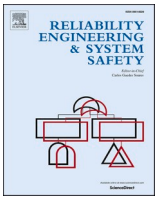
Reliability Engineering and System Safety 225 (2022) 108645

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



An automated machine learning approach for earthquake casualty rate and economic loss prediction

Weiyi Chen [a], Limao Zhang [b,*]

[a] School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[b] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Hongshan District, Wuhan, Hubei 430074, China

ARTICLE INFO

Keywords:
Automated machine learning
Earthquake casualty rate
Economic loss prediction
Seismic loss
Two-step prediction

ABSTRACT

This study presents an automated machine learning (AutoML) framework to predict the casualty rate and direct economic loss induced by earthquakes. The AutoML framework enables automated combined algorithm selection and hyperparameter tuning (CASH), reducing the manual works in the model development. The proposed AutoML framework includes 5 modules: data collection, data preprocessing, CASH, loss prediction, and model analysis. The AutoML models are learned from the dataset that is composed of earthquake information and social indicators. The optimal algorithm and hyperparameter setting of models are determined by the CASH module. A two-step model including a classifier and a regression model is designed for the casualty rate to address zero-casualty cases and also minimize their impacts on data distribution. The proposed AutoML framework is implemented on the seismic loss dataset of mainland China to demonstrate its practicability. A comparison study is conducted to show the high predictive abilities of the AutoML model compared with the traditional seismic risk model and other AutoML models. Models learned from the complete dataset achieve the ultimate performance compared with subsets that are composed of partial features. The model interpretation results indicate that earthquake magnitude, position, and population density are leading indicators for loss prediction.

1. Introduction

Earthquakes have caused numerous losses to societies in terms of human well-being and economics. From 2000 to 2019, earthquakes have resulted in about 721 thousand deaths and $636 billion in economic losses worldwide [1]. To minimize the losses, it is essential to dispatch rescue groups and supplies to stricken areas immediately when earthquakes happen. The allocation of resources will refer to the estimation of loss conditions in affected areas [2]. Therefore, the rapid and accurate prediction of casualties and economic losses after earthquakes is a critical issue for earthquake disaster management.

In order to estimate the earthquake losses, several studies attempted to develop prediction models. These models can be classified into two categories. The first category is the regional-based approach that derives the regression model to predict the earthquake loss. For example, Prompt Assessment of Global Earthquakes for Response (PAGER) system expressed the fatality rate of earthquakes as a lognormal cumulative distribution function of shaking intensity [3]. Guettiche et al. [4] derived linear regression models of casualty numbers with earthquake intensity, and derived the model of economic loss related to the number of damaged and unit Gross Domestic Product (GDP). Cui et al. [5] proposed an ensemble learning model to predict the number of casualties with the input of earthquake information and features of disaster areas. Another category estimated the social and economic loss based on building damage. For example, HAZUS established the table of casualty rates concerning building structure types and damage states [6]. Ceferino et al. [7] estimated the joint probability distribution of casualty rates across different building damage states based on the loss data of the Mw 8.8 Lima earthquake. Generally, the building damage-based approach can provide more detailed estimation results. However, these approaches require spatial distribution of casualties and buildings, which is not available in most areas. Additionally, they also overlook seismic losses that are not related to building damage. Therefore, we focus on the regional-based models in this research.

For most regional-based models, a common limitation is that earthquake cases with zero-casualty are not considered. The overlook of zero-casualty points may result in missing valuable information and overestimation of cases without fatalities [8]. However, a large portion of zero-casualty cases in the dataset will lead to the zero-inflation issue for regression models, where the probability of zero points may be

* Corresponding author.
E-mail address: zlm@hust.edu.cn (L. Zhang).

https://doi.org/10.1016/j.ress.2022.108645
Received 13 August 2021; Received in revised form 29 April 2022; Accepted 2 June 2022
Available online 3 June 2022

underestimated [9]. Additionally, zero points affect the distribution of nonzero points, introducing bias when estimating statistical parameters and their standard deviations, deteriorating the prediction performance [10,11]. To tackle this issue, a two-step machine learning model composed of a classifier and a regression model is proposed for the casualty estimation, where the classifier identifies the presence of casualties, and the regression model predicts the casualty rate for nonzero-casualty cases.

To develop high-performance machine learning models for earthquake loss estimation, combined algorithm selection and hyperparameter tuning (CASH) are crucial steps [12]. In most practical applications, it is difficult to identify the ultimate algorithm directly. The algorithm selection from a pool of machine learning models is required. Moreover, acting as support tools for decision-makers, these loss estimation models are supposed to update automatedly with new cases. Therefore, automated machine learning (AutoML) is introduced to develop prediction models. AutoML refers to a framework that is capable of developing machine learning models with automated algorithm selection and parameter optimization [13]. The AutoML can facilitate data scientists completing the CASH work at a faster speed, which boosts productivity. Some AutoML frameworks have been developed in recent years, such as AutoWEKA [12], Auto-sklearn [14], tree-based pipeline optimization tool (TPOT) [15], and h2o [16]. Although with superior performance, these frameworks are designed for conventional one-step models. Therefore, an AutoML framework that is capable of developing multi-step machine learning models is proposed.

Research questions of this study lie in (1) What are influential indicators that should be selected to construct an effective indicator set to predict seismic loss; (2) How to develop an AutoML framework that meets the requirement of a multi-step machine learning model. This research attempts to propose an AutoML framework for earthquake casualty and economic loss prediction, with both zero-casualty and nonzero casualty earthquake cases. The research contributes to the following two aspects: (a) The state of the knowledge by enabling the proposed AutoML to implement algorithm selection and hyperparameter optimization for multi-step machine learning models. A two-step model is designed for the casualty rate prediction to deal with the zero-inflation issue. (b) The state of practice by developing prediction models for earthquake casualty rate and economic losses using AutoML. The proposed model considered the social loss and economic loss simultaneously. The analysis that involves both casualty and economic loss can deepen the understanding of the correlation between these two variables.

The remainder of the paper is organized as follows. Section 2 describes the construction of an indicator framework for earthquake loss prediction. Section 3 introduces the proposed AutoML framework with the involved techniques. Section 4 presents the collected dataset and the analysis of empirical study results. Section 5 concludes the highlights and limitations of the study.

2. Related studies

The rapid estimation model for seismic losses is able to assist the implementation of research and rescue work, as well as the assignment of relief supplies when earthquakes happen. To achieve a successful machine learning prediction model, it is essential to select appropriate indicators. With the complex mechanism, the seismic losses in the earthquake-stricken areas are affected by indicators from multiple dimensions [17]. In this section, a literature review of earthquake socio-economic loss modeling is presented.

It is widely accepted that the loss caused by earthquakes is related to seismic hazard, exposure, and vulnerability. Most earthquake risk models will consider these 3 elements [18]. For example, The PAGER model includes a vulnerability function that presents the fatality rate with given earthquake intensity. With the input of earthquake intensity (hazard) and exposed population (exposure), the number of fatalities can be estimated (risk) [3]. Jaiswal and Wald [19] derived the model of the economic loss ratio as a function of earthquake intensity. The total economic loss can be estimated by multiplying the loss ratio by the exposed GDP. The earthquake hazard as the source of losses is included in all risk models. Measurements such as the Richter magnitude scale, the Modified Mercalli Intensity, and peak ground acceleration are usually utilized to quantify earthquake hazards [20,21]. The population and GDP are commonly used metrics to present the social and economic exposure [22]. Xing et al. [23] transferred the occurrence time into in-building probability as a factor of exposure.

The establishment of the vulnerability model in the risk prediction varies by assumption. There is a branch of models that estimate the seismic loss based on the building damage, since the building damage and collapse are major sources of social and economic losses. These models are composed of both the building vulnerability model and the socioeconomic vulnerability model. The HAZUS software is an example that firstly predicts the damage state of buildings using the fragility curve of buildings, and then determines the probability of casualties and estimates the cost of building damage [6]. The FEMA-P 58 method establishes the fragility curve of both structural and nonstructural components of building structures, and then estimates the repair cost of different components according to the damage states [24]. Ceferino et al. [25] derived the joint distribution of multiseverity casualties under the given state of building damage based on the historical data of the Mw 8.8 Lima earthquake. These models can map the social and economic loss to building inventories. However, losses that are not related to building are ignored, e.g., destroy of agriculture, damage of infrastructures, and casualties in the outdoor. Furthermore, the detailed distributions of buildings and casualties are required by the building-based risk model. The information, especially the casualty distribution, is not available for most areas. Due to the variation in demographics and living patterns, the relationships between the casualty ratios and the building damage suggested by existing models may be not applicable to other areas.

Another branch of approaches is the regional-based method that estimates the seismic loss of an area directly. Due to the disparity in the physical and social environments of regions, differences exist in the conditional probability of the seismic loss. Local indicators can be adopted to calibrate the vulnerability model for different areas. Xing et al. [23] and Yang et al. [26] adopted binary variables to indicate whether the epicenter is an urban or rural area. Guettiche et al. [4] used latitude and longitude to express the location of the epicenter, since the location implies the vulnerability state of areas. Social and economic indicators can also represent the seismic vulnerability and resistance Huang and Jin [27] introduced population density into the casualty prediction model to calibrate the regression curve. Age structure is also an influential factor considered by some models, since the children and the elder are usually more vulnerable in earthquakes [28]. The dependency ratio that reflects the ratio of people cared for by others is often selected to present the age structure, where the age 15 – 64 is usually defined as the working age, and age $\leq$ 14, and age $>$ 65 are defined as dependent [29]. Wang et al. [30] derived the fatality rate model as a function of peak ground acceleration, where the vulnerability function is calibrated by selected variables from the World Development Indicators to represent the effect of social vulnerability. Jaiswal and Wald [31] introduced Human Development Index and climate features to cluster countries with similar seismic vulnerability, so that the risk models of countries with few earthquake records can be approximated by referring to others with high similarities. Compared with the building-based models, the regional-based models can involve the social indicators more easily.

With the increase in vulnerability indicators, the traditional statistical regression tools can hardly fit the complex and nonlinear data. Therefore, there is a rising trend to apply machine learning or deep learning techniques to develop the risk model [32]. Xing et al. [33] developed the extreme learning model to predict earthquake casualty

Table 1
The descriptions of selected indicators.

