Ensemble of diverse deep neural networks with pseudo-labels for repayment prediction in social lending

Ji-Yoon Kim and Sung-Bae Cho
Department of Computer Science, Yonsei University, Seoul, Korea

Abstract
In peer-to-peer (P2P) social lending, it is important to predict the repayment of borrowers. P2P lending data are generated in real-time, but most of them are pending to decide the repayment because the deadline is not yet expired. Adding the unexpired data with appropriate labels into the training set could improve the performance of a prediction model, but the pseudo-labels cannot be certainly precise. In this paper, we propose an ensemble classifier composed of diverse convolutional neural networks (CNNs) of GoogLeNet, ResNet and DenseNet for the repayment prediction in social lending with the pseudo-labels approximated by an uncertainty handling scheme. The additional data labeled by Dempster-Shafer fusion of two semi-supervised learning methods boost up training of various models of CNNs, which are combined by weighted voting. A diversity measure is applied to constructing a pool of different models of CNNs that extract the effective features in the social lending data with labeling noise and predict the borrower’s loan status. The experiment with the real dataset of 855,502 cases from Lending Club confirms that the diverse ensemble combined with weighted voting achieves the highest performance and outperforms conventional methods.

Keywords
Convolutional neural networks, ensemble classifier, pseudo-label, social lending, FinTech

Introduction
Peer-to-Peer (P2P) social lending is one of the FinTech services that directly match borrowers with lenders.1 In P2P lending, the lenders lend money to the borrowers by selecting them directly. If the borrower fails to complete the repayment, the lenders will suffer a financial loss. It is important to predict repayment in order to reduce the financial risk of the lenders.2,3

Corresponding author:
Sung-Bae Cho, Department of Computer Science, Yonsei University, 50 Yonsei-ro, Seoul 03722, Korea.
Email: sbcho@yonsei.ac.kr
P2P lending platforms provide the borrower information such as demographic information, credit history, applied loan products, and current loan status to address information asymmetry or transparency issues since transactions occur online. The availability and prevalence of P2P social lending data have attracted many researchers. However, most of the data do not include whether the repayment will be completed because the repayment period has not expired. In the Lending Club dataset (https://www.lendingclub.com) for the past three years (2015–2017), the data that the repayment period has expired accounts for about 37% of the labeled data ("fully paid" and "charged off") and the data that the repayment period has not expired account for about 63% of the unlabeled data ("current", "in grace period", "late", and "default"). Here, the data is considered as unlabeled if we do not know the label at the time of constructing a prediction model. The detailed specification of the data used for this study will be given in “Dataset”.

The performance of the model can be improved by utilizing the unlabeled data for training. Many researchers employ the approach that trains the model as supervised learning with labeled and unlabeled data simultaneously by using the pseudo-labels, which selects the class with the highest expected probability for the unlabeled data, as the actual label in relation to the low-density separation assumption between classes in semi-supervised learning. We adopt this approach and design a model for repayment prediction using unlabeled and labeled social lending data. However, this approach implies a problem that the pseudo-labels of unlabeled data may have label noise.

On the other hand, convolutional neural networks (CNN) can provide predictive models for large and complex data and are known to be robust to label noise. The CNN automatically extracts features by stacking the convolutional layer and the pooling layer several times. Discriminative features can be captured by constructing architectures with different depths and widths. Recently, a model for combining convolutional neural networks (CNNs) in various fields has been proposed. The ensemble method is generally employed to improve the performance of the individual base classifiers. The strength of the ensemble method lies in its ability to correct errors caused by some of the members, such as class noise. Several researchers discussed that ensemble members should be diversified in terms of errors. Ensembles with members of high diversity tend to have higher prediction performance.

In this paper, we propose a heterogeneous ensemble composed of diverse CNNs for repayment prediction in social lending with pseudo-labels. We exploit two semi-supervised learning methods to generate pseudo-labels from the unlabeled data and predict the loan status of the borrowers using the ensemble model of the CNNs, where members of the ensemble are constructed based on a diversity measure and combined with weighted voting. Weighted voting also improves performance by weighing high-performance classifiers. The proposed ensemble model that consists of diverse classifiers extracts different features by using unlabeled data with pseudo-labels and predicts the loan status robustly regardless of label noise.

The Lending Club dataset has been used for experiments. We conduct the comparison with the individual classifiers and the homogeneous ensemble methods in order to evaluate the performance of the proposed method in the dataset with some pseudo-labels. The
performance of the ensemble model is presented with diversity measures. The main contributions of this paper are as follows.

- A novel method is proposed for predicting repayment in social lending by adopting the ensemble approach with deep learning models and achieving state-of-the-art performance compared with the conventional methods.
- The idea is confirmed by the real dataset of 855,502 cases from Lending Club to reflect the inherent complexity of the problem.
- We focus on finding a solution to the problem of societal sustainability to encourage economic growth with a new field of financial technology.

