DOI: 10.23977/jceup.2025.070217
ISSN 2616-3969 Vol. 7 Num. 2

Machine Learning-Based Prediction of Concrete Compressive Strength and Interpretability Analysis

Linyuan Tang*

Department of Electronic Information and Computer Engineering, Engineering & Technical College of Chengdu University of Technology, Leshan, 614000, China
**Corresponding author: 18782800050@139.com*

Keywords: Concrete compressive strength prediction; SHAP interpretable model analysis; CatBoost model; RF model; XGBR model

Abstract: In contemporary civil engineering, concrete stands as a preeminent construction material, with its compressive strength serving as a core parameter for the safety assessment of engineering structures. This study proposes constructing multiple integrated regression learning models for predictive analysis, supported by interpretable model frameworks to enhance the accuracy of predicting concrete compressive strength. Leveraging the public dataset from the Heywhale community, a comparative analysis of model architectures reveals that the CatBoost model demonstrates optimal comprehensive performance, achieving an R^2 value of 0.92. By employing the advanced SHAP-based DeepExplainer framework, it is identified that Age and Cement are the primary positive influencing factors. Correspondingly, a three-dimensional parameter optimization system is proposed. This approach not only shortens the testing cycle for concrete compressive strength and optimizes concrete mix design but also provides an efficient and convenient tool for real-time project quality monitoring.

1. Introduction

As the most significant amount of artificial construction material in the world[1], the compressive strength of concrete is a core indicator for assessing the safety and durability of engineering structures. Traditional methods mainly rely on destructive testing of standard laboratory specimens, which have the limitations of being time-consuming, costly, and unable to provide real-time feedback on the actual strength status at the project site[2]. By accurately and efficiently predicting the compressive strength of concrete, it can not only monitor the quality of the project in real time and eliminate potential safety hazards in time, but also significantly improve the efficiency of the project, which is of great significance in the aspects of project quality control, cost management, and green and sustainable development.

The research on concrete compressive strength prediction at home and abroad presents the trend of traditional methods and artificial intelligence technology going hand in hand. The conventional method relies on a laboratory standard specimen destructive test. Strict specifications have been established at home and abroad (e.g., China's Standard for Test Methods of Mechanical Properties of Ordinary Concrete [3], the U.S. ASTM standard). Still, it has the inherent shortcomings of an extended period, high cost, and the inability to reflect the real performance of the engineering site. Domestic scholars use the BP neural network and random forest algorithm [4] to achieve multi-factor high-precision prediction, and foreign countries explore the application of support vector machines, convolutional neural networks, and other models to improve prediction efficiency significantly. However, machine learning is generally faced with the problem of "black box", and domestic and foreign scholars have introduced SHAP [5], LIME, and other interpretable methods to quantify the contribution of features and develop visualisation tools to promote the development of model transparency.

The accuracy of the model prediction and interpretable model analysis still needs to be improved. For this reason, this paper is based on actual engineering data, constructs multidimensional feature sets, achieves synergistic enhancement of prediction accuracy and interpretability through CatBoost parameter optimisation and SHAP in-depth feature profiling, and optimizes the model to improve prediction efficiency. It provides a new path for the dynamic control of concrete quality in intelligent construction.

The technical route of this paper is shown in Figure 1.

