Concrete Compressive Strength Prediction

Introduction

Concrete is the most widely used construction material in civil engineering, and its compressive strength is a indicator of structural strength and durability. Predicting the compressive strength of concrete accurately can help optimize material usage, improve cost-efficiency, and ensure structural safety. The relationship between the concrete's strength and its constituent materials is highly nonlinear, making it a nonlinear regression problem.

In this project, we use the Concrete Compressive Strength dataset from the UCI Machine Learning Repository, which contains 1,030 observations of concrete mixtures. Each record includes eight input variables representing the quantities (in kg/m³) of the concrete ingredients and the curing age (in days), along with the measured compressive strength (in MPa) as the response variable.

Dataset Description

Variable	Type	Unit	Description
Cement	Quantitative	kg/m³	Component 1 – Cement content
Blast Furnace Slag	Quantitative	kg/m³	Component 2 – Slag content
Fly Ash	Quantitative	kg/m³	Component 3 – Fly ash content
Water	Quantitative	kg/m³	Component 4 – Water content
Superplasticizer	Quantitative	kg/m³	Component 5 – Superplasticizer content
Coarse Aggregate	Quantitative	kg/m³	Component 6 – Coarse aggregate content
Fine Aggregate	Quantitative	kg/m³	Component 7 – Fine aggregate content
Age	Quantitative	days (1–365)	Age of the sample at testing
Concrete Compressive Strength	Quantitative	MPa	Measured compressive strength (Target)

The dataset contains no missing values, and all predictors and the response are continuous.

Objective

The goal of this project is to build predictive models for the compressive strength of concrete based on its composition and curing age.

The workflow involves the following steps:

1. Fit and compare multiple regression models, including:
    - Polynomial Regression
    - Spline Regression
    - Regression Tree
    - Random Forest
    - Other candidate models
2. Split the dataset into 70% training and 30% testing subsets (with random_state = 598).
3. Evaluate model performance using Mean Squared Error (MSE) on the test data.
4. Apply cross-validation for hyperparameter tuning.
5. For tree-based models, analyze feature importance to interpret influential predictors.
6. Compare models based on prediction accuracy, interpretability, and model complexity.

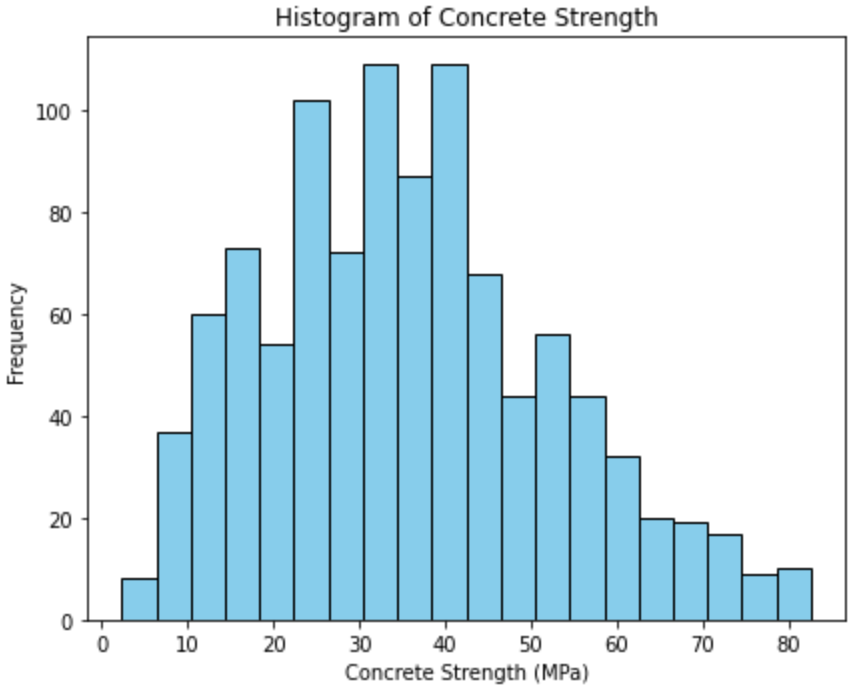
Information about the data and Sample records

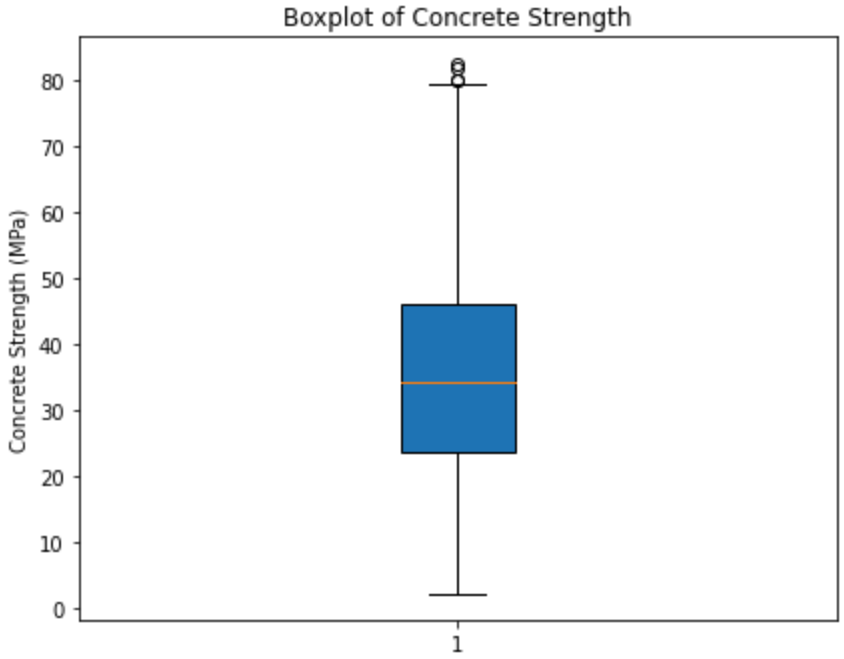
```
Details and sample records from the concrete mixture dataset:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1030 entries, 0 to 1029
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   cement_1              1030 non-null   float64
 1   blast_furnace_slag_2  1030 non-null   float64
 2   fly_ash_3             1030 non-null   float64
 3   water_4               1030 non-null   float64
 4   superplasticizer_5    1030 non-null   float64
 5   coarse_aggregate_6    1030 non-null   float64
 6   fine_aggregate_7      1030 non-null   float64
 7   age_in_days_8         1030 non-null   int64  
 8   concrete_strength     1030 non-null   float64
dtypes: float64(8), int64(1)
memory usage: 72.5 KB
```

	cement_1	blast_furnace_slag_2	fly_ash_3	water_4	superplasticizer_5	coarse_aggregate_6	fine_aggregate_7	age_in_days_8
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360

Exploratory data analysis

Histogram and Box plot of response varaible





Histogram Interpretation

- The histogram appears slightly right skewed
- That is most of the concrete compressive strength either have lower or moderate value(between 30MPa and 40MPa)
- Some of the concrete compressive strength are higher(ard 70MPa to 80MPa), which is stretching the right tail.
- The spread of the concrete compressive strength are continuous and wide indicating a substantial variation in compressive strength

Boxplot Interpretation

- The median, the orange line is roughly in the middle of the boxplot indicating moderate skewness
- The interquartile range(IQR) that is the height of the box, shows where the middle 50% of data lie
- There are few outliers which have high concrete compressive strength

Testing for collinearity between predictors

- To check the collinearity between the predictors we can use correlation matrix
- **Correlation Matrix**

	cement_1	blast_furnace_slag_2	fly_ash_3	water_4	superplasticizer_5	coarse_aggregate_6	fine_aggrega
cement_1	1.000000	-0.275193	-0.397475	-0.081544	0.092771	-0.109356	-0.22
blast_furnace_slag_2	-0.275193	1.000000	-0.323569	0.107286	0.043376	-0.283998	-0.28
fly_ash_3	-0.397475	-0.323569	1.000000	-0.257044	0.377340	-0.009977	0.07
water_4	-0.081544	0.107286	-0.257044	1.000000	-0.657464	-0.182312	-0.45
superplasticizer_5	0.092771	0.043376	0.377340	-0.657464	1.000000	-0.266303	0.22
coarse_aggregate_6	-0.109356	-0.283998	-0.009977	-0.182312	-0.266303	1.000000	-0.17
fine_aggregate_7	-0.222720	-0.281593	0.079076	-0.450635	0.222501	-0.178506	1.00
age_in_days_8	0.081947	-0.044246	-0.154370	0.277604	-0.192717	-0.003016	-0.15
concrete_strength	0.497833	0.134824	-0.105753	-0.289613	0.366102	-0.164928	-0.16

Interpretation of correlation matrix

- From the correlation matrix we can see that almost all of the predictors have correlation score < 0.5 and hence are not correlated.
- Some of the predictors are slightly negatively correlated. ex: **superplasticizer and water have correlation = -0.657464**

