Design and Implementation of an Early Warning System Based on the Risk Measurement Model

Su FENG 1, Jing LI 1, Huan WANG 1, Qi TANG 1, Liqiong LIU 2*, Fengnian YIN 3

1Information Center of the State Administration for Market Regulation, Beijing 100088, China
2Nanning Customs district P.R. China, Nanning, Guangxi 530029, China
3Manzhouli Customs district P.R. China, Manzhouli, Inner Mongolia 021400, China

* Corresponding author’s e-mail: nnhg_liuliqiong@customs.gov.cn

Abstract. In recent years, there are still some defects in the risk management of China's inspection and quarantine work, such as the omission of major factors in the conformity assessment of inbound and outbound commodities and the subjective and unscientific assignment of risk indicators, which affect the ability of risk monitoring. Therefore, this paper constructs an inspection and quarantine risk measurement model based on random forest algorithm, and develops a risk early warning system for inspection and quarantine business. Firstly, according to the target requirements, the training data set is extracted from the original rough data set, and the overall data is cleaned and modified. Secondly, the integrated machine learning model is used to select the feature values with better prediction ability. Then, the risk measurement model is constructed based on random forest algorithm and deployed in multi-model parallel mode. Finally, a complete risk early warning system is developed based on the risk prediction model. Through the operation of this system, the abilities of risk analysis and discrimination for inspection and quarantine have been greatly improved, and the comprehensiveness and accuracy of risk prevention and control have been effectively guaranteed.

1. Introduction
In recent years, with the emergence of a large number of various data and the rapid decline of the cost of computing resources, the information society has begun to enter the era of big data. At present, the inspection and quarantine work of China's customs is mainly aimed at "Guaranteeing public health safety of ports, protecting imported and exported biological safety, commodities quality and food safety". The focal point is to improve the ability of risk management, further rationally allocate big data resources, and build an efficient customs clearance work system [1-3].

The inspection and quarantine supervision department of customs in China has actively explored the risk assessment, classified management and other aspects of all kinds of import and export enterprises, and many results have been achieved. Li et al. [4], analyzed the demands in the field of inspection, quarantine and supervision, and proposed to apply artificial intelligence technology to the relevant supervision field, and solved the lack of supervision of human resources by taking advantage of artificial intelligence technology. Wei et al. [5], made a detailed discussion and analysis on how to optimize credit supervision services, and put forward specific suggestions on how to improve the inspection and quarantine supervision system and strengthen the construction of credit management environment, which played a positive role in promoting the innovation of relevant systems. Based on blockchain
technology. Chen et al. [6], carried out active exploration and research on how to innovate the inspection and quarantine supervision mode. They pointed out that the emergence of blockchain technology may become an advantageous solution to improve cross-border e-commerce regulation. In this paper, they made detailed analysis of the existing problems in the current regulatory issues and how to apply blockchain technology, providing valuable solutions for facilitating more effective regulatory services. Li et al. [7], studied the influence of the rise of cross-border e-commerce on regulation and proposed corresponding countermeasures. They proposed that internal pressure of supervision works can be relieved by optimizing laws and regulations on cross-border e-commerce. Although the ability of risk management on China’s inspection and quarantine work has been improved in somehow in recent years, there are still some defects, such as the omission of major factors in the conformity assessment of import and export commodities, subjective and unscientific risk index assignment, which affect the ability of risk monitoring.

In order to solve the problems existing in the current inspection and quarantine supervision, we built a risk measurement model based on random forest [8-11], and realized a complete set of early warning system on this basis. Our target is to calculate the risk factor weight by using big data technologies, and use historical data to train the model. Besides, taking the conformity assessment of risk assessment model as the core to build a risk early warning system which is suitable for the inspection and quarantine business demands. In addition, it can be used to forecast the risk probability of disqualification about import and export commodities in each batch, and provide scientific, accurate and timely auxiliary support for its risk analysis.

2. Building the Risk Measurement Model
Based on the principle of analyzing specific problems, in accordance with some relevant standard specifications in the industry and mature data mining engineering methodology, combined with the specific business of the customs and the original inspection and quarantine department, we carried out robust and effective works on model research and construction. The specific research flow is shown in Figure 1:

![Figure 1. The construction flow chart of risk measurement model for inspection, quarantine and supervision.](image)

2.1. Data Acquisition, Exploration and Processing
The important part of the development of risk measurement model is the analysis and exploration of data, including outlier detection, data logic verification, missing data analysis, correlation analysis between variables and the exploration of derived variables.

