Received March 30, 2021, accepted May 7, 2021, date of publication May 12, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3079701

Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets

YEN-RU CHEN[1], JENQ-SHIOU LEU[1], (Senior Member, IEEE), SHENG-AN HUANG[1], JUI-TANG WANG[1], (Member, IEEE), AND JUN-ICHI TAKADA[2], (Senior Member, IEEE)

[1]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan
[2]Department of Transdisciplinary Science and Engineering, School of Environment and Society, Tokyo Institute of Technology, Tokyo 226-8503, Japan

Corresponding author: Jenq-Shiou Leu (jsleu@mail.ntust.edu.tw)

This work was supported by the Taiwan Tech-Tokyo Tech Joint Research Program under Grant TIT-NTUST-109-01.

ABSTRACT In the past few years, Peer-to-Peer lending (P2P lending) has grown rapidly in the world. The main idea of P2P lending is disintermediation and removing the intermediaries like banks. For a small business and some individuals without enough credit or credit history, P2P lending is a good way to apply for a loan. However, the fundamental problem of P2P lending is information asymmetry in this model, which may not correctly estimate the default risk of lending. Lenders only determine whether or not to fund the loan by the information provided by borrowers, causing P2P lending data to be imbalanced datasets which contain unequal fully paid and default loans. Imbalanced datasets are quite common in the real worlds, such as credit card fraud in transactions, bad products in the plant and so on. Unfortunately, the imbalanced data are unfriendly to the normal machine learning schemes. In our scenario, models without any adaptive methods would focus on learning the normal repayment. However, the characteristic of the minority class is critical in the loaning business. In this study, we 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. Furthermore, we use the datasets from Lending Club to validate our proposed scheme. The experiment results show that our proposed scheme can effectively raise the prediction accuracy for default risk.

INDEX TERMS Peer-to-Peer lending, imbalanced datasets, re-sampling, machine learning.

I. INTRODUCTION

Peer-to-Peer lending (P2P lending) has been developed in 2005, this application has grown rapidly in the world recently. P2P lending is an approach to get a credit without a money related organization included like banks in the choice procedure and has the likelihood to acquire preferable condition than in the traditional banking system [1]. P2P lending also provides an online platform to connect borrowers and lenders directly. In order to eliminate the brick-and-mortar operating cost, P2P lending can provide a lower interest for borrowers than that of banks. Thus, P2P lending is an alternative way for small businesses and some individuals with no credit history. However, information asymmetry becomes a fundamental problem of P2P lending because lenders only determine the loan based on information that is provided by borrowers.

Normally, Peer-to-Peer lending dataset is imbalanced because fully paid and default loans are not equal. The ratio of fully paid and default loans are around 3.5:1 in our dataset. In the real world, there are various imbalanced datasets such as fraud detection, risk management, medical diagnosis and so on. Hence, it is difficult to make a prediction on such an imbalanced dataset because the classifiers are prone to detecting the majority class rather than to the minority class. Therefore, the output of the classification will be biased. So in this case, addressing the problem in the classification of the imbalanced dataset is highly important.

This study uses under-sampling and cost-sensitive learning for dealing with the imbalanced dataset. Meanwhile, for machine learning schemes, we utilize logistic regression, random forest, and neural network for foreseeing default risk of P2P lending. Furthermore, this paper is organized as follows: Section 2 briefly reviews the related work about predicting default risk of P2P lending and classification of imbalance dataset. Next is Section 3 which presents our methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

Then, Section 4 shows the evaluation metrics and experiment result. The last one is conclusions which are drawn in Section 5.

II. RELATED WORKS

A. CREDIT RISK

Credit risk is the most important issue in the financial field and there are many types of research in the credit risk. Odom and Sharda [2] compared between neural networks model and discriminant analysis method in bankruptcy risk prediction. Then, the result proved that the neural networks model had a higher percentage than that of discriminat analysis. Atiya [3] reviewed the problems of bankruptcy prediction using the neural networks and proposed novel inputs extracted from the equity markets as new indicators to improve the prediction considerably. Moreover, Emekter *et al.* [4] found that higher interest rates charged on the high-risk borrowers were insufficient to make up for a higher likelihood of the loan default. The Lending Club must discover approaches to draw in high FICO score and high-wage borrowers with a specific end goal to support their organizations.

