2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) | 978-1-6654-0825-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/TAAI54685.2021.00064

Predicting Credit Risk in Peer-to-Peer Lending: A Machine Learning Approach with Few Features

Yu-Chieh Cheng
Department of Computer Science and Engineering
Yuan Ze University, Taiwan
s1061439@mail.yzu.edu.tw

Hui-Ting Chang
Department of Computer Science and Engineering
Yuan Ze University, Taiwan
s1061460@mail.yzu.edu.tw

Chia-Yu Lin
Department of Computer Science and Engineering
Yuan Ze University, Taiwan
sallylin0121@saturn.yzu.edu.tw

Heng-Yu Chang
College of Management
School of Business
Chang Gung University, Taiwan
hychang@mail.cgu.edu.tw

***Abstract*—Peer-to-peer (P2P) lending provides borrowers with relatively low borrowing interest rates and gives lenders a channel for investment on an online platform. Since most P2P lending does not require any guarantees, the overdue payment of borrowers results in a massive loss of lending platforms and lenders. Many risk prediction models are proposed to predict credit risk. However, these works build models with more than 50 features, which causes a lot of computation time. Besides, in most P2P lending datasets, the number of non-default data far exceeds the number of default data. These researches ignore the data imbalance issue, leading to inaccurate predictions. Therefore, this study proposes a credit risk prediction system (CRPS) for P2P lending to solve data imbalance issues and only require few features to build the models. We implement a data preprocessing module, a feature selection module, a data synthesis module, and five risk prediction models in CRPS. In experiments, we evaluate CRPS based on the de-identified personal loan dataset of the LendingClub platform. The accuracy of the CRPS can achieve 99%, the recall reaches 0.95, and the F1-Score is 0.97. CRPS can accurately predict credit risk with less than 10 features and tackle data imbalance issues.**

***Index Terms*—Peer-to-peer lending, credit risk prediction, RFECV, data synthesis, borderline-SMOTE, XGBoost**

I. Introduction

Peer-to-peer (P2P) lending has become a growing multi-billion dollar industry that brings borrowers and lenders together on a technology platform and enables both parties to meet their financial needs. Due to the lower handling fees and high execution efficiency, more and more borrowers join the platform to apply P2P lending. However, the overdue repayment of the borrower will cause enormous costs for the lending platform or lenders.

Many studies take the information of borrowers to predict whether the loan transaction will be overdue to reduce the credit risk of P2P lending. Previous studies focused on multivariate statistical analysis. [1] applied the statistical method Cox proportional hazards model to analyze the credit risk in P2P lending. The purpose of the model is applied in predicting survival probabilities at different time periods. [2] selected 58 features out of 100 features by least absolute shrinkage and selection operator (Lasso) based on regression analysis [3]. [4] used datasets of LendingClub platform and adopted Restricted Boltzmann Machine (RBM) for feature selection. After selecting the features, the best subset includes 80 out of 123 features. However, 80 and 58 features are still too many to represent the key factors of the model, and it takes much time to analyze. Finding critical features for P2P lending is the key in the P2P credit risk prediction.

Adopting machine learning models to predict credit risk becomes a trend. [5] build logistic regression (LR), regression tree (RT), bagging neural network (BNN), random forest (RF), gradient boosting decision tree (GBDT), categorical boosting (CatBoost) and, eXtreme gradient boosting (XGBoost) models based on the LendingClub dataset to predict default loans. [6] also adopt data from LendingClub to construct a decision tree, RF, bagging, and extra tree to train the prediction model of peer-to-peer lending data. However, in most P2P lending datasets, the number of non-default data is much more than that of default data. The current research ignore the data imbalance issue, leading to inaccurate predictions.

In order to solve the data imbalance issue, under-sampling and over-sampling are two well-known solutions. Under-sampling uses random under-sampling to delete data and make the dataset more balanced [7]. However, under-sampling may remove some potentially useful majority class samples and lose some critical information, resulting in inaccurately prediction [8], [9]. On the other hand, oversampling solves the data imbalance by changing the distribution of the training data. The fundamental way is to copy the minority samples randomly. The synthetic minority over-sampling technique (SMOTE) [10] is an improved method based on random oversampling, which uses k-nearest neighbor (KNN) to sample the nearest samples, generates new minority samples, and adds them to the dataset. However, samples on the boundary are easily misclassified and make inaccurate oversampling.