Outlier detection and data logic verification can be used to exclude data distortion or data errors that generated in the stored procedure of business data. According to the results from data check procedure, we can conclude the reasons which lead to abnormal data, and then delete or repair corresponding data.
It can avoid abnormal data do harm to the performance of whole model. Missing data verification can exclude variable which has too many missing data, since such variable will decrease the effective information of variables and cannot be regarded as backup variables. The prediction indexes used in the model cannot have very strong relationship with each other, as a result, the correlation verification is required between classified variable and classified variable, continuous variable and continuous variable, continuous variable and classified variable. Furthermore, the variables which have very strong correlation with other variables will be filtered or deleted.

Generally speaking, it is a common problem that most datasets have some missing data, and there are tow frequently used methods. If the number of missing data is relative small, we can use constant, average, mode or other values to fill missing data. If the number of missing data is relative large, we can use random forest or other methods to fill missing data. When there are a small number of samples, if we use random forest method to fill missing data, it can make model better trained with samples and further enhance the generalization and prediction ability of the model.

In order to dig more useful information among variables and relationship between different variables, in the variables exploration process, derived variables are usually utilized to enrich the X variable set of modeling and improve the performance of the model.

2.2 Selecting Feature Engineering

In the feature engineering stage, we do not use traditional correlation test or Chi-square test, instead, we adopt Gradient Boosting Decision Tree (GBDT) which has been widely used in current integrated machine learning model. GBDT is a boosting model which can be used to solve regression problems [12-14], the regression tree and gradient boosting are regarded as classifier and learning algorithms respectively. By using continuous iteration to generate decision tree, and taking GBDT as feature selector can choose the optimized representative feature, which makes feature has strong prediction ability.

The general Boosting Tree algorithm takes decision tree as the basic model, uses forward stepwise algorithm to generate new data iteratively, and finally uses different loss functions to combine decisions, such as the square error loss function commonly used in regression problems and the exponential loss function commonly used in classification problems. In each iteration, the residual data (the difference between the predicted value and the real value) is fitted according to the current model, which is used as new data to generate a new tree. GBDT uses the negative gradient value of the loss function as the approximate value of the residual to fit a regression tree, so that the optimization of each step becomes simple. It is equivalent to that the tree model built each time is in the gradient descent direction of the loss function based on the previous decision tree, and the decreasing part of each model in the gradient direction is regarded as a weak model. GBDT uses the tree complexity as a regularization term in the loss function. The tree complexity is defined as the sum of the square of the number of tree leaf nodes T and the weight of leaf nodes. The loss function is:

$$Obj^{(t)} = \sum_{i=1}^{n} \left( y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)) \right)^2 + \Omega(f_t) + \text{const}$$

$$= \sum_{i=1}^{n} \left[ 2(\hat{y}_i^{(t-1)} - y_i)f_t(x_i) + f_t(x_i)^2 \right] + \Omega(f_t) + \text{const}$$

(1)

Where \( (f_t) \) stands for decision tree, \( \Omega(f_t) \) stands for penalty term. To avoid model become overfitting, that is:

$$f' \propto x' = \sum_{\alpha} T \times \theta_{\alpha}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2$$

(2)

(3)
The detailed steps about Gradient Tree Boosting Algorithm are illustrated in Table 1.

| Algorithm 1 Gradient Tree Boosting Algorithm |
|---------------------------------------------|
| 1. Initializing \( f_0(x) = \arg \min \gamma \sum_{i=1}^{n} L(y_i, \gamma) \). |
| 2. From \( m = 1 \) to \( m = M \) do: |
| a) For each \( i = 1, 2, \ldots, N \) calculates \( \gamma_{im} = -\left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}. \) |
| b) Making regression tree fit to target \( \gamma_{im} \) and the leaf node range is \( R_{jm} \), \( j = 1, 2, \ldots, J_m \). |
| c) For \( j = 1, 2, \ldots, J_m \) calculates \( \gamma_{jm} = \arg \min \gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \). |
| d) Updating Regression Tree \( f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}). \) |
| 3. output \( \hat{f}(x) = f_M(x) \). |

2.3. Constructing Risk Model

After determining the target variable \( Y \) and the multi-dimensional explanatory variable \( X \), we can enter the process of model construction. In this paper, the random forest algorithm is selected to build the training model. Because of its fast training speed and the ability to carry out parallel operations, it is widely used as the basic algorithm to build the model. Random forest algorithm can also randomly select different features and does not need to carry out data normalization operation. As a result, it is superior in processing multi-dimensional and multi-feature data, and greatly improve the efficiency of model training. The specific algorithm flow is shown below.