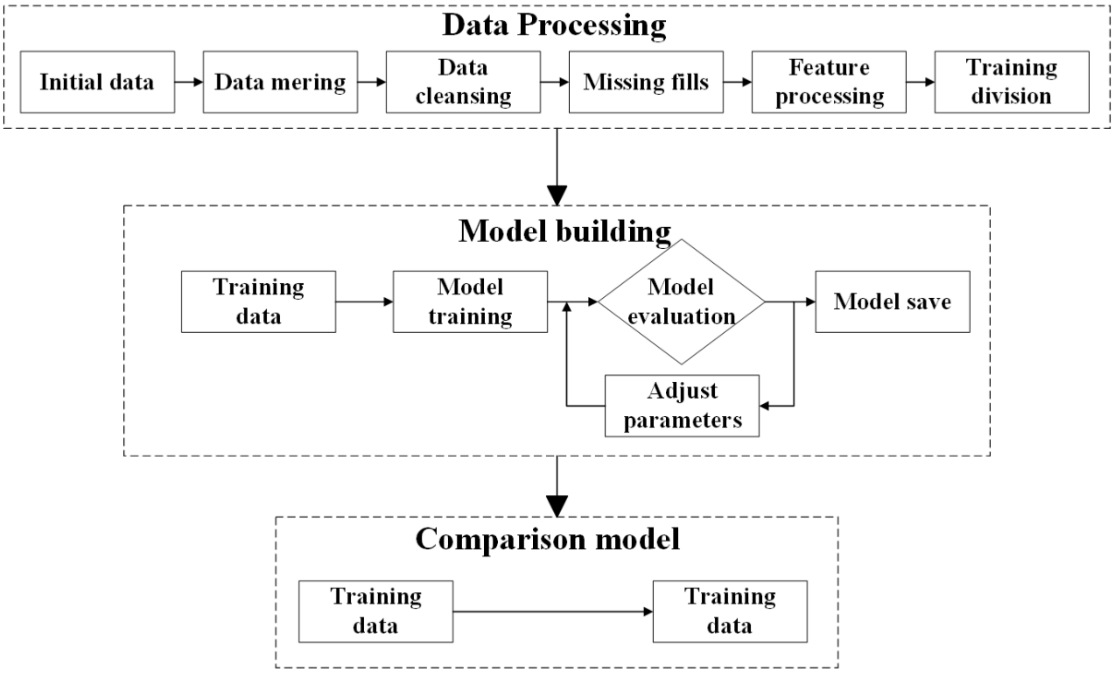

Figure 1 Experimental flow chart

2. Analysis of Data Set and Research Methods

2.1 Data Sources and Analysis

The data for this study was sourced from (https://www.heywhale.com/home).

The dataset used in this study contains the proportioning parameters of 1030 sets of concrete samples and their compressive strength data, involving a total of nine characteristic variables: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age, and Strength.

In this study, the strength of concrete was taken as the dependent variable, and Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age were used as independent variables to construct the concrete strength prediction model.

By performing a statistical description on the independent variable data, the results shown in

Tables 1 and 2 are obtained, indicating that the dataset is complete and contains no missing values.

Table 1 Statistical description table

	Cement **(kg/m ^{3})**	**Blast Furnace Slag** **(kg/m ^{3})**	**Fly Ash** **(kg/m ^{3})**	**Water** **(kg/m ^{3})**
count	1030	1030	1030	1030
mean	281.1678641	73.89582524	54.18834951	181.5672816
std	104.5063645	86.27934175	63.99700415	21.35421857
min	102	0	0	121.8
25%	192.375	0	0	164.9
50%	272.9	22	0	185
75%	350	142.95	118.3	192
max	540	359.4	200.1	247

Table 2 Descriptive table of data statistics (continued)

	Superplasticizer **(kg/m ^{3})**	**Coarse Aggregate** **(kg/m ^{3})**	**Fine Aggregate** **(kg/m ^{3})**	**Age** **(day)**
count	1030	1030	1030	1030
mean	6.204660194	972.918932	773.5804854	45.66213592
std	5.973841392	77.75395397	80.17598014	63.16991158
min	0	801	594	1
25%	0	932	730.95	7
50%	6.4	968	779.5	28
75%	10.2	1029.4	824	56
max	32.2	1145	992.6	365

The data distribution plots for each feature are shown in Figure 2:

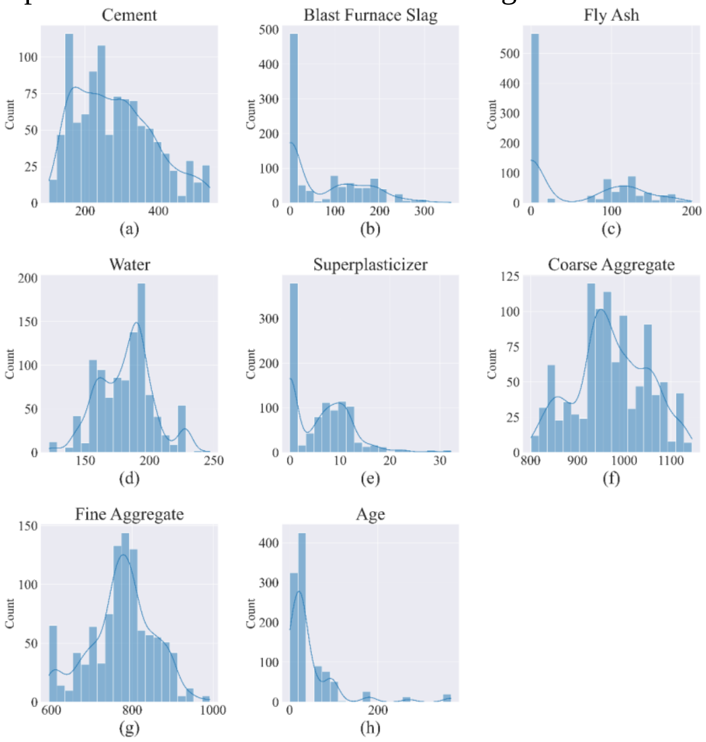

Figure 2 Data distribution chart

2.2 Model Analysis

2.2.1 CatBoost Model

CatBoost is a GBDT framework based on symmetric decision trees as the base learner, featuring fewer parameters, support for categorical variables, and high accuracy [6]. Its primary focus is on efficiently and reasonably handling categorical features. CatBoost constructs balanced trees, where the leaves of the previous tree are split at each step using the same conditions. The feature split pair with the lowest loss is selected and applied to all nodes at all levels. The CatBoost algorithm starts from an initial state, iteratively constructs decision trees, and updates the model's predicted values based on negative gradient information and step size. The results of all iterations are then aggregated to obtain the prediction for new samples. In each iteration, the negative gradient is calculated based on the current model state, and a new decision tree is constructed to fit the negative gradient, thereby gradually optimizing the model's predictive capability. This balanced tree structure facilitates efficient CPU implementation and reduces prediction time. Unlike traditional boosting algorithms, CatBoost uses different data subsets for model training and residual calculation, preventing target leakage and overfitting, effectively countering noise points in the training set, and mitigating gradient estimation bias. CatBoost specific algorithm (1):

$$C(x) = \sum_{m=1}^{M}\left[C_{i,m-1} + \lambda_m T_m(x_i;\theta_m)\right], C_{i,0} = 0 \tag{1}$$

$C(x)$ is the CatBoost prediction result;x is the independent variable; M is the total number of decision trees; m is the number of current iterations; $C_{i,m-1}$ is the predicted value of the sample i at the iteration m; λ_m is the step size at iteration m; $T_m(x_i;\theta_m)$ is the predicted value of the mth decision tree on sale x_i;$\theta_m$is the parameter of the mth decision tree.

2.2.2 RF Model

Random Forest Regression [7] is a classical regression method based on integrated learning, and its core idea is to reduce model variance and improve generalisation performance by constructing multiple discrepancy decision trees and aggregating prediction results. Various subsets of the same size are generated from the original training set using putative sampling, and in the node splitting stage of each decision tree, candidate features are randomly selected from all the features, and the optimal splitting point is determined by maximising the information gain. Each decision tree grows without restriction until the leaf node purity reaches a threshold or the number of samples reaches a minimum to preserve the detailed features of the data. For new samples, the predicted values of all trees are arithmetically averaged as the final output. The specific algorithm for random forest regression (2) is:

$$\mathrm{R(x)} = \frac{1}{\mathrm{B}}\sum_{\mathrm{i=1}}^{\mathrm{B}} \mathrm{T_i(x)} \tag{2}$$

$R(x)$ is the random forest regression result; x is the independent variable; $T_i(x)$ is the predicted value of the ith tree; B is the total number of trees.

2.2.3 XGBR Model

XGBR[8] is a regression model based on the XGBoost (eXtreme Gradient Boosting) algorithm. It fits the target variable by iteratively training multiple regression trees (CART trees). In each iteration, XGBR calculates the residual between the current model's predicted value and the actual value, then constructs a new regression tree to predict this residual. Each tree is designed to reduce the overall model's prediction error. By continuously accumulating the prediction results of new trees, the model gradually approaches the actual value, and the final prediction value is obtained by summing the prediction results of all regression trees. The specific algorithm is:

$$\mathcal{L}(x) = \sum_{m=1}^{n} L(y_m, \hat{y}_m) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

$\mathcal{L}(x)$ is the final output value; $L(y_m, \hat{y}_m)$ is the loss function; $\Omega(f_k)$ is the complexity regularisation term for the kth tree, defined as (4):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{4}$$

T is the number of leaf nodes in the tree; w is the weight of the leaf nodes; γ and λ are hyperparameters that control the complexity of the model.