This paper proposes a credit risk prediction system (CRPS) for P2P lending, which can synthesis data and only require few features to achieve accurate prediction results. CRPS contains

2376-6824/21/$31.00 ©2021 IEEE
DOI 10.1109/TAAI54685.2021.00064

a data preprocessing module, a feature selection module, a data synthesis module, and five risk prediction models. The data preprocessing module fills in the missing values. Feature selection module adopts recursive feature elimination with cross-validated (RFECV) [11], sequential forward selection (SFS) [12], and Lasso [3] to select the best feature sets to save the analyze time and improve model accuracy. In the data synthesis module, we synthesize new samples by borderline synthetic minority oversampling technique (borderline-SMOTE) [13] to solve the data imbalance issue. Finally, we build deep neural network (DNN), LR, RF, CatBoost, and XGBoost as risk prediction models and compare the accuracy of different models to find the best solution. From the experiment results, the proposed CRPS only utilizes 10 features to achieve 99% accuracy, the recall reaches 0.95, and the F1-Score also reaches 0.97.

This contributions of this paper are as following:

- CRPS implements a data synthesis module to solve data imbalance issue and enhance the prediction accuracy.
- CRPS trains the model with less than ten features and predict default accurately.
- CRPS can help lending platforms assess the credit risk of borrowers and intensively decrease the loss.

II. Related Works

Financial information intermediary platforms provide services, such as information collection, information publication, credit evaluation, and loan matchmaking for both supply and demand parties to achieve lending. LendingClub [14] is the largest P2P online lending platform in the United States. This paper utilized the dataset from LendingClub to build the credit risk prediction system. Many researchers also adopted datasets collected by LendingClub to predict P2P risk. [2] analyzed the risk of online lending based on the dataset of LendingClub. Authors adopted Lasso to select 58 variables from 100 variables and developed an LR model to analyze the significance of the variables. [5] built LR, RT, BNN, RF, GBDT, CatBoost and XGBoost models to predict default loans based on the dataset of LendingClub. The CatBoost model got the highest accuracy rate of 79.59%. [6] trained the tree-based classifiers such as decision tree, RF, bagging, and extra tree to analyze the credit risk involved in the P2P lending system of LendingClub with the highest accuracy of 88.5%. These research built models without feature selection and made the prediction accuracy not that high.

Existing research adopted various machine learning models to predict credit risk. [15] utilized a support vector machine (SVM) to covert the features to a high-dimensional space to classify the credit risk. [16] adopted classification, regression tree (CART) and multivariate adaptive regression splines (MARS) to analyze the credit scoring for financial institutions. To investigate the suitability of different neural networks for credit scoring accuracy, [17] utilized five different neural models to benchmark their performance. Both mixture-of-experts and radial basis function neural network models were more suitable for credit scoring. [18] used the characteristics of the sigmoid transfer function to calculate the default probability of the borrower based on LR and classify the credit risk. [19] proposed a credit scoring model through neural network techniques and implemented the model by R language with the nnet, NeuralNetTools, caTools, and ROCR packages to predict credit risk.

These research focused on building machine learning models to determine whether the borrowers default. However, the cases of default and non-default are distributed unevenly in the data. The model with high accuracy may only be able to determine the non-default cases. Thus, dealing with the imbalanced data is an essential process before we train the models for credit risk prediction.

Under-sampling and over-sampling are popular techniques to solve the data imbalance issue. Under-sampling uses randomly delete data and makes the dataset more balanced [7]. Under-sampling may remove some potentially useful majority class samples and lose some critical information [8], [9]. On the other hand, oversampling solves the data imbalance by changing the distribution of the training data. The synthetic minority over-sampling technique (SMOTE) [10] is a typical improvement method based on random oversampling and adopted in many pieces of research. [20]–[22] utilized SMOTE to create data of the minority class to achieve more accurate P2P lending default prediction results. SMOTE takes two observations from the minority class ($(x1, y1)$ and $(x2, y2)$) and creates a random number between 0 and 1, which is called r. The synthetic point will be $(x1 + r * (x2 - x1)$, $y1 + r * (y2 - y1))$. However, the samples on the boundary are easily misclassified. A majority sample may generate if the collected minority samples are too close to the majority samples. If the collected minority samples are too far away from the majority samples, the generated samples may not contain valid information. Borderline-SMOTE [13] improved the shortcomings of SMOTE. It used the minority samples on the boundary to synthesize new samples, thereby improving the class distribution of the samples to construct a risk prediction system with the best accuracy. Therefore, we implement borderline-SMOTE to synthesize data in the proposed CRPS.