Meanwhile, in P2P lending, Bachmann *et al.* [1] and Mateescu [5] reviewed the history of P2P lending and analyzed the advantages and the disadvantages of P2P lending. Then, they introduced how P2P lending works and explained the difference between traditional bank lending and P2P lending. Serrano-Cinca *et al.* [6] used the statistical methods such as Pearson's correlation, point-biserial correlation, and chi-square test to analyze different variable in prediction default risk of P2P lending. For evaluating the most predictive factor of default, they created 7 logistic regression models by using different 7 variables.

Beside the statistic method, there were also some researches that used machine learning algorithms to predict the default risk. Jin and Zhu [7] compared three kinds of machine learning models, namely decision tree, neural networks and support vector machine in P2P lending default risk prediction. They used the dataset of Lending Club from July 2007 until December 2011 and removed the loan data that status is "current". The prediction result was grouped into three classes such as "defaulter", "need attention", and "well paid". Then, they also employed the average percent hit ratio (accuracy) and life curve to evaluation performance. Byanjankar *et al.* [8] used a P2P lending platform Bondora datasets to create neural networks model and utilizing confusion matrix and accuracy to evaluate performance.

In 2016, the authors in [9] proposed a profit scoring scheme. In [9], credit scoring systems mainly focus on loan default probability. By analyzing borrower's interest rate and lenders' profitability, the results indicate that the P2P lending is not a trend in current market. Reference [10] combined cost-sensitive learning and extreme gradient boosting. By doing so, this method can simplify optimization problem to an integer linear programming. Differ from other studies, this research estimates expected profitability in other metrics, such as annualized rate of return (ARR). The metrics used in estimation are designed on the basis of an imbalanced dataset.

Although there were some researches in prediction P2P lending default risk, they did not focus on addressing the problem that imbalanced datasets bring. The main evaluation metrics that they used was accuracy which was not suitable for imbalanced datasets.

B. CLASSIFICATION WITH IMBALANCED DATASETS

Imbalanced datasets are an exceptional case for order issue where the dispersion of class is not equivalent among the classes. Regularly, there are two classes in imbalanced datasets like the majority of negatives class and the minority of positive class. These type of data presume an issue for data mining since standard classification algorithms normally consider a balanced training set and this supposes a bias towards the majority (negative) class [11]. For instance, we have a classifier with 96% of accuracy. It looks very good, but if the 96% data is majority class data, the classifiers will always predict the majority class to get high accuracy.

Veni *et al.* [12] found that why existing classification algorithms are poor performance in imbalanced datasets. Firstly, these algorithms are accuracy driven. The second, the assumption of the distribution of all classes is the same. The third, different classes have the same error cost. They introduced the sampling strategies and cost sensitive learning to address the issue of expectation imbalanced datasets and also used the other performance metrics that were more suitable for imbalanced datasets, such as confusion matrix, precision, F1-score and so on. In addition, to enhance of the sampling strategies, Chawla [13] observed the synthetic minority oversampling technique (SMOTE), ensemble-based method and SMOTE Boost method on imbalanced datasets. Furthermore, Chawla *et al.* [14] also surveyed the issue of imbalanced datasets and its solutions from some researches.

III. PROPOSED SCHEME

This segment depicts the way toward developing advance default expectation models, which is envisioned in Fig. 1.

A. PREPROCESSING

The P2P lending datasets contain many attributes which are empty for most records. Therefore, we remove these attributes and modify the nominal features by using one-hot-encoding technique that can transform nominal features to be a format suitable for classification. For instance, we have a feature "purpose of the loan" which has string value such as "Car", "Business", and "Wedding". Normally, we use ordinal value to encode these to be numbers such as 0, 1, and 2. However, in machine learning schemes, different categories have the same weight. Thus, the ordinal technique cannot be implemented in machine learning because the lowest and the highest value will affect the classification result. One-hot-encoding uses one Boolean column for each category which has different weight. Finally, we use feature scaling to standardize the range of value of each feature.

