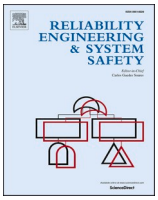
Reliability Engineering and System Safety 225 (2022) 108587

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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



Interpretable boosting tree ensemble method for multisource building fire loss prediction☆

Ning Wang, Yan Xu, Sutong Wang *

School of Economics and Management, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

Keywords:
Boosting tree ensemble
Building fire loss
Multisource
Shapley additive explanation

ABSTRACT

Building fires may cause enormous property loss. Disaster relief organizations and post-disaster recovery efforts benefit from the accurate and interpretable prediction of property loss. To solve this problem, we propose a novel interpretable boosting tree ensemble method (IBTEM), which is of practical significance for providing decision support for dispatching aid and mobilizing recovery resources. First, we fuse multisource datasets including National Fire Incident Reporting System (NFIRS) dataset and National Oceanic and Atmospheric Administration (NOAA) dataset from 2012 to 2016. Second, we construct four variable scenario subsets to select related variables for building fire loss. Third, we adopt Winsorization, logarithmic transformation, recursive feature elimination and weighted voting strategies to create an ensemble of boosting trees. Fourth, we conduct interpretable Shapley additive explanations to analyze the model internal mechanism. The proposed IBTEM is compared with other popular machine learning methods and the experimental results show the IBTEM achieves outstanding superiority. Property value, fire spread and number of stories with minor damage are verified the most effective variables for loss prediction. In conclusion, the IBTEM realizes accurate and interpretable loss prediction of building fires, and assists relevant departments in making disaster relief decisions in a timelier manner.

1. Introduction

Building fire outbreaks lead to enormous property loss and high potential social impacts [1]. The economic loss due to fire reaches 1% of the Gross Domestic Product in developed countries [2,3]. According to the data offered by the National Fire Protection Association (NFPA), in total, 1342,000 fires occurred in the U.S. in 2016, causing an estimated $10.6 billion in direct property loss [4]. Of these fires, building fires remain a critical concern as 475,500 fires (35.4% of all fires) resulted in an estimated $7.9 billion in property loss, accounting for 74.5% of the total loss. Direct property loss is one of the main signs used to measure the severity of fires. Disaster relief organizations and post-disaster recovery efforts benefit from the accurate prediction of property loss. If property loss due to fires is accurately predicted in advance, relevant departments are able to make timelier fire relief decisions, plan investments better and improve the recovery efficiency. However, currently, most studies measuring disaster property loss focus on traditional mathematical statistics for retrospective data analyses, and prospective prediction research is lacking.

Disaster property loss prediction is an intuitive and quantitative estimation of the potential loss caused by a disaster that will occur or has already occurred to provide decision support for the formulation of disaster countermeasures. Disaster property loss prediction is the basis of disaster relief and post-disaster recovery, which benefits from the accurate prediction of property loss. In general, natural, social and economic conditions vary by time and region. When disasters occur, losses vary considerably. Therefore, it is difficult to accurately quantify property loss.

The substantial property loss caused by building fires shows the importance of predicting property loss in building fires. Extensive research has been performed in the field of property loss prediction, but most studies focus on earthquakes, floods, typhoons and other disasters, while few studies investigated fire losses. Yu and Gardoni [5] proposed a probabilistic model to predict road blockage due to building damage following earthquakes, considering considers the relevant factors affecting the road blockage probability. Hsu et al. [6] used the Monte Carlo simulation to predict property loss due to flood events and applied this simulation to problems with the property insurance industry.

☆ This work was supported in part by the National Natural Science Foundation of China under Grant 71774021 and Grant 71874020.
* Corresponding author.
E-mail address: wangsutong@mail.dlut.edu.cn (S. Wang).

https://doi.org/10.1016/j.ress.2022.108587
Received 30 August 2021; Received in revised form 21 March 2022; Accepted 9 May 2022
Available online 11 May 2022

Karhula et al. [7] established a Monte Carlo simulation platform to predict life and property loss in housing fires to evaluate the effectiveness of potential measures to improve housing fire safety. Jagger et al. [8] provided a framework for predicting seasonal and annual property loss due to hurricanes using a Markov chain Monte Carlo (MCMC) simulation. Khan et al. [9] used the Poisson regression method to estimate property loss in highway corridors and discussed the relationship between property loss due to traffic accidents and traffic volume, road length and vehicle mileage. Cao et al. [10] used a self-study discrete regression (SSDR) method to identify the relationship between the intensity of typhoons and the vulnerability of enterprise properties and established a loss rate assessment model of typhoon risk for enterprise property, which provided a basis for typhoon risk avoidance of enterprise property. Westerling and Bryant [11] used a logistic regression to predict property loss due to wildfires by quantifying changes in wildfire risks and property loss under different climate scenarios. Fronstin et al. [12] utilized a two-limit Tobit method to effectively predict property loss due to Hurricane Andrew and then explained the influence of wind speed, housing quality and changes in building codes on the property loss. Hanea et al. [13] presented a probabilistic model based on Bayesian networks for analysis of the Schiphol Cell Complex fire, which could be used prior to the accident. Heatwole and Rose [14] proposed a log-log regression method to conduct a rapid economic consequence estimation of property loss due to U.S. earthquakes, which provided decision support for ex ante risk assessment and ex post disaster loss assessment, resource allocation and mobilization. The above research focused on statistical methods to predict property loss, indicating that statistical methods based on Monte Carlo, linear regression and nonlinear regression are able to predict disaster property loss, but the prediction accuracy is not sufficient. Thus, more scholars introduced the machine learning methods for further improvement in terms of prediction accuracy. Worrell et al. [15] explored the application of machine learning to generate metamodel approximations of a physics based fire hazard model. A k-nearest neighbor metamodel was selected and proved an accurate surrogate to Consolidated Fire and Smoke Transport (CFAST) compared to the 24 other metamodel types. Bhardwaj et al. [16] proposed a Bayesian Network probabilistic framework for risk assessment and probabilistic modeling of fire and explosion accidents in Floating production storage and offloading (FPSO) units. Diaz and Joseph [17] used artificial neural networks to predict property loss due to tornadoes. The results showed that neural networks achieved better prediction accuracy than the statistical method of a zero-inflated log-normal regression. However, neural networks could easily face local optimum and overfitting due to its subjective factors.

In early research, ensemble methods have been proven to possess better accuracy and stronger generalization performance than a single learning method [18–20]. Therefore, scholars were no longer limited to the use of a single learner and transferred to ensemble methods, which were comprised of multiple learners and could constantly improve the predictive performance and adaptability [21,22]. Tyralis et al. [23] proposed a stacking of quantile regression and quantile regression forests through minimizing the interval score of the quantile predictions provided by the ensemble learner. They used machine learning methods to post-process hydrological model simulations to quantify the uncertainty of hydrological predictions. The proposed method was tested on a large dataset and found significant improvement compared to base learner. Mastelini et al. [24] presented a Deep Structure for Tracking Asynchronous Regressor Stacking (DSTARS) by combining multiple stacked repressors into a deep structure, which achieved significative smaller Relative Root Mean Squared Error (RRMSE), than the comparative methods. Ribeiro et al. [25] took use of regression ensembles to develop an effective model for accurate agricultural commodities prices forecasting. The test results verified that ensemble approach presented statistically significant gains, reducing prediction errors for the price series studied.

The application of boosting tree ensemble method to predict property loss due to fire incidents appeared promising although nascent in the current literature. Mohsen et al. [26] reviewed the methods based on geospatial information system (GIS) for modeling forest fires. The comparison results showed the ensemble methods performed more accurate and the data-driven approaches were the most frequent. Agarwal et al. [27] implemented the gradient boosting decision tree (GBT) algorithm to predict the losses due to fire incidents. The results signified the high predictive accuracy obtained by the GBT model. According to Guelman [28], compared with other popular machine learning methods (e.g., neural networks, Support Vector Machines (SVM)), gradient boosting provided interpretable results with little data preprocessing and parameter adjustment. The method had strong robustness when applied to various classification or regression problems. Boosting tree ensemble method could reduce deviation by integrating several weak learners. Then, a good generalized model with small deviation and variance was obtained. Chan and Paelinck [29] evaluated the prediction performance of the AdaBoost, Random Forest (RF), C5.0 tree and Multi-Layer Perceptron (MLP) methods. After 99 trials, Adaboost obtained the best precision, followed by random forest. Chen et al. [30] provided the results of prediction using the extreme gradient boosting (XGBoost) algorithm and compared it with four other machine learning algorithms, including RF, SVM, GBDT. The results demonstrated that the XGBoost algorithm could achieve the highest accuracy in the case of different data sets and the GBDT algorithm ranked the second. Zhang et al. [31] compared the prediction performance and the computational efficiency of light gradient boosting machine (LightGBM) with deep neural networks (DNN), RF, SVM, and XGBoost. The results showed that LightGBM was an effective and highly scalable algorithm achieving the best predictive performance while consuming significantly shorter computational time. Fan et al. [32] evaluated the performance of SVM and four tree-based algorithms, i.e., M5 model tree (M5Tree), RF, XGBoost and gradient boosting with categorical variables support (CatBoost). Comprehensively considering prediction accuracy, computational time and generalization capability, CatBoost was highly recommended to develop general models. The previous research showed that boosting tree ensemble method was competitive among both single learning methods and ensemble learning methods, and was effectively applied to loan default prediction [33], wind speed prediction [34], disease risk prediction [35], Portfolio decision [36], anomaly detection and diagnosis for wind turbines [37] and other fields. Therefore, we make use of the advantages of the boosting tree ensemble method, and establish a building fire loss prediction model based on ensemble learning method. Samson and Khalid [38] consisted that new methods such as the development of a form of machine-assisted interpretation is able to bridge the gap between data sources and the theoretical and practical mechanisms.

The contributions of this study in terms of originality, significance, and performance metrics are listed as follows:

(1) A novel interpretable boosting tree ensemble method (IBTEM) for building fire loss prediction that uses weighted voting strategy to create an ensemble of boosting trees (CatBoost, XGBoost, and LightGBM) based on coefficient of determination (R^2) and interpretable Shapley additive explanations is proposed. Winsorization, logarithmic transformation, recursive feature elimination (RFE), correlation analysis and the sample quantile of order p strategy is applied to improve the model performance.
(2) Four variable scenario subsets (disaster-inducing factors, disaster-affected bodies, disaster-formative environments, emergency activities) for building fire loss prediction based on the regional disaster system theory is constructed and interpretable analysis including the variable importance ranking and partial dependence plot based on Shapley additive explanation is provided.
(3) The model is built based on 10-fold cross-validation of multi-source data on American building fire incidents and their

corresponding weather from 2012 to 2016 and achieves competitive performance in terms of root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), median absolute error (MedAE), coefficient of determination(R^2) for building fire loss prediction. Non-parametric Friedman test and Holm post hoc tests are performed for validating the effectiveness of the proposed method.

Our study provides a systematic approach to quickly predict loss in advance and, thus, assists relevant departments in making timelier decisions regarding dispatching aid and mobilizing resources for recovery and reconstruction, which is of practical significance for rationally planning investments and improving recovery efficiency. Further, it can also provide a reference for constructing related prediction models in other fields.

The remainder of this paper is organized as follows. In Section 2, the data preparation of property loss prediction in building fires is described. Section 3 illustrates the framework and details of the proposed method, interpretable tree ensemble-based method for multisource building fire loss prediction. The comparative experimental results and the improvement analysis of the proposed method are detailed in Section 4. Section 5 presents the discussion of interpretable analysis of the proposed method and conclusion of this study.

2. Data preparation of property loss prediction in building fires

We collect American building fires data of National Fire Incident Reporting System (NFIRS) and their corresponding weather data of Global Historical Climatology Network Daily database (GHCND) from 2012 to 2016, respectively. NFIRS is developed by the U.S. Fire Administration (USFA) that includes information regarding when and where a fire occurred, cause of ignition, heat source, property loss, property value, etc. The original fire dataset contains 115 variables and more than 2 million records within a year. The weather data of GHCND is collected by National Oceanic and Atmospheric Administration (NOAA) which is the world's largest provider of weather and climate data. The original weather dataset covers over weather observations (e.g., temperature, wind, pressure, snowfall) from more than 110,000 weather stations.