III. Credit Risk Prediction System

Fig. 1 is the architecture of the proposed credit risk prediction system (CRPS). Data preprocessing module deals with the missing value of data. The feature selection module chooses the critical features of P2P lending. The data synthesis module solves the data imbalance issue of data. Finally, we build an XGBoost model for risk prediction.

A. Data Preprocessing Module

This study analyzes one million de-identified personal loan data, which were originated from the LendingClub platform [14]. The dataset contains 149 features, including the number of payments on the loan, the total number of credit lines currently in the borrower's credit file, the homeownership status, etc. In the column of loan status, 0 means non-default, 1 means default.

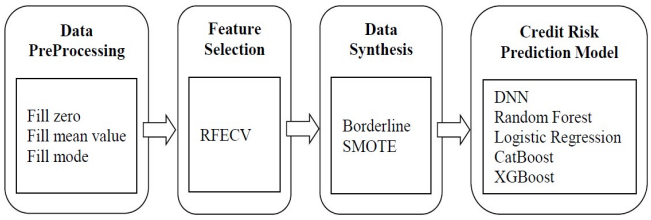

Fig. 1. The system flow of CRPS.

TABLE I
EXAMPLES OF DATA PREPROCESSING.

ID	*loan_amnt*	*home_ownership*	*sub_grade*
1	10,000	$MORTGAGE$	$A5$
2	15,000	Fill Mode	$B1$
3	Fill Mean	$RENT$	$C4$
4	12,000	$MORTGAGE$	Fill Zero

Since some features have a large number of null values, we delete columns with null values more than 80%. We fill the missing values of the remaining data by three methods.

- Fill zero: If the missing value does not correlate with other existing data, we can fill it with $zero$.
- Fill mean value: Some missing values cannot be zero, such as the loan amount which the borrower applies. We fill the mean value of the features in these fields.
- Fill mode: Some missing values cannot be zero, and the average cannot be calculated, such as the categorical features. Therefore, we fill the most frequently occurring category in these fields.

Table I is the example of filling in missing values. The value of $loan\_amnt$ is the loan amount applied by the borrower. It can not be zero or category, so we fill the mean value of the column. $Home\_ownership$ is a column with categorical values, which are $RENT$, OWN, and $MORTGAGE$. Sub_grade is the grade assigned by LendingClub. It is filled in a zero value because it is not easy to evaluate the correlation between this column and other columns.

B. Feature Selection Module

Selecting important features can increase the accuracy and decrease the computation time of the risk prediction process. There are many feature selection methods, such as SFS [12], Lasso [3], and RFECV [11]. SFS starts the training process from an empty feature set. It selects only one feature to add to the feature set based on the currently selected feature set. Each training step adds only one feature, and the selected features cannot be removed. Even if the algorithm finds a better combination without some selected features, the selected features cannot be deleted. As a result, SFS is hard to find the feature set with the highest accuracy.

Lasso adds the L1 penalty function in the objective function of linear regression and optimizes the objective function by reducing the absolute values of the coefficients. Lasso selects the useful features and makes the coefficient of useless features be 0 to discard features. However, Lasso can not be well defined when the number of predictors (p) is larger than the number of observations (n) due to the nature of the convex optimization problem [23].

Recursive feature elimination with cross-validation (RFECV) can retain high-impact features and remove redundant ones by eliminating weak ones in every loop. Cross-validation in RFECV finds the optimal number of features and avoids overfitting during training models. In this way, the training time of RFECV will be shortened a lot, and there is no need to worry about redundant features will be selected. We will show the comparison of SFS, Lasso, and RFECV in the experiments.

