An Accident Prediction Approach Based on XGBoost

Xuehuai Shi, Qianmu Li, Yong Qi*, Tiantian Huang, Jianmei Li
Nanjing University of Science and Technology, School of Computer Science and Technology, China
Nanjing, 210094
Corresponding author: Yong Qi, E-mail:790815561@qq.com

***Abstract*— As an important threat to public security, urban fire accident causes huge economic loss and catastrophic collapse. Predicting and analyzing the interior rule of urban fire accident from its appearance needed to be solved in the field. In this paper, we propose a new urban fire accident prediction approach based on XGBoost. The method determines the predictive indexes in a quantitative and qualitative way from different characteristics in various kinds of fire accidents. For screening the features we need, we adopt the feature selection algorithm based on association rules. For data cleaning, we use a method based on Box-Cox transformation that transforms the continual response variables from the feature space for removing the dependencies on unobservable errors and the predictor variable to some extent. Then we use the data to train the model based on XGBoost to obtain the best prediction accuracy. Experiments show that the method provides a feasible solution to urban fire accident prediction. The method contributes to improving the public security situation, we have added the method and related model to the City in a box™, Shenzhen Aerospace Smart City System Technology Co.,Ltd.**

Keywords: Box-Cox Transformation，XGBoost Algorithm，Accident Prediction

I. INTRODUCTION

Fire accidents happened to factories, warehouses, uptown and superstores keep the public attention of fire prevention growing. On June 14, 2017, 12 people died and at least 74 got injured after a massive fire ravaged a block of flats in central London. As reported by the Daily Mail, an early morning fire ripped through a barn in Pennsylvania on July 3, 2017, which killed at least 100,000 chickens and led to a heavy economic loss.

Urban fire accident prediction analyzes historical data and current state data in the regional fire accidents, taking the correlative factors into consideration to describe the future fire accident information. Through adopting corresponding countermeasures, we can reduce the fire accidents to some extent. The difficulties of fire accident prediction are the modeling of the indicator system and the computation of the level of fire risks. In the past, we put much resources and manpower to construct the indicator system and compute the magnitude of fire risks when the fire accident prediction model was built.

Recently, many researchers have achieved lots of markable achievements on constructing fire risk evaluation system. They build the fire accident prediction model with external data which is independent of the construction process, such as the weather data onto fire etc. For example, as shown in the work of Implementation of Artificial Neural Fuzzy Inference System in a Real Time Fire Detection Mechanism (Sharma D, 2016), in their work, they propose a hardware model that provides new fire detection and control mechanism with the interface of artificial neural network and fuzzy logic. The hardware consists of the temperature sensor, smoke sensor, flame detector and a microcontroller unit. In A Model Integrating Fire accident prediction and Detection for Rural-Urban Interface Area (Ooi C H, Chetty M, Teng S W,2007), they build a model that aims to predict fire risk from weather parameters and to detect smoke using video monitoring systems as smoke is the early sign of fire. The prediction algorithm in the model provides the fire danger index (FDI). In addition, many scholars have made remarkable accomplishments in the field of forest fire accident prediction, such as The Canadian Forest Fire Danger Rating System (Wang X, Wotton B M, Cantin A S, 2016) and The Canadian Forest Fire Behavior Prediction System (Taylor S W, Alexander M E, 2016) etc. However, almost none of them have provided a solution to predict the fire accident by analyzing the internal features of a building and the dynamic data onto a building that combines with the real-time data. The existing fire accident predictions are the macroscopic accident rate predictions that lack of the operability and corporeality.

In this paper, we classify the architectures by its usage, divide the architectures into factory, warehouse, public building and residence, and collect the historical data and current state data in the regional fire accidents, integrate other features of the architecture and real-time weather conditions into the original data. We select related features by the feature selection algorithm based on association rules and clean unobservable errors by a method based on Box-Cox transformation. Finally, we use the data to build the evaluation system based on XGBoost, tune the system parameters to obtain the best prediction accuracy to some degree.

II. Accident Prediction Method Based On XGBoost

XGBoost is the optimized ensemble algorithm based on GBDT (Gradient Boosting Decision Tree). The main idea of the boosting algorithm is that many decision trees perform better than a single one. Each decision tree may do a bad job. When many trees are integrated, the performance gets much better. The overall technical flow chart is shown in Figure 1.

