The Application of Internet Big Data and Support Vector Machine in Risk Warning

Jiang Li¹*, Yaochen Tan², Ao Zhang³

1School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei, 430073
2School of Finance, Zhongnan University of Economics and Law, Wuhan, Hubei, 430073
3College of Business Administration, Hunan University, Changsha, Hunan, 410006

*Corresponding author: jiangli@zuel.edu.cn

Abstract. With the strengthening of macro-control of the real estate industry and the intensification of market competition, it is of great significance for the steady development of the industry to establish an accurate and effective early-warning mechanism for enterprise financing risk. Taking the A-share real estate listed companies in China in 2019 as target, this paper collects financial information of the relevant companies from 2010 to 2019, supplemented the risk sample data from 2005 to 2010, and reduced classification imbalance by removing discrete points and using SMOTE. When the financing risk evaluation system of listed real estate companies is constructed with capital as the entry point, the stochastic forest algorithm is used to finally select five important characteristic dimensions, namely, current ratio, equity financing ratio, operating income, current liability ratio and receivable turnover ratio. This paper establishes SVM early warning model, PSO-GA-BP optimization model and KNN early warning model to predict financing risks of real estate companies. It is concluded that, by comparing their applicability and advantages and disadvantages, SVM after feature screening and sample processing has better performance in the real estate financing risk early warning, which can provide some references for enterprise decision makers, investors and regulatory authorities.

Keywords: Real estate enterprise, Financing risk warning, SVM, PSO-GA-BP, KNN.

1. Introduction
In the past decade, China's real estate industry has experienced great ups and downs. With the increase of social capital leverage, inventory overstocking and the tightening of regulatory policies, real estate companies are confronted with difficulties such as heavy capital turnover pressure, high financing cost, narrow financing channels and great financing risks. Therefore, it is necessary to locate and investigate financing risks of listed real estate companies and formulate relevant response policies, which is crucial to ensure integrity of the company's capital chain, financial security and stability of the entire industry.
2. Literature Reviews
In the field of financing risk research, foreign research has been devoted to financing risk prediction and control, and exploring the optimal financing decision. Aiming at the company's financing risks, Erdogan (2013) [1] used SVM model to give early warning of the company's financial risks, and Sicheng Li (2018) [2] used evolutionary game model and system dynamics methodology to conclude that the company's debt financing risks were mainly investment and interest rate risks. Aiming at the systematic financing risks, the TVP-VAR model constructed by Anzhong Huang (2020) [3] effectively reflects the actual financial stress in China. In addition, Wen-Hsien Tsai (2012) [4] proposed a GDMS model that took into account the company's financial risk and corporate value, so as to provide the best solution for financing decision makers.

Compared with foreign comprehensive financing risk research methods and perspectives, domestic research on financing risk is relatively backward. Wang Xiaoping (2013) [5] used Multiple Regression method and Cluster Analysis method to study the risks of debt financing of listed companies. Shen Wenwen (2018) [6] used regression analysis to conclude that the external financing of listed real estate companies in China is much more affected by the financing structure than the internal financing. In recent years, domestic scholars have also applied artificial intelligence algorithm to the research on financing risks of listed companies. Dai Jun (2016) [7] combined Logistics and DEA models to conclude that equity debt structure efficiency and debt maturity structure efficiency are both in reverse relationship with financial risks. Liu Hanlin (2017) [8] used Logistics and GRA method to construct financing risks early warning matrix composed of warning indexes in financial and environmental dimensions. Xie Liqiang (2018) [9] and Xie Ahong (2019) [10] et al. verified the accuracy and scientficity of BP neural network in the early-warning of financing risks of listed energy or agriculture-related companies. Liu Yumin (2019) [11] and Geng Chengxuan (2020) [12] used genetic algorithm and fuzzy C-means clustering repsectively to optimize the SVM model. Zhang Yitang (2020) [13] used Gradient promotion algorithm to rank the financing risk indexes in the importance of information equipment enterprises.