Correlation of the predictors with concrete compressive strength

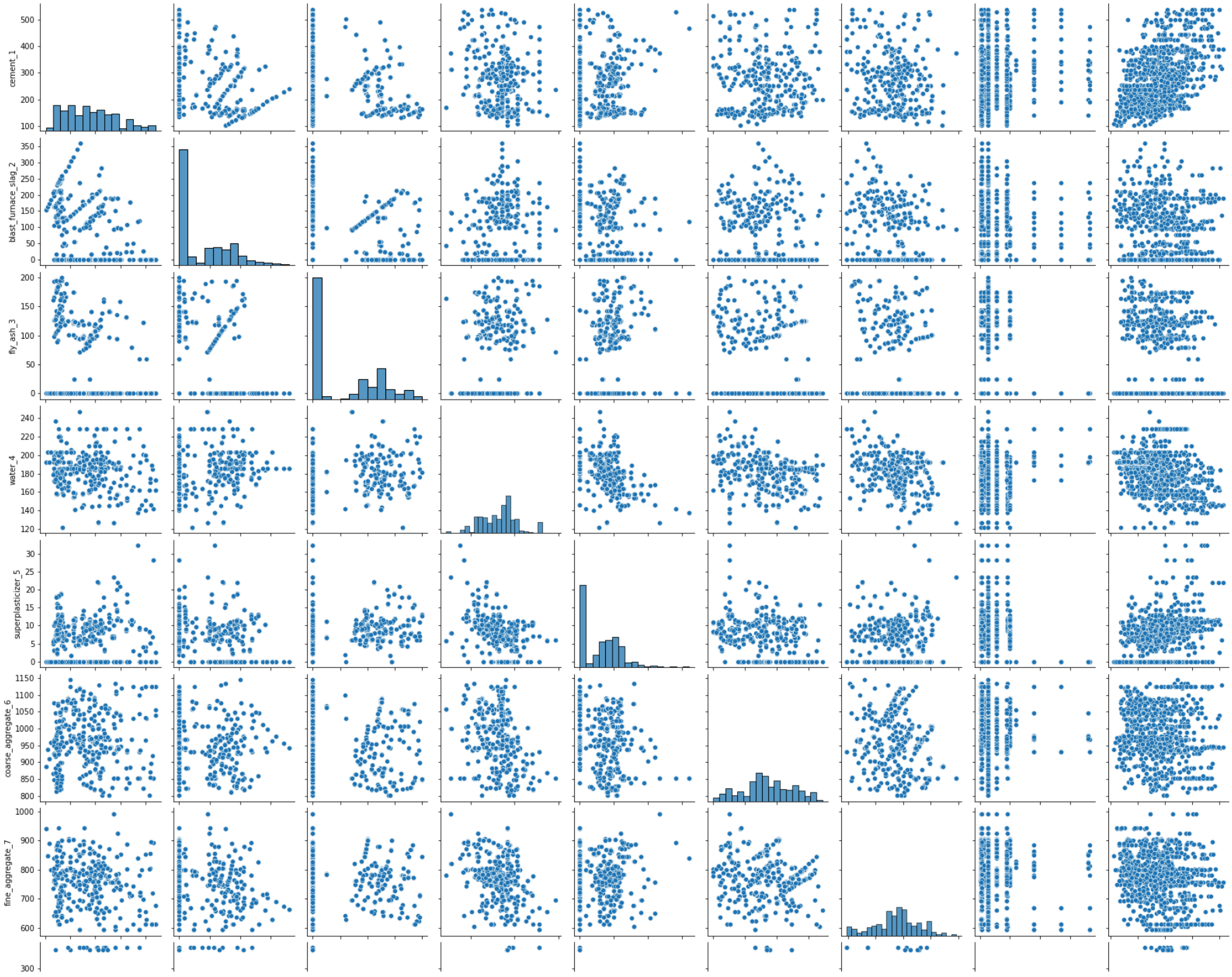
- We can see that the cement component has a moderate positive correlation with the response variable (0.497833)
- Some of the predictors are slightly negatively correlated with the response variable like water (-0.289613), fly_ash, coarse_aggregate and fine_aggregate

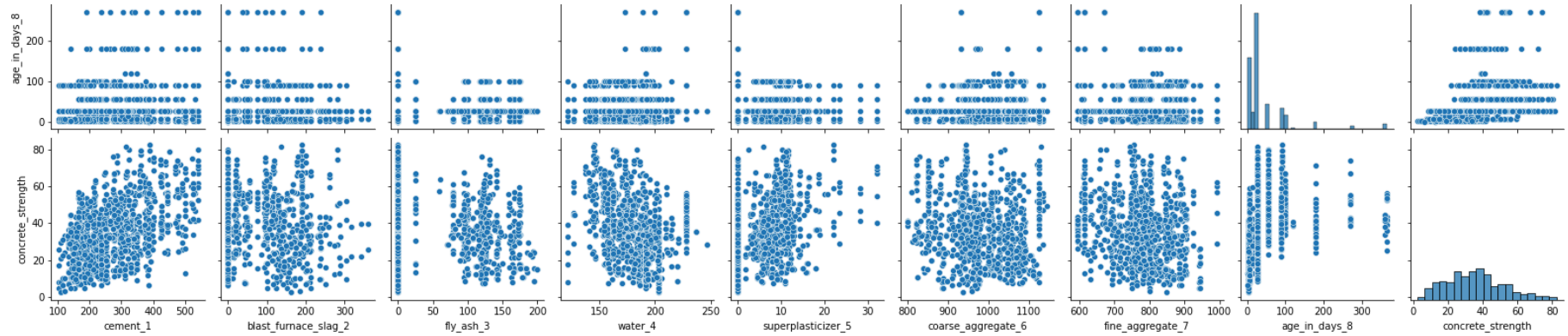
Summary of correlation

- Concrete strength increases mainly with cement, superplasticizer, and age.
- It decreases with water content, fly_ash, coarse_aggregate, and fine_aggregate.

- Few predictors have some degree of multicollinearity
    1. Water and Superplasticizer
    2. Cement and (fly_ash / blast_furnace_slag)

Pair Plot


cement_1
blast_furnace_slag_2
fly_ash_3
water_4
superplasticizer_5
coarse_aggregate_6
fine_aggregate_7



- Same as correlation matrix, we can see that most of the predictors do not have a linear relationship with the response variable except Cement component
- The cement component has a moderate positive correlation with concrete_strength
- And the predictors "Water and Superplasticizer" have some degree of multicollinearity, as discussed above

Variance Inflation factor

	Feature	VIF
0	const	6732.373793
1	cement_1	7.488657
2	blast_furnace_slag_2	7.276529
3	fly_ash_3	6.171455
4	water_4	7.004663
5	superplasticizer_5	2.965297
6	coarse_aggregate_6	5.076044
7	fine_aggregate_7	7.005346
8	age_in_days_8	1.118357

- Several predictors have VIF > 5, especially cement_1, blast_furnace_slag_2, fly_ash_3, water_4, and fine_aggregate_7.
- This suggests moderate to strong multicollinearity among the above mentioned predictors.

Splitting the data into train and test

70% of the available data is split as training data and 30% as test data with seed = 598 for reproducibility

```
Shape of training data(features): (721, 8)
Shape of testing data(features): (309, 8)
Shape of training data(target): (721,)
Shape of testing data(target): (309,)
Sample training data:
```

	cement_1	blast_furnace_slag_2	fly_ash_3	water_4	superplasticizer_5	coarse_aggregate_6	fine_aggregate_7	age_in_days_8
176	379.50	151.20	0.00	153.90	15.90	1134.3	605.00	9
736	238.00	0.00	0.00	186.00	0.00	1119.0	789.00	
83	362.60	189.00	0.00	164.90	11.60	944.7	755.80	
233	213.72	98.05	24.51	181.71	6.86	1065.8	785.38	10
454	250.00	0.00	95.69	191.84	5.33	948.9	857.20	5

