Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Advancing Earthquake Prediction in China: Machine Learning Approaches for Risk Assessment and Magnitude Forecasting

Yaping Pan *

School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China

* Corresponding author: 21064013@suibe.edu.cn

Abstract. Earthquakes pose severe disasters and losses due to their sudden and destructive nature, and despite extensive research, predicting earthquakes remains a significant challenge. Recent advancements in earthquake observation and geophysical methods have highlighted the potential of machine learning technologies in improving prediction accuracy. This study aims to explore earthquake distribution in China through machine learning methods, specifically logistic regression and random forest, to identify high-risk areas and predict earthquake magnitudes. The research identifies five high-risk earthquake zones in China and demonstrates that the random forest model excels in predicting earthquake magnitudes within these zones, outperforming support vector machines and backpropagation models. Notably, the study reveals that the b-value is a crucial factor in earthquake magnitude prediction and should be emphasized in future research. This study not only provides new perspectives and methodologies for earthquake prediction but also offers a scientific basis for earthquake warning and disaster prevention. Future work will focus on incorporating additional seismological and precursor indicators to enhance prediction accuracy and contribute to a deeper understanding and more practical solutions in earthquake forecasting.

Keywords: Earthquake Prediction; Machine Learning; Random Forest.

1. Introduction

Earthquakes are among the most devastating natural disasters, often causing severe damage to infrastructure and significant loss of life. Predicting earthquakes, however, remains a global challenge due to their inherent unpredictability [1]. With the advent of artificial intelligence and machine learning, data-driven methods for earthquake prediction have shown promising potential [2]. This study explores the effectiveness of two specific machine learning techniques: logistic regression and random forest (RF). The choice of these methods rests on their proven capabilities in pattern recognition and classification tasks across various fields. By harnessing the power of these algorithms, this research aims to develop a predictive model that can identify potential earthquake precursors from vast datasets.

In the field of earthquake prediction, various machine learning methods have been employed, including rule-based methods, shallow machine learning, and deep learning [3]. Logistic regression and RF, as two commonly used machine learning algorithms, have been applied in earthquake prediction. Logistic regression has been demonstrated to be effective in analyzing multivariate data to predict the probability of ground surface ruptures caused by earthquakes [4]. It has also been validated for its applicability in real-world assessments, such as evaluating landslide susceptibility following significant seismic events [5]. Furthermore, logistic regression has been used in conjunction with machine learning regression algorithms to provide practical guidance for earthquake prediction at experimental seismic sites in China [6]. On the other hand, the random forest model has been successfully applied to various earthquake related prediction tasks. By integrating multiple decision trees, the RF model demonstrates excellent performance in capturing complex patterns in seismic data. In Budiman et al.'s study, the RF model significantly improved the accuracy of earthquake hazard assessment by optimizing feature selection and model parameters [7]. In addition, Zhang and Wengang et al. applied random forest and extreme gradient boosting algorithms in their

research on landslide susceptibility mapping in the Fengjie area of Chongqing, further demonstrating the effectiveness of RF models in geological hazard prediction [8]. Moreover, Jang et al. used this algorithm to identify hidden control factors in the distribution of earthquake faults, demonstrating how the random forest model can improve its ability to capture the complexity of earthquake data by integrating multiple decision trees [9]. These studies not only demonstrate the application value of logistic regression and random forest models in earthquake prediction, but also highlight their importance in revealing the complexity of earthquake phenomena.

The main objective of this study is to investigate earthquake distribution across China, identify high-risk areas, and predict earthquake magnitudes using machine learning approaches, including logistic regression and RF models. Initially, a logistic regression model is employed to establish decision boundaries, effectively identifying high-risk earthquake zones. This identification facilitates targeted analysis and preventive planning. Subsequently, the RF model is applied within these identified zones to forecast earthquake magnitudes. The performance of the RF model is compared to that of models constructed using support vector machines (SVM) and backpropagation (BP) to evaluate its feasibility and effectiveness in predicting earthquake magnitudes.