This paper is organized as follows. “Related works” reviews the studies related to social lending. “The Proposed Method” presents the structure and description of the proposed method. “Experiments” shows the experimental results to verify the usefulness of the proposed method. Finally, “Conclusions” concludes this paper and discusses future work.

**Related works**

Many researchers have attempted various approaches to predict repayment in P2P social lending. Table 1 shows the studies related to repayment prediction in social lending so far. Most studies use various feature engineering algorithms to improve the performance or predict repayment using a simple model or ensemble model.

Feature engineering has been widely used to improve the performance of repayment prediction models. Lin et al. proposed a credit risk assessment model using Yooli dataset.

**Table 1. Related studies in P2P social lending.**

| Author          | Method                                           | # instances | # features |
|-----------------|--------------------------------------------------|-------------|------------|
| Ma et al.¹⁹     | LightGBM and XGboost                             | 569,339     | 24         |
| Li²⁰            | Ensemble XGboost, logistic regression, deep neural network | 80,000      | 20         |
| Xia et al.²¹    | XGBoost                                          | 49,795      | 17         |
| Lin et al.²²    | Logistic regression                              | 48,784      | 10         |
| Zhang et al.²³  | Logistic regression                              | 193,614     | 21         |
| Chen²⁴          | Logistic regression                              | 3177        | 11         |
| Fu²⁵            | Combination of random forest and neural network  | 1,320,000   | 13         |
| Jiang et al.²⁶  | Logistic regression, naïve Bayes, support vector machine, random forest | 39,538      | 32         |
| Guo et al.²⁷    | Logistic regression                              | 2,016, 4128 | 6, 6       |
| Polena and Regner²⁸ | Regression                               | 70,673      | 14         |
| Serrano-Cinca and Gutiérrez-Nieto⁵ | Linear regression, decision tree | 40,907      | 26         |
P2P lending platform in China. They used the nonparametric statistical method to identify the borrowers’ demographic characteristics and extract the features that affect the default. A total of 10 variables including gender, age, marital status, and loan amount were extracted and a credit risk assessment model was designed using logistic regression analysis. Malekipirbazari et al. used the random forest to assess risk in social lending. Through sophisticated preprocessing, they extracted 15 features and evaluated the performance according to the number of features. As the number of features increases, a higher performance is achieved.

Recently, various ensemble methods have been introduced instead of a single model for predicting repayment. In particular, the random forest, the ensemble model of the decision trees, leads to high performance for repayment prediction. Fu predicted defaults by combining neural networks with random forests to capture non-linearity. Li proposed an ensemble model composed of heterogeneous classifiers in order to effectively predict the imbalanced data and achieved a high performance by weighted fusion of XGBoost, DNN, and logistic regression.

The approach to designing a predictive model that considers the characteristics of social lending data is rare. Existing feature engineering and statistical techniques are not suitable for managing the big data of social lending that are increasing every year, and ensemble models are favored to improve the performance. In this paper, we propose a novel approach to design a predictive model with CNNs adapted with the additional data that the repayment period has not expired to predict the repayment of borrowers.

The proposed method

Figure 1 shows the overall scheme of the proposed method. The unlabeled data that the repayment period has not expired is labeled as "fully paid" or "charged off" by two semi-supervised learning methods with different approaches, and the final label is assigned based on the Dempster-Shafer theory. Unlabeled data with pseudo-labels and labeled data are learned together in feature space by various CNNs. The feature space is used to train the classifier to model the features of the borrower. The social lending data is projected into the representation space learned by diverse base CNNs. In the base classifier pools with four different architectures of CNN, classifiers are selected using Q-statistic as a diversity measure. We combine the probability values from the various classifiers using weighted voting and finally predict the borrowers’ loan status.

Pseudo-labeling

Label propagation and transductive support vector machine (TSVM) assign classes for loan status to unlabeled data, respectively. We then combine the probability values for each class using Dempster-Shafer fusion. Figure 2 shows the pseudo-labeling process. Suppose that $X_L = \{x_1, x_2, \cdots, x_N\}$ and $X_U = \{x_1, x_2, \cdots, x_M\}$ be the labeled data and the unlabeled data of preprocessed social lending, respectively. Unlabeled data $X_U$ and labeled data $X_L$ are learned in the same representation space.
Label propagation. Label propagation is a method of labeling by propagating label information about repayment of the borrowers based on the similarity distance between observations.\(^{33}\) It is estimated based on cluster assumptions that closely located observations belong to the same class.\(^{35}\) Transition matrix \(T_{ij}\) for propagating class probabilities consist of the similarity between observations as follows:

\[
T_{ij} = \frac{W_{ij}}{\sum_{k=1}^{N+M} W_{kj}}
\]  

where \(W_{ij}\) is the weight of the edge between node \(i\) and \(j\), \(N\) is the number of labeled data, and \(M\) is the number of unlabeled data. The weight is calculated as shown in equation (2).
The label is assigned based on the \( k \) observations having high similarity with the class probability. In equation (2), \( x^d \) is the \( d \)-dimensional data and \( \sigma \) is the value specified by the user.

\[
w_{ij} = \exp\left(-\frac{\sum_{d=1}^{D} (x^d_i - x^d_j)^2}{\sigma^2}\right)
\] (2)

**Transductive support vector machine.** TSVM learns to increase the margin of decision boundaries as in the existing support vector machines\(^{34}\) using a large amount of unlabeled data. Salient features of the borrower are transferred by learning feature space considering both labeled and unlabeled data.

