Prediction of Concrete Compressive Strength

Morgan Houston

July 26, 2021

1 Introduction

The goal of this project is to apply the lessons learned through the Harvard edX Data Science program to a new dataset of our choosing.

I have selected the "Concrete Compressive Strength Data Set" from the UCI Machine Learning Repository as the data set for this project.

As described in the dataset description, "Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate" (Yeh 2007).

As stated by PCA, in "its simplest form, concrete is a mixture of past and aggregates, or rocks". Cement is the main ingredient in concrete, and when combined with water, forms a paste. The paste coats the surface of the aggregates in the mixture (fine and/or coarse), and when this paste hardens, it becomes concrete. The proportions of these ingredients are the key to producing a strong and durable concrete. The strength of the paste "depends on the ratio of water to cement" and "high-quality concrete is produced by lowering the water-cement ratio as much as possible" (PCA 2021). This-water cement ratio has been recognized since 1918 as a key predictor of concrete strength.

Superplasticizers, as discussed in "How Super are Superplasticizers", can be used in three ways: "(1) to create flowing, self-leveling concrete without increasing water, without reducing cement and without sacrificing strength; (2) to produce workable, high-strength concrete by reducing the water and thus the water- cement ratio; or (3) to save cement by reducing both the water and cement contents while maintaining the same water-cement ratio and the same workability" (Concrete Construction Staff 2021).

Blast furnace slag is the "nonmetallic product, consisting essentially of silicates and aluminosilicates of calcium and of other bases, that is developed in a molten condition simultaneously with iron in a blast furnace" (American Concrete Institute 2021). Fly Ash is "finely divided residue that results from the combustion of ground or powdered coal and that is transported by flue gases from the combustion zone to the particle removal system" (American Concrete Institute 2021). Both of these ingredients can be used to reduce the amount of cement required in the mixture.

Concrete compressive strength is the most common (and accepted) measurement of concrete strength. It "measures the ability of concrete to withstand loads that will decrease the size of the concrete" (Cor-Tuf 2021) and is measured in pounds per square inch. The minimum for a project usually starts at 2,500-3,000 psi (or 17.2-24.1 MPa). The standards (according to the American Concrete Institute), are that the specified compressive strength for the concrete will be based on the 28 day test results, unless otherwise specified in construction documents. Typically, tests at 3 days and 7 days are used to track early strength gain.

Through this project, I aim to explore the provided dataset and develop a model which uses these features of concrete to predict the compressive strength of the concrete mixture.

2 Methods/Analysis

2.1 Data Analysis

The "Concrete Compressive Strength Data Set" from the UCI Machine Learning Repository consists of eight quantitative input variables and one quantitative output variable, and a total of 1030 rows. The first 10 rows are presented below in Table 1.

Table 1: Concrete Compressive Strength Data Set

Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
540.0	0.0	0	162	2.5	1040.0	676.0	28	79.98611
540.0	0.0	0	162	2.5	1055.0	676.0	28	61.88737
332.5	142.5	0	228	0.0	932.0	594.0	270	40.26954
332.5	142.5	0	228	0.0	932.0	594.0	365	41.05278
198.6	132.4	0	192	0.0	978.4	825.5	360	44.29608
266.0	114.0	0	228	0.0	932.0	670.0	90	47.02985
380.0	95.0	0	228	0.0	932.0	594.0	365	43.69830
380.0	95.0	0	228	0.0	932.0	594.0	28	36.44777
266.0	114.0	0	228	0.0	932.0	670.0	28	45.85429
475.0	0.0	0	228	0.0	932.0	594.0	28	39.28979

Cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate are all recorded in the same units of kg in an m^3 mixture. Age is recorded in days and is the age of the concrete mixture when the compressive strength was tested. Concrete compressive strength is recorded in megapascals (MPa).