- Randomly selecting \( n \) data from the original training dataset to form input training subset, where \( n \) is far less than the entire training data size \( N \). Since this operation can guarantee that a part of data (out-of-bag data) will never be selected as training subset and they can be used directly to test errors without the need for a separate test set or verification set.
- After the input training data has been selected, the next step is to construct decision tree. For each data node, selecting \( m \) features from the entire feature set \( M \), where \( m \) is far less than \( M \).
- When constructing each single decision tree, the feature with the smallest Gini coefficient is used as the split point to construct the decision tree. Other nodes also followed such rule to determine split point. The end of splitting process means that the decision tree reaches to the maximum depth or all the training samples of the node belong to the same category.
- Executing step 2 and step 3 repeatedly. For each input data, the system will generate a new corresponding decision tree, that is, another random forest which can be used to make decision for predicted data.
- Multiple decision trees will be processed simultaneously to make predicted decision for prepared data. And finally predicted decision is determined by majority voting method.

2.4. Model Evaluation and Deployment

After developing the whole model, some statistical indexes will be taken advantages to evaluate the model accuracy. In terms of the binary classification problem studied in this paper, the commonly used model evaluation indexes are Precision and Recall, F1-score [15], ROC (receiver operating
characteristic curve), AUC (Area Under Curve) and so on [16]. Precision and recall are calculated by constructing Confusion Matrix, and F1-score is further calculated to carry out risk model evaluation.

2.4.1. Confusion Matrix.
At first, generating the confusion matrix on the basis of actual and predicted values, as illustrated in Table 2.

| Actual Values | Predicted Values |
|---------------|------------------|
|               | Positive         | Negative        |
| Positive      | TP (True Positive) | FP (False Positive) |
| Negative      | FN (False Negative) | TN (True Negative) |

2.4.2. Precision.
P = \frac{TP}{TP+FP}, is the number of true positives divided by the total number of values labeled as belonging to the positive class.

2.4.3. Recall.
R = \frac{TP}{TP+FN}, is defined as the number of true positives divided by the total number of values that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

2.4.4. F1-score.
Because precision and recall are a pair of contradictory indexes, that is, the increase of one index will lead to the decrease of another one. As a result, the industry uses F1-score to balance the contribution of precision and recall, and it can be interpreted as a weighted average of the precision and recall. The way to calculate F1-score is illustrated as followed:

\[ F1 = \frac{2 \frac{P}{P + R} \frac{R}{P + R}}{\frac{2}{P} + \frac{2}{R}} = \frac{2PR}{P + R} \tag{4} \]

Making the training model to be persisted and running server to load and deploy model. Besides, providing interfaces in REST format or other formats to display predicted results or predicted reports and prepare to be used by other applications.

2.5. Model Monitoring and Reports
In addition, multiple models run in parallel and simultaneously in the model deployment process. By comparing and monitoring the results of the active and standby models, abnormal phenomena and issues can be found as early as possible.

The running state of model will be monitored periodically and corresponding reports including AUC, KS (Kolmogorov-Smirnov) and other core evaluation indexes will be generated to demonstrate the functionality and variation of stability. According the real-time reports, it can correct the rules, iteratively update and continuously improve the performance of the model. In some necessary cases, some measures including model modification or reconstruction should be taken action.
3. The Architecture Design of Early Warning System

3.1. Overall Architecture Design

The overall architecture design of risk early warning system based on the inspection and quarantine business risk measurement model is shown on Figure 2.

**Figure 2. Overall Architecture Design**

Risk early warning system is made up by six parts, including big data center, risk measurement subject database, feature (risk factor) index sets, model construction, model application and model monitoring. Each part has their own functionality, and detailed descriptions are as followed:

3.1.1. Big Data Center.

Using the existing big data resources pool built by inspection, quarantine and supervision department to extract relevant business data for the construction of risk measurement model.