Dimension	Indicator	No.	Description	References
Earthquake information	Magnitude	A_1	The magnitude of the earthquake (Mw)	[5,36]
	Latitude	A_2	The latitude of the epicenter (°)	[4]
	Longitude	A_3	The longitude of the epicenter (°)	[4]
	Depth	A_4	The depth of the hypocenter (km)	[5]
Demographic	Population density	B_1	The number of people per unit area for the earthquake-affected area, calculated by Total population in the disaster area/ Total area of the disaster area (p/km^2)	[4,5,27,30,37]
	Dependency ratio	B_2	The ratio between population who are typical as labor force and population who are typically not labor force, calculated by Population ($15 \leq$ age ≤ 64)/ Population (Age $\leq$14 or Age $\geq$65) $\times$ 100% (%)	[37,38]
	In-building probability	B_3	In-building probability of the disaster area, determined according to Table 2.	[23]
Socioeconomic	GDP per capital	C_1	GDP of earthquake-stricken area/ Population (yuan/p)	[4,39]
	Illiterate rate	C_2	The ratio of the illiterate population for the population of 15 years and above (%)	[40]
	Geological hazard prevention	C_3	The funding for the prevention and treatment of geological hazards (10 thousand yuan)	[41]
	Hospital bed	C_4	The number of hospital beds per 1000 population	[37,42]
Loss	Casualty rate	L_1	The number of injuries and deaths/Total population of the disaster area	
	Normalized direct economic loss	L_2	The economic loss of buildings, infrastructure, and agriculture normalized by GDP of the affected area	

for China. Cui et al. [5] proposed an ensemble learning approach learned from earthquake casualty records in China. Chen and Zhang [34] built an ensemble learning model to estimate the magnitude of the seismic vulnerability, where the built model for the overall building damage in a Nepal case study achieves the highest macro f1 score of 0.72. The shortage of machine learning and deep learning techniques in practical applications is that the model is a black box. It is difficult to implement the interpretation and uncertainty analysis on the model directly. To tackle these issues, SHapley Additive exPlanations (SHAP) and bootstrap techniques are adopted in the proposed model.

Referring to related studies, this research proposes a regional-based seismic risk model with the AutoML technique. An indicator framework consisting of 11 input indicators and 2 target variables is constructed as presented in Table 1. These 11 input indicators are categorized into 3 dimensions: earthquake information, demographic, and socioeconomic indicators, where the earthquake information is related to the hazard, while demographic, and socioeconomic indicators represent the vulnerability. Target variables are casualty rate and normalized direct economic loss. The casualty rate L_1 is determined as the ratio of the injuries and deaths induced by the earthquake to the population of the disaster area, where the disaster area is defined as the area with Modified Mercalli intensity equal to or larger than VI [5]. Direct economic loss is defined as the estimated repair or replacement cost for damaged assets, including buildings, infrastructures, agriculture, etc. [35]. The direct economic loss is normalized by the GDP of the affected area to obtain L_2.

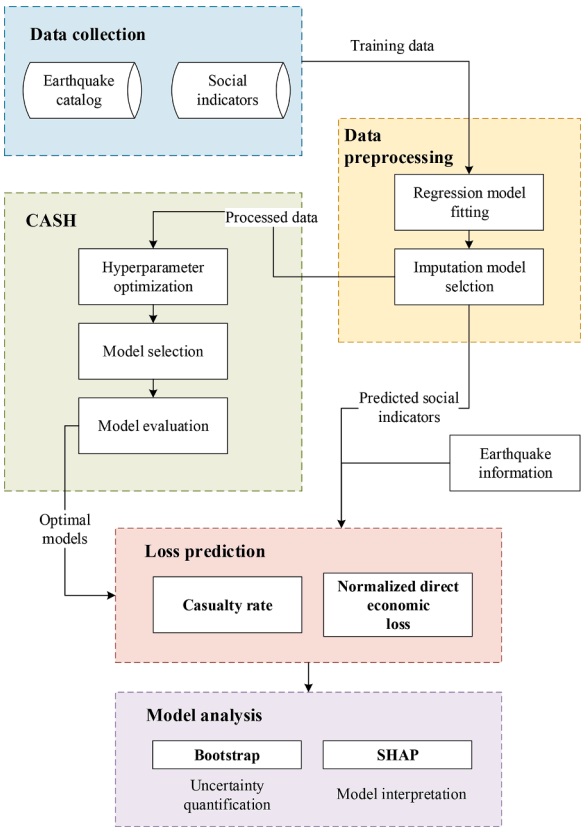


Fig. 1. Flowchart of the proposed AutoML framework for the seismic loss prediction.

3. Methodology

When developing seismic risk models, the dataset of seismic loss will be updated when destructive earthquakes occur. To ensure the prediction models are reliable, a timely update of learning models is required. AutoML techniques enable the automated development of machine learning models with only input of dataset, dispensing with further operations by users. The techniques avoid the cumbersome manual works in the CASH stage, and also lower the threshold for users with little knowledge of machine learning.

In consideration of the adverse impacts of zero-casualty data on the regression model of social loss, a two-step machine learning model composed of a classifier and a regressor is proposed. Specifically, the classifier is learned from the complete training set to distinguish whether casualties are induced. The regression model is trained from the dataset with only non-casualty cases, since zero points may affect the prediction for nonzero points [10]. When predicting casualty rate, earthquake information is firstly input into the classifier to determine the existence of casualties. Then the regression model predicts the casualty rate for non-zero cases. Considering that existing AutoML

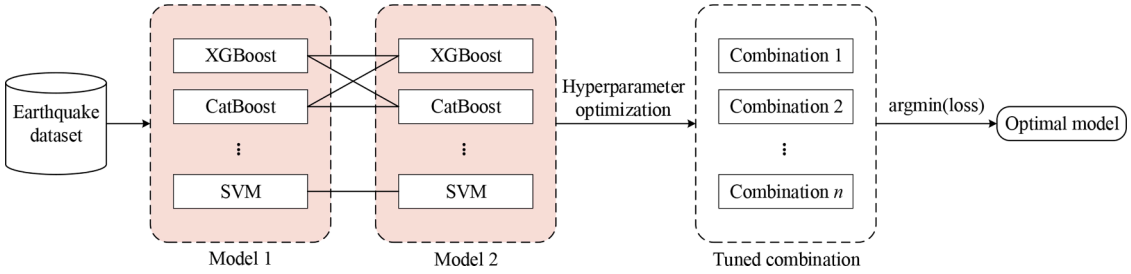


Fig. 2. Flowchart of the CASH module for the proposed framework.

packages are not flexible to meet the requirement of the two-step model, a novel AutoML framework is proposed.

Fig. 1 illustrates the flowchart of the proposed AutoML framework. The framework is composed of 5 modules: data collection, data preprocessing, CASH, loss prediction, and model analysis. In the data collection module, the seismic loss data and social indicators are collected according to Table 1. In the data preprocessing module, regression models are derived for social indicators to impute missing values. In the CASH module, models with the optimal algorithm and hyperparameters combination are selected by Bayesian optimization. In the loss prediction module, seismic losses can be predicted with the input of magnitude of the earthquake, location, and depth of hypocenter, earthquake occurrence time as well as social indicators. Finally, bootstrap and SHAP are implemented for the uncertainty assessment and model interpretation.

3.1. CASH

The model selection and parameter tuning are essential steps for achieving high-performance machine learning models. A CASH module is designed to implement the optimization process automatically. It is a pipeline that enables users to obtain well-developed learning models with only input of training data and essential metadata [43]. Fig. 2 presents the proposed CASH module. Given a set of machine learning algorithms $\boldsymbol{A} = \{A_1, A_2, \ldots, A_m\}$, the combinations of these algorithms $\boldsymbol{C} = \{C_1, C_2, \ldots C_n\ (n = m^2)\}$ are candidates for the optimal two-stage model. These machine learning models are learned from the input dataset D which is split into a training set D_{train} and a validation set D_{val}. The hyperparameter optimization is implemented for all combinations to obtain tunned models for all combinations as presented by Eq. (1) [13]. Then, the optimal combination is selected from tunned combinations by Eq. (2).

$$C_i^* = argmin\ L\left(C_i^j,\ D_{train},\ D_{val}\right) \tag{1}$$

$$C^* = argmin\ L\left(C_i^*, D_{train}, D_{val}\right) \tag{2}$$

where C_i^* is the tunned model for the ith combination, C_i^j is the jth iteration of the hyperparameter optimization, $L(C_i^j,\ D_{train},\ D_{val})$ is the loss function of the model C_i^j, and C^* is the optimal combination with parameter tuning selected for the two-stage model.

3.1.1. Machine learning algorithms

AutoML pipeline selects algorithms from the machine learning algorithm set. The algorithm set of the proposed framework is constituted of support vector machine (SVM), k-nearest neighbors (kNN), random forest (RF), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost, and Light Gradient Boosting Machine (LightGBM). Python packages scikit-learn [44], xgboost [45], catboost [46], and lightgbm [47] are adopted to implement these algorithms. Brief descriptions of these base learners are presented below.

(1) SVM

SVM is an algorithm that is initially developed for linear classification. The basic principle of SVM is to divide points into two classes by a linear plane, where the linear plane can be expressed as $g(x) = w^T x + b$. The plane with the maximal margin between two classes is determined as the optimal classifier. Given a two-class classification problem with a training dataset $D = \{x_i, y_i |\ i = 1, 2, \ldots, n\}$, the goal of SVM is to minimize the objective function given by Eq. (3), with two constraints of Eqs. (4) and (5) [48]. SVM can also be generalized to the regression problem, known as support vector regression (SVR). Similarly, the model of SVR is a plane that can be written as $g(x) = w^T x + b$. The goal of SVR is to find the plane with the narrowest tube (minimum prediction error) to contain all points. The objective function of SVR is given by Eq. (6), with constraints of Eqs. (7)–(9).

$$J(w, b, \xi) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n} \xi_i \tag{3}$$

$$y_i\left[w^T x_i + b\right] \geq 1 - \xi_i \tag{4}$$

$$\xi_i \geq 0 \tag{5}$$

$$J(w, b, \xi) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n} \xi_i + \xi *_i \tag{6}$$

$$y_i - w^T x_i \leq \varepsilon + \xi_i^* \tag{7}$$

$$w^T x_i - y_i \leq \varepsilon + \xi_i \tag{8}$$

$$\xi_i, \xi_i^* \geq 0 \tag{9}$$

where w and b are coefficients of the plane, ξ_i and ξ_i^* are slack variables to allow the existence of data points out of boundaries, ε represents the prediction error of the regression model, C is a regularization term to control the tolerance to the error.

SVM is initially designed for linear problems with one-dimensional features. When facing nonlinear problems with high-dimensional space, kernel trick is introduced. With the kernel function, features are mapped to a hyperplane [49]. The radial basis function (RBF) kernel is used in this study as presented by Eq. (10) [50].

$$\kappa\left(x_i, x_j\right) = \exp\left(-\frac{\| x_i - x_j \|^2}{\theta^2}\right) \tag{10}$$

where $\kappa(x_i, x_j)$ is the kernel function, θ is the kernel parameter.

(1) kNN

KNN is a simple classification algorithm that assigns a label to a data point according to its k-nearest neighbors. When testing data is input into the kNN model, the distances of training data points to the input sample are calculated and k points with the shortest distances are selected as neighbors, and the dominant label of neighbors is assigned to the testing sample [51]. For regression problems, the regression result is determined by averaging the target variable values of neighbors, as

given by Eq. (11) [52]. The Euclidean distance function is utilized to calculate the distance between points, as presented by Eq. (12).

$$\widehat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{11}$$

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n}\left(x_{1j} - x_{2j}\right)^2} \tag{12}$$

where $\widehat{y}$ is the prediction result of the sample, y_i is the target variable value of the ith nearest neighbor of the sample.