For convenience, I will rename the columns in the dataset as follows:

Table 2: Column Renaming

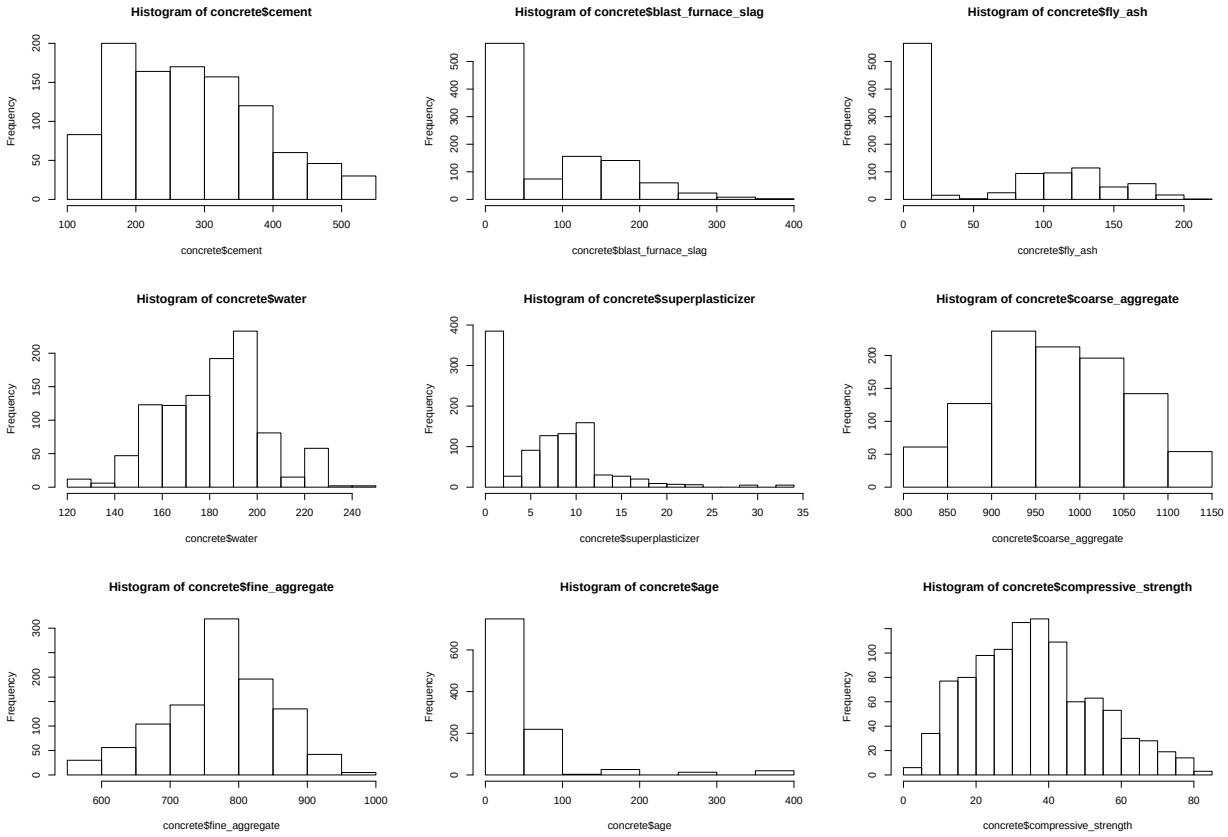
Original	Renamed
Cement (component 1)(kg in a m^3 mixture)	cement
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	blast_furnace_slag
Fly Ash (component 3)(kg in a m^3 mixture)	fly_ash
Water (component 4)(kg in a m^3 mixture)	water
Superplasticizer (component 5)(kg in a m^3 mixture)	superplasticizer
Coarse Aggregate (component 6)(kg in a m^3 mixture)	coarse_aggregate
Fine Aggregate (component 7)(kg in a m^3 mixture)	fine_aggregate
Age (day)	age
Concrete compressive strength(MPa, megapascals)	compressive_strength

The summary statistics for the dataset are presented in Table 3. They confirm that the dataset has no missing values, and that each variable within the dataset varies in scale.