Model Selection and prediction

FULL Linear Model

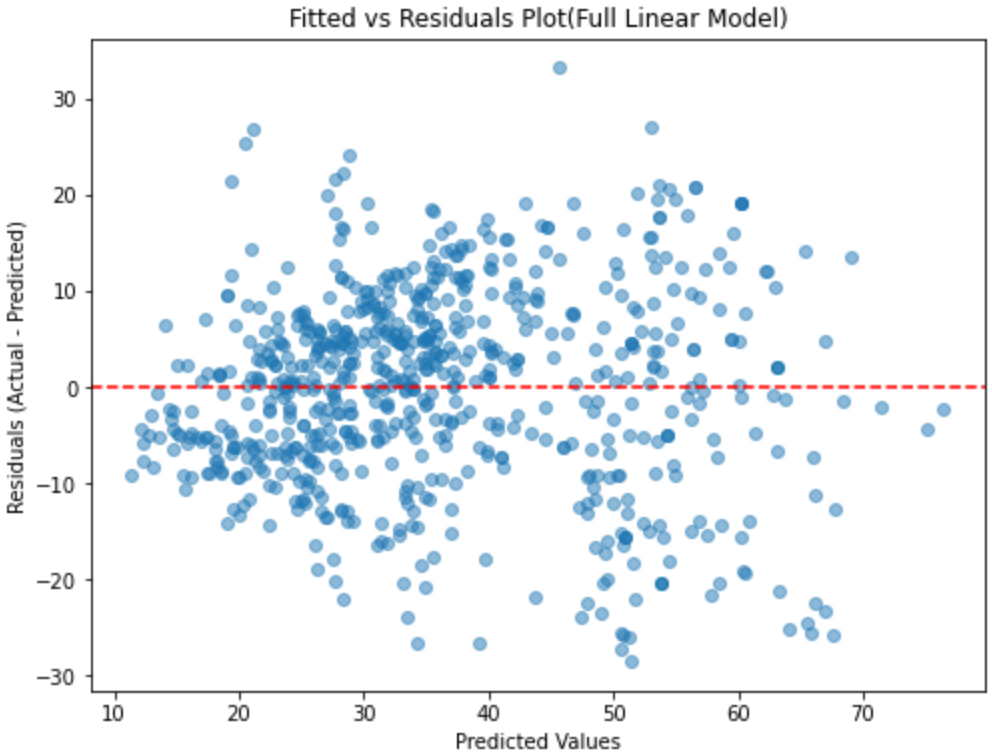
Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      concrete_strength   R-squared:                       0.621
Model:                            OLS   Adj. R-squared:                  0.616
Method:                 Least Squares   F-statistic:                     145.6
Date:                Tue, 21 Oct 2025   Prob (F-statistic):          2.75e-144
Time:                        13:08:45   Log-Likelihood:                -2699.8
No. Observations:                 721   AIC:                             5418.
Df Residuals:                     712   BIC:                             5459.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                    -16.4667     31.546     -0.522      0.602     -78.402      45.468
cement_1                   0.1160      0.010     11.601      0.000       0.096       0.136
blast_furnace_slag_2       0.1012      0.012      8.354      0.000       0.077       0.125
fly_ash_3                  0.0848      0.015      5.692      0.000       0.056       0.114
water_4                   -0.1488      0.047     -3.155      0.002      -0.241      -0.056
superplasticizer_5         0.2706      0.107      2.540      0.011       0.061       0.480
coarse_aggregate_6         0.0147      0.011      1.315      0.189      -0.007       0.037
fine_aggregate_7           0.0178      0.013      1.396      0.163      -0.007       0.043
age_in_days_8              0.1060      0.006     16.948      0.000       0.094       0.118
==============================================================================
Omnibus:                        4.164   Durbin-Watson:                   1.957
Prob(Omnibus):                  0.125   Jarque-Bera (JB):                4.185
Skew:                          -0.186   Prob(JB):                        0.123
Kurtosis:                       2.962   Cond. No.                     1.06e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.06e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- Adj R^2 = 0.616. The linear model explains about 61% of the variability in concrete strength — a moderately good fit.
- F-statistic = 145.6, p < 2.75e-144, this indicates the model as a whole is highly significant, so at least one predictor is related to the response.
- Cement, blast furnace slag, fly ash, water, superplasticizer, and age are significant predictors (p < 0.05).

- Coarse and fine aggregate are not significant (p > 0.05).
- Overall, the model is able to capture only ~61% of the vriability in the response variable.

Fitted vs Residual plot



- The residuals appear to scatter randomly around 0, linearity asssumption is satisfied.
- There is no obvious pattern in errors
- But the errors have slightly increasing variance (heteroscedasticity) as fitted values increase, the spread around 0 widens.
- We can add non-linear term and check if Adj.R^2 improves

Ramsey RESET Test

- The Ramsey RESET test checks whether the linear model is sufficient or if we are missing on non-linear or interaction terms
  - Null hypothesis (H0): The linear model is sufficient to explain the variance of the response variable.
  - Alternative hypothesis (H1): The model requires non linear or interaction terms

```
<Wald test (chi2): statistic=30.29997672082409, p-value=3.7013034634091795e-08, df_denom=1>
```

- Very small p_value = 3.7013034634091795e-08(<0.05) indicates that NULL Hypothesis can be rejected
- That is linear model is not sufficient and it requires polynomial or interaction terms to explain the variance of the response variable

Polynomial Regression(Degree =2)

Model Summary

OLS Regression Results

Dep. Variable:	concrete_strength	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	70.09
Date:	Tue, 21 Oct 2025	Prob (F-statistic):	1.00e-220
Time:	13:08:58	Log-Likelihood:	-2430.5
No. Observations:	721	AIC:	4951.
Df Residuals:	676	BIC:	5157.
Df Model:	44		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
1	-4967.4532	1473.659	-3.371	0.001	-7860.952	-2073.955
cement_1	3.0227	0.925	3.266	0.001	1.206	4.840
blast_furnace_slag_2	2.5078	1.092	2.296	0.022	0.363	4.652
fly_ash_3	1.9140	1.305	1.467	0.143	-0.647	4.475
water_4	14.9623	3.283	4.557	0.000	8.516	21.409
superplasticizer_5	29.6157	8.938	3.313	0.001	12.065	47.166
coarse_aggregate_6	2.8231	1.116	2.531	0.012	0.633	5.013
fine_aggregate_7	4.1489	1.254	3.310	0.001	1.687	6.610
age_in_days_8	0.0306	0.484	0.063	0.950	-0.919	0.980
cement_1^2	-0.0005	0.000	-2.928	0.004	-0.001	-0.000
cement_1 blast_furnace_slag_2	-0.0008	0.000	-2.052	0.041	-0.001	-3.28e-05
cement_1 fly_ash_3	-0.0004	0.000	-0.927	0.354	-0.001	0.000
cement_1 water_4	-0.0050	0.001	-5.240	0.000	-0.007	-0.003
cement_1 superplasticizer_5	-0.0104	0.003	-4.053	0.000	-0.015	-0.005
cement_1 coarse_aggregate_6	-0.0007	0.000	-2.096	0.036	-0.001	-4.71e-05
cement_1 fine_aggregate_7	-0.0011	0.000	-2.878	0.004	-0.002	-0.000
cement_1 age_in_days_8	0.0003	0.000	1.590	0.112	-6.8e-05	0.001
blast_furnace_slag_2^2	-0.0004	0.000	-1.539	0.124	-0.001	9.69e-05
blast_furnace_slag_2 fly_ash_3	-7.772e-05	0.001	-0.146	0.884	-0.001	0.001
blast_furnace_slag_2 water_4	-0.0042	0.001	-3.204	0.001	-0.007	-0.002
blast_furnace_slag_2 superplasticizer_5	-0.0094	0.003	-3.007	0.003	-0.016	-0.003
blast_furnace_slag_2 coarse_aggregate_6	-0.0007	0.000	-1.639	0.102	-0.001	0.000
blast_furnace_slag_2 fine_aggregate_7	-0.0009	0.000	-1.984	0.048	-0.002	-9.41e-06
blast_furnace_slag_2 age_in_days_8	0.0005	0.000	2.599	0.010	0.000	0.001
fly_ash_3^2	0.0002	0.000	0.624	0.533	-0.001	0.001
fly_ash_3 water_4	-0.0047	0.002	-3.080	0.002	-0.008	-0.002
fly_ash_3 superplasticizer_5	-0.0170	0.004	-4.166	0.000	-0.025	-0.009

```
fly_ash_3 coarse_aggregate_6              -0.0004      0.000     -0.826      0.409      -0.001       0.001
fly_ash_3 fine_aggregate_7                -0.0006      0.001     -1.065      0.287      -0.002       0.000
fly_ash_3 age_in_days_8                    0.0009      0.000      2.710      0.007       0.000       0.001
water_4^2                                 -0.0092      0.002     -4.250      0.000      -0.013      -0.005
water_4 superplasticizer_5                -0.0316      0.013     -2.438      0.015      -0.057      -0.006
water_4 coarse_aggregate_6                -0.0050      0.001     -4.103      0.000      -0.007      -0.003
water_4 fine_aggregate_7                  -0.0063      0.001     -4.446      0.000      -0.009      -0.003
water_4 age_in_days_8                  -9.852e-05      0.001     -0.119      0.905      -0.002       0.002
superplasticizer_5^2                      -0.0413      0.018     -2.249      0.025      -0.077      -0.005
superplasticizer_5 coarse_aggregate_6     -0.0093      0.003     -2.747      0.006      -0.016      -0.003
superplasticizer_5 fine_aggregate_7       -0.0122      0.003     -3.616      0.000      -0.019      -0.006
superplasticizer_5 age_in_days_8           0.0017      0.002      0.716      0.474      -0.003       0.006
coarse_aggregate_6^2                      -0.0004      0.000     -1.685      0.092      -0.001    6.13e-05
coarse_aggregate_6 fine_aggregate_7       -0.0011      0.000     -2.412      0.016      -0.002      -0.000
coarse_aggregate_6 age_in_days_8       -9.929e-05      0.000     -0.669      0.504      -0.000       0.000
fine_aggregate_7^2                        -0.0009      0.000     -3.437      0.001      -0.001      -0.000
fine_aggregate_7 age_in_days_8             0.0003      0.000      1.590      0.112    -7.7e-05       0.001
age_in_days_8^2                           -0.0006   4.71e-05    -13.016      0.000      -0.001      -0.001
==============================================================================
Omnibus:                        9.827   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.007   Jarque-Bera (JB):               15.519
Skew:                           0.013   Prob(JB):                     0.000427
Kurtosis:                       3.718   Cond. No.                     7.81e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.81e+09. This might indicate that there are
strong multicollinearity or other numerical problems.


Model Evaluation:
Polynomial regression Train MSE: 49.61370846919852
Polynomial regression Test MSE: 63.13994394780418


Polynomial regression Train RMSE: 7.043699913340895
Polynomial regression Test RMSE: 7.946064683087105


Polynomial regression R^2 on train: 0.8202069247418622
Polynomial regression R^2 on test : 0.7786585978319298
```

