ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

# Transmission Engineering Cost Prediction Based on Data Mining

**Jianqing Li[1*], Chenchen Wang[2], Liming Chen[3]**

*[1]State Grid Anhui Electric Power Co., Ltd. Economic Research Institute, Anhui, China*
*[2] China Energy Engineering Group Anhui Electric Power Design Institute Co., Ltd., Anhui, China*
*[3] State Grid Anhui Electric Power Co., Ltd., Anhui, China*
*Li_Jian_QQing@163.com*
**Corresponding author*

***Abstract:*** *Transmission engineering is an important part of electric power infrastructure engineering, accurate cost prediction can effectively control the project budget and promote the fine management of electric power enterprises. However, the influence factors of transmission engineering cost are complex, and the traditional prediction method cannot meet the requirements of network refinement development. In this paper, the importance of random forest features is used to extract the key influencing factors, and the grid search method is used to select the optimal random forest parameters and train the random forest model. Finally, the random forest model is tested by actual engineering data and compared with the prediction results of the supporting vector machine model. The results show that the relative error of the stochastic forest model is low, which provides a certain reference value for the cost budget of transmission engineering.*

***Keywords:*** *Transmission Line, Engineering Cost, Data Mining, Random Forest*

## 1. Introduction

With the continuous growth of power load demand, the scale of power investment continues to expand, and the cost of power grid engineering continues to rise [1]. Power grid companies are facing the rapid growth of construction cycle pressure and capital pressure, power grid investment management mode needs to gradually from extensive management to fine management transition. As an important power supply carrier in China, the power transmission project has witnessed the expansion of capital investment and construction scale year by year, which directly affects the operation level and benefit of power grid enterprises. Therefore, the accurate prediction of transmission engineering is of great significance to improve the fine management level of electric power enterprises [2].

The traditional cost prediction of transmission engineering is usually based on the quota index system and the experience of front-line workers. However, the cost of transmission engineering is affected by social and natural factors and has the characteristics of trans-regional, large scale, long construction period, and high total investment. Therefore, the traditional cost prediction has been unable to meet the requirements of the precision of transmission engineering cost prediction under the background of refinement [3]. As an important part of the prediction model, the key factors are the basis of accurate prediction. At present, related scholars mainly analyze the influencing factors of transmission engineering cost from two perspectives: the characteristics of engineering itself and the external environment of engineering. From the analysis of the characteristics of the project itself, Wang Jiao [4] screened the total capacity of the main transformer, reactive power compensation capacity, the number of high-voltage side outlet loops, and the capacity of high-reactance device as the main influencing factors of the project cost through T-test and sensitivity analysis. Hao Yaogang [5] analyzed the path, soil quality, terrain, transport distance, compensation, and materials of the transmission line, and proposed targeted dynamic cost control measures for the whole process. From the analysis of the external environment of the project, Lu Yanchao et al. [6] divided the influence of the external environment of the power transmission and transformation project into four aspects including policy factors, economic factors, social and natural factors, and technical factors by using PEST method and comprehensively analyzed the influence degree and change trend of each influencing factor based on historical data.

With the development of artificial intelligence, the prediction method is improved compared with

ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

the traditional cost prediction, and machine learning algorithm is gradually applied to the cost prediction of transmission engineering in power grid enterprises. Peng Guangjin et al. [7] applied correlation analysis, cluster analysis, and support vector machine theory, and other data mining techniques to study the construction cost estimation of power transmission engineering, and proposed a new prediction model for transmission line cost estimation. Wang Jiao et al. [8] proposed a method based on Grey Relation Analysis (GRA) and Particle Swarm Optimization (PSO). The combined prediction model of Support Vector Regression (SVR) of PSO is used to predict the project cost, which improves the prediction accuracy again. Ling Yunpeng et al. [9] proposed a method of transmission line project cost prediction based on Back Propagation artificial neural network (BP) neural network and established a three-layer BP neural network prediction model. The results show that the model can accurately and quickly predict the transmission line project cost. Xu Li et al. [10] constructed a cost prediction model based on factor analysis and BP neural network, providing a new idea and implementation method for management optimization of UHV engineering in the whole life cycle. Geng Pengyun et al. [11] established the cost prediction model of power transmission and transformation project based on data mining technology and applied the support vector machine model to predict the cost of power transmission and transformation project, which provided a certain reference value for the cost budget.

