Construction of An Integrated Machine Learning Model for Cost Risk Prediction in Construction Projects

Weiyue Liu

Shenyang University, Key Projects Construction Office,Shenyang 110001, Liaoning, China
Corresponding author: Weiyue Liu (e-mail: LWY40203456@163.com)

ABSTRACT Cost overruns are commonly occurred in construction project. Single prediction model always has low accuracy and generalization ability. It is hard to explore complex cost risk factors accurately. This paper develops an ensemble machine learning model based on stacking strategy to solve the problem. Firstly, collect data from 458 construction projects and extract 27 feature variables, including construction period, material price volatility and construction complexity. Then Principle Component Analysis (PCA) reduces dimensionality into 15 critical features. Finally, the base layer applies Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and Support Vector Machine (SVM) for initial prediction. The metalearner applies logistic regression to integrate output of base model. The optimal parameter combination is searched out by grid search based on 10-fold cross validation. On the test set, the model achieves a 6.1 percentage point improvement in accuracy compared to the best-performing GBDT, an 18.8% reduction in RMSE, and a precision increase to 91.2%. Feature importance analysis shows that the average SHAP value of material price volatility is 0.23 across all samples, construction complexity is 0.19, and the number of design changes is 0.15. This integrated model provides an effective predictive tool for construction cost risk management.

INDEX TERMS construction cost risk, ensemble learning, stacking strategy, machine learning, predictive model

I. INTRODUCTION

COST overruns affect corporate profitability and the return on investment of projects. In recent years, unpredictable prices of construction materials, increasing labor costs, and complicated and dynamic construction environments have caused an increasing number of construction projects to suffer from cost overruns. According to statistics, the average ratio of cost overruns for all large construction projects is still on the rise, and there are even cases where the cost is more than 30% higher than the budget [1,2]. Cost risk prediction depends on human experience and judgment, which has obvious subjectivity and low responsiveness. When faced with complex risk situations, the response is slow and unable to cope with the joint action of multiple factors [3,4]. Although single machine learning models can achieve automated predictions, their generalization ability is still insufficient, and they are greatly affected by the data distribution, and the prediction accuracy cannot meet the needs of engineering decision-making.

It is urgently necessary to develop an accurate and robust intelligent prediction tool to support cost risk management.

Ensemble learning improves overall performance by combining the prediction results of multiple base learners. Stacking strategy uses meta-learners to integrate the outputs of heterogeneous models and has been widely used in areas such as financial risk control and medical diagnosis. This paper constructs a Stacking ensemble model for construction cost risk prediction, takes random forest, gradient boosting decision tree and support vector machine as base learners, and takes logistic regression as a meta-learner. Principal component analysis is used to reduce feature dimensionality. The SHAP values are used to explain the model's decision-making mechanism and to explore the key risk factors affecting the overrun [3,4]. The experimental results show that the ensemble model has higher prediction accuracy and generalization ability than single algorithms. It can provide accurate cost risk warnings for construction companies, assist management in formulating corresponding risk response strategies, and reduce the probability of cost overruns.

II. RELATED WORK

Research on construction cost risk prediction has evolved from qualitative assessment to quantitative modeling. Early research mainly explored influencing factors of cost overrun by statistical method. Researchers conducted interviews and questionnaires with experts to obtain the list of risk factors such as material price fluctuation, design change, construction conditions and management level. These methods