(1) RF

RF is an ensemble learner of decision trees (DTs). The ensemble of DTs reduces the variance compared to the base learner. The final prediction result is the aggregation of all decision trees. In the case of the regression problem, the average value of all trees is utilized, as presented by Eq. (13), while for the classification problem, the majority voting is applied [53].

$$\widehat{y_i} = \frac{1}{T}\sum_{t=1}^{T} \widehat{y_{it}} \tag{13}$$

where $\widehat{y_i}$ is the final prediction result for the ith sample, and $\widehat{y_{it}}$ is the expected result of the ith sample obtained by the tth tree.

(1) XGBoost

XGBoost proposed by Chen and Guestrin [45] is an ensemble learner of regression trees. Based on the gradient boosting decision tree (GBDT), the new regression model is trained in an additive manner. The new tree is learned from residual values of the previous iteration, where the predictive output is the summation of all trees, as presented by Eq. (14). When applying to a classification problem, multiple ensemble trees are generated for all classification categories. The probabilities of the input sample belonging to different categories are determined by ensemble trees, where the category with the highest probability is determined as the predicted label, as presented by Eq. (15). The new regression tree is generated by minimizing the objective function as shown in Eq. (16). Compared with the ordinary loss function, a regularization term is introduced to avoid the overfitting issue by penetrating complex models [54]. The details of the regularization term are presented by Eq. (17).

$$\widehat{y_i} = F(x_i) = \sum_{t=1}^{T} \widehat{y_{it}} \tag{14}$$

$$\widehat{y_i} = \arg\max_k F_k(x_i) \tag{15}$$

$$L(F) = \sum_i l(\widehat{y_i}, y_i) + \sum_t \Omega(f_t) \tag{16}$$

$$\Omega(f) = \gamma\ell + \frac{1}{2}\lambda \| \omega \|^2 \tag{17}$$

where F denotes the function of the ensemble model, $L(F)$ represents the objective function of F, k refers to the kth classification label for the classification problem, $\sum_i l(\widehat{y_i}, y_i)$ is the term of the loss function, and $\sum_t \Omega(f_t)$ is the term of regularization term. ℓ is the number of leaves of the decision tree, γ and λ are assigned coefficients, and $\| \omega \|^2$ is the score of leaves.

(1) CatBoost

CatBoost, proposed by Prokhorenkova et al. [55], is also a GBDT-based ensemble learner. Compared with other gradient boosting algorithms, the significance of CatBoost falls in handling categorical features. The common approach to dealing with categorical features is to convert categorical features into numerical values by encoding them in the data preprocessing stage. CatBoost adopts a more efficient transfer strategy in the training time [56]. The strategy computes statistics for label values of the training data, and then substitutes categorical features with the mean value of the category. Considering that the use of the whole training dataset may lead to overfitting, CatBoost performs a random permutation before substitution. Given a dataset with n samples, $D = \{x_i, y_i | \ i = 1, 2, \ldots, n\}$, a random permutation is implemented for each sample, $\tau = \{\tau_i | \ i = 1, 2, \ldots, n\}$. Assuming that the mapping of the ith sample (x_i, y_i) in the permutation τ_i is the pth sample $(x_{\tau p}, y_{\tau p})$, and the ith sample has a categorical feature k, the category is replaced by the mean value of label values for the samples that are ahead of $x_{\tau p}$ and with the sample category. The computation of substitution value is given by Eq. (18).

$$\widehat{x}_{ik} = \frac{\sum_{j=1}^{p-1}\left[x_{\tau_{j,k}} = x_{\tau_{p,k}}\right] Y_{\tau_j} + qP}{\sum_{j=1}^{p-1}\left[x_{\tau_{j,k}} = x_{\tau_{p,k}}\right] + q} \tag{18}$$

where $\widehat{x}_{ik}$ stands for the substitution value for the categorical feature k of sample x_i, $[x_{\tau_{j,k}} = x_{\tau_{p,k}}]$ finds the sample also with the categorical feature k, P is a prior value to reduce noises induced from categories with low frequencies, q refers to the weight of the prior value.

Furthermore, CatBoost attempts to combine categorical features to improve performance by the greedy algorithm when determining splitting for trees.

(1) LightGBM

LightGBM, proposed by Ke et al. [47], is designed to handle the dataset with large samples and high dimensions. This algorithm achieves a good balance between accuracy and computational efficiency by reducing the number of samples and number of features using Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques, respectively.

GOSS reduces the number of samples by removing part of instances with a small gradient since it is believed that instances with small gradients have been well trained. Instances are firstly ranked by their absolute gradients. The top $a \times 100\%$ instances are identified as the large gradient to form an instance set A, and then the remaining $(1\text{-}a) \times 100\%$ instances are small gradients. $b \times 100\%$ instances are random samples from the small gradient instances to form an instance set B. Given a dataset $D = \{x_i, y_i \mid i = 1, 2, \ldots, n\}$, the absolute gradients of the loss function for instances are denoted as $\{G_i | i = 1, 2, \ldots, n\}$. The estimated information gain of a candidate split point x^s with GOSS is determined by Eq. (19).

$$\widetilde{V}_j(x^s) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} G_i + \frac{1-a}{b}\sum_{x_i \in B_l} G_i\right)^2}{n_l^j(x^s)} + \frac{\left(\sum_{x_i \in A_r} G_i + \frac{1-a}{b}\sum_{x_i \in B_r} G_i\right)^2}{n_r^j(x^s)}\right) \tag{19}$$

where $A_l = \{x_i \in A, \text{ and } x_{ij} < \mathrm{d}\}$, $A_r = \{x_i \in A, \text{ and } x_{ij} > x^s\}$, $B_l = \{x_i \in B, \text{ and } x_{ij} < x^s\}$, $B_r = \{x_i \in B, \text{ and } x_{ij} > x^s\}$, $n_l^j(x^s)$ and $n_r^j(x^s)$ are numbers of instances on the left and right sides of the split point, respectively.

LightGBM further reduces the size of the dataset by bundling features together. Considering a common characteristic of high-dimensional data is that many features are mutually exclusive, the EFB technique is proposed to bundle exclusive features. In this way, the dimension of the dataset is reduced, so that the computational efficiency is improved.

Algorithm 1
Sequential model-based optimization.

Input: *Space*, *f*, M_0, *Acq*
$R = \varnothing$
For t in 1 to T
  $h_t^* = \underset{h \in Space}{\operatorname{argmax}} Acq(h, M_{t-1})$
  Evaluate $f(h_t^*)$
  $R = R \cup (h_t^*, f(h_t^*))$
  $FitModel(M_t, R)$
Return h_T^*

3.1.2. Hyperparameter optimization

Normally, the training of machine learning models requires users to set up necessary hyperparameters. The manual searching for optimal hyperparameter settings needs a large amount of human effort. The automated hyperparameter optimization is able to reduce repetitive work, enhance the performance of models, and improve the reproductivity of the research. Assuming that h is a combination of hyperparameters with the searching space H, the objective function is f, the goal of the hyperparameter optimization is to determine the optimal hyperparameter configuration by Eq. (20) [43].

$$h^* = \underset{h \in H}{\operatorname{argmin}} f(h) \tag{20}$$

Considering the non-convex nature of the hyperparameter optimization problem, global optimization is preferred [57]. Bayesian optimization as a state-of-the-art global optimization technique has shown outstanding performance compared with other global optimization methods [58]. Therefore, Bayesian optimization is adopted to determine the optimal hyperparameter configuration. The Bayesian optimization of hyperparameters is implemented by the Python package Hyperopt [59].

The basic principle of Bayesian optimization is to develop a surrogate model M for hyperparameter h and corresponding $f(h)$ based on point estimation of f and prior information, and then provide suggestions of hyperparameter configurations for the next step [60]. The implementation of Bayesian optimization for hyperparameter optimization can be realized by the Sequential model-based optimization (SMBO) algorithm as presented by Algorithm 1 [61]. The algorithm requires inputs of configuration space of hyperparameters, objective function f, acquisition function Acq, and initial surrogate model M_0. In the algorithm, the T-iteration loop is implemented to find the optimal hyperparameter configuration. In each iteration, a promising hyperparameter setting h_t* is determined by the acquisition function with the surrogate model $M_{t\text{-}1}$. A dataset R records the promising hyperparameter and the corresponding objective function (h_t*, $f(h_t$*)). Then, the surrogate model M_t is fitted by the updated R.

In this research, the expected improvement (EI) is adopted as the acquisition function which is given by Eq. (21). Tree Parzen Estimator (TPE) is adopted to estimate the EI. Rather than calculating $p(f|h)$ directly, TPE obtains the value of EI by calculating $p(h|f)$ and $p(f)$. Herein, $p(h|f)$ is defined by the density of samples by Eq. (22).

$$Acq(h, M_t) = EI(h) = \int_{-\infty}^{\infty} \max\left(f^{min} - f, 0\right) \cdot p_{M_t}(f|h) df \tag{21}$$

$$p_{M_t}(h|f) = \begin{cases} g_l(h) & if\ f < f^{min} \\ g_r(h) & if\ f \geq f^{min} \end{cases} \tag{22}$$

where f^{min} is the best value obtained by the objective function in previous iterations by observing R, p_{Mt} is the probability distribution of surrogate model M_t. $g_l(h)$ is the density distribution formed by samples with corresponding y smaller than f^{min}, and $g_r(h)$ is the density distribution of samples with corresponding y equal to or larger than f^{min}.

3.2. Data imputation

For prediction models with social indicators, a common issue is raised that the information for the current year is not available most of the time. Moreover, due to unexpected situations, part of the data may be not recorded, resulting in the problem of missing data in the model development process. To resolve these problems, missing values of social indicators are imputed by regression models derived from the collected dataset. Social indicators (B_1, B_2, C_1, C_2, C_3, and C_4) are expressed by mathematical functions in terms of the year. The linear regression and polymer regression are selected as presented by Eqs. (23) and (24), respectively. Considering the distribution of data, log transformation may be required for independent variables or dependent variables to achieve good fitting. The best-fit regression curves are identified by minimizing the loss function metrics, as presented by Eq. (25).

$$y = u + v \cdot x \tag{23}$$

$$y = u + v \cdot \mathrm{x}^2 \tag{24}$$

$$M^* = argmin\ L(M_i, D) \tag{25}$$

where u and v are parameters that are to be determined by model fitting, M^* refers to the optimal regression curve, M_i is the ith candidate regression model, and D is the dataset used for curve fitting.

In order to determine the best parameter set that fits the dataset, the least square method is adopted. The goal of the least square method is to minimize the squared error, as presented by Eq. (26) [62].

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{26}$$

where y_i is the true observed value, and $\hat{y}_i$ refers to the predicted value by the regression model.