Interpretation of the Polynomial Regression results

- The train R^2 ~0.82, that is the model explains ~82% of the variation in Train data and ~77% of the variation in Test data.
- The model moderately generalizes well.(Train MSE ~49.6 and TEST MSE ~63.13)
- F-statistic = 70.09 and p-value is very small (1.00e-220), indicating the model is significant, and atleast one beta coefficient is not equal to zero.
- Though the model has got a better R^2 when compared to the linear model, the model complexity is very high with 44 parameters
- And the condition number is large, 7.81e+09 which indicates high multicollinearity
- To reduce the model complexity and to resolve multicollinearity, we can try polynomial ridge regression model

Polynomial Feature Standardization

	1	cement_1	blast_furnace_slag_2	fly_ash_3	water_4	superplasticizer_5	coarse_aggregate_6	fine_aggregate_7	age_in_d
0	0.0	0.911123	0.903200	-0.838329	-1.264101	1.562592	2.051352	-2.060177	0.6
1	0.0	-0.429040	-0.850950	-0.838329	0.215309	-1.022405	1.858670	0.165808	-0.6
2	0.0	0.751061	1.341738	-0.838329	-0.757138	0.863505	-0.336393	-0.235837	-0.6
3	0.0	-0.658999	0.286579	-0.452425	0.017593	0.092883	1.188691	0.122014	0.8
4	0.0	-0.315387	-0.850950	0.668284	0.484460	-0.155862	-0.283500	0.990874	0.1

5 rows × 45 columns

Polynomial Ridge Regression

```
Best lambda (Ridge): 0.001
Minmum and Maximum Lambda: 0.001 1000.0


Model Evaluation:
Ridge Train MSE: 49.790271426160736
Ridge Test MSE: 63.19918801947087


Ridge Train RMSE: 7.0562221780610574
Ridge Test RMSE: 7.94979169660884


Ridge R^2 on train: 0.8195670855121739
Ridge R^2 on test : 0.7784509136771309


Ridge Predictors and Coefficients:
```

	Predictors	Coefficients
0	1	0.000000
1	cement_1	176.804857
2	blast_furnace_slag_2	84.604784
3	fly_ash_3	8.358780
4	water_4	222.084990
5	superplasticizer_5	151.770477
6	coarse_aggregate_6	92.861230
7	fine_aggregate_7	187.603111
8	age_in_days_8	4.318149
9	cement_1^2	-17.519062
10	cement_1 blast_furnace_slag_2	-6.491799
11	cement_1 fly_ash_3	2.193753
12	cement_1 water_4	-71.812684
13	cement_1 superplasticizer_5	-21.980602
14	cement_1 coarse_aggregate_6	-25.287506
15	cement_1 fine_aggregate_7	-45.899882
16	cement_1 age_in_days_8	6.523282
17	blast_furnace_slag_2^2	-1.816626
18	blast_furnace_slag_2 fly_ash_3	2.923244
19	blast_furnace_slag_2 water_4	-41.385375
20	blast_furnace_slag_2 superplasticizer_5	-6.949228
21	blast_furnace_slag_2 coarse_aggregate_6	-8.823268
22	blast_furnace_slag_2 fine_aggregate_7	-18.344331

	Predictors	Coefficients
23	blast_furnace_slag_2 age_in_days_8	3.604376
24	fly_ash_3^2	5.702371
25	fly_ash_3 water_4	-31.687975
26	fly_ash_3 superplasticizer_5	-9.665972
27	fly_ash_3 coarse_aggregate_6	15.088761
28	fly_ash_3 fine_aggregate_7	7.145891
29	fly_ash_3 age_in_days_8	2.672384
30	water_4^2	-54.288289
31	water_4 superplasticizer_5	-24.854195
32	water_4 coarse_aggregate_6	-74.920243
33	water_4 fine_aggregate_7	-70.582813
34	water_4 age_in_days_8	-2.351143
35	superplasticizer_5^2	-4.750986
36	superplasticizer_5 coarse_aggregate_6	-42.612579
37	superplasticizer_5 fine_aggregate_7	-51.750692
38	superplasticizer_5 age_in_days_8	0.482565
39	coarse_aggregate_6^2	-10.562303
40	coarse_aggregate_6 fine_aggregate_7	-38.912905
41	coarse_aggregate_6 age_in_days_8	-7.402839
42	fine_aggregate_7^2	-69.601407
43	fine_aggregate_7 age_in_days_8	15.205786
44	age_in_days_8^2	-13.012625

- In the above ridge regression, I see that RidgeCV always picks the smallest alpha in the logspace.
- That is if the logspace is
  - alphas = np.logspace(-1, 1, 100) - best alpha was 0.1 (Minmum and Maximum Lambda: 0.1 10.0)
  - alphas = np.logspace(-2, 2, 100) - best alpha was 0.01 (Minmum and Maximum Lambda: 0.01 100.0)
  - alphas = np.logspace(-3, 3, 100) - best alpha was 0.001 (Minmum and Maximum Lambda: 0.001 1000.0)
- **It means that the model doesn't require regularization, i.e., Ridge is behaving like OLS because the data doesn't benefit from penalizing coefficients for its complexity.**
- **And the R^2 of the ridge polynomial regression is close to the R^2 of the normal polynomial regression with degree 2. That is both the models can explain ~80% of the variation in concrete compressive strength**
- **Although VIF analysis indicated moderate multicollinearity among a few predictors, the RidgeCV results suggest that multicollinearity does not affect model performance.**

Orthogonal Polynomial regression

- Apart from Ridge regression this is another way to solve multicollinearity that occurs in polynomial regression,which is using orthogonal polynomial regression.

Model Summary

OLS Regression Results

Dep. Variable:	concrete_strength	R-squared:	0.890
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	64.86
Date:	Tue, 21 Oct 2025	Prob (F-statistic):	2.71e-257
Time:	13:09:22	Log-Likelihood:	-2252.8
No. Observations:	721	AIC:	4668.
Df Residuals:	640	BIC:	5039.
Df Model:	80		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
1_Legendre_0	2.605e+12	3.06e+12	0.852	0.395	-3.4e+12	8.61e+12
1_Legendre_1	5.684e+08	6.67e+08	0.852	0.395	-7.42e+08	1.88e+09
1_Legendre_2	-1.022e+12	1.2e+12	-0.852	0.395	-3.38e+12	1.33e+12
cement_1_Legendre_0	2.818e+12	3.31e+12	0.852	0.395	-3.68e+12	9.31e+12
cement_1_Legendre_1	4.14e+11	4.86e+11	0.852	0.395	-5.4e+11	1.37e+12
cement_1_Legendre_2	5.143e+10	6.04e+10	0.852	0.395	-6.71e+10	1.7e+11
blast_furnace_slag_2_Legendre_0	1.459e+12	1.71e+12	0.852	0.395	-1.9e+12	4.82e+12
blast_furnace_slag_2_Legendre_1	3.37e+11	3.96e+11	0.852	0.395	-4.4e+11	1.11e+12
blast_furnace_slag_2_Legendre_2	1.32e+11	1.55e+11	0.852	0.395	-1.72e+11	4.36e+11
fly_ash_3_Legendre_0	2.227e+12	2.61e+12	0.852	0.395	-2.91e+12	7.36e+12
fly_ash_3_Legendre_1	3.298e+11	3.87e+11	0.852	0.395	-4.3e+11	1.09e+12
fly_ash_3_Legendre_2	1.311e+11	1.54e+11	0.852	0.395	-1.71e+11	4.33e+11

