Research on breach prediction for big data through hybrid ensemble learning and logistic regression

Xiaohan Zhang¹, Xu Chen²,*

¹School of Finance Southwestern University of Finance and Economics China
²School of Statistics Southwestern University of Finance and Economics China

*Corresponding author: 41833012@smail.swufe.edu.cn

Abstract. The fast-growing online lending sector faces the risk of borrowers defaulting. Network lending data are often unbalanced. On the basis of EasyEnsemble algorithm, we proposed a hybrid ensemble algorithm, which combines the advantages of Bagging and Boosting. By comparing the performance of different learning models, the experimental results show that the hybrid ensemble algorithm has good classification accuracy and generalization ability. Further, we find that social capital has a negative effect on borrower default rate. This paper provides a relatively reliable method for the credit risk assessment of online borrowers, and also provides a new idea for risk monitoring in the field of online lending.

Keywords: Default prediction; class-imbalance learning; hybrid ensemble algorithm; logistic regression.

1. Introduction
COVID-19 is the biggest black swan event of 2020, an international health emergency affecting more than 200 countries and regions around the world, with profound economic impacts. Under the influence of the epidemic, fintech ushered in excellent development opportunities, but also exposed its security shortcomings. This is a new challenge. We have noticed that the epidemic has exacerbated the rise of non-performing loan ratio, which in turn has pushed up the incidence of "lightning strikes" on online lending platforms. Due to the overdue loan of a large number of customers, the lending platform has insufficient cash flow, which leads to the withdrawal difficulties, severe runs and even bankruptcy.

P2P online lending platforms first appeared in 2015. Prosper, the most famous online peer-to-peer lending platform in the United States, was launched in February 2006 and opened its transaction data to the public in 2007. Its operating model is a financial model in which individuals provide small loans to other individuals through third-party platforms on the premise of charging a certain amount of interest. There are two types of customer groups: one is small and micro businesses, and the other is the general group with emergency capital needs. The most important measure of a platform's performance is the default rate. Default of customers will not only lead to varying degrees of capital losses for investors and lending platforms, but also cause investors to lose confidence in the platform, which increases the difficulty for customers to raise money on the platform and hinders the development of the platform. Therefore, we expect that advanced machine learning algorithms such as
Bagging and Boosting can be used to improve the default probability of customers and thus prevent systemic risks.

2. Hybrid ensemble learning for default risk

2.1. Overviews of ensemble learning

In 1979, Dasarathy and Sheela first proposed the idea of ensemble learning [1]. In 1990, Hansen and Salamon demonstrated an integrated model based on neural network [2], which has lower variance and better generalization ability. In the same year, Schapire [3] proved that weak classifiers can be combined into a strong classifier through Boosting method, which laid the theoretical foundation of ensemble learning. In 1991, Jacobs proposed a hybrid expert model [4]. In 1994, Wolpert proposed the stacking generalization model [5]. In 1995, Freund and Schapire proposed the AdaBoost algorithm [6], which is highly efficient and can be applied to practical problems very easily. In 1996, Breiman proposed a technique similar to Boosting, called Bagging [7]. In 2001, Breiman proposed a special Bagging algorithm, the random forest algorithm [8].

With the development of The Times, more ensemble learning algorithms are proposed, and the generalization ability of ensemble is usually much stronger than that of a single learner, which makes the ensemble approach very attractive.

2.2. Hybrid ensemble classification algorithm

Only a small number of customers overdue, the majority of customers can be in accordance with the provisions of the normal repayment. So, it is an unbalanced data set. In practical application, the cost of Type II error is much higher than that of Type I. In other words, the loss caused by misjudgment of a few categories is often higher than that of misjudgment of a large number of categories, and the loss caused by a small number of defaults is extremely high. However, many classification algorithms usually assume that the training data set is balanced. When dealing with unbalanced data sets, the performance of the traditional classification algorithm, which takes the overall classification accuracy as the learning goal, will decline [9].

We note that with the development of ensemble learning technology, more and more studies show that ensemble learning performs well on unbalanced dataset [10-12]. Due to the pros of ensemble learning in classifying unbalanced data, we use it for modeling and analyzing loan data.

Bagging and Boosting are the most widely used ensemble learning algorithms. In the Bagging algorithm, we randomly select a number of samples from the original training set to train the basic learner. After several rounds of training, a prediction function sequence can be obtained, and the final prediction function adopts voting method. Selecting different training sets for several times increases the difference degree of the base learner, thus improving the generalization ability of the final ensemble classifier. AdaBoost is the concrete implementation of Boosting algorithm. In AdaBoost, each training sample is given equal weight initially. After each training, the training samples that failed in training were given greater weight, that is, the learning algorithm was asked to pay more attention to the training samples that were difficult to classify in the subsequent learning. Similarly, after several rounds of training, a prediction function sequence can be obtained, and finally an ensemble classifier can be obtained through weighted voting. The pseudocode of the AdaBoost Algorithm is shown in Figure 1.
Base learning algorithm including AdaBoost; 
Number of learning rounds $T$.

