Study on insurance in earthquake-prone areas based on BP neural network and logistic regression model

Bohong Wang [1, *, #], Xuanye Tao [1, #], Yan Min [2, #]

[1] School of Finance, Shanghai University of International Business and Economics, Shanghai, China, 201620

[2] School of Management, Shanghai University of International Business and Economics, Shanghai, China, 201620

* Corresponding author: johnhopkins0624@gmail.com

Abstract. The frequent occurrence of extreme events poses major challenges to the insurance industry and affects its sustainable development. To provide insurance recommendations for extreme disaster areas, this paper combines BP neural network model, principal component analysis method and Logistic regression model to establish our insurance model. First, we processed the data. Then, the BP neural network mode used to correlate 33 kinds of earthquake correlation indexes with earthquake probability and make prediction. To make the data more representative, we conducted dimension reduction prediction through principal component analysis, screened out 8 important factors, and identified 3 principal components. We then performed Logistic regression analysis based on these three principal components and used the Logistic model to decide whether to conduct insurance. After the experiment, the accuracy of the model can reach 78.9%, which can provide a reliable prediction of the disaster in the region, to help the insurance company to make a decision whether to insure in the region.

Keywords: BP Neural Network, Logistic Regression, Insurance.

1. Introduction

With the escalating frequency of climate change, there is a concurrent rise in extreme weather events. According to the global extreme data for 2023, the incidence of such severe weather phenomena is projected to increase. This presents a formidable challenge for the insurance industry, which must not only expand its coverage of catastrophe insurance but also fortify its own resilience. To address this issue, we have developed an insurance model that aids insurance companies in assessing whether they should provide coverage in specific areas [1-2]. We employ a BP neural network model to establish correlations between 33 earthquake-related indicators and earthquake frequency for predictive purposes. In order to enhance data correlation, we utilize principal component analysis for post-dimension reduction prediction and identify eight crucial factors along with three principal components. Subsequently, regression analysis is conducted based on these three principal components followed by employing dichotomy to determine the insurance [3].

2. Data processing

2.1. Data Collection

To predict the earthquake probability more accurately and ensure the stability of the model, we search a large number of different types of data. Through the database and the official website of China, we collected 33 types of data of 50 earthquakes in Chongqing over the past three years, amount to 1650 pieces of data [4-5].

2.2. Data Filling

Data availability is an important criterion since models cannot provide a valid assessment if they are based on unreliable data. Therefore, ensuring the continuity and authenticity of research data plays

an important role in the establishment of the model. Because of the amount of data is huge, in the process of our data collection, there are still a small part of the data can ’t be found through official website and database [6].

In order to improve this situation, we used the following two methods to improve the data. The details are as follows:

If the indicator value is smooth, we replace it with the previous data.

If the data before and after are available, the average is the missing.

3. Comprehensive Insurance Model

Because there are many related factors to the occurrence of extreme weather events, in order to improve the prediction accuracy, we comprehensively used BP neural network model、principal component analysis and binary classification model to build an insurance model.

3.1. Establish the Earthquake Probability Prediction Model Based on BP Neural Network

3.1.1. The Construction of BP Neural Work

We first establish a threshold and divide earthquakes into 2 categories, high and low. In the beginning, we used 33 metrics data that might be relevant to seismic and underwriting decisions as input features to the BP neural network, and we selected 70% of the data as a training set and 30% as a test set to train the neural network and test its effectiveness. Given the small amount of data, we choose to build a low -complexity net- work that consists of one input layer, ten hidden layers, and one output layer. By training BP neural network and associating it with earthquake probability, the probability of Chongqing earthquake in the next ten years can be predicted [7].

3.1.2. The Evaluation of BP Model

The evaluation of the test set is shown as followed:

On the test set, the R of the neural network =0.91846, which is close to 1, indicates that the model fits well.

The result of MSE is 0.2980, so we believe that this model has a high accuracy.

