ICIC Express Letters
Part B: Applications
Volume **10**, Number **11**, November **2019**

ICIC International ©2019 ISSN 2185-2766
pp. 1013–1019

A ROTATION FOREST-BASED SMOTE FOR ASSESSING P2P LENDING RISKS

Francis Joseph Costello[1] and Kun Chang Lee[2,3,*]

[1]School of Business Administration
[2]Global Business Administration
[3]Department of Health Sciences & Technology
Samsung Advanced Institute for Health Sciences & Technology (SAIHST)
Sungkyunkwan University
25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea
joe.costehello@gmail.com; *Corresponding author: kunchanglee@gmail.com

Received May 2019; accepted August 2019

Abstract. *In this research our aim was to identify good borrowers within the context of social lending. We investigated effective ways to conduct this credit risk assessment of these borrowers. We implemented a classification approach in order to make our analysis and came upon the problem of a major class imbalance. In literature, there are few studies on solving this problem in the social lending setting. In this sense, we propose the implementation of an over-sampling technique known as Synthetic Minority Over-Sampling Technique (SMOTE). Furthermore, we wanted to explore the use of the ensemble technique Rotation Forest which has had little attention in the literature within this setting. Thus, our research looked to compare this model to more classically used models in social lending such as Support Vector Machines and Logistic Regression. We also included the use of a deep neural network model as deep learning has proven its worth in complicated problems of late. The data sample used in our analysis was retrieved from the publicly available LendingClub website, which is a platform that facilitates P2P lending and also shares data on these lenders. Our results showed that the implementation of the Rotation Forest classifier alongside SMOTE gave significantly increased ability to identify good borrowers within our data sample, including more modern techniques such as deep learning.*
Keywords: P2P lending, Credit risk assessment, Classification, Rotation forest, SMOTE

1. **Introduction.** Peer-to-Peer (P2P) lending or otherwise known as social lending is an online service whereby lending is from one person to another via an online agent. P2P lending is slowly growing into a sizable portion of the lending environment thanks to the widespread access of the Internet, with year-on-year market growth seen since its inception in 2012, see Figure 1. By engaging with P2P lending, a mutual benefit is incurred: borrowers can obtain loans which have previously been rejected by mainstream banks or that are cheaper comparatively; with lenders profiting from higher interest rates based on the borrowers credit rating [1-3].

Till now, leading companies in this field, such as Prosper, LendingClub (LC) and Kiva, have actively tried to seek and engage individuals who want to directly lend money to other individuals when small amounts is necessary for the borrower. On top of this, traditional banks are reluctant to engage in similar types of lending and thus the rise of these companies and their lending platforms are on the rise [4,7]. More specifically, prospective lenders can fund listings made by potential borrowers who must specify the loan amount for a prospective lender to fulfil these listings [5]. However, loans are usually

DOI: 10.24507/icicelb.10.11.1013

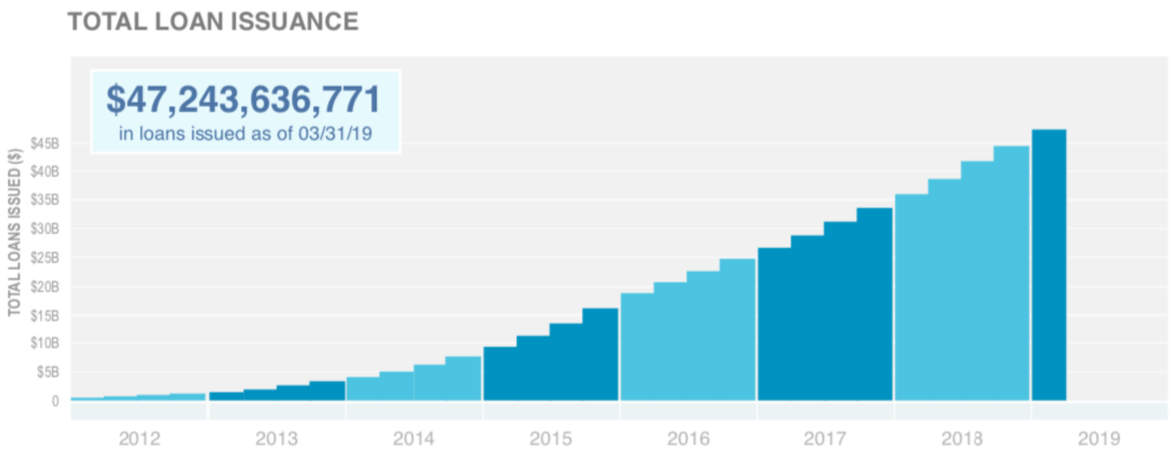

FIGURE 1. LendingClub loans issued online since 2012 in dollars, Source [7]

uncollateralized leading to lenders seeking higher returns for the financial risk they incur [6].