Table 3: Concrete Summary Statistics

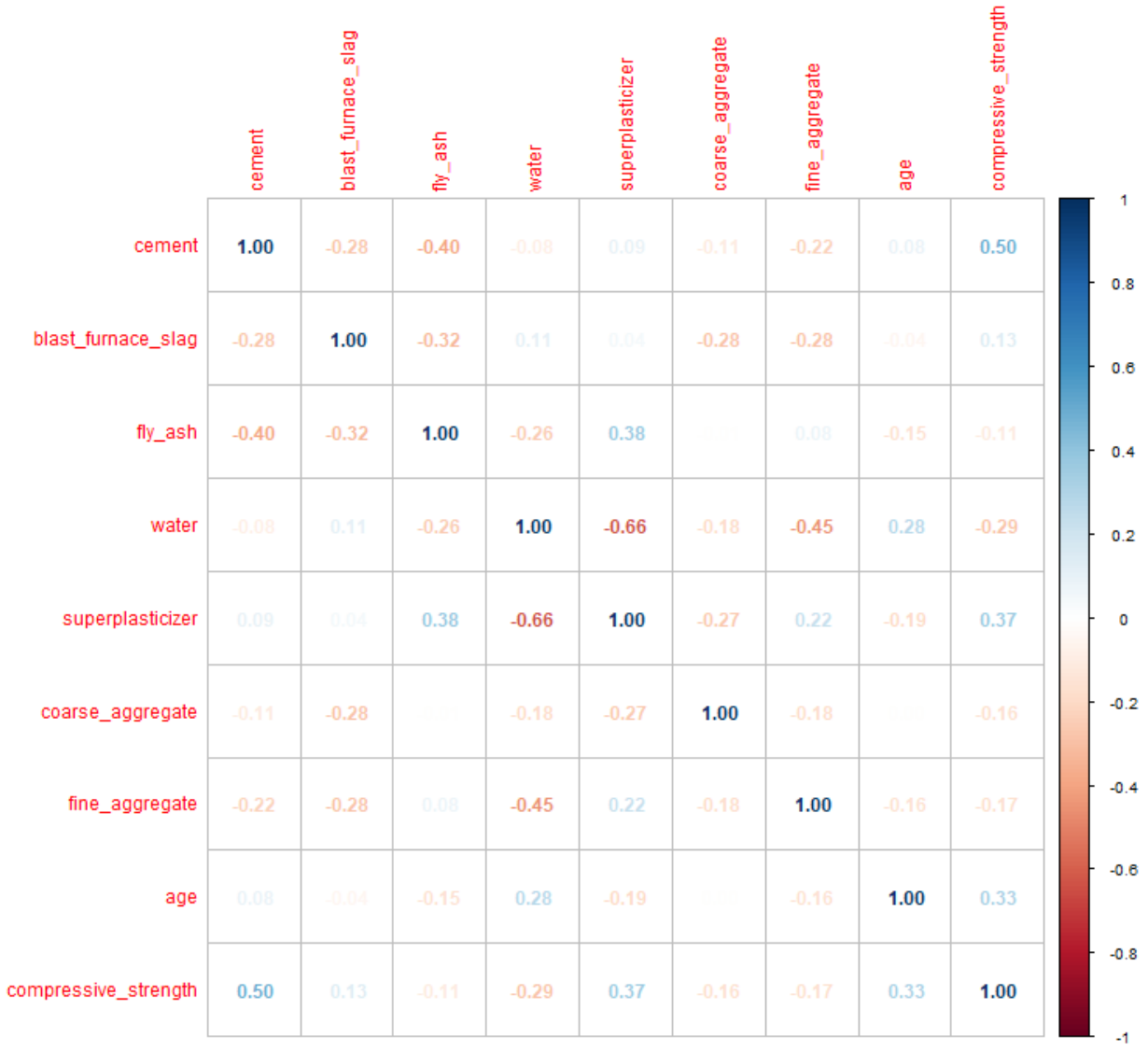
cement	Min. :102.0	1st Qu.:192.4	Median :272.9	Mean :281.2	3rd Qu.:350.0	Max. :540.0
blast_furnace_slag	Min. : 0.0	1st Qu.: 0.0	Median : 22.0	Mean : 73.9	3rd Qu.:142.9	Max. :359.4
fly_ash	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 54.19	3rd Qu.:118.27	Max. :200.10
water	Min. :121.8	1st Qu.:164.9	Median :185.0	Mean :181.6	3rd Qu.:192.0	Max. :247.0
superplasticizer	Min. : 0.000	1st Qu.: 0.000	Median : 6.350	Mean : 6.203	3rd Qu.:10.160	Max. :32.200
coarse_aggregate	Min. : 801.0	1st Qu.: 932.0	Median : 968.0	Mean : 972.9	3rd Qu.:1029.4	Max. :1145.0
fine_aggregate	Min. :594.0	1st Qu.:731.0	Median :779.5	Mean :773.6	3rd Qu.:824.0	Max. :992.6
age	Min. : 1.00	1st Qu.: 7.00	Median : 28.00	Mean : 45.66	3rd Qu.: 56.00	Max. :365.00
compressive_strength	Min. : 2.332	1st Qu.:23.707	Median :34.443	Mean :35.818	3rd Qu.:46.136	Max. :82.599

First, we will examine the distribution of each variable within the dataset.



