Study on Concrete Strength Prediction Model Based on Gradient Boosting Regressor

Mingzhen Li

China Nuclear Power Engineering Co., Ltd., Xudapu Project Department, Xingcheng, Liaoning, China

Abstract: Concrete strength is a core indicator for evaluating the quality of construction projects, and its accurate prediction holds great significance for engineering design and construction quality control. However, traditional prediction methods (e.g., empirical formulas and linear regression) struggle to accurately capture the nonlinear relationships between multiple influencing factors and concrete strength. This study proposes a concrete strength prediction model based on Gradient Boosting Regressor (GBR), systematically elaborating its mathematical principles and parameter design logic. Twelve key features—including cement dosage, fly ash content, water-binder ratio, and age—are utilized to construct the prediction model. Through training and validation on 131 sets of concrete mix proportion and strength data from an actual engineering project, the model demonstrates excellent predictive performance: its coefficient of determination (R^2) reaches 0.9063, root mean squared error (RMSE) is 5.0697 MPa, and mean absolute error (MAE) is 4.0340 MPa. The results indicate that the established GBR model can effectively capture the nonlinear relationships between concrete components and strength, providing a scientific basis for concrete mix proportion design and a reference for similar nonlinear prediction problems.

Keywords: Concrete Strength Prediction; Machine Learning; Gradient Boosting Regressor; Mix Proportion Optimization

1. Introduction

As one of the most widely used materials in modern construction engineering, the strength performance of concrete is directly associated with the safety and durability of buildings [1]. Traditional concrete strength prediction primarily relies on empirical formulas and concrete specimen tests, which suffer from drawbacks such as long cycles, high costs, and limited accuracy [2]. With the advancement of artificial intelligence technology, machine learning methods have exhibited significant potential in the field of material performance prediction. These methods can learn latent patterns from historical data to achieve rapid and accurate prediction of concrete strength [3].

In recent years, scholars at home and abroad have attempted to apply various machine learning algorithms to concrete strength prediction. Approaches like Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) have all achieved favorable prediction results in relevant studies [4-6]. As an ensemble learning method, Gradient Boosting Regressor (GBR) iteratively constructs multiple weak learners and combines them with weights. It can effectively handle nonlinear data and avoid overfitting, thereby demonstrating unique advantages in regression prediction tasks [7].

Based on concrete mix proportion and strength data from actual engineering projects, this study constructs a concrete strength prediction model using GBR. Key influencing factors are extracted through feature engineering, model parameters are optimized, and the prediction results are analyzed in depth to verify the model’s effectiveness. This work aims to provide a new technical tool for quality control in concrete engineering.

2. Review of Related Research

Research on concrete strength prediction models has become a hot topic at the intersection of civil engineering and artificial intelligence. Early studies mostly adopted linear regression methods; for instance, the empirical formula for concrete strength proposed by Paul et al. [8] can only reflect the linear relationships between a few factors and strength, resulting in limited prediction accuracy. With the development of machine learning technology, nonlinear models have gradually become the mainstream of research.

Neural network models are widely used in concrete strength prediction due to their strong nonlinear fitting capabilities. Li et al. [9] established a concrete strength prediction model based on 7 raw material parameters using a BP Neural Network, and the results showed that its prediction accuracy was significantly higher than that of traditional empirical formulas. However, neural networks face issues such as numerous parameters and unstable training.

Support Vector Machine (SVM) exhibits good generalization ability in small-sample scenarios. Zhang et al. [10] applied an improved SVM model to the strength prediction of high-performance concrete, determining the optimal parameters via the Particle Swarm Optimization (PSO) algorithm. The average relative error of the model was controlled within 5%. Nevertheless, SVM has low computational efficiency when dealing with large-scale datasets.

Ensemble learning methods improve overall performance by integrating the prediction results of multiple models. Wang et al. [11] compared the performance of three ensemble algorithms—Random Forest, AdaBoost, and Gradient Boosting—in concrete strength prediction, and found that the Gradient Boosting model outperformed the other two methods in all evaluation indicators.