The experiments show that RFECV selects the fewest features for models and achieves the highest accuracy. Thus, we choose RFECV as the primary feature selection technique in the proposed CRPS. Based on the dataset of the LendingClub platform [14], RFECV selects ten critical features for the model, as shown in Table II. The top three important feature are $recoveries$, $Last\_range\_fico\_high$, and $loan\_amnt$. $recoveries$ is a payment received from the debt that was written off and considered uncollectible. It can be seen that $recoveries$ is closely related to the repayment behavior of the borrower. Thus, the most noteworthy feature is $recoveries$, which is accounting for 30.20% of importance. $Last\_range\_fico\_high$ is the debtor's recent Fair Isaac Corporation (FICO) score evaluation ceiling. Lenders use the FICO score to help make accurate, reliable, and fast credit risk decisions across the customer lifecycle. Thus, the influence is second-highest, accounting for 24.63% of importance. $loan\_amnt$ is the loan amount applied by the borrower, accounting for 10.73% of importance.

C. Data Synthesis Module

There is a data imbalance issue since most loans are in a non-default state. A lot of researchers adopted SMOTE [10], a typical improvement method based on random oversampling, to increase the number of minority categories. However, the samples on the boundary are easily misclassified in SMOTE. Increasing minority samples may increase into majority samples or increase the number of samples that are already easy to distinguish, resulting in inaccurate oversampling. Therefore, we use borderline-SMOTE [13] to generate default data, which only generates new samples for the boundary samples of the minority class.

There are three steps of borderline-SMOTE.

- Step1: First, borderline-SMOTE chooses k nearest neighbors of a randomly selected point p in the minority sample. Among these nearest neighbors, the numbers of minority sample points are x, and majority sample points are y.
- Step2: We must ensure that the random sampling point p is the minority sample surrounded by the majority. Thus, we set the upper and lower bound to compare with $x/(x+y)$. When the $x/(x+y)$ is smaller than the lower bound, p is directly regarded as noise, and borderline-SMOTE

TABLE II
TOP TEN IMPORTANT FEATURES.

Feature Name	*Description*	*Importance*
recoveries	Total recovery after creditors bad debts.	30.20%
last_fico_range_high	The debtor's recent FICO evaluation ceiling.	24.63%
loan_amnt	The loan amount applied by the borrower.	10.73%
last_pymnt_amnt	The total payment received recently by the creditor.	7.76%
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.	5.42%
debt_settlement_flag	Mark whether the debtor with bad debts is performing the liquidation process.	4.86%
out_prncp	The total outstanding principal of the debtor.	4.02%
total_rec_prncp	The principal received by creditors to date.	2.72%
hardship_flag	Flags whether or not the borrower is on a hardship plan.	1.32%
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	1.31%

TABLE III
THE PERCENTAGE OF DEFAULT DATA BEFORE/AFTER BORDERLINE-SMOTE.

Before borderline-SMOTE		*After borderline-SMOTE*	
Loan status	Percentage	Loan status	Percentage
0 (non-default)	86.17%	0 (non-default)	50%
1 (default)	13.83%	1 (default)	50%

does not operate on it. If the $x/(x+y)$ is bigger than the upper bound, p is considered as safe sample and does not need to be synthesized. For $x/(x+y)$ between the upper and lower bound, we add p and minority sample points x to the danger set.

- Step3: Finally, borderline-SMOTE synthesizes data based on the samples in danger set. As shown in Table III, after borderline-SMOTE, the amount of default data has increased to balance the amount of non-default data.

D. Credit Risk Prediction Model

Machine learning models can be adopted to predict the risk of P2P. According to [5], [17], [18], [22], we implement DNN, RF, LR, CatBoost, and XGBoost as credit risk prediction models. After the preliminary experiments, we find that XGBoost [24] achieves the highest accuracy.

XGBoost uses classification and regression trees (CART) to generate a linear classifier, which is an improvement and extension based on gradient boosted decision tree (GBDT). The operation of XGBoost [25] is to keep the original model $y_i^{(t-1)}$ unchanged in each iteration of the tree model and add a new function $f_t(x_i)$ to correct the loss of the previous tree and improve the overall model. As shown in Eq. (1), the prediction result at round t is $y_i^{(t)}$. The objective function is to minimize the difference between the predicted value and the actual value, as shown in Eq. (2). In Eq. (2), $2(y_i^{(t-1)} - y_i)$ is the residual of the previous round and $\Omega(f_t z)$ is the regularization in the boosting process. $\Omega(f_t z)$ shows in Eq. (3), where T is the number of leaves in the tree, γ, λ, and w_j^2 are used to regularization.

$$y_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = y_i^{(t-1)} + f_t(x_i) \tag{1}$$

$$Obj^{(t)} = \sum_{i=1}^{n} [2(y_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + const \tag{2}$$

$$\Omega(f_t) = \gamma(T) + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

Overall, XGBoost utilizes regularization to reduce overfitting and uses cross-validation to obtain the optimal number of boosting iterations. Thus, XGBoost is a highly efficient, flexible, and optimized distributed gradient boosting model.