2.1. Data cleaning

Data preprocessing costs around 80% work of data mining process, which has a great impact on experimental results [39]. With regard to the fire data of NFIRS, the downloaded dataset is composed of basic incident and fire incident from two files by year. We join the two files together indexed by "State", "Fire Department ID", "Incident Date", "Incident Number" and "Exposure Number", which can locate the fire incident uniquely. We extract "YEAR", "MONTH", "HOUR" from "Alarm Date and Time" and calculate response time (RESP), number of apparatus (APP), number of personnel (PER), total loss (LOSS) and original value (VAL) through the following equations:

$$RESP = ARRIVAL_{TIME} - ALARM_{TIME} \tag{1}$$

$$APP = SUPPRESSION_{APP} + EMS_{APP} + OTHER_{APP} \tag{2}$$

$$PER = SUPPRESSION_{PER} + EMS_{PER} + OTHER_{PER} \tag{3}$$

$$LOSS = PROPERTY_{LOSS} + CONTENTS_{LOSS} \tag{4}$$

$$VAL = PROPERTY_{VAL} + CONTENTS_{VAL} \tag{5}$$

Property loss is greatly affected by its own value (the target variable's own value determines the upper limit of its property loss). Items of high property value, even if only slightly damaged, will lose more than items of low property value that are completely destroyed, which cannot be generalized. Thus, the property loss rate provides an accurate measure of fire loss. In terms of the building fire loss prediction, it is considered the dependent variable to measure loss, i.e., the ratio between the total property loss and the original property value. In terms of the original weather data downloaded from the file "ghcnd-daily" in GHCND, they record the weather condition every day around the world. The samples of original GHCND data are shown in Table 1. We first filter the variable station identifier with "US" to limit the scope of area. Then we transpose the original dataset and select Average temperature (TAVG), Maximum temperature (TMAX), Minimum temperature (TMIN), Precipitation (PRCP), Average wind (AWND), Snowfall (SNOW), Snow depth (SNWD), Wind direction (WDF5), Wind speed (WSF5) as weather variables. The variable "Measurement Date" of original data belongs to string format and is transferred into "%month% day%year" date format to match that of NFIRS. According to the file "ghcnd-stations", we match the weather station to the state at which it is located indexed by "Station Identifier". The weather data are grouped by "Measurement Date", "State" and "Measurement Type" and concatenated together from 2012 to 2016.

Table 1
Samples of original GHCND data.

Station Identifier	Measurement Date	Measurement Type	Measurement Flag
US009052008	20,120,101	TMAX	−10
US009052008	20,120,101	TMIN	−67
US10adam002	20,120,101	PRCP	0
US10adam002	20,120,101	SNOW	0

We fuse the multisource data from NFIRS and GHCND based on "Date", "State" variables and drop some unrelated variables such as "Version", "Exposure Number", "Incident Number", etc. We remove the variables with missing values more than 70%.

To ensure that the analytical results of the fire loss prediction are more meaningful, it is important to select incidents that can represent building fires to the greatest extent possible. When filtering the incident type of structure fires, we exclude confined fires with no flame damage to structure or its contents, such as chimney fires and rubbish fires. When filtering the structure type of buildings involved in fires, structures, such as bridges, telephone poles, fences, tents, etc., are excluded because they are unrepresentative as the economic value is too low or the duration of the fire is often too short.

Before analyzing data and modeling, it is necessary to clean and preprocess data to improve data quality. We perform an initial processing of the obtained dataset through the following 4 steps.

(1) Eliminate duplicate data.
(2) Remove outliers whose loss rate is less than 0 or greater than 1, and turn infinity into 0.
(3) Delete missing values and records encoded as "Undetermined" for categorical variables, while filling in the missing values with the mean value for numerical variables.
(4) Remove outliers whose "RESP" exceeds 24 h.

Thus, 8124 fire incident samples are finally obtained.

2.2. Scenario subsets construction

The variables that affect the accuracy of building fire loss prediction are various, therefore, it is crucial to select the relevant variables. The independent variables affecting the loss are selected from relevant articles [40–43] while considering other metrics in NFIRS and GHCND. According to the Regional Disaster System Theory [44], disasters can be regarded as the process of interaction between disaster-inducing factors and disaster-affected bodies in specific disaster-formative environments. Meanwhile, emergency activities also have an important impact on the system. Specifically, disaster-inducing factors are the factors that lead to

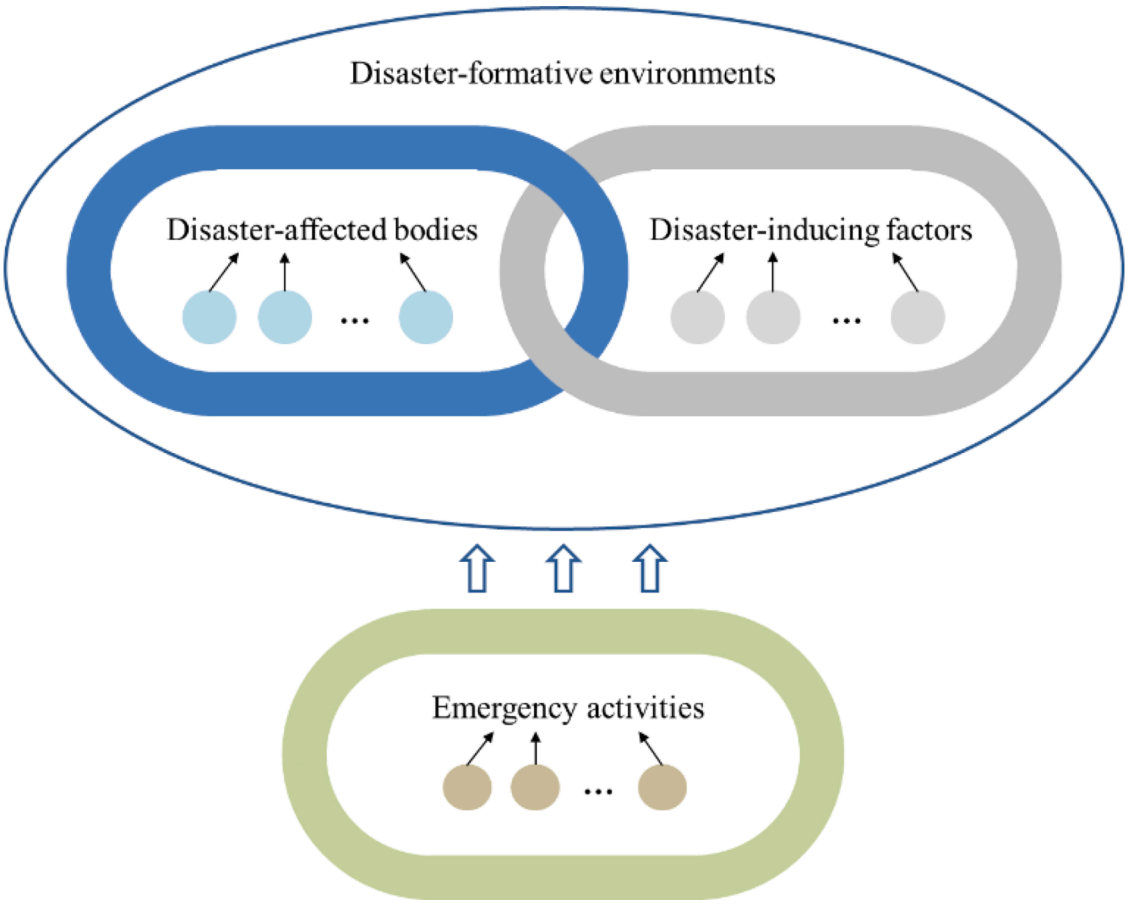


Fig. 1. Structure of regional disaster system.

the occurrence of disasters or promote the development of disasters, which can have adverse effects on human life, property, or various activities. Disaster-affected bodies are the objective entities directly affected by the disaster. Disaster-formative environments are the external environments of the formation of disaster-inducing factors and the whole process of emergencies. Emergency activities refer to the activities taken by relevant departments to prevent, respond to and recover from emergencies to reduce disaster loss. Here, we group the variables into four scenario subsets, i.e., disaster-affected bodies scenario, disaster-inducing factors scenario, disaster-formative environments scenario and emergency activities scenario. The first step is to construct the disaster-affected bodies scenario of target buildings. The required information related to the building includes the occupancy of the building at all hours, the number of residential units and the number of buildings involved in the fire, etc. The second scenario considers the effect of disaster-inducing factors, which may affect the conditions of fire spread. Cause of ignition, heat source, burned area, presence of fire detectors and so on have been acknowledged. The third disaster-formative environments scenario takes in weather observations such as temperatures, wind speed, snowfall. In the final emergency activities scenario, the major factors influencing fire department intervention include the response time, number of firefighters and number of suppression apparatus, etc. Since decreasing the response time to building fires leads to less property loss, we select response time rather than rescue time, which is highly correlated to personnel and apparatus. Thus, the variables affecting the property loss rate in building fires can be summarized as Fig. 1, and detailed descriptions on each variable are explained in Table A1. In the subsequent experiments, the 53 filtered variables are taken as input samples, and the property loss rate is the output.

2.3. Data transformation

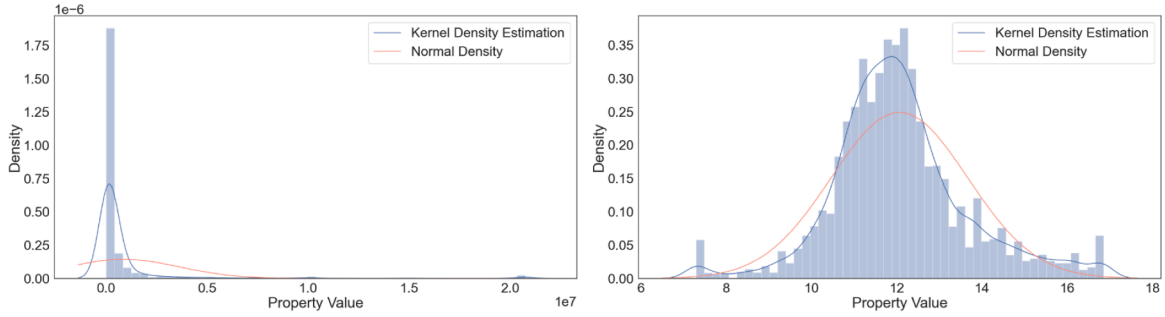
The statistics of the selected data samples are presented in Table 2. We carry out label encoding for categorical variables so that the discontinuous numbers or text are numbered.

As shown in Table 2, the distribution of many numeric variables can be heavily influenced by outliers. When the sample data is large enough, continuous variables are usually winsorized in order to eliminate the influence of some extreme values on the research. Winsorizing is a statistical transformation that reduces the impact of possible spurious outliers by limiting extreme values in the statistics. A typical strategy is to set all outliers to a specified percentile of the data, that is, to replace data that exceeds a specified percentile with adjacent values reserved for that specified percentile. For example, a 90% Winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile. Winsorized estimators are usually more robust to outliers than their standard forms [45]. Here we set the bottom 1% to the 1st percentile and the top 1% to the 99th percentile.

The skewed distribution of data has many negative effects: (1) the asymmetry of frequency distribution causes the concentration to be tilted to one side; (2) the increase of outliers has great influence on the mean value, which leads to the deviation of the mean value to the direction of maximum or minimum value; (3) many models assume that the data is normally distributed, which affects the predicted results. Therefore, we perform logarithmic transformation on numeric variables including loss rate. Logarithmic transformation [46] processes the original data by formula (1) to change the skewness of them so as to make the data more consistent with normal distribution. Thus, the validity of the data is guaranteed. Taking property value as an example, it can be seen from Fig. 2 that the probability density function after

Table 2
Statistics of the data samples.