2.2.4 Evaluation Parameters

Mean Absolute Error (MAE) is a core indicator used to evaluate the accuracy of prediction models or estimation methods. It intuitively reflects the average deviation between predicted values and actual observed values, defined as (5):

$$\mathrm{MAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} |\tilde{y}_{\mathrm{i}} - \hat{y}_i|}{\mathrm{N}} \tag{5}$$

Mean Squared Error (MSE) is a classic indicator for evaluating model fit. It quantifies the deviation between the predicted values and the actual observed values by calculating the average of the squares of the differences between the predicted values and the actual observed values, and is defined as (6):

$$\mathrm{MSE} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{(\hat{y}_{\mathrm{i}} - \tilde{y}_{\mathrm{i}})^2}{\mathrm{N}} \tag{6}$$

The Coefficient of Determination (R2) is a key statistic in regression analysis that measures the effectiveness of model fit and takes a value ranging from 0 to 1. The closer the value of R2 is to 1, the stronger the model's ability to explain the variability of the data; conversely, the closer it is to 0, the weaker the model's ability to explain the variability, defined as (7):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\tilde{y}_{\mathrm{i}} - \hat{y}_i)^2}{\sum_{i=1}^{N} (\tilde{y}_{\mathrm{i}} - \bar{y})^2} \tag{7}$$

In equations (6), (7) and (8) above, the specific meanings of each element are explained as follows: $\hat{y}_i$ represents the predicted value; $\tilde{y}_{\mathrm{i}}$ represents the actual observed value; N represents the total number of samples; $\bar{y}$ is the mean value of the actual observed values.

3 Research Process and Results

3.1 Model Construction Process Analysis

The corresponding parameters must be set reasonably to construct the three models accurately. To optimise the model parameters and obtain the best prediction results, we use a combination of random search and grid search to determine the parameters.

In constructing a CatBoost model, four parameters, n_estimators, l2_leaf_reg, learning_rate, and depth, play a key role in the prediction effect of the model, and therefore need to be adjusted and optimised. The training set was comprehensively analysed using the grid search method. The optimal parameters of the model were finally determined, and the optimal parameters of the model were obtained by randomly searching the training set as follows Table 3:

Table 3 CatBoost Parameter Settings

Parameters	Values
n_estimators	368
learning_rate	0.34
depth	4
l2_leaf_reg	1.96

In the construction of model 2, the random forest model, the three parameters n_estimators, max_depth, and bootstrap need to be adjusted. Through grid search of the training set, the optimal parameters of the model are obtained as shown in Table 4:

Table 4 RF Parameter Settings

Parameters	Values
n_estimators	400
max_depth	71
bootstrap	False

When constructing the XGBR model, its performance is closely related to many parameters. To achieve the best prediction results, grid search was used to tune the key parameters colsample_bytree, learning_rate, max_depth, min_child_weight, gamma, and n_estimators. The optimal parameters for the model are shown in Table 5:

Table 5 XGBR Parameter Settings

Parameters	Values
n_estimators	150
learning_rate	0.3
colsample_bytree	0.9
max_depth	3
min_child_weight	1
gamma	0

3.2 Analysis of Projected Results

Each model predicts the concrete compressive strength of the validation set using the model trained above, and the results are shown in Figure 3.

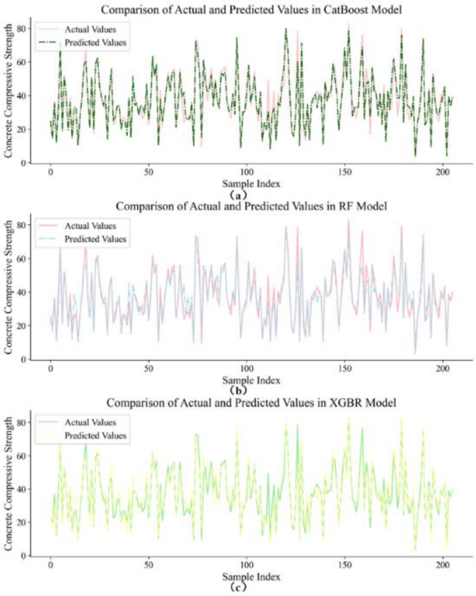

Figure 3 Distribution chart of prediction results

As seen from the observation in Figure 3, the CatBoost model predicts the best results, and there is not much difference between the RF model and the XGBR model. All three models can fit the concrete compressive strength data to a certain extent, but all have different degrees of prediction errors.