2. Methodology

2.1. Dataset Description and Preprocessing

This study employs a dataset provided by the China Earthquake Networks Center, which encompasses earthquake events in China from January 14, 1990, to December 2, 2023 [10]. The dataset consists of 1,496 records, each characterized by seven attributes: magnitude, depth, longitude, latitude, earthquake time, province, and reference location. The extensive nature of this dataset allows for a detailed analysis of earthquake occurrences across various provinces, aiding in both regional earthquake impact assessment and magnitude prediction.

To enhance the accuracy of the results, the dataset underwent several preprocessing steps. Firstly, records with missing data were removed, leaving 842 entries for analysis. The dataset was then divided, allocating 80% for training and 20% for testing, to train predictive models effectively while retaining a subset for model evaluation. Additionally, to balance the influence of each numeric variable in predictive modeling, data standardization was applied, normalizing features to have zero mean and unit variance. These preprocessing measures ensure that the analysis is based on robust and accurate data, setting a strong foundation for identifying high-risk zones and predicting earthquake magnitudes.

2.2. Proposed Approach

The primary goal of this study is to apply machine learning techniques to attempt to delineate high-risk earthquake zones in China and conduct magnitude prediction, aiming to provide new insights and more accurate prediction tools for the field of earthquake prediction. Following the process in Fig. 1, initially, exploratory data analysis is conducted on the selected dataset to understand the frequency and magnitude of earthquakes in different regions. Based on this analysis, the risk level was defined and a logistic regression model was used to delineate high-risk earthquake areas. For the five designated high-risk areas, based on the Gutenberg Richter law, the RF model is used to further analyze earthquake data with magnitudes greater than 4.0 to predict earthquake magnitudes. The effectiveness of the BP model was verified by comparing the predictions of the RF model with those of the SVM and BP models. After verifying the effectiveness of the model, key factors affecting earthquake magnitude were further identified and analyzed.

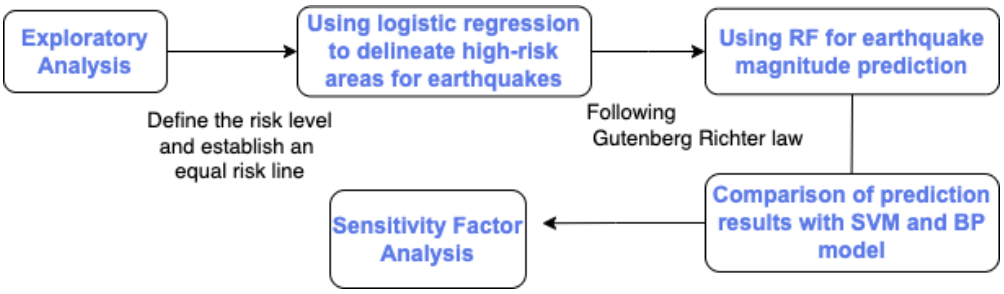

Figure 1. Flow chart process

2.2.1. Exploratory Data Analysis (EDA).

EDA refers to the exploration of obtained data with fewer prior assumptions, using visual or statistical methods to conduct preliminary checks on the data in order to better understand the distribution of data and the correlation between variables. It mainly includes steps such as data clarity, data segmentation, visualization, and statistical calculation. This article uses EDA to analyze earthquake data in various regions of China. Firstly, a thorough review of the data is conducted through data clarity steps to ensure its accuracy and completeness, identify and handle missing values, outliers, or data entry errors, and confirm the definitions and measurement units of each variable in the dataset. Next, the dataset is reasonably divided into training set, validation set, and testing set through data segmentation steps. This segmentation is crucial for evaluating the performance of machine learning models, as it allows this study to evaluate the model on different subsets of data, avoid overfitting, and ensure the model's generalization ability. By conducting EDA on earthquake data, this study can gain a preliminary understanding of the frequency and magnitude distribution of earthquakes in various regions of China, identify areas with concentrated seismic activity, which not only help to understand the spatial distribution characteristics of earthquakes, but also provides important information and guidance for subsequent dataset partitioning and model training, laying a solid foundation for establishing earthquake prediction models.