Category	Count	Unique	Top	Frequency	Category	Count	Unique	Top	Frequency
ADD_WILD	8124	1	N	8124	FIRST_IGN	8124	74	76	1191
AID	8124	6	N	6153	TYPE_MAT	8124	50	41	1351
SHIFT	8124	36	C	1870	CAUSE_IGN	8124	6	2	5552
ALARMS	8124	22	1	4095	FACT_IGN_1	8124	49	NN	2557
APP_MOD	8124	2	Y	5195	HUM_FAC_1	8124	2	N	7599
RESOU_AID	8124	2	N	7718	STRUC_STAT	8124	8	2	7743
DET_ALERT	8124	2	2	4418	FLAME_SPRD	8124	2	Y	7007
HAZ_REL	8124	10	N	8040	DETECTOR	8124	2	1	6401
MIXED_USE	8124	12	NN	6254	AES_PRES	8124	3	N	7203
PROP_USE	8124	105	419	5114	state	8124	49	FL	1285
NOT_RES	8124	2	N	6393	FIRE_SPRD	8124	5	2	3501
LESS_1ACRE	8124	2	N	7662	YEAR	8124	5	2012	2065
AREA_ORIG	8124	74	24	2212	MONTH	8124	12	1	862
HEAT_SOURC	8124	33	12	1645	HOUR	8124	24	18	516
Numeric	Count	Mean	Std	Min	25%	50%	75%	Max	–
PROP_VAL	8124	1,632,521	19,521,232	0	70,000	150,000	350,000	999,999,999	–
CONT_VAL	8124	413,552.6	4,664,524	0	6000	20,000	75,000	250,000,000	–
NUM_UNIT	8124	5.468378	28.82109	0	1	1	1	1200	–
BLDG_INVOL	8124	0.944599	0.331642	0	1	1	1	10	–
ACRES_BURN	8124	2.437853	3.338749	0	0	0	7.011373	7.011373	–
BLDG_ABOVE	8124	1.807607	2.820981	0	1	1	2	131	–
BLDG_BELOW	8124	0.289513	0.595344	0	0	0	1	20	–
TOT_SQ_FT	8124	13,775.77	244,090.2	0	1000	1600	3236.5	15,000,000	–
FIRE_ORIG	8124	1.214693	1.287998	−1	1	1	1	75	–
ST_DAM_MIN	8124	0.61556	0.548067	0	0	1	1	13	–
ST_DAM_SIG	8124	0.149765	0.381906	0	0	0	0	3	–
ST_DAM_HVY	8124	0.067242	0.272498	−1	0	0	0	3	–
ST_DAM_XTR	8124	0.051054	0.267796	0	0	0	0	4	–
APP	8124	8.597858	27.91296	0	5	8	10	2220	–
PER	8124	16.54985	10.80137	0	9	16	23	200	–
AWND	8124	33.17927	14.02386	3.181818	23.125	31.05409	40.66667	158.6296	–
PRCP	8124	31.82477	61.39682	0	0.628535	6.911084	37.09932	890.6138	–
SNOW	8124	1.317576	7.464792	0	0	0	0	245.2738	–
SNWD	8124	16.71648	71.77126	0	0	0	0	1146.651	–
TAVG	8124	146.9754	99.23026	−238.831	76.90741	164.5085	228.0139	325.4167	–
TMAX	8124	206.0346	103.56	−183.211	135.2169	228.5205	288.4902	401.875	–
TMIN	8124	87.31945	100.4165	−296.654	11.75758	98.41469	171.6671	258.35	–
WDF5	8124	192.0671	65.14961	24.70588	145.8824	192.1714	240.5882	346	–
WSF5	8124	107.2792	28.53793	42.32	87.41176	103.2016	123.1364	234.4444	–
RESP	8124	5.675283	5.476679	0	4	5	7	283	–



(a) Probability density distribution before logarithmic transformation

(b) Probability density distribution after logarithmic transformation

Fig. 2. Comparison of probability density distribution.

logarithmic transformation is closer to normal distribution. Finally, we restore the predicted smooth data with formula (7).

$$y = log(x+1) \tag{6}$$

$$x = e^{y} - 1 \tag{7}$$

The selected variables have different dimensions, which affect the data processing results. To eliminate a dimensional influence between the indexes and avoid interference of such a difference in the model training, variable normalization processing needs to be carried out to solve the comparability problem among the performance metrics. Common normalization methods include the min-max method and the z-score method. Here, we use the z-score method to scale the raw data such that the value is mapped to a specific range, and the conversion formula [47] is as follows:

$$x^{*} = \frac{x-\mu}{\sigma} \tag{8}$$

where μ is the mean of the sample, σ is the standard deviation of the sample, and the original value x of the variable is normalized to x^{*}. Based on the above formula, the processed data have a mean of 0 and a standard deviation of 1.

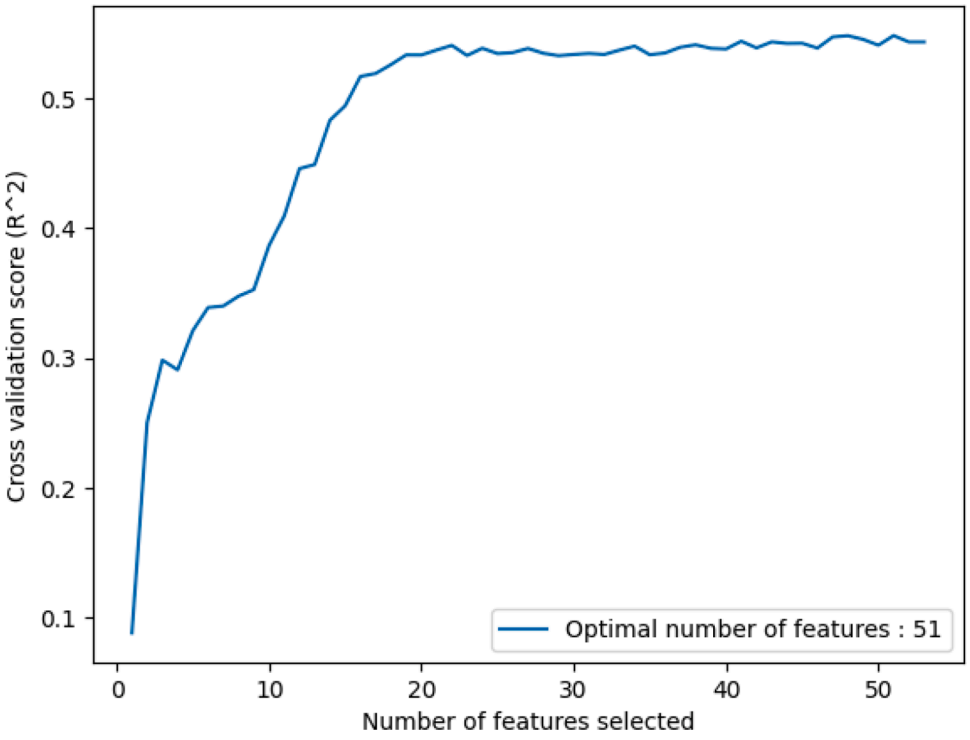


Fig. 3. Variable selection based on RFE.

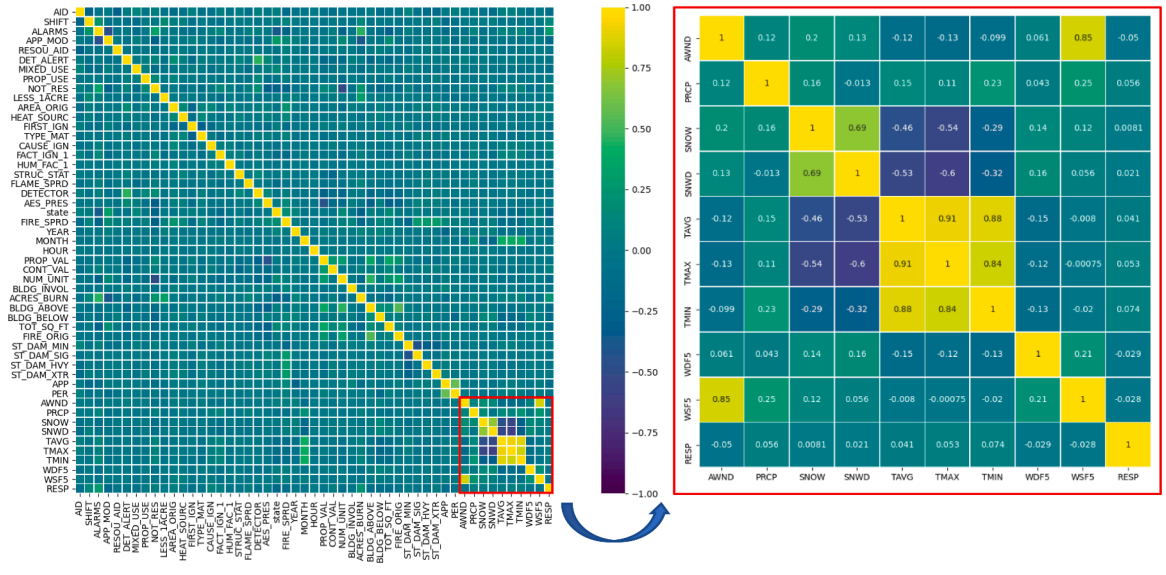


Fig. 4. Correlation heatmap of the filtered variables.

2.4. Variable selection and correlation analysis

2.4.1. Variable selection based on RFE

After data preprocessing, we need to select meaningful variables and input them into machine learning algorithms for training. But it is unknown which variable is valid for a particular learning algorithm. Therefore, it is necessary to select the relevant variables that are beneficial to the learning algorithm from all the variables. Variable selection methods provides us a way of improving prediction performance, reducing model complexity, speeding up computation, and a better understanding of data characteristics or underlying structure [48].

There are three main variable selection ways: Filter, Wrapper and Embedded methods. Recursive feature elimination (RFE) belongs to Wrapper method and is a representative of variable selection. It builds the model recursively and select the best (or worst) variables according to the coefficient until all the variables have been traversed. The order in which variables are eliminated during this process is the sorting of variables. Therefore, it is a greedy algorithm to find the optimal variable subset.

The stability of RFE depends on the base estimator used for iteration to a great extent. Here we select Catboost algorithm as the base estimator and take R^2 as cross validation score. RFE automatically adjusts the number of selected variables by cross validation. From Fig. 3, the optimal number of variables turns out to be 51, while the eliminated variables are ADD_WILD and HAZ_REL.

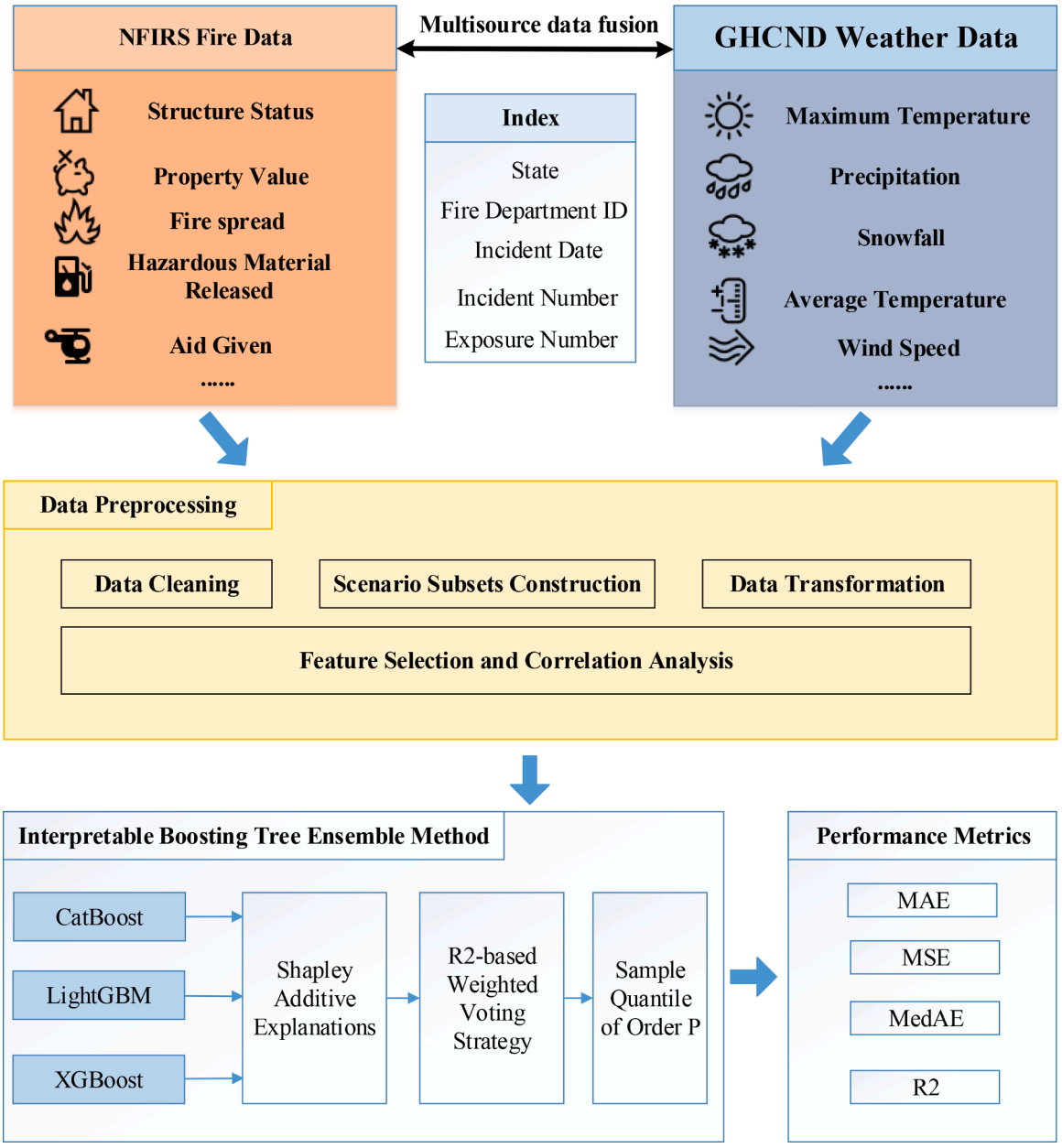


Fig. 5. The framework of the proposed interpretable boosting tree ensemble method.

