J Inf Process Syst, Vol.10, No.2, pp.314~333, June 2014
http://dx.doi.org/10.3745/JIPS.04.0002

pISSN 1976-913X
eISSN 2092-805X

Default Prediction for Real Estate Companies with Imbalanced Dataset

Yuan-Xiang Dong *, Zhi Xiao*, and Xue Xiao**

Abstract—When analyzing default predictions in real estate companies, the number of non-defaulted cases always greatly exceeds the defaulted ones, which creates the two-class imbalance problem. This lowers the ability of prediction models to distinguish the default sample. In order to avoid this sample selection bias and to improve the prediction model, this paper applies a minority sample generation approach to create new minority samples. The logistic regression, support vector machine (SVM) classification, and neural network (NN) classification use an imbalanced dataset. They were used as benchmarks with a single prediction model that used a balanced dataset corrected by the minority samples generation approach. Instead of using prediction-oriented tests and the overall accuracy, the true positive rate (TPR), the true negative rate (TNR), G-mean, and F-score are used to measure the performance of default prediction models for imbalanced dataset. In this paper, we describe an empirical experiment that used a sampling of 14 default and 315 non-default listed real estate companies in China and report that most results using single prediction models with a balanced dataset generated better results than an imbalanced dataset.

Keywords—Default prediction, Imbalanced dataset, Real estate listed companies, Minority-sample generation approach

1. INTRODUCTION

The fluctuation in the real estate market in 2008 caused a financial crisis that seriously impacted the economies of many countries. Evaluating the real estate default rate is an important issue to avoid credit risks being made by financial institutions. The default of a real estate company will have a serious financial impact on other companies involved with that industry. Thus, predicting the default rate of real estate companies has become an important research topic in recent years [1-3].

There is abundant amount of literature on corporate default in all sectors [1-4]. However, different industries face different levels of competition. They follow different accounting practices [5] and have different characteristics [6]. Therefore, the default prediction models for all sectors tend to be too general and may not adequately address the real estate industry. Few

※ The authors are extremely grateful to the anonymous referees and Professor Gholamreza Torkzadeh for their insightful comments and valuable suggestions, which greatly improved the quality of this paper. This paper was supported by the National Natural Science Foundation of China (Grant No. 71171209) and by the Major Consulting Research Project of the Chinese Academy of Engineering (Grant No. 2012-ZD-15).
Manuscript received April 17, 2013; accepted September 20, 2013.
Corresponding Author: Zhi Xiao (xiaozhicqu@163.com)
* School of Economics and Business Administration, Chongqing University, Chongqing, PR China (fox_dyx@163.com, xiaozhicqu@163.com)
** Department of Real Estate, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore (sherlyxx1990725@sina.com)

Copyright ⓒ 2014 KIPS

researchers have made the efforts to develop default prediction models specifically for real estate [7-10].

Most previous studies use traditional prediction models such as the KMV (Moody's KMV) approach, Multivariate Discriminant Analysis (MDA), the Z-Score model, and the Logistic model. However, the frames of these models are assumed and require corresponding data that is not available in places like China. The parameters in these models are likely to need periodical adjustment due to the changes in the economic environment and market trends. Machine learning models provide an appropriate model to work with non-linear pattern classification for predicting default rates. Machine learning models have been used to predict financial distress situations in general cases [11-14]. Furthermore, most studies on default prediction for real estate companies are based on imbalanced datasets. For example, Patel and Vlamis [7] collected 11 insolvent and 101 solvent real estate companies listed on the London Stock Exchange (LSE) between 1980 and 2001. Patel and Pereira [8] used a sample of 52 real estate companies listed on the LSE, of which only about 15% were insolvent. Such imbalances in the dataset is called a two-class imbalance [15]. Machine learning models tend to be overwhelmed by imbalanced large solvent cases versus insolvent cases. There is a need for a method to reliably predict the default rate for small insolvent real estate firms.

The main purpose of this paper is to build a framework for conducting a default prediction for real estate companies. The logistic regression, support vector machines (SVM) classification, and neural networks (NN) classification use an imbalanced dataset. They were used as benchmarks with a single prediction model that used a balanced dataset corrected by the minority samples generation approach. Two separate criteria (TPR and TNR) were used to measure the prediction power of default and non-default samples. Then, two comprehensive criteria (G-mean and F-score) established by TPR and TNR were used to measure the performance of default prediction models.

The remainder of this paper is laid out as follows: Section 2 provides a review of default prediction studies. Section 3 gives a brief description of logistic regression, neural networks, and support vector machines. Section 4 describes the data and analyzes variable selection. Section 5 compares the empirical results using an imbalanced dataset with a balanced dataset that was corrected by the minority sample generation approach. The conclusion and discussion of future work are presented in Section 6.

2. Review of Previous Studies

Predicting defaults and business failures have been a major preoccupation of researchers and practitioners for a long time, and the approaches they have used are becoming more and more deeply. Beaver [16] was the first to apply a univariate model on financial ratios to predict corporate bankruptcy. However, his analysis is very simple in that it is based on one financial ratio at a time and uses a cutoff threshold for each ratio. Under the assumption that the two classes have Gaussian distributions with equal covariance matrices, Altman [17] proposed the method of multiple discriminant analysis (MDA) based on applying the Bayes classification procedure to perform bankruptcy predictions. However, MDA has been widely criticized because the validity of its results hinges on restrictive assumptions [12,18]. Later, Ohlson [19] introduced a logistic regression (Logit) model to predict financial distress. Zmijewski [20]

proposed a new financial distress prediction method of Probit.