Blast furnace slag, fly ash, superplasticizer, and age all show a left skew in the distribution. Cement, coarse aggregate, fine aggregate, and compressive strength appear to approximate a normal distribution.

We also examine the correlation between each of these variables by using the corr() function to create the correlation matrix for all variables within the dataset.

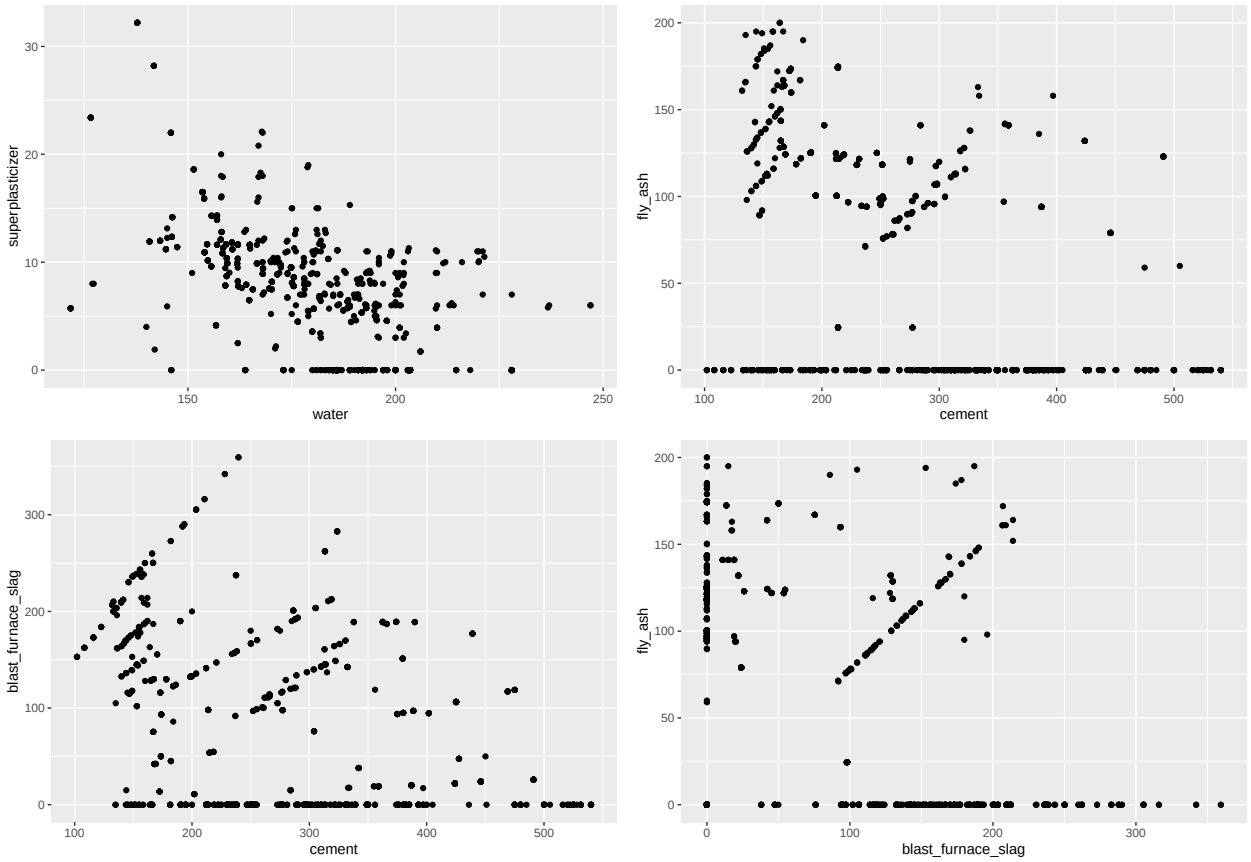


	cement	blast_furnace_slag	fly_ash	water	superplasticizer	coarse_aggregate	fine_aggregate	age	compressive_strength
cement	1.00	-0.28	-0.40	-0.08	0.09	-0.11	-0.22	0.08	0.50
blast_furnace_slag	-0.28	1.00	-0.32	0.11	0.04	-0.28	-0.28	-0.04	0.13
fly_ash	-0.40	-0.32	1.00	-0.26	0.38	-0.01	0.08	-0.15	-0.11
water	-0.08	0.11	-0.26	1.00	-0.66	-0.18	-0.45	0.28	-0.29
superplasticizer	0.09	0.04	0.38	-0.66	1.00	-0.27	0.22	-0.19	0.37
coarse_aggregate	-0.11	-0.28	-0.01	-0.18	-0.27	1.00	-0.18	-0.00	-0.16
fine_aggregate	-0.22	-0.28	0.08	-0.45	0.22	-0.18	1.00	-0.16	-0.17
age	0.08	-0.04	-0.15	0.28	-0.19	-0.00	-0.16	1.00	0.33
compressive_strength	0.50	0.13	-0.11	-0.29	0.37	-0.16	-0.17	0.33	1.00

The greatest correlation is between water and superplasticizer - since superplasticizer is used within concrete to reduce the amount of water required, this makes sense. Cement, fly ash, and blast furnace slag all have a weak correlation as well, which again makes sense given that we know that fly ash and blast furnace slag are both materials that can be used to reduce the amount of cement required in concrete. Fine aggregate and water also appear to have a weak correlation.