IV. EXPERIMENTS

In this section, we will evaluate the accuracy of the proposed CRPS and discuss the performance of different modules. We build the CRPS based on a dataset from the LendingClub platform [14]. The dataset contains one million de-identified personal loans and 149 features, including the number of payments on the loan, the total number of credit lines currently in the borrower's credit file, the homeownership status, and so on. In the data, the number of fully paid loans is $936,523$, whereas the number of default loans is $146,852$. We used 80% of the data as the training set and 20% as the test set.

A. Evaluation Metrics

We adopt accuracy (4), recall (5), precision (6), and F1-Score (7) as the evaluation metrics. In P2P risk prediction, we predict the lenders who will be in default on loans. Thus, we should focus on recall and F1-Score.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + +FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F1 - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

B. Accuracy of CRPS

We compare the accuracy of the CRPS to the models which had high accuracy in previous research, such as DNN in [17], RF in [22], LR in [18], and CatBoost in [5]. Each model adopts the best feature set in Table II to train the model.

From Table IV, LR has lowest F1-Score, which is 0.65. It seems that the ten features we select may not be suitable for LR. The F1-Score of XGBoost is 0.97, and the recall of

TABLE IV
ACCURACY OF CRPS.

Machine Learning Algorithms	*Accuracy(%)*	*Recall*	*Precision*	*F1-Score*
Deep Neural Network	98.8	0.94	0.97	0.95
Random Forest	98.8	0.95	0.96	0.95
Logistic Regression	85.7	0.97	0.49	0.65
CatBoost	99.0	0.94	0.99	0.97
XGBoost	99.0	0.95	0.99	0.97

TABLE V
THE RESULT OF DATA SYNTHESIS.

Data Synthesis	*Accuracy(%)*	*Recall*	*Precision*	*F1-Score*
Without Data Synthesis	98.9	0.93	0.99	0.95
SMOTE	99.0	0.94	0.99	0.96
borderline-SMOTE	99.0	0.95	0.99	0.97

XGBoost is 0.95. CatBoost also has higher F1-Score, but the recall of CatBoost is lower than XGBoost. That is, XGBoost can predict more default cases than other models.

C. Ablation Study:Data Synthesis

We adopt borderline-SMOTE to solve the data imbalance issue of loan data. In Table V, if we train the model without synthesizing data, the recall and F1-Score are low. Borderline-SMOTE gets higher recall and F1-Score compared to SMOTE.

D. Ablation Study: Combinations of Feature Selection Methods and Different Models

In this experiment, we try different combinations of feature selection methods and models to evaluate the performance. Since DNN automatically selects features, we build RF, LR, CatBoost, and XGBoost models with SFS, LASSO, and RFECV to compare the performance. In Table VI, XGBoost combines with RFECV achieves the recall 0.95, and the F1-Score 0.97. Although the F1-Score of CatBoost with RFECV is also 0.97, the recall of CatBoost is lower than XGBoost. The number of features selected by RFECV for CatBoost is 87, which is much more than RFECV for XGBoost. In addition, XGBoost combines with SFS selects 15 features to make F1-Score be 0.95. XGBoost with Lasso chooses 28 features to get F1-Score be 0.96. XGBoost combines with RFECV only requires 10 critical features to make the F1-Score be 0.97, and the recall is 0.95. That is, XGBoost with RFECV can utilize the fewest number of features to achieve the best result.