2.2.2. Logistic regression model.

Logistic regression model is a type of generalized linear model mainly used to describe and infer the relationship between two or more categorical dependent variables and a set of explanatory variables. This model mainly uses the Sigmoid function to map the linear combination of independent variables to the [0,1] interval, and its output can be understood as the probability of event occurrence. For a certain region alone, earthquakes cannot be continuous and uninterrupted natural events. Therefore, this model can be used to determine the probability of high-risk earthquakes occurring in a certain area. As shown in formula 1, this study chose the most basic logistic regression model, where the response variable y represents whether the event occurred (0 or 1), and the feature vector x contains the input features for predicting the event. The model parameter β includes intercept term β0 and coefficient βj, which are used to measure the impact of each feature on the probability of event occurrence.

$$P(Y=0\,|\,x)=\frac{1}{1+e^{(\beta_0+x\beta)}} \tag{1}$$

In this study, logistic regression models were used to delineate decision boundaries and identify five high-risk earthquake zones. The identification of these high-risk areas provides a foundation for the subsequent use of RF models for earthquake magnitude prediction. In this way, logistic regression models not only provide tools for visualizing earthquake risks, but also provide important preliminary data preparation for further magnitude prediction analysis.

2.2.3. Gutenberg richter law.

The Gutenberg richter law, also known as the earthquake frequency magnitude distribution law, is widely used in the study of earthquake magnitude. The law states as shown in formula 2:

$$N = 10^{a-bM} \tag{2}$$

where N represents the number of earthquakes with a magnitude greater than or equal to M, i.e. the cumulative frequency; M represents the magnitude of the earthquake; A and b are constants, where a reflects the seismic activity of a region, while b reflects the seismic structure and describes the rate at which the number of earthquakes decreases with increasing magnitude [11]. Based on this law, this study comprehensively considers the following six conventional physical factors as the main factors affecting earthquake magnitude: cumulative frequency of earthquakes, cumulative released energy, average magnitude, η value, b value, and magnitude of relevant areas.

2.2.4. RF model.

RF is a versatile ensemble learning method for both classification and regression, known for its robustness and high accuracy in handling various types of data. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of individual trees for classification, or averaging the predictions of individual trees for regression. It can also capture complex nonlinear relationships without the need for explicit kernel tricks, as it inherently model's nonlinearity through the structure of the decision trees. It is effective for seismic data analysis, where it can handle the intricate patterns and interactions between different variables that influence earthquake magnitude and frequency. By combining the predictions of many trees, Random Forest model can better handle the relationship between seismic activity characteristics and earthquake magnitude.

To verify the effectiveness of RF model in earthquake prediction, the prediction results of SVM and BP models can be compared. SVM model is a sophisticated supervised learning model that finds the optimal hyperplane for classification or regression, adept at capturing complex relationships with kernel functions, making it suitable for predicting earthquake magnitudes by analyzing seismic activity features. BP model is a training algorithm for artificial neural networks that learns patterns in data through iterative weight adjustments, effective for modeling the intricacies of seismic events and improving the precision of earthquake forecasts. This study compares these three models on the same dataset to demonstrate the practical application value of SVM in earthquake magnitude prediction.

2.2.5. Sensitivity factor analysis method.

Sensitivity factor analysis is a technique used to quantify the degree of influence of input variables on output results. Through this analysis, it is possible to determine which factors have a significant impact on the prediction results of the model, which can help optimize the input features of the model and improve its performance. For the sensitivity of factors, this study determines it based on the average relative error and mean square error of the predicted results, as shown in formulas 3 and 4, where R1i and R2i represent the accuracy sensitive factor and dispersion sensitive factor, respectively; active-resistive exercises (ARE) and Multiscale entropy (MSE) respectively represent the average relative error and mean square error after model prediction in the absence of factor i. If R1i>R1j, it indicates that the influence factor of i is more sensitive to the accuracy of earthquake magnitude than the influence factor of j; If R2i>R2j, it indicates that the influence factor of i is more sensitive to the stability of earthquake magnitude results than the influence factor of j. If the values of R1i and R2i are greater than or close to 1, it indicates that the influencing factor has a strong impact on earthquake magnitude. If their values are less than 1, it indicates that the factor has a small impact on predicting earthquake magnitude.

$$R_{1i} = \frac{ARE_i}{ARE} \tag{3}$$

$$R_{2i} = \frac{MSE_i}{MSE} \tag{4}$$

In the study, by conducting sensitivity factor analysis on the prediction results of RF, that can understand the different degrees of influence of earthquake indicators on magnitude, and further study the mechanisms of these factors.