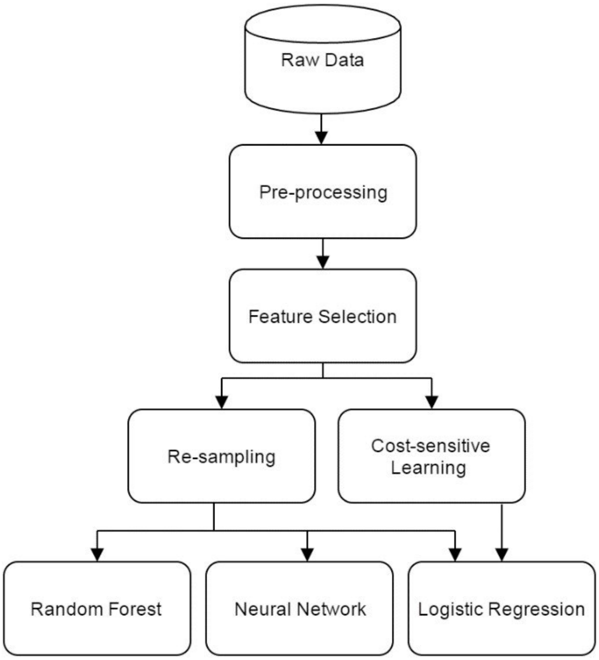


FIGURE 1. Data processing flow.

TABLE 1. Selecting feature for building models.

Category	Attribute	Value Type
Loan characteristics	Loan amount Loan purpose	Real-value One-hot-encoding
Borrower characteristics	Annual income Home ownership status Employment length LC assigned loan grade	Real-value One-hot-encoding Real-value One-hot-encoding
Borrower assessment	Interest rate Installment Term	Real-value Real-value One-hot-encoding

If some features have a wide range of values, it will influence the performance of machine learning models. Moreover, feature scaling also makes gradient descent converges much faster.

B. FEATURE SELECTION

In this section, we describe features used to prediction. First, we intuitively choose relevant features, such as loan amount, installment, and so on. The relevant features are shown in Table 1. Secondly, we distinguish the residence of borrowers by the three-digit of zip code. If we encode the zip code by using one-hot-encoding, data dimensions will be too high. Thus, we decide to calculate the mean and the median income of each state and add these two features into data. Original features contain words that describe loan applications. In general, words cannot be numerical features. Here, we analyze the words first. The words related to default loans and fully-paid loans are illustrated in Fig. 2 and 3, respectively. The sizes of words are proportional to frequency of occurrence. Obviously, some common words, such as "credit", "card", and "loan", are shown in two word clouds. That is, the positive samples and negative samples contains common words. In classification works, these common words would result in lower accuracy. Hence, we remove the common words from our features. Finally, we encoded the rest of words to numerical features.

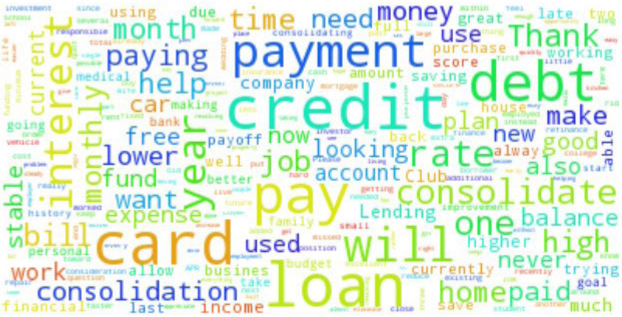


FIGURE 2. A word loud for default loans.