E. Feature Attention

Through the above experiments, we find that both XGBoost and CatBoost have good prediction results. XGBoost uses RFECV to select the 10 best features, as shown in Table II. CatBoost uses RFECV to select 87 features. The number of features of XGBoost and CatBoost is quite different. Therefore, we conduct an additional experiment to evaluate the impact of each feature on the prediction results. Table VII shows that each feature has different importance in XGBoost and CatBoost. To compare the performance of the model with the different number of features, we sort the features by the importance in Table VII and delete the least important features one by one in each experiment, as shown in Table VIII. Some selected features of XGBoost and CatBoost are overlapped. The description of the selected features of XGBoost is in Table II. $Collection\_recovery\_fee$, Out_prncp_inv, $Funded\_amnt$, and $Installment$ are critical features only for CatBoost. $Collection\_recovery\_fee$ means the post charge off collection fee. Out_prncp_inv means the remaining outstanding principal for portion. $Funded\_amnt$ means the total amount committed to that loan. $Installment$ means the monthly payment owed by the borrower.

When we select features whose importance is higher than 7.5%, XGBoost and CatBoost will get acceptable prediction results. XGBoost retains the four most critical features and makes recall be at least 83%, and the F1-Score be at least 87%. CatBoost retains the five most critical features and keeps recall be at least 89%, and the F1-Score be at least 92%. That is, CRPS obtains stable results with the least number of features.

TABLE VI
COMPARISON OF DIFFERENT FEATURE SELECTION METHODS AND DIFFERENT MODELS.

Algorithms	*Feature Selection*	*Number of Features*	*Accuracy (%)*	*Recall*	*Precision*	*F1-Score*
RF	SFS	15	98.7	0.93	0.97	0.95
	Lasso	28	98.8	0.93	0.99	0.96
	RFECV	15	98.6	0.93	0.97	0.95
LR	SFS	15	98.4	0.94	0.94	0.94
	Lasso	28	98.7	0.91	0.99	0.95
	RFECV	22	99.0	0.93	0.99	0.96
CatBoost	SFS	15	99.0	0.94	0.99	0.96
	Lasso	28	99.0	0.94	0.99	0.96
	RFECV	87	99.0	0.94	0.99	0.97
XGBoost	SFS	15	98.7	0.94	0.96	0.95
	Lasso	28	99.0	0.94	0.98	0.96
	RFECV	10	99.0	0.95	0.99	0.97

TABLE VII
THE IMPORTANCE RATIO OF FEATURES OF XGBOOST AND CATBOOST.

	XGBoost		*CatBoost*	
Importance	*Feature Name*	*Percentage*	*Feature Name*	*Percentage*
1	recoveries	30.20%	total_rec_prncp	17.35%
2	last_fico_range_high	24.63%	last_pymnt_amnt	12.67%
3	loan_amnt	10.73%	recoveries	9.25%
4	last_pymnt_amnt	7.76%	collection_recovery_fee	8.45%
5	funded_amnt_inv	5.42%	out_prncp_inv	7.78%
6	debt_settlement_flag	4.86%	funded_amnt_inv	7.22%
7	out_prncp	4.02%	out_prncp	6.93%
8	total_rec_prncp	2.72%	funded_amnt	5.85%
9	hardship_flag	1.32%	loan_amnt	5.92%
10	term	1.31%	installment	4.21%

V. CONCLUSION

We proposed a credit risk prediction system (CRPS) for P2P lending, which synthesized data and only required few features to achieve accurate results. In CRPS, a data preprocessing module, a feature selection module, a data synthesis module,

TABLE VIII
XGBOOST AND CATBOOST WITH DIFFERENT NUMBER OF FEATURES.

	XGBoost				*CatBoost*			
Number of Features	***Accuracy (%)***	***Recall***	***Precision***	***F1-Score***	***Accuracy (%)***	***Recall***	***Precision***	***F1-Score***
10	99.0	0.95	0.99	0.97	98.9	0.93	0.98	0.96
9	99.0	0.94	0.99	0.96	98.9	0.93	0.98	0.96
8	99.0	0.94	0.98	0.96	98.9	0.93	0.98	0.90
7	97.3	0.87	0.93	0.90	98.9	0.93	0.98	0.90
6	97.0	0.84	0.93	0.88	98.9	0.93	0.98	0.96
5	96.9	0.84	0.93	0.88	97.8	0.89	0.94	0.92
4	96.9	0.83	0.93	0.87	95.6	0.72	0.96	0.82
3	95.8	0.75	0.94	0.83	95.6	0.71	0.96	0.82
2	92.7	0.73	0.95	0.82	87.4	0.18	0.68	0.28
1	95.2	0.66	1.00	0.79	86.2	0.03	0.71	0.06

and five risk prediction models were implemented. The data preprocessing module dealt with the missing value of data. Feature selection module adopted RFECV to select the critical features for prediction models. In the data synthesis module, we balanced the default and non-default data by borderline-SMOTE. Finally, we built DNN, LR, RF, CatBoost, and XGBoost models to predict credit risk.

