
| KNN | 85.16 | 84.00 | 85.16 | 84.58 |
| SVM | 88.85 | **90.20** | 88.85 | **89.52** |
| LR | 88.91 | 89.59 | 88.91 | 89.25 |

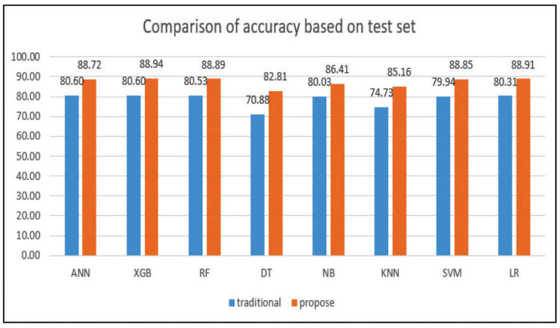

Fig. 3. Comparison of accuracy based on test set

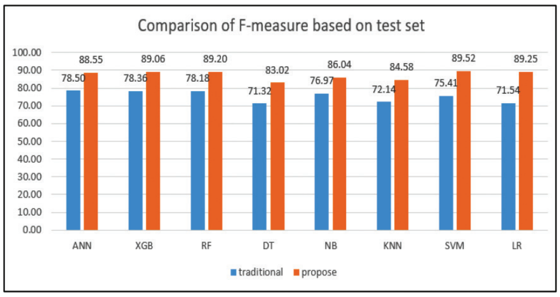

Fig. 4. Comparison of F-measure based on test set

The findings of this research showed that the proposed framework significantly outperformed the original framework in all algorithms presented in Table II to Table III and Fig. 3. The experimental results in this research found that when using the ANN algorithm, accuracy increased by 8.12%; using the XGB algorithm, accuracy increased by 8.34%; using the RF algorithm, accuracy increased by 8.36%; and using the DT algorithm, accuracy increased by 11.93%. Using the NB algorithm, accuracy increased by 6.38%; using the KNN algorithm, accuracy increased by 10.43%; using the SVM algorithm, accuracy increased by 8.91%; and using the LR algorithm, accuracy increased by 8.91%. As shown in Fig. 4, comparing F-Measure indicators, it was found that when using ANN, efficiency increased by 10.05%; using XGB, efficiency increased by 10.71%; using RF, efficiency increased by 11.02%; using DT, efficiency increased by 11.71%; using NB, efficiency increased by 9.07%; using KNN, efficiency increased by 12.43%; using SVM, efficiency increased by 14.11%; and using LR, efficiency increased by 17.71%.

The experiment of this research framework in Table III found the XGB algorithm provided the most effective classification accuracy of 88.94%, followed by the LR algorithm with an accuracy of 88.91%, the RF algorithm with an accuracy of 88.89%, the SVM algorithm with an accuracy of 88.85%, the ANN algorithm with an accuracy of 88.72%, the NB algorithm with an accuracy of 86.41, the KNN algorithm with an accuracy of 85.16%, and the DT algorithm with an accuracy of 82.81%, respectively.

The results of the experiments in this study are consistent with the findings of Cao [12] and Ma et al. [24] investigation into the performance evaluation of machine learning approaches for credit scoring. The results of this research indicate that the Boosting classifier has better performance in predictive analytics compared with the other classifier. The experimental findings from this study are also consistent with earlier research by Brown [17] and Chen [21] which involved experimental evaluations of classification algorithms for unbalanced credit-scoring datasets. The model evaluation yields consistent outcomes with high accuracy and optimal performance.

## IV. Conclusion

This study tried to combine the benefits of both feature selection and feature engineering to improve the performance of the credit scoring model. The contribution of this research is to provide a framework for the development of credit scoring models using feature engineering and machine learning techniques consisting of artificial neural networks (ANN), XGBoost (XGB), random forest (RF), logistic regression (LR), and support vector machines (SVM), naive Bayes (NB), k-nearest neighbor (KNN), and decision tree (DT).