2.4.2. Correlation analysis based on Pearson

Fig. 4 reports the correlation analysis results of filtered variables in the obtained 8124 samples. For clarity, we have magnified the figure part containing the variables with strong correlations on the right. By calculating the Pearson correlation coefficients of the variables, we can determine whether the selected variables are reasonable. If the correlation coefficient of the variables is greater than 0.7, multiple co-linear problems may exist, indicating that the variable selection is not reasonable. Otherwise, the variable selection is reasonable [49].

The heatmap is a symmetric structure. The correlation coefficient of the same variable on the diagonal line is 1. The closer the color of the subgraph is to yellow, the stronger the correlation is, which tends to be positive. The closer the color of the subgraph is to purple, the stronger the correlation is, which tends to be negative. The color at 0 represents an irrelevant correlation. The closer the color of the subgraph is to that at 0, the weaker the correlation is. As shown in Fig. 4, the absolute values of correlation coefficients for most variables are less than 0.7. Exceptionally, the correlation coefficients of TMIN, TMAX and TAVG are greater than 0.7, as well as WSF5 and AWND. The high correlation among TMIN, TMAX and TAVG is easy to understand because there exist strong linear relations among the minimum, maximum and average of temperature. When dealing with this pair of variables with high correlation, we consider that the average temperature TAVG and the average wind speed AWND are more representative. So, we eliminate the three redundant variables, TMIN, TMAX and WSF5. Thus, the number of final variables in use are reduced to 48.

3. Interpretable boosting tree ensemble method

3.1. Framework

The entire process of the proposed interpretable boosting tree ensemble method (IBTEM) is depicted in Fig. 5.

As shown in Fig. 5, the framework mainly consists of three parts. The first part refers to multi-source data fusion. The data from NFIRS and GHCND are fused indexed by “State”, “Fire Department ID”, “Incident

Date", "Incident Number" and "Exposure Number". The second part is data preprocessing, which incorporates data cleaning, scenario subsets construction, data transformation, variable selection and correlation analysis. The raw data are then converted to the compatible format for model construction. The third part corresponds to interpretable boosting tree ensemble model trained and tested with performance metrics MAE, MSE, MAPE and R^2. We aim to build an interpretable boosting tree ensemble method to improve the accuracy of property loss prediction due to building fires. Tree ensemble methods have advantages in small deviation, variance and good generalization through integrating several base learners. Therefore, we design a heterogeneous ensemble learner, which is a combination of Catboost, XGBoost and LightGBM by R^2-based weighted voting strategy. Sample quantile of order P is used to improve the model performance. Shapley additive explanations is an effective way to visualize the internal mechanism in machine learning models and improve the model interpretability effectively.

3.2. Boosting tree ensemble algorithm

Ensemble learning methods usually achieves higher accuracy and generalization performance compared with single learners. In recent years, tree ensemble models have been widely used in machine learning. The boosting algorithm [50] is a representative of the ensemble learning method, which has various superiority in fitting ability and generalization performance. Gradient boosting decision tree (GBDT), proposed in [51], is a classic algorithm of boosting algorithm, which iteratively combines various weak learners into a strong learner through training the decision trees. It adopts negative gradient of loss function to continuously adjust the sample weight and gradually reduce the deviation. Besides, the algorithm forces the subsequent learner to pay more attention to the samples misclassified by the previous learner to determine the weight of each sample according to the precision of the previous overall classification. Then, the new dataset with modified weights is transmitted to the next learner for training. Finally, the learner of each training is integrated for use as the final decision learner.

Gradient boosting with categorical features support (Catboost) [52], eXtreme gradient boosting (XGBoost) [53], Light Gradient Boosting Machine (LightGBM) [54] are improved versions of GBDT, which have their own advantages in prediction tasks.

Catboost introduces the categorical variable processing algorithm and ordered boosting to solve the problems of gradient bias and prediction shift, so as to reduce the occurrence of overfitting and improve the accuracy and generalization ability of the algorithm. It adopts ordered boosting instead of the traditional gradient estimation to obtain the unbiased estimation of the statistical value of the target variable and gradient value. CatBoost usually shows superiority in dataset with categorical variables.

XGBoost puts more focus on prediction accuracy and computing efficiency. GBDT only uses the first-order derivative to approximate the cost function, while XGBoost performs a second-order Taylor expansion on the cost function. At the same time, XGBoost adds regularization to penalize the complex models, such as the number of leaf nodes and the depth of the tree in the cost function. All instance values of each variable are boxed with histograms to discretize the data. The approximate algorithm, parallelization support and cache access optimization speed up the construction of the tree, thus improving the efficiency.

LightGBM is constructed with Gradient-based One-Side Sampling (GOSS) algorithm and Exclusive Feature Bundling (EFB) based on GBDT algorithm, which shows higher training efficiency and lower memory utilization. It is an efficient and scalable way to reduce the amount of data and high-dimensional features when they are complex and cost lots of resources. Goss retained the samples with large gradients and randomly sampled the instances with small gradients to reduce the computation. Thus, the purpose of improving efficiency is achieved on the premise of ensuring accuracy. EFB is a way to reduce feature dimensions by bundling mutually exclusive features, thus improving computing efficiency without missing information.

In this paper, we denote the Catboost, XGBoost, LightGBM models as h_{cat}, h_{xgb}, h_{lig}, respectively. We construct a property loss model of building fires based on Catboost, XGBoost and LightGBM as base learners. Here, the R^2-based weighted voting strategy and sample quantile of order P are used in the iterative process of the tree ensemble training. The entire method process of boosting tree ensemble for building fire loss prediction is expressed as Algorithm 1.

As Algorithm 1, three models are established by the Catboost, XGBoost, LightGBM algorithms to predict property loss due to building fires. We use the Catboost, XGBoost, LightGBM models as base learners to predict the loss rate. The final output of the IBTEM model is the weighted sum of the predicted results learnt by the boosting tree ensemble models above. The coefficient of determination (R^2) of the predicted results are used as weight for model integration. The weights ω_{model} of the output of Catboost, XGBoost, LightGBM are adjusted by the mean values of R^2 validated on validation set V. The final value of the predicted loss rate is calculated by assigning weights to the models to maximize R^2. The weights of integrated models satisfy the following formula:

$$\sum_{i=1}^{M} \omega_{\text{model}} = 1 \tag{9}$$

3.3. Interpretable Shapley additive explanations

Interpretability is the degree to which humans can consistently predict model results. The higher the interpretability of the machine learning model, the easier it is for people to understand the reasons for making certain decisions or predictions.

The traditional variable importance evaluation method can only calculate which variable is more important, but cannot know the influence of each variable on the prediction result. The Shapley Additive Explanations (SHAP) method [55] is inspired by the Shapley value of cooperative game theory, and constructs an additive explanation model based on the Shapley value. The Shapley value can measure the marginal contribution of each participators in the entire cooperation. We compare the variables used for fire loss prediction to participants in SHAP methods. When a new variable is added to the model, the marginal contribution of the variable can be calculated through SHAP method, and then the different marginal contributions of the variable are considered under different variable orders.

Assuming that x_{ij} represents the jth variable of the ith sample of a set of sample data, the model predicted value of this sample is y_i, and the average value of the model predicted value of the entire data is y_{base}. The SHAP value of $SHAP_i$ is expressed as follows:

$$SHAP_i = f(x)_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ij}) = f(x)_{base} + \sum_{j=1}^{m} f(x_{ij}) \tag{10}$$

where $f(x_{ij})$ represents the SHAP value of x_{ij}, m represents the number of variables, and $f(x)_{base}$ represents the average value of all $f(x_{ij})$. When $f(x_{ij}) > 0$, the jth variable of the ith sample has a positive effect on the prediction result y_i, and vice versa. It truly reflects the positive and negative effects of the variable on the prediction results, the building fire loss.

A partial dependence plots (PDP) is a scatter plot that reflects the effect of one variable to the final model output. Suppose that the variable in PDP is not related to other variables, PDP can perfectly show how this variable affects the predicted value on average. The partial dependency function used for regression is defined as follows:

$$\widehat{f_{x_S}}(x_S) = E_{x_C}[\widehat{f}(x_S, x_C)] = \int \widehat{f}(x_S, x_C) dP(x_C) \tag{11}$$

where x_S corresponds to the variable used for plotting the partial de-

pendency function, and x_C represents the other variables used in the machine learning model. The set S contains only one or two variables and the variables in S are used to explore their influence on the building fire loss rate. The total variable space x is composed of variable vectors x_S and x_C. The partial dependence plot marginalizes the output of the machine learning model on the distribution of variables in the set C, so that the function only displays the relationship between the variables in the set S of interest and the predicted results. By marginalizing other variables, we can obtain functions that only depend on the variables in S and the interaction with other variables. The partial dependency function $\widehat{f_{x_S}}$ can be estimated by calculating the mean value in the dataset, also known as the Monte Carlo method:

$$\widehat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{\{i=1\}}^{n}\left(x_S, x_C^{(i)}\right) \tag{12}$$

where $x_C^{(i)}$ is the actual value of the variable that we are not concerned about in the dataset, and n is the number of samples in the data set. The partial dependency function predicts the average marginal effect according to the given values of the variable S.

However, the partial dependency plot has great limitations in terms of independence assumptions and heterogeneous effects. The independence assumption is the biggest limitation of the partial dependency graph and is difficult to satisfy in actual situations. It assumes that the partial dependency variables in C are not related to other variables in S. If this assumption is violated, the average calculated by the partial dependency plot will contain extremely unlikely data points. When the variables are related, new data points with very low actual probability are created in the variable distribution area. Heterogeneous effects may be hidden because the partial dependency graph only shows the average marginal effect. If half of the data in a variable has a positive correlation with the predicted target, and the other half has a negative correlation with the predicted target, the partial dependence plot will become a horizontal line, and the two halves of the dataset will cancel each other out, thus hiding the heterogeneous effects in the data. Alternatively, accumulated local effect (ALE) plots can also describe how variables influence the prediction of a model on average even when variables are correlated. It incorporates many small fluctuations in a high number of intervals. PDP and ALE are often used at the same time, but they can only show the average effect.

The SHAP-based partial dependency plot [56] is a combination of the partial dependency graph (PDP) and the accumulated local effect (ALE), which reflects the average and variance of model output clearly based on PDP and ALE. It shows the marginal influence of one or two variables on the prediction results of the machine learning model and reflects the linear and monotonic relationship between the predicted target and the variables. It also shows the variance of the Shapley values and highlights the interaction. The partial dependency plot based on SHAP is displayed in the form of scattered points $\{(x_j^{(i)}, \phi_j^{(i)})\}_{i=1}^{n}$, indicating the influence of a single variable on the model prediction. Each point in the plot is a single prediction from the dataset. For each sample, a point for the variable value is drawn on the x-axis, and the Shapley value corresponding to the variable is drawn on the y-axis, indicating the influence degree of the variable to the model output of the sample. The interaction of the second most relevant variable can be expressed through color distinction, and the intensity of the interaction can be indicated by the lightness and darkness of the color.

Table 3
Parameter settings of base learners of IBTEM.

Model	Parameter	Optimized value
Catboost	*iteration*	500
	learning_rate	0.1
	max_*depth*	10
LightGBM	*n_estimator*	1000
	max_*bin*	200
	n_leaves	6
	learning_rate	0.05
	bagging_fraction	0.8
	feature_fraction	0.2
XGBoost	*learning_rate*	0.05
	n_estimator	1000
	max_*depth*	4
	gamma	0.6
	subsample	0.7
	colsample_bytree	0.7
	scale_pos_weight	1

The interaction effect is an additional combined variable effect after considering the effect of a single variable. The Shapley interaction index in the game theory is defined as:

$$\phi_{i,j} = \sum_{S\subseteq\backslash\{i,j\}} \frac{|S|!(m-|S|-2)!}{2(m-1)!}\delta_{ij}(S) \tag{13}$$

When i ≠ j, the $\delta_{ij}(S)$ can be defined as follows:

$$\delta_{ij}(S) = f_x(S\cup\{i,j\}) - f_x(S\cup\{i\}) - f_x(S\cup\{j\}) + f_x(S) \tag{14}$$

In order to obtain the pure interaction after considering each effect, the main effect of the variable is subtracted from the formula. The values in all possible variable set S are averaged and each sample gets a matrix of $m*m$ dimensional variables when calculating the Shapley values of all variables.

