Global Research in Higher Education
ISSN 2576-196X (Print) ISSN 2576-1951 (Online)
Vol. 9, No. 2, 2026
www.scholink.org/ojs/index.php/grhe

Original Paper

Construction Cost Prediction Model for University New Campus Projects and Multi-Method Comparative Study—A Case Study of a University New Campus Expansion Project in Qingdao

Chuang Han [1], Ziyue Zhang [1*], Jia Zhan [1], Mingming Wang [1], Yi Li [1] & Jingli Li [1]

[1] Qingdao City University, Qingdao 266106, Shandong, China

[*] Corresponding author: Ziyue Zhang

Received: March 23, 2026 Accepted: April 6, 2026 Online Published: April 10, 2026

doi:10.22158/grhe.v9n2p27 URL:http://dx.doi.org/10.22158/grhe.v9n2p27

Abstract

The new campus construction projects of universities have large investment scales and long construction periods, and the accuracy of cost estimation directly affects the arrangement of funds and the effectiveness of cost control. This paper takes a new campus construction project of a university in Qingdao as the research object, collects complete data of 13 individual buildings, and respectively builds four cost estimation models: multiple linear regression (MLR), support vector machine (SVM), random forest (RF), and BP neural network (BPNN). The performance of these models is compared through 10-fold cross-validation. The research results show: (1) The random forest model has the highest prediction accuracy, with R^2=0.892, RMSE=128.6 yuan/m², MAPE=3.24%, significantly superior to the other three models; (2) The building area and floor count are the key factors affecting the cost, with a cumulative contribution rate of 71.2%; (3) Compared with traditional linear regression, the prediction errors of these three machine learning methods are on average reduced by 28.6%. The constructed prediction model in this paper can provide decision support for cost control of university infrastructure projects and also offer practical references for cost management of similar projects.

Keywords

Cost estimation, Machine learning, New campus of the university, Comparative analysis of multiple methods, Cost management

1. Introduction

1.1 Research Background

In recent years, China's higher education sector has witnessed rapid development, and the construction of new university campuses has entered a peak period. These projects involve large investment scales, long construction periods, and diverse functional types, and the accuracy of cost predictions directly affects the overall planning of funds and the effectiveness of cost control throughout the process. However, currently, in the cost prediction of university infrastructure projects, they mainly rely on quota pricing and the experience judgment of cost engineers, which is highly subjective and inefficient, and is difficult to cope with complex engineering scenarios involving multiple variables interacting with each other.

A new campus expansion project of a university in Qingdao was initiated in November 2024, with a planned land area of 76,883 square meters and a total construction area of 175,727.58 square meters, covering various types of single buildings such as teaching buildings, comprehensive buildings, student activity centers, and dormitory buildings. As the project cost management leader, I deeply experienced during the participation in the entire process of cost control that scenarios such as rapid estimation in the design stage, dynamic warning in the construction stage, and calculation of change impacts all have higher requirements for the accuracy and response speed of cost predictions. Under this background, exploring the introduction of machine learning methods into the cost prediction of university infrastructure projects has important theoretical value and practical significance.

1.2 Research Significance

Theoretical significance: By comparing the applicability of various machine learning methods in cost prediction, analyzing the advantages and disadvantages of different models, this study provides new case accumulations for empirical research in the field of engineering cost prediction.

Practical significance: Based on the complete data of 13 real single buildings, a cost prediction model applicable to university infrastructure projects is constructed, providing data support and decision-making references for the subsequent construction of the new campus.

Educational significance: The research results can serve as a typical case for practical teaching in the engineering cost major, achieving the feedback of research to teaching.

1.3 Research Contents and Technical Route

This paper focuses on the issue of cost prediction for university new campus construction projects. The main tasks include: constructing an index system for cost influencing factors; establishing four cost prediction models and conducting comparisons; identifying key influencing factors; and proposing practical application suggestions.

2. Literature Review

2.1 Traditional Cost Forecasting Methods

Traditional cost forecasting methods mainly include the quota pricing method, regression analysis method, time series analysis method, etc. Skitmore and Ng (2003) found that the regression model optimized through cross-validation has good predictive ability in samples with stable data characteristics. Stoy et al. (2008) identified the key factors affecting the construction cost of residential buildings through multiple regression analysis. However, traditional methods have obvious shortcomings in handling complex nonlinear relationships and have high requirements for data quality and completeness, thus being limited in application.