Currently, although there are a lot of domestic research existing on listed companies financing risk early warning, but most of the real estate listed companies financing researches only use traditional analysis methods, and fail to properly use artificial intelligence algorithms and compare the pros and cons of different methods in actual problems, leading to the domestic real estate financing risk early warning model of listed companies relatively backward. And the applicability and accuracy of the models need to be improved.

3. Index Selection and Data Sources

3.1. Index Selection
In this paper, ten financial indexes of listed real estate companies, such as asset-liability ratio, are selected to preliminarily establish the financing risk evaluation system. Each types, names and relationship with financing risks are shown in Table 1.
Table 1. Financing risk evaluation index

| Index Type           | Index Code | Index Name                  | Relationship with Financing Risk |
|----------------------|------------|-----------------------------|---------------------------------|
| Capital structure    | X1         | Asset-liability ratio       | +                               |
|                      | X2         | Ratio of current liabilities| +                               |
| Financing structure  | X3         | Equity financing ratio      | uncertainty                     |
|                      | X4         | Inventory turnover          | -                               |
|                      | X5         | Accounts receivable turnover| -                               |
|                      | X6         | Current asset turnover      | -                               |
| Operation capacity   | X7         | Return on total asset       | -                               |
| Profitability        | X8         | Current ratio               | -                               |
| Debt paying ability  | X9         | Quick ratio                 | -                               |
| Growth ability       | X10        | Year-on-year growth rate of operating income | - |

3.2. Data Sources

This paper collects the financial data disclosed by China’s A-share real estate listed companies in 2010-2019 from Juling Financial Services Platform, and uses the three-year average values of each indicator as one sample. The ST refers to the domestic listed companies to be de-listed risk warning due to operating losses for two consecutive years. Therefore, this paper uses ST treatment as determination criteria of a single company being at financial risk. Excluding companies with incomplete financial disclosure, a total of 866 data containing 118 ST records were collected and 58 companies have been processed by ST.

4. Early Warning Model Construction

4.1. Data Processing and SVM Early Warning Model

Considering that the magnitude of the ten features in the evaluation index system is not consistent, and the amplitude of each feature is large, the input vector needs to be normalized in order for the model to be able to learn and train effectively. Set a single sample data \( x_i^j \), \( i = 1, 2, ..., n, j = 1, 2, ..., 10 \). Then calculate the normalized data through the equation(X) shown in follow,

\[
x_i^{j'} = \frac{x_i^j - x_{i\text{min}}}{x_{i\text{max}} - x_{i\text{min}}}
\]

Where, \( x_{i\text{max}} \) is the maximum of \( x_i \), \( x_{i\text{min}} \) is the minimum of \( x_i \). The normalized data is used as the input vector of the model.

The small number of listed real estate companies and incomplete disclosure of financial data of some companies lead to less sample data, in order to avoid influences of the existence of data imbalance and too many outlier points on model effects, this paper first used the SVM model to test the original data [14].

As a classifier, the advantages of SVM are mainly reflected in small samples and its applicability. The ultimate goal of SVM is to find the optimal superplane to classify the targets. In this paper, a nonlinear support vector machine is used to make \( \varphi(x) \) represents the feature vector that maps \( x \) to \( n - 1 \)-dimensional space, where the hyperplane used to divide the dataset is:
The minimize function can be obtained as:

\[
g(x) = \omega^T \varphi(x) + b
\]

The minimize function can be obtained as:

\[
\min \frac{1}{2} ||\omega||^2
\]

\[s.t. \ g_i (\omega^T \varphi(x_i) + b) \geq 1 (i = 1, 2, ..., n)\]

Turn a problem into a pair of problems when KKT conditions is met:

\[
\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i^T) \varphi(x_j)
\]

\[s.t. \ \sum_{i=1}^{n} \alpha_i y_i = 0, \ 0 < \alpha_i < C, \ i = 1, 2, ..., n\]

And get:

\[
g(x) = \omega^T \varphi(x) + b = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b
\]

Where, \(K(x_i, x_j)\) stands for the nuclear function.

