Construction and Empirical Analysis of Intelligent Prediction Model for Electric Grid Project Cost

Peng An¹*, Jiaguo Liu², Peng Rong³, Xiang Qi³, Qing Wang³, Yong Dai³, Baijian Yu¹, Fei Yuan², Yunxing Gao³

¹State Grid Shandong Electric Power Company, Shandong Province, Jinan; 250000, China
²State Grid Shandong Electric Power Company Taian Power Supply Company, Shandong Province, Taian; 271000, China
³Corresponding author’s e-mail: 12536060@qq.com

Abstract. China's economic situation has entered a new normal development period, the growth rate of national economic development has fallen, and the state has comprehensively implemented various measures such as price reduction and fee reduction, which makes the operation of electric power enterprises face greater pressure. In this paper, Shandong power transmission and transformation project is taken as the research object, using big data analysis technology and current popular intelligent algorithm to build the intelligent prediction model of power grid project cost, to assist in investment decision-making, reduce the balance rate, strengthen the fine control of investment, improve the precise control level of investment, and enhance the competitiveness of enterprises.

1. Research background
China's economic situation has entered a new normal development period, and the growth rate of national economic development has fallen, which has a serious negative impact on the scale of power demand market and the investment of electric power enterprises. At the same time, in order to support the development of the real economy, the state comprehensively implemented various measures such as price reduction and fee reduction, and the transmission and distribution prices are facing the risk of reduction, which makes the operation of electric power enterprises face greater pressure. Therefore, scientific and accurate determination and effective project cost control and management can take the initiative to adapt to the changes of power demand market, better meet the challenges brought by the "new normal" of economic and energy consumption to the operation of electric power enterprises, which is of great significance to the rational overall investment planning of electric power enterprises and improve the level of investment fine control.

2. Construction framework of intelligent prediction model for power grid project cost
Based on the historical data related to the construction cost of 35kV and above electricity grid construction projects in Shandong Province, through the collection, preprocessing and index dimensionality reduction of the historical cost data, the information input can be used to build the intelligent prediction model of power project cost. Based on the input cost related data and index information, this paper uses neural network, support vector machine, step-by-step regression, random...
forest and other intelligent algorithms to build a combined forecasting model to predict the project cost level. The specific process framework is shown in the figure below:

![Construction flow chart of cost prediction model.](image)

The specific model construction steps are as follows:

- Collect the data of power grid infrastructure projects that have been put into operation in 2013-2018 for 35 kV and above, consider the impact of replacing business tax with value-added tax, exclude the data from 2013-2015, and retain the data from 2016-2018.
- Clean, integrate and reassign the collected data.
- Principal component analysis is used to reduce the dimension of cost factors and eliminate the multicollinearity among variables.
- BP neural network, support vector machine, stepwise regression, random forest and other algorithms are used to build the combined forecasting model.
- The model predicted cost is compared with the actual cost level of the project.

3. Introduction of model construction method

3.1 Principal component analysis

In the study of a variable, it is found that there are many influencing factors, so there is likely to be multiple collinearity between these influencing factors, that is, some of the information contained in these variables is repetitive. If all the influencing factors are used in the process of building the model, not only the human and material resources are increased, but also the results are inaccurate. Therefore, the principal component analysis method can be used to analyze multiple influencing factors in the process of dimensionality reduction, a few independent comprehensive variables which contain most of the information of the original data are the main components.

Suppose that there are n samples and T influence variables, then a matrix of n x T is formed:

\[
X = \begin{bmatrix}
x_{11} & \cdots & x_{1p} \\
\vdots & \ddots & \vdots \\
x_{n1} & \cdots & x_{np}
\end{bmatrix} = (X_1, X_2, \ldots, X_T) \quad X_i = \begin{bmatrix}
x_{i1} \\
\vdots \\
x_{ni}
\end{bmatrix}
\]

The idea of principal component analysis is to find the linear combination of the initial influence variables, so that the following equation can be satisfied:

\[
F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{t1}X_t \\
F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{t2}X_t \\
\vdots \\
F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ti}X_t \\
\vdots \\
F_T = a_{1T}X_1 + a_{2T}X_2 + \cdots + a_{tT}X_t
\]

\(F_i\) is called the main component, \(a_{ij}\) is the coefficient of the main component. From the above equation, it can be obtained that the principal component analysis is to transform the original multiple
indicators into several independent principal components with decreasing variance in order to achieve
the purpose of reducing dimensions.