3.3. Model evaluation

The AutoML and regression models are selected by comparing objective function values among candidate models. Mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2) are commonly used evaluation metrics for regression models. Mathematical definitions of these 3 metrics are presented by Eqs. (27)–(29), respectively [5]. Among these metrics, R^2 describes the explanatory ability of the derived model, which is not frequently used as the objective function [27]. RMSE is more appropriate for models with errors following normal distribution and sensitivity to extreme errors, while MAE performs higher resistance to errors [63]. Since the goal of this study is to develop seismic loss prediction models for all earthquake scenarios, MAE that can relieve impacts of extreme values due to catastrophic earthquakes is selected as the objective function. RMSE and R^2 are also computed as quantitative measurements for the prediction abilities of models.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{27}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{28}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{29}$$

where $\bar{y}$ is the mean value of observations.

Algorithm 2
Bootstrap resampling.

Input: n, D_0
For η in 1 to H
  D_η= Sample n data from the dataset D_0 randomly with replacement
  f^η = Train machine learning model according to D_η
  $\hat{y}_i^{\mathrm{H}} = f(x_i)^{\mathrm{H}}$
Return $\{\hat{y}_i\}_{\eta=1}^{\mathrm{H}}$

3.4. Uncertainty quantification

Uncertainty exists in the AutoML model development process. Since the machine learning models are composed of complex nonlinear models, it is difficult to determine the error of parameter estimation using traditional methods. In this case, the bootstrap technique is commonly used to estimate the distribution of errors [64]. Bootstrap generates sampling datasets from the original dataset. New machine learning models can be learned from the sampling datasets. Then the error of the model can be approximated based on the prediction error of multiple models. The general procedure of the Bootstrap sampling is presented by Algorithm 2 [65]. It is assumed that the expected prediction value of the ith point is the average value of the prediction results of all B models, as presented by Eq. (30). Under this assumption, the model misspecification variance can be computed using Eq. (31). The confidence interval (CI) can be computed using the model misspecification. The variance of model errors can be estimated by Eq. (32) to construct the prediction interval (PI) [66]. Finally, the (1 - α)% CI and PI of the prediction results can be obtained by Eqs. (33) and (34), respectively.

$$\hat{y}_i = \frac{1}{Z}\sum_{\eta=1}^{\mathrm{H}} \hat{y}_i^{\eta} \tag{30}$$

$$\sigma^2_{\hat{y}_i} = \frac{1}{\mathrm{H}-1}\sum_{\eta=1}^{\mathrm{H}} \left(\hat{y}_i^{\eta} - \hat{y}_i\right)^2 \tag{31}$$

$$\sigma^2_{\hat{\epsilon}_{ii}} \simeq E\left\{(y_i - \hat{y}_i)^2\right\} - \sigma^2_{\hat{y}_i} \tag{32}$$

$$CI = \hat{y}_i \pm z_{1-\alpha/2}\sigma_{\hat{y}_i} \tag{33}$$

$$PI = \hat{y}_i \pm z_{1-\alpha/2}\sigma_{\epsilon_i} \tag{34}$$

3.5. Model interpretation

Other than the prediction output, there is also valuable information held behind the well-trained model. Therefore, SHAP, which can be implemented by shap package [67] in Python, is adopted to interpret developed models. SHAP is an additive feature attribution method to measure the feature importance for prediction outputs. It captures the impacts on predictors induced by small changes of inputs for all instances [68]. The SHAP algorithm observes the effects of features on the prediction results by controlling the presence of features. If a feature is present, the true value will be used, while if it is absent, a random value is assigned to this feature from the dataset [69]. For a machine learning model with J input features, SHAP defines the explanation model for the objective machine learning model as a linear function with J binary features. The linear function is denoted by g as shown in Eq. (35).

$$g\left(z_j'\right) = \phi_0 + \sum_{j=1}^{J} \phi_j z_j' \tag{35}$$

where z_j' is a binary variable that represents the feature j being present ($z_j' = 0$) or absent ($z_j' = 1$), ϕ_0 is the base value of the prediction, and ϕ_j is the contribution of the jth feature.

To estimate the contributions of features, the output of a tree under the condition of a feature subset S_j is defined as $f_x(S_j) = [\mathrm{E}(f(x)|x_{Sj})]$. The contribution of the feature, also called the SHAP value, is computed with the conditional output of subsets based on the game theory as presented by Eq. (36).

$$\phi_i = \sum_{S_j \subseteq S\setminus\{j\}} \frac{|S_j|!(|S| - |S_j| - 1)!}{|S|!}\left[f_x\left(S_j \cup \{j\}\right) - f_x\left(S_j\right)\right] \tag{36}$$

where S is the set of all features, S_j refers to the feature subset without the jth feature, f_x ($S_j \cup \{j\}$) refers to the expected outputs of models with the jth feature, and f_x (S_j) refers to the expected outputs of models without the jth feature.

4. Empirical study and results

4.1. Data acquisition and description

Affected by the collision of the Indian Plate with the Eurasian Plate, China is exposed to high seismic hazards [70]. With the growth of population density and rapid development of urbanization, the vulnerability of societies also increases [71]. The combination of high seismic hazard and seismic vulnerability results in heavy seismic losses. As one of the most active seismic countries, China suffered about one-third of the continental earthquakes and over 50% of earthquake deaths globally [72]. Seismic risk management is essential for China to minimize earthquake-induced losses. In this research, the proposed AutoML framework is implemented on the seismic loss dataset of mainland China.

The dataset is composed of 211 destructive earthquakes in mainland China from 2001 to 2019. 198 earthquake events of 2001 – 2018 are selected as the training set. 13 events of 2019 are selected as the testing set to validate the performance of the AutoML. The information about earthquakes (e.g. location, magnitude, occurrence time) and losses are collected from the earthquake catalog and the loss summary reports published by the China Earthquake Administration, where the occurrence time of earthquakes is transferred to in-building probability (B_3) according to Table 2. For events whose affected populations are not reported, the data were estimated with ShakeMap published by USGS [73] and the population density map published by the Center for International Earth Science Information Network [74]. The ShakeMap includes the distribution of earthquake intensity, by which the map of the affected area was obtained. The number of affected people was determined by multiplying the map of the disaster area by the spatial distribution of the population density. The social indicators were extracted from the China Statistical Yearbook published by National Bureau of Statistics of China [75]. The social indiactors are provincial-level data that are collected every year. The data source of the

Table 2
In-building probability of China according to Xing et al. [23].

Place	Day	Time								
		0–6	6–7	7–8	12–14	14–17	17–18	18–19	19–22	22–24
Urban	Workday	1.00	0.84	0.15	0.95	0.95	0.14	0.79	0.79	1.00
	Holiday	1.00	1.00	0.73	0.73	0.73	0.73	0.73	0.73	1.00
Rural	Workday	1.00	1.00	0.30	0.30	0.30	0.30	0.90	1.00	1.00
	Holiday	1.00	1.00	0.75	0.75	0.75	0.75	0.75	1.00	1.00

Table 3
Summary of the data source for the collected dataset.

Source	Data name	Description	Spatial resolution	Temporal resolution	Related indicator
China Earthquake Administration	Earthquake Damage Losses in Mainland of China	Earthquake catalog including earthquake information and loss	Point data of epicenter/ regional data of disaster area	Per earthquake	A_1, A_2, A_3, A_4, B_3, L_1, L_2
USGS [73]	ShakeMap	Spatial distribution of the shaking field	5 km	N.A.	L_1
Center for International Earth Science Information Network [74]	Gridded Population of the World5	Estimated population density distribution	30 arc-seconds	5 year	B_1, L_1
National Bureau of Statistics of China [75].	China Statistical Yearbook	China statistics of population, agriculture, economics, medical service, etc.	Provincial-level data	1 Year	B_1, B_2, C_1, C_2, C_3, L_2

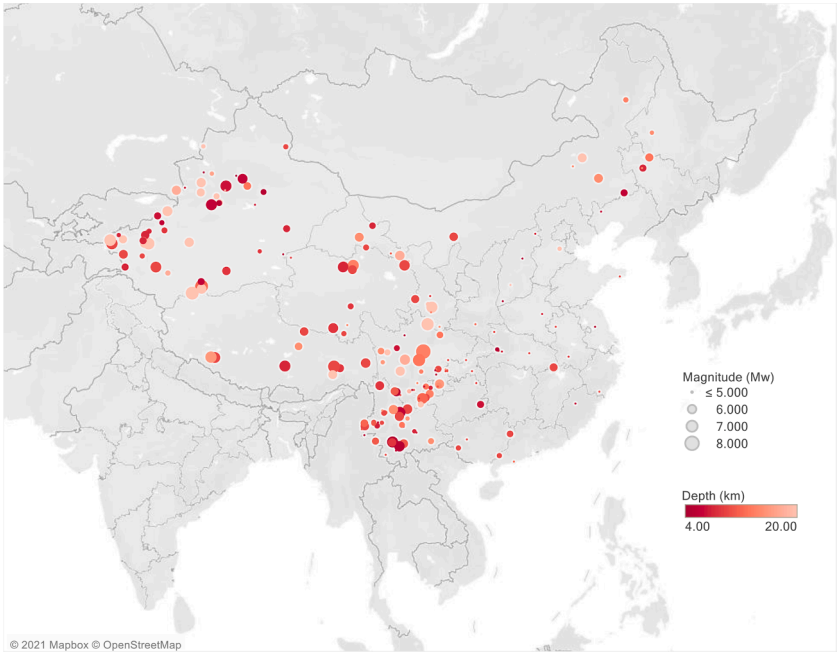


Fig. 3. Distribution of destructive earthquakes in mainland China from 2001 to 2019.

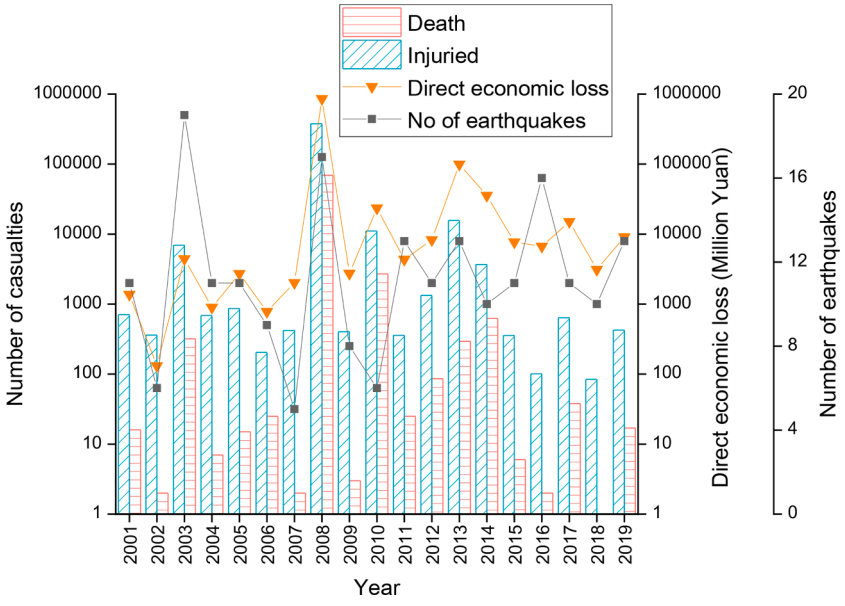


Fig. 4. Number of destructive earthquakes recorded, casualties, and direct economic loss caused by earthquakes in mainland China from 2001 to 2019.