4. Experimental results and analysis

4.1. Performance metrics

Referred to some literatures on regression methods in [17,57] etc., we employ the following indicators to evaluate the performance of learning models for sake of comparison. The robustness of models is inspected with 10-fold cross-validation. Suppose that n denotes the number of samples, y_i represents the real value of the i sample; $y_i^{'}$ represents the predicted value of the i sample and $\bar{y}$ reflects the mean value of sample.

Root Mean Square Error (RMSE) can be defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{'} - y_i\right)^2} \tag{15}$$

Mean Square Error (MSE) can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{'} - y_i\right)^2 \tag{16}$$

Mean Absolute Error (MAE) is often used in machine learning model and can be defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{'} - y_i\right| \tag{17}$$

Median Absolute Error (MedAE) can be defined as:

$$MedAE = median\left(\left|y_1 - y_1^{'}\right|, \ldots, \left|y_n - y_n^{'}\right|\right) \tag{18}$$

Coefficient of determination (R^2) reflects the fitting degree and can be defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{19}$$

Where n represents the number of samples; RMSE, MAE, MSE, and MedAE are used to measure prediction error, and the smaller these indicators are, the better the model is. R^2 is used to measure the prediction accuracy and the closer it is to 1, the more accurate it is.

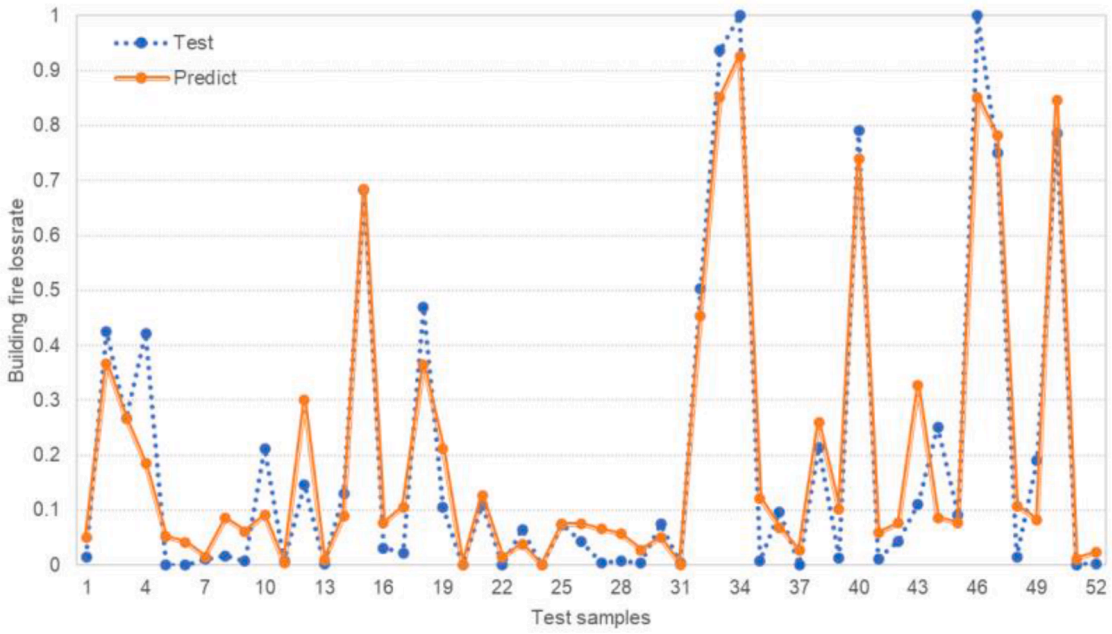


Fig. 6. Fitting renderings of partial test samples.

4.2. Parameter settings

The most important step of model training is parameter optimization, which directly affects the model performance. In this experiment, the filtered variables are selected as input, and the property loss rate is the output. Then, the samples are evaluated with ten-fold cross-validation. In the process of model training, iterations and learning rate are necessary. In terms of iterations, as the number of iterations increases, the cost function converges, and the error is gradually minimized. It is appropriate to set iteration value for each base model when performance improvement is not obvious. If the learning rate is set too low, the training will progress slowly. In contrast, oscillation will occur if the learning rate is too high. Combined with previous study [58] and hyper parameters tuning with grid search, the primary parameter settings of base learners of IBTEM are listed as Table 3.

In Table 3, the parameters *iteration* and *n_estimator* denote the maximum number of used trees and are usually set from 0 to 1000. The parameter max_*bin* means the maximum number of bins. The parameter *n_leaves* corresponds to the number of leaves in a tree. The parameter *learning_rate* stands for the speed of model training. The parameters *bagging_fraction* and *feature_fraction* should be set together. They select part of data randomly without resampling, which can be used to speed up training and handle overfitting. The parameter max_*depth* corresponds to the max depth of the tree. The parameter *gamma* denotes the minimum loss reduction required to make a further partition on a leaf node of the tree. The parameter *Subsample* means the secondary sampling ratio of the training set. The parameter *colsample_bytree* denotes the proportion of variables sampled when establishing the tree. The parameter *scale_pos_weight* is used to make the algorithm converge faster.

4.3. Ensemble prediction results

Based on the above parameter settings, the building fire loss prediction model, IBTEM, is trained and tested. Fig. 6 shows the actual loss rate and the predicted loss rate of partial test samples. From Fig. 6, we can see the specific gap between the predicted value and the actual value of each sample. Overall, the IBTEM model shows good generalization ability in the test data.

We compared our method with popular single learning regression methods such as decision tree (DT), multilayer perceptron (MLP),

Table 4
Comparison results with popular regression methods.

Model	Prediction Time/s	RMSE	MSE	MAE	MedAE	R^2
DT	0.024 (+/- 0.002)	0.290 (+/- 0.003)	0.087 (+/- 0.005)	0.187 (+/- 0.007)	0.091 (+/- 0.007)	0.113 (+/- 0.028)
MLP	0.068 (+/- 0.010)	0.275 (+/- 0.038)	0.065 (+/- 0.015)	0.177 (+/- 0.024)	0.122 (+/- 0.032)	0.340 (+/- 0.163)
Dummy	**0.004** (+/- **0.000**)	0.315 (+/- 0.003)	0.099 (+/- 0.008)	0.257 (+/- 0.008)	0.235 (+/- **0.001**)	−0.003 (+/- **0.004**)
Ridge	0.032 (+/- 0.002)	0.255 (+/- 0.033)	0.061 (+/- 0.009)	0.176 (+/- 0.006)	0.127 (+/- 0.004)	0.383 (+/- 0.107)
Lasso	0.020 (+/- 0.001)	0.278 (+/- 0.003)	0.077 (+/- 0.006)	0.220 (+/- 0.007)	0.171 (+/- 0.003)	0.218 (+/- 0.006)
Elastic net	0.020 (+/- 0.001)	0.274 (+/- 0.003)	0.075 (+/- 0.005)	0.215 (+/- 0.006)	0.167 (+/- 0.003)	0.238 (+/- 0.007)
Bayesian	0.021 (+/- 0.001)	0.250 (+/- 0.022)	0.061 (+/- 0.009)	0.176 (+/- 0.006)	0.127 (+/- 0.004)	0.383 (+/- 0.106)
PA	0.016 (+/- 0.002)	0.370 (+/- 0.010)	0.129 (+/- 0.005)	0.331 (+/- 0.008)	0.378 (+/- 0.013)	−0.313 (+/- 0.134)
BAG	0.118 (+/- 0.007)	0.214 (+/- 0.002)	0.047 (+/- 0.003)	0.147 (+/- 0.007)	0.092 (+/- 0.009)	0.526 (+/- 0.020)
RF	0.495 (+/- 0.016)	0.205 (+/- 0.002)	0.043 (+/- 0.003)	0.142 (+/- 0.007)	0.090 (+/- 0.007)	0.566 (+/- 0.018)
ADB	0.069 (+/- 0.009)	0.242 (+/- 0.002)	0.059 (+/- 0.003)	0.197 (+/- 0.007)	0.153 (+/- 0.012)	0.404 (+/- 0.017)
GBDT	0.056 (+/- 0.004)	0.259 (+/- 0.003)	0.067 (+/- 0.007)	0.184 (+/- 0.009)	0.110 (+/- 0.003)	0.322 (+/- 0.017)
IBTEM (Ours)	0.138 (+/- 0.010)	**0.140** (+/- **0.001**)	**0.020** (+/- **0.001**)	**0.085** (+/- **0.001**)	**0.047** (+/- **0.001**)	**0.621** (+/- 0.006)

Table 5
Non-parametric Friedman test for various methods ($\alpha = 0.05$).

Indicator	Statistic	p-value
RMSE	53.2814	0.0000
MSE	53.2814	0.0000
MAE	175.5634	0.0000
MedAE	143.2305	0.0000
R^2	53.2814	0.0000

Dummy regression, Ridge regression, Lasso regression, Elasticnet, Bayesian network, Passive Aggressive (PA), and ensemble learning regression methods, bagging (BAG), random forest (RF), adaboost (ADB), gradient boost decision tree (GBDT). The ten-fold cross validation is used in the comparative experiments. A series of experiments are conducted and optimal parameters are set for comparative methods. The mean value and standard deviation of each model performance in terms of RMSE, MSE, MAE, MedAE, R^2 are computed and shown in Table 4.

From Table 4, our proposed method IBTEM achieves the best performance with 0.02 MSE, 0.047 MedAE, 0.085 MAE and 0.621 R^2, which realize 70.15% improvement in MSE, 53.8% improvement in MAE, 57.27% improvement in MedAE and 92.86% improvement in R^2 compared with the representative boosting algorithm, Gradient boosting decision tree (GBDT). The performance of IBTEM is superior to popular single learning methods like decision tree (DT), regression methods like ridge regression and ensemble methods like random forest (RF). Dummy performs better in standard deviation and prediction time. However, it cannot achieve satisfying imitative effect and prediction error. IBTEM truly costs a little more time in prediction stage but 0.138 second for nearly 800 instances prediction is negligible. Moreover, it has around 72.12% improvement compared with representative ensemble learning method Random forest in prediction time. Thus, we can conclude that IBTEM is a promising model for prediction analyses since it has the ability to solve complex problems with good performance in interpretable ways.

To further validate the prediction performance of the proposed method, Friedman test is conducted to perform hypothesis testing according to the non-parametric testing with the significant level $\alpha = 0.05$ in Table 5. We analyze the statistical significance of the results through computing statistics and p-value in terms of different performance indicators on 10-fold cross-validation datasets. Null hypothesis (H_0) means that the methods perform equally well. According to the results of the Friedman test in Table 5, the obtained p-value of RMSE, MSE, MAE, MedAE and R^2 are 0.0000, which are less than α, so we reject the null-hypothesis H_0. Then the *post hoc* Holm test procedure is conducted to achieve more accurate statistical analysis of our proposed method with other methods. In Table 6 we report the results of the *post hoc* Holm tests, which show that our method is significantly superior to methods Adaboost, GBDT, multilayer perceptron, dummy regression, ridge regression, lasso regression, elastic net regression etc. in terms of RMSE, MSE, MAE, MedAE and R^2.

To verify the efficiency of applying a series of strategies, i.e. R^2-based weighted voting, we randomly select one-fold of dataset to compare the method before and after introducing sample quantile of order P (Quantile), winsorization (Winsorize), logarithmic transformation (Log), recursive feature elimination (RFE) and Pearson correlation coefficients (Pearson). The corresponding performance comparison results are shown in Table 7.

From the Table 7, we can find IBTEM achieves the best performance in terms of RMSE, MAE, MSE, MedAE and R^2 which adopts R^2-based weighted voting, winsorization, logarithmic transformation, recursive feature elimination and Pearson correlation coefficients. Catboost, XGBoost and LightGBM achieve different degrees of improvement after introducing various strategies. It can be observed that the MSE and MAE of Winsorize-Log-RFE-Pearson-Catboost-Quantile are reduced by 4.43% and 6.31%, respectively. The MSE, MAE and MedAE of Winsorize-Log-

Table 7
Performance of methods before and after introducing various strategies.