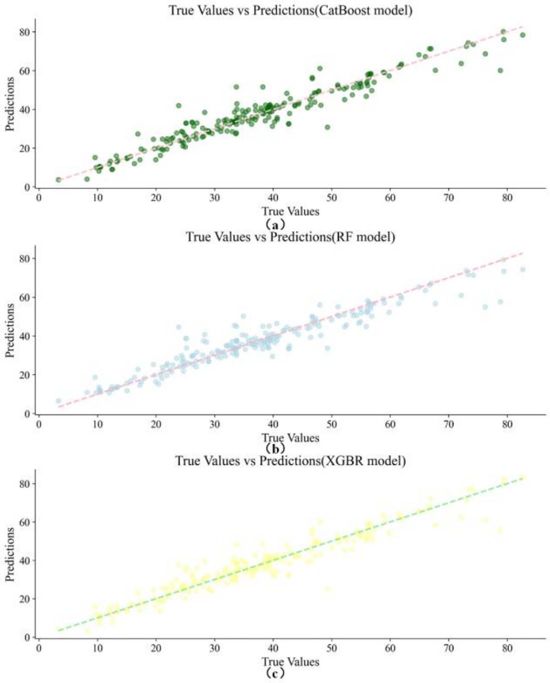

Figure 4 Scatterplot of prediction results

As can be seen from Figure 4, the scatter of the CatBoost model is relatively more concentrated near the fitting line, which may be slightly better in terms of overall predictive stability; the scatter of the RF and XGBR models is relatively more discrete.

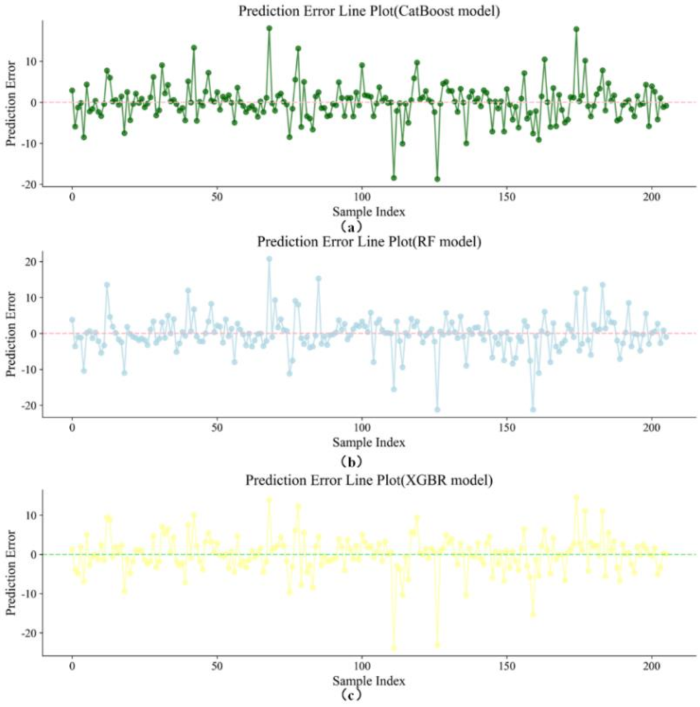

Figure 5 Error line graphs

As can be seen in the Error line graphs in Figure 5, the prediction errors of all three models fluctuate around 0, but the CatBoost model significantly outperforms the RF and XGBR models.

The model evaluation parameters are shown in Table 6 below:

Table 6. Parameters for model evaluation

Models	MAE	MSE	R^2
CatBoost	0.18	0.07	0.92
RF	0.20	0.09	0.90
XGBR	0.20	0.08	0.91

A comparison of the above evaluation parameters reveals that the CatBoost model performs better in the three indicators of mean absolute error, mean square error, and coefficient of determination. It also has a relatively better prediction accuracy and fitting effect on the concrete compressive strength data. Therefore, this paper uses the CatBoost model to predict the concrete compressive strength with the best effect.