3.2 BP neural network
BP neural network is the most widely used algorithm model in artificial neural network. The
construction process is mainly divided into two stages: the first stage is the input signal, which passes
through the hidden layer, and finally reaches the output layer; the second stage is to transmit the error
from the output layer to the hidden layer, and finally to the input layer.

For neuron structure, $x_1, x_2, \ldots, x_n$ as the input, $W_{ij}$ as the weight, the input data is weighted by the
weight to get the output $u_i$ of the middle hidden layer, $\theta_i$ as the threshold, the data is processed by the
threshold to get $v_i$, then processed by the activation function, and finally output $y_i$.

The design formula of neuron structure is as follows:

$$u_i = \sum_{j=1}^{N} w_{ij} x_j$$
$$v_i = u_i + \theta_i$$
$$y_i = f(v_i)$$

After the combination of these neuron structures, a complete neural network is formed. For BP
neural network, the commonly used activation function is sigmoid function. Its mathematical form is
as follows:

$$f(x) = \frac{1}{1 + e^{-ax}}^{-1}, \alpha \text{ is constant}$$

Because this function and its derivative are continuous, so it is more convenient in processing.

3.3 Support vector machine
Support vector machine is the most commonly used in the study of binary classification, and now it is
also often used for prediction, and the prediction effect is good. Its basic principle is to classify the
data correctly by constructing a hyperplane. The standard of classification is to maximize the interval.
Finally, the data can be classified or predicted by solving a convex quadratic programming problem.
Data can be divided into linear separable and linear non separable: for linear separable problems, in
two-dimensional space, it is to construct a straight line to classify data; in three-dimensional space, it
is to construct a plane; in multi-dimensional space, it is to construct a hyperplane to classify data. For
the linear indivisible problem, the kernel function is usually used to solve the convex quadratic
programming problem. As for the selection of kernel function, different problems have different
selection principles.

3.4 Multiple linear regression
Multivariate linear regression mainly studies the influence of multiple independent variables on one
dependent variable, that is, there are more than one variable in the research problem, but multiple, and
the linear regression equation established at this time is called multiple linear regression equation. The
expressions of the multiple linear regression equation are as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The main problem of multiple linear regression is that the demand is obtained by $\beta_0, \beta_1, \ldots, \beta_n$. After
obtaining the value of the coefficient, the multivariate linear regression equation can be obtained by
substituting the above multiple linear regression equation expression. The solution to $\beta_0, \beta_1, \ldots, \beta_n$ is
mainly solved by the ordinary least squares method.

3.5 Stepwise regression
Multivariate linear regression is to form the coefficient before the independent variable into a system
of equations and solve them together, while stepwise regression is to introduce variables one by one, and
each introduced variable passes the t-test, that is to say, if the coefficient before the variable fails
the t-test, the variable will be eliminated.
3.6 Random forest
Random forest is mainly the combination of cart algorithm and bagging algorithm. It is mainly through the construction of multiple decision trees, which constitute a random forest. Finally, the results of each decision tree are summarized, and the voting method is adopted to solve the problem studied.

For the construction of random forest, there are four steps. The first step is to use the bagging algorithm to form different training sets. Assuming that the sample size of the original training set is N, then N samples are extracted from the original training set in each round. After K times of extraction, K new independent sample sets can be obtained: $T_1, T_2, \ldots, T_K$.

The second step is to split the internal nodes in the process of generating the decision tree. Assuming that there are M features in the process of building the decision tree, then in each process of generating the decision tree, m-trees will be randomly selected from M features in each internal node, and then the nodes will be split to generate K decision trees.

The third step is to let each decision tree grow freely without pruning.

The fourth step is to combine the decision trees. If it is dealing with the classification problem, we can give the same weight to the generated K decision trees, and then use the simple majority voting method, using the result with the most votes as the final result; if it is dealing with the regression problem, we can use the mean value of the decision output as the final result.

3.7 Combined forecasting model
The combined model is based on the combination of a number of model algorithms according to a certain weight to predict the research variables. Generally speaking, the prediction of variables using a single model may not be very significant, and the error is large, which is not conducive to the study of the problem. However, the combination model can change the shortcomings of a single model. It can reduce the error rate and improve the accuracy of prediction by combining several models according to a certain weight to predict variables.