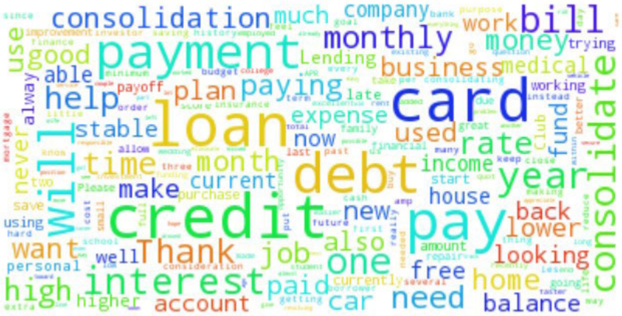


FIGURE 3. A word cloud for fully paid loans.

C. RE-SAMPLING

The re-sampling method makes the datasets becoming balance by changing the distribution class. It is divided into two types, the first is called under-sampling which renders the larger class to reach a size close to that of the smaller class. Meanwhile, the second type is over-sampling which performs the small class to reach a size close to that of the larger class [15].

1) UNDER-SAMPLING

Under-sampling is to select a subsample of the majority class that its size matches the set of minority class to make the datasets balance. However, it potentially brings another issue because it removes certain important data. Another type method of under-sampling is random under-sampling which eliminates data of the majority class randomly until the class distribution balance. In our research, we applied Tomek as under-sampling method. Tomek links is also another way to do under-sampling. Tomek links is also considered as a pair of the minimum distance of nearest neighbors of opposite class. In the under-sampling, Tomek link method eliminates data of the majority class that belongs to Tomek link.

2) OVER-SAMPLING

The over-sampling method produces additional data of minority class to make the datasets distribution balance. The random over-sampling method is a simple way to expand minority class data by randomly duplicating the data. SMOTE [16] which stands for synthetic minority oversampling technique is another method to perform over-sampling. Take a feature vector x_i of minority examples and m is the nearest neighbor minority examples in feature space. Then, the interpolation between m and x_i is done to produce new data of minority class until distribution balance. Borderline SMOTE is the new form of SMOTE over-sampling method which only perform over-sampling the borderline data of minority class [17]. If the number of x_i's nearest neighbor that belong to majority class which fit $\frac{m}{2} < |x_j \cap majority| < m$, define the x_i near the borderline and form the new data.

3) COST-SENSITIVE LEARNING

Practically, the ratio of the positive samples to negative samples is not 1:1. For instance, the number of murderer would lower than kind people. The loans are also imbalanced data. Thus, the traditional cost function would suffer from imbalanced data. To overcome this obstacle, we set a scalar α in Eq. 1. In this way, the loss from fewer negative samples would increase the term behind addition operator. In this study, we perform experiment with α from 1 to 4.8. Compared to other methods that applied to deal with imbalanced datasets, Eq. 1 is an intuitive approach for machine learning models with imbalanced datasets. In this way, the model can classify the targets better than one without the adaptive cost function.

$$cost = \sum_{i=0}^{m}(1 - y^{(i)}) \log(1 - h(x^{(i)})) + \alpha y^{(i)} \log(h(x^{(i)})), \quad (1)$$

where h is an activation function, such as sigmoid function.

4) MACHINE LEARNING SCHEME

In section III-C3, we introduce the concept of cost-sensitive learning. Next, we would illustrate the machine learning models in this study. Simultaneously, the cost function in the logistic regression is replaced with Eq. (1) Re-sampling is used to the random forest and neural networks.

a: LOGISTIC REGRESSION

In this section, we discuss machine learning models which are employed in this study. Firstly, we utilize logistic regression which is suitable for binary classification. The logistic regression model is figured in Eq. (2) which transform the linear regression into non-linear. The output of the logistic regression model is as probabilities between 0 and 1. Normally, the logistic regression threshold is 0.5. However, if the result greater than 0.5, it will be predicted as the true value. In this research, we applied Eq. 1 to our training strategy. Other parameters are set in default values by scikit learn framework.