3. Result and Discussion

After conducting exploratory analysis, logistic regression model, RF model, and sensitivity factor analysis to explore the principles and model construction steps, this chapter proceeded to build the model. Firstly, through exploratory data analysis, this article constructed a scatter plot of earthquake events nationwide, and based on this, formulated the definition of high-risk earthquake areas; Secondly, using logistic regression models to attempt to divide seismic and non-seismic regions; Subsequently, following the Gutenberg Richter law, this study selected earthquake data from specific regions, calculated key indicators, and trained RF, SVM, and BP models for magnitude prediction; Finally, in order to evaluate the accuracy and sensitivity of the main influencing factors in the RF model, this study used sensitivity factor analysis to exclude each factor one by one and measure the sensitivity changes of the model.

3.1. EDA Analysis

To explore the characteristics of the spatial distribution of earthquakes in China, this paper employs exploratory analysis. By integrating earthquake data with geographical information, a scatter plot containing national earthquake events is constructed using historical earthquake location parameters (latitude and longitude), which intuitively displays the location and intensity of historical earthquakes. The square of the earthquake magnitude is used as the basis for the size of the scatter points, while the depth of the scatter point color measures the magnitude of the earthquake. As shown in Fig. 2, this paper finds that earthquakes in China frequently occur in areas of active crustal movement. For example, Yunnan Province, located in the eastern part of the Himalayan seismic belt, shows frequent seismic activity, especially in the Hengduan Mountains area. Based on this, this paper preliminarily defines high-risk earthquake areas as regions with active crustal movement, that is, the higher the frequency of earthquakes, the more energy the strata have to release, and the higher the possibility of future earthquakes. In addition, based on Fig. 2, this paper can clearly see the sparsity and density of the distribution of earthquakes, and based on this, this paper attempts to use a logistic regression model to find a decision boundary to divide the earthquake area and the non-earthquake area.