water_4_Legendre_0	2.226e+12	2.61e+12	0.852	0.395	-2.9e+12	7.36e+12
water_4_Legendre_1	9.918e+10	1.16e+11	0.852	0.395	-1.29e+11	3.28e+11
water_4_Legendre_2	3.956e+09	4.64e+09	0.852	0.395	-5.16e+09	1.31e+10
superplasticizer_5_Legendre_0	1.803e+12	2.12e+12	0.852	0.395	-2.35e+12	5.96e+12
superplasticizer_5_Legendre_1	2.046e+11	2.4e+11	0.852	0.395	-2.67e+11	6.76e+11
superplasticizer_5_Legendre_2	6.671e+10	7.83e+10	0.852	0.395	-8.7e+10	2.2e+11
coarse_aggregate_6_Legendre_0	1.84e+12	2.16e+12	0.852	0.395	-2.4e+12	6.08e+12
coarse_aggregate_6_Legendre_1	3.053e+11	3.58e+11	0.852	0.395	-3.98e+11	1.01e+12
coarse_aggregate_6_Legendre_2	8.318e+09	9.76e+09	0.852	0.395	-1.09e+10	2.75e+10
fine_aggregate_7_Legendre_0	1.732e+12	2.03e+12	0.852	0.395	-2.26e+12	5.72e+12
fine_aggregate_7_Legendre_1	-1.681e+11	1.97e+11	-0.852	0.395	-5.56e+11	2.19e+11
fine_aggregate_7_Legendre_2	-5.976e+09	7.01e+09	-0.852	0.395	-1.97e+10	7.8e+09
age_in_days_8_Legendre_0	1.653e+12	1.94e+12	0.852	0.395	-2.16e+12	5.46e+12
age_in_days_8_Legendre_1	9.705e+10	1.14e+11	0.852	0.395	-1.27e+11	3.21e+11
age_in_days_8_Legendre_2	4.55e+10	5.34e+10	0.852	0.395	-5.94e+10	1.5e+11
cement_1^2_Legendre_0	1.917e+12	2.25e+12	0.852	0.395	-2.5e+12	6.34e+12
cement_1^2_Legendre_1	-4.602e+11	5.4e+11	-0.852	0.395	-1.52e+12	6.01e+11
cement_1^2_Legendre_2	3.3654	0.742	4.535	0.000	1.908	4.823
cement_1 blast_furnace_slag_2_Legendre_0	5.223e+12	6.13e+12	0.852	0.395	-6.82e+12	1.73e+13
cement_1 blast_furnace_slag_2_Legendre_1	-25.9962	8.244	-3.153	0.002	-42.186	-9.807
cement_1 blast_furnace_slag_2_Legendre_2	1.7210	0.730	2.358	0.019	0.288	3.154

```
cement_1 fly_ash_3_Legendre_0                -7.899e+11   9.27e+11     -0.852      0.395   -2.61e+12    1.03e
+12
cement_1 fly_ash_3_Legendre_1                  -15.4593      7.385     -2.093      0.037     -29.961      -0.
958
cement_1 fly_ash_3_Legendre_2                    1.5523      0.792      1.959      0.051      -0.004       3.
108
cement_1 water_4_Legendre_0                  -7.762e+11   9.11e+11     -0.852      0.395   -2.57e+12    1.01e
+12
cement_1 water_4_Legendre_1                    -96.0546     21.042     -4.565      0.000    -137.374     -54.
735
cement_1 water_4_Legendre_2                      1.6348      1.279      1.279      0.202      -0.876       4.
145
cement_1 superplasticizer_5_Legendre_0       -7.818e+11   9.18e+11     -0.852      0.395   -2.58e+12    1.02e
+12
cement_1 superplasticizer_5_Legendre_1         -8.3525      7.695     -1.085      0.278     -23.462       6.
757
cement_1 superplasticizer_5_Legendre_2         -1.2467      0.553     -2.256      0.024      -2.332      -0.
162
cement_1 coarse_aggregate_6_Legendre_0       -7.838e+11    9.2e+11     -0.852      0.395   -2.59e+12    1.02e
+12
cement_1 coarse_aggregate_6_Legendre_1        -131.5536     36.430     -3.611      0.000    -203.090     -60.
017
cement_1 coarse_aggregate_6_Legendre_2           5.4411      1.445      3.765      0.000       2.603       8.
279
cement_1 fine_aggregate_7_Legendre_0          -7.84e+11    9.2e+11     -0.852      0.395   -2.59e+12    1.02e
+12
cement_1 fine_aggregate_7_Legendre_1           -95.6109     30.459     -3.139      0.002    -155.424     -35.
798
cement_1 fine_aggregate_7_Legendre_2             4.6946      1.403      3.345      0.001       1.939       7.
450
cement_1 age_in_days_8_Legendre_0            -7.767e+11   9.12e+11     -0.852      0.395   -2.57e+12    1.01e
+12
cement_1 age_in_days_8_Legendre_1              -2.6395      4.428     -0.596      0.551     -11.335       6.
056
cement_1 age_in_days_8_Legendre_2              -0.0372      0.249     -0.149      0.881      -0.527       0.
452
blast_furnace_slag_2^2_Legendre_0            -7.778e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.01e
+12
blast_furnace_slag_2^2_Legendre_1            -5.344e+11   6.27e+11     -0.852      0.395   -1.77e+12    6.97e
+11
blast_furnace_slag_2^2_Legendre_2                0.3637      0.503      0.723      0.470      -0.624       1.
351
```

```
blast_furnace_slag_2 fly_ash_3_Legendre_0           -7.79e+11   9.14e+11     -0.852      0.395   -2.57e+12    1.02e
+12
blast_furnace_slag_2 fly_ash_3_Legendre_1             -0.4319      3.012     -0.143      0.886      -6.346       5.
482
blast_furnace_slag_2 fly_ash_3_Legendre_2             -0.1205      0.219     -0.549      0.583      -0.551       0.
310
blast_furnace_slag_2 water_4_Legendre_0            -7.805e+11   9.16e+11     -0.852      0.395   -2.58e+12    1.02e
+12
blast_furnace_slag_2 water_4_Legendre_1              -61.6254     22.092     -2.789      0.005    -105.008     -18.
243
blast_furnace_slag_2 water_4_Legendre_2                3.0397      2.531      1.201      0.230      -1.931       8.
011
blast_furnace_slag_2 superplasticizer_5_Legendre_0 -7.785e+11   9.14e+11     -0.852      0.395   -2.57e+12    1.02e
+12
blast_furnace_slag_2 superplasticizer_5_Legendre_1    -5.0046      3.339     -1.499      0.134     -11.560       1.
551
blast_furnace_slag_2 superplasticizer_5_Legendre_2     0.3506      0.347      1.012      0.312      -0.330       1.
031
blast_furnace_slag_2 coarse_aggregate_6_Legendre_0 -7.785e+11   9.14e+11     -0.852      0.395   -2.57e+12    1.02e
+12
blast_furnace_slag_2 coarse_aggregate_6_Legendre_1   -83.4047     30.753     -2.712      0.007    -143.795     -23.
015
blast_furnace_slag_2 coarse_aggregate_6_Legendre_2     5.1959      3.147      1.651      0.099      -0.983      11.
375
blast_furnace_slag_2 fine_aggregate_7_Legendre_0   -7.777e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.01e
+12
blast_furnace_slag_2 fine_aggregate_7_Legendre_1     -31.4469     30.355     -1.036      0.301     -91.055      28.
161
blast_furnace_slag_2 fine_aggregate_7_Legendre_2       1.2102      2.566      0.472      0.637      -3.830       6.
250
blast_furnace_slag_2 age_in_days_8_Legendre_0       -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
blast_furnace_slag_2 age_in_days_8_Legendre_1          2.7699      1.687      1.642      0.101      -0.544       6.
083
blast_furnace_slag_2 age_in_days_8_Legendre_2         -0.2480      0.073     -3.382      0.001      -0.392      -0.
104
fly_ash_3^2_Legendre_0                             -7.781e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3^2_Legendre_1                             -4.535e+11   5.32e+11     -0.852      0.395    -1.5e+12    5.92e
+11
fly_ash_3^2_Legendre_2                                -0.1025      0.674     -0.152      0.879      -1.426       1.
221
```

```
fly_ash_3 water_4_Legendre_0                -7.779e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3 water_4_Legendre_1                  -32.8877     18.886     -1.741      0.082     -69.974       4.
199
fly_ash_3 water_4_Legendre_2                   -2.3885      2.269     -1.053      0.293      -6.844       2.
067
fly_ash_3 superplasticizer_5_Legendre_0     -7.779e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3 superplasticizer_5_Legendre_1        -6.6912      2.870     -2.332      0.020     -12.326      -1.
056
fly_ash_3 superplasticizer_5_Legendre_2         0.1898      0.340      0.559      0.577      -0.477       0.
857
fly_ash_3 coarse_aggregate_6_Legendre_0      -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3 coarse_aggregate_6_Legendre_1       -16.9047     30.854     -0.548      0.584     -77.491      43.
682
fly_ash_3 coarse_aggregate_6_Legendre_2        -5.5252      3.936     -1.404      0.161     -13.255       2.
204
fly_ash_3 fine_aggregate_7_Legendre_0       -7.779e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3 fine_aggregate_7_Legendre_1          21.3929     27.140      0.788      0.431     -31.901      74.
687
fly_ash_3 fine_aggregate_7_Legendre_2          -6.9136      2.977     -2.322      0.021     -12.760      -1.
068
fly_ash_3 age_in_days_8_Legendre_0           -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fly_ash_3 age_in_days_8_Legendre_1             3.6385      1.213      2.999      0.003       1.256       6.
021
fly_ash_3 age_in_days_8_Legendre_2             -0.7528      0.144     -5.210      0.000      -1.037      -0.
469
water_4^2_Legendre_0                         -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
water_4^2_Legendre_1                        -1.001e+11   1.18e+11     -0.852      0.395   -3.31e+11    1.31e
+11
water_4^2_Legendre_2                           4.9933      1.524      3.277      0.001       2.001       7.
986
water_4 superplasticizer_5_Legendre_0        -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
water_4 superplasticizer_5_Legendre_1          -29.8408     15.478     -1.928      0.054     -60.234       0.
552
water_4 superplasticizer_5_Legendre_2          -3.2726      2.081     -1.572      0.116      -7.360       0.
815
```

```
water_4 coarse_aggregate_6_Legendre_0              -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
water_4 coarse_aggregate_6_Legendre_1              -341.1010     65.094     -5.240      0.000    -468.926    -213.
276
water_4 coarse_aggregate_6_Legendre_2                9.9030      2.728      3.630      0.000       4.546      15.
260
water_4 fine_aggregate_7_Legendre_0                -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
water_4 fine_aggregate_7_Legendre_1                 -43.3845     36.019     -1.205      0.229    -114.113      27.
344
water_4 fine_aggregate_7_Legendre_2                  -1.1263      1.401     -0.804      0.422      -3.877       1.
624
water_4 age_in_days_8_Legendre_0                   -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
water_4 age_in_days_8_Legendre_1                    -46.4748     12.887     -3.606      0.000    -71.780     -21.
169
water_4 age_in_days_8_Legendre_2                     4.6866      1.289      3.635      0.000       2.155       7.
218
superplasticizer_5^2_Legendre_0                    -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
superplasticizer_5^2_Legendre_1                   -3.462e+11   4.06e+11     -0.852      0.395   -1.14e+12    4.52e
+11
superplasticizer_5^2_Legendre_2                      0.3775      0.391      0.965      0.335      -0.391       1.
146
superplasticizer_5 coarse_aggregate_6_Legendre_0   -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
superplasticizer_5 coarse_aggregate_6_Legendre_1    -34.4155     21.512     -1.600      0.110    -76.658       7.
827
superplasticizer_5 coarse_aggregate_6_Legendre_2     -2.4998      2.686     -0.931      0.352      -7.774       2.
774
superplasticizer_5 fine_aggregate_7_Legendre_0     -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
superplasticizer_5 fine_aggregate_7_Legendre_1      -54.9594     21.316     -2.578      0.010    -96.817     -13.
102
superplasticizer_5 fine_aggregate_7_Legendre_2       1.7664      1.748      1.010      0.313      -1.667       5.
200
superplasticizer_5 age_in_days_8_Legendre_0        -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
superplasticizer_5 age_in_days_8_Legendre_1          0.7886      1.102      0.716      0.474      -1.375       2.
952
superplasticizer_5 age_in_days_8_Legendre_2          -0.3505      0.094     -3.739      0.000      -0.535      -0.
166
```

```
coarse_aggregate_6^2_Legendre_0                    -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
coarse_aggregate_6^2_Legendre_1                   -3.055e+11   3.59e+11     -0.852      0.395   -1.01e+12    3.99e
+11
coarse_aggregate_6^2_Legendre_2                      2.2491      2.249      1.000      0.318      -2.168       6.
666
coarse_aggregate_6 fine_aggregate_7_Legendre_0     -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
coarse_aggregate_6 fine_aggregate_7_Legendre_1      -51.5303     68.544     -0.752      0.452    -186.129      83.
069
coarse_aggregate_6 fine_aggregate_7_Legendre_2       -0.6874      2.249     -0.306      0.760      -5.105       3.
730
coarse_aggregate_6 age_in_days_8_Legendre_0        -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
coarse_aggregate_6 age_in_days_8_Legendre_1         -33.8774     14.377     -2.356      0.019     -62.109      -5.
646
coarse_aggregate_6 age_in_days_8_Legendre_2            4.7778      1.266      3.775      0.000       2.293       7.
263
fine_aggregate_7^2_Legendre_0                      -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fine_aggregate_7^2_Legendre_1                      1.664e+11   1.95e+11      0.852      0.395   -2.17e+11     5.5e
+11
fine_aggregate_7^2_Legendre_2                          4.6955      1.473      3.188      0.002       1.804       7.
587
fine_aggregate_7 age_in_days_8_Legendre_0          -7.78e+11   9.13e+11     -0.852      0.395   -2.57e+12    1.02e
+12
fine_aggregate_7 age_in_days_8_Legendre_1           -23.2215     11.084     -2.095      0.037     -44.988      -1.
455
fine_aggregate_7 age_in_days_8_Legendre_2              1.9253      0.756      2.547      0.011       0.441       3.
410
age_in_days_8^2_Legendre_0                        -7.443e+11   8.74e+11     -0.852      0.395   -2.46e+12    9.71e
+11
age_in_days_8^2_Legendre_1                        -3.424e+11   4.02e+11     -0.852      0.395   -1.13e+12    4.47e
+11
age_in_days_8^2_Legendre_2                             1.9384      0.227      8.536      0.000       1.492       2.
384
==============================================================================
Omnibus:                       25.061   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.391
Skew:                           0.116   Prob(JB):                     3.45e-13
Kurtosis:                       4.362   Cond. No.                     5.62e+16
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.1e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.


Model Evaluation:
Orthogonal Polynomial regression Train MSE: 30.298050507992755
Orthogonal Polynomial regression Test MSE: 48.54606903555299


Orthogonal Polynomial regression Train RMSE: 5.504366494701525
Orthogonal Polynomial regression Test RMSE: 6.967500917513609


Orthogonal Polynomial regression R^2 on train: 0.8902041422978837
Orthogonal Polynomial regression R^2 on test : 0.8298184268430774
```