TSVM generates the initial classifier by performing inductive learning using the labeled data. It then uses the classifier to assign labels to unlabeled data, exchange classes of two observations and train the classifier until a pair of observations in each class with a slack value more than zero does not exist.

**Dempster-shafer fusion.** Dempster-Shafer’s theory is a probabilistic inference method in uncertain situations.\(^{31,32}\) Dempster-Shafer fusion (DSF) combines the class probabilities of two semi-supervised learning to carefully determine the classes for the loan status. This theory computes the class probabilities for all possible cases (\{"fully paid"\}, \{"charged off"\}, \{"fully paid", "charged off"\}) based on the power set of classes. DSF inforces more precise labeling by considering the case that two classes are mixed together.

\[
m^M(c) = \frac{p^M(c)}{\sum_{c' \in C} p^M(c')}
\] (3)

The belief \( m^M(c) \) for each case is computed using the class probability of the classifier as equation (3). \( p^M(c) \) means the probability of the loan status class \( c \) for the classifier \( M = \{LP, TSVM\} \). \( C = \{\emptyset, "unknown", "fully paid", "charged off"\} \) means a set of all classes. Class probabilities are combined based on Dempster’s rule as equations (4) and (5). Equation (4) means independent probability combination rule, and equation (5) shows a general combination rule.

\[
m^{\text{Fusion}}(C) = \sum_{C_1 \cap C_2 = C} m^{LP}(C_1) m^{TSVM}(C_2)
\] (4)

\[
m(c|x) = \frac{\sum_{C_1 \cap C_2 = C} m^{LP}(C_1|x) m^{TSVM}(C_2|x)}{1 - \sum_{C_1 \cap C_2 = \emptyset} m^{LP}(C_1|x) m^{TSVM}(C_2|x)}
\] (5)

If the maximum probability in the class probability calculated by this rule is "unknown", it means that the class is congested and the observation is removed. Also, if the probabilities of classes of "fully paid" and "charged off" are the same, the observation is removed. Filtering of observation points improves labeling performance.
**Ensemble of CNNs**

**Diverse CNNs.** This section presents the four models of CNN employed in the ensemble as a base classifier. Each model has a different network topology and extracts diversified features. Some models except plain CNN have a bottleneck layer as a $1 \times 1$ convolution, and ResNet$^{13}$ and DenseNet$^{14}$ use shortcut connections. Figure 3 shows the structure of each network employed in the ensemble.

**Convolutional Neural Network.** Unlabeled data with pseudo-labels and labeled data are used as input to the CNNs. CNNs extract discriminative features through a local connection in the borrowers and loan product information of a large amount of social lending data. CNNs perform convolution operations instead of matrix multiplication and generate output values from social lending data through the convolutional layer as the following equation (6).$^{36}$ $y_{i}^{l,j}$ is calculated by the output vector $x$ of the previous layer and the weight $w$. $l$ is the index of the layer, $j$ is the index of the feature-map, $K$ is the filter size, $k$ is the index of the filter, and $\sigma$ is the activation function. Activation functions often follow the convolutional layer and extract nonlinear features.

$$y_{i}^{l,j} = \sigma \left( \sum_{k=0}^{K} w_{i}^{l,j,k} x_{i+k+1} + b_{j} \right)$$  \hspace{1cm} (6)$$

$$f_{i}^{l,j} = \max_{r \in R} x_{i \times T+r}^{l,j}$$  \hspace{1cm} (7)$$

The pooling layer combines similar features extracted from the previous convolutional layer.$^{37}$ We use a pooling layer to extract the representative features from social lending data. One feature-map is distorted and reduced the dimensionality by calculating the value of the local patch. Equation (7) means the process of extracting the maximum value from the $l$th pooling layer. $R$ is a pooling size, and $T$ means the stride to move the pooling window.

---

**Figure 3.** The structures of base classifiers employed in the ensemble.
Several convolutional and pooling layers can be stacked up to play a role of feature extraction hierarchically. The feature-maps generated by repeating the convolutional layer and the pooling layer from the social lending data are arranged one-dimensionally through the fully connected layer, and the loan status is predicted through the softmax classifier.

The structure of the CNN is stacked twice with alternating convolutional and pooling layers. Followed by two fully connected layers and a dropout with a probability of 0.25 to prevent overfitting. Finally, the class probability for each loan status is output through the softmax classifier.