In recent decades, various machine learning models were used to predict default and financial distress. Frydman et al. [11] used classification trees for financial distress prediction. Odom and Sharda [12] were two of the first to apply a neural network (NN) approach to the bankruptcy prediction problem. They used Altman's financial ratios as inputs to the NN, as well as to the MDA, as a way to compare 128 solvent and insolvent US firms. By contrasting a multilayer perceptron neural network with linear discriminant analysis for a set of firms labeled viable or distressed, Coats and Fant [21] found that the neural network is more accurate than linear discriminant analysis, especially in financial distress. Ahn et al. [13] established a hybrid intelligent system that predicts bankruptcy by combining the rough set approach with a neural network. They compared the prediction performance with traditional discriminant analysis and the neural network approach. Hardle et al. [14] used the smooth support vector machine (SSVM) to predict the default risk of companies, and investigated how the factors, including the oversampling and the selection of appropriate accounting ratios (predictors), the length of the training period, and structure of the training sample, affect the precision of a prediction.

Machine learning models show their best performance when they are used to predict the default rate for all sectors with balanced datasets. When dealing with the imbalanced dataset, Machine learning models are known to have a shortage. It shows a good classification rate for the majority class, but has an unsatisfactory classification rate for the minority class. It is worth noting that these datasets for real estate companies are always imbalanced. Since the minority class, which represents the firms with a high probability of default, is more important in practice, the minority instances should be paid more attention to. Many researchers have worked to improve the performance of the minority class and to solve the imbalance problem [22-24].

The approaches which deal with the imbalanced problem include imbalanced learning algorithms and re-sampling methods [25]. The former is a process that modifies the method by assigning distinct costs to the classification errors or by recognizing the classification result. However, there is a limit to this approach, as it is not valid for straightforward dealings with other existing classifiers. The latter can be integrated with classifiers by directly adjusting initial data, rather than modifying the learning algorithm. Re-sampling methods can be divided into two classes: the under-sampling method and the over-sampling method. Under-sampling the majority class removes some majority samples until the classes are approximately equally represented. Zmijewski [20] suggested that the sampling procedure, such as the under-sampling method, will lead to a choice-based sample bias, because the financial distress attribute of a firm determines the quantity of the samples. Better stated, a majority sample (a non-defaulted firm) is less likely to be selected. Yet, a minority sample（a defaulted firm）has a greater chance of being part of a sample set. Therefore, due to purposefully selecting samples, the under-sample will be biased, and the resulting predictions might be unreliable. On the contrary, over-sampling helps to increase the sensitivity of a classifier to the minority class by randomly duplicating instances into the minority class. Nonetheless, the simple replication may over-emphasize the noise examples in the region that combined with mostly negative samples and several minority samples. It will lead the learning algorithm to learn more and more specific regions of the minority class. This phenomenon is the so-called overfitting problem. For this reason, Chawla et al. [22] proposed the synthetic minority over-sampling technique (SMOTE), which has received wide acceptance. Based on the same theory of generating synthetic examples, Li and Sun [26] described a new over-sampling approach (MSGA-RPD-NN) to create new minority samples by

generating synthetic instances in a feature space instead of in a data space. The technique for saving an imbalanced problem has proven very effective in forecasting business failures. Therefore, to avoid overfitting, we chose MSGA-RPD-NN, rather than ordinary over-sampling, which simply appends replicated instances. To take choice bias, which is caused by initial sampling, into consideration, this paper adopted all firm-years sample available during the sample period to the default prediction models. Furthermore, considering that initial sampling causes choice bias, as with many recent studies [27-29], all the firm-years available during the sample period were applied to the default prediction models for real estate companies.

3. DESCRIPTION OF THE MODELS

This section demonstrates a brief summary of logistic regression, support vector machine (SVM), and Neural Networks (NN). Then, a brief introduction about the minority-samples generation approach based on a random percentage distance to the nearest neighbor (MSGA-RPD-NN) is given. In the end, this section describes the default prediction process based on an imbalanced dataset.

3.1 Prediction models

Logistic Regression

Given a training set of N data points $D = \{(x_i, y_i)\}_{i=1}^{N}$, with input data $x_i \in R^n$ and corresponding binary class labels $y_i \in \{0,1\}$, the logistic regression approach to classification tries to estimate the probability $P(y=1|x)$ is as follows:

$$P(y=1|x) = \frac{1}{1+\exp(w_0 + w^T x)} \tag{1}$$

where $x \in R^n$ is an n-dimensional input vector, w is the parameter vector, and the scalar w_0 is the intercept. The parameters w_0 and w are then typically estimated using the maximum likelihood procedure.

Neural Networks

A Neural network, which is an information processing paradigm inspired by the biological nervous system, works just like the brain processes information. It includes the input layer, hidden layer, and output layer (Fig. 1). It is composed of a large number of processing elements that are interconnected with unidirectional signal channels called connections. The nodes/neurons of the input layer are the feature values of an instance, and the output nodes/neurons represent a discriminator between this class and all of the other classes. The goal of the training process was to find one model from the set of allowed models that minimizes some of the overall error measures, such as the sum of squared errors (SSEs) and mean squared errors (MSEs). Hence, the network training is actually for minimizing specific error measures. There are numerous algorithms available for training neural network models, and most of them can be viewed as a straightforward application of the optimization theory and statistical estimation. In this study, we applied a three-layer neural network to our default prediction model.

First, we chose the initial values of the parameters of the network (i.e., the connection weights and the neuron residual error values). Second, the financial ratios of every validation were selected as inputs, and the default rates were selected as the outputs, lying in the range [0, 1]. Third, the network was trained and tested.

Fig. 1. Overview of neural networks

Support Vector Regression

Support Vector Machines are binary classifiers and their main purpose is to find a separating hyperplane with a margin as large as possible to minimize classification errors. Although the SVM classifier is binary, it is not directly suitable for the prediction of default rates. Since we are focusing on the default rate in this paper, a simple method is to use SVM regression to directly build a forecast model.

Support vector regression (SVR) is a regression technique that utilizes kernel functions. This subsection briefly introduces SVR, which performs nonlinear mapping to forecast the default rate.