Method	RMSE	MSE	MAE	MedAE	R^2
Catboost	0.1425	0.0203	0.0903	0.0504	0.6077
Catboost-Quantile	0.1421	0.0202	0.0882	0.0489	0.609
Winsorize-Catboost-Quantile	0.1421	0.0202	0.0879	0.0517	0.609
Winsorize-Log-Catboost-Quantile	0.1418	0.0201	0.0866	**0.0471**	0.6122
Winsorize-Log-RFE-Catboost-Quantile	0.1418	0.0201	0.0857	0.0494	0.6106
Winsorize-Log-RFE-Pearson-Catboost-Quantile	0.1393	**0.0194**	**0.0846**	0.0487	**0.6254**
XGBoost	0.1435	**0.0206**	0.0916	0.0570	**0.6018**
XGBoost-Quantile	0.1439	0.0207	**0.0902**	0.0566	0.5997
Winsorize-XGBoost-Quantile	0.1449	0.021	0.0906	**0.0542**	0.5947
Winsorize-Log-XGBoost-Quantile	0.1456	0.0212	0.0917	0.0543	0.5899
Winsorize-Log-RFE-XGBoost-Quantile	0.1456	0.0212	0.0919	0.0557	0.5904
Winsorize-Log-RFE-Pearson-XGBoost-Quantile	0.1442	0.0208	0.0912	0.0556	0.5982
LightGBM	0.1456	0.0212	0.0932	0.0550	0.5903
LightGBM-Quantile	0.1456	0.0212	0.0908	0.0531	0.5895
Winsorize-LightGBM-Quantile	0.1453	0.0211	0.0901	0.0522	0.5921
Winsorize-Log-LightGBM-Quantile	0.1432	0.0205	0.0874	0.0517	0.6038
Winsorize-Log-RFE-LightGBM-Quantile	0.1425	0.0203	0.0869	0.0528	0.6078
Winsorize-Log-RFE-Pearson-LightGBM-Quantile	0.1407	**0.0198**	**0.0851**	**0.0487**	**0.6165**
IBTEM (Ours)	**0.1382**	**0.0191**	**0.0844**	0.0478	**0.6314**

Table 6
Holm post hoc tests for IBTEM ($\alpha = 0.05$).

Method	RMSE		MSE		MAE		MedAE		R^2	
	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
DT	5.3398	0.0000	5.3398	0.0000	3.1005	0.0097	0.9761	0.6580	5.3398	0.0000
MLP	4.1340	0.0003	4.1340	0.0003	4.1340	0.0003	4.3063	0.0001	4.1340	0.0003
Dummy	6.0862	0.0000	6.0862	0.0000	6.2584	0.0000	6.2584	0.0000	6.0862	0.0000
Ridge	2.6986	0.0245	2.6986	0.0245	2.1531	0.1252	3.3876	0.0042	2.6986	0.0245
Lasso	4.7082	0.0000	4.7082	0.0000	5.6269	0.0000	5.5694	0.0000	4.7082	0.0000
Elastic Net	4.1340	0.0003	4.1340	0.0003	4.9953	0.0000	4.9953	0.0000	4.1340	0.0003
Bayesian	2.8134	0.0245	2.8134	0.0245	2.0957	0.1252	3.2728	0.0053	2.8134	0.0245
PA	6.8326	0.0000	6.8326	0.0000	6.8900	0.0000	6.8900	0.0000	6.8326	0.0000
BAG	1.1483	0.5017	1.1483	0.5017	1.1483	0.5017	1.7225	0.2549	1.1483	0.5017
RF	0.5742	0.5659	0.5742	0.5659	0.5742	0.5659	0.9187	0.6580	0.5742	0.5659
ADB	2.8134	0.0245	2.8134	0.0245	4.2489	0.0002	4.1340	0.0003	2.8134	0.0245
GBDT	3.5024	0.0028	3.5024	0.0028	3.5599	0.0022	2.3541	0.0743	3.5024	0.0028

Table 8
Comparisons with methods for disaster prediction in other literatures.

Method	Target Value	RMSE	MSE	MAE	R^2
Artificial neural network [17]	Tornado property damage	–	0.0918	–	0.432
Propensity Score Matching (PSM) [59]	Flood losses	–	–	–	Nearest Neighbor (0.474) Kernel Matching (0.521)
GA-Elman neural networks +SVR +generalized regression neural networks (GRNN) [60]	Direct economic losses of tropical cyclone	Combined model (3.30) GA–Elman (5.05) SVR (7.85) GRNN (3.82)	–	–	–
log-log regression [14]	Earthquake property damage	–	–	–	0.60
IBTEM (Ours)	Building fire property loss rate	**0.14**	**0.02**	0.085	**0.621**

RFE-Pearson-LightGBM-Quantile are also reduced by 6.6%, 8.69% and 11.46%, respectively. In terms of R^2, Catboost and LightGBM increased by 2.91% and 4.44% through strategies. The final IBTEM has lower MSE, MAE and MedAE compared with original Catboost, XGBoost and LightGBM and higher performance in R^2, which illustrates the advantages of the proposed IBTEM loss prediction method. The computational cost in the prediction stage of Catboost, XGBoost, LightGBM and IBTEM are all 0.0001 s for each sample, where the difference can be ignored.

To validate the effectiveness of our proposed method in build fire loss prediction, we perform it based on 10-fold cross-validation procedure and compare the mean value with the state-of-the-art methods in recent literatures on disaster loss prediction. The methods in literatures are not evaluated with all the performance indicators, so the missing value of corresponding indicator is replaced with "–". The comparative results are shown in Table 8.

Diaz and Joseph use artificial neural networks to predict tornado-induced property damage, omitting tornado events which cause no property damage. Conditional on damage caused by tornadoes, neural network predicts the amount of damage with results of MSE = 0.0918 and R^2 = 0.432. However, model fitting requires log-scale data which leads to large natural-scale residuals. Allaire use propensity score matching (PSM) to determine if a causal relationship exists between social media and flood loss reduction. Regression of flood losses on key covariates is done to estimate average treatment effect on the treated (ATT), using regression analysis and PSM matched samples. The R^2 of nearest neighbor and kernel matching turn out to be 0.474 and 0.521, respectively. Chen et al. conduct three models, GA-Elman neural networks, support vector regression (SVR) and generalized regression neural networks (GRNN), to predict direct economic losses of tropical cyclone, obtaining the root mean square error (RMSE) value of 5.05, 7.85 and 3.82, respectively. Then they make use of exhaustive attack method, which takes 0.01 as the grad to search for the optimal weighted group, to combine the three models into a comprehensive evaluation model with the RMSE value of 3.30. Heatwole and Rose estimate a log-log regression equation for property damage from significant U.S. earthquakes, with the adjusted R^2 value of 0.60. The comparative result proves that the proposed IBTEM is able to precisely predict the loss rate of property in building fires and perform better than those in other literatures.

Table 9
Effects of multi-source data fusion.

Data source	RMSE	MSE	MAE	MedAE	R^2
fire	0.1694	0.0287	0.1021	0.0595	0.4980
basic	0.2020	0.0408	0.1299	0.0763	0.2107
weather	0.2421	0.0586	0.1653	0.1151	−0.0254
fire + basic	0.1393	0.0194	**0.0842**	**0.0449**	0.6246
fire +weather	0.1691	0.0286	0.1020	0.0574	0.4985
basic + weather	0.2030	0.0412	0.1319	0.0798	0.2031
fire + basic + weather (ours)	**0.1382**	**0.0191**	0.0844	0.0478	**0.6314**

Algorithm 1
Boosting Tree Ensemble Algorithm for Building Fire Loss Prediction.

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, K-fold K, Base learner h_{cat}, h_{xgb}, h_{lig}
Output: Final model $G(x)$
1: **For** $k = 1$ to K **do**
2: Split D into training set T and validation set V with K-fold cross-validation
3: Initialize the parameters settings of h_{cat}, h_{xgb}, h_{lig} and train models on T
4: Calculate the coefficient of determination (R^2) of each base learner R^2_{cat}, R^2_{xgb}, R^2_{lig} on V
5: **End for**
6: Calculate the mean values of R^2 as $mR^2$for each base learner
7: Compute the weight of each base learner through formula $\omega_{\text{model}} = \frac{mR^2_{\text{model}}}{\sum_{i=1}^{M} mR^2_i}$, where M denotes the number of base learners
8: Construct a linear combination of the basic classifiers $G(x) = \omega_{\text{cat}} h_{\text{cat}} + \omega_{\text{xgb}} h_{\text{xgb}} + \omega_{\text{lig}} h_{\text{lig}}$
9: **End**

4.4. Multi-source data fusion experiment

Multi-source data provides a wealth of cross-information. Through fusion, the quality of data can be further improved and more valuable information can be mined. We conduct a variety of experiments to verify the improvement through multi-source data fusion. The comparison results are shown in Table 9.

As shown in Table 9, the model constructed based on multi-source data including fire data, basic data and weather data achieves the better results with 0.0191 in MSE and 0.6314 in R^2. The quality of the data determines the upper limit of the effect. Thus, we can conclude that the multi-source fusion realizes the complementarity of valid information of data and provides valuable information for data mining based on novel model IBTEM. Thus, these results convincingly prove the scalability and superiority of the proposed IBTEM.

5. Discussion

The influence of each indicator weighing on property loss is of great concern. It is of great significance to objectively evaluate the contribution of each indicator to reduce property loss due to building fires. As mentioned in Section 3.3, we use SHAP method to visualize the factors of building fire loss. We calculate the Shapley values per variable and order them according to their values. The variable with higher Shapley values possesses more important roles in predictions.

SHAP decision plots show how complex models arrive at their predictions. Fig. 7 is the SHAP decision plot introduced in Section 3.3, which shows the variable importance ranking with SHAP values.

In Fig. 7, we can find the variable importance for fire loss prediction ranks from the top to the bottom, which are PROP_VAL (Property Value), FIRE_SPRD (Fire Spread), ST_DAM_MIN (Number of Stories with Minor Damage), ST_DAM_HVY (Number of Stories with Heavy Damage), ST_DAM_SIG (Number of Stories with Significant Damage), ST_DAM_XTR (Number of Stories with Extreme Damage), AID (Aid Given or Received), APP (Number of Suppression, EMS and Other Apparatus), CONT_VAL (Contents Value), PROP_USE (Fixed Property