**GoogLeNet.** The key idea of GoogLeNet is to construct a sparse structure in the conventional CNN model to process the dense elements. Several inception modules are used to combine convolution operations with filters of various sizes in parallel. Using different size filters can extract features from various aspects of social lending data and form discriminative feature representation.

The inception module consists of $1 \times 1$ convolution, $1 \times 3$ convolution, $1 \times 5$ convolution, and $1 \times 2$ pooling layer. The $1 \times 1$ convolution used in GoogLeNet has a role to merge similar features from multiple feature maps and reduce the number of feature maps. The structure of GoogLeNet initially controls the number of feature maps with the convolutional layer and pooling layer, and two inception modules are applied. Followed by two fully connected layers, and the final probability is output through the softmax classifier.

**ResNet.** Residual network learns the difference between input and output and performs convolution operation with the value added to the input using shortcut connections. The general CNN considers the output of the $l$th layer as the input of the $(l + 1)$th layer as shown in equation (8), but ResNet introduces a shortcut connection to perform the identity mapping as shown in equation (9).

\[
x_{l+1} = f(x_l) \quad (8)
\]
\[
x_{l+1} = f(x_l, W_l) + x_l \quad (9)
\]

where $x$ represents the input. ResNet increases the abstraction ability and the representation power for social lending by repeatedly stacking several residual blocks to increase the depth. The ResNet structure consists of one convolutional layer to adjust the number of initial feature maps, one convolution block and three residual blocks. It is arranged in one dimension through a fully connected layer and outputs final probability through a softmax classifier.

**DenseNet.** It has direct connections to all subsequent layers as dense connectivity. The low-level and high-level features extracted in the social lending data are reused using dense connectivity. Suppose that the output through the $l$th layer from the social lending data is $x_l$, and the output of the $l$th layer is represented by equation (10).

\[
x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \quad (10)
\]

where $[x_0, x_1, \cdots, x_{l-1}]$ means the concatenation of the feature-map generated in the convolutional layer and $H_l(\cdot)$ is the composite function composed of the activation function and the $1 \times 3$ convolution.

The DenseNet initially passes through a convolutional layer and a pooling layer to resize the feature-map. The structure consists of two dense blocks and one transition
layer therebetween. The dense block includes a directly connected convolutional layer, and diversified features are extracted from the characteristics of the borrower or inherent property of the loan product in the social lending data. The transition layer for the down-sampling layer consists of a convolutional layer and a pooling layer, merging the representative features of social lending data. Each feature-map passes through two fully connected layers and outputs a probability value through the softmax classifier.

**Diversity and similarity.** One of the important steps in designing an ensemble model is ensemble selection, and it is also important to construct diverse classifiers. The selection criteria can be diversity, similarity, and data handling, which can be measured in different ways: Pairwise and global diversity measures. The pairwise diversity measure is computed for each pair of classifiers and averaged over two or more classifiers. The pairwise diversity is calculated for two repayment prediction algorithms from the relative amounts between the correct and incorrect predicted values. Table 2 shows the notation for the percentage of data instances in the classified dataset.

Here, $i$ and $j$ mean classifiers, $c_{ij}$ is the number of instances in which classifiers $i$ and $j$ are correct, while $c_{ij}$ is the number of incorrect instances. Similarly, $c_{ij}$ and $c_{ij}$ are the number of instances in which one classifier has correctly classified the instances while the other classifier has misclassified these instances.

Two pairwise diversity measures are introduced. Q-statistic $Q_{ij}$ and disagreement $D_{ij}$ are defined as follows:

$$Q_{ij} = \frac{c_{ij} \cdot c_{ij} - c_{ij} \cdot c_{ij}}{c_{ij} \cdot c_{ij} + c_{ij} \cdot c_{ij}}$$  \hspace{1cm} (11)

$$D_{ij} = \frac{c_{ij} + c_{ij}}{c_{ij} + c_{ij} + c_{ij} + c_{ij}}$$  \hspace{1cm} (12)

$Q$ has a value between $-1$ and $1$, and the closer the value is to zero, the greater the diversity is. $D$ is the most intuitive diversity factor, which means the proportion of two predictors disagree. The higher the value of $D$, the greater the diversity.

We introduce an entropy measure as a global diversity measure. The entropy measure assumes the most diverse when the correct member prediction is equal to $\lfloor L/2 \rfloor$ and the incorrect member prediction is equal to $\lfloor L - L/2 \rfloor$. Entropy $E$ is calculated as follows:

$$E = \frac{1}{N} \cdot \frac{2}{L - 1} \sum_{i=1}^{N} \min\{l(y_i), (L - l(y_i))\}$$  \hspace{1cm} (13)

| Table 2. The notation of the predicted values from two classifiers. |
|---------------------------------------------------------------|
| $C_i$ correct          | $C_i$ incorrect   |
| $C_j$ correct          | $C_j$            |
| $C_j$ incorrect        | $C_j$            |
where \( N \) is the number of instances, \( L \) is the number of classifiers, and \( l(y_i) \) is the number of classifiers from the correctly recognized instance \( y_i \).

