Concrete Compressive Strength Prediction

Project 3

Trey Schulman (treys3)

10/27/2025

Table of Contents

1. Introduction *2*

2. Data and Preprocessing *2*

3. Methodology *3*

3.1 Polynomial Regression 3

3.2 Regression Splines 4

3.3 Smoothing Splines 4

3.4 Regression Tree 4

3.5 Random Forest 4

4. Results *5*

5. Discussion *6*

1. Introduction

The purpose of this project is to build and evaluate predictive models for estimating concrete compressive strength using mixture composition and curing age. Concrete strength prediction is an essential problem in materials engineering, as accurate models can inform quality control and optimize mix design. Each ingredient contributes differently to the final strength, and these effects often interact in nonlinear ways.

The task is particularly suited for nonlinear modeling because high-performance concrete is highly complex and contributions of different materials to overall concrete strength are nonlinear and exhibit diminishing returns. These relationships suggest that linear models may be insufficient.

The goal of this project was therefore to compare several regression approaches of increasing flexibility, including Polynomial Regression, B-splines, Smoothing Splines, Regression Trees, and Random Forests. Each model was trained on a 70/30 train-test split with a fixed random seed of 598, and performance was evaluated using the Mean Squared Error (MSE) on the test set as the primary success metric.

2. Data and Preprocessing

The dataset used for this project is the UCI Concrete Compressive Strength dataset, containing 1,030 observations. The predictors include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age. The response variable is compressive strength measured in megapascals. Each record corresponds to a specific mixture and curing condition.

The dataset contained no missing values and was entirely numeric, which simplified preprocessing. Exploratory data analysis confirmed that cement and age were positively correlated with strength, while water showed a negative relationship. The correlation heatmap (Figure 1) also revealed that some material components exhibited moderate multicollinearity, which reinforced the motivation for using flexible models. The variable age did exhibit a strong right skew (skewness=3.27) with many small values and a long tail. Therefore, age was log-transformed to assist with modeling and creating a more stable relationship with the response.