SVR is expressed formally, as follows: given a training set $(\mathbf{x}_i, y_i), i = 1, 2, \cdots, m$, where the input variable $\mathbf{x}_i \in R^n$ is a n-dimensional vector, and the response variable $y_i \in R$ are continuous values. SVR builds the linear regression function as demonstrated by the following equation:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b \tag{2}$$

Based on the Vapnik's linear ε -Insensitivity loss (error) function [see Eq. (2)], the linear regression $f(\mathbf{x}, \mathbf{w})$ is estimated by simultaneously minimizing $\|\mathbf{w}\|^2$ and the sum of the linear ε -Insensitivity losses [as shown in Eq. (3)]. The constant C , which influences a trade-off between an approximation error and the weights vector norm $\|\mathbf{w}\|$, is a design parameter chosen by the user.

$$\left|y_i - f(\mathbf{x}_i, \mathbf{w}, b)\right| = \begin{cases} 0, \text{if} \left|y_i - f(\mathbf{x}_i, \mathbf{w}, b)\right| < \varepsilon \\ \left|y_i - f(\mathbf{x}_i, \mathbf{w}, b)\right| - \varepsilon, \text{others} \end{cases} \tag{3}$$

Minimizing the following risk:

$$\min T(\mathbf{w}, \xi^{(*)}, \varepsilon) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{l}\sum_{i=1}^{l}\left|y_i - f(\mathbf{x}_i, \mathbf{w})\right|_{\varepsilon} \tag{4}$$

Under constraints:

$$\begin{aligned} &(\mathbf{w}^T\phi(\mathbf{x}_i) + b) - y_i \le \varepsilon + \xi_i, \\ &y_i - (\mathbf{w}^T\phi(\mathbf{x}_i) + b) \le \varepsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \ge 0, i = 1, \cdots, l, \varepsilon \ge 0 \end{aligned} \tag{5}$$

Where ξ_i and ξ_i^* are slack variables, one is for exceeding the target value by more than ε and the other is for being more than ε below the target.

3.2 Minority-sample generation approach

This paper integrated the minority-sample generation approach with classifiers by directly adjusting initial data rather than by modifying the learning algorithm. To avoid overfitting, we chose MSGA-RPD-NN, which generates new minority samples in a feature space, instead of over-sampling by simply appending replicated instances as usual.

The algorithm of MSGA-RPD-NN is presented as described below.

Step 1. Separate the actual real estate companies' dataset into defaulted (minority) and non-defaulted (majority) samples.

Step 2. Input the original defaulted sample set, $num - maj$ represents the number of non-defaulted samples, and $num - min$ represents the number of defaulted samples.

Step 3. Construct array $Sample[][]$ to keep original defaulted samples and new synthetic samples.

Step 4. Randomly choose a defaulted sample, rs , from the defaulted sample set. Compute the Manhattan distance between rs and all other defaulted samples. Then set the sample nn with the minimum Manhattan distance as the nearest neighbor for rs , and use $dif(rs, nn)$ to keep the distance.

Therefore, the $dif(rs, nn)$ can be calculated as follows:

$$dif(rs, nn) = \sum_{i=1}^{n} w_i \times \left|f_i(rs) - f_i(nn)\right| \tag{6}$$

Where w_i is the weight of the i th financial ratio, and $f_i(rs)$ and $f_i(nn)$ represents the values of the rs and nn company on the i th financial ratio. In this paper, the weight of each financial ratio is set equally after feature selection.

Step 5. Choose a random number between 0 and 1 and call it gap .

Step 6. Generate a new defaulted sample, ns , and add ns into the array $Sample[][]$. Add one to the number of defaulted samples, $num - min$.

The value of new defaulted sample on the i th financial ratio can be calculated by the following formula:

$$f_i(ns) = f_i(nn) + gap \times dif \tag{7}$$

Then the new defaulted sample can be synthesized as follows:

$$Sample[ns][financial\ ratio] = Sample[ns][financial\ ratio] + gap \times dif \tag{8}$$

Step 7. Repeat Steps 4 through 6 until $num - min$ equal $num - maj$. Finally, generating new minority samples in the feature space forms a balanced dataset about real estate companies.

3.3 Prediction process

The research process for default prediction will now be presented. To prevent the sample choice bias of the traditional sample method, this paper used all of the firm-year data that was available during the sample period. We first corrected the imbalance samples to improve the performance of the prediction model, and then we input the financial ratios into each prediction model to predict the default rate. Finally, we made decisions according to the outputs of single predictions based on the training samples.

More specifically, the steps of default prediction as based on the proposed method are described as explained below.

Step 1. Normalize the data.

The data obtained from the real world usually differs from each other in unit and scale due to the criteria. Therefore, it is of great importance to normalize the data to eliminate the difference before applying the data to single prediction. The function of normalization is defined as follows:

$$f_{ij}' = \frac{f_{ij} - \min_i}{\max_i - \min_i} \tag{9}$$

where f_{ij} means the i th financial ratio values for the j th company. $\min_i$ and $\max_i$ represent the maximal value and minimal value of the i th financial ratio cross all companies, respectively

Step 2. Determine whether a dataset belongs in an imbalanced dataset or not.

Before using a forecasting method such as logistic regression, NN, and SVM to predict the probability of default, we needed to calculate the ratio of the majority against the minority in order to determine the properties of the dataset. The ratio can be calculated as follows:

$$Ratio = \frac{num - maj}{num - min} \tag{10}$$

If the ratio was less than 1.2, the dataset is balanced. In that case, the user could skip Step 3

and proceed directly to Step 4. If the ratio is more than 1.2, the dataset is an imbalanced dataset, and the user should go to Step 3

Step 3. Create a balanced dataset.

MSGA-RPD-NN is employed to generate new minority samples so that a balanced dataset is produced. With the MSGA-RPD-NN, the difference between the number of the new minority samples and the number of the majority samples is less than 5%.

Step 4. Predict the default probabilities.

The logistic regression, NN, and SVM methods were used as forecasting techniques to evaluate the default rate of real estate companies. Let y represents the output of default rate prediction.

Step 5. Make decisions.

A criterion needs to be set to diagnose whether or not the real estate companies will default. Suppose CR is the criteria.

If $y \geq CR$, we say that the j th real estate company is in default.

If $y \leq CR$, we say that the j th real estate company is not in default.

In this paper, we set 0.5 as the criterion that diagnoses whether one company is in default or not.