**Input**: Data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;
Base learning algorithm $L$ including AdaBoost; 
Number of learning rounds $T$.

**Process**:

1. $D_0(i) = 1/m$. %Initialize the weight distribution
2. For $t = 1, 2, \ldots, T$:
   a. $h_t = L(D_t, D_t)$; %Train a base learner $h_t$ from distribution $D_t$
   b. $e_t = \Pr_{x,y} [h_t(x, y) = \neq y]$; 
   c. $a_t = \frac{1}{2}\ln\frac{1 - e_t}{e_t}$; %Calculate the weight parameter $a_t$
   d. $Z_t = \sum_s D_t(s)\exp(-a_t y h_t(x_s))$ %$Z_t$ is a normalization factor
   e. $D_{t+1}(i) = \frac{D_t(i)\exp(-a_t y h_t(x_i))}{Z_t}$ %Update the distribution $D_{t+1}$

3. **Output**: $H(x) = \text{sign}(f(x)) = \text{sign}\sum_{t=1}^{T}a_t h_t(x)$

**Fig. 1.** The AdaBoost algorithm

By using different data rebalance sampling strategies, a series of ensemble classification algorithms for handling unbalanced data are generated. Zhou [13] proposed the EasyEnsemble algorithm, which adopts Bagging as the main integration method, and each basic learner is trained by AdaBoost method. This is a kind of integration algorithm, which combines the advantages of AdaBoost to effectively reduce the bias and the characteristics of Bagging to effectively reduce the variance, thus improving the performance of the model to deal with unbalanced data.

We extend the algorithm based on EasyEnsemble. We introduced SMOTE over-sampling technology to deal with unbalanced data, and made the base learner lay more emphasis on positive samples by adding synthetic positive samples. Basic learners include AdaBoost, Decision Tree, Logistic Regression and KNN. Based on AdaBoost, other basic learners are added to improve the difference of the basic learners and the final classification discriminant index AUC. The pseudocode of the Hybrid Ensemble Algorithm is shown in Figure 2.

**Input**: Data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;
Base learning algorithm $L$ including LR, KNN, Decision Tree, AdaBoost; 
Number of learning rounds $T$.

**Process**:

1. $D_t = \text{Bootstrap}(D)$; %Generate a sample $D_t$ from $D$ by SMOTE
2. $h_t = L(D_t)$ %Train a base learner from the sample

3. **Output**: $H(x) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} I(y = h_t(x))$

**Fig. 2.** The Hybrid ensemble algorithm

### 3. Experiments

#### 3.1. Prosper dataset

We studied loan data published by Prosper between 2005 and 2014. This data set consists of 81 variables and 113,937 observations, including information of loan status, customer credit rating, customer occupation, customer monthly income, social capital and other dimensions. In the experiment, each data set was randomly divided into 70% of the training data set and 30% of the test data set. All subsets maintain the same category probability distribution as the original dataset.
3.2. Evaluation criteria

After processing the Prosper dataset, we get an unbalanced dataset with an imbalance ratio of 10:1. The performance of classifier cannot be measured by classification accuracy alone when learning unbalanced data. In this paper, we use G-mean, F-measure, and AUC as performance evaluation measures. These indices are widely used for comparison, and all indices are functions of the confusion matrix.

\[ \text{True Positive Rate (Acc.)} = \frac{TP}{TP+FN} \]
\[ \text{True Negative Rate (Acc.)} = \frac{TP}{TN+FP} \]
\[ \text{Precision} = \frac{TP}{TP+FP} \]
\[ \text{Recall} = \frac{TP}{TP+FN} \]
\[ \text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
\[ \text{G-mean} = \sqrt{\text{Acc.} \times \text{Acc.}} \]

ROC curve is a graphical method to show the tradeoff between true and false positivity in classification models. True positive ratio (TPR), also known as sensitivity, represents the ratio of the number of positive samples correctly classified to the number of positive samples, namely TP/(TP+FN); False-positive ratio (FPR), also known as specificity, is the ratio of the number of negative samples misclassified to the number of negative samples, i.e. FP/(FP+TN). For unbalanced data, the area under the ROC curve (AUC) is proved to be a reliable performance measure. The larger the AUC value, the better the model. The ideal model has an AUC of 1, while the random guess model has an AUC of 0.5.

3.3. Experimental settings

For the processed Prosper Dataset, we used the hybrid integration algorithm to test. Then we performed a tenfold Stratified Cross Validation to calculate Accuracy, F-measure, G-mean, AUC and other indicators by averaging. We compared the performance of the 7 methods, and the specific models used are as follows.