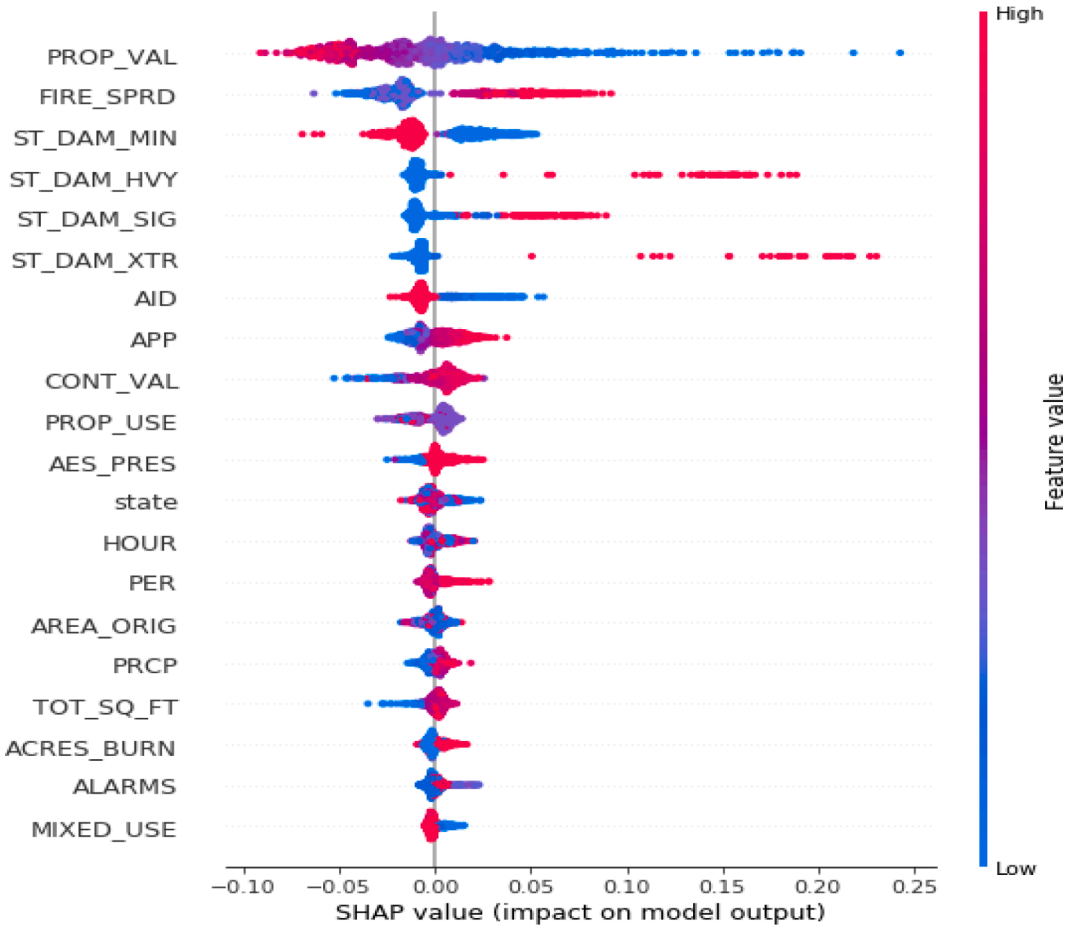
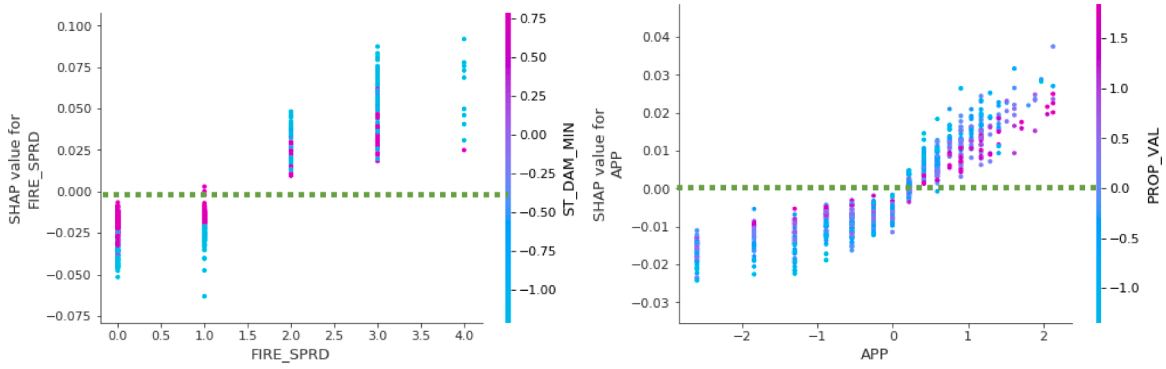


Fig. 7. The variable importance ranking with SHAP values.



(a) (b)

Fig. 8. The example of partial dependence plot with SHAP values.

Use), AES_PRES (Presence of Automatic Extinguishing System), state (The state where the incident occurred), HOUR (Incident Time (Hour)). We are supposed to focus on the important factors according to the variable importance ranking when the new building fire incident occurs.

Besides, the red value reflects the positive correlation and the blue value indicates the negative correlation between the variable and the building fire loss rate. The FIRE_SPRD, ST_DAM_HVY, ST_DAM_SIG, ST_DAM_XTR, APP, CONT_VAL, AES_PRES indicate the positive correlation while PROP_VAL, ST_DAM_MIN, AID show the negative correlation with building fire property loss rate.

To reduce property loss due to building fires, we need to focus on the high-valued property and install adequate fire protection equipment. We also ought to put emphasis on the fire confined to building of origin or beyond building of origin, which usually cause the huge property loss. Rescue aid and apparatus should be dispatched reasonably, and the response speed should be improved. In future fire control management, the laying scope of fire prevention facilities inside buildings should be enlarged, and its effectiveness should be checked regularly to ensure that the facilities are in an effective state when accidents occur.

In Fig. 8, we have listed a series of partial dependence plots (PDP) with Shapley values of top important variables. In the plot, each dot corresponds to a single prediction from the dataset. The x-axis represents the value of the variable and the y-axis is the SHAP value for the variable, indicating the effect of variable value change degree on the

Table A1
Building fire loss rate prediction variable set.

No.	Variable	Data Type	Description	Code
1	ADD_WILD [c)]	Category	Address on Wildland Flag	Yes or No
2	AID [d)]	Category	Aid Given or Received	1 Mutual aid received 2 Automatic aid received 3 Mutual aid given 4 Automatic aid given …
3	SHIFT [d)]	Category	Shift responds to an incident	–
4	ALARMS [d)]	Category	The level of alarms	–
5	APP_MOD [d)]	Category	Apparatus/ Personnel Module Used	Yes or No
6	RESOU_AID [d)]	Category	Resources Include Mutual Aid	Yes or No
7	DET_ALERT [b)]	Category	Detector Alerted Occupants	1 Detector alerted occupants 2 Detector did not alert them U Unknown
8	HAZ_REL [b)]	Category	Hazardous Material Released	1 Natural gas 2 Propane gas 3 Gasoline, vehicle fuel tank or portable container 4 Kerosene …
9	MIXED_USE [a)]	Category	Mixed Use Property	10 Assembly use 20 Education use 33 Medical use 40 Residential use …
10	PROP_USE [a)]	Category	Fixed Property Use	161 Restaurant or cafeteria 215 High school, junior high, middle school 311 Nursing homes licensed by the State 419 1- or 2-family dwelling …
11	NOT_RES [a)]	Category	Not Residential Flag	Yes or No
12	LESS_1ACRE [b)]	Category	Less than one Acre	Yes or No
13	AREA_ORIG [b)]	Category	Area of Origin	21 Bedroom for less than five people 24 Cooking area, kitchen 47 Vehicle storage area: garage, carport. 74 Attic …
14	HEAT_SOURC [b)]	Category	Heat Source	13 Electrical arcing 43 Hot ember or ash 61 Cigarette 73 Lightning discharge …
15	FIRST_IGN [b)]	Category	Item First Ignited	17 Structural member or framing 25 Appliance housing or casing 76 Cooking materials 81 Electrical wire, cable insulation …
16	TYPE_MAT [b)]	Category	Type of Material	27 Cooking oil, transformer oil, lubricating oil 41 Plastic, regardless of type. 63 Sawn wood 71 Fabric, cotton, blends, rayon, wool, finished goods …
17	CAUSE_IGN [b)]	Category	Cause of Ignition	1 Intentional 2 Unintentional 3 Failure of equipment or heat source 4 Act of nature …
18	FACT_IGN_1 [b)]	Category	Factors Contributing To Ignition	12 Heat source too close to combustibles 34 Arc from faulty contact, broken conductor 53 Equipment unattended 62 Storm …
19	HUM_FAC_1 [b)]	Category	Human Factors	1 Asleep N None
20	STRUC_STAT [a)]	Category	Structure Status	1 Under construction 2 In normal use 3 Idle, not routinely used 4 Under major renovation …
21	FLAME_SPRD [b)]	Category	No Flame Spread	Yes or No
22	DETECTOR [b)]	Category	Presence of Detectors	1 Present. N None present. U Undetermined
23	AES_PRES [b)]	Category	Presence of Automatic Extinguishing System	1 Present. 2 Partial System Present. N None present. U Undetermined
24	STATE [c)]	Category	The State where the incident occurred	CA California FL Florida NC North Carolina TX Texas …
25	FIRE_SPRD [b)]	Category	Fire Spread	1 Confined to object of origin. 2 Confined to room of origin. 3 Confined to floor of origin. 4 Confined to building of origin. 5 Beyond building of origin.
26	YEAR [c)]	Category	Incident Date (Year)	From 2012 to 2016
27	MONTH [c)]	Category	Incident Date (Month)	From 1 to 12
28	HOUR [c)]	Category	Incident Time (Hour)	From 0 to 23
29	RESP [d)]	Numeric	Response Time	–
30	APP [d)]	Numeric	Number of Suppression, EMS and Other Apparatus	–
31	PER [d)]	Numeric	Number of Suppression, EMS and Other Personnel	–
32	PROP_VAL [a)]	Numeric	Property Value	–
33	CONT_VAL [a)]	Numeric	Contents Value	–
34	NUM_UNIT [a)]	Numeric	Number of Residential Units	–

(*continued on next page*)

Table A1 (*continued*)

35	BLDG_INVOL [a)]	Numeric	Number of Buildings Involved	–
36	ACRES_BURN [b)]	Numeric	Acres Burned	–
37	BLDG_ABOVE [a)]	Numeric	Building Height: Stories Above Grade	–
38	BLDG_BELOW [a)]	Numeric	Building Height: Stories Below Grade	–
39	TOT_SQ_FT [a)]	Numeric	Total Square Feet	–
40	FIRE_ORIG [b)]	Numeric	The story of fire origin	–
41	ST_DAM_MIN [b)]	Numeric	Number of Stories with Minor Damage	–
42	ST_DAM_SIG [b)]	Numeric	Number of Stories with Significant Damage	–
43	ST_DAM_HVY [b)]	Numeric	Number of Stories with Heavy Damage	–
44	ST_DAM_XTR [b)]	Numeric	Number of Stories with Extreme Damage	–
45	AWND [c)]	Numeric	Average Wind	–
46	PRCP [c)]	Numeric	Precipitation	–
47	SNOW [c)]	Numeric	Snowfall	–
48	SNWD [c)]	Numeric	Snow Depth	–
49	TAVG [c)]	Numeric	Average Temperature	–
50	TMAX [c)]	Numeric	Maximum Temperature	–
51	TMIN [c)]	Numeric	Minimum Temperature	–
52	WDF5 [c)]	Numeric	Wind Direction	–
53	WSF5 [c)]	Numeric	Wind Speed	–

a) indicates this variable belongs to disaster-affected body, b) indicates this variable belongs to disaster-inducing factor, c) indicates this variable belongs to disaster-formative environment, and d) indicates this variable belongs to emergency activities.

predicted results of the model. The color corresponds to a second variable that may have the interaction effect with the first variable. In the presence of the interact variable, the variable with the largest interaction effect with the first variable will show up as a distinct vertical pattern of coloring.

In Fig. 8(a), the green line represents the dividing line of the Shapley value. It is better to have the lower of the property loss rate. The samples above the line indicates that FIRE_SPRD has positive impact on property loss rate and vice versa. The FIRE_SPRD is a categorical value and the meaning of each value can be referenced in Table A1. The variable ST_DAM_MIN has an obvious interaction effect on FIRE_SPRD. The plot indicates the fire loss confined to floor, building and beyond building of origin result in higher property loss risk. Meanwhile, the lower value of ST_DAM_MIN results in higher fire loss rate. The fire loss confined to object and room of origin contributes less to fire property loss and the larger value of ST_DAM_MIN results in higher fire loss rate.

Similarly, for the example in Fig. 8(b), the building loss rate rises with the increasing of APP and large APP with a higher PROP_VAL has less impaction on the property loss rate, which suggests an interaction effect between APP and PROP_VAL.

6. Conclusions

In view of recent studies investigating property loss prediction due to building fires, efforts to improve prediction models realized by machine learning remains lacking. Therefore, we propose an interpretable boosting tree ensemble method (IBTEM) to realize property loss prediction due to building fires with multisource of NFIRS and NOAA from 2012 to 2016. An R^2-based weighted voting strategy is designed by integrating the weights of the base learners to improve the prediction performance of property loss due to building fires. The interpretable Shapley additive explanations are applied for visual internal mechanism analysis of boosting ensemble methods for building fire loss predictions, which give great decision support for emergency management department. Based on the comparative results experimented in this study, the following can be concluded:

(1) We fused multisource data on American building fire incidents and their corresponding weather from 2012 to 2016, and constructed four variable scenario subsets for building fire loss prediction based on the regional disaster system theory.
(2) We applied Winsorization, logarithmic transformation, recursive feature elimination (RFE), correlation analysis and other strategies to improve data quality and add the sample quantile of order p strategy and R^2-based weighted voting strategy to improve the model performance.
(3) It was effective to predict the property loss rate due to building fires based on the proposed boosting tree ensemble method, which was superior to the comparative methods and considerably improves the predictive performance of based predictors including CatBoost, XGBoost, and LightGBM.
(4) Interpretable analysis including the variable importance ranking and partial dependence plot based on Shapley additive explanation was provided. According to the variable importance assessed by Shapley values, Property Value, Fire Spread, Number of Stories with Minor Damage, Apparatus etc. were shown to be the most influential in property loss prediction.