As with any lending, reducing the risk of one's loans remains a key area of interest. In P2P lending, an investor can look to a traditional economic model in an attempt to evaluate potential borrowers. Another viable method is with the use of machine learning models [8,9]. Lenders actively engaged in P2P lending interact with credit score models that are made based on traditional credit score risk assessment. Although such metrics are useful, prior research on P2P lending has shown that it works under different dynamics when compared to more traditional lending methods [4]. This has meant that traditional lending risk assessment models have not yielded desirable risk assessments [3].

Due to this reason in this paper we look to expand upon previous research and look to find a model that is able to take into consideration the dynamics at play within P2P lending by exploring the use of a Rotation Forest (RoF) model. The rest of this paper is organized as follows: 1) a look at recent trends in P2P lending including common machine learning techniques, 2) a description of our methodology will be presented, 3) experimental results will be presented, and 4) lastly conclusion as well as future recommendation and limitations will be explored.

2. Related Work.

In P2P lending identification of so called "good borrowers", i.e., those who will pay back their loan in full within due time is of great importance for investors participating in social lending. In turn, reducing risk allows for more profitability of social investors which is a critical component in continued interest in social lending as well as the overall sustainability of the social lending market. However, recent work in this area has found that common risk assessment models, such as FICO, are not sufficient in dealing with the dynamics of P2P lending, proclaiming "traditional financial score metrics are not well-equipped to capture the non-conventional dynamics prevalent in social lending" [3]. Consequently, the way to approach risk within this problem requires a readjustment of traditional credit risk models. The reasons for this are the following.

1) P2P lending platforms provide considerable amounts of data on borrowers. Little is known of borrowers credit history, leading to lenders suffering from potential information asymmetry. This lack of knowledge on a borrower could deter potential lending behavior. Companies such as LC try to provide as much data as possible on borrowers to help promote a fair loaning system. This makes P2P lending considerably more transparent [2].
2) Loans on a P2P website is more like an auction process. A borrower puts up a listing for money they would like and then lenders bid for the loan. Loans that attract attention

can attract other lenders who get distracted from other viable loan options on the website [3].

3) Traditional credit risk metrics, such as FICO, are not always the best indicators of a good borrower. Recent literature has found that FICO and grades given by financial institutions are not always enough in dealing with the dynamics of P2P lending [4], proclaiming "traditional financial score metrics are not well-equipped to capture the non-conventional dynamics prevalent in social lending" [3]. Hence identification of "good borrowers", i.e., those who will pay back their loan in full within due time, is of great importance. Additionally, assessing this risk can help continuing the profit of social investors a critical component in sustaining the P2P market.

2.1. **Data imbalances and resampling.** When it comes to credit risk assessment, datasets used to study loans have meant working with unbalanced classes, with the minority class normally representing defaulted or written off loans [10]. In order to tackle the problem of class imbalances we implement an over-sampling technique known as Synthetic Minority Over-Sampling Technique (SMOTE) [11]. SMOTE has been successful in previous credit risk assessment and thus will be implemented in this research [12,13]. SMOTE creates instances by using a kNN algorithm to produce instances from the minority class. $S_{\min}$, S_{maj}, S_{syn} are the classes for minority, majority and synthetic. 1) Determine kNN for a sample $x_i \in S_{\min}$ and determine the value of S_{syn}. 2) Next, it chooses a random sample x_i $(i = 1, 2, \ldots, k)$ from k nearest neighbors of sample $x_i \in S_{\min}$. 3) Implement (1) to create synthetic sample x_n [11]. This can be represented in the following way:

$$x_{nj} = x_{ij} + gap * (x_{tj} - x_{ij}) \tag{1}$$

where *gap* can be a random number between 0 and 1 and $i = 1, 2, \ldots, |S_{\min}|$, $t = 1, 2, \ldots, k$, $j = 1, 2, \ldots, m$. With SMOTE you can generate as many synthetic instances in a dataset as what is in the minority class [11].

2.2. **Models.** To test our approach, we have implemented the following classification models: firstly, two standard classifiers in Support Vector Machines (SVMs) and Logistic Regression (LG), next, we implemented one ensemble technique, Rotations Forest (RoF), and lastly, we implemented two Deep Neural Network (DNN) models. One of these models consisted of two hidden layers and the other one consisted of four hidden layers.