Looking at compressive strength, the dependent variable, the variable with the greatest correlation is cement, with superplasticizer and age the second and third, respectively. We note that this may indicate these are predictors of the concrete strength, and will explore this further in a later section.

We can visualize these relationships using the ggplot functionality.



An interesting feature of the blast furnace slag and fly ash relationship becomes apparent when graphed - we can see three distinct linear trends; two are present on the axes when either is zero, but there is a third clearly linear relationship when both variables are nonzero. This implies that when both ingredients are used in the concrete mixture, they are included in a standard proportion.

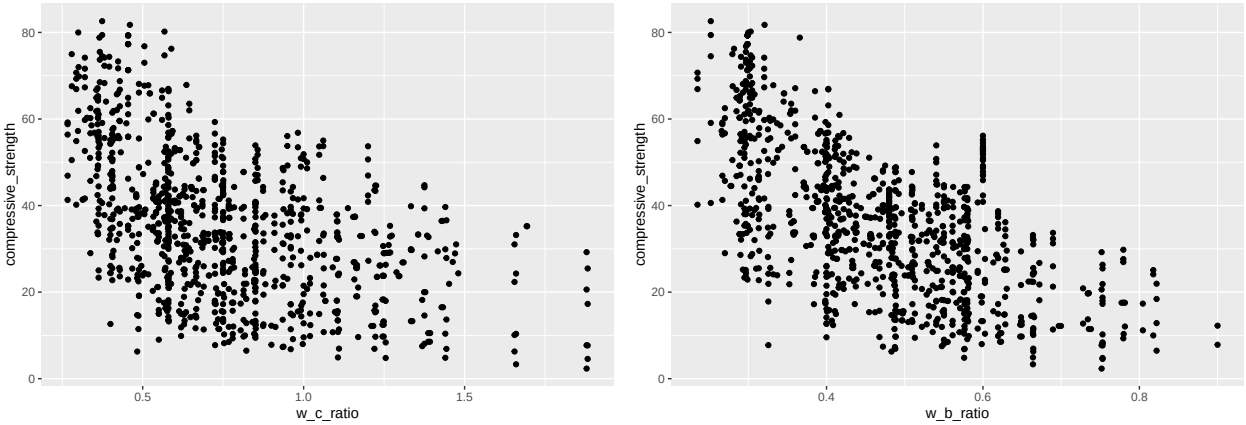
This dataset does not include the well-known water-cement ratio, which was presented by Duff Abrams in 1918, and states that the lower the water to cement ratio, the greater the strength of the concrete (American Concrete Institute 2021). We will calculate this using the water and cement features included in the dataset and adding a new column, w_c_ratio.

Additionally, we will extend this concept to calculate a water-binder ratio, where the binder includes cement, blast furnace slag, and fly ash, and add a new column, w_b_ratio.