In recent years, deep learning methods have begun to be applied in this field. Chen et al. [12] proposed a concrete strength prediction model based on Convolutional Neural Network (CNN), which reduces the influence of human factors by automatically extracting features. However, this method requires a large amount of data to exert its advantages.

Comprehensively, GBR balances prediction accuracy and computational efficiency on medium-scale datasets, making it suitable as a basic model for concrete strength prediction. Building on this, this study further enhances prediction performance by optimizing feature engineering and model parameters.

3. Mathematical Principles of the Gradient Boosting Regressor Model

3.1 Core Framework of the Model

The essence of GBR is to minimize the loss function through gradient descent iteration and incrementally optimize the prediction model. Its core logic is as follows: the initial model is set to the mean value of the target variable; in each iteration, a new weak learner is trained to fit the prediction error (residual, i.e., the negative gradient of the loss function) of the current model; finally, the prediction results are output by weighted combination of all weak learners. The mathematical expression is:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m \tag{1}$$

Where:

$F_M(x)$ denotes the final prediction model after M iterations;

$F_0(x)$ is the initial model, which takes the mean value of the target variable ($F_0(x) = \bar{y} = \sum_{i=1}^{N} y_i / N$, where N is the total number of samples);

$h_m(x)$ represents the weak learner trained in the m-th iteration (decision trees are used in this study to adapt to nonlinear data);

γ_m is the weight coefficient of the m-th weak learner (solved by minimizing the loss function).

3.2 Mathematical Logic of Iterative Training

For the continuous regression task of concrete strength prediction, Mean Squared Error (MSE) is selected as the loss function. Its continuous differentiability facilitates gradient calculation, and the square term causes the error loss to grow quadratically—this effectively penalizes large errors, which aligns with engineering requirements for strength prediction accuracy. The mathematical expression of MSE is:

$$L(y, F(x)) = \frac{1}{2}(y - F(x))^2 \tag{2}$$

Where:

y is the actual concrete strength, $F(x)$ is the strength predicted by the model.

The pseudo-residual is calculated as follows: before the m-th iteration, the current model is $F_{m-1}(x)$, and its prediction error for samples is x_i represented by the negative gradient of the loss function. Taking the partial derivative of Equation (2) yields:

$$\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\Big|_{F(x_i)=F_{m-1}(x_i)} = -(y_i - F_{m-1}(x_i)) \tag{3}$$

By taking the negative gradient (the direction of steepest descent), the pseudo-residual of the m-th iteration is obtained, which represents the error direction that the current model needs to correct:

$$r_{im} = y_i - F_{m-1}(x_i) \tag{4}$$

The training objective of the m-th weak learner $h_m(x)$ is to fit to $\{(x_i, r_{im})\}_{i=1}^N$ achieve error correction.

After obtaining $h_m(x)$, the weight γ_m is determined by minimizing the loss function. Substituting $F_m(x) = F_{m-1}(x) + \gamma h_m(x)$ into Equation (2), the optimization objective is constructed as:

$$\hat{\gamma}_m = \arg\min_{\gamma} \sum_{i=1}^{N} \frac{1}{2}(y_i - [F_{m-1}(x_i) + \gamma h_m(x_i)]^2 \tag{5}$$

Taking the derivative of with respect to γ and setting the derivative to 0, the optimal weight is solved as:

$$\hat{\gamma}_m = \frac{\sum_{i=1}^{N} r_{im} h_m(x_i)}{\sum_{i=1}^{N} [h_m(x_i)]^2} \tag{6}$$

A learning rate υ (shrinkage coefficient) is introduced to control the contribution of a single tree and avoid overfitting. Referring to the parameter setting experience of Liu et al. (2022) in recycled concrete prediction [13], the final weight is:

$$\gamma_m = \upsilon \hat{\gamma}_m \tag{7}$$

After each iteration, the model is updated according to $F_m(x) = F_{m-1}(x) + \gamma h_m(x)$. The processes of calculating residuals, training weak learners, solving weights, and updating the model are repeated until the number of iterations reaches *M* or the loss of the validation set converges.