$$g(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

$$h_0 = g(\theta^T x) \quad (3)$$

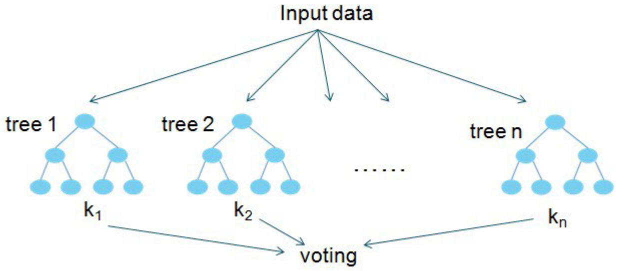


FIGURE 4. Random forest model.

b: RANDOM FOREST

The next machine learning that we use is random forest (RF) which is an ensemble of decision trees as described in Fig. 4. It constructs multitude of decision trees at training time and creates multiple different models by bagging selection training data and randomly selection features. Finally, deciding the final result is done by using majority voting. We utilize the methodology of CART (classification and regression tress) to build the trees of the random forest. CART algorithm is a binary decision tree which uses Gini index for impurity measurement.

$$Gini(S) = \sum_{j=1}^{n} P_j^2, \quad (4)$$

where P_j is the purity of jth data.

$$Gini_A(S) = \frac{|S_1|}{|S|}Gini(S_1) + \frac{|S_2|}{|S|}Gini(S_2), \quad (5)$$

where S_1 and S_2 are subsets of S.

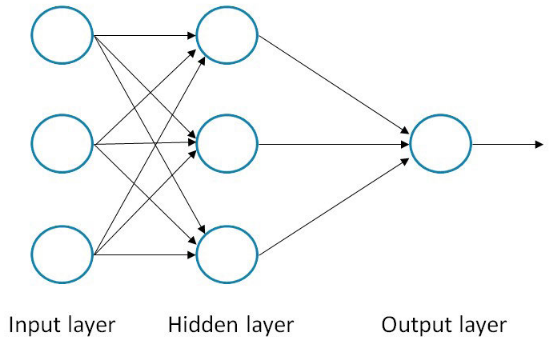
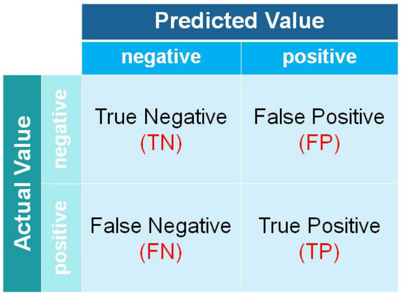


FIGURE 5. A simplified three layer of neural networks model.

c: NEURAL NETWORKS

Neural networks are inspired by biological neural networks. Fig. 5 shows a simplified three layer of neural networks. The input layer passes different features and communicates to one or more hidden layers. The node is called neuron which presents an activation function. Every connection presents a weight. The weight value is different from one to others. These weights and non-linear activation function



		Predicted Value	
		negative	positive
Actual Value	negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

FIGURE 6. Confusion matrix.

produce complex relationships. In our study, the model consists 64 input neurons, two hidden layers, and one output. In order to prevent our model from overfitting, we set dropout rate to 0.5.

IV. RESULTS AND DISCUSSION

A. EVALUATION METRICS

Accuracy is not enough to evaluate imbalanced datasets. Therefore, we use another evaluation metrics which called confusion matrix to evaluate the performance of machine learning schemes. The confusion matrix is a specific table layout which can show the classifier result as explained in the Fig. 6. The examples of a classifier which predicts the correctness are called TP (true positive) and TN (true negative). Meanwhile, the examples of a classifier which predicts the incorrectness are called FP (false positive) and FN (false negative). Then, recall is called sensitivity because its representative predicts the positive rate in all actual positive data, while the precision representative is the correct rate when it predicts positive. Specificity is also called the true negative rate. In order to combine recall and precision, we use the F1-score which is the harmonic mean of precision and recall and G-mean which combines sensitivity and specificity.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{9}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{10}$$