Using the sklearn module in Python to implement SVM to fit the data, and the analysis of visual classification prediction results shows that there is a serious imbalance in the distribution of data categories, and the SVM model is overfitted.

![Figure 1. SVM prediction results](image)

Adding risk sample data from 2005 to 2010 to the original sample and excluding outlier points, the final sample size is determined to be 855. Important features are screened by random forest algorithm, and the results of features significance obtained are shown in Table 2.

| Feature | X8 | X3 | X10 | X2 | X5 | X6 | X7 | X1 | X9 | X4 |
|---------|----|----|-----|----|----|----|----|----|----|----|
| Significance (10^-2) | 12.2 | 11.8 | 11.4 | 11.1 | 10.2 | 9.5 | 8.8 | 8.7 | 8.2 | 8 |

Table 2. Priority of feature importance based on random forest algorithm
In order to eliminate the unbalanced characteristics of sample data classification as far as possible, this paper uses SMOTE (Synthetic Minority Oversampling Technique) to further process the data, that is, to analyze and simulate the few categories of samples, and to combine the new samples of artificial simulation with the original data to generate a new training set.

Take the first five features in Table 2, and fit the processed data again, where the test set ratio is 0.4, and the core function is RBF function. And then, using 10-fold cross-validation to obtain the optimal parameter gamma and C are 10 and 1, respectively. According to the fitting results shown in Figure 2, the model prediction is more accurate with the accuracy 0.953 and the data separation effect is better.

![Figure 2. Confusion matrix and ROC curve](image)

For further analysis of the prediction effect, the visualized classification results are output four times, as shown in Figure 3. Combined with ROC curve, it can be seen that after data processing and use of random forest screening variables, the model goodness of fit has been greatly improved and the overfitting problem has been alleviated to a great extent.

![Figure 3. SVM prediction results](image)
4.2. BP Neural Network Early Warning Model

Forward propagation and error reverse updates are the two cores of BP neural network, as shown in Figure 4 of the algorithm flowchart:

![BP Neural Network Algorithm Flow Chart](image)

**Figure 4.** BP neural network algorithm flow chart

In forward propagation, set $w_{ij}^{(p)}$ as the connection weight between the $j^{th}$ cell of $p^{th}$ layer and the $i^{th}$ cell in the $(p-1)^{th}$ layer, and $b_i^{(p)}$ as the bias of the first cell of $p^{th}$ layer, then:

$$l_i^{(p)} = f(v_i^{(p)})$$

$$v_i^{(p)} = \sum_{j=1}^{b_i} w_{ij}^{(p)} z_j^{(p-1)} + b_i^{(p)}$$

Where $f$ is the activation function, and in the use of support vector machine to predict real estate financing risk, the activation function is the sigmoid function. $l_i^{(p)}$ is the first element in the p-layer output vector.

After a forward propagation, a reverse update is made. In this paper, three layers of neural network is constructed. The prediction error for the output layer is:

$$err = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{t} (H_i^{(o)} - y_i^{(o)})^2$$

Where, the predicted output for the first input sample is $H (i)$. The weights and bias update values of the output layer and the implied layer are:

$$W^{(0)} = W^{(0)} - \alpha \frac{\partial}{\partial W^{(0)}} err$$

$$b^{(0)} = b^{(0)} - \alpha \frac{\partial}{\partial b^{(0)}} err$$

Where, the debugging range for the learning rate is set to 0.01-0.1.

After the initial fit using the basic BP neural network, it is found that the prediction results are all positive or completely reverse. In addition, the prediction accuracy is unstable and the data separation
effect is poor. Therefore, the genetic algorithm and particle group algorithm are used to optimize relevant parameters.