Although at present, domestic researchers have used the neural network represented by BP, support vector regression machine and random forest, and other prediction models, but each model in the actual project cost prediction has shortcomings. BP neural network requires a large amount of training sample data, resulting in a long training time of network model, and prone to local optimization problems. Although SVR can solve practical problems such as small sample size, high dimension, nonlinearity, and local optimization well, given the particularity of transmission engineering prediction, when using support vector regression machine to model cost prediction, there is blindness in model parameter setting, which leads to large prediction error. The important parameters of the random forest model, such as the number of decision trees and candidate feature attributes, have a great influence on the model performance, and the randomness of random forest algorithm sampling samples and selecting feature attributes will lead to certain fluctuations in the error value.

To sum up, this paper proposes a stochastic forest prediction model based on grid search optimization. The model uses random forest feature extraction technology to screen the key factors of transmission engineering and uses a grid search algorithm to optimize the random forest model. The accuracy of the model was verified by comparing it with a support vector machine. The model of transmission engineering cost prediction constructed in this paper provides a new idea and a feasible method for transmission line engineering cost prediction and transmission engineering optimization control.

## 2. Methodology

### *2.1 Grid Search*

The grid search method is a search calculation method that divides the area to be searched into a series of grids with equal intervals and then traverses each grid. Specifically, using the grid search algorithm, the program will automatically use the exhaustive method to run all the parameters used and get the combination of parameters with the minimum prediction error, that is, the optimal parameter value. The grid search steps are as follows [12].

(1) Determine the initial regional search area;

(2) The region is divided into several grids according to the standard interval $I_1$(KM), and the first-level search is carried out. The adaptive function of each grid point is calculated, and the smallest grid point $X_1^2$ is taken as the current optimal solution.

(3) The next level search area is constructed according to the standard interval $I_2$ $(I_2 < I_1)$, and the smallest lattice point $X_2^2$ is taken as the current optimal solution.

(4) Repeat the previous step, stop the search when convergence conditions $|X_{i+1}^2 - X_2^2| < \partial$ are reached, and the center position of the last lattice point is the optimal target position.

ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

### *2.2 Random Forest*

To improve calculation accuracy and efficiency, Breiman proposed a random forest algorithm (RF) in 2001. The core idea of the random forest lies in the decision tree and bagged sampling. The essence of the decision tree is a tree-like classifier, and the bag sampling method is a common method to reduce the variance of statistical learning methods. During the training process, the decision tree is constructed independently from the sample samples of the original data set, and the optimal variables in the subset of the prediction model are used to separate each node. After the final prediction model is established, the prediction result is determined by voting or the average value of each decision tree.

Compared with other machine learning methods, such as support vector machine (SVM), backpropagation neural network (BPNN), the RF model has a lower error rate, can better reduce the impact of noise, thus reducing the over-fitting phenomenon. The specific steps of random forest model construction are as follows:

(1) Data sampling to generate training sets for each decision tree. Training each decision tree requires a corresponding training set, so the same number of data sets need to be generated from the original full set.

(2) Build a decision tree. Multiple decision trees are established by using the generated training set and the generated decision tree algorithm. Decision trees in random forests do not need pruning.

(3) Formation of random forest and implementation of the algorithm. Multiple decision trees established through the above two steps will be combined into a random forest. The decision tree will participate in the decision by voting, and the category with the most votes will be the final output result of the random forest algorithm.

## 3. Forecasting Model of Transmission Engineering Cost

The number of decision trees (n_estimators) and candidate feature attributes (max_features) in stochastic forest algorithms have a great influence on model performance. To reduce the influence of its uncertainty on parameter selection and prediction results, the grid search method is adopted to select the optimal parameters. The grid range is set to traverse the two parameters and the global optimal solution of the number of decision trees and candidate feature attributes is found through the highest accuracy. The specific steps are as follows:

(1) The data set was divided into the training set and test set according to 8.5:1.5, and the training set was divided into K folds, that is, K pieces on average;

(2) Determine the number of decision trees and the range of candidate feature attributes. The above two parameters are the coordinate system to establish a two-dimensional grid, and the intersection node of the grid is the corresponding PARAMETER combination of s and max_features;

(3) Determine any K-1 data in the training set;

(4) Select a set of parameters at the intersection of network search, and extract the sample data from K-1 data as the sample of a decision tree;

(5) Predict the remaining data and calculate the root mean square error of all trees on the remaining training samples;

(6) Repeat steps (4) and (5) until the prediction effect of K-1 training samples is traversed;

(7) The corresponding decision tree is constructed by traversing all the parameter combinations at grid intersections to form a random forest, and steps (3) ~ (6) are repeated to select the optimal hyperparameter combination;

(8) According to the optimal parameter combination, all samples in the test set are repeated to establish a random forest model;

(9) Input the test set data into each tree to obtain the prediction effect of the improved model;

(10) The static investment of transmission engineering is predicted by the trained model. The optimal performance of the prediction model is improved by determining the parameter combination with the minimum rmSE or the highest accuracy.

ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

## 4. Empirical Analysis

### *4.1 Raw Data and Data Preprocessing*

This paper takes 588 sets of 110kV and 220kV overhead transmission lines settled by a power grid company from 2015 to 2021 as the original data. However, because the overhead line cost data involves many technical and economic indicators and has a large amount of data, there are inevitable problems such as data loss, ambiguity, inconsistency, noise, and redundancy. All these deficiencies will lead to the failure of the subsequent model prediction. Therefore, samples need to be preprocessed according to the actual situation, and the specific steps are as follows:

(1) Data preprocessing

The problems of data loss, noise, and redundancy are found in transmission engineering data collection. Because of the problems existing in the data, the following processing is carried out in this paper.

1) Data cleaning Delete irrelevant, duplicate, and seriously missing attributes. Data with little impact on cost prediction, repeated data, empty data, and serious missing data will be directly deleted.

2) Missing value filling. In this paper, three methods are used to fill the missing value: ①Fill the fixed value: according to the engineering practice, get the qualitative relationship between the attributes, and fill the fixed value. ②Fill in the statistical value: fill in the mean, median, or mode of the data in the current year according to the attribute characteristics. ③Use random forest regression to fill: select the appropriate number of missing features, use the random forest to fit, build the model to get the predicted value, and then fill.

3) Outlier processing. Using the principle to identify outliers in numerical data, that is, the samples whose distance is greater than the mean value are judged as outliers. After the outliers were removed, the missing value filling method of numerical data was used to fill the above data.

(2) Data conversion. The purpose of data conversion is to convert data into a format suitable for data mining. Several related attributes can be transformed into a higher-level attribute to replace the original attributes through some combination method, to achieve the purpose of transformation, effectively reduce the data space, simplify the mining process.

(3) Feature coding. For non-numerical data, feature coding is needed to encode and quantify it, which is convenient for model calculation. There are many methods of feature coding, such as tag coding, unique thermal coding, binary coding, and so on. Because the unique thermal coding solves the problem that the classifier cannot deal with attribute data, it also plays a role in expanding features to a certain extent. Therefore, in this paper, the ordered discrete variables are numerically characterized by label coding, and the disordered discrete variables are individually coded.

### *4.2 Dataset*

In this paper, the collected transmission engineering data are segmented and 85% of engineering data are selected as training samples to form training sets. The data in the training set is processed, and the identification method commonly used in data mining is used to transform the data. The random forest algorithm is used for classification learning. The remaining 15% of engineering data were used as test samples to test *OOB* estimation and test samples to test the effect of learning. After passing the test, according to the known main cost factors affecting the transmission project cost, as the input variable of the random forest prediction model, the empirical analysis is carried out by using the random forest prediction model.

### *4.3 Evaluation Index*

In this paper, the prediction accuracy is used to evaluate the prediction effect of the model, and its calculation formula is

$$P = (1 - MAPE) \times 100\% \tag{1}$$

Where $P$ is the prediction accuracy, MAPE is the average absolute percentage error (MAPE), and MAPE is calculated by

ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2}$$

Where $n$ represents the number of samples, $y_i$ represents the true value of sample $i$, and $\hat{y}_i$ represents the predicted value of sample $i$.

### *4.4 Result Analysis*

In this paper, the support vector machine model is selected by comparing models. The prediction accuracy of the random forest model and support vector machine model in the test set is shown in Table 1. It can be seen that the prediction accuracy of the random forest model is significantly higher than that of the support vector machine model. The results show that for imbalanced data sets, the random forest model can balance the errors and maintain the accuracy even if some features are missing. In addition, according to the comparison of prediction model results in Figure 1, compared with the SUPPORT vector machine model, the random forest model has better tolerance to outliers and noises, with more input variables and fast convergence speed, and controllable generalization error, which is superior to the support vector machine model. Nearly half of transmission project cost prediction accuracy is higher than 80% by using a random forest model prediction test set, which can be used for the analysis and prediction of the project cost in the next year.