4. Empirical analysis
In this study, 35kV and above power transmission and transformation project in Shandong Province is taken as the research object. Based on the data of planning, infrastructure management and control, ERP and other systems, the main technical conditions, construction characteristics, budget estimates, final accounts and other relevant data of historical engineering projects are collected, so as to build the corresponding basic database. In the basic database, the data of transformer project and voltage class of 110kV are screened, and the data are preprocessed, i.e. the data are cleaned, integrated and the variables are re assigned. Finally, 14 variables that affect the cost of the transformer project are obtained, a total of 38 valid data. The 14 influence variables are as follows:
As for the selection of indicators, the construction scale of the substation mainly includes the capacity of the main transformer, the number of main transformers in this phase and the number of main transformers in the long term. The basic information of the substation includes the total construction area, the main control building construction area, the total land acquisition area, the main control building construction cost, the site leveling cost, the foundation treatment fee, and the retaining wall and slope protection cost indicators. Through research, these indicators have a certain impact on static investment.

The preprocessed 14 index data were input into R software for principal component analysis, and the results of KMO and Bartlett tests were obtained. The test results can verify whether the data is colinear and suitable for principal component analysis.

| Table 2. KMO and Bartlett's tests. |
|-----------------------------------|
| Kaiser-Meyer-Olkin metric of sampling sufficiency | 0.529 |
| Approximate chi-square | 403.537 |
| Bartlett's sphericity test | df | 91 |
| Sig. | 0.000 |

Regarding the value of KMO, the closer the value of KMO is to 1, the stronger the correlation between variables. It can be seen from table 2 that the KMO value of the data in this study is 0.529, which means that the correlation between the influence variables is relatively strong, and there are multiple collinearity effects among the variables. Therefore, for the cost prediction problem, the original data can not be directly used for prediction, but the influence of multiple collinearity between the data should be eliminated first, so this study is suitable to use the principal component analysis method to the original dimension reduction of initial index. The results of principal component analysis using SPSS software are as follows.

| Table 3. Total variance of interpretation. |
|-------------------------------------------|
| Component | Initial eigenvalue | Extract square sum load | Rotate square sum load |
|            | Total | Percent of variance | Cumulative percentage | Total | Percent of variance | Cumulative percentage | Total | Percentage of variance | Cumulative percentage |
| 1           | 4.014 | 28.668                     | 28.668                     | 4.014 | 28.668                     | 28.668                     | 2.128 | 15.202                     | 15.202                     |
| 2           | 2.372 | 16.942                     | 45.611                     | 2.372 | 16.942                     | 45.611                     | 2.022 | 14.440                     | 29.641                     |
| 3           | 1.718 | 12.270                     | 57.881                     | 1.718 | 12.270                     | 57.881                     | 1.966 | 14.040                     | 43.682                     |
| 4           | 1.294 | 9.240                      | 67.121                     | 1.294 | 9.240                      | 67.121                     | 1.609 | 11.495                     | 55.176                     |
| 5           | 1.078 | 7.697                      | 74.817                     | 1.078 | 7.697                      | 74.817                     | 1.314 | 9.384                      | 64.561                     |
| 6           | .874  | 6.243                      | 81.061                     | .874  | 6.243                      | 81.061                     | 1.093 | 7.805                      | 72.365                     |
| 7           | .702  | 5.011                      | 86.071                     | .702  | 5.011                      | 86.071                     | 1.075 | 7.678                      | 80.043                     |
| 8           | .586  | 4.184                      | 90.255                     | .586  | 4.184                      | 90.255                     | 1.025 | 7.320                      | 87.362                     |
| 9           | .494  | 3.532                      | 93.787                     | .494  | 3.532                      | 93.787                     | .900  | 6.425                      | 93.787                     |
| 10          | .360  | 2.569                      | 96.357                     |        |                           |                           |       |                           |                           |
| 11          | .207  | 1.477                      | 97.834                     |        |                           |                           |       |                           |                           |
| 12          | .155  | 1.109                      | 98.943                     |        |                           |                           |       |                           |                           |
| 13          | .090  | .645                       | 99.588                     |        |                           |                           |       |                           |                           |
| 14          | .058  | .412                       | 100.000                    |        |                           |                           |       |                           |                           |

It can be seen from the above table that nine principal components are extracted in this study, which explain 93.787% of the information of the original variables. Therefore, these nine principal
components can be used to replace the original 14 variables as the input of the cost prediction model, and these nine principal components are independent of each other.