4.2.1. **Number of Hidden Nodes Optimization in the Neural Network Based on PSO Algorithm.** The Particle Group (PSO) algorithm is an optimized algorithm based on group intelligence global search. The number of neuron nodes in the hidden layer is the key factor related to the accuracy of neural network prediction, since too many numbers of neurons in the implicit layer needs to constantly train the network while too few will make the information contained in the network too one-sided. This paper first uses PSO to optimize the neural network, of which the specific optimization steps are shown in Figure 5:

![Optimized neural network node flow chart by PSO](image)

**Figure 5.** Optimized neural network node flow chart by PSO

Using the PSO-BP algorithm to fit the data, the optimal implied layer node is determined to be 11. Figure 6 shows that when the number of iterations reaches 80, the model average square error is minimized, 0.253, and the accuracy of the model has been improved to a certain degree.

![MSE curve of PSO-BP](image)

**Figure 6.** MSE curve of PSO-BP
4.2.2. **Network Weights Optimization Based on GA Algorithm.** Genetic algorithm (GA) is another optimization algorithm based on natural selection and natural genetic mechanism, and its algorithm flow is shown in Figure 7:

![Diagram showing the optimization process of GA algorithm](image)

**Figure 7.** Optimized neural network weight flow chart by GA

When using GA algorithm to find the optimal network weight, first read the class labels of data features and filter the characteristics according to the standard deviation. The next, build ANN architecture to generate the initial scheme, and then calculate adaptabilities of all schemes as well as select the best parents and cross and mutate, and finally create a new group. In this paper, it is assumed that there are 8 populations with mutation rate 10%, and number of termination evolutions 100 times. The hidden layer node is taken 11, according to the optimization results of PSO-BP algorithm. And it can be known from Figure 8 that after five iterations, the model accuracy reaches 0.915.

It can be found that using PSO or GA optimization single parameters to improve model accuracy by 49.4% and 22.5%, respectively, while using PSO-GA-BP optimization algorithm to optimize the number and weight of nodes in neural networks can improve the model accuracy to 83%. Therefore, PSO-GA-BP algorithm has the better applicability in real estate listed companies financing risk warning.

![Fitness curve for GA algorithm](image)

**Figure 8.** GA algorithm fitness curve
4.3. **KNN Classified Warning Model**

Compared with SVM and BP neural network, KNN algorithm has the advantages of unnecessarily estimating parameters and training, so this paper compares KNN model with the first two models to explore whether it has certain applicability in the prediction of financing risk in real estate enterprises. In the course of this study, the RBF nuclear function and cross-validation is used to determine the optimal K value.

![Figure 9. KNN error line graph](image)

According to Figure 9, the optimal K value is 12. The optimal confusion matrix and classification results are obtained from the ten-fold cross-validation, as shown in Figure 10. It can be concluded that the model has a certain degree of overfit with goodness of fit 0.893.

Through oversampling techniques, random forest feature selection and parameter optimization methods, SVM and BP neural networks can eliminate the effects of sample classification imbalance as much as possible, while KNN's over-learning training set results in poor prediction of extra-sample data. Actually, there are a number of potential risk companies in real estate industry, so the KNN model is not suitable for financing risk warning for listed companies in the industry.

![Figure 10. Confusion matrix and prediction results by KNN](image)
5. Conclusions
After using random forest to screen importance characteristics, this paper compares the applicabilities and relative merits of three artificial intelligence algorithms in financing risk prediction of real estate listed companies, and draws the following conclusions:

1. Removing discrete points, increasing the number of risk samples, and using SMOTE oversampling technology can effectively improve the unbalanced characteristics of raw data classification.

2. When using BP neural network directly to predict risk, it is likely to have unstable classification accuracy and poor prediction effects. Using PSO algorithm alone to optimize the implicit layer network nodes only improves the stability of classification accuracy, not prediction effects. On PSO-BP basis, combined with GA algorithm optimizing network weights, the prediction effect has been greatly improved to 78.3%.

3. Although KNN is convenient not to need to estimate parameters, when the sample is unbalanced, the prediction accuracy of the rare category is low, because when the sample size of one class is large whereas the other sample size is very small, it may lead to majority proportion of the large volume class samples in the K adjacent points when a new sample is entered. For financing risk problems, sample imbalance often occurs, so the KNN model is not suitable for financing risk warning.