4. Application of the Minority-Sample Generation Apeoaxh in the Real Estate Industry

4.1 Data

Default is defined as the nonpayment of any scheduled payment, interest, or principal. However, it is very difficult to observe such a default for listed companies according to the available public data in China. On account of there being a short time horizon of credit data, a comprehensive credit information database has not been set up. Moreover, there have been few existing credit data announced. Up until now, none of listed real estate companies have gone bankrupt or been delisted. Given these reasons, the criteria applied in those previous studies [7,29] to recognize the defaulted firms may not be suitable for the default prediction of Chinese listed real estate companies. In this paper, the defaulted firms are identified as companies that are given special treatment (ST) by the China Securities Regulatory Commission (CSRC). A company will be treated as special if it has had a negative net profit continuously in the two consecutive recent years or if it has purposely published financial statements with serious misstatements. In our study, we consider a ST company as being a company that has had a negative net profit in two consecutive recent years. According to the benchmark for whether a listed real estate company has been given special treatment (ST), we categorized listed real estate companies into two classes: default and non-default companies. If a company is treated as special, it is prone to defaults and vice versa.

Data was collected from the China Stock Market and Accounting Research Database (CSMAR). This paper focuses on the real estate companies with December fiscal year-ends, which were chosen from *The Listed Company Industry Classification Guidance* published in 2001 by China Securities Regulatory Commission, with codes between *J01* and *J09*. To make the adjustment in the category easier, these codes take the jumping –like encoding pattern. The

sample real estate companies include three categories:

J01: real estate development industries

J05: real estate management industries

J09: real estate intermediary services industries

Given the significant differences of current fund employment rates and loan amounts, the companies engaged in real estate development are riskier than those engaged in real estate management and intermediary services. The number of real estate development companies makes up a large proportion of all the companies engaged. Therefore, we picked 107 developers, and ignored the samples in real estate management and intermediary services industries. In order to ensure that input variables clearly explained the financial characteristics of the listed companies, the developers, which stayed in the real estate industry less than two years, were removed from the sample. To avoid choice bias, this paper used every firm-year in which the data was available in our analysis. By eliminating the sample companies in case of missing financial ratios data, the final combined sample consists of 329 firm-year observations between 2005 and 2009, including 14 default and 315 non-default samples, from 107 individual real estate companies.

In this paper, we followed the approach of selecting the cross section date proposed by Su and Li [30,31]. Their work used data in the year (*t-2*), which represents two years before, to predict financial distress in the year (*t-0*).

4.2 Input variables selection

The first stage in deriving a financial ratio default prediction model is selecting the financial variables. This paper selects 34 initial financial ratios in Table 1.

These initial variables are selected for the following reasons: first, most of the variables are commonly used in previous studies (Altman [17], Tam and Kiang [32] and Xiao et al. [4]). Second, based on the CSMAR Database, the initial variables encompass a broad cross-section of accounting ratios. These ratios describe a developer's enterprise survival, development, and profitability. As a group, these ratios capture the finical characteristics and performance of the real estate industry.

Table 1. Definition of variables

Variable	Index
	Short-Term Liquidity
1	Current ratio
2	Quick ratio
3	Liquidity ratio
4	Working capital/ Total assets
	Operational Capacity
5	Account receivable turnover
6	Inventories turnover
7	Account payable turnover
8	Working capital turnover

9	Current assets turnover
10	Fixed assets turnover
11	Total assets turnover
	Long-Term Solvency
12	Total debt/Total assets
13	Long-term asset-liabilities ratio
14	Long term debt/Total assets
15	Equity to liability ratio
16	Owner's equity ratio
17	Current assets/Current liabilities
18	Fixed assets ratio
19	Current liabilities/ Total liabilities
	Profitability
20	Operating profit margin
21	Rate of return on total assets
22	Profit margin on net assets
23	Return on invested capital
	Risk Level
24	Comprehensive leverage
	Shareholders profitability
25	Earnings per share
26	Net asset value per share
27	Price to book ratio
28	Business income per share
	Cash Flow Ability
29	Cash flow ratio
30	Operating income per share
31	Net operating cash flow per share
32	Net cash flow per share
	Development
33	Capital maintenance and appreciation
34	Total assets growth rate

According to the prediction process, MSGA-RPD-NN was used to generate 286 default samples to obtain a balanced dataset. In order to minimize the influence of the variability of the training set, 10-fold cross-validation was applied 10 times to Chinese real estate listed companies datasets, including 329 firm-year observations. More specifically, the dataset was partitioned into 10 subsets with similar sizes and numbers. Then, the union of 9 subsets was used as the training set while the remaining subset was used as the test set, which was repeated 10 times so that every subset was used as the test set once. The imbalanced dataset and

generated balanced dataset is summarized in Table 2.

Table 2. Number of default and non-default samples, respectively, in balanced and imbalanced datasets using 10-fold cross-validation

Dataset	Validation	Number of Default Samples in a Training Set	Number of Non-Default Samples in a Training Set	Number of Default Samples in a Training Set	Number of Non-Default Samples in a Training Set
Imbalanced Dataset	1/2/5/7/9	13	283	1	32
	3/4/6/10	12	284	2	31
	8	13	284	1	31
Balanced Dataset	1/2/5/7/9	270	283	30	32
	3/4/6/8/10	270	284	30	31

The T-test and stepwise logistic regression were used to reduce features from 34 initial financial ratios in Table 1 using the software SPSS 16.0 based on training datasets. We first used the t-test to remove some features on the significance at 5%. Then we utilized the stepwise logistic regression to further reduce the remaining features. The selected features of 10-fold cross-validation using initial balanced and imbalanced datasets, which are listed in Table 3.