The similarity is measured in the output of the classifier. Euclidean distance \( C = \{c_1, \ldots, c_M\} \), respectively, is calculated in pairs of instances of the validation set. Euclidean distance ranges from zero to a high number, with a smaller number meaning a similarity. A representative measure is cosine distance that is calculated using the inner product of the vector using the predictor of the classifier. The cosine similarity has a value from \(-1\) to 1, and the closer to 1, the more similar.

**Ensemble based on diversity using weighted voting.** This section describes the proposed ensemble method. The proposed ensemble algorithm consists of two parts. First, we select the classifier subsets to construct the final ensemble in the four base classifier pools using the diversity measure. Second, we combine selected classifiers using weighted voting. Figure 4 shows the overall procedure for the ensemble of base classifiers.

First, we set an arbitrary initial weight for locally connected edges of CNNs with four architectures using the training set. The backpropagation algorithm is used to perform training in a number of epochs and the weights of CNNs are adjusted. Then, we construct a homogenous ensemble in four base classifiers pool using \( Q \)-statistic as a pairwise diversity measure. The diversity \( Q \) is calculated for the validation set. The four models are combined by increasing the number of classifiers. The number of ensemble members is fixed when the \( Q \) value is closest to zero (when diversity is the greatest). We improve the prediction performance using the ensemble with a large diversity, and

![Figure 4. Overall procedure for the ensemble of CNNs.](image)
the ensemble with four different structures extract diverse features from social lending data.

Second, we combine the ensemble models of the four CNNs to determine the final prediction. Heterogeneous ensembles are more diverse and known to provide better results. It has also been shown that ensembles of heterogeneous models in credit scoring improve performance. In this paper, a weighted voting scheme is utilized, in which weights are assigned according to the classification accuracy over validation data. Weighted voting is applied to the output generated from the ensemble of CNNs. The ensemble method based on weighted voting generates the final output for the loan status, and the weight is determined by the accuracy \( p_i \) in the validation set. Weights are defined as follows:

\[
w_j \propto \log\left( \frac{p_i}{1 - p_i} \right)
\]

\[
class(x_i) = \arg\max\left( \sum_{j=1}^{M} v_{ij}w_j \right)
\]

where \( M \) is the number of classifiers. For the instance \( x_i \), if the ensemble model of the CNN sets its class label to \( y = \text{class}(x_i) \), the binary variable \( v_{ij} \) is set to 0 or 1. According to the voting rule, the class of \( x_i \) is predicted as equation (15).

**Experiments**

**Dataset**

In this paper, we use P2P social lending transaction data provided by Lending Club. The data consist of 111 attributes such as loan amount, amount of payment, loan period, and loan status as a predictor variable. Only the "Current" data is used for data that the repayment period has not expired, and "fully paid" and "charged off" are used for data that the repayment period has expired. A total of 855,502 data are collected from 2015 to 2016.

63 attributes and 679,596 data are used by excluding the attributes that cannot be used for prediction such as borrowers’ ID, URL, description of loans, attributes with more than 80% of missing values, duplicate attributes, and attributes that are filled after the borrower starts to repay. Table 3 shows the used variables.

CNN has an input format in the range \([0, 1]\). The 63 attributes used for prediction are preprocessed. Categorical variables are encoded as dummy variables, and continuous variables are min-max normalized by removing 1% of outliers.

\[
X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}
\]

The instances consist of 357,060 labeled data and 322,537 unlabeled data. The number of "fully paid" is 270,751, accounting for about 76% of the total labeled data. In the labeled data, the training set and the test set are split by 7:3, and 249,942 training set and 107,116 test set.
Results

In this section, before presenting other experimental results, we briefly describe the results of pseudo-labeling. Then, we present the performance of ensemble models according to diversity. Experiments show that the greater the diversity, the better the performance. In addition, we demonstrate the performance comparison with single classifiers, an ensemble of homogeneous classifiers, and an ensemble of heterogeneous classifiers. The ensemble model achieves higher performance, especially when constructed with heterogeneous classifiers, outperforming homogeneous ensembles.

We conduct LP, TSVM, and DSF using 322,536 unlabeled data and 249,942 labeled data. About 26% of pseudo-labels were labeled as "unknown," and about 94% except "unknown" was labeled "Fully Paid." Based on the pseudo-labeling results, the training set used for repayment prediction is 488,012 (249,942 labeled data + 238,070 unlabeled data with pseudo-labels). Table 4 shows the pseudo-labeling results.