3.1.2. Risk Measurement Subject Database.

Collect, manage and process the original data, and finally form the subject database of risk measurement.
3.1.3. Feature (risk factor) Index Set.
Based on subject database and according to the business rules in enterprise, commodity and customs areas, extracting corresponding features from dataset. In addition, using some statistical methods on extracted feature to form the business indexes and related statistical indexes.

3.1.4. Model Construction.
The model is constructed by strictly following the explicit requirements of model training. After the construction process is completed, specific tasks such as feature selection, model training and testing are carried out.

3.1.5. Model Application.
After training process finished, the risk measurement model will analyze new business data at first and then generate corresponding analysis results including risk early warning and risk prediction. As a result, these analysis results can be applied to some actual operations in real life, such as querying or application.

3.1.6. Model Monitoring.
Monitoring the state of the model can help administrators to control the whole system. When some problems occur, real-time monitoring can make problem be located as soon as possible, such as some features or client changes will lead to some problems. It can help administrators quickly decide whether to rebuild the model or simply optimize existing model.

3.2. Model Architecture Design
The implementation architecture design for the risk model is shown in flow chart Figure 3.

As illustrated in Figure 3, the first step is to collect data. In this paper, we propose that the required data for model construction comes from big data center, after these data been integrated and normalized, they will be divided into training data and test data. The training data is used to train the model, no test data will be existed in the training process. And test data will be utilized in the test process to verify the model accuracy.
The next important step is feature extraction. By using feature engineering algorithm of data mining, all impact factors are digitized and principal component analysis is carried out in the same spatial dimension, the factors and coefficients are sorted. Meanwhile, data noise reduction and pruning are processed to establish the pure mathematical model of risk factors. Then, through research data, reference materials, business understanding and other methods to identify the dimension of impact factors and analyze the business in a interpretable manner. In the end, choosing and analyzing business risk features.

After acquiring the expected features, the next step is to construct model. In this paper, random forest algorithm is taken as the main framework for the model. Firstly, using acquired training data to train the constructed model until a risk prediction model with optimal parameters is obtained. And then, evaluating the performance and accuracy of the model to determine whether it is a good model or not. If it can achieve the predefined evaluation index, it will be released and used in the official environment after passing the approval process. In addition, if satisfactory experimental results can be obtained, an early warning system based on the risk measurement model of inspection and quarantine can be constructed. If the expected evaluation index is not achieved, the model will be trained again.

4. The Realization of Early Warning System Functions
The purpose of training a model with optimal parameters is to build the system. We have successfully trained an inspection and quarantine risk measurement model with optimal parameters, and developed a complete inspection and quarantine risk early warning system based on the model. The following is the partial demonstration of the risk early warning system.

4.1. Real-time Monitoring
In the inspection and quarantine business, the quantity of declaration, quantity of inspection, quantity of seizure, rate of inspection, rate of seizure, and rate of inspection and seizure in entry-exit regions shall be monitored in real time, all results will be displayed in corresponding charts. The system can remind the high-risk enterprises and commodities periodically and repeatedly, which is very intelligent and safe. The details are shown in Figure 4.

![Early Warning Diagram of High-risk Enterprises and Commodities](image-url)
4.2. Model Monitoring
Risk early warning system keeps real-time monitoring on the running state of the model. The prediction ability is measured by confusion matrix, various ratio indexes, ROC graph, the comparison between predicted values and actual values. And the final results for different measurements mentioned before will be visualized, as shown in Figure 5.

![Figure 5. Real-time Monitoring of the Early Warning System](image)

4.3. Risk Prediction
Making prediction for the newly declared inspection form, including the predicted values of the failed rate of inspection and quarantine. The results will be demonstrated in the form of chart. This function provides a more intuitive visualization, as shown in Figure 6.