Observations from the Orthogonal polynomial results

- The model is explaining ~89% of the variability of concrete_strength on train data and ~82% of the variability in test data.
- This is a strong fit, indicating that the combination of features (linear, quadratic, interaction, and Legendre terms) successfully captures the trend in the training data.
- The condition number is very high 5.62e+16. This indicate that there are strong multicollinearity in the design matrix
- And the model is very complex and not interpretable

SPLINES - Piecewise polynomial regression

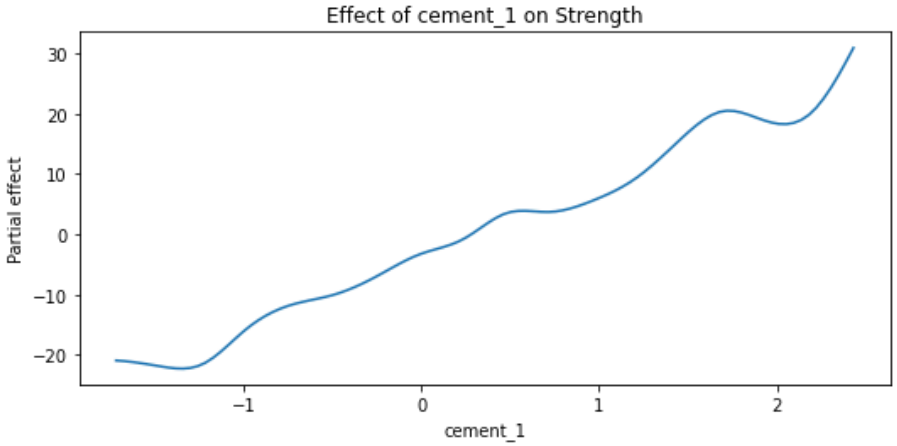
- Splines allow model to be **locally flexible** without overfitting globally
- Splines are piecewise polynomial joined at specific points called knots
- Since concrete strength is nonlinear with its features, data modeling with splines may help
- Instead of manually choosing degree of spline and the number of knots and applying a ridge penalty, we can use PyGAM
- pyGAM (Generalized Additive Models) automatically
  - Fits smooth spline functions for each predictor.(By default, **pyGAM uses cubic splines degree=3**)
  - Tunes the smoothing penalty lambda automatically using cross-validation.

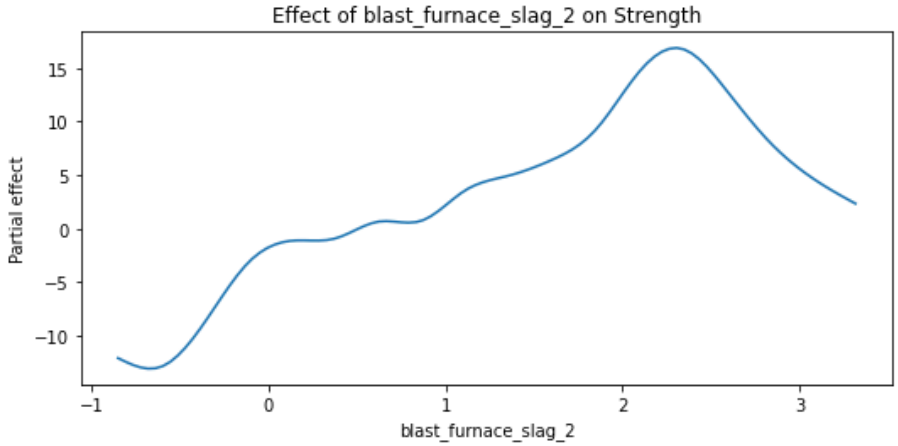
- That is **pyGAM** can be used for **Splinetransformer + RidgeCV** (which automatically chooses the number of knots and penalty)
- Since GAM uses penalty term to control the overall smoothness of the function, **the predictors has to be standardized**, so that penalty is applied across all predictors evenly.