Table 4
Range and sample data of the collected dataset.

Dimension	Indicator	No.	Range	Sample no.			
				1	2	16	17
Earthquake information	Magnitude (Mw)	A_1	[4, 8]	5.3	5.7	7	6.6
	Latitude (°)	A_2	[21.7, 49]	34.56	45.27	33.2	42.11
	Longitude (°)	A_3	[73.9, 125]	96.53	124.71	103.82	83.43
	Depth (km)	A_4	[4, 33]	9	13	20	6
Demographic	Population density (p/km^2)	B_1	[2.2, 738.8]	8.35	144.60	170.79	14.68
	Dependency ratio (%)	B_2	[27.19, 50.45]	37.25	32.73	42.37	43.19
	In-building probability	B_3	[0.3, 1]	0.3	1	1	0.3
Socioeconomic	GDP per capital (yuan/p)	C_1	[4467, 72,730]	45,738	41,516	45,768	46,089
	Illiterate rate (%)	C_2	[1.69, 54.86]	10.24	2.8	7.05	3.19
	Geological hazard prevention (10 thousand yuan)	C_3	[70, 445,280]	–	–	207,000	28,236
	Hospital bed	C_4	[1.87, 7.19]	6.49	6.18	6.79	6.85
Loss	Casualty rate	L_1	[0, 0.0595]	0	0	0.00018	0.00016
	Normalized direct economic loss	L_2	[0.001, 6.462]	0.586	0.064	0.056	0.407

Note: "-" stands for missing data. Earthquakes referred by sample no. are 1. 2018 Qinghai Chindu Earthquake, 2. 2018 Jilin Songyuan Earthquake, 16. 2017 Sichuan Aba Earthquake, 17. 2017 Xinjian Bortala Earthquake.

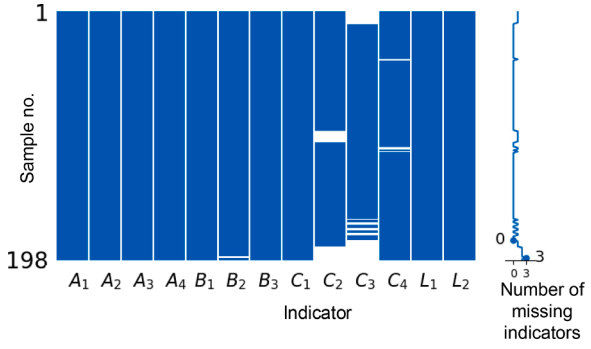


Fig. 5. Visualization of missing data.
Note: The blank indicates that the data is missing for that position. For example, C_3 is missing for sample no. 1. C_2 and C_3 are missing for sample no. 198.

collected dataset is presented in Table 3.

An overview of the earthquake loss for China can be obtained from the collected dataset. The distribution of selected earthquakes is presented in Fig. 3. There is a higher frequency of destructive earthquakes in the western region compared with the eastern region. The summarized seismic losses of the selected earthquakes are plotted by year as presented in Fig. 4. The annual mean deaths, injures, and direct economic loss caused by the selected earthquakes are 346.5, 1987.4, and 527.9 billion yuan, respectively.

The range of each indicator and sample data of the dataset is presented in Table 4. Due to historical or other reasons, part of social indicator records is not available. The locations of missing values are visualized in Fig. 5. In the figure, each column represents an indicator or target variable. The blank in the column implies the indicator is missing for the corresponding sample. For example, it can be observed that C_3 of sample no. 1 is missing; C_2 and C_3 of sample no. 198 are missing. The

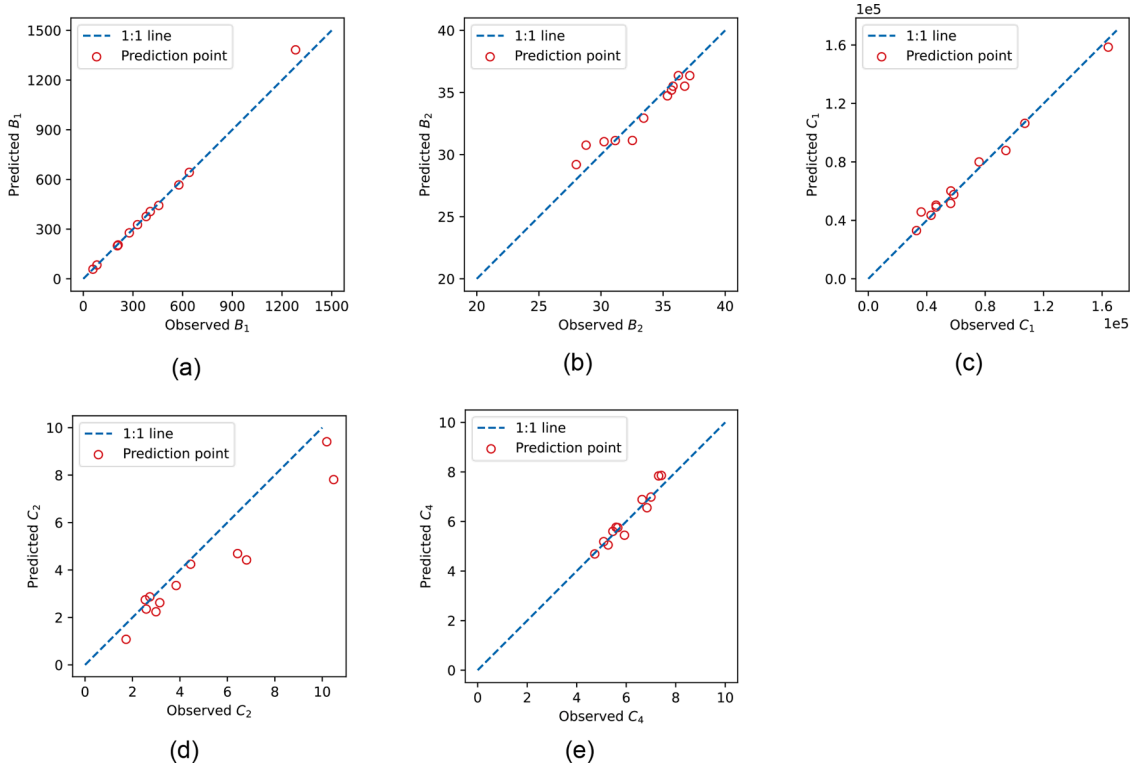


Fig. 6. Imputation results versus the observed results by data imputation models for (a) B_1, (b) B_2, (c) C_1, (d) C_2, (e) C_4.

Table 5
Hyperparameters and searching spaces of algorithms involved in the AutoML.

Algorithm	Hyperparameter and searching spaces
SVM	C = {0.1, 1, 10, 100, 1000} *gamma* = {0.0001, 0.001, 0.1, 1} *eplison** = [0.1, 0.5]
kNN	*n_neighbors* = {2, 3, 4, 5, 6, 7, 8, 9, 10} *leaf_size* = [10, 50] p = {1, 2}
RF	*max_depth* = {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16} *n_estimators* = {20, 30, 40, …, 500}
XGBoost	*max_depth* = {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16} *learning_rate* = [0.01, 0.5] *n_estimators* = {20, 30, 40, …, 500} *gamma* = [0, 0.5]
Catboost	*max_depth* = {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16} *learning_rate* = [0.01, 0.5] *n_estimators* = {20, 30, 40, …, 1000}
LightGBM	*max_depth* = {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16} *learning_rate* = [0.01, 0.5] *n_estimators* = {20, 30, 40, …, 500} *min_child_weight* = [0.0001, 0.1]

missing data issue is present for B_2, C_2, C_3, and C_4 The maximum number of missing indicators for a sample is 3.

4.2. Model development

In order to develop the rapid estimation model with the collected dataset, the AutoML model is developed as presented in the following section. The training dataset is composed of 198 earthquake events from 2001 to 2018. 75 events (37.9%) are zero-casualty cases. The development of AutoML models follows the procedures demonstrated in the Methodology section. Data preprocessing is firstly implemented to deal with the data missing problem. Regression models of social vulnerability indicators including B_1, B_2, C_1, C_2, C_3, and C_4 with respect to year and location are derived referring Eqs. (23) and (24). Social indicators of 2001 – 2018 are used to derive the regression models. To validate the data imputation, observed social indicators and predicted ones for 2019 earthquake events are compared. The comparison results of B_1, B_2, C_1, C_2, and C_4 are plotted in Fig. 6. C_3 is not presented since the data is not published in 2019. The scatter plots of the prediction points of regression models are located close to the 1:1 line, verifying the effectiveness of the imputation models.

After the imputation of the missing data, the CASH module is implemented to develop the optimal AutoML model. The auto-hyperparameter tuning is implemented by Bayesian optimization. The hyperparameters and corresponding searching spaces for machine learning algorithms are presented in Table 5. 5-fold cross-validation is implemented and the average MAE of the 5 folds is evaluated. A 300-iteration Bayesian optimization is conducted to determine the optimal setting of hyperparameters. Fig. 7 presents part of the tuning process for different candidate algorithms. Since the Bayesian optimization attempts to update hyperparameter settings according to the training history, there is a certain trend for the values of hyperparameters over iterations. This optimization strategy is more efficient than the random search and grid search [58]. A set of local optimal models are determined with the Bayesian optimization. The validation set is input into these trained models to obtain the corresponding MAE.