Table 4: Concrete Summary Statistics with Added Features

cement	Min. :102.0	1st Qu.:192.4	Median :272.9	Mean :281.2	3rd Qu.:350.0	Max. :540.0
blast_furnace_slag	Min. : 0.0	1st Qu.: 0.0	Median : 22.0	Mean : 73.9	3rd Qu.:142.9	Max. :359.4
fly_ash	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 54.19	3rd Qu.:118.27	Max. :200.10
water	Min. :121.8	1st Qu.:164.9	Median :185.0	Mean :181.6	3rd Qu.:192.0	Max. :247.0
superplasticizer	Min. : 0.000	1st Qu.: 0.000	Median : 6.350	Mean : 6.203	3rd Qu.:10.160	Max. :32.200
coarse_aggregate	Min. : 801.0	1st Qu.: 932.0	Median : 968.0	Mean : 972.9	3rd Qu.:1029.4	Max. :1145.0
fine_aggregate	Min. :594.0	1st Qu.:731.0	Median :779.5	Mean :773.6	3rd Qu.:824.0	Max. :992.6
age	Min. : 1.00	1st Qu.: 7.00	Median : 28.00	Mean : 45.66	3rd Qu.: 56.00	Max. :365.00
compressive_strength	Min. : 2.332	1st Qu.:23.707	Median :34.443	Mean :35.818	3rd Qu.:46.136	Max. :82.599
w_c_ratio	Min. :0.2669	1st Qu.:0.5333	Median :0.6753	Mean :0.7483	3rd Qu.:0.9350	Max. :1.8824
additions	Min. : 0.00	1st Qu.: 94.11	Median :124.45	Mean :128.08	3rd Qu.:186.85	Max. :382.00
binder	Min. :200.0	1st Qu.:336.4	Median :391.4	Mean :409.2	3rd Qu.:483.7	Max. :640.0
w_b_ratio	Min. :0.2351	1st Qu.:0.3839	Median :0.4721	Mean :0.4692	3rd Qu.:0.5612	Max. :0.9000

We can visualize the relationships between these additional variables and concrete compressive strength.



2.2 Modeling

First, we will split the concrete dataset into a training dataset, which we will use to develop our model(s) on, and a test dataset, which we will use to test the performance of our model(s). We do this so that the test set is unknown to our model(s) and only the train set is used for learning. A generally accepted split is 80:20 train:test for larger datasets; as ours is just over 1000 rows, we will select a slightly larger split at 70:30 to allow our test set to be representative of the training set, and allow sufficient variability within the train set.

We can check the summary statistics for both the train and test sets to verify that they appear similar.

Table 5: Train Set - Summary Statistics

cement	Min. :102.0	1st Qu.:198.6	Median :277.0	Mean :284.8	3rd Qu.:358.2	Max. :540.0
blast_furnace_slag	Min. : 0.00	1st Qu.: 0.00	Median : 20.00	Mean : 72.59	3rd Qu.:142.20	Max. :359.40
fly_ash	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 53.51	3rd Qu.:118.16	Max. :200.10
water	Min. :121.8	1st Qu.:164.9	Median :185.7	Mean :181.6	3rd Qu.:192.9	Max. :247.0
superplasticizer	Min. : 0.000	1st Qu.: 0.000	Median : 6.165	Mean : 6.257	3rd Qu.:10.300	Max. :32.200
coarse_aggregate	Min. : 801.0	1st Qu.: 932.0	Median : 968.0	Mean : 970.6	3rd Qu.:1028.1	Max. :1145.0
fine_aggregate	Min. :594.0	1st Qu.:732.7	Median :780.1	Mean :774.8	3rd Qu.:825.0	Max. :992.6
age	Min. : 1.00	1st Qu.: 7.00	Median : 28.00	Mean : 44.66	3rd Qu.: 56.00	Max. :365.00
compressive_strength	Min. : 4.565	1st Qu.:23.707	Median :34.443	Mean :36.104	3rd Qu.:46.147	Max. :82.599
w_c_ratio	Min. :0.2669	1st Qu.:0.5221	Median :0.6667	Mean :0.7363	3rd Qu.:0.8798	Max. :1.8824
additions	Min. : 0.0	1st Qu.: 76.0	Median :123.4	Mean :126.1	3rd Qu.:187.0	Max. :382.0
binder	Min. :200.0	1st Qu.:343.0	Median :390.6	Mean :410.9	3rd Qu.:484.5	Max. :640.0
w_b_ratio	Min. :0.2351	1st Qu.:0.3716	Median :0.4721	Mean :0.4676	3rd Qu.:0.5612	Max. :0.9000