depend on domain experts to construct evaluation index system and calculate the level of risk by AHP or fuzzy comprehensive evaluation. The methods are subjective and unable to deal with large scale project data [5,6]. Statistical regression methods try to establish mathematical relationships between costs and influencing factors. They fit historical data by multiple linear regression or logistic regression in training set and then predict future project cost deviation. However, the method assumes linear relationship between variables and has poor fitting ability for nonlinear complex system. Machine learning technology brings new solutions to cost risk prediction. Support vector machines accomplish nonlinear classification by mapping samples into high dimensional space through kernel function. The method has advantages over small sample scenario. However, the method is sensitive to parameter selection and has high computational complexity. Decision trees and their ensemble variants, random forests and gradient boosting trees, can automatically learn interaction relationship between features, are able to deal with mixed type data and have strong interpretability. However, single decision tree method is likely to overfit. Neural network model is able to learn deep nonlinear relationship between variables. However, they need a large amount of labeled samples in training process and have black box characteristics which limit their application. Ensemble learning method improves the performance of whole system by ensemble multiple base models. Stacking method learns the optimal combination weights of base models by another learner. Most of the existing research focuses on improving single algorithm. There are few heterogeneous model fusion schemes and interpretability explorations which are customized for characteristics of construction cost risk [7,8].

III. METHODS

A. DATA ACQUISITION AND PREPROCESSING

This study collected complete cost data of 458 completed projects in six provinces of China for residential projects, commercial complex projects, and public building projects. The data were extracted from the project contract documents, construction log, material purchase files, and completion settlement data. Due to the existence of missing values in the original data, for the features with missing value rate less than 5%, the K-nearest neighbor algorithm was used to fill the missing values; and for the features with missing value rate higher than 15%, the feature was directly removed. The fluctuation degree of material price was calculated by dividing the standard deviation of unit price of steel, cement and aggregate by their mean value over the life of the project. The construction complexity indicator was calculated as the weighted sum of three subindicators: number of layers, geological condition indicator, and proportion of non-regular components. Data preprocessing. Dimensional differences were removed by using the Z-score standardization method. The variables with different ranges, such as time to complete a project (from 30 to 850 days) and contract price (from 5 million to 830 million RMB), were converted to a normal distribution with a mean of 0 and a standard deviation of 1. Box plots were used to identify and remove the 23 extreme outliers. Pearson correlation coefficient matrix was used to remove highly correlated features; if the correlation coefficient between two features was larger than 0.85, more features with larger variances were kept.

Cost overrun risk was defined as follows: projects with an overrun greater than or equal to 10% ($\geq$10%) were classified as high-risk; overruns between 5% (inclusive) and 10% (exclusive) (5% $\leq$ overrun < 10%) were classified as medium-risk; and projects with an overrun less than 5% (<5%) were classified as low-risk. Finally, a balanced dataset with 137 high-risk samples, 186 medium-risk samples and 135 low-risk samples was constructed.

B. FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION

Twenty-seven feature variables were extracted from the raw data, namely, eight feature variables of project attributes (construction area, type of structure, period of construction, etc.), twelve feature variables of cost composition (ratio of labor cost, ratio of material cost, ratio of machinery cost, etc.) and seven feature variables of risk factors (volatility of material price, number of design changes, number of subcontractors, etc.). To prevent data leakage, the correlation between each feature and the cost risk label was calculated using the mutual information method exclusively on the training set (comprising 320 samples). Finally, the scores of all features were calculated and sorted, and the top scores were 0.341 for material price volatility, 0.298 for construction complexity and 0.276 for number of design changes.

PCA was used to reduce the dimension of 27-dimensional feature space. The eigenvalues and eigenvectors of the covariance matrix were calculated to build principal components. It can be seen from Fig. 8 that the cumulative variance contribution rate of the first 5 principal components approached 68.4%, and the cumulative variance contribution rate of the first 10 principal components approached 85.7%, and the cumulative variance contribution rate of the first 15 principal components approached 93.2%. Therefore, the 15 principal components were selected as the dimensionality reduction results, which retained sufficient information and avoided the curse of dimensionality. Interpretability analysis was conducted on each principal component. As illustrated in Fig. 8, the first principal component (explaining 28.6% of the variance) is primarily contributed to by material cost volatility (loading 0.52), though other factors such as material ratio also contribute. The second principal component (explaining 18.3% of the variance) shows a significant contribution from construction management complexity (loading 0.48), while the third principal component (explaining 12.4% of the variance) is largely influenced by the project scale factor (loading 0.44) [9,10].