4.3. Results analysis

The AutoML models are developed based on earthquake data of 2001 – 2018. To validate the proposed AutoML framework, the prediction results of seismic loss for 2019 earthquakes are compared with the actual values. Fig. 8 presents model prediction results and actual values of the testing data. The 95% CI and PI of the testing set are plotted in Fig. 9. The prediction results of the proposed model and one-step model

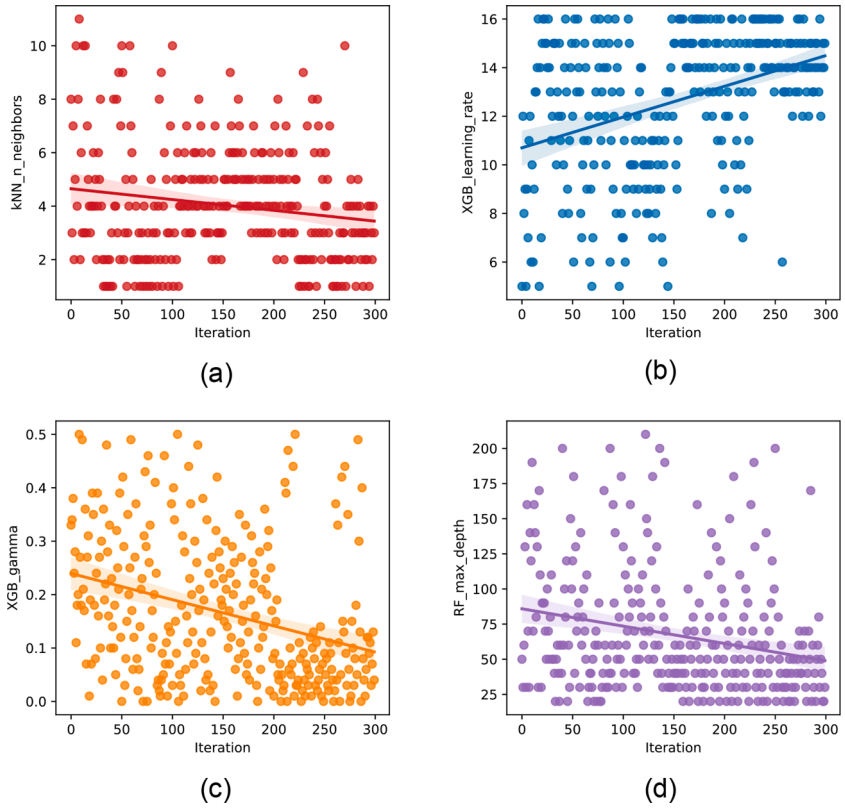


Fig. 7. Hyperparameter tuning of (a) 'n_estimators' of kNN-CatBoost model for casualty rate prediction, (b) 'learning_rate' of kNN-XGBoost model for casualty rate prediction, (c) 'gamma' of XGBoost model direct economic loss prediction, and (d) 'max_depth' of RF model for direct economic loss prediction.

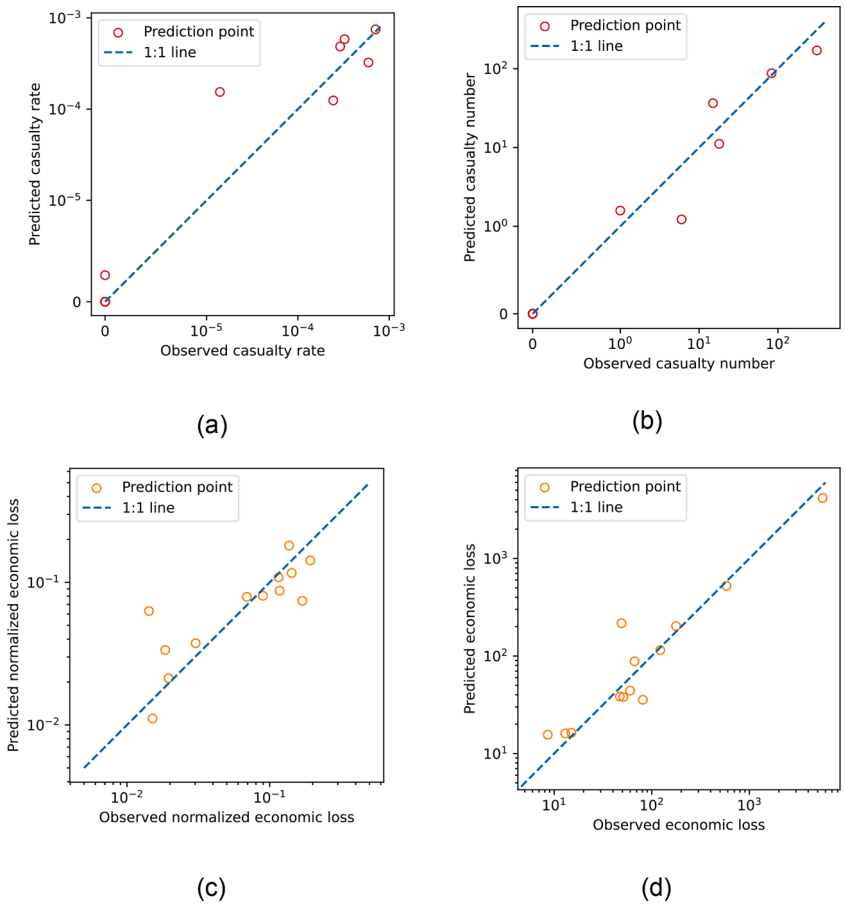


Fig. 8. Predicted results and observed values for (a) casualty rate, (b) casualty number, (c) normalized economic loss, and (d) observed economic loss.

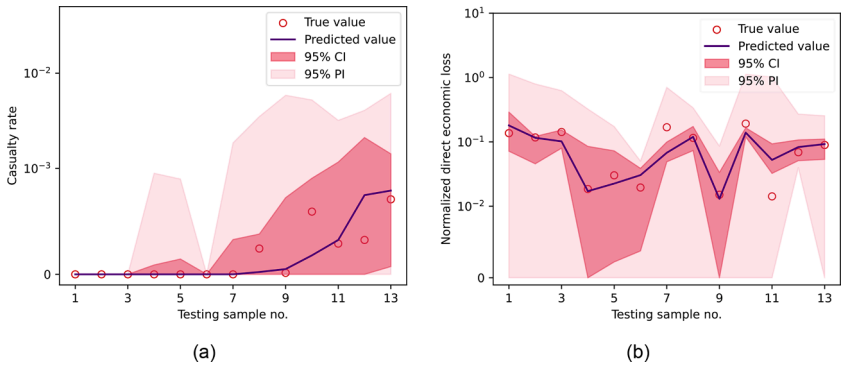


Fig. 9. Confidence interval and prediction interval for (a) casualty rate, and (b) normalized direct economic loss.

are plotted in Fig. 10 to investigate the advantage of the two-step model to deal with the zero inflation problem. A comparative study with other models is implemented, including the one-step model for casualty rate, other AutoML packages (h2o and TPOT), and the traditional seismic loss prediction method (PAGER model). Furthermore, the prediction performances of models learned from data with partial indicators are compared with models with complete information. The prediction metrics of these models for casualty rate and normalized direct economic loss are presented in Tables 6 and 7, respectively. In addition to the predictive analysis, more meaningful information may be implied inside the model. Therefore, model interpretation is conducted for the AutoML models by SHAP analysis. The contribution of features to seismic losses and the model differences between the two target variables are investigated. The results of SHAP analysis are plotted in Figs. 11 and 12. All the results are analyzed in detail as follows.

(1) The prediction points of seismic losses are close to the one-to-one line, as presented in Fig. 8, validating the prediction abilities of the developed models. The prediction results of L_1 and L_2 versus the true values for the testing set are presented in Fig. 8(a) and (c), respectively. The number of casualties is computed by multiplying L_1 by people in the affected area, while the direct economic loss is determined by multiplying L_2 by the GDP of the affected area, where the computation results are plotted in Fig. 8 (b) and (d). The prediction points are proximate to the 1:1 line for both casualty and economic loss, implying a good alignment between the predictions and observations. The prediction errors are acceptable in practical applications. The uncertainty existing

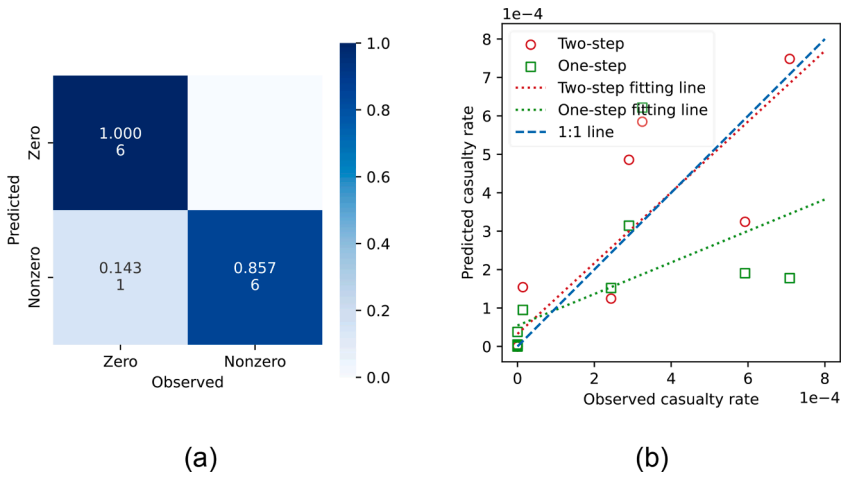


Fig. 10. (a) Confusion matrix of the classification model of the two-step regression model, and (b) prediction results of the casualty rate with one-step regression model and two-step regression model.

Table 6
Prediction metrics of models for L_1.

Metric	Model							
	Proposed	One-step	h2o	TPOT	PAGER	Subset 1	Subset 2	Subset 3
MAE	0.000077	0.000118	0.000128	0.000159	0.000224	0.000121	0.000117	0.000079
RMSE	0.000162	0.000227	0.000215	0.000277	0.000246	0.000226	0.000186	0.000165
R^2	0.656615	0.335250	0.272937	0.515113	0.232813	0.216267	0.436830	0.553222

Table 7
Prediction metrics of models for L_2.

Metric	Model						
	Proposed	h2o	TPOT	Traditional model [19]	Subset 1	Subset 2	Subset 3
MAE	0.024690	0.042493	0.050167	0.059999	0.047870	0.035032	0.037559
RMSE	0.037624	0.053013	0.060031	0.073859	0.052574	0.048364	0.046621
R^2	0.633274	0.469006	0.607067	0.102753	0.328623	0.411107	0.450973

in the prediction results is presented in Fig. 9 in the forms of CI and PI. Generally, although the relationship between uncertainty and variables is complex, there is a tendency that the higher seismic loss is corresponding to a larger uncertainty interval (i.e., a higher degree of uncertainty).

(2) The proposed two-step regression model for L_1 results in better performance compared with the model developed with only a regression algorithm. Considering the effect of zero-point on the distribution of data, a two-step regression model consisting of a classifier to distinguish zero-point samples and a regressor to estimate the value of casualty rate. As observed from Fig. 10(a), the developed classifier shows high accuracy when distinguishing between zero and nonzero cases. The precision for zero cases is 0.857, and 1.0 for nonzero cases. Finally, the two-step model performs higher predictive ability compared with the one-step model consisting of a regression model, as presented by Table 6. The prediction results of the two-step model and the one-step model are plotted in Fig. 10(b) to investigate the difference between these two models more deeply. It can be observed that the fitting line of prediction points of the two-step model is close to the 1:1 line. However, the slope of the fitting line for the one-step model is much smaller than 1. For zero-casualty cases, a small value will be assigned to the prediction results by the one-step model. Meanwhile, the one-step model tends to underestimate the results for high casualty cases, due to the impact of zero-point on the distribution. The combination of classification and regression models relieves this weakness of the one-step model, where the prediction results of the classifier are corrected by the regression model. This strategy is not limited to the casualty prediction of earthquakes, but also other datasets with a large number of zero points.