Table 6: Test Set- Summary Statistics

cement	Min. :102.0	1st Qu.:178.0	Median :251.9	Mean :272.8	3rd Qu.:342.0	Max. :540.0
blast_furnace_slag	Min. : 0.00	1st Qu.: 0.00	Median : 38.00	Mean : 76.89	3rd Qu.:148.70	Max. :359.40
fly_ash	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 55.74	3rd Qu.:119.29	Max. :195.00
water	Min. :121.8	1st Qu.:164.9	Median :181.9	Mean :181.4	3rd Qu.:192.0	Max. :237.0
superplasticizer	Min. : 0.000	1st Qu.: 0.000	Median : 6.575	Mean : 6.080	3rd Qu.:10.000	Max. :32.200
coarse_aggregate	Min. : 801.0	1st Qu.: 932.0	Median : 968.0	Mean : 978.2	3rd Qu.:1047.0	Max. :1134.3
fine_aggregate	Min. :594.0	1st Qu.:727.8	Median :777.8	Mean :770.8	3rd Qu.:821.0	Max. :992.6
age	Min. : 3.00	1st Qu.: 7.00	Median : 28.00	Mean : 47.97	3rd Qu.: 56.00	Max. :365.00
compressive_strength	Min. : 2.332	1st Qu.:23.721	Median :34.426	Mean :35.159	3rd Qu.:46.005	Max. :79.297
w_c_ratio	Min. :0.2938	1st Qu.:0.5448	Median :0.7259	Mean :0.7759	3rd Qu.:0.9915	Max. :1.8824
additions	Min. : 0.00	1st Qu.: 95.68	Median :125.18	Mean :132.63	3rd Qu.:185.25	Max. :379.00
binder	Min. :200.0	1st Qu.:333.0	Median :393.0	Mean :405.4	3rd Qu.:480.0	Max. :616.0
w_b_ratio	Min. :0.2351	1st Qu.:0.3931	Median :0.4678	Mean :0.4729	3rd Qu.:0.5612	Max. :0.9000

We will also define a function to calculate RMSE, which we will use as a measure of model performance. RMSE (root-mean-square-error) is a measure of the difference between the values predicted by the model and the actual observed values in the data. Trying to predict the compressive strength of concrete (a quantitative, continuous variable) from various numeric features is a regression problem, and as such, RMSE is an appropriate measure for model performance.

The RMSE is defined as:

$RMSE = \sqrt{\frac{1}{N} * \sum_{i=1}^{n} (y_i - \hat{y_i})^2}$

where y_i is the observed value for the i-th data point and $\hat{y_i}$ is the predicted value.

2.2.1 Linear Model

We will start with a simple linear model that predicts the compressive strength based on the amount of cement in the mixture, as we saw earlier that cement had the highest correlation with compressive strength in the correlation matrix. We can implement this using the lm() function from the caret package.

We will also create a table to store the RMSE results in for comparison.

Table 7: RMSE Results

method	RMSE
Linear - Cement	13.86355

Next, we try a linear model using water-cement ratio (previously discussed) as our predictive variable.

Table 8: RMSE Results

method	RMSE
Linear - Cement	13.86355
Linear - W/C Ratio	13.89652

We can also try to predict concrete strength based on the water-binder ratio, as this extends the concept of the water-cement ratio by looking at the whole cementitious mixture.

Table 9: RMSE Results

method	RMSE
Linear - Cement	13.86355
Linear - W/C Ratio	13.89652
Linear - W/B Ratio	12.83594

Next, we look at modeling the compressive strength based on all of the original features in the dataset (without the water-cement ratio or the water-binder ratio.)

Table 10: RMSE Results

method	RMSE
Linear - Cement	13.863554
Linear - W/C Ratio	13.896515
Linear - W/B Ratio	12.835943
Linear - Original Features	9.625567

Finally, we look at introducing the water-cement ratio and the water-binder ratio into the model. Because we are introducing these engineered features, we will drop the water, cement, blast_furnace_slag, and fly_ash features.