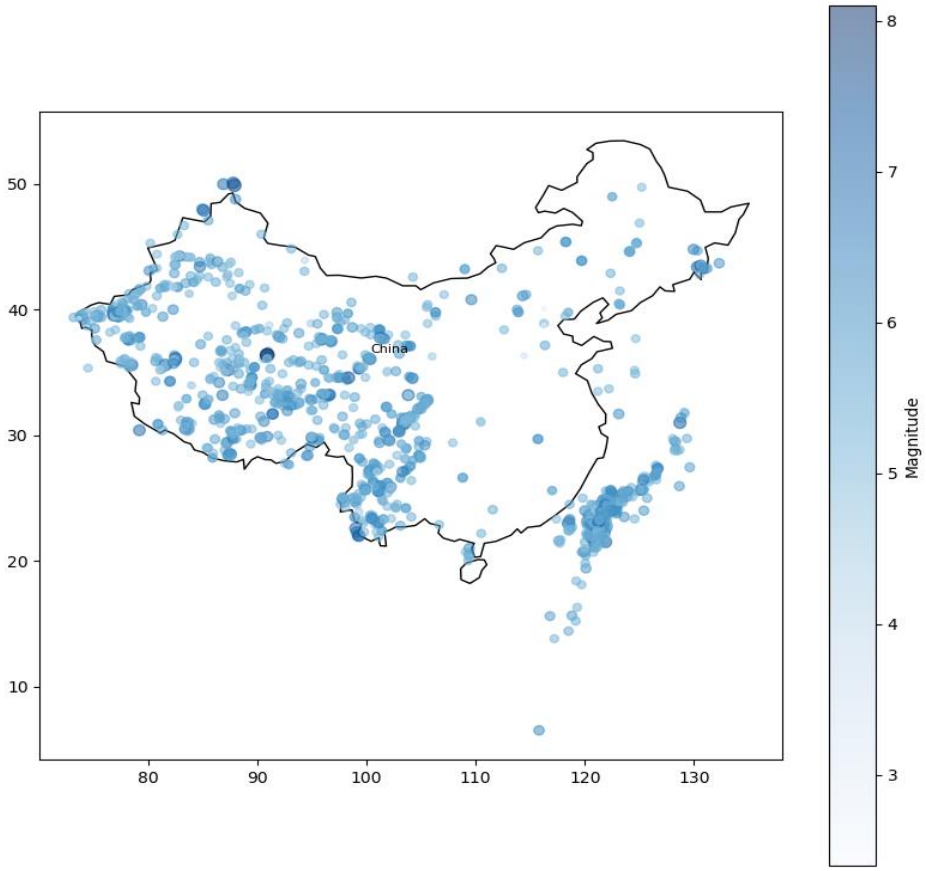

Figure 2. Scatter plot of earthquake events

3.2. Logistic Regression Model Analysis

In order to construct the decision-making boundary, this paper uses the logical regression model to determine the high-risk areas, divide the seismic areas of different levels, and finally determine the five areas of Sichuan, Guizhou, Qinghai, Yunnan and Xizang as high-risk areas. With a risk level of 60, the coefficients and intercept terms of the logistic regression model are [0.01389908 -0.19463403] and 2.7077, respectively, and the model accuracy is 0.9315. As shown in Fig. 3, The logistic regression model demonstrated excellent discriminatory ability, with an area under the Receiver Operating Characteristic (ROC) curve (AUC) of 0.916, indicating its high accuracy in classifying the earthquake risk areas.

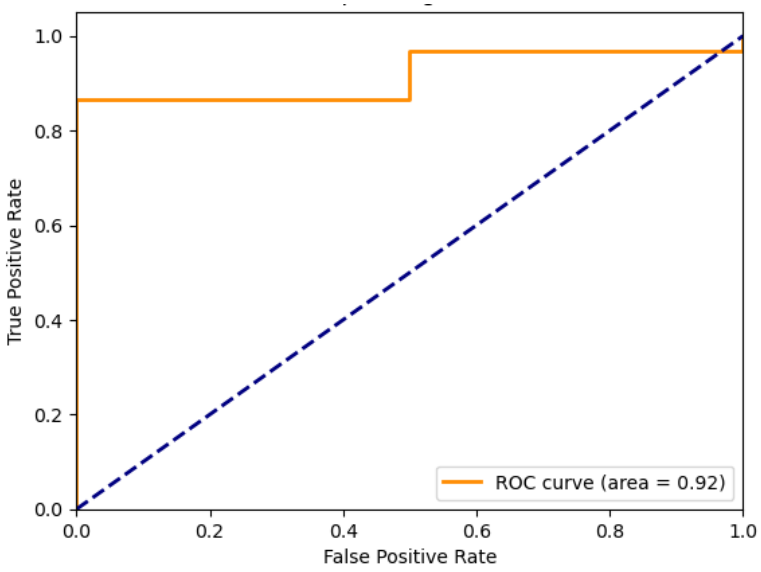

Figure 3. ROC chart

The high ROC results indicate that the high-risk earthquake area division in this study has certain reference significance. However, due to the fact that the division of this high-risk area only considers location parameters and does not introduce some geological variables, it is only based on a relatively pure empirical approach, using historical earthquake frequency to analogize the prior probability in Bayesian statistics, and using linear logistic regression to provide a rough inference of the posterior distribution of earthquakes. Therefore, in the following content of this article, this paper will introduce some geological concepts to further predict the magnitude of earthquakes.

3.3. RF Model Analysis

Following the Gutenberg-Richter law, this study selected earthquake data of magnitude 4.0 and above from five high-risk areas as the foundation. Based on this, the cumulative frequency of earthquakes, cumulative released energy, average magnitude, η value, b value, and related regional magnitude were calculated as six key indicators. These indicators were used as input data to construct and train RF, SVM and BP models for earthquake magnitude prediction.