![Figure 6. Predicted Results of Failed Rate](image)

5. Conclusion
In this paper, we build a risk measurement model for inspection and quarantine qualification assessment based on random forest algorithm. In addition, a complete early warning system is realized on this model. This risk early warning system is based on big data center, by collecting relevant data from big data center and process these datasets to fit the algorithm and train the model. After implementing feature
selection, model training, model evaluation, model release and other processes, a complete risk measurement model is constructed. This model is made up by two parts including algorithm model and data-based rules. These two parts are balanced and fused by weight or rule-based methods. The system will process risk analysis and generate corresponding risk early warning and prediction once there have some new business data. And then, user can query or apply to the risk analysis results, which provides auxiliary support for risk assessment and improves the ability of risk prevention and control. According to practical experiments in real life, it can provide early warning services to high risk enterprises and commodities by running the risk measurement model based on the inspection and quarantine business risk early warning system. Furthermore, the detection rate of failed commodities have been greatly improved. Meanwhile, it significantly improves the performance and accuracy of the risk analysis and prevention ability on inspection and quarantine business. In the future, we will further optimize the decision-making model by integrating multiple classifiers. At the same time, we will further enhance the trend analysis function on the basis of the existing risk warning and risk prediction of the system.

Acknowledgements
This research was funded by the Science and technology research project of Special Program for Technical Support of the State Administration for Market Regulation (2020YJ037).

References
[1] Fei G P.（2020） Further expanding the development and management of special Customs Supervision areas in the context of opening up -- A Case Study of Anhui Province [J]. Heilongjiang Finance., (08): 75-78.
[2] Wen R.（2020） The Model Design and Policy Research of Customs Supervision of Service Trade in New Area of Shanghai Pilot Free Trade Zone [J]. Scientific Development., (08): 40-49.
[3] Liu Y.（2020） The Significance, Impact and Suggestion about New Policy and Supervision on Customs Cross-border B2B E-commerce Export [J]. Computer & Network., 46(14): 9-11.
[4] Li J, Xing L, Zhao H D.（2020） Analysis of the Application of Artificial Intelligence in Customs Supervision [J]. Technology and Economic Guide., 28(28): 22-21.
[5] Wei Y X.（2020） How to optimize credit supervision service? [J]. China Customs,(09): 84-86.
[6] Chen Y B, Zhou X.（2020） Research on Customs Control Innovation of China-Russia Cross-Border E-Commerce Based on Block Chain [J]. Foreign Economic Relations & Trade, (09): 37-41.
[7] Li J.（2020） On the Impact of Cross-border E-commerce on China’s International Trade and Countermeasures [J]. Journal of Shanxi Institute of Economic Management, 28(03): 43-46.
[8] OuYang Y Z, Zeng Y T, Guo W Q, et al.（2020） Mass Spectrometric Discrimination of Human Lung Tumors under Ambient Conditions Based on Random Forest Algorithm [J]. Chinese Journal of Analytical Chemistry, 48(08): 1012-1022.
[9] Xie K, Rong Y T, Hu F P, et al.（2019） Random Forest Algorithm Based on Data Integration[J/OL]. Computer Engineering: 1-15 [2019-12-09]. https://doi.org/10.19678/j.issn.1000-3428.0055891.
[10] Su M Y, Liu H S, Lin H X, et al.（2020） Machine-Learning Model for Predicting the Rate Constant of ProteinLigand Dissociation [J]. Acta Physico-Chimica Sinica, 36(01): 179-187.
[11] Xu Z Y, Kang Y, Cao Y,（2019） et al. Man-machine verification of mouse trajectory based on the random forest model [J]. Frontiers of Information Technology & Electronic Engineering, 20(07): 925-930.
[12] Duan D G, Gai X X, Han Z M, et al.（2018） Micro-blog misinformation detection based on gradient boost decision tree [J]. Journal of Computer Applications, 38(02): 410-414.
[13] Ke G L.（2016） Methodology Research on Parallel Algorithm of Gradient Boosting Decision Tree [D]. Xiamen: Xiamen University.
[14] Liu H F. (2020) Research of Optimizing Gradient Boosting Decision Tree Performance [D]. Guang Zhou: Guangdong University of Foreign Studies.

[15] Zhang H Y, Xie Y M, Yuan Z X, et al. (2016) A Method of CHI-square Feature Selection Based on Probability [J]. Computer Engineering, (08): 194-198.

[16] Wang J H, Zhao B J. (2020) Research on Feature Selection Algorithm Based on Unbalanced Data [J/OL]. Computer Engineering :1-9 [2020-10-29]. https://kns.cnki.net/kcms/detail/31.1289.TP.20201028.1745.001.html.