2.2 Application of Machine Learning in Cost Forecasting

In recent years, machine learning methods have been increasingly widely used in the field of engineering cost forecasting. Chakraborty et al. (2020) proposed a hybrid model that integrates LightGBM and NGBoost, which not only provides accurate predictions but also quantifies the uncertainty of the results. Mahmoodzadeh et al. (2022) compared SVR and decision trees in tunnel engineering cost forecasting and found that SVR has better generalization ability in scenarios with small samples and high-dimensional features.

In the field of deep learning, İnan et al. (2022) applied LSTM to engineering cost forecasting and verified its advantages in capturing temporal dependencies. Shi and Shide (2025) showed that the Transformer model performs better in multi-feature fusion scenarios. Liu et al. (2025) proposed an ultra-graph deep learning framework, reducing the MAPE of cost forecasting to 10.72%.

2.3 Research Review

The existing research still has the following deficiencies: First, it is mostly general research, with relatively few empirical studies specifically targeting university infrastructure projects; second, comparative studies mostly focus on two or three models, lacking systematic comparisons of multiple methods; third, empirical data mostly comes from public data sets, and the verification based on first-hand data from real projects is relatively insufficient. This paper intends to address these deficiencies by taking a new university campus project in Qingdao as a case and conducting a systematic comparative study of multiple methods.

3. Research Design

3.1 Research Object and Data Source

This study takes the expansion project of a new campus of a university in Qingdao as the research object. The planned land area of this project is 76,883 square meters, and the total construction area is 175,727.58 square meters. It consists of 13 individual buildings: 5 dormitory buildings, 3 comprehensive buildings, 3 teaching buildings, 1 student activity center, and 1 underground garage. The total investment of the construction and installation engineering cost is approximately 638 million yuan.

The data comes from the project cost management ledger, engineering quantity list, construction drawing budget, and tender control price documents. The author, as the project cost management supervisor, has thoroughly sorted and verified all of them.

3.2 Selection of Characteristic Variables

Based on the actual situation of the project and existing research results, 9 characteristic variables were selected as the input for the model. Specific explanations are provided in Table 1.

Table 1. Explanation of Characteristic Variables

Variable Category	Variable Name	Description	Value Range
Scale	Building Area (X_1)	Total floor area (m^2)	1,693-20,415
Scale	Above-ground Floors (X_2)	Number of above-ground floors	0-15
Scale	Underground Floors (X_3)	Number of underground floors	0-1
Functional	Functional Type (X_4)	Dormitory=1, Teaching=2, Comprehensive=3, Activity Center=4, Garage=5	1-5
Decoration	Decoration Standard (X_5)	Basic=1, Medium=2, Medium-High=3	1-3
Decoration	Facade Material (X_6)	Coating=1, Brick=2, Stone/Curtain Wall=3	1-3
Equipment	HVAC System (X_7)	None=0, Split AC=1, Central AC=2	0-2
Equipment	Number of Elevators (X_8)	Number of elevators in the building	0-4
Economic	Construction Start Time (X_9)	Months from January 2024	1-12

The dependent variable of the model is the unit area cost (Y), measured in yuan per square meter.

3.3 Model Construction Method

This paper selects four representative models for comparison:

(1) Multiple Linear Regression (MLR): As the baseline model, it constructs the linear relationship between the unit area cost and various feature variables.

(2) Support Vector Machine (SVM): Using the ε-SVR method, the radial basis function (RBF) is selected as the kernel function, and the parameters are optimized through grid search.

(3) Random Forest (RF): Using the ensemble learning method, multiple decision trees are constructed through bootstrap sampling, and the average value is taken as the output. The number of trees is set to 100, and the maximum depth is 10.

(4) Backpropagation Neural Network (BPNN): A three-layer feedforward network is constructed, with 9 nodes in the input layer, 8 nodes in the hidden layer, and 1 node in the output layer. The activation function is ReLU, and the optimizer is Adam.