4. Data Processing and Model Establishment

4.1 Affiliations

The data used in this study is derived from the test reports of concrete mix proportion and strength of a large-scale construction project, totaling 131 sets. The classification of concrete strength grades is presented in Table 1. The original features in the dataset include: cement dosage, fly ash dosage, water dosage, fine aggregate dosage, coarse aggregate dosage, admixture dosage, age, and cubic compressive strength under standard curing conditions.

Table 1. Classification of Concrete Strength Grades

Category	Strength Grade	Number of Original Samples
I	＜30 MPa	23
II	30 MPa~50 MPa	56
III	＞50 MPa	52

To improve the generalization ability of the model, the following preprocessing steps were performed on the original data:

1.Missing value handling: A small number of missing values were filled using the K-Nearest Neighbors (KNN) algorithm.

2.Outlier detection: Outlier samples were identified and removed using the Z-score method.

3.Feature standardization: All input features were standardized to have a mean of 0 and a standard deviation of 1.

The comparison between data before and after processing is shown in Figure 1 and Figure 2.

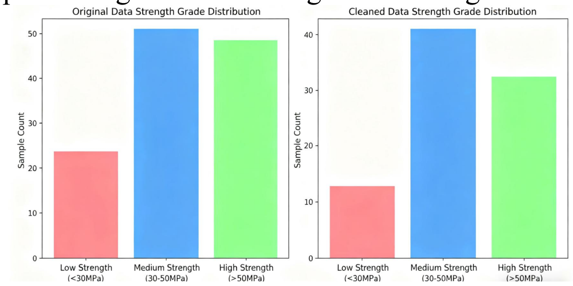

Figure 1. Comparison of the Distribution of Strength Grades Between Original Data and Cleaned Data

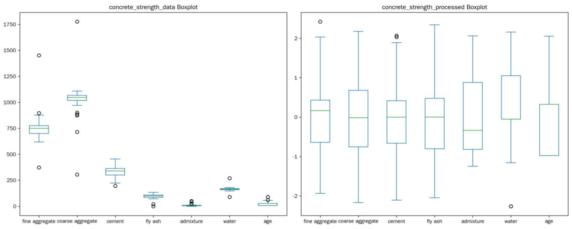

Figure 2. Comparison of the Distribution of Strength Grades in Original Data and the Distribution of Data Features after Standardization

A total of 39 outlier samples were removed during data preprocessing, and the cleaned data was subjected to feature standardization—this ensured all features had a mean close to 0 and a standard deviation of 1, laying a solid foundation for subsequent modeling and analysis.

4.2 Feature Engineering Indicators Affecting Concrete Strength

There are 7 factors in the concrete mix

proportion that influence strength, namely: cement dosage (mc), fly ash dosage (mf), water dosage (mw), fine aggregate (sand) dosage (ms), coarse aggregate (gravel) dosage, admixture dosage, and age.

To investigate the nonlinear effects of the combination of various factors on concrete strength, 5 additional feature engineering indicators were introduced in this study: total binder content (BinderTotal), water-binder ratio (WaterBinderRatio), fly ash content ratio (FlyAshRatio), total aggregate content (AggregateTotal), and sand ratio (SandRatio). Details are provided in Table 2. Finally, the feature set used for model training included 12 features, covering the key influencing factors of concrete mix proportion.

Table 2. Added Feature Engineering Indicators for Concrete Strength

Feature Indicator Name	Mathematical Symbol	Calculation Formula	Description
Total Binder Content	m_b	$m_b = m_c + m_f$	Sum of cement and fly ash dosages
Water-Binder Ratio	W/B	$W/B = m_\omega / m_b$	Ratio of water dosage to total binder content
Fly Ash Content Ratio	β_f	$\beta_f = m_f / m_b$	Ratio of fly ash dosage to total binder content
Total Aggregate Content	m_a	$m_a = m_s + m_g$	Sum of fine aggregate and coarse aggregate dosages
Sand Ratio	β_s	$\beta_s = m_s / (m_s + m_g)$	Ratio of fine aggregate dosage to total aggregate content

4.3 Model Construction

In this study, the GBR algorithm was employed to construct the prediction model. This algorithm generates a series of decision trees iteratively and minimizes the loss function via the gradient descent method. The dataset was divided into a training set and a test set at a ratio of 8:2, and 5-fold cross-validation was adopted for model parameter optimization. The main parameter settings of the model are listed in Table 3.