In the comparative analysis, this paper selected predictive data from 10 samples for evaluation. As shown in Table 1, the RF model demonstrated excellent performance in predicting the 42nd sample, with a relative error of only -0.01, a precision superior to the minimum relative errors achieved by the SVM and BP models. Although the RF model had the maximum relative error of -0.085 on the 20th sample, this maximum error value is still significantly smaller than those of the SVM and BP models. Specifically, the SVM model had the maximum relative error of -0.5671 on the 13th sample, while the BP model had the maximum relative error of 0.5409 on the 36th sample.

Table 1. Comparison of Prediction Results of Three Models

Sample	Actual valur	RF	Relative error	SVM	Relative error	BP	Relative error
1	0.088235	0.089706	0.0167	0.100824	0.1427	0.097589	0.1060
10	0.235294	0.232647	-0.0113	0.319414	0.3575	0.246496	0.0476
13	0.029412	0.027941	-0.0500	0.012733	-0.5671	0.031667	0.0767
15	0.441176	0.472647	0.0713	0.495452	0.1230	0.452499	0.0257
20	0.058824	0.053824	-0.0850	0.049818	-0.1531	0.048439	-0.1765
22	0.588235	0.610294	0.0375	0.496761	-0.1555	0.401743	-0.3170
27	0.235294	0.232647	-0.0113	0.319414	0.3575	0.246496	0.0476
30	0.088235	0.090588	0.0267	0.126597	0.4348	0.130657	0.4808
36	0.058824	0.057941	-0.0150	0.086848	0.4764	0.090640	0.5409
42	0.058824	0.058235	-0.0100	0.059649	0.0140	0.055679	-0.0535

In addition, as shown in Fig. 4, after visualizing the results of the table, this study found that the normalized prediction points of the RF model matched the actual observation points quite well, which further confirmed the accuracy of its prediction.

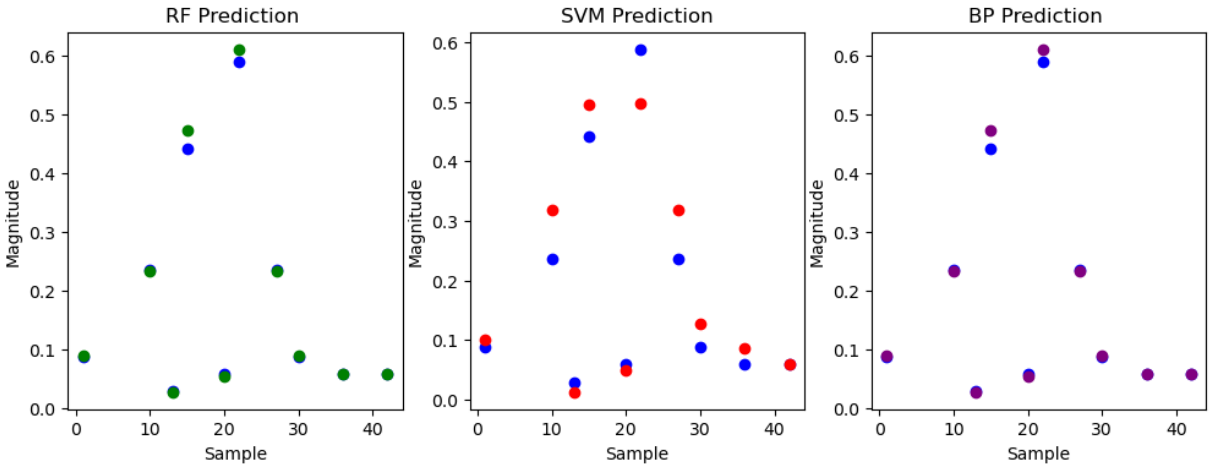

Figure 4. Comparison of normalized prediction results of three models

Next, in order to compare the distribution characteristics of the predicted results of the three models more intuitively, this paper compares the predicted results of the RF, SVM and BP models with the actual earthquake magnitude. As shown in Fig. 5, although there are some discrepancies between the predicted values of the RF model and the BP model, they are generally close to the true values and have their own advantages in different observation numbers. On the other hand, the predicted values of the SVM model are relatively poor and deviate from the true values.