Table 3. Selected features

Validation	Features of **10-fold cross-validation** using initial imbalanced datasets	Features of **10-fold cross-validation** using balanced datasets
1	Quick ratio (2) Current assets/Current liabilities (17) Profit margin on net assets (22) Earnings per share (25) Net asset value per share (26)	Quick ratio (2) Liquidity ratio (3) Account receivable turnover (5) Current assets/Current liabilities (17) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Earnings per share (25) Price to book ratio (27) Capital maintenance and appreciation (33) Total assets growth rate (34)
2	Equity to liability ratio (15) Rate of return on total assets (21) Profit margin on net assets (22)	Quick ratio (2) Account receivable turnover (5) Owner's equity ratio (16) Fixed assets ratio (18) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Net asset value per share (26) Capital maintenance and appreciation (33)
3	Long-term debt/Total assets (14) Current assets/Current liabilities (17) Price to book ratio (27) Business income per share (28)	Quick ratio (2) Account receivable turnover (5) Long-term debt/Total assets (14) Current assets/Current liabilities (17) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Price to book ratio (27) Capital maintenance and appreciation (33)

4	Quick ratio (2) Current assets/Current liabilities (17) Rate of return on total assets (21) Profit margin on net assets (22) Net asset value per share (26)	Current ratio (1) Quick ratio (2) Long-term debt/Total assets (14) Current assets/Current liabilities (17) Profit margin on net assets (22) Earnings per share (25) Price to book ratio (27) Business income per share (28) Capital maintenance and appreciation (33)
5	Long-term debt/Total assets (14) Equity to liability ratio (15) Rate of return on total assets (21) Net asset value per share (26)	R Quick ratio (2) Owner's equity ratio (16) Fixed assets ratio (18) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Net asset value per share (26) Capital maintenance and appreciation (33)
6	Long-term debt/Total assets (14) Rate of return on total assets (21) Profit margin on net assets (22)	Quick ratio (2) Current assets/Current liabilities (17) Current liabilities/ Total liabilities (19) Profit margin on net assets (22) Earnings per share (25) Net asset value per share (26) Business income per share (28) Capital maintenance and appreciation (33)
7	Long-term debt/Total assets (14) Equity to liability ratio (15) Current assets/Current liabilities (17) Rate of return on total assets (21) Net asset value per share (26)	Quick ratio (2) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Net asset value per share (26) Capital maintenance and appreciation (33)
8	Quick ratio (2) Equity to liability ratio (15) Current assets/Current liabilities (17) Rate of return on total assets (21) Profit margin on net assets (22) Earnings per share (25) Net asset value per share (26)	Quick ratio (2) Working capital turnover (8) Current assets/Current liabilities (17) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Cash flow ratio (29) Net asset value per share (26) Capital maintenance and appreciation (33)
9	Long-term debt/Total assets (14) Rate of return on total assets (21) Profit margin on net assets (22)	Quick ratio (2) Inventories turnover (6) Fixed assets ratio (18) Profit margin on net assets (22) Return on invested capital (23) Earnings per share (25) Net asset value per share (26) Cash flow ratio (29) Capital maintenance and appreciation (33)
10	Long-term debt/Total assets (14) Equity to liability ratio (15) Rate of return on total assets (21)	Account receivable turnover (5) Long-term debt/Total assets (14) Equity to liability ratio (15) Current liabilities/ Total liabilities (19) Rate of return on total assets (21) Profit margin on net assets (22) Price to book ratio (27) Capital maintenance and appreciation (33)

5. RESULTS

In this section we explain the metrics for the performance measurement. Logistic regression, NN classification, and SVM classification are separately implemented for two groups of datasets, which are imbalance datasets in the real world and balanced datasets with synthetic samples of listed real estate companies. We then compared the performances of the two group datasets above individually.

5.1 Performance criteria for the imbalance problem

In the two-class imbalanced problem, the minority instances belong to the positive class, while the majority instances belong to the negative class. In order to estimate the performance of default prediction models, 10-fold cross-validation is used under multiple criteria, including the true positive rate (TPR), the true negative rate (TNR), the G-mean [33], and F-score [34]. These criteria are commonly adopted as the performance metrics for evaluating imbalanced learning classifiers (see for example Hwang et al. [24] and Gao et al. [35]). Many prior studies on the prediction of business defaults relied on prediction-oriented tests and the overall accuracy to measure the performance of the classifier. However, the influence of the negative samples is much higher than that of positive samples because the size of the negative samples is much larger [24]. Therefore, accuracy of default prediction models is unfair and unreliable for conducting a performance assessment of imbalanced datasets. Additionally, the prediction-oriented tests have another shortcoming in the presence of imbalanced datasets. There is an assessing principle in which the costs of each type's classification error are valued equally. This isn't reasonable in the real world, in fact, Type I errors are generally more costly than Type II errors. Given that the prediction-oriented test and accuracy are not suitable to represent the performances of default prediction models, this paper employed the true positive rate (TPR) and the true negative rate (TNR), as well as the G-mean and F-score to indicate the classifier generalization capability. These criteria can be calculated with respect to a confusion matrix, as shown in Table 4.

Table 4. Confusion matrix

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

The TPR, TNR, and G-mean are defined as shown in Eqs. (11) - (13)

$$TPR = TP / (TP + FN) \tag{11}$$

$$TNR = TN / (FP + TN) \tag{12}$$

$$G - mean = \sqrt{TPR + TNR} \tag{13}$$

The F-score can be defined by two parameters called Recall and Precision, where Precision and Recall are defined as follows:

$$precision = \frac{TP}{TP + FP} \tag{14}$$

$$recall = \frac{TP}{TP + FN} \tag{15}$$

As a composite metric, the F-score is desirable for increasing the recall without a sacrifice in precision.

$$f - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{16}$$

5.2 Estimating Prediction Models

To evaluate the prediction performance of the method based on MSGA-RPD-NN, we used the single prediction models with imbalanced datasets as benchmarks in carrying out default prediction of the real estate industry. These single prediction models are also corrected by MSGA-RPD-NN with balanced datasets. For testing, we obtained a rational ratio of negative samples against positive samples by generating 286 default samples based on MSGA-RPD-NN. We used 10-fold cross-validation to perform our estimation. The initial imbalance samples and corrected balance samples were randomly divided into 10 datasets. The 10 datasets were divided into two parts: 9 datasets for training and 1 dataset for testing. On each validation, the logistic regression, NN, and SVM classifiers were compared in terms of TPR, TNR, Accuracy, G-mean, Precision, and F-score.