Table 3. Initial Parameter Settings of the Model

Parameter Name	Symbol	Initial Value
Number of iterations (number of weak learners)	M	300
Learning rate	υ	0.1
Subsample ratio	η	0.8
Maximum tree depth	d	5
Minimum number of samples for splitting	s	10
Minimum number of samples at leaf nodes	t	5

4.4 Evaluation Indicators

Three indicators were selected to assess the model performance:

1.Coefficient of determination (R^2): Measures the model’s ability to explain the variation of the dependent variable. Its value range is (-∞, 1], and a value closer to 1 indicates a better fitting effect

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

2. Root Mean Squared Error (RMSE): Reflects the average deviation between predicted values and actual values, with the same unit as the dependent variable. A smaller value indicates a smaller difference between predicted and actual values.

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{9}$$

3. Mean Absolute Error (MAE): Reflects the average absolute deviation between predicted values and actual values, and is insensitive to outliers. A smaller value indicates a smaller difference between predicted and actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{y}_i - y_i \right| \tag{10}$$

5. Results and Analysis

5.1 Model Performance Evaluation

The prediction results of the GBR model on the test set were compared with the actual values, and the model performance indicators are shown in Table 4.

Table 4. Model Performance Evaluation Indicators

Evaluation Indicator	Value
R^2	0.9063
RMSE	5.0697 MPa
MAE	4.0340 MPa

As indicated in Table 4, the R^2 of the model reaches 0.9063, which means the model can explain 90.63% of the variation in concrete strength and exhibits high goodness of fit. The

RMSE and MAE are 5.0697 MPa and 4.0340 MPa, respectively—this demonstrates that the deviation between predicted and actual values is within an acceptable range. The model has good prediction accuracy and can meet the prediction needs in engineering practice.

5.2 Visual Analysis of Prediction Results

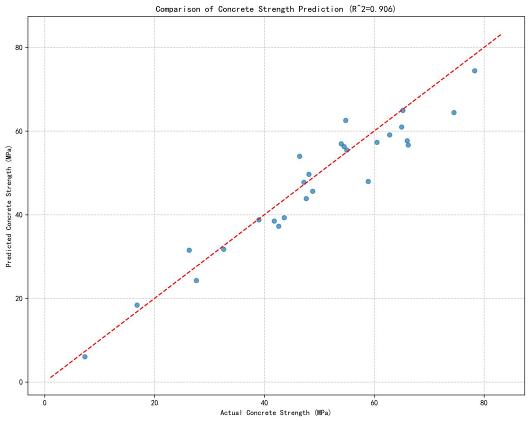

Figure 3. Comparison between Predicted and Measured Values of Concrete Strength

Figure 3 presents a scatter plot of the predicted and actual concrete strength values in the test set. The diagonal line in the figure represents the ideal prediction result (predicted value equals actual value). It can be observed from the figure that most data points are distributed near the diagonal line, indicating good consistency between predicted and actual values and verifying the reliability of the model.

5.3 Feature Importance Analysis

The permutation importance method was used to calculate the influence of each feature on the model's prediction results, and the results are illustrated in Figure 4.

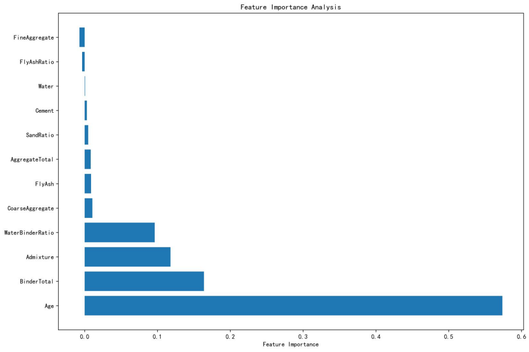

Figure 4. Ranking of the Importance of Various Features to Concrete Strength

The analysis results reveal that age (Age) is the most critical factor affecting concrete strength, with an importance score as high as 0.5738. This is consistent with the basic characteristic that concrete strength increases with age—cement hydration reaction becomes more complete over time, leading to continuous strength development.