2.2.1. *Support vector machines.* SVM is a classification and regression model where there exists a hyperplane as the decision plane separated by the positive (+1) and negative (−1) classes [15]. Given a training data, $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, in which $x_i \in R^d$ signifies vectors in a d-dimensional hyperplane, and $y_i \in [-1, +1]$ is a class label given to the data. SVM are then represented by morphing the input vectors into a new, higher dimensional analogue plane indicated as: Φ: $R^d \to H^f$ in which $d < f$. Thereafter, an optimum hyperplane is formed by a kernel function $K(x_i, x_j)$, which is the product of the input vectors x_i and x_j, in which $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ [14]. In this paper we implement the polynomial SVM, where p is the degree of polynomial:

$$K_{poly}(x_i, x_j) = (x_i \cdot x_j + 1)^p \tag{2}$$

2.2.2. *Logistic regression.* LG is popular and widely used model in credit risk assessment [10] and P2P lending [1]. The approach for LG can be seen for binary classification in the following formula, where $\acute{\beta}_l$ is calculated though the maximum likelihood method:

$$\hat{f}_{LR}(x) := \frac{e^{\acute{\beta}x}}{1 + e^{\acute{\beta}x}} \tag{3}$$

2.2.3. *Rotation forest.* Rotation forest is a tree-based ensemble with some key differences to random forest. Firstly, it uses all attributes for each tree, rather than sampling. Attributes are split into r random sets of a given size f, and the transformation is built independently for each set of attributes. Next RoF discards instances of a given class and then groups them in a sample with replacement to include a give proportion of cases. A Principle Component Analysis model is then built on this reduced data set, and the model is then applied to all instances to generate f new attributes for that particular set. The new attributes are then assembled to form a new data set with $m = r * f$ attributes [16]. Finally, the classification can be expressed as:

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^{L} d_{i,j}\left(xR_i^a\right), \quad j = 1, \dots, c \tag{4}$$

where $d_{i,j}\left(xR_i^a\right)$ represents the probability assigned by the classifier D_i to the hypothesis that x comes from class w_j, a class label.

2.2.4. *Deep neural network.* DNNs are a type of NN that have come to prominence in academia of late. DNNs are based on feedforward networks which can be represented by composing together many different functions say $\left(f^{(1)}, f^{(2)}, f^{(3)}\right)$ are all interconnected within a given chain:

$$f(x) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(x)\right)\right) \tag{5}$$

Here, $f^{(1)}$ represents the input layer, $f^{(2)}$ represents a deep (hidden layer), and $f^{(3)}$ represents an output layer. DNNs allow for a nonlinear transformation represented by ϕ. ϕ is a way of describing x based on a number of features within a given DNN, whereby learning ϕ is the ultimate goal [17]. The model $y = f(x : \theta, w) = \phi(x : \theta)$ can be seen as θ parameters that are implemented to learn ϕ from a broad class of functions, and parameters w that go from $\phi(x)$ to an output desired by the user. ϕ can be viewed as the deep (hidden) layer of a given DNN meaning a user only needs to find the right general function rather than needing a precise definition of a given model [18].

3. **Methodology.** In this section we will describe the methodology that we have used in our study in order to study P2P risk assessment. This includes the dataset, pre-processing steps and evaluation metrics used to achieve the final results seen in the next section.

3.1. **LendingClub dataset.** The dataset in this study was retrieved from LC as it is made publicly available for free. The data we obtained was based upon all loan requests in the year 2016. Following recommendations from previous researchers [3] we applied the same steps to creating the final feature list for learning. This had 18 features in total including one class and all pre-processes and data manipulation followed the procedure seen in [3]. Standard features selected included: 1) loan status (class attribute), 2) annual income, 3) credit age, 4) delinquencies (last two years), 5) employment length, 6) home ownership, 7) inquiries (credit inquiries in the last six months), 8) loan amount, 9) loan purpose, 10) open accounts (currently opened credit lines), 11) total accounts (total number of credit line held), 12) term (length of loan).

The next features are ratios seen within [3]: 13) DTI (Debt to Income Ratio), 14) Income to Payment Ratio: this ratio represents the loan's monthly payments to monthly income, 15) Revolving Utilization Rate, 16) Revolving to Income Ratio: revolving credit balance to the borrower's monthly income. The last two are based upon scores provided by LC. 17) FICO Score: this is a standard credit line that is used in the majority of lending decisions in the US [4]. It is based on financial attributes from the borrowers' credit records [3]. 18) LC Grade: this is a grade given by LC themselves. These are A1-G5, with A representing a less risky loan.