*Table 1: Precision table of transmission engineering prediction model*

| Model | Random forests | Radial kernel support vector machines | Linear support vector machines |
|---|---|---|---|
| Prediction accuracy | 77.83% | 73.68% | 54.53% |

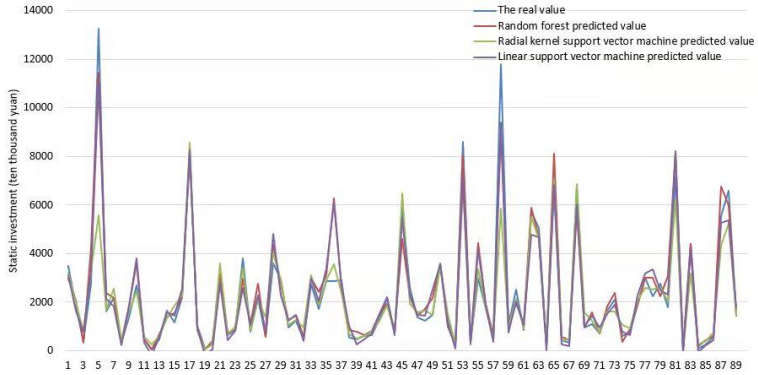


*Fig 1: Result diagram of transmission engineering prediction model*

## 5. Conclusion

Because of the present power engineering cost complex factors, is difficult for the power transmission project cost estimation problem, this paper used the characteristics of random forest importance screening key factors, then adopts the grid search method to select the optimal parameters of random forests, which will extract the key factors affecting the transmission of the project cost as the model input training model. Finally, the proposed transmission engineering cost measurement model is tested with actual engineering data, and the prediction accuracy is compared with the support vector machine model. The experimental results show that the relative error of the prediction results of the random forest model constructed in this paper is low, which is suitable for evaluating the advantages and disadvantages of the comparison scheme in the early stage of the project and can provide a reference for the cost audit and control of transmission engineering.

## References

*[1] Wei Xuelin.Guangxi Economy, 2008(11):36-37. (in Chinese)*
*[2] Abdel-Khalek H., Schafer M., Vasquez R., et al. Forecasting cross-border power transmission*

ISSN 2706-655X Vol.4, Issue 1: 1-6, DOI: 10.25236/IJFET.2022.040101

*capacities in Central Western Europe using artificial neural networks[J]. Energy Informatics, 2019, 2(1) (in Chinese)*

*[3] Zhang Yuchen, ZHANG Yulin, FENG Chunfei et al.Research on transmission line cost Prediction Model based on BP neural network model [J]. Electric Power Big Data, 2020, 23(06):35-42 (in Chinese)*

*[4] Wang Jiao. Analysis of main Factors influencing the cost of 500KV substation Project [J].Journal of Engineering Management, 2012, 26(06):66-69 (in Chinese)*

*[5] Hao Yaogang.Analysis on the Influencing Factors of transmission line Construction Cost [J]. 2011:5 (in Chinese)*

*[6] Lu Yanchao, ZHENG Yan, Zhao Biao.Analysis of the external environmental impact of power transmission and transformation Project [J]. China Electric Power, 2012, 45(10):100-103 (in Chinese)*

*[7] Peng Guangjin, YU Jihui, Cui Rong et al.Transmission engineering cost Estimation Based on Data Mining Technology [J]. Industrial Engineering and Management, 2009, 14(03):90-95 (in Chinese)*

*[8] Wang Jiao, LIU Yanchun.Pso-SVR Project Cost Prediction Model based on Grey Correlation Analysis [J]. Journal of Huaqiao University (Natural Science Edition), 2016, 37(06):708-713 (in Chinese)*

*[9] Ling Yunpeng, YAN Pengfei, Han Changzhan et al.Power Transmission Line Project cost Prediction Model based on BP neural network [J]. China Electric Power, 2012, 45(10):95-99 (in Chinese)*

*[10] Xu Li, Li Zhuoran.Research on Construction Cost Prediction Model of UHV Transmission Line -- Based on Factor Analysis and BP Neural Network [J].Industrial Technology Economics, 2017, 36(07):18-26 (in Chinese)*

*[11] Geng Pengyun, AN Lei, Wang Xin.Construction and Implementation of transmission Engineering Cost Prediction Model Based on Data Mining Technology [J]. Modern Electronic Technique, 2018, 41(04):157-160 (in Chinese)*

*[12] Zhang Yixiao, GUO Wenpu, Kang Kai et al.TDOA/DOA Joint Positioning Method Based on Clustering and Grid Search [J]. Tactical Missile Technology, 2020(01):105-112 (in Chinese)*