Because the sample data size is ten million in the fire accident prediction model, neural networks or other boosting algorithms based on decision tree are computationally intractable. Aiming to handle massive data, XGBoost provides cache aware pre-read technology, distributed memory computing technology, AllReduce fault-tolerant tools for improving the calculation speed of existing boosting tree algorithms, which can not handle data with the size of millions efficiently.

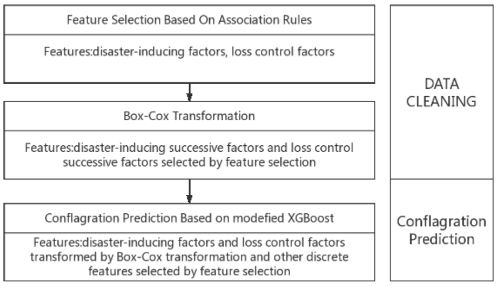

Figure 1. Overall Technical Flow Chart

A Feature Selection Based On Association Rules

Fire accident prediction association mining aims to discover the underlying associations among data features. We can obtain the minimum feature subset through selecting related features which prefer short rules as the association rules can discover the correlations among features. Therefore, we use a feature selection algorithm based on association rules. It contains two steps, including the generation of the rule set and the construction of the feature set.

The fire accident index set is denoted as $\mathrm{I} = \{i_1, i_2, \ldots, i_m\}$. The collection of the fire accident related database transactions is denoted as D. Each transaction, denoted as T, is a collection of items and $\mathrm{T} \in \mathrm{I}$. Each transaction has its identifier called TID, which denotes it uniquely. A is the collection of items. T includes A only when $\mathrm{A} \subseteq \mathrm{T}$. An association rule of fire accident related data is described as A⇒B, where $\mathrm{A} \subset \mathrm{I}, \mathrm{B} \subset \mathrm{I}, \mathrm{A} \cap \mathrm{B} = \emptyset$. The association rule A⇒B can be described by the following parameters based on the current theory of association rules.

We explain some related metrics as follows.

Support defines the possibility that A and B are in D at the same time (see Equation (1)), $P(I)$ denotes the number of elements in set I, $P(A \cup B)$ denotes the number of elements that include A and B in set I. In the context of fire accident prediction, A and B are the collections of items. It can evaluate the importance of association rules that denotes the representativeness of the rule among all rules. If the item set satisfies the minimum support (min_sup), we call it frequent item set.

$$\text{Support}(\text{A} \Rightarrow \text{B}) = \frac{P(A \cup B)}{P(I)} \quad (1)$$

In the collection of fire accident related database transactions D, confidence evaluates the percentage of B in D when A is in D(see Equation (2)). It represents the accuracy of association rules. It reflects the posterior probability of B when A is given. In the context of fire accident prediction, B gets more closely related to A when $\text{Confidence}(\text{A} \Rightarrow \text{B})$ gets bigger.

$$\text{Confidence}(\text{A} \Rightarrow \text{B}) = \frac{\sup(A \cup B)}{\sup(A)} \quad (2)$$

Lift can be called Interest. It describes the ratio that the possibility of A and B happen at the same time divided by the possibility of A and B happen at the same time while they are independent of each other(see Equation (3)). In the context of fire accident prediction, lift represents the level of the correlations between A and B and the independence level of them. If lift<<1, A and B retrain from each other. The rule can be better if lift>>1.

$$\text{Lift}(\text{A} \Rightarrow \text{B}) = \frac{\sup(A \cup B)}{\sup(A) \times \sup(B)} \quad (3)$$

The main steps of the algorithm are listed as follows.

Step 1: generate a strongly-associated rule set, discretize the feature vectors if the feature in training set is continual.

Step 2: Choose the rules whose consequence is the prediction label, then form the new rule set. Calculate the lift of each rule and remove the rules with lift<1.0. Therefore, we can get the effective rule set.

Step 3: Sort the rules in the effective rule set by shortest length, maximum lift, maximum support.

Step 4: Print the feature in the feature set and quit if cycle index is greater than the given number. Otherwise, we get the first rule in the effective rule set and add its consequent to the feature set, remove the rule from the effective rule set. We do not need to add the feature which already exists on the feature set.

Step 5: Delete samples that satisfy the rule, calculate lift and support in the remaining training set.

Step 6: Sort the rules in the effective rule set by maximum lift, maximum support, go to Step 4.