3.2. **Pre-processing.** In line with previous research the raw numerical figures had to be pre-processed based on the logarithm function (transformed features included 2, 3, 14, and 16) [3]. We then had to convert the nominal text data into binary numerical figures. The idea with this method is if one nominal feature, say 'home ownership [yes; no]', is present in the dataset, this will be turned into two separate features whereby non-presence of the 'yes' is represented by a zero and presence of 'yes' is represented by a one. In our dataset we had two nominal features (6 and 9) with a total of fourteen values within them. Thus, the final dataset went from eighteen features to a final feature number of thirty features (18 original features plus 14 new binarized features (32) minus the original 2 features which have been split (30)). Furthermore, all data was standardized to allow for all features to have a zero mean and unit standard deviation.

3.3. **Cost sensitive analysis.** In credit risk assessment finding bad borrowers is paramount to its success and therefore misclassification holds a greater risk [3]. For this reason, we implemented the use of a cost-sensitive analysis [3]. Based on Schenbesch & Stecking [19], Malekipirbazari & Aksakalli suggest that the cost sensitive ratio should be 5 : 1 [3]. However, with the implementation of SMOTE, the minority class is boosted and thus punishing the classifier by 5 : 1 for wrongly predicting a bad borrower seems too harsh. Therefore, we tested cost sensitivity from 2 : 1 to 5 : 1 with the SMOTE dataset based on the RoF classifier. With cost-sensitive analysis there is a trade-off between the accuracy and precision and finding a balance is key.

3.4. **Model description.** All tests were performed using the WEKA tool (https://www.cs.waikato.ac.nz/ml/weka). All the models used WEKA's original settings. Lastly, the implemented DNN models were created with 4 (32, 64, 128 and 256) and 2 (64, 64) dense layers. The output layer was shaped with a Softmax output layer. The activation function used was the rectified linear unit (ReLu).

4. **Results.** In this paper, we investigated whether SMOTE can successfully identify bad borrowers based on two main metrics: accuracy (ACC) and the area under the roc curve (AUC). 10-fold cross validation was used so that training data was split into 10 subsets of equal size. The mean results of each fold were then analyzed with a t-test to find significance in the results.

4.1. **Cost-sensitive analysis parameter analysis.** As can be seen in these results the best trade-off lies within a 3 : 1 cost sensitive analysis. This attempts to punish the classifier for falsely classifying a bad borrower as good; without drastically reducing the accuracy too much while retaining a high precision.

Table 1. Cost-sensitivity analysis of RoF

Cost-sensitivity analysis with RoF			
	ACC	AUC	Precision
RoF 2 : 1	82.79%	0.85	0.85
RoF 3 : 1	**80.24%**	**0.84**	**0.87**
RoF 4 : 1	77.60%	0.84	0.88
RoF 5 : 1	74.05%	0.84	0.89

4.2. **Empirical results.** This section shows analysis from the tests made. Accuracy and AUC are standard measures within credit risk scoring and will also be implemented. Further, a paired means test will be applied to significantly testing the performance of our proposed method.

Accuracy and AUC. First, we have to compare the original dataset to that of the implemented SMOTE RoF model. From these models, RoF provided the greatest accuracy with a score of 80.24% and the greatest AUC with 0.84. Surprisingly the deep models performed somewhat bad despite their known ability in complex problem solving. Next, a t-test was performed to confirm the significance of our proposed methodology for credit risk assessment. Compared to the original dataset and cost-sensitive settings seen in [3] our methodology statistically outperforms their method. Next, t-tests with RoF as the baseline classifier show significance in the results against other implemented models. Thus, we can reject the null hypothesis of the methodology seen in [3], and also reject the other classification models implemented against RoF.

TABLE 2. Experimental results & t-test

		Original data		**SMOTE**	
		ACC	AUC	ACC	AUC
RoF		77.60%	0.67	**80.24%**	**0.84**
SVM		59.96%	0.65	64.90%	0.65
LG		64.95%	0.71	64.95%	0.71
DNN (4)		59.16%	0.71	61.36%	0.70
DNN (2)		65.75%	0.70	56.70%	0.71
t-test					
	Metric	**Mean**		**t**	**Sig. (2-tailed)**
Pair 1* RoF SMOTE – RoF**	Accuracy	2.63933		13.542	.000
	AUC	.17116		82.727	.000
Pair 2* RoF SMOTE – SVM SMOTE**	Accuracy	15.33963		95.988	.000
	AUC	.19609		126.835	.000
Pair 3* RoF SMOTE – LG SMOTE**	Accuracy	15.29157		91.418	.000
	AUC	.13952		76.808	.000
Pair 4* RoF SMOTE – DNN (4) SMOTE**	Accuracy	18.87949		27.064	.000
	AUC	.14300		109.051	.000
Pair 5* RoF SMOTE – DNN (2) SMOTE**	Accuracy	23.54743		8.738	.000
	AUC	.13869		126.049	.000