The 38 sample data processed by the principal component analysis method were divided into two parts, of which 35 samples were used as input for the combined model for training, and 3 samples were used for testing.

4.1 An empirical analysis of BP neural network cost prediction

The R software is used to build a neural network cost prediction model. The results of using this model to predict static investment are as follows:

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|-----------------------------------------------|------------|
| 1      | 2653.62                                       | 2766.04                                       | -4.24%     |
| 2      | 2834.38                                       | 2600.668                                      | 8.25%      |
| 3      | 2738.14                                       | 2600.678                                      | 5.02%      |

From the above table, we can see that the error rates of the three prediction samples are -4.24%, 8.25% and 5.02% respectively, the error rates are less than 10%, and the average error is 5.83%. Therefore, the accuracy of neural network algorithm used in this study is better.

4.2 An empirical analysis of support vector machine cost prediction

Using R software to build the support vector machine cost prediction model, the results of using this model to predict the static investment are as follows:

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|-----------------------------------------------|------------|
| 1      | 2653.62                                       | 2596.595                                      | 2.15%      |
| 2      | 2834.38                                       | 2562.737                                      | 9.58%      |
| 3      | 2738.14                                       | 2613.951                                      | 4.54%      |

From table 5, it can be seen that the error rates of the three prediction samples are 2.15%, 9.58% and 4.54% respectively, the error rates are all less than 10%, and the average error of the test samples is 5.42%, so the effect of using SVM to predict the cost is better.

4.3 Empirical analysis of multiple linear regression cost prediction

The multiple linear regression model is constructed by R software, and the results are as follows:

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|---------|
| (Intercept)   | 2470.72  | 122.63     | 20.147  | <2e-16  *** |
| data2[, 1]    | 133.79   | 117.65     | 1.137   | 0.2662  |
| data2[, 2]    | -60.20   | 108.96     | -0.552  | 0.5855  |
| data2[, 3]    | -22.39   | 92.67      | -0.242  | 0.8110  |
| data2[, 4]    | -186.34  | 145.93     | -1.270  | 0.2158  |
| data2[, 5]    | -321.64  | 137.50     | -2.339  | 0.0276  * |
| data2[, 6]    | 206.29   | 119.61     | 1.725   | 0.0969  |
| data2[, 7]    | -14.31   | 111.65     | -0.397  | 0.6948  |
| data2[, 8]    | -234.74  | 317.46     | -0.739  | 0.4665  |
| data2[, 9]    | 51.51    | 119.72     | 0.433   | 0.6689  |

Figure 2. Multiple linear regression results.

It can be seen from the above figure that except for the intercept term, the fifth variable and the sixth variable, all other variables failed to pass the test. Through multiple linear regression results, the prediction results of static investment are as follows:
Table 6. Static investment prediction results of multiple linear regression.

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|-----------------------------------------------|------------|
| 1      | 2653.62                                       | 3446.65                                       | -29.88%    |
| 2      | 2834.38                                       | 1994.95                                       | 29.62%     |
| 3      | 2738.14                                       | 2662.65                                       | 2.76%      |

It can be seen from the table that the prediction error rate of static investment using multiple linear regression is relatively large, with an average error rate of 20.75%, which is greater than 20%. Therefore, multiple linear regression can not be used as a part of the portfolio model, and then stepwise regression is used to correct multiple linear regression.

4.4 An empirical analysis of gradual regression cost prediction

The R software is used to build a stepwise regression model. The model results are as follows:

**Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 2467.54    | 97.05   | 25.424   | <2e-16 *** |
| data2[, 1]  | 198.06     | 99.88   | 1.983    | 0.0563    |
| data2[, 5]  | -305.15    | 127.19  | -2.399   | 0.0226    *
| data2[, 6]  | 240.44     | 101.84  | 2.361    | 0.0247    *

Figure 3. Stepwise regression results.

It can be seen from the above figure that through the use of stepwise regression, variables 1, 5 and 6 are finally left behind, and the results of stepwise regression are similar to those of multiple linear regression. These three variables and intercept terms are used to predict the static investment, and the results are as follows:

Table 7. Results of stepwise regression static investment forecast.