B. LENDING CLUB DATASETS

We use the Lending club datasets which were collected from 2007 until 2015 and contains 887,379 loan records. This data set is from website https://www.lendingclub.com/investing. Then, we remove the loans data which has status "Current" and the remaining number of loans data is 269,668 as explained in Fig. 7. Fig. 8 depicts the status of all loans in the dataset. We allocate 70% of the data for training (N = 188, 767) and 30% (80,901) of data for testing. The raw data contains 73 features for each record. Afterward, we classify the loan data into two categories, namely default and fully paid. Default label contains value as default, charge off, and late to payment loan which is classified as positive examples whereas a fully paid label is classified negative examples. In short, the ratio of fully paid and default loans are around 3.5:1 in our dataset.

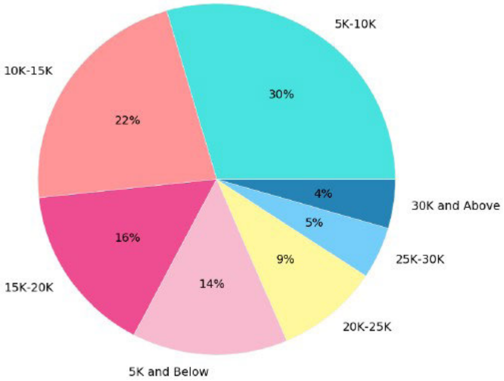


FIGURE 7. Percentage of loan amount overall.

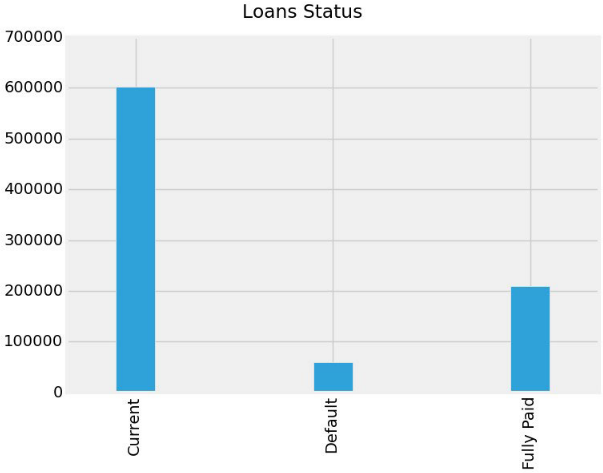


FIGURE 8. Loan status.

TABLE 2. Classification result of random forest.

Re-sampling	Accuracy	Recall	F1	G-mean
Original ratio	77.688	9.260	15.606	30.108
Random under-sampling	64.721	61.369	43.662	63.488
Tomek links	77.426	13.361	20.868	35.775
Random over-sampling	76.583	18.605	26.144	41.641
SMOTE	77.390	12.212	19.398	34.253
Borderline SMOTE	77.168	11.896	18.840	33.772
SMOTE + Tomek link	77.377	10.997	17.802	32.560

C. EVALUATION RESULT

In this part, we discuss the results. Firstly, we try different sampling methods on three machine learning algorithms. Table 2 shows the classification result of random forest.

TABLE 3. Classification result of neural networks.

Re-sampling	Accuracy	Recall	F1	G-mean
Original ratio	77.738	0.177	0.354	4.213
Random under-sampling	64.061	65.220	44.706	64.470
Tomek links	77.906	4.255	7.903	20.528
Random over-sampling	65.729	63.744	45.316	65.008
SMOTE	37.927	91.932	39.758	45.429
Borderline SMOTE	39.281	90.261	39.842	47.188
SMOTE + Tomek link	49.220	80.263	41.322	56.890

TABLE 4. Classification result of logistic regression.

Re-sampling	Accuracy	Recall	F1	G-mean
Original ratio	77.885	6.414	11.443	25.119
Random under-sampling	63.628	66.246	44.773	64.519
Tomek links	77.761	10.659	17.597	32.154
Random over-sampling	63.367	66.618	44.758	64.493
SMOTE	63.946	65.714	44.814	64.566
Borderline SMOTE	64.065	65.425	44.787	64.545
SMOTE + Tomek link	69.943	63.711	44.742	64.499

TABLE 5. The result of cost-sensitive learning for logistic regression.