The proposed framework is tested on P2P lending datasets and measures performance with accuracy. This experiment in this research found the XGB algorithm provided the most effective classification accuracy of 88.94%. Therefore, the proposed research framework of this research, working with feature engineering, feature selection, and machine learning techniques, is suitable and effective for credit scoring problem analysis.

## References

[1] H. Wang, K. Chen, W. Zhu et al., “A Process Model on P2P Lending,” *Financial Innovation*, vol. 1, no. 3, pp. 2-9, Jun. 2015.

[2] M. L. Challa, V. Malepati, and S. N.R. Kolusu, “Forecasting Risk Using Auto Regressive Integrated Moving Average Approach,” *Financial Innovation*, vol. 4, no. 24, pp. 1-17, Oct. 2018.

[3] M. Malekipirbazari and V. Aksakalli, “Risk Assessment in Social Lending Via Random Forests,” *Expert Systems with Applications*, vol. 42, no. 10, pp. 4621-4631, Jun. 2015.

[4] D. Selvamuthu, V. Kumar, and A. Mishra, “Indian Stock Market Prediction Using Artificial Neural Networks on Tick Data,” *Financial Innovation*, vol. 5, no. 16, pp. 1-12, Mar. 2019.

[5] X. Zhong and D. Enke, “Predicting the Daily Return Direction of the Stock Market Using Hybrid Machine Learning Algorithms,” *Financial Innovation*, vol. 5, no. 16, pp. 1-20, Dec. 2019.

[6] X. Ye, L. Dong, and D. Ma, “Loan Evaluation in P2P Lending Based on Random Forest Optimized by Genetic Algorithm with Profit Score,” *Electronic Commerce Research and Applications*, vol. 32, pp. 23-36, Nov. 2018.

[7] 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, Oct. 2015.

[8] X .Yao, J. Crook, and G. Andreeva, “Support Vector Regression for Loss Given Default Modelling,” *European Journal of Operational Research*, vol. 240, no. 2, pp. 528-538, Oct. 2015.

[9] R. Emekter, Y. Tu, B. Jirasakuldech et al., “ Evaluating Credit Risk and Loan Performance in Online Peer-To-Peer Lending,” *Applied Economics*, vol. 47, pp. 54-70, Oct. 2014.

[10] A. Bagherpour, *Predicting Mortgage Loan Default with Machine Learning Methods*. Berkeley, CA: University of California Riverside, 2017, pp. 1-29.

[11] A. Byanjankar, M. Heikkilä, and J. Mezei, “Predicting Credit Risk in Peer-To-Peer Lending: A Neural Network Approach,” in *Proc. IEEE Symposium Series on Computational Intelligence*, 2015, pp. 719-725.

[12] A. Cao, H. He, Z. Chen et al., “Performance Evaluation of Machine Learning Approaches for Credit Scoring,” *International Journal of Economics Finance and Management Science*, vol. 6, no. 6, pp. 255-260, Dec. 2018.

[13] H. Kvamme, N. Sellereite, K. Aas et al., “Predicting Mortgage Default Using Convolutional Neural Networks,” *Expert Systems with Applications*, vol. 102, pp. 207-217, Jul. 2018.

[14] A. Kim and S. B. Cho, “An Ensemble Semi-Supervised Learning Method for Predicting Defaults in Social Lending,” *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 193-199, May. 2019.

[15] Y. Tang, A. Moro, S. Sozzo et al., “Modelling Trust Evolution within Small Business Lending Relationships,” *Financial Innovation*, vol. 4, no. 19, pp.1-18, Sep. 2018.

[16] S. Moradi and F. M. Rafiei, “A Dynamic Credit Risk Assessment Model with Data Mining Techniques: Evidence from Iranian Banks,” *Financial Innovation*, vol. 5, no. 15, pp.1-27, Mar. 2019.

