2020_1124_Exploratory_Data_Analysis

November 25, 2020

1 Predicting Concrete Compressive Strength - Exploratory Data Analysis

1.1 Dataset Citation

This dataset was retrieved from the UC Irvine Machine Learning Repository from the following URL: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

The dataset was donated to the UCI Repository by Prof. I-Cheng Yeh of Chung-Huah University, who retains copyright for the following published paper: 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). Additional papers citing this dataset are listed at the reference link above.

1.2 Civil Engineering Domain Knowledge

1.2.1 Concrete Mix Design

The challenge of the dataset is to be able to predict the compressive strength of concrete given initial quantities of its components and the age after mixing and installation. The engineering term for the relative amounts of each material contained within a concrete mix is called the "mix design." The following materials comprise a mix design: * Cementitious materials (e.g. Portland cement, fly ash, etc.) * Coarse Aggregate (e.g. crushed rock, stone, gravel, etc.) * Fine aggregate (i.e. sand) * Water * Admixtures (materials to increase plasticity, prevent freezing, prevent corrosion, etc.)

Fly ash is a cementitious material that is typically lower cost but has different engineering characteristics than Portland cement. In this dataset, we are also provided data for "superplasticizer," an admixture, and "blast furnace slag," a byproduct of the concrete manufacturing process.

The data is provided in raw scientific quantities (kilograms of each material per cubic meter of mixed concrete). While this is useful from a scientific perspective, it is not in the standard engineering format that is used in the United States. Engineers typically specify concrete mix designs by certain minimum or maximum values for the following quantities: * Maximum water-to-cement (w/c) ratio – This is the ratio of all cementitious materials (Portland cement + fly ash) to the amount of water in the mix, by weight. It is widely accepted that the w/c ratio has the highest impact on the compressive strength of concrete. The lower the w/c ratio, the higher the compressive strength of the cured concrete should be. * Minimum sacks per cubic yard – A sack (sk) is a unit of measure for the weight of one cubic foot of Portland cement, usually around 94 pounds. * Maximum fly ash percentage – This is the ratio of fly ash to total cementitious materials, expressed as a percentage by weight. Engineers typically specify "no more than 25% fly ash substitution," for example.

1.2.2 Concrete Mixing, Curing, and Testing

Once the components are mixed, the water reacts with the cementitious materials in a process called hydration. The water-cementitious materials mixture is a paste that binds with the coarse and fine aggregates to create the high-strength material that we know as concrete. The period of time after which the components are mixed and during which hydration occurs is known as curing.

If the concrete is kept moist following mixture and installation (by, for example, covering it with moistened cloth or burlap), the compressive strength of the material will increase, following approximately a logarithmic curve as a function of time for a given mix design. However, installers rarely keep the concrete moist for longer than a week or a month due to water and labor costs.

We are not provided the length of time that the concrete was kept moist in this dataset. We do know, however, that the testing information provided was from concrete cylinder sampling. During construction, engineers typically sample concrete in cylinders and send them to a state-certified engineering laboratory. The lab holds the cylinders until the specified testing time, then they test them for compressive strength. Given this industry-standard practice, we assume in this analysis that the concrete was not moist-cured for any length of time.

The industry standard for concrete compressive strength is a 28-day cure. In this analysis, we expect that the concrete gains compressives trength rapidly from day of mixing until day 28, following which it will increase only very slightly over time.

1.3 Data Preprocessing

From an engineering and constructability perspective, it does not make sense to list the data as static kg/m^3 quantities. Rather, it is more practical to express quantities in terms of an engineering mix design; that is, to express each component as a percentage of the entire mix by weight. Furthermore, we expect the w/c ratio and fly ash-to-cement ratios also to play an important role in the overall compressive strength performance of the mixture.

For this reason, we have two separate data files. The first, 2020_1123_Concrete_Data_Loaded_Original.csv, contains the original data in an unaltered state, with only the column names shortened for ease-of-use in Python.