C. BUILDING AN INTEGRATION MODEL BASED ON STACKING

The ensemble model consists of two layers. The base layer utilizes three heterogeneous learners: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Support Vector Machine (SVM). The RF layer uses 500 decision trees as hidden units, each of them with max depth of 12, check 4 features every split and at least split samples 5. The GBDT layer uses learning rate 0.05, 300 iterations, subsampling rate 0.8 and max tree depth of 8 layers to avoid over-fitting. The SVM layer uses kernel of type radial basis function, and tune the penalty parameter C to be 100 via grid search, and gamma to be 0.01.

This paper randomly split 458 samples into training set of 320 samples and test set of 138 samples with 7:3 ratio. For the base layer, we use 5-fold cross validation to get the meta-features. We split the training set into 5 parts. Each time we use 4 parts to train the base model and predict the remained part. We do this for 5 times and get the complete prediction result for these 320 samples. The 3 base models output probability high, medium, low respectively. We have 9 probability values for each sample and concatenate them into the meta-feature vector.

D. MODEL TRAINING AND PARAMETER OPTIMIZATION

The parameter optimization consists of a grid search with a strategy of 10-fold cross-validation: We split the training set in 10 subsets and use one of the subsets for validation in order to assess the performance of a model built with the remaining subsets. For RF we search in steps of 100 for the number of trees between 100 and 1000; for the maximum depth we search in steps of 8 to 16; for the minimum number of split samples we test 2, 5, and 10 and thereby get 150 different combinations. For GBDT we search in steps of 0.01, 0.05, 0.1 and 0.2 for the learning rate in steps between 0.01 and 0.2; for the number of iterations we search in steps of 100 between 100 and 500; for the subsampling rate we test 0.6, 0.8 and 1.0 and thereby arrive at 60 combinations. For SVM we search in steps of 10 on a logarithmic scale between 0.01 and 1000 for the penalty parameter C; for the gamma parameter we search in steps of 8 between 0.001 and 1 and thereby get 80 combinations. A 10-fold approach is employed here to maximize the use of the training data for hyperparameter stability, whereas a 5-fold cross-validation is utilized in the stacking process to maintain computational efficiency while generating robust meta-features.

After performing 10-fold cross-validation for each parameter combination, the average accuracy and F1 score were calculated. The parameter configuration with the highest F1 score was selected as the optimal solution. The optimal parameters for RF were 500 trees, a maximum depth of 12, and a minimum number of split samples of 5, with a cross-validation F1 score of 0.874. The optimal parameters for GBDT were a learning rate of 0.05, 300 iterations, and a subsampling rate of 0.8, with a cross-validation F1 score of 0.891. The optimal parameters for SVM were C = 100 and gamma = 0.01, with a cross-validation F1 score of 0.852.

IV. RESULTS AND DISCUSSION

A. COMPARATIVE ANALYSIS WITH A SINGLE MODEL

Test set includes 138 architectural project samples that have never been used in the training set, includes 47 high-risk, 51 medium-risk and 40 low-risk projects, and the class distribution of the test set is consistent with the training set. All samples were standardized as , that is, all samples were standardized with the mean and standard deviation parameters of the training set. Finally, the three models in the base layer independently predicted on the test set, and the output probability were input into the meta-learner for final classification. The accuracy, precision, recall and RMSE of the three models were recorded , and the results are shown in

TABLE 1. Comparison Results

Model	Accuracy	Precision	Recall rate	RMSE	F1 score
RF	83.6%	81.4%	82.8%	0.187	0.821
GBDT	86.2%	84.9%	85.5%	0.165	0.852
SVM	81.9%	79.7%	80.3%	0.203	0.800
Stacking Integration Model	92.3%	91.2%	89.6%	0.134	0.904

The ensemble model achieved a better performance than single models on all four evaluation indicators. Accuracy was improved by 6.1 percentage points and RMSE was decreased by 18.8% compared with the best GBDT. Precision of 91.2% means that the model was more accurate in predicting high-risk items, and false positive rate was only 8.8%. Recall of 89.6% means that we successfully identified 89.6% of real high-risk items (only 10.4% were missed). And F1 score of 0.904 shows that the precision and recall are both relatively reasonable.