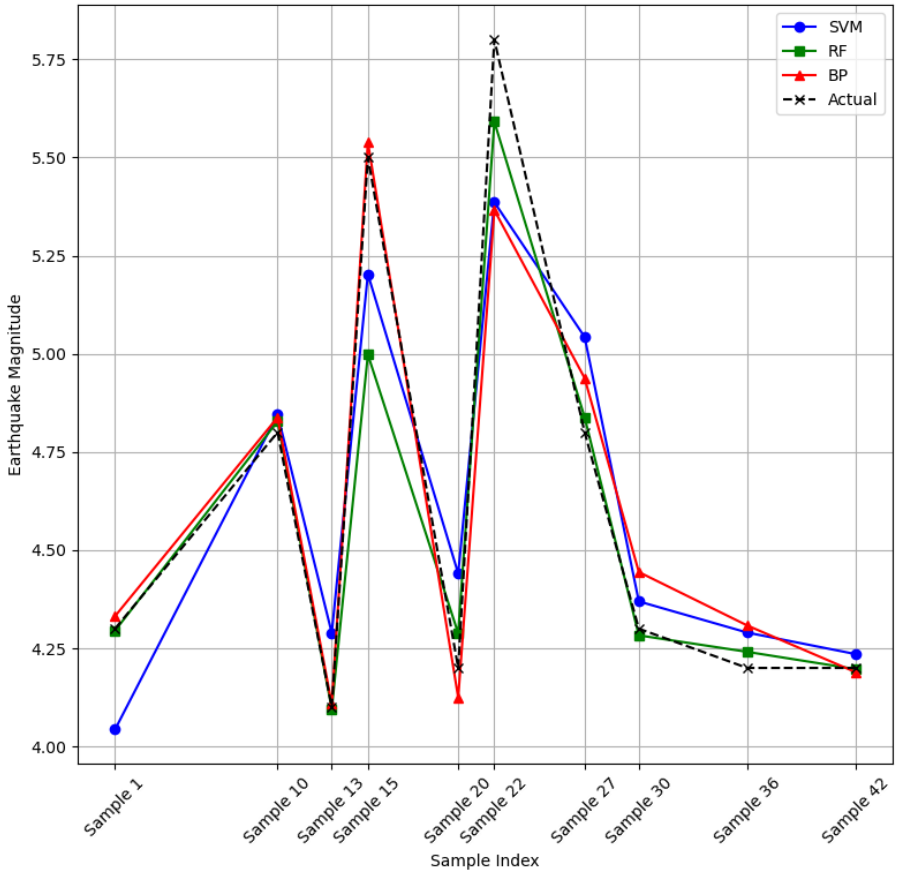

Figure 5. Comparison of actual magnitude of three models

Taking all factors into consideration, although the RF model and BP model have similar predictive effects on actual earthquake magnitudes, the RF model outperforms the BP model in terms of relative error. Therefore, based on the accuracy and stability of model predictions, this article will further explore the RF model.

3.4. Sensitivity Factor Analysis

To further explore the accuracy of the main influencing factors selected by the RF prediction model and the sensitivity of these factors to the earthquake magnitude, this study employed a sensitivity factor analysis method for evaluation. In this paper, the six main influencing factors in the RF model were excluded one by one. For each exclusion scenario, an RF model containing the remaining five factors was constructed, and the sensitivity of these models was measured. The results are summarized in Table 2.

As shown in Table 2, this study finds that the b value has the highest sensitivity index among all factors for earthquake magnitude, indicating that the b-value has the most significant impact on magnitude. In addition, the influence of the earthquake cumulative frequency and average magnitude is relatively close, while the sensitivity indices of the earthquake cumulative frequency and η-value are relatively lower. It is worth noting that the sensitivity index of the related regional magnitude is the smallest among all factors, which may imply that in the prediction model, the influence of the related regional magnitude on the magnitude is relatively minor.