[17] I. Brown and C. Mues, “An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets,” *Expert Systems with Applications*, vol. 39, no. 3, pp. 3346-3453, Feb. 2012.

[18] Y. W. Li, “Research on Credit Score of P2P Online Lending,” *Advances in Social Science, Education and Humanities Research (ASSEHR)*, vol. 181, pp. 883-886. Sep. 2018.

[19] T. Zhang, W. Zhang, W. Xu et al., “Multiple Instance Learning for Credit Risk Assessment with Transaction Data,” *Knowledge-Based Systems*, vol. 161, no. 1, pp. 65-77, Dec. 2018.

[20] V. B. Djeundje and J. Crook, “Identifying Hidden Patterns in Credit Risk Survival Data Using Generalised Additive Models,” *European Journal of Operational Research*, vol. 277, no. 1, pp. 366-376, Aug. 2019.

[21] Y. Chen, J. Leu, S. Huang et al., “Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets,” *IEEE Access*, vol. 9, pp. 73103-73109, May. 2021.

[22] K. Masmoudi, L. Abid, and A. Masmoudi, “Credit Risk Modeling Using Bayesian Network with a Latent Variable,” *Expert Systems with Applications*, vol. 127, pp. 157-166, Aug. 2019.

[23] M. Papouskova and P. Hajek, “Two-Stage Consumer Credit Risk Modelling Using Heterogeneous Ensemble learning,” *Decision Support Systems*, vol. 118, pp. 33-45, Mar. 2019.

[24] X. Ma, J. Sha, D. Wang et al., “Study on a Prediction of P2p Network Loan Default Based on the Machine Learning Lightgbm and Xgboost Algorithms According to Different High Dimensional Data Cleaning,” *Electronic Commerce Research and Applications*, vol. 31, pp. 24-39, Sep. 2018.

[25] A. Coser, M. Maer-matei, and C. Albu, “Predictive Models for Loan Default Risk Assessment,” *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 149-165, Apr. 2019.

[26] P. Cho, W. Chang, and J. W. Song, “Application of Instance-Based Entropy Fuzzy Support Vector Machine in Peer-to-Peer Lending Investment Decision,” *IEEE Access*, vol. 7, pp. 16925-16939, Feb. 2019.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[28] N. Chirawichitchai, "Developing Term Weighting Scheme Based on Term Occurrence Ratio for Sentiment Analysis," *Information Science and Applications Lecture Notes in Electrical Engineering*, vol. 339, pp. 737-744, Jan. 2015.

[29] N. Chirawichitchai, "Emotion Classification of Thai Text Based Using Term Weighting and Machine Learning Techniques," in *Proc. The 11th The International Joint Conference on Computer Science and Software Engineering*, 2014, pp. 91-96.

[30] N. Chirawichitchai, "Sentiment Classification by a Hybrid Method of Greedy Search and Multinomial Naïve Bayes Algorithm," in *Proc. International Conference on ICT and Knowledge Engineering ICT & Knowledge Engineering*, 2013, pp. 1-4.

**Chonlada Muangthanang** received M. Sc. in Information Technology, Naresuan University, Thailand. She currently works as a lecturer at the Faculty of Management Sciences, Phetchabun Rajabhat University.

**Surasak Mungsing** received his B.S. in Engineering in Structures Materials and Fluids from the University of South Florida, his M.Sc. in Computer Science, from the Naval Postgraduate School, U. S. A and a D. Eng degree in Computer Science from the Asian Institute of Technology. He currently works as a graduate lecturer at the Faculty of Information Technology, Sripatum University.

**Nivet Chirawichitchai** received his B.B.A. in Industrial Management from Ramkhamhaeng University, M. Sc. in Information Technology, and his Ph. D. in Information Technology from King Mongkut's Institute of Technology North Bangkok, Thailand. He currently has the rank of director, Master of Engineering Program in Engineering and Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management.