Although GBDT achieved a better performance than RF and SVM alone, the ensemble strategy can offset the bias of single model judgment by adding the prediction result of three models together, especially the accuracy of high-risk and medium-risk items were greatly improved.

B. FEATURE IMPORTANCE ANALYSIS

In the calculation process, SHAP explanatory vectors are output for all 138 samples in the test set. The marginal contribution of a feature is reflected in the difference in the prediction of the model before and after its removal. Because of the 15 principal components in the input, the importance of the principal components

was back-mapped to the original 27-dimensional feature space to the 27 original feature spaces. The SHAP value of each principal component is assigned to the corresponding original feature according to its loading coefficient, and then the absolute values of the feature contributions of all 138 samples are averaged and normalized to weight distribution. The average SHAP value of material price volatility, construction complexity, design change frequency, number

of subcontractors, and contract amount of all samples are 0.23, 0.19, 0.15, 0.12, and 0.09, respectively. Finally, the top ten features are extracted according to weight to present the analysis results; the data is shown in Table 2.

TABLE 2. Results of Feature Importance Analysis

Feature Name	Importance weight	Impact on high risks	Impact of medium risk	Impact on low risk
Material price volatility	0.23	+0.31	+0.08	-0.39
Construction complexity	0.19	+0.27	+0.05	-0.32
Number of design changes	0.15	+0.22	+0.11	-0.33
Number of subcontractors	0.12	+0.18	+0.06	-0.24
Contract amount	0.09	+0.13	+0.04	-0.17
Construction period	0.07	+0.10	+0.03	-0.13
Labor cost percentage	0.06	+0.09	+0.02	-0.11
Geological condition score	0.04	+0.06	+0.01	-0.07
Proportion of irregularly shaped components	0.03	+0.04	+0.02	-0.06
Building area	0.02	+0.03	+0.01	-0.04

C. VALIDATION OF MODEL GENERALIZATION ABILITY

The hold-out strategy randomly selected 20% of the original 458 samples as an independent validation set of 92 samples, and retrained the model using the remaining 366 samples. The time-series validation used 63 of the most recent projects completed in the second half of 2024 as the validation set, and trained the model using 395 projects from before the first half of 2024 to simulate real-world prediction scenarios for future projects. Both validation methods recorded accuracy, precision, recall, and RMSE metrics to evaluate model stability. Table 3 presents the model's capability validation results.

TABLE 3. Verification Results

Verification method	Sample size	accuracy	Precision	Recall rate	RMSE	F1 score
Original test set	138	92.3%	91.2%	89.6%	0.134	0.904
Leave out the validation set	92	90.2%	88.9%	87.4%	0.148	0.882
Time series validation set	63	88.9%	87.1%	85.7%	0.159	0.864

The accuracy of validation set was 90.2%, decreased by 2.1 percentage points than that of original test set, and the RMSE increased by 0.014, which demonstrated the robustness of the model to the changes of training data distribution.

The accuracy of the time-series validation set was 88.9%, which represents a decrease of 3.4 percentage points compared to that of the original test set, but still relatively high, and it demonstrated that the predictive ability of model for new projects not included in the time span was effective. The RMSE of validation set in time series was 0.159, increased by 18.7% than that of original test set, and it demonstrated that the difficulty of prediction was increased by the drastic fluctuations of construction material prices and policy adjustment in the second half of 2024.

The F1 value of three kinds of validation sets were larger than 0.86, and the absolute value of the difference between precision and recall was controlled within 4 percentage points, which demonstrated that the model did not have serious phenomenon of overfitting, and it had reliability for engineering cost risk early warning in practice.