(3) Compared with other models, the proposed AutoML models achieve the optimal prediction performance. The developed L_1 model by the proposed AutoML framework shows the best MAE, RMSE, and R^2 among all models, as presented in Table 6. The L_2 model by the proposed framework also possesses the smallest MAE and RMSE, and the highest R^2. Since the developed L_1 models of other AutoML packages are also one-step models, similar one-inflation problems also exist. Furthermore, the candidate machine learning algorithms and hyperparameter tunning approaches are limited for the existing AutoML packages, which results in the worse performances of other AutoML packages. Even though worse than the proposed framework, the prediction abilities of AutoML-based models are higher than traditional models. Traditional models are derived without the calibration of social and economic indicators, leading to biased results when predicting.

(4) The AutoML models learned from the complete dataset perform higher predictive abilities than subsets. To validate the effectiveness of the selected indicators, 3 subsets that are composed of partial indicators are constructed, where Subset 1 only involves earthquake information, Subset 2 is a combination of earthquake

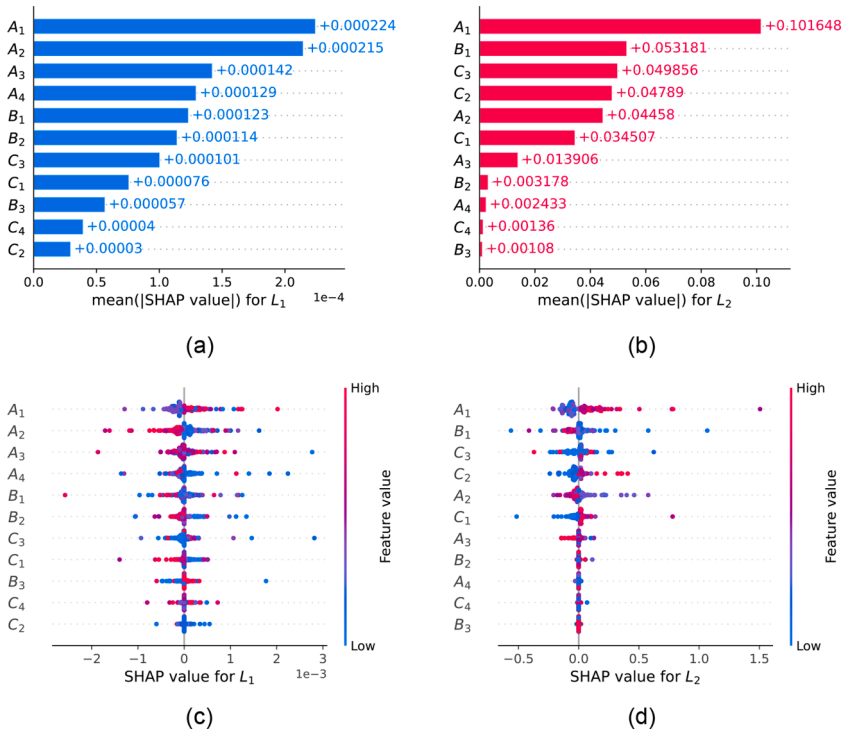


Fig. 11. Mean absolute SHAP values of indicators for (a) L_1 and (b) L_2, and distribution of SHAP values versus indicator values for (c) L_1 and (d) L_2.

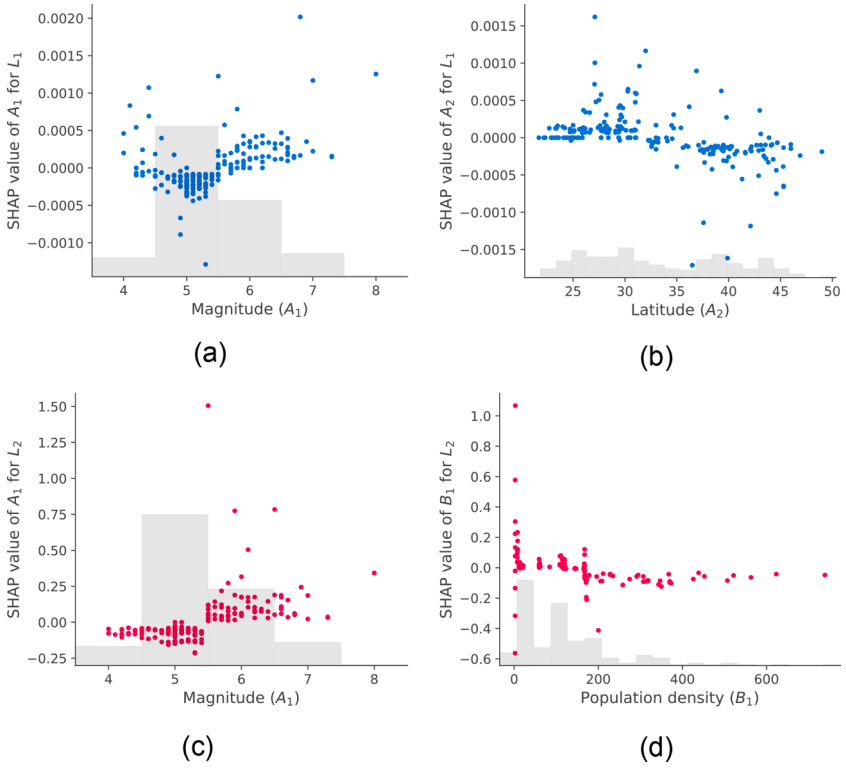


Fig. 12. SHAP values of (a) A_1 for L_1, (b) A_2 for L_1, (c) A_1 for L_2, and (d) B_1 for L_2.

information and demographic indicators, and Subset 3 is composed of earthquake information and socioeconomic indicators. Evaluation metrics of models developed from these datasets are presented in Tables 6 and 7. It can be observed that, for both target variables, models learned from complete indicators result in the lowest MAE, verifying the high effectiveness of the selected feature set. Furthermore, the prediction metrics of Subset 2 and Subset 3 models are better than Subset 1, implying that the introduction of partial social indicators can improve the predictive abilities. For L_1, the improvement by introduction of socioeconomic indicators (Subset 3) is more significant than the demographic indicators (Subset 2), as observed in Table 6.

However, the impacts of the demographic and socioeconomic indicators are close for L_2 model (see Table 7).

(5) The model interpretation results show that the A_1 (Magnitude) is of most importance for both target variables, and the social loss is more sensitive to earthquake information features, while the economic loss is more sensitive to demographic and socioeconomic features. Observed in Fig. 11(a) and (b), A_1 presents the highest mean absolute SHAP values for both casualty rate and direct economic loss model, indicating its dominant position in seismic loss estimation. According to Fig. 11(c) and (d), the SHAP value raises with the increase of A_1, showing the positive impact of the earthquake magnitude on the output. The scatter plots of SHAP values of A_1 for L_1 and L_2 are presented in Fig. 12(a) and (c), respectively. There is a positive trend for both target variables. This trend changes from slight to steep with the increase of the A_1, which agrees with the real situation that the seismic loss of catastrophic earthquakes is dramatically higher than small earthquakes. The earthquake information features take up the first four places for L_1 (see Fig. 11(a)) indicating the importance of earthquake hazards when predicting the social loss. A_2 (Latitude) is the second important feature for L_1. According to Fig. 12 (b), there is a trend that southern China suffers higher economic loss than the northern. As for L_2, the leading features except for A_1 (B_1, C_3, C_2) all belong to demographic and socioeconomic categories. As observed in Fig. 12(d), the impact of B_1 is significant when the value is small. SHAP value becomes steady when B_1 becomes large. Overall, the area with a higher population density shows a higher resilience against seismic economic loss.

5. Conclusions and future works

This research presents a novel AutoML framework to develop prediction models for rapid seismic loss estimation. Rather than determine the optimal model configurations by manual comparison, the AutoML framework identifies the optimal machine learning algorithm and hyperparameter setting automatically, reducing the repetitive work effectively. AutoML models to predict seismic loss are learned from the dataset integrating earthquake catalogs and social indicators. The missing data of the dataset constructed by the data collection stage is imputed by the data preprocessing module. Then, AutoML models for casualty rate and direct economic loss prediction are developed, respectively, where a two-step model is designed for casualty rate to eliminate the impacts of zero-casualty cases. The estimated loss can be obtained by inputting earthquake information and the corresponding social indicators. Bootstrap and SHAP techniques are introduced to quantify uncertainty and interpretation models. To validate the effectiveness of the proposed AutoML framework, seismic loss prediction models are developed based on the dataset of 2001 – 2018 earthquake events in China, and prediction results of 2019 earthquake cases are evaluated.

The results of the empirical study show that the AutoML framework is capable of developing the model with the highest predictive abilities compared with other AutoML models and traditional seismic loss estimation models. The proposed two-step model for casualty rate estimation results in higher performance compared with the single regression model. Models learned from the complete dataset achieve the ultimate performance compared with subsets that are composed of partial features, implying the effectiveness of the selected features. The model interpretation conducted by the SHAP analysis shows that the earthquake magnitude is the dominant factor for the seismic loss. Overall, the social loss is more sensitive to earthquake information features while the economic loss is more affected by social vulnerability features.

Although this research has contributed to some theoretical and practical implications, some limitations still exist in the framework. For one thing, the proposed models only estimate the total losses of the disaster area, rather than the distribution. If more spatial information is provided, such as the distribution of building inventory, the distribution of social and economic losses can be estimated. In future research, we will attempt to derive models for spatial loss prediction. The estimated spatial distribution of loss can help the government to dispatch the search and rescue team as well as assign resources more effectively. For another, this research adopted the social indicators of epicenters directly. However, the disaster areas may consist of multiple administrative areas. A weighted method may be required to aggregate the information of multiple areas in the future study. From the perspective of technique, the computation efficiency of the proposed AutoML can be improved by discarding inferior algorithms earlier during training. The suitable criterion to discard will be explored in future studies. More data preprocessing techniques can be involved in the framework to improve the performance, such as feature engineering, standardization. Furthermore, multi-objective optimization algorithms [76] will be explored to develop informed-based decisions and strategies to mitigated social and economic loses in seismic areas.

CRediT authorship contribution statement

Weiyi Chen: Writing – original draft, Methodology, Visualization, Data curation, Investigation, Validation, Formal analysis. **Limao Zhang:** Conceptualization, Supervision, Methodology, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The Start-Up Grant at Huazhong University of Science and Technology (No. 3004242122) is acknowledged for its financial support of this research.