α	Accuracy	Recall	F1	G-mean
1	77.885	6.414	11.443	25.119
1.2	77.736	11.807	19.113	33.779
1.4	77.342	18.044	26.189	41.258
1.6	76.734	24.275	31.735	47.199
1.8	75.754	30.251	35.728	51.829
2	74.580	35.667	38.466	55.298
2.2	73.360	41.282	40.843	58.378
2.4	71.992	46.587	42.565	60.771
2.6	70.599	51.242	43.710	62.466
2.8	68.984	55.210	44.230	63.455
3	67.357	58.628	44.450	63.998
3.2	65.770	61.929	44.631	64.353
3.4	64.268	65.159	44.826	64.583
3.6	62.742	67.761	44.760	64.451
3.8	61.231	70.119	44.623	64.147
4	59.848	72.589	44.612	63.869
4.2	58.376	74.525	44.373	63.289
4.4	56.965	76.278	44.125	62.634
4.6	55.716	78.087	43.997	62.047
4.8	54.485	79.669	43.815	61.365

By using the original data, the F1-socre is 15.606 but when we try to balance dataset by implementing different sampling methods, the F1-score increases. The same thing also happens to the neural networks and logistic regression which the outcomes appear on Table 3 and 4. From Table 2, 3,and 4, we can evidently discover that the logistic regression with re-sampling or the cost-sensitive learning outperforms the random forest and neural networks. In view of the outcomes, the best sampling method is random under-sampling because it has the highest F1-score than that of other sampling methods. Furthermore, Table 5 shows the result of logistic regression with cost-sensitive learning. We tried 20 different α and found that the better α are 3, 3.2, and 3.4 because the accuracy and accuracy default have slight differences. In the feature selection, it is an important issue how many features are selected to obtain an optimum performance. Therefore, we select the first ten important features which are yield by performing random forest. Moreover, we also use another feature set that uses only the first three important features to experiment with the random under-sampling method and the result shows in Table 6. In addition, in order to analyze the correlation between the forecasting results and the amount distribution, we take two range of loan amount. The first loan amount is less than \$5,000 and the second is greater than \$30,000. The results are shown in Table 7 and 8. As a whole, the outcomes of two loan amount distribution range data are not so different from the result of whole data. It means that the amount of the loan is not too affected the forecasting result. Overall, in this research, the cost-sensitive learning and re-sampling improve the quality of prediction task. Especially, random under-sampling can effectively assist machine learning models attain better results than original ones. Although Random under-sampling makes the datasets balanced and decrease the size of datasets, the accuracy and recall of machine learning models are kept on a high level.

TABLE 6. The result of using three important features.

Random under-sampling	Accuracy	Recall	F1	G-mean
Random forest	63.931	60.881	42.923	62.812
Neural networks	63.559	66.463	44.830	64.567
Logistic regression	63.236	66.146	44.494	64.247

TABLE 7. The result of using a loan amount below \$5,000.

Random under-sampling	Accuracy	Recall	F1	G-mean
Random forest	62.941	58.368	37.526	61.128
Neural networks	60.840	65.276	38.865	62.475
Logistic regression	61.237	67.182	39.795	63.403

TABLE 8. The result of using a loan amount below \$30,000.

Random under-sampling	Accuracy	Recall	F1	G-mean
Random forest	65.550	57.830	57.491	59.719
Neural networks	58.294	67.641	48.610	60.686
Logistic regression	60.686	67.169	50.00	62.531

V. CONCLUSION

Peer-to-peer (P2P) lending is a solution to lend money without involving financial institutions and allows borrowers to connect to lenders directly. However, P2P lending has a fundamental problem because its dataset is imbalanced. Therefore, it makes the classifiers are prone to majority class rather than minority class. In this study, we employ a various machine learning algorithm to predict the default risk of P2P lending, use re-sampling and cost-sensitive mechanisms to processing imbalanced datasets. We get the dataset from Lending Club to validate our proposed scheme. In the experiment results, random under-sampling shows the best performance in different classifiers. Then after doing preprocessing and feature selection, the proposed scheme can effectively raise the prediction accuracy for default risk.