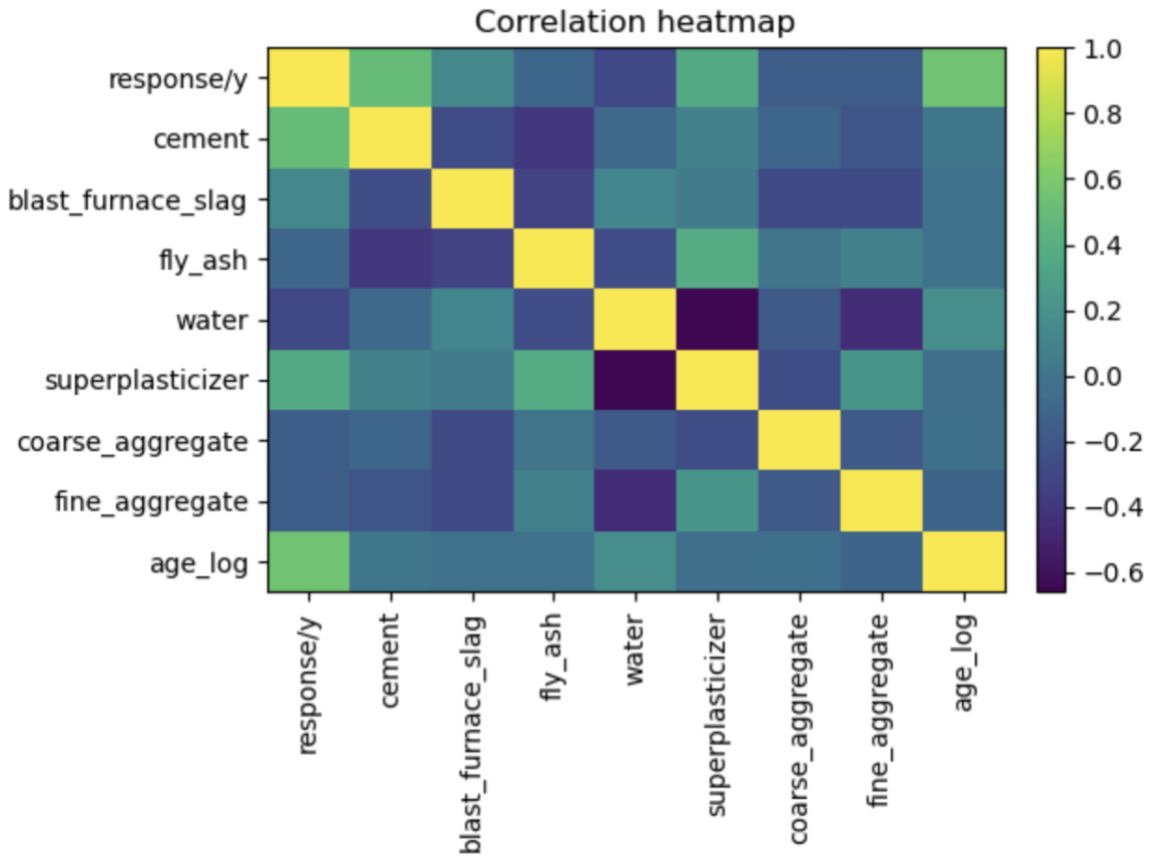

Figure 1: Correlation Heatmap

3. Methodology

All models were trained on the 70 percent training set and evaluated once on the 30 percent test set. A five-fold cross-validation framework was used throughout model selection to tune hyperparameters in a consistent manner across methods. This ensured that all comparisons were fair and performance differences were due to modeling capabilities rather than data partitioning.

3.1 Polynomial Regression

Polynomial Regression was implemented using a scikit-learn pipeline consisting of a StandardScaler, PolynomialFeatures, and a LinearRegression estimator. Polynomial degrees of one through three were evaluated with cross-validation guiding selection. The model with degree two produced the lowest validation error. The model's ability to incorporate feature interactions was particularly valuable since four of the model's top eight largest feature coefficients came from interaction terms, which helped capture relationships like the combined effect of cement and water on the response. This configuration was retained as the polynomial benchmark.

3.2 Regression Splines

To enable more localized flexibility than global polynomials, regression splines were constructed using the patsy bs() function to generate B-spline bases for each predictor. Cross-validation was performed across a grid of degrees of freedom {3, 4, 5, 6, 7, 8} and spline degrees {1, 2, 3}. This ensured linear, quadratic, and cubic splines were explored with varying degrees of freedom. The optimal model used six degrees of freedom and quadratic spline bases, which provided a balance between smoothness and flexibility. Models with very low degrees of freedom underfit curvature, while those with higher values risked overfitting and instability. The B-spline approach successfully captured nonlinearity but tended to produce higher test error than smoother penalized alternatives, which is possibly due to collinearity among spline basis functions.

3.3 Smoothing Splines

Smoothing splines were implemented through a Generalized Additive Model (GAM) using the PyGAM library. Each predictor entered the model through an additive smoothing term s(x), and the level of smoothness was selected automatically via internal grid search. The GAM effectively balanced bias and variance by penalizing excessive curvature, rather than pre-specifying degrees of freedom as in B-splines. This approach allowed the model to learn appropriate smoothness per feature directly from the data. The resulting partial-dependence plots confirmed plausible nonlinear relationships between the features and response. The GAM also outperformed the polynomial and spline regressions in test MSE, which indicates that automated smoothness selection improved generalization while maintaining interpretability.

3.4 Regression Tree

A single regression tree was fit using the training data and then pruned via the cost-complexity pruning path. Cross-validation across candidate ccp_alpha values identified the optimal pruning constant of approximately 0.0015. This process actually had no real impact on the train or test MSE, and both values actually very slightly increased. This demonstrates that even with pruning the regression tree is still too weak a learner and greatly overfits the data. The model provided useful qualitative insights from observing cement and age at the top nodes, but it lacked the predictive stability required for this regression task.