In this study, a large number of building fire data samples are used for the experiments, and our proposed method achieves promising performance. Therefore, this method can be extended to loss prediction due to building fires to assist relevant departments in making timelier decisions regarding dispatching aid and mobilizing resources. Furthermore, we compensate for the lack of accuracy of traditional predicting methods and propose a novel ensemble method that is more suitable to solve the problem of property loss prediction due to building fires. This method is able to predict loss more accurately in advance, which is of practical significance for rationally planning investment and improving recovery efficiency. In future, more updated data can be included for model improvement and building fire loss prediction based on time series forecasting deserves to be studied.

CRediT authorship contribution statement

Ning Wang: Supervision, Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing, Data curation. **Yan Xu:** Data curation, Writing – original draft, Methodology. **Sutong Wang:** Methodology, Visualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 71774021, Grant 71533001 and Grant 71874020.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ress.2022.108587.

References

[1] Himoto K, Suzuki K. Computational framework for assessing the fire resilience of buildings using the multi-layer zone model. Reliab Eng Syst Saf 2021;216. https://doi.org/10.1016/j.ress.2021.108023.

[2] Ioannou I, Aspinall W, Rush D, Bisby L, Rossetto T. Expert judgment-based fragility assessment of reinforced concrete buildings exposed to fire. Reliab Eng Syst Saf 2017;167:105–27. https://doi.org/10.1016/j.ress.2017.05.011.

[3] Rahardjo HA, Prihanton M. The most critical issues and challenges of fire safety for building sustainability in Jakarta. J Build Eng 2020;29. https://doi.org/10.1016/j.jobe.2019.101133.

[4] Evarts B. Fire loss in the United State during 2017. Natl Fire Prot Assoc 2018;5:1–20.

[5] Yu YC, Gardoni P. Predicting road blockage due to building damage following earthquakes. Reliab Eng Syst Saf 2022;219:108220. https://doi.org/10.1016/J.RESS.2021.108220.

[6] Hsu WK, Huang PC, Chang CC, Chen CW, Hung DM, Chiang WL. An integrated flood risk assessment model for property insurance industry in Taiwan. Nat Hazards 2011;58:1295–309. https://doi.org/10.1007/s11069-011-9732-9.

[7] Teemu K, Topi S, Simo H, Olavi KR. A monte carlo simulation platform of housing fires in Finland forecasting life and property loss. In: 11th Int. Probabilistic Saf. Assess. Manag. Conf. Annu. Eur. Saf. Reliab. Conf. 2012. 2. PSAM11 ESREL; 2012. p. 1584–93.

[8] Jagger TH, Elsner JB, Burch KK. Climate and solar signals in property damage losses from hurricanes affecting the United States. Nat Hazards 2011;58:541–57. https://doi.org/10.1007/s11069-010-9685-4.

[9] Khan S, Shanmugam R, Hoeschen B. Injury, fatal, and property damage accident models for highway corridors. Transp Res Rec 1999:84–92. https://doi.org/10.3141/1665-12.

[10] Cao WH, Huang CF, Zhao HP, Zhao SJ. Typhoon risk assessment of enterprise property: a case study on Taizhou city of Zhejiang province. Xitong Gongcheng Lilun Yu Shijian/System Eng Theory Pract 2012;32:425–32.

[11] Westerling AL, Bryant BP. Climate change and wildfire in California. Clim Change 2007;87. https://doi.org/10.1007/s10584-007-9363-z.

[12] Fronstin P, Holtmann AG. The determinants of residential property damage caused by hurricane Andrew. South Econ J 1994;61:387. https://doi.org/10.2307/1059986.

[13] Hanea DM, Jagtman HM, Ale BJM. Analysis of the Schiphol Cell Complex fire using a Bayesian belief net based model. Reliab Eng Syst Saf 2012;100:115–24. https://doi.org/10.1016/j.ress.2012.01.002.

[14] Heatwole N, Rose A. A reduced-form rapid economic consequence estimating model: application to property damage from U.S. earthquakes. Int J Disaster Risk Sci 2013;4:20–32. https://doi.org/10.1007/s13753-013-0004-z.

[15] Worrell C, Luangkesorn L, Haight J, Congedo T. Machine learning of fire hazard model simulations for use in probabilistic safety assessments at nuclear power plants. Reliab Eng Syst Saf 2019;183:128–42. https://doi.org/10.1016/j.ress.2018.11.014.

[16] Bhardwaj U, Teixeira AP, Guedes Soares C, Ariffin AK, Singh SS. Evidence based risk analysis of fire and explosion accident scenarios in FPSOs. Reliab Eng Syst Saf 2021;215:107904. https://doi.org/10.1016/j.ress.2021.107904.

[17] Diaz J, Joseph MB. Predicting property damage from tornadoes with zero-inflated neural networks. Weather Clim Extrem 2019;25:100216. https://doi.org/10.1016/j.wace.2019.100216.

[18] Tsai CF, Lin YC, Yen DC, Chen YM. Predicting stock returns by classifier ensembles. Appl Soft Comput J 2011;11:2452–9. https://doi.org/10.1016/j.asoc.2010.10.001.

[19] Aburomman AA. Ibne Reaz M Bin. A novel SVM-kNN-PSO ensemble method for intrusion detection system. Appl Soft Comput J 2016;38:360–72. https://doi.org/10.1016/j.asoc.2015.10.011.

[20] Zou Q, Chen S. Resilience-based Recovery Scheduling of transportation network in mixed traffic environment: a deep-ensemble-assisted active learning approach. Reliab Eng Syst Saf 2021;215. https://doi.org/10.1016/j.ress.2021.107800.

[21] Cárdenas-Gallo I, Sarmiento CA, Morales GA, Bolivar MA, Akhavan-Tabatabaei R. An ensemble classifier to predict track geometry degradation. Reliab Eng Syst Saf 2017;161:53–60. https://doi.org/10.1016/j.ress.2016.12.012.

[22] Li Z, Wu D, Hu C, Terpenny J. An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. Reliab Eng Syst Saf 2019;184:110–22. https://doi.org/10.1016/j.ress.2017.12.016.

[23] Tyralis H, Papacharalampous G, Burnetas A, Langousis A. Hydrological post-processing using stacked generalization of quantile regression algorithms: large-scale application over CONUS. J Hydrol 2019;577. https://doi.org/10.1016/j.jhydrol.2019.123957.

[24] Mastelini SM, Santana EJ, Cerri R, Barbon S. DSTARS: a multi-target deep structure for tracking asynchronous Regressor stacking. Appl Soft Comput J 2020;91. https://doi.org/10.1016/j.asoc.2020.106215.

[25] Ribeiro MHDM, dos Santos Coelho L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput J 2020;86. https://doi.org/10.1016/j.asoc.2019.105837.

[26] Naderpour M, Rizeei HM, Khakzad N, Pradhan B. Forest fire induced Natech risk assessment: a survey of geospatial technologies. Reliab Eng Syst Saf 2019;191. https://doi.org/10.1016/j.ress.2019.106558.

[27] Agarwal P, Tang J, Narayanan ANL, Zhuang J. Big data and predictive analytics in fire risk using weather data. Risk Anal 2020;40:1438–49. https://doi.org/10.1111/risa.13480.

[28] Guelman L. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst Appl 2012;39:3659–67. https://doi.org/10.1016/j.eswa.2011.09.058.

[29] Chan JCW, Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens Environ 2008;112:2999–3011. https://doi.org/10.1016/j.rse.2008.02.011.

[30] Chen Z, Jiang F, Cheng Y, Gu X, Liu W, Peng J. XGBoost classifier for DDoS attack detection and analysis in SDN-Based cloud. In: Proc. - 2018 IEEE Int. Conf. Big Data Smart Comput. BigComp; 2018. p. 251–6. https://doi.org/10.1109/BigComp.2018.00044.

[31] Zhang J, Mucs D, Norinder U, Svensson F. LightGBM: an Effective and scalable algorithm for prediction of chemical toxicity–application to the Tox21 and mutagenicity data sets. J Chem Inf Model 2019;59:4150–8. https://doi.org/10.1021/acs.jcim.9b00633.

[32] Fan J, Wang X, Zhang F, Ma X, Wu L. Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data. J Clean Prod 2020;248:119264. https://doi.org/10.1016/j.jclepro.2019.119264.

[33] Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron Commer Res Appl 2018;31:24–39. https://doi.org/10.1016/j.elerap.2018.08.002.

[34] Jiajun H, Chuanjin Y, Yongle L, Huoyue X. Ultra-short term wind prediction with wavelet transform, deep belief network and ensemble learning. Energy Convers Manag 2020;205:112418. https://doi.org/10.1016/j.enconman.2019.112418.

[35] Song Y, Liu X, Zhang L, Jiao X, Qiang Y, Qiao Y, et al. Prediction of double-high biochemical indicators based on lightGBM and XGBoost. ACM Int. Conf. Proceeding Ser. 2019:189–93. https://doi.org/10.1145/3349341.3349400.

[36] Nobre J, Neves RF. Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. Expert Syst Appl 2019;125:181–94. https://doi.org/10.1016/j.eswa.2019.01.083.

[37] Zhang C, Hu D, Yang T. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. Reliab Eng Syst Saf 2022;222:108445. https://doi.org/10.1016/J.RESS.2022.108445.

[38] Tan S, Moinuddin K. Systematic review of human and organizational risks for probabilistic risk analysis in high-rise buildings. Reliab Eng Syst Saf 2019;188:233–50. https://doi.org/10.1016/j.ress.2019.03.012.

[39] Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. Appl Soft Comput 2020;86:105941. https://doi.org/10.1016/j.asoc.2019.105941.

[40] Xin J, Huang C. Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management. Fire Saf J 2013;62:72–8. https://doi.org/10.1016/j.firesaf.2013.09.022.

[41] Chu G, Sun J. Decision analysis on fire safety design based on evaluating building fire risk to life. Saf Sci 2008;46:1125–36. https://doi.org/10.1016/j.ssci.2007.06.011.

[42] Yi GW, Qin HL. Fuzzy comprehensive evaluation of fire risk on high-rise buildings. Procedia Eng 2011;11:620–4. https://doi.org/10.1016/j.proeng.2011.04.705.

[43] Sund B, Jaldell H. Security officers responding to residential fire alarms: estimating the effect on survival and property damage. Fire Saf J 2018;97:1–11. https://doi.org/10.1016/j.firesaf.2018.01.008.

[44] Shi PJ. Theory and practice on disaster system research in a fifth time. J Nat Disasters 2009;18:1–9.

[45] Hastings C, Mosteller F, Tukey JW, Winsor CP. Low moments for small samples: a comparative study of order statistics. Ann Math Stat 1947;18:413–26. https://doi.org/10.1214/aoms/1177730388.

[46] Leydesdorff L, Bensman S. Classification and powerlaws: the logarithmic transformation. J Am Soc Inf Sci Technol 2006;57:1470–86. https://doi.org/10.1002/asi.20467.

[47] Stern MD, Bajpai AC, Mustoe LR, Walker D. Advanced engineering mathematics. Math Gaz 1991;75:246. https://doi.org/10.2307/3620303.

[48] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

[49] Twa M.D., Parthasarathy S., Raasch T.W., Bullimore M.A. Decision tree classification of spatial data patterns from Videokeratography using zernike polynomials, 2003, p. 3–12. https://doi.org/10.1137/1.9781611972733.1.

[50] Freund Y. Boosting a weak learning algorithm by majority. Inf Comput 1995;121:256–85. https://doi.org/10.1006/inco.1995.1136.

[51] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.

[52] Dorogush A.V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support. arXiv 2018.

[53] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[54] 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;2017:3147–55.

[55] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60. https://doi.org/10.1038/s41551-018-0304-0.

[56] Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 2020;82:1059–86. https://doi.org/10.1111/rssb.12377.

[57] Zheng J, Zhang L, Gong J, Wang W. Feature analysis and comparison of prediction methods for fire accidents. Int J Saf Secur Eng 2020;10:707–12. https://doi.org/10.18280/ijsse.100516.
[58] Cui S, Wang Y, Yin Y, Cheng TCE, Wang D, Zhai M. A cluster-based intelligence ensemble learning method for classification problems. Inf Sci (Ny) 2021;560:386–409. https://doi.org/10.1016/j.ins.2021.01.061.
[59] Allaire MC. Disaster loss and social media: can online information increase flood resilience? Water Resour Res 2016;52:7408–23. https://doi.org/10.1002/2016WR019243.
[60] Chen S, Tang D, Liu X, Chunhua H. Assessment of tropical cyclone disaster loss in Guangdong province based on combined model. Geomatics. Nat Hazards Risk 2018;9:431–41. https://doi.org/10.1080/19475705.2018.1447024.