1) Logistic Regression (abbreviated as LR): It uses the entire data set to train a single classifier.
2) Random Forest (abbreviated as RF): It uses the entire data set to train a Bagging classifier. Decision Tree is used to train weak classifiers.
3) GBDT: Gradient Boosting Decision Tree. It is a kind of Boosting algorithm, which uses the entire data set.
4) XGBoost: Extreme Gradient Boosting. It is a kind of Boosting algorithm based on GBDT, which uses the entire data set.
5) lightGBM: Light Gradient Boosting Machine. It is a kind of Boosting algorithm based on GBDT, which uses the entire data set.
6) Bagging#1: It uses the entire data set to train a bagging classifier, which includes LR, KNN, Decision Tree. Both KNN and Decision Tree are weak classifiers.
7) Bagging#2: It uses the entire data set to train a bagging classifier, which includes LR, KNN, Decision Tree and AdaBoost. This model combines Bagging and Boosting.

3.4. Results and analyses

We tested the proposed methods on the dataset and the results are summarized in Table I. The AUC obtained by the logistic regression model was the lowest among the seven models, and even after variable screening, its AUC was still around 0.71. In the random forest model of Bagging's special algorithm, the AUC obtained is 0.04 higher than that of logistic regression. In Boosting algorithm, the AUC obtained by using GBDT, XGBoost and LightGBM can improve by about 0.03 compared with
random forest. Among them, the results of XGBoost and LightGBM are the evaluation indexes obtained by training after using the built-in parameter adjustment to get the balanced data set.

Whereas, compared with the optimal XGBoost among the single classifiers, the AUC predicted by Bagging#1 after bagging Logistic Regression, KNN and Decision Tree classifiers improved by about 0.12. However, the AUC of the test set obtained by the Bagging#2 model with AdaBoost added after training is 0.03 higher than Bagging#1, which is the best default prediction effect among the seven models. Meanwhile, its F-measure and G-mean are also much higher than other classifiers.

### Table 1 Index of Models

| Model Index | Logistic Regression | Random Forest | GBDT | XGBoost | lightGBM | Bagging#1 | Bagging#2 |
|-------------|---------------------|---------------|------|---------|----------|-----------|-----------|
| Accuracy    | 0.922               | 0.924         | 0.914| 0.923   | 0.922    | 0.904     | 0.934     |
| Precision   | 0.619               | 0.620         | 0.432| 0.433   | 0.590    | 0.930     | 0.936     |
| Recall      | 0.011               | 0.103         | 0.226| 0.114   | 0.104    | 0.874     | 0.933     |
| F-measure   | 0.021               | 0.177         | 0.296| 0.296   | 0.177    | 0.906     | 0.934     |
| G-mean      | 0.105               | 0.253         | 0.312| 0.222   | 0.248    | 0.902     | 0.934     |
| AUC         | 0.699               | 0.755         | 0.786| 0.787   | 0.773    | 0.904     | 0.934     |

#1-(LR, KNN, Decision Tree); #2-(LR, KNN, Decision Tree, AdaBoost)

We also depicted the ROC curve of the hybrid integration algorithm K-fold cross-validation, as shown in Figure 2. The data set was randomly divided into 6 folds, trained on 5 folds, and tested on the last fold. After 6 iterations of the training model, the average AUC of Bagging#2 model is 0.93, indicating that the model has better generalization ability. Except for the low AUC of the first segmentation, the AUC of the other 5 times were all stable at 0.96, but 0.8 was generally higher than the index of a single classifier. The variance of the model is small and the model has certain stability.

![Cross-Validation ROC of Bagging](image)

**Fig. 2.** ROC for K-fold cross validation of Bagging#2
3.5. Exploration of social capital

After completing the prediction of default risk, we hope to further explore the influence factors of customer default. In the study of online lending, Lin [14] regarded social capital as the information about risks provided by the customer's social network, which can alleviate the problem of information asymmetry in the network market. Based on this, we adopted the model of "social capital variable + hard information variable" to study the impact on default prediction. The designed model is:

\[
\ln \frac{P(\text{Default} = 1)}{P(\text{Default} = 0)} = \xi + \eta F_i + \lambda H_i + \delta S_i + \epsilon_i
\]

\(\text{Default}\) means whether the customer is in default; \(F\) represents borrowing characteristics; \(H\) represents characteristics; \(S\) represents social capital, including \(\text{Commend}, \text{Group}, \text{and Friends}\).

We put social capital into Logistic Regression model. The regression coefficients of these three equations are all negative, and the P-value of each model is less than 0.05, the results are all very statistically significant. The coefficients ranged from \(-0.1729\) to \(-0.3792\). Therefore, with other variables unchanged, when social capital increases by one unit, the ratio of occurrence of default events ranged from \(e^{-0.3792}\) to \(e^{-0.1729}\).