Table 5 and Table 6 provide the prediction results of the training datasets and the corresponding testing datasets by using the initial imbalanced datasets at year (t-2). The prediction results of the training datasets and the corresponding testing datasets using balanced datasets at year (t-2) are summarized in Table 7 and Table 8.

In the aspect of TPR, the single prediction shows better performance in most cases. It used MSGA-RPD-NN to create a balanced real estate dataset for training and testing datasets. The TNRs are generally consistent with the logistic regression, NN, and SVM models that used balanced and imbalanced datasets. Compared with the models that were based on imbalanced datasets and balanced datasets can achieve a higher value of the G-mean. The reason for this is that the over-sampling method not only improves the sensitivity of discriminatory to minority samples, but it also does not affect the forecasting performance of majority samples.

For training datasets, classifiers using balanced samples had the highest precision during all the classifiers except during the NN classifier, and the F-score that used balanced samples was also the highest. Since the imbalance degree of the majority samples compared with the minority samples was approximately 30:1, and only one or two defaulted samples were in the testing set, the prediction methods were overwhelmed by non-defaulted samples and were not susceptible to the default samples. Once the models failed to predict a default when it occurred, or non-default forecasts are exactly right, the value of TP and FP was 0. In this case, there was a zero denominator for the Precision and F-score, as shown in Table 6. Even so, based on the TPR and G-mean, the costs of a prediction failure for defaulted real estate companies is much larger than that for healthy companies. We were also able to draw the conclusion that the methods using MSGA-RPD-NN to create a balanced real estate datasets demonstrated competitive test performance on the prediction of default samples.

Table 5. Training results of 10-fold cross-validation using initial imbalanced datasets at year (t-2)

Classifier	TPR	TNR	Accuracy	G-mean	Precision	F-score
Logit						
1	0.692	1.000	0.986	1.301	1.000	0.818
2	0.615	0.993	0.976	1.268	0.801	0.696
3	0.333	0.993	0.966	1.152	0.668	0.444
4	0.833	0.996	0.990	1.352	0.898	0.864
5	0.538	0.993	0.973	1.237	0.779	0.637
6	0.333	0.996	0.970	1.153	0.779	0.466
7	0.846	1.000	0.993	1.359	1.000	0.917
8	1.000	1.000	1.000	1.414	1.000	1.000
9	0.231	0.989	0.956	1.105	0.491	0.314
10	0.500	0.986	0.970	1.219	0.601	0.546
NN						
1	0.308	1.000	0.970	1.144	1.000	0.471
2	0.231	1.000	0.966	1.110	1.000	0.375
3	0.168	1.000	0.966	1.081	1.000	0.288
4	0.333	1.000	0.973	1.155	1.000	0.500
5	0.154	1.000	0.963	1.074	1.000	0.267
6	0.082	1.000	0.962	1.040	1.000	0.152
7	0.385	1.000	0.970	1.177	1.000	0.556
8	0.385	0.997	0.970	1.176	0.855	0.531
9	0.077	0.989	0.960	1.032	0.243	0.117
10	0.250	1.000	0.970	1.118	1.000	0.400
SVM						
1	0.615	1.000	0.983	1.271	1.000	0.762
2	0.692	1.000	0.987	1.301	1.000	0.818
3	0.333	1.000	0.973	1.155	1.000	0.500
4	0.750	1.000	0.990	1.323	1.000	0.857
5	0.923	1.000	0.997	1.387	1.000	0.960
6	0.250	1.000	0.970	1.118	1.000	0.400
7	0.385	1.000	0.973	1.177	1.000	0.556
8	0.539	1.000	0.980	1.241	1.000	0.700
9	0.154	1.000	0.963	1.074	1.000	0.267
10	0.500	0.997	0.976	1.224	0.876	0.637

Table 6. Testing results of 10-fold cross-validation using initial imbalanced datasets at year (t-2)

Classifier	TPR	TNR	Accuracy	G-mean	Precision	F-score
Logit						
1	0.000	1.000	0.970	1.000	——	——
2	0.000	0.967	0.939	0.983	0.000	——
3	0.500	1.000	0.970	1.225	1.000	0.667
4	1.000	1.000	1.000	1.414	1.000	1.000
5	0.000	1.000	0.970	1.000	——	——
6	1.000	1.000	1.000	1.414	1.000	1.000
7	0.000	1.000	0.970	1.000	——	——
8	0.000	0.968	0.936	0.984	0.000	——
9	0.000	1.000	0.970	1.000	——	——
10	0.500	1.000	0.970	1.225	1.000	0.667
NN						
1	0.000	1.000	0.970	1.000	——	——
2	0.000	0.969	0.939	0.984	0.000	——
3	0.000	1.000	0.939	1.000	——	——
4	0.000	1.000	0.939	1.000	——	——
5	0.000	1.000	0.970	1.000	——	——
6	0.000	1.000	0.939	1.000	——	——
7	0.000	1.000	0.970	1.000	——	——
8	0.000	1.000	0.970	1.000	——	——
9	0.000	1.000	0.970	1.000	——	——
10	0.000	1.000	0.939	1.000	——	——

SVM						
1	0.000	0.969	0.939	0.984	0.000	——
2	0.000	0.969	0.939	0.984	0.000	——
3	0.000	1.000	0.939	1.000	——	——
4	0.000	1.000	0.939	1.000	——	——
5	0.000	0.969	0.939	0.984	0.000	——
6	0.000	1.000	0.939	1.000	——	——
7	0.000	1.000	0.970	1.000	——	——
8	1.000	1.000	1.000	1.414	1.000	1.000
9	0.000	1.000	0.970	1.000	——	——
10	0.000	1.000	0.939	1.000	——	——

Table 7. Training results of 10-fold cross-validation using balanced datasets at year (t-2)