4. SHAP Model Interpretability Analysis

4.1 Introduction to SHAP Model

SHAP (Shapley Additive Explanations) [9] is an approach to model interpretation based on the game theory Shapley Value. The core idea is to decompose the model's predicted value into a weighted sum of features while considering the interactions between features to provide an intuitive and theoretically rigorous explanation. SHAP improves interpretability in machine learning, assists in feature engineering and model optimisation, enhances credibility and fairness, and provides a basis for decision-making. And its computational principle (8) is as follows:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{8}$$

ϕ_0 is the baseline value;ϕ_i is the SHAP value of feature i, which is calculated as follows (9):

$$\mathrm{\phi_i} = \sum_{S \subseteq \{1,\dots,M\} \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[f\left(x_{S \cup \{i\}}\right) - f(x_S) \right] \tag{9}$$

M is the number of features; S is the subset of features that does not include feature i; x_S is the feature value retained only in subset S; $f(x_S)$ is the predicted value based on subset S; $\frac{|S|!(M-|S|-1)!}{M!}$ is the weight of all possible feature permutations and combinations.

4.2 Visual Analyses of Feature Significance

Given the excellent performance of the CatBoost model in concrete compressive strength prediction, this paper introduces the DeepExplainer framework for SHAP to improve the prediction accuracy. The importance of features is dissected by calculating and ranking each input feature's absolute mean SHAP values. As shown in Figure 6, the horizontal coordinates of the graph are different features, and the vertical coordinates are the mean absolute SHAP values. The results show that Age and Cement are the core features affecting the prediction of concrete compressive strength; features like Water and Blast Furnace Slag also have a role to play, while Superplasticizer, Coarse Aggregate, Fine Aggregate, and Fly Ash are less influential. When optimising the model, the focus should be on the key features while considering the combined effect of the other features.

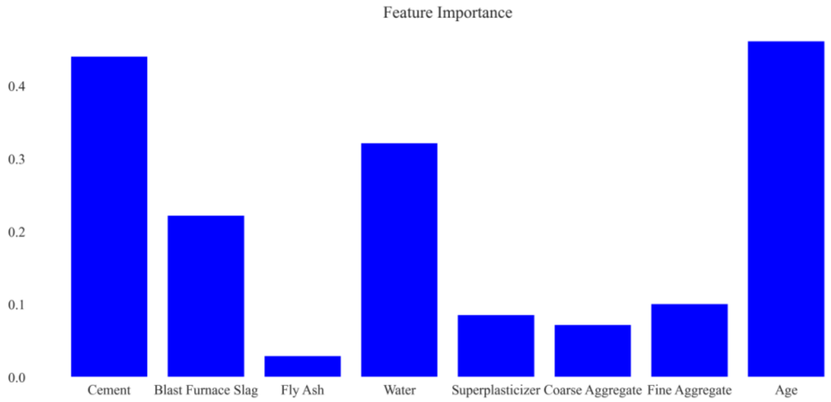

Figure 6 Characteristic Importance Chart

As shown in Figure 7, this SHAP value summary plot systematically demonstrates the effect of eight key features, such as Age, Cement, Water, Blast Furnace Slag, Fine Aggregate, Superplasticizer, Coarse Aggregate, and Fly Ash, on the concrete compressive strength of the SHAP value distribution characteristics. The SHAP values of each feature are distributed bilaterally and symmetrically along the baseline, where the left side of the baseline characterises the adverse effect of the feature on compressive strength and the right side characterises the positive impact. The model generates a scatter distribution of the corresponding feature values for each sample, and the gradient of the chromatogram varies from blue to red, which accurately maps the monotonous incremental law of the independent variable values.

The three features, Age, Cement, and Water, showed significant feature importance through quantitative analysis of SHAP values. An in-depth analysis of the influence mechanism of each feature showed that:

The Age eigenvalue is positively correlated with the SHAP value. The SHAP value shifts to the right when its eigenvalue increments, indicating that prolonging the curing age can significantly enhance the predicted value of concrete compressive strength. This phenomenon aligns with the theory of the time effect of the hydration reaction of cement. It verifies the age parameter's time-varying characteristics in the strength development [10]. The impact of Cement eigenvalue on the output of the model exhibits a significant linear law, with the SHAP valuemonotonically increases with the increase of cement dosage, which coincides with the dominant role of cement cementitious material in the three-phase system of concrete, confirming that cement dosage is the core factor determining the 28d compressive strength. The Water feature exhibits a nonlinear response characteristic. The change of its eigenvalue shows a bidirectional moderating effect on the strength prediction: a moderate amount of water promotes the complete hydration of cement, and the SHAP value is positively shifted; however, the SHAP value is positively shifted when water-cement ratio exceeds. However, when the water-cement ratio exceeds the critical threshold, the excess free water leads to an increase in porosity, and the SHAP value turns to be negatively shifted, which is consistent with the theory of pore structure formation in concrete.