We constructed an ensemble model based on a diversity measure. As new classifiers are added to the ensemble for four single models, we measure the diversity to check the contribution of the last classifier. If diversity increases, we construct an ensemble until the maximum size of the ensemble is reached. Figure 5 shows the relationship between diversity and performance according to the number of new classifiers in the validation set. The ensemble method employs weighted voting. CNN and DenseNet show that the greater the diversity, the better the performance. We combine the classifiers with the largest diversity.

| Type                | Name                                                                 | # variables |
|---------------------|----------------------------------------------------------------------|-------------|
| Borrower information| Loan status, Annual income, Employment length, Home ownership, Address state, zipcode | 5           |
| Loan characteristics| Loan amount, Term, Interest rate, Installment, Purpose, Verification status, Initial list status, DTI, Application type, Last credit pull day | 10          |
| Credit history      | Total current balance, Total bankcard limit, Account now delinquent, Revolving line utilization rate, Number of satisfactory accounts, and so on | 48          |

Table 4. The result of pseudo-labeling.

| Method  | Class   | LP     | TSVM   | DSF    |
|---------|---------|--------|--------|--------|
| Fully paid | 306,824 | 229,829 | 226,107 |
| Charged off | 15,713  | 92,708 | 11,963  |
| Unknown    | 0       | 0      | 84,467  |

Results

In this section, before presenting other experimental results, we briefly describe the results of pseudo-labeling. Then, we present the performance of ensemble models according to diversity. Experiments show that the greater the diversity, the better the performance. In addition, we demonstrate the performance comparison with single classifiers, an ensemble of homogeneous classifiers, and an ensemble of heterogeneous classifiers. The ensemble model achieves higher performance, especially when constructed with heterogeneous classifiers, outperforming homogeneous ensembles.

We conduct LP, TSVM, and DSF using 322,536 unlabeled data and 249,942 labeled data. About 26% of pseudo-labels were labeled as "unknown," and about 94% except "unknown" was labeled "Fully Paid." Based on the pseudo-labeling results, the training set used for repayment prediction is 488,012 (249,942 labeled data + 238,070 unlabeled data with pseudo-labels). Table 4 shows the pseudo-labeling results.

We constructed an ensemble model based on a diversity measure. As new classifiers are added to the ensemble for four single models, we measure the diversity to check the contribution of the last classifier. If diversity increases, we construct an ensemble until the maximum size of the ensemble is reached. Figure 5 shows the relationship between diversity and performance according to the number of new classifiers in the validation set. The ensemble method employs weighted voting. CNN and DenseNet show that the greater the diversity, the better the performance. We combine the classifiers with the largest diversity.

| Table 3. The list of used attributes. |
|--------------------------------------|
| Type                  | Name                                                                 | # variables |
| Predictor             | Loan status, Employment length, Home ownership, Address state, zipcode | 1           |
| Borrower information  | Annual income, Employment length, Home ownership, Address state, zipcode | 5           |
| Loan characteristics  | Loan amount, Term, Interest rate, Installment, Purpose, Verification status, Initial list status, DTI, Application type, Last credit pull day | 10          |
| Credit history        | Total current balance, Total bankcard limit, Account now delinquent, Revolving line utilization rate, Number of satisfactory accounts, and so on | 48          |

Results

In this section, before presenting other experimental results, we briefly describe the results of pseudo-labeling. Then, we present the performance of ensemble models according to diversity. Experiments show that the greater the diversity, the better the performance. In addition, we demonstrate the performance comparison with single classifiers, an ensemble of homogeneous classifiers, and an ensemble of heterogeneous classifiers. The ensemble model achieves higher performance, especially when constructed with heterogeneous classifiers, outperforming homogeneous ensembles.

We conduct LP, TSVM, and DSF using 322,536 unlabeled data and 249,942 labeled data. About 26% of pseudo-labels were labeled as "unknown," and about 94% except "unknown" was labeled "Fully Paid." Based on the pseudo-labeling results, the training set used for repayment prediction is 488,012 (249,942 labeled data + 238,070 unlabeled data with pseudo-labels). Table 4 shows the pseudo-labeling results.

We constructed an ensemble model based on a diversity measure. As new classifiers are added to the ensemble for four single models, we measure the diversity to check the contribution of the last classifier. If diversity increases, we construct an ensemble until the maximum size of the ensemble is reached. Figure 5 shows the relationship between diversity and performance according to the number of new classifiers in the validation set. The ensemble method employs weighted voting. CNN and DenseNet show that the greater the diversity, the better the performance. We combine the classifiers with the largest diversity.
The base classifiers for constructing the ensemble include 24 CNNs, 24 GoogLeNets, 6 ResNets, and 25 DenseNets. Table 5 shows the comparison of the performance with single models, ensemble models of the homogenous classifiers, and a combination model of four ensemble models using weighted voting. Here, \( n \) denotes the number of classifiers to be constructed. The proposed method achieves higher performance than the other models, except for the recall that it is the second-best followed by DenseNet. Even though the difference is slight, we need further analysis of the results because this kind of decision support system can be ethically controversial.
We measure the diversity, similarity, and performance according to a combination of heterogeneous ensembles using weighted voting in the validation set. Table 6 shows the performance for all combinations of heterogeneous ensembles. The darker the color, the greater the diversity, the lower the similarity, and the higher the performance. Experimental results show that the performance of the four heterogeneous models outperforms the other ensemble models and the combination of the models is diverse. The combination of CNN, GoogLeNet, and ResNet is the most diverse and the combination of ResNet and DenseNet are the most similar. The combination of the four models is presented as achieving higher performance than the combination of CNN, GoogLeNet, and ResNet models because DenseNet has achieved the highest performance among single models.