References

[1] CRED, UNISDR. The human cost of disasters 2000-2019. 2020.
[2] Zhang S, Yang K, Cao Y. GIS-based rapid disaster loss assessment for earthquakes. IEEE Access 2019;7:6129–39.
[3] Jaiswal KS, Wald DJ, Earle PS, Porter KA, Hearne M. Earthquake casualty models within the USGS prompt assessment of global earthquakes for response (PAGER) system. In: Spence R, So E, Scawthorn C, editors. Human casualties in earthquakes: progress in modelling and mitigation. Dordrecht: Springer Netherlands; 2011. p. 83–94.
[4] Guettiche A, Guéguen P, Mimoune M. Economic and human loss empirical models for earthquakes in the Mediterranean region, with particular focus on Algeria. Int J Disaster Risk Sci 2017;8:415–34.
[5] Cui S, Yin Y, Wang D, Li Z, Wang Y. A stacking-based ensemble learning method for earthquake casualty prediction. Appl Soft Comput 2021;101:107038.
[6] Federal emergency management agency. HAZUS-MH 2.1 earthquake model technical manual2013.
[7] Ceferino L, Kiremidjian A, Deierlein G. Regional multiseverity casualty estimation due to building damage following a Mw 8.8 earthquake scenario in Lima, Peru. Earthquake Spectra 2019;34:1739–61.
[8] Wang Y, Gardoni P, Murphy C, Guerrier S. Predicting fatality rates due to earthquakes accounting for community vulnerability. Earthquake Spectra 2019;35:513–36.
[9] Guikema SD, Quiring SM. Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. Reliab Eng Syst Saf 2012;99:178–82.
[10] Noh H-Y, Kiremidjian A, Ceferino L, So E. Bayesian updating of earthquake vulnerability functions with application to mortality rates. Earthquake Spectra 2020;33:1173–89.
[11] Mahmood T, Xie M. Models and monitoring of zero-inflated processes: the past and current trends. Qual Reliab Eng Int 2019;35:2540–57.
[12] Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA: automatic model selection and hyperparameter optimization in WEKA. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated machine learning: methods, systems, challenges. Cham: Springer International Publishing; 2019. p. 81–95.
[13] Elshawi, R., Maher, M., Sakr, S. Automated machine learning: state-of-the-art and open challenges. arXiv preprint arXiv:190602287. 2019.

[14] Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn: efficient and robust automated machine learning. Automated machine learning. Cham: Springer; 2019. p. 113–34.
[15] Olson, R., Bartley, N., Urbanowicz, R., Moore, J. Evaluation of a tree-based pipeline optimization tool for automating data science2016.
[16] LeDell E, Poirier S. H2o AutoML: scalable automatic machine learning. In: Proceedings of the AutoML workshop at ICML; 2020.
[17] Wu Y, Hou G, Chen S. Post-earthquake resilience assessment and long-term restoration prioritization of transportation network. Reliab Eng Syst Saf 2021;211:107612.
[18] Silva V, Amo-Oduro D, Calderon A, Costa C, Dabbeek J, Despotaki V, et al. Development of a global seismic risk model. Earthquake Spectra 2020;36:372–94.
[19] Jaiswal K, Wald DJ. Estimating economic losses from earthquakes using an empirical approach. Earthquake Spectra 2013;29:309–24.
[20] Wu S, Jin J, Pan T. Empirical seismic vulnerability curve for mortality: case study of China. Nat Hazards 2015;77:645–62.
[21] Firuzi E, Amini Hosseini K, Ansari A, Izadkhah YO, Rashidabadi M, Hosseini M. An empirical model for fatality estimation of earthquakes in Iran. Nat Hazards 2020;103:231–50.
[22] Lian F, Zhang C, Pan H, Li M, Yang W, Meng Y, et al. Mapping environments and exposures of the world. In: Shi P, Kasperson R, editors. World atlas of natural disaster risk. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015. p. 3–21.
[23] Xing H, Zhonglin Z, Shaoyu W. The prediction model of earthquake casuailty based on robust wavelet v-SVM. Nat Hazards 2015;77:717–32.
[24] Federal Emergency Management Agency. Seismic performance assessment of buildings 2018.
[25] Ceferino L, Kiremidjian A, Deierlein G. Probabilistic model for regional multiseverity casualty estimation due to building damage following an earthquake. ASCE-ASME J Risk Uncertain Eng Syst Part A 2018;4:04018023.
[26] Yang, H.-C., Geng, L., Xue, Y., Hu, G. Spatial weibull regression with multivariate log gamma process and its applications to china earthquake economic loss. arXiv preprint arXiv:191203603. 2019.
[27] Huang X, Jin H. An earthquake casualty prediction model based on modified partial Gaussian curve. Nat Hazards 2018;94:999–1021.
[28] Frigerio I, Ventura S, Strigaro D, Mattavelli M, De Amicis M, Mugnano S, et al. A GIS-based approach to identify the spatial variability of social vulnerability to seismic hazard in Italy. Appl Geogr 2016;74:12–22.
[29] Cissé JD, Barrett CB. Estimating development resilience: a conditional moments-based approach. J Dev Econ 2018;135:272–84.
[30] Wang YV, Gardoni P, Murphy C, Guerrier S. Worldwide predictions of earthquake casualty rates with seismic intensity measure and socioeconomic data: a fragility-based formulation. Nat Hazards Rev 2020;21:04020001.
[31] Jaiswal K, Wald D. An empirical model for global earthquake fatality estimation. Earthquake Spectra 2010;26:1017–37.
[32] Zhang L, Pan Y. Information fusion for automated post-disaster building damage evaluation using deep neural network. Sustain Cities Soc 2022;77:103574.
[33] Xing H, Junyi S, Jin H. The casualty prediction of earthquake disaster based on extreme learning machine method. Nat Hazards 2020;102:873–86.
[34] Chen W, Zhang L. Building vulnerability assessment in seismic areas using ensemble learning: A Nepal case study. J Clean Prod 2022;350:131418.
[35] Cavallo E, Powell A, Becerra O. Estimating the direct economic damages of the earthquake in Haiti. Econ J 2010;120:F298–312.
[36] Sauti NS, Daud ME, Kaamin M, Sahat S. GIS spatial modelling for seismic risk assessment based on exposure, resilience, and capacity indicators to seismic hazard: a case study of Pahang, Malaysia. Geomat Nat Hazards Risk 2021;12:1948–72.
[37] Zhang W, Xu X, Chen X. Social vulnerability assessment of earthquake disaster based on the catastrophe progression method: a Sichuan Province case study. Int J Disaster Risk Reduct 2017;24:361–72.
[38] Shapira S, Novack L, Bar-Dayan Y, Aharonson-Daniel L. An integrated and interdisciplinary model for predicting the risk of injury and death in future earthquakes. PLoS ONE 2016;11:e0151111.
[39] Moudi M, Yan S, Bahramimianrood B, Li X, Yao L. Statistical model for earthquake economic loss estimation using GDP and DPI: a case study from Iran. Qual Quant 2018;53:583–98.
[40] Guo X, Kapucu N. Assessing social vulnerability to earthquake disaster using rough analytic hierarchy process method: a case study of Hanzhong City, China. Saf Sci 2020;125:104625.
[41] Gao Z, Ding M, Huang T, Hu X. Geohazard vulnerability assessment in Qiaojia seismic zones, SW China. Int J Disaster Risk Reduct 2021;52:101928.
[42] Shapira S, Aharonson-Daniel L, Shohet IM, Peek-Asa C, Bar-Dayan Y. Integrating epidemiological and engineering approaches in the assessment of human casualties in earthquakes. Nat Hazards 2015;78:1447–62.
[43] Tsamardinos I, Fanourgakis GS, Greasidou E, Klontzas E, Gkagkas K, Froudakis GE. An automated machine learning architecture for the accelerated prediction of metal-organic frameworks performance in energy and environmental applications. Microporous Mesoporous Mater 2020;300:110160.
[44] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[45] Chen, T., Guestrin, C. Xgboost: a scalable tree boosting system. 2016. p. 785–94.
[46] Dorogush, A.V., Ershov, V., Gulin, A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:181011363. 2018.
[47] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–54.
[48] Awad M, Khanna R. Support vector machines for classification. In: Awad M, Khanna R, editors. Efficient learning machines: theories, concepts, and applications for engineers and system designers. Berkeley, CA: Apress; 2015. p. 39–66.
[49] Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res 2018;267:687–99.
[50] Ahmad I, Basheri M, Iqbal MJ, Rahim A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access 2018;6:33789–95.
[51] Zeineddine H, Braendle U, Farah A. Enhancing prediction of student success: automated machine learning approach. Comput Electr Eng 2021;89:106903.
[52] Song Y, Liang J, Lu J, Zhao X. An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing 2017;251:26–34.
[53] Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random forest. J Mach Learn Res 2017;18:6673–90.
[54] Parsa AB, Movahedi A, Taghipour H, Derrible S, Mohammadian AK. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid Anal Prev 2020;136:105405.
[55] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:170609516. 2017.
[56] Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J Hydrol 2019;574:1029–41.
[57] Feurer M, Hutter F. Hyperparameter optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated machine learning: methods, systems, challenges. Cham: Springer International Publishing; 2019. p. 3–33.
[58] Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter optimization for machine learning models based on bayesian optimizationb. J Electron Sci Technol 2019;17:26–40.
[59] Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. PMLR 2013:115–23.
[60] Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2012;25:2951–9.
[61] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst 2011;24:1–9.
[62] Fan C, Liu C. A novel algorithm for circle curve fitting based on the least square method by the points of the Newton's rings. In: 2015 International conference on computers, communications, and systems (ICCCS); 2015. p. 256–60.
[63] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 2014;7:1247–50.
[64] Trichakis I, Nikolos I, Karatzas GP. Comparison of bootstrap confidence intervals for an ANN model of a karstic aquifer response. Hydrol Process 2011;25:2827–36.
[65] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141–64.
[66] Khosravi A, Nahavandi S, Creighton D, Atiya AF. Comprehensive review of neural network-based prediction intervals and new advances. IEEE Trans Neural Netw 2011;22:1341–56.
[67] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: 31st Conference on neural information processing systems; 2017. p. 4768–77.
[68] Guo K, Zhang L. Adaptive multi-objective optimization for emergency evacuation at metro stations. Reliab Eng Syst Saf 2022;219:108210.
[69] Zhang L, Lin P. Multi-objective optimization for limiting tunnel-induced damages considering uncertainties. Reliab Eng Syst Saf 2021;216:107945.
[70] Rong Y, Xu X, Cheng J, Chen G, Magistrale H, Shen Z-K. A probabilistic seismic hazard model for Mainland China. Earthquake Spectra 2020;36:181–209.
[71] Holzer TL, Savage JC. Global earthquake fatalities and population. Earthquake Spectra 2013;29:155–75.
[72] Li X, Li Z, Yang J, Liu Y, Fu B, Qi W, et al. Spatiotemporal characteristics of earthquake disaster losses in China from 1993 to 2016. Nat Hazards 2018;94:843–65.
[73] Wald DJ, Worden BC, Quitoriano V, Pankow KL. ShakeMap manual: technical manual, user's guide, and software guide. Tech Methods. Version 1.0 2005.
[74] Center for International Earth Science Information Network CCU. Gridded population of the world, version 4 (GPWv4): population density, revision 11. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC); 2018.
[75] National Bureau of Statistics of China. China statistical yearbook. 2021. http://www.stats.gov.cn/tjsj/ndsj/.
[76] Guo K, Zhang L. Multi-objective optimization for improved project management: Current status and future directions. Autom Constr 2022;139:104256.