5. **Conclusion.** In this paper we have analyzed recent trends in social lending (P2P) using a classification methodology. Specifically, we implemented the use of Synthetic Minority Over-Sampling Technique (SMOTE) [11] in order to help reduce the burden of an imbalanced dataset and found greater results in the SMOTE model when applied alongside a Rotation Forest ensemble classifier. Also the use of a less strict cost-sensitive analysis also helped to improve the overall performance against the originally implemented amount seen in [3].

This research has introduced the idea of synthetically enhancing the minority class in the dataset within P2P credit risk analysis. Exploring a similar methodology when it comes to researching P2P lending could be fruitful research endeavor. Also exploring alternative methods in researching imbalances such as under sampling techniques like Random under-sampling or the One-Sided Selection technique is also a potential research area [20].

Limitations of this research can be seen in the fact that we only used data from 2016. Results from other years could help to validate whether this methodology is valid or not. Another limitation is that this methodology may be limited to the LC data only and therefore this model should be evaluated on another P2P lending dataset in order to validate these results.

REFERENCES

[1] H. Zhao, Y. Ge, Q. Liu, G. Wang, E. Chen and H. Zhang, P2P lending survey: Platforms, recent advances and prospects, *ACM Trans. Intelligent Systems and Technology (TIST)*, vol.8, no.6, 2017.
[2] 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, *Decision Support Systems*, vol.89, pp.113-122, 2016.
[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, 2015.
[4] R. Emekter, Y. Tu, B. Jirasakuldech and M. Lu, Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending, *Applied Economics*, vol.47, no.1, pp.54-70, 2015.
[5] Y. Guo, W. Zhou, C. Luo, C. Liu and H. Xiong, Instance-based credit risk assessment for investment decisions in P2P lending, *European Journal of Operational Research*, vol.249, no.2, pp.417-426, 2016.
[6] A. Byanjankar, M. Heikkilä and J. Mezei, Predicting credit risk in peer-to-peer lending: A neural network approach, *IEEE Symposium Series on Computational Intelligence*, pp.719-725, 2015.
[7] *LendingClub.com*, https://www.lendingclub.com/info/statistics.action, accessed on June 15th 2019.
[8] J. Abellán and J. G. Castellano, A comparative study on base classifiers in ensemble methods for credit scoring, *Expert Systems with Applications*, vol.73, pp.1-10, 2017.
[9] S. Lessmann, B. Baesens, H. V. Seow and L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research*, vol.247, no.1, pp.124-136, 2015.
[10] 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.3446-3453, 2012.
[11] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.321-357, 2002.
[12] J. Sun, J. Lang, H. Fujita and H. Li, Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates, *Information Sciences*, vol.425, pp.76-91, 2018.
[13] L. Zhang and W. Wang, A re-sampling method for class imbalance learning with credit data, *International Conference of Information Technology, Computer Engineering and Management Sciences*, vol.1, pp.393-397, 2011.
[14] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, *Proc. of the 5th Annual Workshop on Computational Learning Theory*, pp.144-152, 1992.
[15] H. Byun and S. W. Lee, A survey on pattern recognition applications of support vector machines, *International Journal of Pattern Recognition and Artificial Intelligence*, vol.17, no.3, pp.459-486, 2003.
[16] J. J. Rodriguez, L. I. Kuncheva and C. J. Alonso, Rotation forest: A new classifier ensemble method, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.28, no.10, pp.1619-1630, 2006.
[17] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature*, vol.521, pp.436-444, 2015.
[18] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.
[19] K. B. Schebesch and R. Stecking, Support vector machines for credit scoring: Extension to non-standard cases, in *Innovations in Classification, Data Science, and Information Systems*, Berlin, Heidelberg, Springer, 2005.
[20] V. García, A. I. Marqués and J. S. Sánchez, Improving risk predictions by preprocessing imbalanced credit data, *Proc. of the 19th International Conference on Neural Information Processing (ICONIP)*, pp.68-75, 2012.