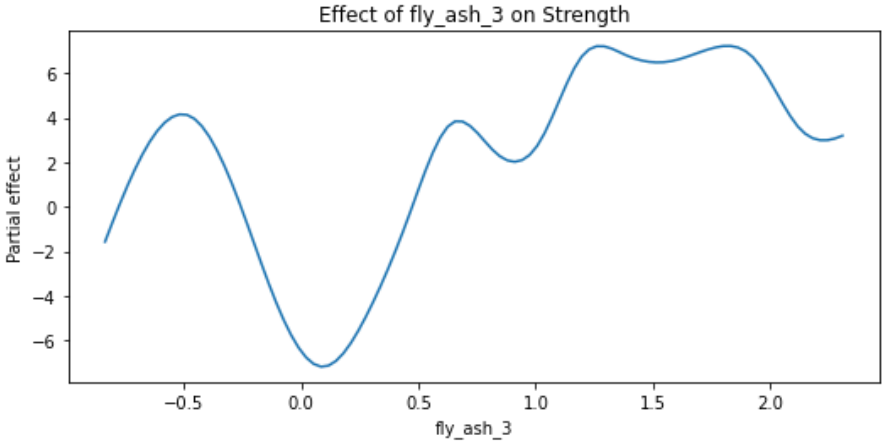
```
  0% (0 of 11) |                        | Elapsed Time: 0:00:00 ETA:  --:--:--
  9% (1 of 11) |##                      | Elapsed Time: 0:00:00 ETA:   0:00:01
 18% (2 of 11) |####                    | Elapsed Time: 0:00:00 ETA:   0:00:01
 27% (3 of 11) |######                  | Elapsed Time: 0:00:00 ETA:   0:00:01
 36% (4 of 11) |########                | Elapsed Time: 0:00:00 ETA:   0:00:01
 45% (5 of 11) |##########              | Elapsed Time: 0:00:00 ETA:   0:00:00
 54% (6 of 11) |############            | Elapsed Time: 0:00:00 ETA:   0:00:00
 63% (7 of 11) |##############          | Elapsed Time: 0:00:01 ETA:   0:00:00
 72% (8 of 11) |################        | Elapsed Time: 0:00:01 ETA:   0:00:00
 81% (9 of 11) |##################      | Elapsed Time: 0:00:01 ETA:   0:00:00
 90% (10 of 11) |####################   | Elapsed Time: 0:00:01 ETA:   0:00:00
100% (11 of 11) |######################| Elapsed Time: 0:00:02 Time:  0:00:02
Model Evaluation:
Spline GAM Train MSE: 21.561392496281627
Spline GAM Test MSE: 32.25018695590052


Spline GAM Train RMSE: 4.643424651728681
Spline GAM Test RMSE: 5.6789248063256235


Spline GAM R^2 on train: 0.921864557531294
Spline GAM R^2 on test : 0.8869447586633521
```

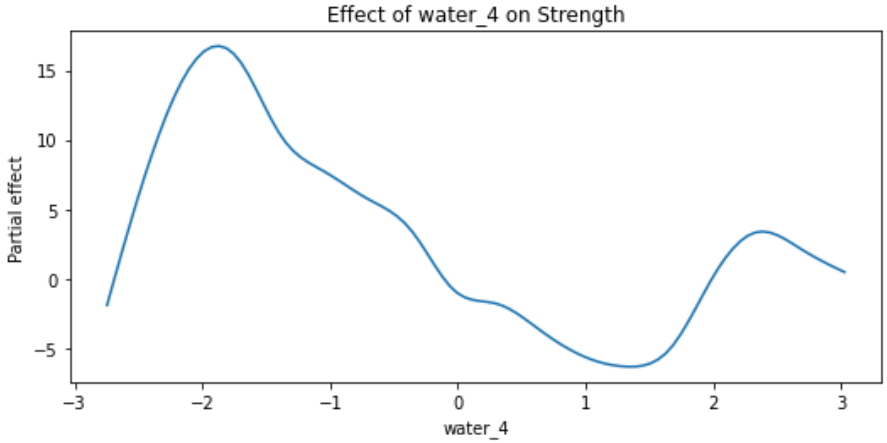
Observations from spline polynomial regression(with Ridge) using pyGAM

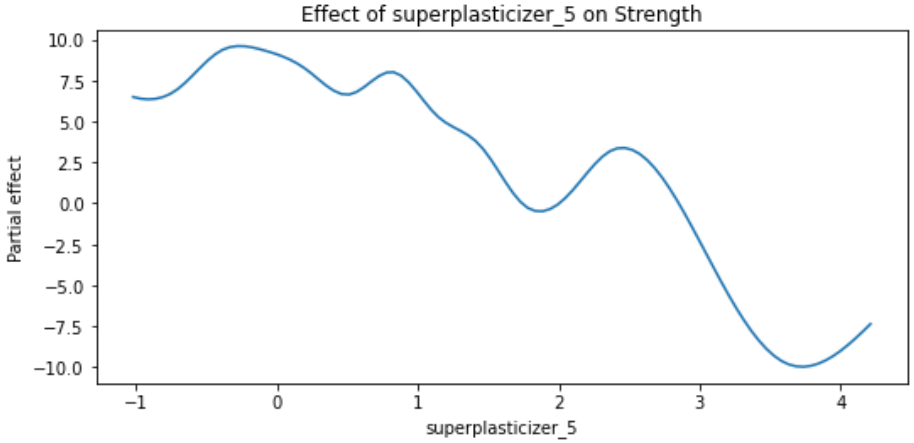
- The R^2 values (0.92 on train, 0.89 on test) indicate that GAM explains ~89% of the variance in the test data which is a strong predictive performance.
- Though there is a slight increase in test MSE and RMSE when compared to train MSE and RMSE, the model generalization is still strong.
- **Hence we can see that the GAM model was able to capture the non-linear relationship between the predictors and the response variable with local flexibility using splines.**

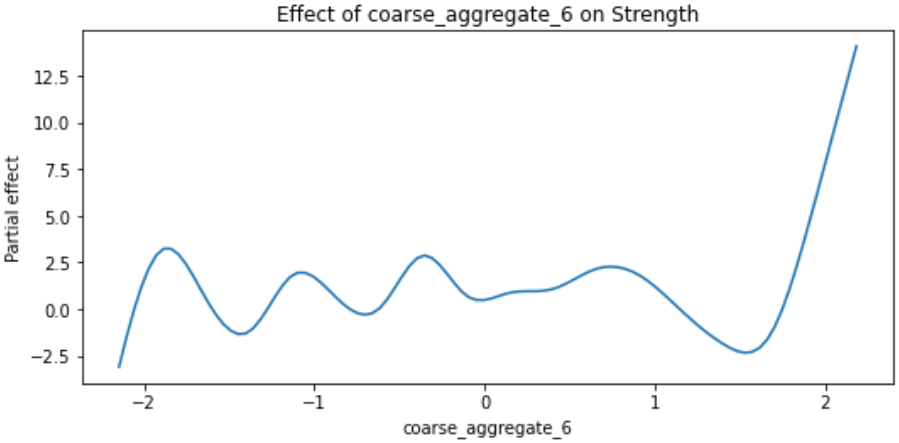

Effect of cement_1 on Strength
Partial effect
cement_1

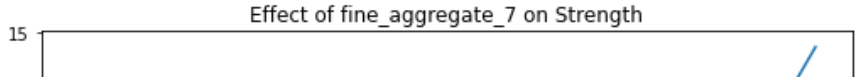

Effect of blast_furnace_slag_2 on Strength
Partial effect
blast_furnace_slag_2

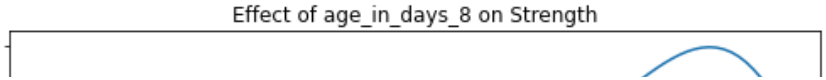

Effect of fly_ash_3 on Strength
Partial effect
fly_ash_3

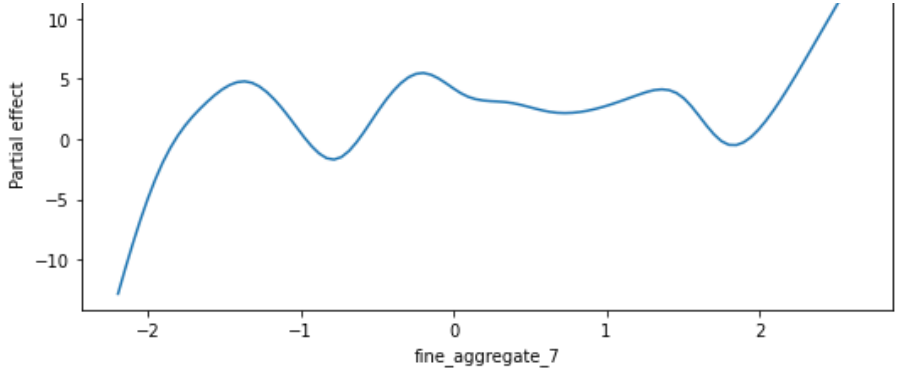

Effect of water_4 on Strength
Partial effect
water_4

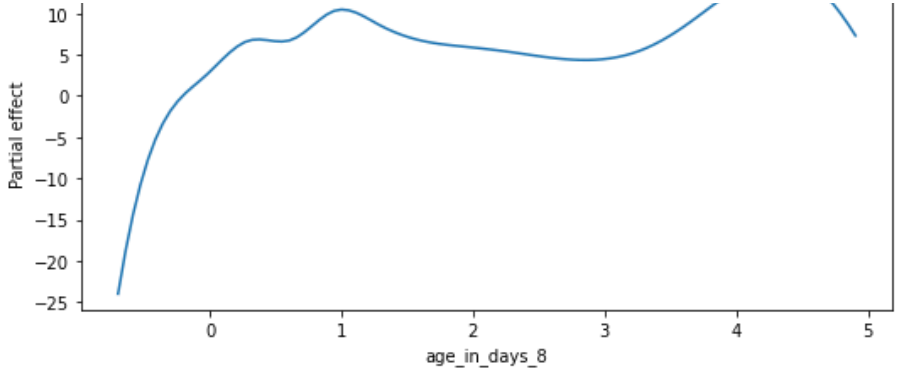

Effect of superplasticizer_5 on Strength
Partial effect
superplasticizer_5


Effect of coarse_aggregate_6 on Strength
Partial effect
coarse_aggregate_6


Effect of fine_aggregate_7 on Strength


Effect of age_in_days_8 on Strength





About the partial dependence plot

- Each plot corresponds to one predictor variable
- The x-axis represents the predictor's standardized value
- The y-axis represents the partial effect of each predictor on the predicted response, with all the other variables constant