Total binder content (BinderTotal) ranks second in importance score (0.1639), indicating that the total dosage of binder has a significant impact on concrete strength. Sufficient binder is the foundation for forming high-strength concrete.

The importance score of admixture (Admixture) ranks third (0.1179), reflecting that in the concrete mix proportions covered by this dataset, admixtures play a crucial role in strength development. This may be attributed to the functions of admixtures in improving concrete workability and promoting hydration reactions.

The importance score of water-binder ratio (WaterBinderRatio) is 0.0964, which is also an important factor affecting concrete strength. This result is consistent with the basic theory of concrete material science—the water-binder ratio directly influences the compactness and strength of cement paste.

It is worth noting that some features (e.g., fly ash content ratio (FlyAshRatio) and fine aggregate (FineAggregate)) have negative importance scores. This indicates that the random permutation of these features in the model slightly improves the prediction effect. This phenomenon may be due to multicollinearity between these features and other features, or their minimal impact on strength within the range of the current dataset.

5.4 Analysis of the Influence of Key Factors

5.4.1 Influence of Water-Binder Ratio on Concrete Strength

With the age fixed at 28 days, the relationship between water-binder ratio and concrete strength was analyzed, and the results are shown in Figure 5.

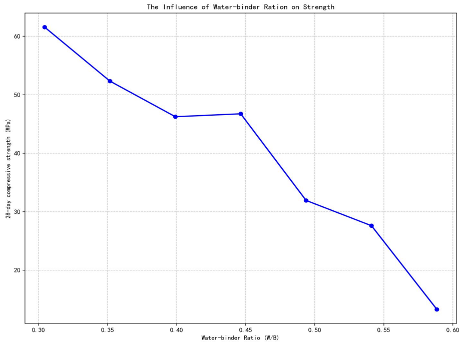

Figure 5. Effect of Water-Binder Ratio on 28-Day Concrete Strength

As shown in Figure 5, within the range of mix proportions studied in this work, concrete strength exhibits a significant negative correlation with water-binder ratio: as the water-binder ratio increases, concrete strength decreases gradually. This result verifies the correctness of the "water-binder ratio rule" in classical concrete theory—under the same other conditions, a smaller water-binder ratio leads to higher concrete strength.

5.4.2 Influence of Fly Ash Content on Concrete Strength

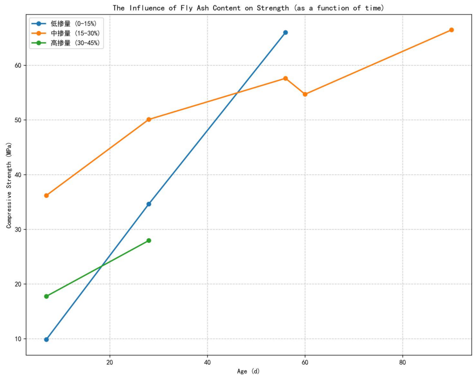

Figure 6. Variation of Concrete Strength with Age under Different Fly Ash Dosages

The influence of different fly ash contents (low content: 0-15%, medium content: 15-30%, high content: 30-45%) on concrete strength at different ages was analyzed, and the results are presented in Figure 6.

As can be seen from Figure 6, at all ages, concrete with different fly ash contents shows distinct strength development patterns. Concrete with low fly ash content has relatively higher early-age strength (3-7 days), which is because the hydration reaction of fly ash is slow, and it mainly plays a filling role in the early stage. With the increase of age, the strength growth rate of concrete with medium and high fly ash content accelerates, reflecting the late-age strength contribution of fly ash. This is associated with the pozzolanic effect and micro-aggregate effect of fly ash.

6. Conclusions and Future Outlook

6.1 Conclusions

In this study, a concrete strength prediction model based on GBR was constructed. Through training and validation on actual engineering data, the following conclusions were drawn:

1.The GBR model demonstrates excellent performance in concrete strength prediction. With an R^2 of 0.9063, its prediction accuracy meets the requirements of engineering applications, providing a reliable tool for concrete strength prediction.