Based on the above characteristic interaction law, it is suggested to establish a multi-objective constraint system in the optimisation of the mixing ratio to achieve the synergistic enhancement of the compressive strength and workability of the concrete by controlling the three-dimensional parameter space of the age of the curing period (≥28d), optimising the cement dosage (350-450 kg/m ^{3}) and adjusting the water-cement ratio (0.4-0.5) to provide a theoretical basis for the preparation of high-performance concrete.

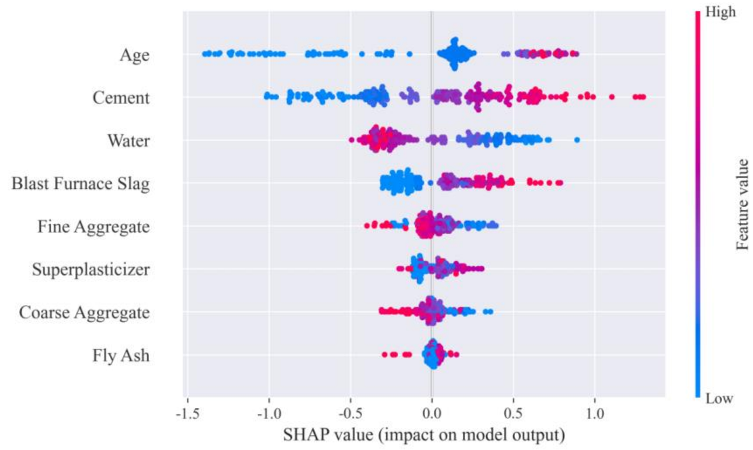

Figure 7 Summary of eigenvalues

The analysis of the Figure 8 reveals that Age and Blast Furnace Slag variables make a significant positive contribution to the prediction of concrete compressive strength, and Fine Aggregate, Cement, and Water variables make a significant negative contribution to the prediction of concrete compressive strength.