Classifier	TPR	TNR	Accuracy	G-mean	Precision	F-score
Logit						
1	1.000	1.000	1.000	1.414	1.000	1.000
2	1.000	1.000	1.000	1.414	1.000	1.000
3	1.000	1.000	1.000	1.414	1.000	1.000
4	0.996	0.996	0.996	1.411	0.996	0.996
5	1.000	1.000	1.000	1.414	1.000	1.000
6	1.000	1.000	1.000	1.414	1.000	1.000
7	0.989	0.986	0.987	1.405	0.985	0.987
8	1.000	1.000	1.000	1.414	1.000	1.000
9	1.000	1.000	1.000	1.414	1.000	1.000
10	0.993	0.993	0.993	1.409	0.993	0.993
NN						
1	0.985	0.919	0.951	1.380	0.921	0.952
2	0.970	0.834	0.901	1.343	0.848	0.905
3	0.959	0.954	0.957	1.383	0.952	0.955
4	0.956	0.947	0.951	1.379	0.945	0.950
5	0.933	0.979	0.957	1.383	0.977	0.954
6	0.963	0.975	0.969	1.392	0.973	0.968
7	0.974	0.933	0.953	1.381	0.933	0.953
8	0.959	0.965	0.962	1.387	0.963	0.961
9	0.982	0.958	0.969	1.393	0.957	0.969
10	0.919	0.944	0.931	1.365	0.940	0.929
SVM						
1	1.000	1.000	1.000	1.414	1.000	1.000
2	1.000	1.000	1.000	1.414	1.000	1.000
3	1.000	1.000	1.000	1.414	1.000	1.000
4	1.000	1.000	1.000	1.414	1.000	1.000
5	1.000	1.000	1.000	1.414	1.000	1.000
6	1.000	1.000	1.000	1.414	1.000	1.000
7	0.996	1.000	0.998	1.413	1.000	0.998
8	1.000	1.000	1.000	1.414	1.000	1.000
9	1.000	1.000	1.000	1.414	1.000	1.000
10	1.000	1.000	1.000	1.414	1.000	1.000

Table 8. Testing results of 10-fold cross-validation using balanced datasets at year (t-2)

Classifier	TPR	TNR	Accuracy	G-mean	Precision	F-score
Logit						
1	0.933	1.000	0.968	1.390	1.000	0.965
2	1.000	1.000	1.000	1.414	1.000	1.000
3	1.000	1.000	1.000	1.414	1.000	1.000
4	1.000	0.935	0.967	1.391	0.937	0.968
5	1.000	1.000	1.000	1.414	1.000	1.000
6	1.000	1.000	1.000	1.414	1.000	1.000
7	1.000	1.000	1.000	1.414	1.000	1.000
8	1.000	1.000	1.000	1.414	1.000	1.000
9	1.000	1.000	1.000	1.414	1.000	1.000
10	0.500	1.000	0.970	1.225	1.000	0.667
NN						
1	0.933	0.938	0.936	1.368	0.934	0.933
2	1.000	0.781	0.887	1.335	0.811	0.895
3	0.967	1.000	0.984	1.402	1.000	0.983
4	0.967	0.903	0.934	1.367	0.906	0.936
5	0.967	0.906	0.936	1.369	0.906	0.936
6	1.000	0.936	0.968	1.391	0.938	0.968
7	0.967	0.969	0.968	1.391	0.967	0.967
8	1.000	0.968	0.984	1.403	0.968	0.984
9	0.900	0.906	0.903	1.344	0.900	0.900
10	0.900	1.000	0.950	1.378	1.000	0.947
SVM						
1	0.967	0.969	0.968	1.391	0.967	0.967
2	1.000	0.969	0.984	1.403	0.968	0.984
3	1.000	0.969	0.984	1.403	0.969	0.984
4	1.000	0.968	0.984	1.403	0.968	0.984
5	1.000	0.969	0.984	1.403	0.968	0.984
6	1.000	0.936	0.967	1.391	0.938	0.968
7	1.000	0.969	0.984	1.403	0.968	0.984
8	1.000	1.000	1.000	1.414	1.000	1.000
9	1.000	0.969	0.984	1.403	0.968	0.984
10	1.000	1.000	1.000	1.414	1.000	1.000

6. CONCLUSION

To the best of our knowledge, most studies on default prediction for real estate companies are based on imbalanced datasets. But, an imbalanced dataset creates an enormous hindrance for machine learning models.

In this paper we tried to introduce a minority-sample generation approach and performance criteria for imbalanced datasets. We also attempted to provide a new framework for predicting the default rate of real estate companies. To avoid the choice bias caused by initial sampling, we used all firm-years sample of Chinese real estate companies over the period 2005 to 2009 and used 14 default and 315 non-default samples. Then we applied MSGA-RPD-NN [26], which generates new minority samples in a feature space instead of in a data space, to predict the default rate for an imbalanced dataset of defaulted and non-defaulted Chinese real estate companies. For performance criteria, previous studies mostly applied the prediction-oriented test, which ignores the unequal costs of defaulted and non-defaulted cases. Thus, we chose TPR, TNR, the G-mean, and the F-score to evaluate the performance of prediction models with an imbalanced dataset. In order to identify the necessity of employing the minority-sample generation approach to correct an imbalanced dataset, we compared the prediction power of

single prediction models using both imbalanced and balanced datasets. The results indicated that machine learning models, as well as the logistic models with a balanced dataset, had a higher G-mean and F-score and a higher true positive rate (TPR) without losing the true negative rate (TNR).

A further extension of this research would be to search for an optimal cut-off point so that the extent of misclassifications of defaulted and non-defaulted companies are lower at the same time. If a predicted default rate is close to 0.5, the slight fluctuation of the cut-off point will affect the diagnoses of a defaulted company. Consequently, further studies should design a sensitivity analysis for the cut-off point. Meanwhile, besides choosing an optimization of parameters, non-financial indicators, such as geographic regions and regional macroeconomic environment factors, need to be adopted in the models. In addition, housing, immovable property, is not clearly distinguished from the products in other industries. Hence, the correlation between the default rate and regional economic factors needs further analysis.