Table 11: RMSE Results

method	RMSE
Linear - Cement	13.863554
Linear - W/C Ratio	13.896515
Linear - W/B Ratio	12.835943
Linear - Original Features	9.625567
Linear - Engineered Features	9.798686

We see an improvement in the linear regression approach as we refine the selected features used in the model.

Future work could introduce further feature refinement, as well as regularization to reduce the impact of noisy estimates.

2.2.2 Regression Tree

We know from the industry that concrete strength is a non-linear function of the concrete ingredients and the concrete age, and have demonstrated through several linear models that none perform exceptionally well. Next, we look at the application of a regression tree to this problem. A regression tree allows us to approach this problem with a more flexible, nonlinear approach. Where a linear regression model is a single formula for the entire dataset, a regression tree essentially breaks down the data into smaller sets so that we can fit a formula to each localized data set.

We can implement this using the rpart library.

Table 12: RMSE Results

method	RMSE
Linear - Cement	13.863554
Linear - W/C Ratio	13.896515
Linear - W/B Ratio	12.835943
Linear - Original Features	9.625567
Linear - Engineered Features	9.798686
Regression Tree	8.422840

2.2.3 Random Forest

We also explore the application of a random forest regression to this problem. A random forest is built on the same principle as a regression tree model, but instead of being built on the entire dataset like a regression tree, a random forest uses randomly selected subsets of the data to build multiple decision trees and then averages the results.

We can implement this using the randomForest library.

Table 13: RMSE Results

method	RMSE
Linear - Cement	13.863554
Linear - W/C Ratio	13.896515
Linear - W/B Ratio	12.835943
Linear - Original Features	9.625567
Linear - Engineered Features	9.798686
Regression Tree	8.422840
Random Forest	4.610495

3 Results

We see that the linear regression models using only cement, only water-cement ratio, and only water-binder ratio, are the worst performing models of the ones presented in this project. As we know that the compressive strength of concrete is a non-linear function of all ingredients, this model makes sense in context. The addition of the other features in the dataset improved model performance by about 25%.

We saw further improvement by introducing the regression tree technique, although the magnitude of performance increase was not as great (about 14% improvement over linear regression with engineered features).

After exploring multiple models to predict the compressive strength of concrete, we found the best RMSE was produced by the Random Forest model. This gives us an RMSE of approximately 4.61, which was almost a 50% performance improvement from the regression tree model. We know that RMSE is on the same scale as our dependent variable (compressive strength in MPa), which ranges from 2.33 MPa to 82.60 MPa in our dataset. This is a significant improvement over the other models explored and seems to best fit the non-linear nature of the problem.

4 Conclusion and Recommendations

This project has explored the "Concrete Compressive Strength Data Set" from the UCI Machine Learning Repository, and applied several models and techniques, as learned through the HarvardX Data Science program, to predict the compressive strength of concrete.

We think future improvments on this project would be to explore polynomial regression to capture the relationship of the various features to compressive strength. Additionally, there could be significantly more work done in the feature selection and model tuning for the regression tree models and random forest model, which may yield improved performance. We would suggest using cross validation in the model tuning and parameter selection, to offset the challenges presented by a small dataset. Additionally, we would explore and add evaluation metrics to capture whether any model is overfitting the data.

5 References

1. Irizarry, Rafael A (May 24,2021). *Introduction to Data Science,Data Analysis and Prediction Algorithms with R*. https://rafalab.github.io/dsbook/
2. Yeh, Prof. I-Cheng (August 3, 2007). *Concrete Compressive Strength Data Set*. http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength
3. I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)
4. American Concrete Institute (2021). *Topics In Concrete.*(26 July 2021). https://www.concrete.org/topicsinconcrete.aspx
5. PCA (2019). *How Concrete is Made*. (26 July 2021). https://www.cement.org/cement-concrete/how-concrete-is-made
6. Concrete Construction Staff (2021). *How Super are Superplasticizers*. (26 July 2021). https://www.concreteconstruction.net/how-to/materials/how-super-are-superplasticizers_o
7. Cor-Tuf UHPC (2021). *Everything You Need to Know About Concrete Strength*. (26 July 2021). https://cor-tuf.com/everything-you-need-to-know-about-concrete-strength/