The second file, 2020_1123_Concrete_Data_Loaded_Transformed.csv, contains the following columns, which are related to the original by the relationships listed below. The total mass was calculated as cement + blast furnace slag + fly ash + water + superplasticizer + coarse aggregate + fine aggregate. The columns are the following: * Cementitious_Ratio = (Cement + Fly Ash)/(Total Mass) * Slag_Ratio = (Blast Furnace Slag)/(Total Mass) * Fly_Ash_Ratio = (Fly Ash)/(Cement + Fly Ash) * Water_to_Cementitious_Ratio = (Water)/(Cement + Fly Ash) * Superplasticizer_Ratio = (Superplasticizer)/(Total Mass) * Coarse_Aggregate_Ratio = (Coarse Aggregate)/(Total mass) * Sand_Ratio = (Fine Aggregate)/(Total Mass) * The Age and Concrete Compressive Strength columns retain the same data as the original file

The purpose of converting the original data into these quantities is just to gain deeper insights into the engineering properties of the materials during exploratory data analysis. But since the quantities are calculated based on the total mass (a quantity derived from the sum of the features), there is multicollinearity; therefore, in any modeling, we will need to use the raw scientific values in kg/m^3 (with scaling).

1.4 Exploratory Data Analysis

1.4.1 Import the Relevant Libraries

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set()
```

1.4.2 Import the Data

[2]:
```
df1 = pd.read_csv('2020_1123_Concrete_Data_Loaded_Original.csv')
df2 = pd.read_csv('2020_1123_Concrete_Data_Loaded_Transformed.csv')
original_data = df1.copy()
ratio_data = df2.copy()
```

1.4.3 Exploring the Original Data

[3]:
```
original_data.head()
```

[3]:
```
   Cement  Blast_Furnace_Slag  Fly_Ash  Water  Superplasticizer  \
0   540.0                 0.0      0.0  162.0               2.5
1   540.0                 0.0      0.0  162.0               2.5
2   332.5               142.5      0.0  228.0               0.0
3   332.5               142.5      0.0  228.0               0.0
4   198.6               132.4      0.0  192.0               0.0

   Coarse_Aggregate  Fine_Aggregate  Age  Compressive_Strength
0            1040.0           676.0   28                 79.99
1            1055.0           676.0   28                 61.89
2             932.0           594.0  270                 40.27
3             932.0           594.0  365                 41.05
4             978.4           825.5  360                 44.30
```

[4]:
```
original_data.describe()
```

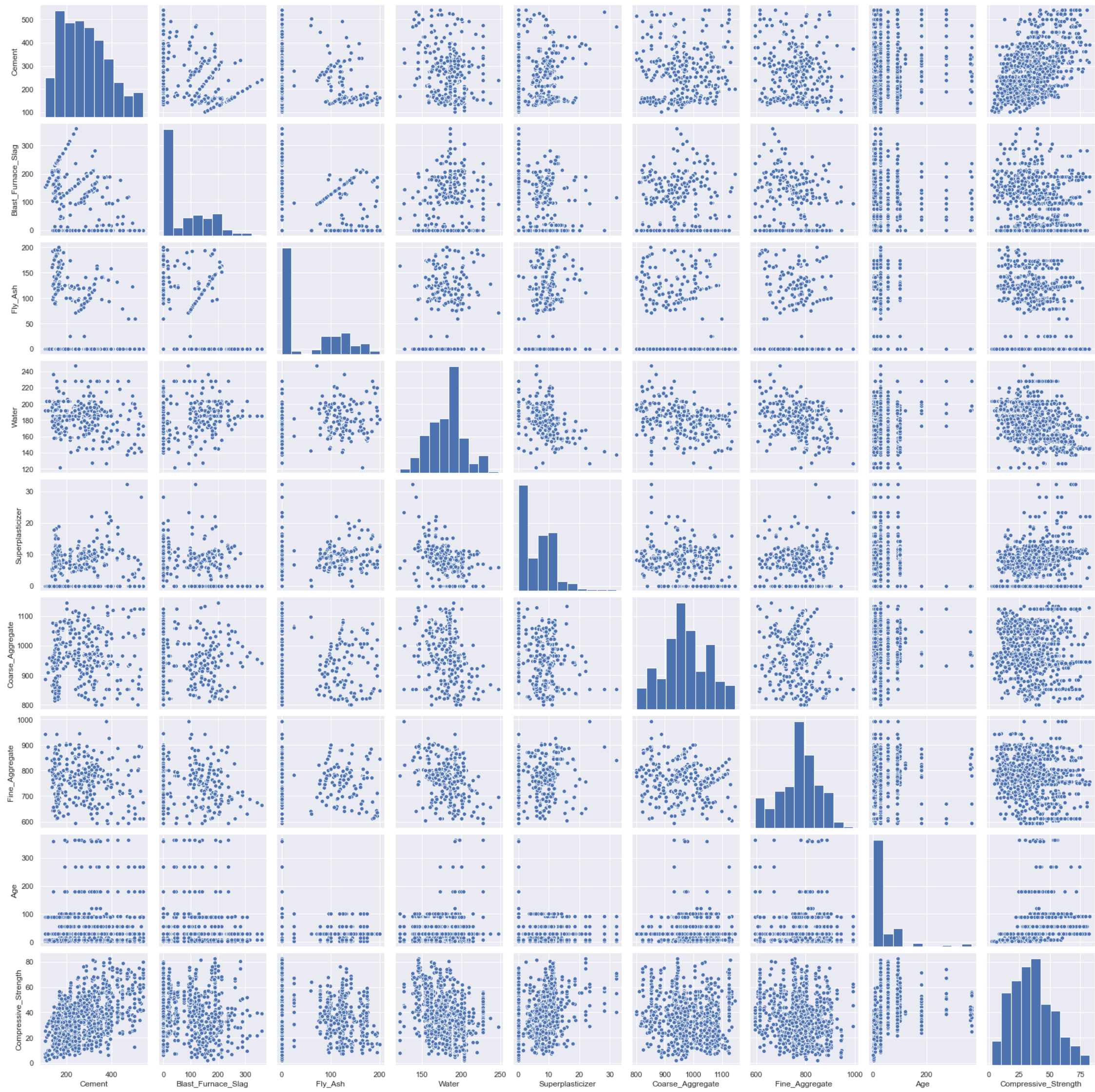
[4]:
```
            Cement  Blast_Furnace_Slag      Fly_Ash        Water  \
count  1030.000000         1030.000000  1030.000000  1030.000000
mean    281.167864           73.895825    54.188350   181.567282
std     104.506364           86.279342    63.997004    21.354219
min     102.000000            0.000000     0.000000   121.800000
25%     192.375000            0.000000     0.000000   164.900000
50%     272.900000           22.000000     0.000000   185.000000
75%     350.000000          142.950000   118.300000   192.000000
max     540.000000          359.400000   200.100000   247.000000
```