REFERENCES

[1] A. Camara, I. Popova, and B. Simkins, "A comparative study of the probability of default for global financial firms," *Journal of Banking & Finance*, vol. 36, no. 3, pp. 717-732, 2012.

[2] P. Gharghori, H. Chan, and R. Faff, "Default risk and equity returns: Australian evidence," *Pacific-Basin Finance Journal*, vol. 17, no. 5, pp. 580-593, 2009.

[3] M. Xu and C. Zhang, "Bankruptcy prediction: the case of Japanese listed companies," *Review of Accounting Studies*, vol. 14, no. 4, pp. 534-558, 2009.

[4] Z. Xiao, X. Yang, Y. Pang, and X. Dang, "The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory," *Knowledge-Based Systems*, vol. 26, pp. 196-206, 2012.

[5] S. Chava and R. A. Jarrow, "Bankruptcy prediction with industry effects," *Review of Finance*, vol. 8, no. 4, pp. 537-569, 2004.

[6] M. C. Gupta and R. J. Huefner, "Cluster analysis study of financial ratios and industry characteristics," *Journal of Accounting Research*, vol. 10, no. 1, pp. 77-95, 1972.

[7] K. Patel and P. Vlamis, "An empirical estimation of default risk of the UK real estate companies," *Journal of Real Estate Finance and Economics*, vol. 32, no. 1, pp. 21-40, 2006.

[8] K. Patel and R. Pereira, "Expected default probabilities in structural models: Empirical evidence," *Journal of Real Estate Finance and Economics*, vol. 34, no. 1, pp. 107-133, 2007.

[9] H. Shen and Y. Jiang, "Logit model for pre-warning financial distress of listed real-estate companies in China," in *Proceedings of the International Conference on Management and Service Science*, Wuhan, China, 2010.

[10] Y. H. Kang and X. Li, "Research on financial distress prediction of China real estate public companies based on Z-Score model," in *Proceedings of the 18th Annual International Conference on Management Science and Engineering*, Rome, Italy, 2011, pp. 1166-1173.

[11] H. Frydman, E. I. Altman, and D. L. Kao, "Introducing recursive partitioning for financial classification: the case of financial distress," *Journal of Finance*, vol. 40, no. 1, pp. 269-291, 1985.

[12] M. D. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA, 1990, pp. 163-168.

[13] B. S. Ahn, S. S. Cho, and C. Y. Kim, "The integrated methodology of rough set theory and artificial neural network for business failure prediction," *Expert Systems with Applications*, vol. 18, no. 2, pp. 65-74, 2000.

[14] W. Hardle, Y. J. Lee, D. Schafer, and Y. R. Yeh, "Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies," *Journal of Forecasting*, vol. 28, no. 6, pp. 512-534, 2009.

[15] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429-449, 2002.

[16] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of Accounting Research*, vol. 4, pp. 71-111, 1966.

[17] E. I. Altman, "Financial ratios, discriminant analysis and prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589-609, 1968.

[18] E. I. Altman and R. A. Eisenbeis, "Financial applications of discriminant-analysis: clarification," *Journal of Financial and Quantitative Analysis*, vol. 13, no. 1, pp. 185-195, 1978.

[19] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109-131, 1980.

[20] M. E. Zmijewski, "Methodological issues related to the estimation of financial distress prediction models," *Journal of Accounting Research*, vol. 22, pp. 59-82, 1984.

[21] P. K. Coats and L. F. Fant, "Recognizing financial distress patterns using a neural-network tool," *Financial Management*, vol. 22, no. 3, pp. 142-155, 1993.

[22] 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.

[23] R. Barandela, J. S. Sanchez, V. Garcia, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849-851, 2003.

[24] J. P. Hwang, S. Park, and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8580-8585, 2011.

[25] N. Japkowicz, "The class imbalance problem: significance and strategies," in *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas, NV, 2000, pp. 111-117.

[26] H. Li and J. Sun, "Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples: evidence from the Chinese hotel industry," *Tourism Management*, vol. 33, no. 3, pp. 622-634, 2012.

[27] P. Brockman and H. J. Turtle, "A barrier option framework for corporate security valuation," *Journal of Financial Economics*, vol. 67, no. 3, pp. 511-529, 2003.

[28] S. A. Hillegeist, E. K. Keating, D. P. Cram, and K. G. Lundstedt, "Assessing the probability of bankruptcy," *Review of Accounting Studies*, vol. 9, no. 1, pp. 5-34, 2004.

[29] H. P. Tserng, G. F. Lin, L. K. Tsai, and P. C. Chen, "An enforced support vector machine model for construction contractor default prediction," *Automation in Construction*, vol. 20, no. 8, pp. 1242-1249, 2011.

[30] J. Sun and H. Li, "Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers," *Expert Systems with Applications*, vol. 35, no. 3, pp. 818-827, 2008.

[31] J. Sun and H. Li, "Financial distress prediction based on serial combination of multiple classifiers," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8659-8666, 2009.

[32] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management Science*, vol. 38, no. 7, pp. 926-947, 1992.

[33] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[34] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 225-252, 2008.

[35] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456-3466, 2011.

Yuan-Xiang Dong

She received BS degree in Information management from Harbin University of Commerce in 2008. Currently, she is a Ph.D. candidate in the School of Economics and Business Administration at Chongqing University. Her research interests include intelligent decision-making, information intelligent analysis and business Intelligence.

Zhi Xiao

He received BS degree in Department of Mathematics from Southwest Normal University in 1982. And he received his Ph.D. degree in Technology Economy and Management from Chongqing University in 2003. He is currently a Professor in the School of Economics and Business Administration at Chongqing University. His research interests are in the areas of information intelligent analysis, data mining, and business Intelligence.

Xue Xiao

She went to Singapore for study with a government full scholarship since 2005. Currently she is pursuing her BSc (Hon.) degree of Real Estate in National University of Singapore in the 4th year with intention of specializing in real estate finance. She has particular research interest in topics relating to listed real estate companies' behavior, commercial real estate rental dynamics and alternative financing methods of real estate companies.