B Box-Cox Transformation

In fire accident prediction, some continual response variables are not normally distributed in the sample set. The performance of the model is not good enough if we use the data to build the model directly. Hence, we consider transforming the continual response variables by Box-Cox Transformation that can remove the dependencies on unobservable errors and the predictor variable to some extent.

For a continual response variable in the fire accident related feature space, we denote the continual response variable $(y^{(0)}, y^{(2)}, \ldots, y^{(n)})$, for any $y^{(i)}$ in it. The transformation of $y^{(i)}$ is shown in Equation (4) when $y^{(i)} > 0$:

$$y^{(i)^{(\lambda)}} = \begin{cases} \frac{y^{(i)^{(\lambda)}} - 1}{\lambda}, \lambda \neq 0; \\ log y^{(i)^{(\lambda)}}, \lambda = 0 \end{cases} \quad (4)$$

$y^{(i)}$ is the original data, the transformations vary with different λ. We use the expensive formula shown in Equation (5) when $y^{(i)}$<=0:

$$y^{(i)^{(\lambda)}} = \begin{cases} \frac{(y^{(i)^{(\lambda)}} + a)^{\lambda} - 1}{\lambda}, \lambda \neq 0; \\ \log(y^{(i)^{(\lambda)}} + a) y^{(i)^{(\lambda)}}, \lambda = 0 \end{cases} \quad (5)$$

We need to estimate λ whether we use the fundamental formula or the expensive formula. An optimal λ can transform the distribution of data is not normally into the normal one. We adopt Maximum Likelihood Estimation to estimate the parameter λ. $y^{(\lambda)} \sim N(X\beta, \sigma^2 I)$, we can get the likelihood function of β, σ^2 for the fixed λ shown in Equation (6):

$$\begin{cases} \text{L}(\beta, \sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \times e^{\left\{-\frac{1}{2\sigma^2}\left(Y^{(\lambda)} - X\beta\right)^T \left(Y^{(\lambda)} - X\beta\right)\right\}J} \\ \text{J} = \prod_{i=1}^{n} \left| \frac{d_{y_i}{}^{(\lambda)}}{d_{y_i}} \right| = \prod_{i=1}^{n} y_i{}^{(\lambda-1)} \end{cases}$$

(6)

Taking the partial derivatives of β, σ in $L(\beta, \sigma^2)$, we can get the MLE $\hat{\beta}(\lambda), \sigma^2(\lambda)$ as shown in Equation (7) and Equation (8):

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T Y^{(\lambda)} \qquad (7)$$

$$\begin{cases} \sigma^2(\lambda) = \frac{1}{n} Y^{(\lambda)^T} (I - X^T (X^T X)^{-1} X) Y^{(\lambda)} = \frac{1}{n} RSS(\lambda, Y^{(\lambda)}) \\ RSS(\lambda, Y^{(\lambda)}) = Y^{(\lambda)^T} \left(I - X^T \left(X^T X \right)^{-1} X \right) Y^{(\lambda)} \end{cases}$$

(8)

The maximum value of the likelihood function shown in Equation (9):

$$L_{max}(\lambda) = L\left(\hat{\beta}(\lambda), \sigma^2(\lambda)\right)$$

$$= (2\pi e)^{-\frac{n}{2}} J\left(\frac{RSS(\lambda, Y^{(\lambda)})}{n}\right)^{-\frac{n}{2}} \qquad (9)$$

We can confirm λ by getting the maximum value of the likelihood function. Thus, the question is transformed into getting the maximum value of $In(L_{max}(\lambda))$, and removing the constant terms that are not related to λ, as shown in Equation (10):

$$In\left(L_{max}(\lambda)\right) = -\frac{n}{2} Ln\left(RSS(\lambda, Y^{(\lambda)})\right) + lnJ$$

$$= -\frac{n}{2} RSS(\lambda, Z^{(\lambda)}) \qquad (10)$$

$Z^{(\lambda)}$ is shown in Equation (11):

$$Z^{(\lambda)} = \begin{cases} \frac{y_i^{(\lambda)}}{(\prod_{i=1}^{n} y_i)^{\frac{\lambda-1}{n}}}, \lambda \neq 0; \\ (ln y_i)(\prod_{i=1}^{n} y_i)^{\frac{1}{n}}, \lambda = 0; \end{cases} \qquad (11)$$