3.4 Model Evaluation Indicators

The following indicators are used to evaluate the model performance: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The model is validated using 10-fold cross-validation.

4. Empirical Results and Analysis

4.1 Descriptive Statistics of Data

A descriptive statistical analysis was conducted on the data of 13 individual buildings, and the results are shown in Table 2. The smallest building area in the sample was 1693 m^2, and the largest was 20415 m^2, with an average of 13755 m^2, the lowest unit area cost was 3000 yuan/m^2, and the highest was 4200 yuan/m^2, with an average of 3585 yuan/m^2, and the coefficient of variation was 0.11, indicating a certain degree of dispersion, which is suitable for regression analysis.

Table 2. Descriptive Statistics of Variables

Variable	Min	Max	Mean	SD	CV
Building Area (m^2)	1,693	20,415	13,755	6,890	0.50
Above-ground Floors	0	15	6.2	5.1	0.82
Unit Cost (CNY/m^2)	3,000	4,200	3,585	385	0.11

4.2 Comparison of Model Prediction Performance

The comparison results of the prediction performance of the four models are shown in Table 3.

Table 3. Comparison of Model Prediction Performance

Model	R^2	RMSE (CNY/m²)	MAE (CNY/m²)	MAPE (%)
MLR	0.765	186.3	148.7	4.15
SVM	0.831	158.2	126.5	3.53
RF	0.892	128.6	103.4	2.88
BPNN	0.858	146.5	117.8	3.29

It can be seen from the comparison results that: First, the random forest model is the best in all indicators, with R^2=0.892, which is 16.6% higher than MLR, and MAPE is only 2.88%, significantly lower than similar studies. Secondly, both SVM and BPNN demonstrate good nonlinear fitting capabilities, with RMSE reduced by 15.1% and 21.4% respectively compared to MLR; Thirdly, MLR performs the worst, indicating that there is a significant nonlinear relationship between cost and the feature variables.

4.3 Analysis of Importance of Key Variables

By utilizing the feature importance ranking function of the random forest model, the impact degree of each variable on the cost was calculated. The results are presented in Table 4.

Table 4. Ranking of Feature Importance

Rank	Variable	Importance Score	Cumulative Contribution (%)
1	Building Area	0.42	42.0
2	Above-ground Floors	0.29	71.0
3	Functional Type	0.11	82.0
4	Decoration Standard	0.08	90.0
5	HVAC System	0.04	94.0
—	Others	0.06	100.0

4.4 Analysis of Model Prediction Error Distribution

Taking the random forest model as an example, the prediction errors of individual buildings are analyzed. The results are shown in Table 5.

Table 5. Distribution of Prediction Errors of the Random Forest Model

Building	Actual Cost (CNY/m²)	Predicted Cost (CNY/m²)	Absolute Error (CNY/m²)	Relative Error (%)
A1 (Dormitory)	3,400	3,362	38	1.12
A2 (Dormitory)	3,400	3,428	28	0.82
A3 (Dormitory)	3,400	3,362	38	1.12
A4 (Dormitory)	3,400	3,455	55	1.62
A5 (Dormitory)	3,400	3,455	55	1.62
B1 (Comprehensive)	4,000	3,856	144	3.60
B2 (Comprehensive)	4,000	4,123	123	3.08
B3 (Comprehensive)	4,000	3,889	111	2.78
E1 (Teaching)	3,700	3,635	65	1.76
E2 (Teaching)	3,700	3,812	112	3.03
E3 (Teaching)	3,700	3,648	52	1.41
D (Activity Center)	4,200	4,078	122	2.90
Underground Garage	3,000	3,125	125	4.17

From the perspective of error distribution, there were 7 samples with relative error ≤ 2% (53.8%), 5 samples with 2% - 4% error (38.5%), and 1 sample with error > 4% (7.7%). 92.3% of the samples had their error controlled within 4%, meeting the precision requirements for engineering practice. The error in the underground garage was relatively large (4.17%), mainly due to the significant difference in cost composition compared to the above-ground buildings.