2.Age is the most critical factor affecting concrete strength, with an importance score of 0.5738. It is followed by total binder content, admixture, and water-binder ratio. The combined importance of these four factors exceeds 95%, offering a clear direction for mix proportion optimization.

3.Visual analysis of the model verifies the negative correlation between water-binder ratio and concrete strength, as well as the influence law of fly ash content on strength development. These results are consistent with the theory of concrete material science, indicating that the model has good interpretability.

4.Some features (e.g., fly ash content ratio and fine aggregate) have negative importance scores, suggesting that their impact on strength is minimal within a specific mix proportion range or that they have multicollinearity with other features. This provides a reference for feature selection in subsequent studies.

6.2 Future Outlook

Future research can be expanded in the following aspects:

1. Expand the dataset scale by incorporating more environmental factors (e.g., curing temperature and humidity) and material property parameters to enhance the model's generalization ability.

2. Attempt to combine the GBR model with other machine learning methods to construct a hybrid prediction model, further improving prediction accuracy.

3. Conduct in-depth research on features with low importance, optimize feature engineering methods, and explore the potential relationships between these features and concrete strength.

4. Develop a Web-based concrete strength prediction platform to realize the engineering application of the model and provide a convenient tool for concrete mix proportion design.

5. Combine optimization algorithms to achieve intelligent optimization of concrete mix proportion based on the prediction model, ensuring strength while reducing costs and carbon emissions.

This paper is supported by National Natural Science Foundation of China (No. XXX).

References

[1] Li, J., Zhang, L., and Wang, Q. Machine learning-based prediction of concrete strength: A review. Construction and Building Materials, 2023, 367:130123.

[2] Wang, Y., Liu, X., and Chen, H. Comparative study of machine learning models for predicting compressive strength of recycled aggregate concrete. Journal of Cleaner Production, 2022, 357:131945.

[3] Zhang, S., Li, M., and Zhang, J. Prediction of concrete strength using gradient boosting decision tree and SHAP value analysis. Structural Concrete, 2021, 22(5):2345-2360.

[4] Liu, H., Zhang, Y., and Wu, C. A hybrid machine learning model for predicting the compressive strength of high-performance concrete. Automation in Construction, 2020, 117:103287.

[5] Chen, J., Yu, R., and Li, T. Prediction of concrete strength using random forest algorithm optimized by particle swarm optimization. Materials Research Express, 2023, 10(4):045701.

[6] Zhao, X., Liu, J., and Wang, Y. Compressive strength prediction of recycled aggregate concrete using artificial neural network: A comparative study. Journal of Materials in Civil Engineering, 2022, 34(8):04022184.

[7] Peng, Y., Zhang, L., and Li, Q. Gradient boosting regression for predicting the compressive strength of concrete with supplementary cementitious materials. Materials and Structures, 2021, 54(3):109.

[8] Paul, S. C., Singh, B., and Reddy, K. H. Prediction of compressive strength of concrete using gene expression programming. Advances in Cement Research, 2020, 32(6):245-258.

[9] Li, C., Wang, Q., and Zhang, Y. BP neural network optimized by genetic algorithm for predicting the compressive strength of self-compacting concrete. Neural Computing and Applications, 2023, 35(15):10897-10910.

[10]Zhang, H., Liu, Y., and Chen, W. Support vector machine for predicting the compressive strength of high-strength concrete: A comparative study. KSCE Journal of Civil Engineering, 2022, 26(2):765-776.

[11]Wang, Z., Li, H., and Zhang, J. Comparative study of ensemble learning methods for predicting the compressive strength of concrete. Journal of Testing and Evaluation, 2021, 49(2):1123-1140.

[12]Chen, L., Zhang, Y., and Liu, J. A convolutional neural network approach for predicting the compressive strength of concrete. Automation in Construction, 2023, 147:104604.

[13]Liu, X. Y., Liu, F. Y., Wang, Z. X., et al. Prediction of sulfate resistance of recycled concrete using four ensemble learning methods. Construction and Building Materials, 2022, 314:125637.