REFERENCES

[1] A. Bachmann, A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, P. Tiburtius, and B. Funk, "Online peer-to-peer lending—A literature review," *J. Internet Banking Commerce*, vol. 16, no. 2, pp. 1–19, Aug. 2011.

[2] M. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 2, Jul. 1990, pp. 163–168.

[3] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," *IEEE Trans. Neural Netw.*, vol. 12, no. 4, pp. 929–935, Jul. 2001.

[4] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, "Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending," *Appl. Econ.*, vol. 47, no. 1, pp. 54–70, Jan. 2015.

[5] A. Omarini, "Peer-to-peer lending: Business model analysis and the platform dilemma," 2018.

[6] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of default in P2P lending," *PLoS ONE*, vol. 10, no. 10, Oct. 2015, Art. no. e0139427.

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

[8] A. Byanjankar, M. Heikkila, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 719–725.

[9] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decis. Support Syst.*, vol. 89, pp. 113–122, Sep. 2016.

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

[11] N. Rout, D. Mishra, and M. K. Mallick, "Handling imbalanced data: A survey," in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Singapore: Springer, Jan. 2018, pp. 431–443.

[12] C. V. KrishnaVeni and T. S. Rani, "On the classification of imbalanced datasets," *IJCST*, vol. 2, pp. 145–148, 2011.

[13] N. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, vol. 5. Jan. 2005, pp. 853–867.

[14] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 1–6, Jun. 2004.

[15] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, Feb. 2004.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[17] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing* (Lecture Notes in Computer Science), vol. 3644. Berlin, Germany: Springer, Sep. 2005, pp. 878–887.

YEN-RU CHEN received the B.S. degree in electrical engineering from Fu Jen Catholic University and the M.S. degree from the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, in 2016. Her research interests include machine learning and data analysis.

JENQ-SHIOU LEU (Senior Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1991 and 1993, respectively, and the Ph.D. degree on a part-time basis in computer science from National Tsing Hua University, Hsingchu, Taiwan, in 2006. From 1995 to 1997, he was with Rising Star Technology, Taiwan, as an Research and Development Engineer. From 1997 to 2007, he worked in the telecommunication industry with Mobital Communications and Taiwan Miobile, as an Assistant Manager. In February 2007, he joined the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, as an Assistant Professor. From February 2011 to January 2014, he was an Associate Professor and he has been a Professor, since February 2014. His research interests include mobile services over heterogeneous networks and heterogeneous network integration.

SHENG-AN HUANG received the B.S. degree from the Department of Electronic and Computer Engineering, Taipei, Taiwan, in June 2019. He is currently a Teaching Assistant for calculus and a graduate student with the National Taiwan University of Science and Technology. His research interests include machine learning, algorithms design, social networks analytics, and numerical methods for engineering.

JUI-TANG WANG (Member, IEEE) received the master's degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, in 2000, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, in 2008. In February 2019, he joined the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, as an Assistant Professor. His research interests include wireless communication and security protocol.

JUN-ICHI TAKADA (Senior Member, IEEE) received the Ph.D. degree in electrical and electronic engineering from the Tokyo Institute of Technology (Tokyo Tech), in 1992. After serving as a Research Associate with Chiba University, from 1992 to 1994, and as an Associate Professor with Tokyo Tech, from 1994 to 2006, he has been a Professor with Tokyo Tech, since 2006. From 2003 to 2007, he was a Researcher with the National Institute of Information and Communication Technology (NICT), Japan. His current research interests include radio-wave propagation and channel modeling for mobile and short range wireless systems, applied measurement using radio waves, and ICT applications for international development. He is a fellow of the Institute of Electronics, Information and Communication Engineering (IEICE), Japan, and a member of the Japan Society for International Development.