3.5 Random Forest

Finally, a Random Forest model was trained using an ensemble of 900 decision trees. The grid search tuned the number of estimators across {300, 600, 900, 1200}, the maximum number of features sampled at each split across {sqrt, 0.3, 0.5, 0.7}, and the minimum number of samples per leaf across {1, 3, 5}. The best-performing configuration used 900 trees, a feature sampling fraction of 0.7, and a minimum leaf size of 1. Random Forests

average multiple decorrelated trees, which effectively reduces variance without increasing bias. This model produced the lowest cross-validated and test MSE among all methods. Feature importance rankings confirmed that age, cement, and water were the dominant predictors, aligning well with engineering intuition.

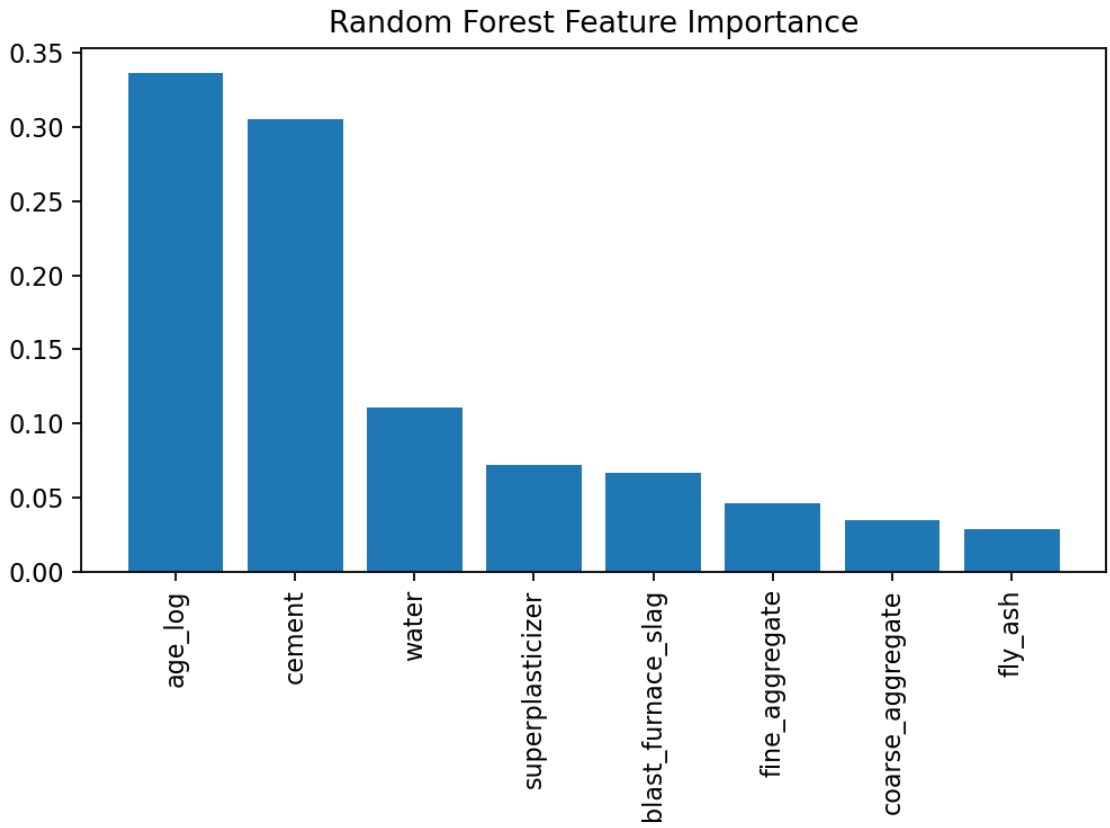

Figure 2: RF Feature Importance

4. Results

Table 1 summarizes the final test MSE for each model. The Random Forest achieved the lowest test error at 29.6, followed closely by the GAM at 32.31. The Polynomial Regression performed moderately well, while B-splines and especially the single Regression Tree produced higher errors. These outcomes reflect the bias–variance tradeoff across modeling families where simpler models underfit curvature, and overly flexible ones overfit noise unless regularized or aggregated. The strong performance of the Random Forest is also show in Figure 3 where predicted values closely match actual values with some larger residuals and heteroscedasticity at high strengths.