In addition, we describe the relationship between social and default samples as shown in Table II. We found that in default samples, the mean of \(\text{Commend}, \text{Group}, \text{and Friends}\) is lower than the sample has been completed. This shows borrowers join the group, increase the number of friends among investors, and increase the number of times their loan list is recommended, the probability of default decreases.

| Loan status | Commend mean | Group mean | Friends mean | Sample size |
|-------------|--------------|------------|--------------|-------------|
| Completed   | 0.0916       | 0.8427     | 0.0487       | 46857       |
| Defaulted   | 0.0846       | 0.6421     | 0.0310       | 4091        |
| Mean change | {}           | {}         | {}           | {}          |
| Total       | 0.0910       | 0.8266     | 0.0473       | 50948       |

We believe that under the group mechanism, customers in the group will generally choose to repay on time under the pressure of reborrowing and social punishment imposed by the group. Under the friend mechanism, customers will not choose to default under the pressure of reborrowing, supervision and social punishment imposed by friends. Under the referral mechanism, customers are under pressure from supervision and social punishment imposed by the group leader or friends, and will not default in order to get a chance to reborrow.

4. Conclusion

Default is a key concern in the operation of online lending platforms. Its essence is moral hazard and adverse selection caused by information asymmetry. It is of great significance to study the relationship between borrower characteristics and loan defaults, especially in the context of rising default rates during the epidemic period.

In this paper, a hybrid ensemble algorithm is proposed based on EasyEnsemble algorithm. We choose the transaction data of Prosper to build the early warning model of the overdue risk of network loan. We compared and analyzed the applicability of different classification methods for default prediction. On the basis of data cleaning and feature engineering, evaluation indexes (accuracy, F-Measure, AUC, etc.) were used to compare the performance of different learning models. The results showed that the hybrid ensemble algorithm had better warning ability than other learning models. This is related to the advantage that it can increase sample differentiation and solve the problem of data imbalance. The empirical results prove the effectiveness of the hybrid ensemble algorithm and its
better ability of risk screening. This provides a relatively reliable method for the credit risk assessment of online borrowers, and also provides a reference for the risk monitoring in the field of online lending.

Furthermore, through analysis and empirical test, we prove that the introduction of social capital can significantly reduce the default risk of borrowers. Joining the group, increasing the number of friends among investors or increasing the number of times the loan list is recommended can increase the social capital of the borrower, thus reducing the possibility of default. By increasing the mechanism design of customers' social capital, online lending platforms can have a more comprehensive understanding of customers' information and reduce the default risk of customers.

The limitation of this paper is that the data of only one platform is used as the research object in sample selection. Due to the difficulty in obtaining data of online lending platforms and certain differences in information disclosed by different platforms, it is difficult to unify the selection of variables. Follow-up studies can further consider the impact of economic development level on the existence of borrowers' defaults, and consider introducing variables reflecting the macroeconomic level. It can also expand the number of samples to build a more intelligent credit risk identification model, such as using deep learning algorithm for modeling.

References
[1] Dasarathy B V, Sheela B V. A composite classifier system design: Concepts and methodology[J]. Proceedings of the IEEE, 1979, 67(5):708-713.
[2] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990, 12(10):0-1001.
[3] Robert E, Schapire R E. The Strength of Weak Learnability [J]. Machine Learning,1990,5(2).
[4] Jacob J. Adaptive mixtures of local experts [J]. Neural Computation, 1991, 3.
[5] Wolpert D H. Stacked Generalization [J]. Neural Networks, 1992, 5(2):241-259.
[6] Freund Y, Schapire R E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting [J]. Journal of Computer & System Sciences, 1997, 55(1):119-139.
[7] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24:148-156.
[8] Breiman L. Random Forests[J]. Machine Learning, 2001.
[9] Chawla N, Japkovic N, Kotcz A, et al. Editorial: Special issues on learning from imbalanced data sets[J]. Annals of Nuclear Energy, 2004, 36(3):255-257.
[10] Sun Y, Kamel M S, Wong A, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12):3358-3378.
[11] Galar M, A Fernández, Barrenechea E, et al. EUSBoost: Enhancing ensembles for highly imbalanced datasets by evolutionary undersampling[J]. Pattern Recognition, 2013, 46(12):3460–3471.
[12] Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1):30.
[13] Liu X Y, Wu J, Zhou Z H. Exploratory Undersampling for Class-Imbalance Learning[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 39(2):539-550.
[14] Lin M, Prabhala N R, Viswanathan S, et al. Social Networks as Signaling Mechanisms: Evidence from Online Peer-to-Peer Lending. Wise, 2009