To get the maximum value of $InL_{max}(\lambda)$, we only need the minimum value of $RSS(\lambda, Z^{(\lambda)})$. It is not easy to get the explicit analytic expression of $RSS(\lambda, Z^{(\lambda)})$, but we can give a series of λ and get the least-squares estimation of a linear regression model, consequently, we can easily get the corresponding value of $RSS(\lambda, Z^{(\lambda)})$. In conclusion, the process of the feature selection algorithm based on Association Rules are as follows:

Step 1: Calculate $Z_i^{(\lambda)}$ for each given value of λ;

Step 2: Calculate $RSS(\lambda, Z_i^{(\lambda)})$;

Step 3: Make the curve about λ and $RSS(\lambda, Z_i^{(\lambda)})$, find out the point that represents the minimum value of $RSS(\lambda, Z_i^{(\lambda)})$;

Step 4: Use Step 3, get $Y^{(\lambda)}$.

C Fire accident Prediction Based On Modified XGBoost

In the context of fire accident prediction, for a given dataset with n examples and m features $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m)$, a tree ensemble model uses K additive functions to predict the output, as shown in Figure 2 left.

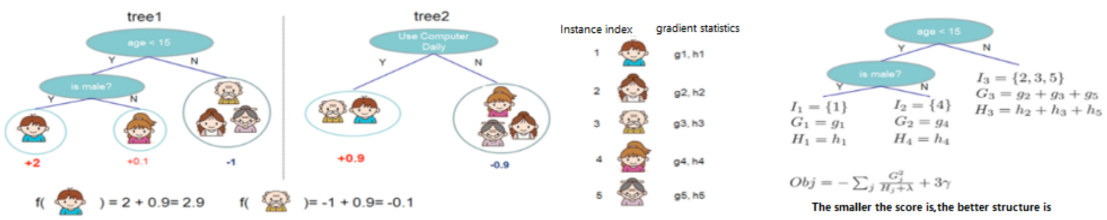

Figure 2. Left: tree ensemble model. Right: The structure score calculation of a single tree

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \qquad (12)$$

$$F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^m)$$

(13)

F is the space of regression trees (see Equation (13)), q denotes the structure of each tree. w denotes the weight vector of each leaf. For a given sample, we will use the decision rules in the trees (given by q) to partition them recursively into leaves and calculate the final prediction by summing up the scores in the corresponding leaves (given by w), the equation of final prediction value $\hat{y}_i$ is in Equation (12). To learn a tree ensemble, we optimize the following regularized objective as shown in Equation (14):

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \Omega(f)$$

$$= r^T + \lambda ||w||^2 \qquad (14)$$

The objective calculates the quality of $\hat{y}_i$ in the training set, Ω calculates the complexity of model to avoid overfitting. The model accumulates regression trees and appends new

Table I. ARCHITECTURE FEATURES

Residence	Public	Factory	Warehouse
Disaster-inducing Factors			
Energy Dissipation Mode	Content Fire Load	Content Flammability	Content Flammability
Population Density	Population Density	Quality OF Goods	Quality OF Goods
/	/	Storage Mode	Storage Mode
/	/	Production Safety Level	Warehouse Scale Level
Electrical Equipment Safety Level	Electrical Equipment Safety Level	Electrical Equipment Safety Level	Electrical Equipment Safety Level
Interior Safety Level	Interior Safety Level	Interior Safety Level	/
Building Structure	Building Structure	Building Structure	Building Structure
Building Height	Building Height	Building Height	Building Height
Age OF Building	Age OF Building	Age OF Building	Age OF Building
Loss Control Factors			
Fire Resistance Level	Fire Resistance Level	Fire Resistance Level	Fire Resistance Level
Fire Isolation Level	Fire Isolation Level	Fire Isolation Level	Fire Isolation Level
Safety Evacuation Level	Safety Evacuation Level	Safety Evacuation Level	Safety Evacuation Level
/	Explosion-proof Level	Explosion-proof Level	Explosion-proof Level
Monitoring Level	Automatic Alarm AND Linkage Level	Automatic Alarm AND Linkage Level	Automatic Alarm AND Linkage Level
Hydrant Equipment Level	Hydrant Equipment Level	Hydrant Equipment Level	Hydrant Equipment Level
/	Ventilating Unit Level	Ventilating Unit Level	Ventilating Unit Level
Fire Extinguisher Equipment Level	Fire Extinguisher Equipment Level	Fire Extinguisher Equipment Level	Fire Extinguisher Equipment Level

optimized object in each cycle. The score calculation of a newly added trees given in Equation (15).