Model	Test MSE
Random Forest	29.60
Smoothing Splines	32.31
Polynomial Regression	38.46
Regression Splines	49.45
Regression Tree	63.08

Table 1: Test MSE Comparison

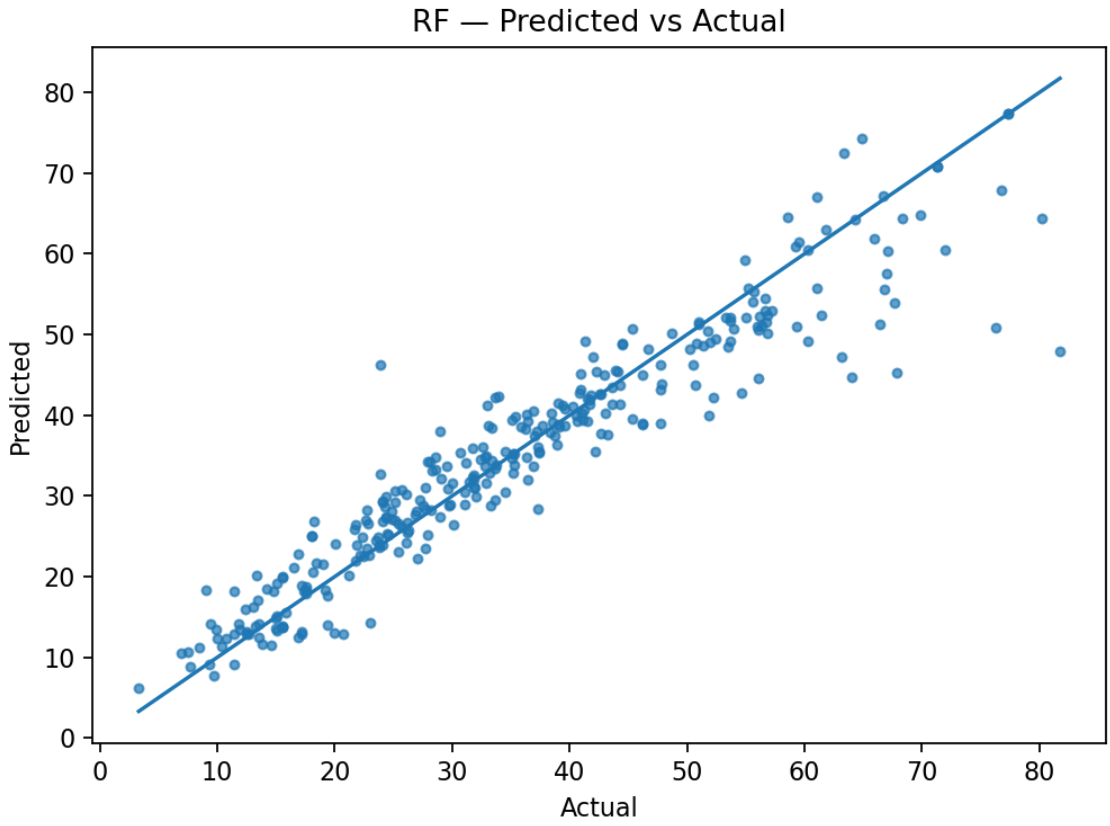

Figure 3: Predicted vs. Actual (Random Forest)

5. Discussion

The comparative results demonstrate that concrete strength is governed by nonlinear and interactive processes that are best captured through flexible models. The Random Forest achieved the best predictive performance and suggests that nonlinear ensemble methods generalize effectively in this context. The GAM model's results, though slightly less accurate, added interpretability by illustrating the specific shapes of each relationship. Together, these two models provide a robust and complementary understanding of the data.

The polynomial and spline regressions improved upon linear modeling by capturing curvature, but they lacked the adaptability to model complex interactions beyond simple additive or global effects. The single regression tree offered interpretability but performed poorly in accuracy and confirmed the need for ensembles like Random Forests to mitigate variance.

From an engineering perspective, the findings align well with known material science principles. Cement content and curing age are the primary drivers of strength, while excessive water negatively impacts it. The results suggest that data-driven models can serve as powerful decision-support tools for optimizing mix design and evaluating quality control in concrete production.