Observations from the partial dependence plot

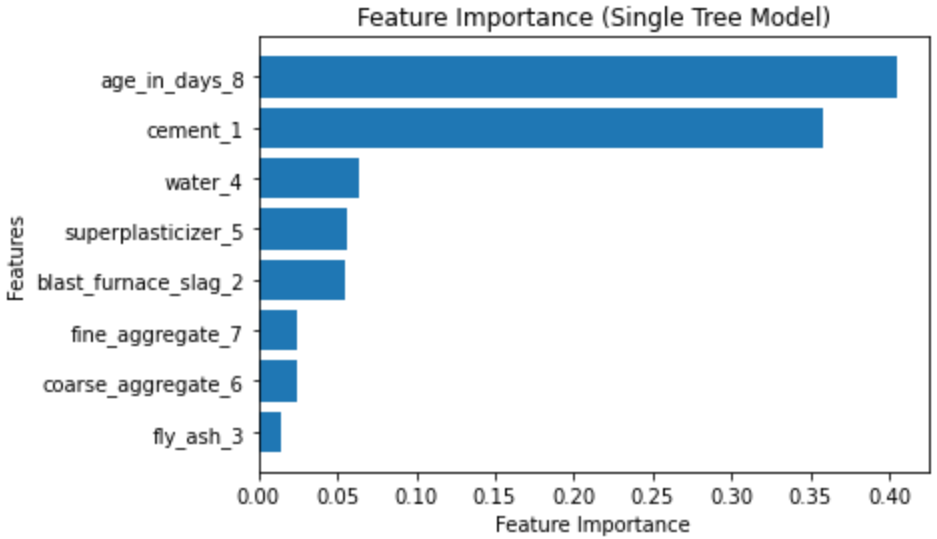
- As Cement content increases, the concrete compression strength rises
- Small content of Superplasticizer implies moderate concrete compression strength, but increasing the Superplasticizer beyond a level decreases the concrete compression strength
- As the number of days increases the concrete compression strength increases and the graph flatten as the days increases further
- For smaller values of water the concrete compression strength is good, but as the water level increases the concrete compression strength decreases

Tree based Models

Prediction using DecisionTreeRegressor (max_depth=6)

```
Model Evaluation:
Single Tree Train MSE: 31.54487289511842
Single Tree Test MSE: 69.99670815569486


Single Tree Train RMSE: 5.6164822527199725
Single Tree Test RMSE: 8.366403537703334


Single Tree R^2 on train: 0.8856858339875668
Single Tree R^2 on test : 0.754621740824819
```

Feature Importance for DecisionTreeRegressor model



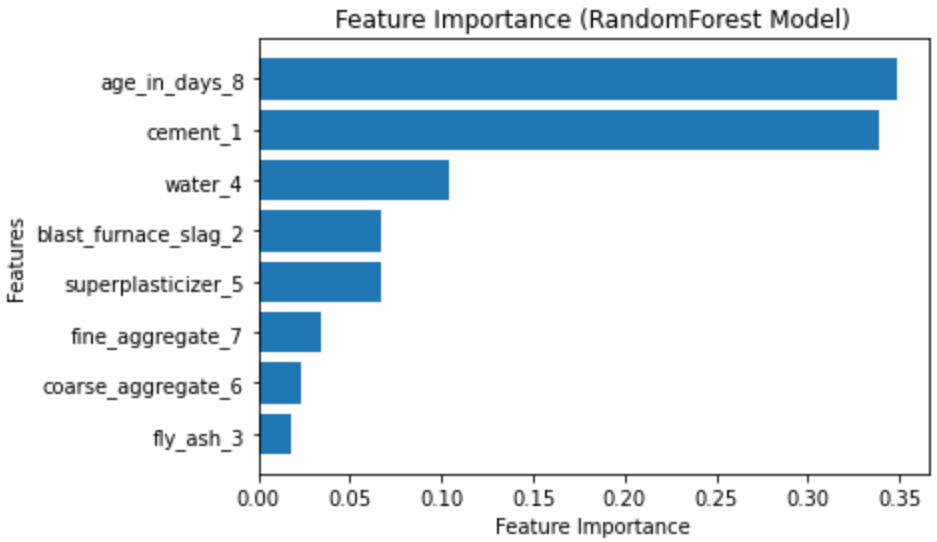
Observations from the results of Single Regression Tree

- The tree fits the training data well with R^2 ~0.89, for max_depth=6
- The test R^2 is ~0.77 which is less then the train R^2, but still the model generalizes reasonably well with slight overfitting.
- The prediction results can improve with different values of max_depth or min_samples_leaf
- From the feature importance plot, we can see that the top 3 predictors are "age_in_days" , "cement" and "water"

Prediction using RandomForestRegressor(Hyperparameters tuned using GridSearchCV)

```
Best Parameters: {'max_depth': 7, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators':
300}
Model Evaluation:
Random Forest Train MSE: 17.337294874157617
Random Forest Test MSE: 40.76697563964452


Random Forest Train RMSE: 4.1638077374150715
Random Forest Test RMSE: 6.384902163670523


Random Forest R^2 on train: 0.9371720909752768
Random Forest R^2 on test : 0.8570885720505255
```

Feature Importance for RandomForestRegressor model



Observations from the Prediction result of RandomForestRegressor (using GridSearchCV)

- The tuned Random Forest model (with n_estimators = 200, max_depth = 7, max_features = 6, min_samples_split = 2, and min_samples_leaf = 1) shows strong predictive performance.
- The model explains 94.8%(Train R^2=0.9475) of the variance in the training data and 86.7%(Test R^2=0.8667) of the variance in the test data.
- The model fits the training data very well(Train MSE=14.4) and generalizes reasonably well to unseen data, with only a moderate increase in Test MSE(38.0233).
- Though the prediction results are better than Single Tree model, in random forest, the interpretability of the model is less compared to Single tree model. This is because, each tree has its own structure (different splits and thresholds) and the final prediction is an average of all trees.
- From the feature importance plot, we can see that the top 3 predictors are "age_in_days", "cement" and "water"

Model Comparison for Concrete Compressive Strength Prediction

Model	Train MSE	Test MSE	Train R^2	Test R^2	Observations
Polynomial Regression (Degree = 2)	49.61	63.14	0.820	0.779	Explains ~78% variance on test data. Has high multicollinearity and overfitting due to 44 predictors.
Polynomial Ridge Regression	49.79	63.20	0.820	0.778	Regularization didn't improve results — in different range of logspace, always smallest alpha was chosen, indicating Ridge behaves like OLS.
Orthogonal Polynomial Regression	30.30	48.55	0.890	0.830	Better generalization than regular polynomial; was still complex to interpret and had multicollinearity.
Spline Regression (PyGAM)	21.56	32.25	**0.922**	**0.887**	Strong nonlinear predictive performance, captures local trends; best smooth generalization among regression models.
Single Regression Tree (max_depth = 6)	31.54	72.08	0.886	0.747	Slight overfitting; less predictive performance when compared to spline; interpretable; key predictors were age, cement and water.
Random Forest (Best params via GridSearchCV)	14.47	38.02	**0.948**	**0.867**	strong predictive model; less interpretable but robust. key predictors were age, cement and water.

- Among all the models evaluated, **the Random Forest and Spline GAM models achieved the best predictive performance**, capturing the nonlinear relationships between concrete strength and its composition.
- The Spline GAM model was smooth, interpretable and captured nonlinear trends and performed comparably well with high Train and Test R^2. And the generalization was also good, with a small difference in Train and Test MSE.

- Random forest also provided good prediction performance but its interpretability was low when compared to spline GAM and single tree models.
- Simpler models like Polynomial regression and Single tree model, had moderate accuracy, but had mulitcollinearity and slight overfitting.
- Overall,
  - **The spline GAM model and**
  - **Random Forest model**
    had good predictive performance.
- **Spline GAM model was more interpretable than Random Forest and Spline GAM also had highest accuracy(R^2) on TEST data among rest of the models analyzed.**