$$\begin{cases} \widetilde{L^{(t)}}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i+\lambda} + r^T \\ g_i = \partial_{\widehat{y_i}^{(t-1)}} l\left(\widehat{y_i}, y_i^{(t-1)}\right) \\ h_i = \partial_{\widehat{y_i}^{(t-1)}}^{2} l\left(\widehat{y_i}, y_i^{(t-1)}\right) \end{cases} \quad (15)$$

The structure scores calculation of a newly added tree evaluates the decision tree, as shown in the right part of Figure 2. We only need all the samples in leaves, Equation 15 calculates the score of a tree, the algorithm would iteratively search over the possible split candidates and find the best split to add to the tree until a maximum depth is reached, as shown in Figure 3. As the algorithm would traverse the dataset, it would consume plenty of time while handling massive data. We use an approximate algorithm in which we can get potential better split points, and get the cumulative histogram by estimating the split points.

Input: I, instance set of current node
Input: $I_k = \{i \in I | x_{ik} \neq \text{missing}\}$
Input: d, feature dimension
$gain \leftarrow 0$
$G \leftarrow \sum_{i\in I} g_i, H \leftarrow \sum_{i\in I} h_i$
for $k = 1$ ***to*** m ***in parallel*** **do**
  $G_L \leftarrow 0,\ H_L \leftarrow 0$
  for *j in sorted(*I_k*, ascent order by* $\mathbf{x}_{jk}$*)* **do**
    $G_L \leftarrow G_L + g_j,\ H_L \leftarrow H_L + h_j$
    $G_R \leftarrow G - G_L,\ H_R \leftarrow H - H_L$
    $gain \leftarrow \max(gain, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$
  end
end
Output: Split and default direction with max gain

Figure 3. Parallel Tree Split Finding Algorithm on Single Machine

III. IMPLEMENTATION OF FIRE ACCIDENT PREDICTION MODEL

A Data Preparation

In this paper, we collect the internal features of architectures, the weather conditions and the causes of the fire from 2011 to 2017 in China. The internal features of architectures is shown in Table

Table II. DATA AFTER TRANSFORMATION

ID	r_weat	r_temp	r_wind	r_EDM	r_PD	r_EESL		r_ISL	r_BS
0	5	4	2	0.4	2	1		1	2
1	0	2	1	0.05	2	2		4	1
2	3	3	2	0.4	2	2		1	2
3	3	2	3	0.03	3	2		4	2
4	5	2	0	0.5	3	2		4	1
…	…	…	…	…	…	…		…	…
3321231	3	3	2	0.80061	3	2		1	3
ID	**r_BH**	**r_AOB**	**r_FRL**	**r_FIL**	**r_SEL**	**r_ML**	**r_HEL**	**r_FEEL**	**Label**
0	0	3	2	0	2	11.856	1.1127	9.821004	3
1	2	1	2	0	0	12.282	1.1127	11.7669	2
2	0	0	1	0	1	10.806	1.1127	7.871179	2
…	…	…	…	…	…	…	…	…	…
3321231	2	2	3	3	2		12	7.871179	0

I. The weather features include Weather (shower, split, moderate rain, heavy rain, cloudy, clear, snow), Temperature, Wind (discretized to 1-5) and Humidity. The prediction label denotes the causes of fire disaster with 0.Normal 1.Natural Cause 2.Electric Cause 3. Combustible Cause. It is not easy to get related data about all resources of fire occurrence factors. What's more, fire occurrence factors differ from the research target. For all the reasons above, we use the residence type of architecture as the experimental data. Adopting feature selection algorithm based on Association Rules, we get the feature space of fire accident prediction model that includes weather features and residence features.

Table III. DESCRIPTIONS OF PARAMETERS IN XGBOOST

PARAMETER CATEGORY	PARAMETER	DESCRIPTION
General	***booster***	Select the type of model to run at each iteration: gbtree, gblinear.
	silent	Silent mode is activated is set to 1.
	nthread	Used for parallel processing, set the number of cores.
Booster	***eta***	Learning rate.
	min_child_weight	Define the minimum sum of weights of all observations required in a child.
	max_depth	The maximum depth of a tree.
	gamma	Specify the minimum loss reduction required to make a split.
	subsample	Denotes the fraction of observations to be randomly sampled for each tree.
	colsample_bytree	Denotes the fraction of columns to be randomly sampled for each tree.
	lambda	L2 regularization term on weights.
	alpha	L1 regularization term on weights.
	scale_pos_weight	Be used in case of high-class imbalance as it helps in faster convergence.
Learning Task	***objective***	Define the loss function to be minimized.
	eval_metric	Be used for validation data.