Table 2. Sensitivity Comparison

Measuring indicators	Average value	Cumulative frequency	Accumulated release of energy	Average magnitude	η value	b value	Related area magnitude
ARE	0.01787	0.03098	0.03088	0.03052	0.02778	0.03972	0.02315
MSE	0.00322	0.00264	0.00371	0.00332	0.00261	0.00595	0.00187
R1i		1.73364	1.72829	1.70768	1.55434	2.22260	1.29568
R2i		0.81848	1.14989	1.02900	0.80903	1.84471	0.58094
factor ranking		4	2	3	5	1	6

Based on the analysis results of Table 2, this study can conclude that the b value is a key factor affecting earthquake magnitude, while other factors such as cumulative frequency, average magnitude, η value, and magnitude of related areas also have some influence, but their impact is relatively small compared to the b-value. These findings provide important reference for further optimizing RF models and help better understand the roles and importance of various factors in earthquake magnitude prediction.

4. Conclusion

This study explores the application of machine learning models for earthquake prediction in China, focusing on logistic regression and RF approaches. Despite the inherent challenges and uncertainties in earthquake forecasting, this research offers novel insights by applying these techniques to analyze earthquake distribution, identify high-risk seismic zones, and predict earthquake magnitudes. The primary contributions include advancements in pinpointing high-risk areas and demonstrating the utility of machine learning in seismic analysis. Specifically, the study highlights the b value as a critical factor for earthquake magnitude prediction, suggesting its importance for future investigations.

However, the study faces several limitations. The selection of indicators is restricted to six seismic activity parameters, primarily based on earthquake catalogs, which limits the scope of geological factors considered. Additionally, the use of empirical Bayesian statistical methods simplifies the analysis and overlooks more complex data nuances. While the RF model shows promise, the study did not fully compare its performance with the BP model, which also demonstrated advantages in certain aspects. To address these limitations, future research should include a broader array of seismological indicators and precursor signals. Expanding the range of data and models used could lead to more accurate earthquake forecasting and a better understanding of seismic phenomena.

References

[1] Koronovskii N.V. Zakharov V.S. and Naimark A.A. The unpredictability of strong earthquakes: New understanding and solution of the problem. Moscow University Geology Bulletin, 2021, 76: 366 - 373.

[2] Liu R. et al. The performance quality of LR, SVM, and RF for earthquake-induced landslides susceptibility mapping incorporating remote sensing imagery. Arabian Journal of Geosciences, 2021, 14: 1 - 15.

[3] Ridzwan N. S. M. Yusoff S.H.M. Machine learning for earthquake prediction: a review (2017–2021). Earth Science Informatics, 2023, 16 (2): 1133 - 1149.

[4] Bo J.S. et al. A prediction method for strong earthquake surface rupture based on logistic regression analysis. Earthquake Engineering and Engineering Vibration, 2019, 4: 1 - 7.

[5] Xu C. and Xie X.W. Application and verification of logistic regression model in the risk assessment of Yushu earthquake landslide. Journal of Engineering Geology, 2012, 20 (3): 326 - 333.

[6] Shi X.Y. Research on earthquake prediction based on machine learning regression algorithm and its application in China Earthquake Science Experimental Site MS Thesis. Institute of Earthquake Prediction, China Earthquake Administration, 2021.

[7] Budiman K. Ifriza Y.N. Analysis of earthquake forecasting using random forest. Journal of soft computing exploration, 2021, 2 (2): 153 - 162.

[8] Zhang W. He Y. Wang L. et al. Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing. Geological Journal, 2023, 58 (6): 2372 - 2387.

[9] Jang J, So B.D. Yuen D.A. A machine learning algorithm with random forest for recognizing hidden control factors from seismic fault distribution. Geosciences Journal, 2023, 27 (1): 113 - 126.

[10] Historical Inquiry, “China Earthquake Networks Center”, 2024, www.ceic.ac.cn/history.

[11] Zhang X.B. Wu H. Xie X.N. et al. Performance Comparison of Neural Network Prediction Models in Seismic Response of High-speed Railways. Journal of Xiangtan University (Natural Science Edition), 2023, 45 (04): 107 - 117.