4. Principal Component Analysis

After trying to use the BP neural network model for data prediction, we found that some data showed low correlation, so we combined the principal component analysis method to reduce the dimension of indicators and then forecast, eliminating the redun- dancy among indicators and making the model more perfect. After data review and ver- ification and the survey of earthquake in other areas, we selected eight indicators as representative indicators and carried out principal component analysis on them, so that the data could retain as much seismic information reflected by various factors as possible [8].

4.1. Principal Component Analysis Steps

4.1.1. Set the initial matrix.

With n samples and p index, a sample matrix x of size $n \times p$ can be formed.

$$x = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \left(x_1, x_2, x_3, x_4 \right) \tag{1}$$

4.1.2. Data standardization processing

After normalization of the initial matrix, the correlation matrix R can be obtained:

$$R=\frac{\sum_{k=1}^{n} x_i\left(x_{ki}-\overline{x_i}\right)\left(x_{kj}-\overline{x_j}\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}-\overline{x_i}\right)^2\sum_{k=1}^{n}\left(x_{kj}-\overline{x_j}\right)^2}} \tag{2}$$

4.1.3. Calculate the eigenvector value.

By using MATLAB, we can get the following table 1:

Table 1. Principal component analysis of each factor load matrix

Eigenvector	a1	a2	a3
X1: Historical claim	0.401	-0.077	0.415
X2: Earthquake probability	0.132	0.749	0.332
X3: Humidity	0.375	0.065	-0.442
X4: Frequency of vibration	0.32	0.345	-0.478
X5: Density of rock	0.388	-0.232	0.279
X6: Temperature	0.406	0.027	-0.31
X7: Focal depth	0.326	-0.496	-0.034
X8: Casualty	0.396	0.096	0.345
Eigenvalue	5.098	1.352	0.574
Contribution value	0.637	0.169	0.072
Cumulative contribution value	0.637	0.806	0.878

(1) a_1 - local aggregate housing conditions component

The first principal component a_1 has approximately equal positive loads on all variables (except for a slightly smaller load on x2), reflecting the local aggregate data, so the first principal component can be called the local aggregate housing conditions [9-10].

(2) a_2 - natural disaster intensity component

The second principal component a_2 has a high positive load on variable x2, a moderate positive load on variable x4, and a negative or small positive load on the remaining variables, so it can be considered that this principal component measures the frequency and probability of earthquakes, so the second component can be called the intensity of natural disasters.

(3) a_3 - financial expenditure estimate component

The third component a_3 has a high positive load on variable 1 and a moderate positive load on variable 2. Therefore, it is judged that principal component 3 mainly reflects property-related data, which is called the financial expenditure estimate.

4.1.4. Calculate principal component contribution rate and cumulative contribution rate:

$$Contribution\ value=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}\left(i=1,2,\ldots,p\right) \tag{3}$$

$$Cumulative\ contribution\ value=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}\left(i=1,2,\ldots,p\right) \tag{4}$$

All of the 3 eigenvectors` cumulative contribution rate close to85%. So a1, a2, a3 can be called as the principal component. Based on the F-matrix where the principal component index is located, we

use SPSS cluster analysis to calculate the values of the first, second and third principal components, which are regarded as three new indicators, namely comprehensive living conditions, natural disaster intensity and financial expenditure forecast, in order briefly summarize the previous impact.

4.2. Meaning Analysis

Through principal component analysis, we cluster eight indicators so that they can basically represent all the data related to earthquake and insurance.

5. Logistic Model Based on PCA

To determine whether to insure, we use a binary model. In this model, the dependent variable is whether it is covered or not and is written as 1 if it is covered and 0 if it is not covered.

We will analyze the data inputs of the resulting three principal components in a principal component analysis method, training them so that the conditional probability of a given input sample best predicts the class to which it belongs. (We have already set the level before the use of BP neural network).