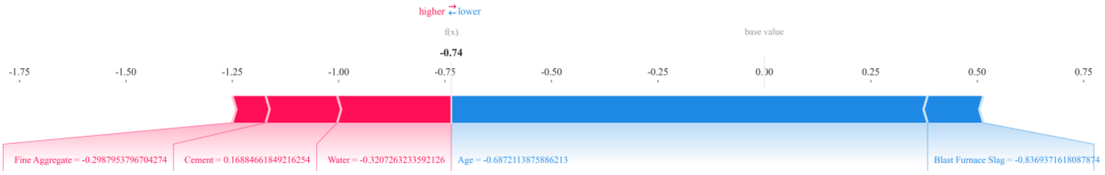

Figure 8 Force Diagrams

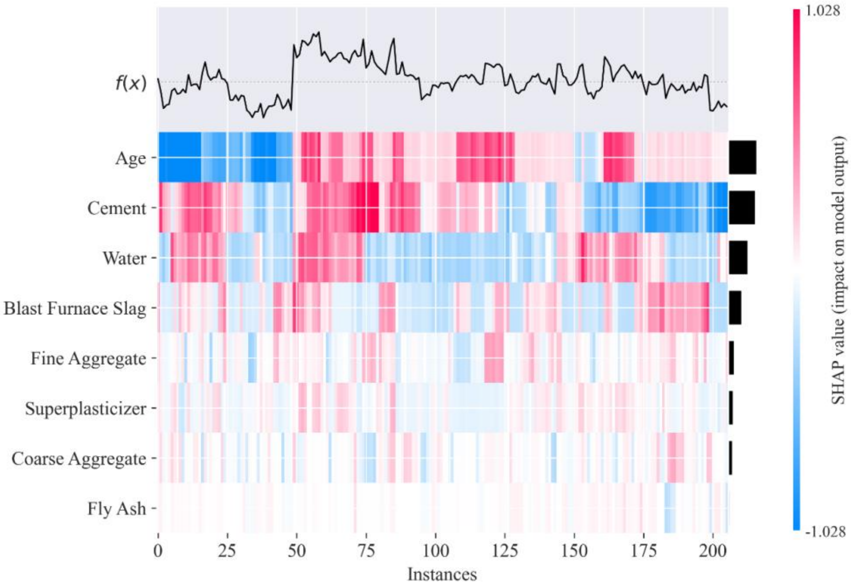

Figure 9 SHAP Dependency Diagram

As shown in Figure 9, it can be seen that these features of Age, Cement, and Blast Furnace Slag contribute positively to the prediction of concrete compressive strength. At the same time, Water, Cement, Age, Fine Aggregate, Superplasticizer, Coarse Aggregate, and Fly Ash negatively contribute to the prediction of concrete compressive strength.

5. Conclusions and Outlook

5.1 Conclusion

This study focuses on the prediction of concrete compressive strength using machine learning, with the main conclusions as follows:

The CatBoost model demonstrates excellent predictive performance. SHAP analysis indicates that Age and Cement are key contributing features. An increase in age promotes cement hydration and enhances structural density, showing a significant positive correlation with strength. As the core cementitious material, the amount and quality of cement directly determine strength development, making a prominent contribution to model predictions.

Removing the Fly Ash feature increased the R ^{2} value in the validation set from 0.92 to 0.93, suggesting that this feature has a limited contribution and may interfere with prediction accuracy.

In summary, age and cement are core factors in strength prediction, while fly ash plays a secondary role and poses a potential interference risk. The findings provide a basis for optimizing concrete mix proportions and selecting features for predictive models.

5.2 Outlook

The multi-feature prediction model based on the CatBoost algorithm performs well in the concrete compressive strength prediction study. However, there are still problems, such as a single channel of data collection, the limitation of dynamic monitoring due to the insufficient sample size, and the weak generalization ability of the model under complex working conditions. In the future, the Transformer architecture can be introduced to strengthen the time-varying feature analysis capability and capture the intensity evolution pattern using its sequence modelling advantage. This study can provide technical support for engineering quality control and optimizing building materials with theoretical innovation and practical value.

References

[1] Shi Y , Li J , Zhang Y ,et al. Interpretable Machine Learning Method for Compressive Strength Prediction and Analysis of Pure Fly Ash-based Geopolymer Concrete[J]. Journal of Wuhan University of Technology-Mater. Sci. Ed. 2025, 40(1):65-78.DOI:10.1007/s11595-025-3041-8.

[2] Ouyang B. Concrete's Strength Prediction Using Machine Learning Method[J]. Engineering Computer, 2024, 000(000):138.

[3] Shahab M , Shakeel M , Gulaly L ,et al. Predicting the compressive strength of self-compacting concrete containing recycled aggregate and supplementary cementitious materials using machine learning techniques[J]. Progress in Engineering Science, 2025, 2(2). DOI:10.1016/j.pes.2025.100077.

[4] Jesse W ,Brennan B ,Marc M . Creating a Universal Depth-to-Load Conversion Technique for the Conterminous United States Using Random Forests[J].Journal of Cold Regions Engineering,2022,36(1).

[5] Ji Y ,Shang H ,Yi J , et al.Machine learning-based models to predict type 2 diabetes combined with coronary heart disease and feature analysis-based on interpretable SHAP[J]. Acta Diabetologica,2025, (prepublish):1-16.

[6] Yixiao Z ,Zhongguo Z ,Jianghua Z . CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China[J]. Journal of Hydrology,2020,588(prepublish)

[7] Geng C ,Chew W K ,Yeo S K , et al. Physically-based RF model for metal-oxide-metal capacitors[J]. Electronics Letters,2000,36(5):425-427.

[8] Zhang Y, Chen L, Tian Y. A Method for Evaluating the Interpretability of Machine Learning Models in Predicting Bond Default Risk Based on LIME and SHAP[J]. Progress in Computer Science, 2025.

[9] Celal C ,Yaren A ,Gebrail B , et al. Interpretable Predictive Modelling of Basalt Fiber Reinforced Concrete Splitting Tensile Strength Using Ensemble Machine Learning Methods and SHAP Approach.[J]. Materials (Basel, Switzerland),2023,16(13).

[10] Meiting Z ,Haichen G ,Yinghao X .Exploration of Teaching Reform in the Course of Concrete Structure Design Principles in the Context of the New Era[J]. World Journal of Educational Research,2024,11(6).