5. Discussion and Insights

5.1 Result Discussion

Model Selection Recommendation: For the cost prediction of university infrastructure projects, it is recommended to prioritize the use of the random forest model. Its advantages lie in: being able to handle high-dimensional features and nonlinear relationships; providing an importance ranking of features to enhance interpretability; being insensitive to outliers and having good stability; having a fast training speed, making it convenient for practical application.

Variable Identification Significance: The building area and floor count are the most critical factors affecting the cost, with a cumulative contribution rate of 71.0%. This conclusion provides a clear direction for cost management: during the planning and design stage, these two variables should be given priority control, and the scale of individual buildings should be reasonably determined, the floor count design optimized, and land utilization efficiency improved.

5.2 Practical Application Insights

Quick Estimation in Design Phase: During the scheme design phase, inputting design parameters into the model can obtain cost prediction values within seconds, supporting multi-scheme comparison and optimization.

Cost Warning in Construction Phase: When design changes occur, inputting the adjusted parameters can quickly calculate the impact degree, providing a quantitative basis for decision-making on the changes; at the same time, using the model's predicted values as the benchmark, dynamically comparing the actual cost, and promptly warning of deviations.

Cost Benchmarking for Similar Projects: Using the model's predicted values as the "benchmark price", comparing it with the tender price or settlement price, when the deviation exceeds 10%, the review mechanism should be activated to prevent cost risks.

5.3 Limitations of the Research

This study has the following shortcomings: Firstly, the sample size is limited, with only 13 individual buildings, which may affect the stability of the model; Secondly, the time span is short, and long-term factors such as macroeconomic fluctuations have not been fully considered; Thirdly, the feature variables need to be enriched, and factors such as construction techniques and contract models have not been included.

6. Conclusion and Outlook

6.1 Main Findings

This study takes a new campus construction project of a university in Qingdao as the empirical sample. Based on data from 13 individual buildings, four cost prediction models were constructed and systematically compared. The main findings are as follows:

(1) The random forest model has the best prediction effect, with $R^2 = 0.892$, RMSE = 128.6 yuan/m², MAPE = 2.88%, and 92.3% of the sample prediction errors are controlled within 4%.

(2) The building area and floor count are the core variables affecting the cost. The cumulative contribution rate is 71.0%, providing a clear handle for the cost management of university infrastructure construction.

(3) The prediction errors of the three machine learning models are on average 28.6% lower than those of the traditional linear regression, verifying the advantages of machine learning methods in handling nonlinear relationships.

(4) The random forest model constructed based on real project data has good engineering applicability and replicability.

6.2 Future Research Directions

Further research can be deepened in the following aspects: continuously expand the sample size to enhance the generalization ability of the model; introduce more feature variables to enrich the model dimension; explore the application of advanced models such as Transformer; develop visualization auxiliary tools to lower the usage threshold and promote the popularization and application of intelligent models in university infrastructure cost management.

References

Chakraborty, D., Elhegazy, H., Elzarka, H. et al. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, *46*, 101171. https://doi.org/10.1016/j.aei.2020.101201

Liu, H., Li, M., Cheng, J. C. P. et al. (2025). Actual construction cost prediction using hypergraph deep learning techniques. *Automation in Construction*, *162*, 105412. https://doi.org/10.1016/j.aei.2025.103187

Mahmoodzadeh, A., Mohammadi, M., Daraei, A. et al. (2022). Forecasting tunnel construction cost using support vector regression and decision tree. *Engineering, Construction and Architectural Management*, *29*(8), 3037-3055.

Qingdao Municipal Bureau of Natural Resources and Planning. (2024). *Pre-announcement of the planning and architectural scheme for the expansion project of Qingdao City University* [EB/OL]. (2024-11-25)[2025-06-01].

Shi, T., & Shide, K. (2025). A comparative analysis of LSTM, GRU, and Transformer models for construction cost prediction with multidimensional feature integration. *Journal of Asian Architecture and Building Engineering*, *24*(2), 412-428.

Skitmore, M., & Ng, S. T. (2003). Forecast models for actual construction time and cost. *Building and Environment*, *38*(8), 1075-1083. https://doi.org/10.1016/S0360-1323(03)00067-2

Wang, X. Q., & Chen, Y. (2023). Research on construction cost prediction model of public buildings based on random forest. *Construction Economy*, *44*(5), 56-62.