5.1. Model Evaluation

As it can be seen in Table 2, the significance of the local comprehensive living conditions and the estimated financial expenditure are both less than 0.05, showing a good performance. The OR value of comprehensive living conditions is the strongest, so it is the factor that has the strongest influence on insurance decision -making. The significance of natural disaster intensity is relatively low, that is, the influence of natural disaster intensity on insurance decision -making is relatively low.

Because we assume that the insurance company is risk averse, it may have an impact on this type of indicator.

Table 2. variables in the equation

	B	Standard error	Wald	Degree of freedom	Significance	Exp (B)
Estimated financial expenditure	-0.059	0.022	7.382	1	0.007	0.942
Comprehensive living conditions	4.442	1.782	6.212	1	0.013	84.906
Natural disaster intensity	8.561	5.434	2.483	1	0.115	5224.652
	-29.392	13.225	4.940	1	0.026	0.000

Table 3. Binary table on underwriting decisions

Actual measurement	Underwriting decision=underwriting			Correct percentage
Underwriting decision=underwriting		0	1	
	0	15	4	78.9
	1	5	14	73.7
Overall percentage				76.3

As it can be seen from the above table 3, among the 19 examples of non -coverage, 4 are predicted to be covered and 15 are not covered, with an accuracy rate of 78.9%. The same is true for the underwriting example. Therefore, we believe that the accuracy of this model is 78.9%.

6. Conclusion

Aiming at the extreme weather of earthquakes, we selected Tokyo, Japan and California, USA, two regions with frequent earthquake outbreaks to demonstrate our model. (Again, we assume the insurer is risk averse.) Based on the model, we suggest that insurance companies can continue coverage in California and stop coverage in Tokyo, Japan.

As it can be seen from the above table2, among the 19 examples of non -coverage, 4 are predicted to be covered and 15 are not covered, with an accuracy rate of 78.9%. The same is true for the underwriting example. Therefore, we believe that the accuracy of this model is 78.9%.

Thus, this model can be used to predict the possibility of the occurrence of the extreme weather in the area and according to the data obtained the model will calculate if the insurance company can profit here and help the insurance company to make decisions helping insurance company to maintain a relative development.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic [J]. Transactions of China Electrotechnical Society, 2004, 19 (10): 53 - 58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network [D]. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability [J]. IEEE Transactions on Power Systems, 2001, 16 (4): 798 - 805.

[4] Li Lixiao, et al. Analysis of extreme natural disaster risk and disaster-causing model in typical urban areas under the background of urban size and population growth. Industrial buildings 1 - 11.

[5] Wang Shen, Li Xinguang and Zhan Jun et al. Short-term traffic flow prediction based on BP neural network with improved Sparrow search algorithm [J]. Journal of Qingdao University of Technology, 2024, 45 (01):126 - 133+140.

[6] Han Lei. Research on evaluation model of academic papers based on BP neural network [J]. Modern Information, 2019, 44 (02):170 - 177.

[7] Zhao Tianwei, Chen Huida. National population forecast based on linear regression model, Malthusian population growth model and logistic model [J]. Journal of Guangdong Medical University, 2019, 41 (06): 623 - 627.

[8] Lin Cunjie, Xiao Feng, Qiao Nan. Multi-source public loan default forecasting method based on distributed Logistic model [J/OL]. Mathematical statistics and management, 1 - 18 [2024-03-14]. https://doi.org/10.13860/j.cnki.sltj.20231030 - 009.

[9] Chen Yanjun, Li Feng. An empirical analysis of grass-roots work evaluation system in universities based on fuzzy evaluation model [J]. Theoretical Research and Practice of Innovation and Entrepreneurship, 2022, 5 (08): 4 - 6.

[10] Guo Chunmei, Gao Xiang. Based on multi-level fuzzy evaluation model of mathematics curriculum remote live study [J]. Journal of xiangtan university (natural science edition), 2021 lancet (5): 79 - 88. The DOI: 10.13715 / j. carol carroll nki nsjxu. 2021. 05. 010.