V. CONCLUSION

This paper develops an ensemble machine learning model based on a stacking strategy to solve the problem of cost risk prediction in construction projects. Principal component analysis (PCA) is employed to alleviate the dimensionality of high-dimensional feature space. The model combines the advantages of random forests, gradient boosting decision trees and support vector machines, and employs a logistic regression meta-learner to integrate the results generated by heterogeneous models. Finally, accurate classification of high, medium and low risk projects is achieved. Subsequently, the SHAP value analysis shows that material price volatility and construction complexity are the main factors affecting the occurrence of cost overruns, which provide references for risk management decision-making. Hold-out validation and time-series validation show that the model has good generalization ability and can be used for cost early warning in engineering projects. The limitations of this study are that the sample comes from only six provinces in China, so the regional representativeness needs to be expanded; and the dynamic impact of policy and regulations on costs are not considered. In the future, deep learning technology can be introduced to capture the nonlinear interactions between features, combine real-time market data to construct a dynamic prediction system, and further explore the transfer learning application of the model in different regions and engineering types to improve the practicability and application value of the prediction system.

AUTHOR CONTRIBUTIONS

Liu Weiyue designed, collected and analyzed the data, and drafted the manuscript. Liu Weiyue conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Liu Weiyue participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. M. A. Jezzini, R. H. Assaad, et al., "Modeling Framework to Quantify and Gauge Project Cost Risks due to Construction Material Price Volatilities Using Predictive Probabilistic Deep-Learning Algorithms and Stochastic Risk Modeling," Journal of Construction Engineering and Management, vol. 151, no. 7, 2025, DOI: 10.1061/JCEMD4.COENG-16055.

[2] L. B. Sihombing and B. Christin, "Analyzing Contingency Cost Risks in a Pipeline EPC Project Using a Monte Carlo Simulation," Journal of Pipeline Systems Engineering and Practice, 2023, DOI: 10.1061/jpsea2.pseng-1382.
[3] M. H. U. Akbar and Y. Latief, "Risks in the use of BIM 5D in the estimated cost of the project tender phase," AIP Conference Proceedings, vol. 2926, no. 1, pp. 7, 2024, DOI: 10.1063/5.0182843.
[4] A. Lapidus, D. Topchiy, and T. C. O. Kuzmina, "Influence of the Construction Risks on the Cost and Duration of a Project," buildings, vol. 12, no. 4, 2022, DOI: 10.3390/buildings12040484.
[5] M. Sadeghi and M. Lu, "Time-Cost Trade-off Optimization Incorporating Accident Risks in Project Planning," 2021, DOI: 10.1007/978-3-030-51295-8_45.
[6] L. Bai, M. Yang, T. Pan, et al., "Project portfolio selection and scheduling incorporating dynamic synergy," Kybernetes, vol. 54, no. 2, 2025, DOI: 10.1108/K-04-2023-0694.
[7] P. Srinuvasarao, M. Chakkaravarthy, and C. Bhattacharjee, "Risks with Cost Overruns and Project Completion in Public-Private Partnerships: An Examination of the Sondu-Miriu Hydropower Project," Journal of Progress in Civil Engineering, vol. 6, no. 12, pp. 1-7, 2024, DOI: 10.53469/jpce.2024.06(12).01.
[8] T. Yuan, P. Xiang, H. Li, et al., "Identification of the main risks for international rail construction projects based on the effects of cost-estimating risks," Journal of Cleaner Production, vol. 274, Art. no. 122904, 2020, DOI: 10.1016/j.jclepro.2020.122904.
[9] Plebankiewicz E , Wieczorek D .Prediction of Cost Overrun Risk in Construction Projects.Sustainability, 2020, 12(22):9341, DOI: 10.3390/su12229341.
[10] A. N. Ephraem and D. M. M. Nyawira, "Cost-Related Risks and Completion of Commercial Building Construction Projects in Kisumu County, Kenya," International Journal of Research and Innovation in Social Science, 2023, DOI: 10.47772/ijriss.2023.70510.