B Nonlinear Transformation Analysis

The residence fire accident prediction model uses the detailed data onto residence from 2011 to 2017 in China on the basis of feature selection algorithm based on association rules, the original data is marked as follows. weather—r_weat, Temperature—r_temp, wind—r_wind, purpose—r_use, energy dissipation mode—r_EDM, population density—r_PD, electrical equipment safety level—r_EESL, interior safety level—r_ISL, building structure—r_BS, building height—r_BH, age of building—r_AOB, fire resistance level—r_FRL, fire isolation level—r_FIL, safety evacuation level—r_SEL, monitoring level—r_ML, hydrant equipment level—r_HEL, fire extinguisher equipment level—r_FEEL.

In the residence fire accident prediction model, there are some continual response variables which are not normally distributed in the sample set. The result of model prediction would not be good enough if we use the data to build the model directly. Hence, we consider transforming the continual response variables by Box-Cox Transformation that can remove the dependencies on unobservable errors and the predictor variable to some extent. The variables after transformation are denoted as bc_r_temp, bc_r_popu, bc_r_heig, and bc_r_age, respectively. For getting the similar data comparing to the original data after transformation, we tune λ many times.

First, we transform the original data for better performance. Then we do the correlation analysis and regression analysis. We use expensive formula based on Box-Cox for the portion of data whose values are 0. For bc_r_temp $a_1 = 30$, for bc_r_heig $a_2 = 1$, for bc_r_age $a_3 = 1$, we get the optimized values of λ for r_temp, r_heig, r_age are 4.51039571649, -0.775774467944, 0.988957448167. The distribution of r_popu is not ideal after transformation so we keep it unchanged. Maximum likelihood method of r_temp to confirm λ would make the series of data much bigger than other data, thus we enumerate λ to get the appropriate value and the value is 0.5.

C Experimental Results

We train the model based on XGBoost with clean data. In this paper, we use the toolkit in the XGBoost python module[5]. There are three types of parameters in the XGBoost python module. 1. General Parameters, which guide the overall functioning. 2. Booster Parameters, which guide the individual booster (tree/regression) at each step. 3. Learning Task Parameters, which guide the optimization process. The description of each parameter in the XGBoost python module are shown in Table III

Get the optimized parameters by adding parameter tuning in each iteration, loss value decreases along with the increase of iteration, the accuracy of model would be better. Thus we can add the iteration to get optimized parameters.We tune (max_depth, min_child_weight), (gamma), (subsample, colsample_bytree), (reg_alpha), (learning rate) by order. Finally, we can apply these parameters in the model, the report about fire accident prediction model is shown in Figure 4. In Figure 4, we can know prediction accuracy is 0.9 based on cross validation, the mean of error rates is 0.1000000 and the standard deviation of error rates is 0.044721. The screenshot of fire accident prediction system is shown in Figure 5.