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|-----------------------------------------------|------------|
| 1      | 2653.62                                       | 2879.12                                       | -8.50%     |
| 2      | 2834.38                                       | 2624.37                                       | 7.41%      |
| 3      | 2738.14                                       | 2670.98                                       | 2.45%      |

It can be seen from the above table that the prediction accuracy of stepwise regression is higher than that of multiple linear regression, and the average accuracy of stepwise regression is 6.12%, less than 20%. Therefore, stepwise regression can be used as a part of the combination model.

4.5 Empirical analysis of random forest cost forecast

Using R software to build a random forest model, the static investment prediction results are as follows:

Table 8. Prediction results of random forest static investment.

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|-----------------------------------------------|------------|
| 1      | 2653.62                                       | 2626.685                                      | 1.02%      |
| 2      | 2834.38                                       | 2723.994                                      | 3.89%      |
| 3      | 2738.14                                       | 2580.96                                       | 5.74%      |

The static investment is predicted by the random forest model, the error rate is 1.02%, 3.89%, 5.74% respectively. The error rate of the three test samples is small, and the average error rate of the model is 3.55%.

4.6 An empirical analysis on the cost prediction of combination model

In order to make the prediction results more accurate, this paper will use the combination model to predict the static investment, which includes four models: BP neural network, support vector machine,
stepwise regression and random forest. The reciprocal of the square sum of errors is taken as the weight, and the weight results are as follows:

Table 9. Combined model weights.

| Model                  | Sum of squares of errors | Weight          |
|------------------------|--------------------------|-----------------|
| BP neural network      | 86155.3568               | 0.272901698     |
| Support vector machine | 92464.6778               | 0.292886809     |
| Stepwise regression    | 99464.9157               | 0.315060437     |
| Random forest          | 37616.1156               | 0.119151057     |

Multiply the predicted value by the weight of each algorithm in the above table, and finally get the predicted result of static investment as follows:

Table 10. Static investment prediction results of portfolio model.

| Sample | Actual value of static investment (10000 yuan) | Static investment forecast value (10000 yuan) | Error rate |
|--------|-----------------------------------------------|------------------------------------------------|------------|
| 1      | 2653.62                                       | 2735.435                                       | -3.08%     |
| 2      | 2834.38                                       | 2611.72                                        | 7.86%      |
| 3      | 2738.14                                       | 2624.365                                       | 4.16%      |

The error rate of static investment is -3.08%, 7.86% and 4.16% respectively, and the average error rate of the model is 5.03%. Although the average error rate of the combined model is higher than that of the random forest model, the combined model is the aggregation of the results of various models, and the model is more convincing.

5. Conclusion

Based on the historical cost data of electricity grid construction projects of 35kV and above in Shandong Province, this paper focuses on the construction and analysis method of intelligent prediction model of electricity grid construction projects of 35kV and above. Through the model, the project cost prediction value can be obtained in the early stage of project construction. Through the scientific evaluation and prediction of the investment cost level of different types of projects, it can assist the investment decision-making and improve the level of investment fine management.

References

[1] Lu, Y., (2016) Research on dynamic management model of power grid project cost based on target control. Beijing: North China Electric Power University.

[2] Xie, W.B., (2017) Research on the design of power grid engineering cost information system. Hubei University of Technology.

[3] He, G.P., (2017) Research on cost management of intelligent substation based on life cycle theory. Beijing: North China Electric Power University.

[4] Yu, C., (2016) Study on the influencing factors and evaluation model of power grid project cost. Beijing: North China Electric Power University.

[5] Yang, K., Yu, B., Xiao, Y.L., He, Y.P., Wang, F.X., (2019) Construction cost forecast of distribution network based on GA-BP Neural Network. Process Automation Instrumentation.

[6] Guo, B., Zhou, W.W., Li, Z.W., (2019) Identification of key cost factors of power grid technical transformation project based on grey correlation degree method. Financial Practices.

[7] Liu, Y.H., Zhang, H.Z., Shi, S.F., (2019) Study on the cost difference of technological renovation projects in Beijing and Shanghai based on big data analysis. Electricity and Energy.

[8] Kang, Y.F., Li, D.P., Li, X.Y., Ge, G.W., (2019) Analysis on the influence of power transmission and transformation project cost based on AHP under the background of new technology situation. China Management Informationization.