4. Empirical results show that the SVM model combined with random forest and SMOTE technology can predict the financing risk of real estate listed companies more accurately and efficiently. It is of great reference significance for enterprises to take relevant measures in time, improve financing credit evaluation system and strengthen risk management. In the face of inventory increase, financing difficulties and other issues, real estate companies should actively broaden financing channels. And the government should further improve relevant policies and regulations and improve the coordination between policies to do a good job of protection for enterprise financing.

6. Lack of Research
This paper optimizes neural node number and network weights for BP neural network to improve the prediction accuracy of the model, but the number of hidden and output layers and the weight of the sample data are not optimized, so there is also some room for improvement in the model.

Machine learning and deep learning have great potential in the study of financing risks of listed companies. This paper only tries to uses SVM, BP neural networks and KNN optimization models, and other methods such as LSTM, CNN and SVR model can also be applied to subsequent studies.

Acknowledgements
We acknowledge the data provider. The data used in this paper is from Juling Financial Service Platform, which can be downloaded from the website, http://terminal.chinaef.com/dataextract/extractData!enterStock.action.

References
[1] Birsen Eygi Erdogan. Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. Journal of Statistical Computation and Simulation, 2013, 83(8).
[2] Sicheng Li,Bin Wang. Evolutionary game simulation of corporate investing and financing behavior from a risk perspective [J]. Cluster Computing,2019,22(3).
[3] Anzhong Huang,Lening Qiu,Zheng Li. Applying deep learning method in TVP-VAR model under systematic financial risk monitoring and early warning [J]. Journal of Computational and Applied Mathematics, 2021, 382.
[4] Wen-Hsien Tsai,Ching-Chien Yang,Jun-Der Leu,Ya-Fen Lee,Chih-Hao Yang. An Integrated Group Decision Making Support Model for Corporate Financing Decisions [J]. Group Decision and Negotiation, 2013, 22(6).
[5] Wang Yiping, Zhao Yunshan, Pan Weishuang. Empirical study of the risk of debt financing of listed companies - Take Shanxi Province listed companies as an example. Friends of Accounting, 2013 (02): 47-51.

[6] Shen Wenwen. Research on Financing Mode and Financing Structure of Listed Real Estate Companies in China [D]. University of International Business and Economics, 2018.

[7] Dai Jun. The Efficiency and Risk Analysis of the Financing Structure of Listed Companies in Guangxi, 2016 (03): 112-114.

[8] Liu Hanlin. Research on financing risk early warning of real estate listed companies. Wuhan University of Technology, 2017.

[9] Xie Liqiang. Early Warning Research on Financing Risks of Coal Listed Companies Based on BP Neural Networks. Shanxi University of Finance and Economics, 2018.

[10] Xie Ahong, Xue Weiyu, Bao Jianhua, Zhu Jiaming. Early Warning Of Financing Risks of Agricultural-related Listed Enterprises Based on Entropy Rights-BP Neural Networks, Journal of Qiqihar University (Natural Sciences Edition), 2019, 35 (05): 78-83 plus 94.

[11] Liu Yumin, Liu Li, Ren Guangqian. Financing Risk Prediction of Chinese Listed Companies Based on GA-SVR Model, Journal of Beijing Polytechnic University (Social Sciences Edition), 2019, 21 (04): 73-81.

[12] Geng Chengxuan, Li Xiaoxuan. Sample-Weighted SVM Early Warning Study of The Risks of Technology-based Enterprise Financing, Industrial Technology Economy, 2020, 39 (07): 56-64.

[13] Zhang Youxuan, Cao Yaowei. Research on financing risk early warning and monitoring of listed companies in information equipment manufacturing industry, 2020 (14): 50-57.

[14] Zhao Dandan, Ding Jianchen. China Banking Systemic Risk Early Warning Study - Modeling Analysis Based on SVM Modeling [J] International Business (Journal of the University of International Economics and Trade), 2019 (04): 100-113.