**Conclusions**

In this paper, we have presented CNNs with pseudo-label data for repayment prediction in P2P social lending and proposed an ensemble model which consists of diverse CNNs. Large amounts of unlabeled data are available in social lending. In order to improve the performance of the repayment prediction, diverse CNN models have demonstrated high performance compared with other methods by adding unlabeled data with pseudo-labels combining the two semi-supervised learning methods. The proposed heterogeneous ensemble model, which is composed of diverse classifiers, extracts additional features by adding the information of the borrowers whose repayment period has not expired and can help to select the borrower for future lenders.

The deep learning model has difficulty understanding the internal operation. Future studies should be conducted to explain the factors that have a significant impact on repayment prediction. Alternatively, an approach like case-based reasoning might be another option that frames the problem differently to predict a risk category to be associated with a loan request and addresses some of the implied ethical concerns.49 In addition,
the ensemble approach accompanies the overhead of computation for inference as well as training. We need to reduce the computational cost to deploy the method of practice. Another research issue is that social lending is traded online so we can use the additional dataset from social networks like Facebook and Twitter. We expect that combining borrowers’ social lending data with social network data will improve the performance of repayment prediction to encourage economic growth.

**Declaration of conflicting interests**
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub; No. 2022-0-00113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework).

**ORCID iD**
Sung-Bae Cho [https://orcid.org/0000-0002-7027-2429](https://orcid.org/0000-0002-7027-2429)

**Notes**
1. “Charged off” means that the lender is not payed back the loan in the deadline.  
2. “Default” is a very final state associated to money loss.  
3. See the table 4 for all the attributes in the dataset.  
4. Pseudo-label is predicted for unlabeled data by using the labeled data model. Usually, the dataset containing labels is used to train a model that generates pseudo-labels for the unlabeled dataset.

**References**
1. Basha SA, Elgammal MM and Abuzayed BM. Online peer-to-peer lending: a review of the literature. *Electron Commer Res Appl* 2021; 48: 19.  
2. Suryono RR, Purwandari B and Budi I. Peer to peer (P2P) lending problems and potential solutions: a systematic literature review. *Procedia Comput Sci* 2019; 161: 204–214.  
3. Chengeta K and Mabika ER. Peer to peer social lending default prediction with convolutional neural networks. In: *IEEE international conference on artificial intelligence, big data, computing and data communication systems*, 2021, p.10.  
4. Yan J, Yu W and Zhao JL. How signaling and search costs affect information asymmetry in P2P lending: the economics of big data. *Financial Innovation* 2015; 1: 19.  
5. Serrano-Cinca C and Gutiérrez-Nieto B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decis Support Syst* 2016; 89: 113–122.  
6. Grandvalet Y and Bengio Y. Entropy regularization. *Semi-supervised Learning*. 2006, pp.151–168.  
7. Lee DH. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. *Int. Conf. on Machine Learning Workshop* 2013; 3: 2.
8. Kim A and Cho SB. An ensemble semi-supervised learning method for predicting defaults in social lending. *Eng Appl Artif Intell* 2019; 81: 193–199.

9. Chawla NV. Learning from labeled and unlabeled data: an empirical study across techniques and domains. *J Artif Intell Res* 2005; 23: 331–366.

10. Rolnick D, Veit A, Belongie S, et al. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694, 2017.

11. Jan B, Farman H, Khan M, et al. Deep learning in big data analytics: a comparative study. *Comput Electr Eng* 2019; 75: 275–287.

12. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *IEEE conference on computer vision and pattern recognition*, 2015, pp.1–9.

13. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition*, 2016, pp.770–778.

14. Huang G, Liu Z, Matten L, et al. Densely connected convolutional networks. In: *IEEE conference on computer vision and pattern recognition*, 2017, pp.4700–4708.

15. Liu W, Luo Z and Li S. Improving deep ensemble vehicle classification by using selected adversarial samples. *Knowl Based Syst* 2018; 160: 167–175.

16. Dietterich TG. Ensemble learning. *The handbook of brain theory and neural networks*. Cambridge, MA, USA: MIT Press, 2002, pp.405–408.

17. Kuncheva LI, Whitaker CJ, Shipp CA, et al. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications* 2003; 6: 22–31.

18. Kim KJ and Cho SB. Evolutionary ensemble of diverse artificial neural networks using specialization. *Neurocomputing* 2008; 71: 1604–1618.

19. Ma X, Sha J, Wang D, et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning. *Electron Commer Res Appl* 2018; 31: 24–39.

20. Li W, Ding S, Chen Y, et al. Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *IEEE Access* 2018; 6: 54396–54406.

21. Xia Y, Liu C and Liu N. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electron Commer Res Appl* 2017; 24: 30–49.