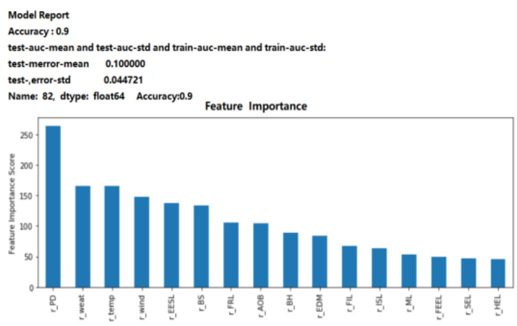

Figure 4. Report about Model

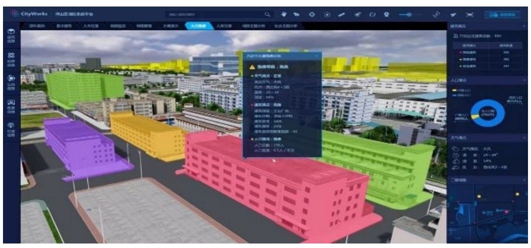

Figure 5. Screenshot of Fire accident prediction System

IV. CONCLUSION

In this paper, we propose a new fire accident prediction method based on XGBoost. We use feature selection techniques based on association rules that can remove unnecessary features, to reduce feature space and improve the efficiency. The method transforms continual response variables that are not normal distribution in the sample set based on Box-Cox to reduce the dependencies on unobservable errors and the predictor variable to some extent. In the fire accident prediction system, neural networks or other boosting algorithms based on decision tree are time-consuming since the sample data size is ten million. Aiming to handle massive data, XGBoost provides cache aware pre-read technology, distributed memory computing technology, All Reduce fault-tolerant tools for improving the calculation speed of boosting trees. The accuracy of the prediction method based on XGBoost after parameter tuning is satisfactory. Then we apply the model to fire accident prediction system and hope it would offer help to fire prevention, as well, fire prevention is the important part in the constructions of the smart city that fire prevention moves towards intelligent.

In our future work, we plan to apply multiple data processing models for deep data cleaning. Through analyzing the correlation between fire-related data and fire accident-caused reasons, we aim to extract more information features in the time dimension and add the characteristics into fire accident prediction to improve the prediction accuracy.

ACKNOWLEDGEMENTS

This paper is supported by The Fundamental Research Funds for the Central Universities (No. 30916015104), National key research and development program: key projects of international scientific and technological innovation cooperation between governments No. 2016YFE0108000), CERNET next generation internet technology innovation project (NGII20160122), The Project of ZTE Cooperation Research(2016ZTE04-11), Jiangsu province key research and development program: Social development project (SBE2017741114), Jiangsu province key research and development program: Industry outlook and common key technology projects (SBE2017030519).

REFERENCES

[1] Sharma D, Singh K, Aggarwal S. Implementation of Artificial Neural Fuzzy Inference System in a Real Time Fire Detection Mechanism. International Journal of Computer Applications, 2016, 146(10).

[2] Ooi C H, Chetty M, Teng S W. Differential prioritization in feature selection and classifier aggregation for multiclass microarray datasets. Data Mining and Knowledge Discovery, 2007, 14(3): 329-366.

[3] Wang X, Wotton B M, Cantin A S, et al. cffdrs: an R package for the Canadian Forest Fire Danger Rating System. Ecological Processes, 2017, 6(1):5

[4] Taylor S W, Alexander M E. Field guide to the Canadian Forest Fire Behavior Prediction (FBP) System. (BINDER)[M]. 2016.

[5] XGBoost Python Package, https://pypi.python.org/pypi/xgboost, 2017/7/16

[6] Gao D, Lin H, Jiang A, et al. A forest fire accident prediction system based on rechargeable wireless sensor networks// IEEE International Conference on Network Infrastructure and Digital Content. IEEE, 2015:405-408.

[7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016:785-794.

[8] Bicego M, Baldo S. Properties of the Box-Cox Transformation for Pattern Classification. Neurocomputing, 2016, 218:390-400.

[9] Stojanova D, Kobler A, Ogrinc P, et al. Estimating the risk of fire outbreaks in the natural environment. Data Mining and Knowledge Discovery, 2012, 24(2): 411-442

[10] Bandyopadhyay S, Wolfson J, Vock D M, et al. Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. arXiv preprint arXiv:1404.2189, 2014.

[11] Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering, 2014, 40(1): 16-28.

[12] Ooi C H, Chetty M, Teng S W. Differential prioritization in feature selection and classifier aggregation for multiclass microarray datasets. Data Mining and Knowledge Discovery, 2007, 14(3): 329-366.

[13] Alamgir N, Boles W, Chandran V. A Model Integrating Conflagration prediction and Detection for Rural-Urban Interface// International Conference on Digital Image Computing: Techniques and Applications. IEEE, 2016:1-8.

[14] Konda P, Kumar A, Ré C, et al. Feature selection in enterprise analytics: a demonstration using an R-based data analytics system. Vldb Demo, 2013, 6(12):1306-1309.

[15] Pibiri G E, Venturini R. Clustered Elias-Fano Indexes. ACM Transactions on Information Systems (TOIS), 2017, 36(1): 2.

[16] Alamgir N, Boles W, Chandran V. A Model Integrating Fire accident prediction and Detection for Rural-Urban Interface// International Conference on Digital Image Computing: Techniques and Applications. IEEE, 2016:1-8.