In the experiments, we evaluated CRPS based on personal loan data from the LendingClub platform. We found that the XGBoost with RFECV only required 10 features and obtained the accuracy 99%, the recall rate 0.95, and the F1-Score 0.97. We also saw that borderline-SMOTE can synthesize data and achieve higher accuracy and recall than without synthesizing data. Besides, we analyzed the impact of features on the prediction results. When we selected features with importance higher than 7.5%, XGBoost only required four critical features and achieved the accuracy be at least 96.9%, the recall be above 83%, and F1-Score be above 87%. In other words, CRPS not only solved the data imbalance issue but could achieved accurate credit risk prediction with few features. CRPS provided a reliable risk control indicator and reduced the costs for lending platforms.

REFERENCES

[1] Ajay Byanjankar, "Predicting credit risk in peer-to-peer lending with survival analysis," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017.

[2] Christophe Croux, Julapa Jagtiani, Tarunsai Korivi, and Milos Vulanovic, "Important factors determining fintech loan default: Evidence from a lendingclub consumer platform," *Journal of Economic Behavior & Organization*, vol. 173, pp. 270–296, 2020.

[3] Valeria Fonti and Eduard Belitser, "Feature selection using lasso," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1–25, 2017.

[4] Van-Sang Ha, Dang-Nhac Lu, Gyoo Seok Choi, Ha-Nam Nguyen, and Byeongnam Yoon, "Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning," in *IEEE International Conference on Advanced Communication Technology (ICACT)*, 2019.

[5] Yufei Xia, Lingyun He, Yinguo Li, Nana Liu, and Yanlin Ding, "Predicting loan default in peer-to-peer lending using narrative data," *Journal of Forecasting*, vol. 39, no. 2, pp. 260–280, 2020.

[6] Vinod Kumar, S Natarajan, S Keerthana, KM Chinmayi, and N Lakshmi, "Credit risk analysis in peer-to-peer lending system," in *IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*, 2016.

[7] Show-Jane Yen and Yue-Shi Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[8] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.

[9] Krystyna Napierala and Jerzy Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 563–597, 2016.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] Puneet Misra and Arun Singh Yadav, "Improving the classification accuracy using recursive feature elimination with cross-validation," *Int. J. Emerg. Technol*, vol. 11, no. 3, pp. 659–665, 2020.

[12] Alexis Marcano-Cedeño, J Quintanilla-Domínguez, MG Cortina-Januchs, and Diego Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *Annual Conference of the IEEE Industrial Electronics Society*, 2010, pp. 2845–2850.

[13] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005.

[14] "Lendingclub," https://www.lendingclub.com/.

[15] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert systems with applications*, vol. 33, no. 4, pp. 847–856, 2007.

[16] Tian-Shyug Lee, Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," *Computational Statistics & Data Analysis*, vol. 50, no. 4, pp. 1113–1130, 2006.

[17] David West, "Neural network credit scoring models," *Computers & operations research*, vol. 27, no. 11-12, pp. 1131–1152, 2000.

[18] Derrick N Joanes, "Reject inference applied to logistic regression for credit scoring," *IMA Journal of Management Mathematics*, vol. 5, no. 1, pp. 35–43, 1993.

[19] Ajay Byanjankar, Markku Heikkilä, and Jozsef Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," in *IEEE Symposium Series on Computational Intelligence*, 2015.

[20] Luis Eduardo Boiko Ferreira, Jean Paul Barddal, Heitor Murilo Gomes, and Fabrício Enembreck, "Improving credit risk prediction in online peer-to-peer (p2p) lending using imbalanced learning techniques," in *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017.

[21] Shuhui Chen, Qing Wang, and Shuan Liu, "Credit risk prediction in peer-to-peer lending with ensemble learning framework," in *IEEE Chinese Control And Decision Conference (CCDC)*, 2019.

[22] Haojiang Cai, "Analysis of p2p online lending default based on random forest," *Journal of Physics: Conference Series*, vol. 1237, no. 2, pp. 022046, 2019.

[23] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[24] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al., "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[25] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.