22. Lin X, Li X and Zheng Z. Evaluating borrower’s default risk in peer-to-peer lending: evidence from a lending platform in China. *Appl Econ* 2017; 49: 3538–3545.

23. Zhang Y, Li H, Hai M, et al. Determinants of loan funded successful in online P2P lending. *Procedia Comput Sci* 2017; 122: 896–901.

24. Chen Y. Research on the credit risk assessment of Chinese online peer-to-peer lending borrower on logistic regression model. In: *DEStech transactions on environment, energy and Earth science*, 2017, pp.216–221.

25. Fu Y. Combination of random forests and neural networks in social lending. *Journal of Financial Risk Management* 2017; 06: 418–426.

26. Jiang C, Wang Z, Wang R, et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Ann Oper Res* 2018; 266: 511–529.

27. Guo Y, Zhou W, Luo C, et al. Instance-based credit risk assessment for investment decisions in P2P lending. *Eur J Oper Res* 2016; 249: 417–426.

28. Polena M and Regner T. Determinants of borrowers’ default in P2P lending under consideration of the loan risk class. *Jena Economic Research Papers* 2016; 23: 1–30.

29. Malekpirbazari M and Aksakalli V. Risk assessment in social lending via random forests. *Expert Syst Appl* 2015; 42: 4621–4631.

30. Chen CL and Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inf Sci (Ny)* 2014; 275: 314–347.
31. Dempster AP. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 1967; 38: 325–339.
32. Shafer G. *A mathematical theory of evidence*. Princeton, NJ, USA: Princeton University Press, 1976.
33. Zhu X and Ghahramani Z. Learning from labeled and unlabeled data with label propagation, 2002.
34. Joachims T. Transductive inference for text classification using support vector machines. In: *International conference on machine learning*, 1999, pp.200–209.
35. Blum A and Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: *International conference on machine learning*, 2001, pp.19–26.
36. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86: 2278–2324.
37. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
38. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929–1958.
39. Rokach L. Taxonomy for characterizing ensemble methods in classification tasks: a review and annotated bibliography. *Comput Stat Data Anal* 2009; 53: 4046–4072.
40. Augusto S, Filho L, Canuto AMP, et al. Investigating the impact of selection criteria in dynamic ensemble selection methods. *Expert Syst Appl* 2018; 106: 141–153.
41. Yule G and Pearson K VII. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philos Trans R Soc London Ser A* 1900; 194: 257–319.
42. Skalak DB. The sources of increased accuracy for two proposed boosting algorithms. In: *American association for artificial intelligence, integrating multiple learned models workshop*, vol. 1129, 1996, p.1133.
43. Kuncheva LI and Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 2003; 51: 181–207.
44. Hsu KW and Srivastava J. Diversity in combinations of heterogeneous classifiers. In: *Paciﬁc-Asia conference on knowledge discovery and data mining*, 2009, pp.923–932.
45. Gashler M, Christophe GC and Tony M. Decision tree ensemble: small heterogeneous is better than large homogeneous. In: *International conference on machine learning and applications*, 2008, pp.900–905.
46. Xia Y, Liu C, Da B, et al. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Syst Appl* 2018; 93: 182–199.
47. Zhou ZH. *Ensemble methods: foundations and algorithms*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2012.
48. Vinod JH, Natarajan S, Keerthana S, et al. Credit risk analysis in peer-to-peer lending system. In: *IEEE international conference on knowledge engineering and applications*, 2016, pp.193–196.
49. Uddin MJ, Vizzari G, Bandini S, et al. A case-based reasoning approach to rate microcredit borrower risk in online Kiva P2P lending model. *Data Technol Appl* 2018; 52: 58–83.

**Author biographies**

Ji-Yoon Kim received the MS degree in computer science from Yonsei University, Seoul, Korea, in 2019. Her research interests include probabilistic recognition models, and deep learning.

Sung-Bae Cho received the BS degree in computer science from Yonsei University, Seoul, Korea and the MS and PhD degrees in computer science from KAIST (Korea Advanced Institute of Science and Technology), Taejeon, Korea. He was an invited researcher of Human Information
Processing research laboratories at ATR (Advanced Telecommunications Research) institute, Kyoto, Japan from 1993 to 1995, and a visiting scholar at University of New South Wales, Canberra, Australia in 1998. He was also a visiting professor at University of British Columbia, Vancouver, Canada from 2005 to 2006, and at King Mongkut’s University of Technology at Thonburi, Bangkok, Thailand in 2013. Since 1995, he has been a professor in the Department of Computer Science, Yonsei University, and a Underwood distinguished professor from 2021. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life. He was the recipient of the Richard E. Merwin prize from the IEEE Computer Society in 1993. He received several Distinguished Investigator Awards from Korea Information Science Society in 2005, and Gaheon Sindoricoh in 2017. He is also a recipient of Service Merit Medal from Korean government in 2022.