```
       Superplasticizer  Coarse_Aggregate  Fine_Aggregate          Age  \
count       1030.000000       1030.000000     1030.000000  1030.000000
mean           6.204660        972.918932      773.580485    45.662136
std            5.973841         77.753954       80.175980    63.169912
min            0.000000        801.000000      594.000000     1.000000
25%            0.000000        932.000000      730.950000     7.000000
50%            6.400000        968.000000      779.500000    28.000000
75%           10.200000       1029.400000      824.000000    56.000000
max           32.200000       1145.000000      992.600000   365.000000

       Compressive_Strength
count           1030.000000
mean              35.817961
std               16.705742
min                2.330000
25%               23.710000
50%               34.445000
75%               46.135000
max               82.600000
```

Let us examine the original data file by plotting pair plots. This will show us any potential relationships between the different dataset features and will help us give a cursory check for multicollinearity.

[5]:
```
sns.pairplot(original_data)
```

[5]:
```
<seaborn.axisgrid.PairGrid at 0x1acd7c169a0>
```



We can see that there is a clear positive linear relationship between cement and compressive strength, and an inverse linear relationship between water and compressive strength, both of which were expected. There may be relationships between superplasticizer and fly ash, but this will be further explored in the transformed data analysis.

There may be direct multicollinear relationships between blast furnace slag and cement and slag and fly ash. This makes sense, considering that the blast furnace slag is a byproduct of the concrete manufacturing process.

1.4.4 Exploring the Transformed Data

[8]:
```
ratio_data.head()
```

[8]:
```
   Cementitious_Ratio  Slag_Ratio  Fly_Ash_Ratio  Water_to_Cementitious_Ratio  \
0            0.205086    0.000000            0.0                     0.400000
1            0.167391    0.000000            0.0                     0.483117
2            0.058291    0.087436            0.0                     1.375358
3            0.145726    0.000000            0.0                     0.550143
4            0.085350    0.056900            0.0                     0.966767

   Superplasticizer_Ratio  Coarse_Aggregate_Ratio  Sand_Ratio  Age  \
0                     0.0                0.461444    0.251436    1
1                     0.0                0.420000    0.331739    1
2                     0.0                0.437179    0.336924    3
3                     0.0                0.437179    0.336924    3
4                     0.0                0.420474    0.354764    3

   Compressive_Strength
0             12.638095
1              6.267337
2              8.063422
3             15.049193
4              9.131420
```

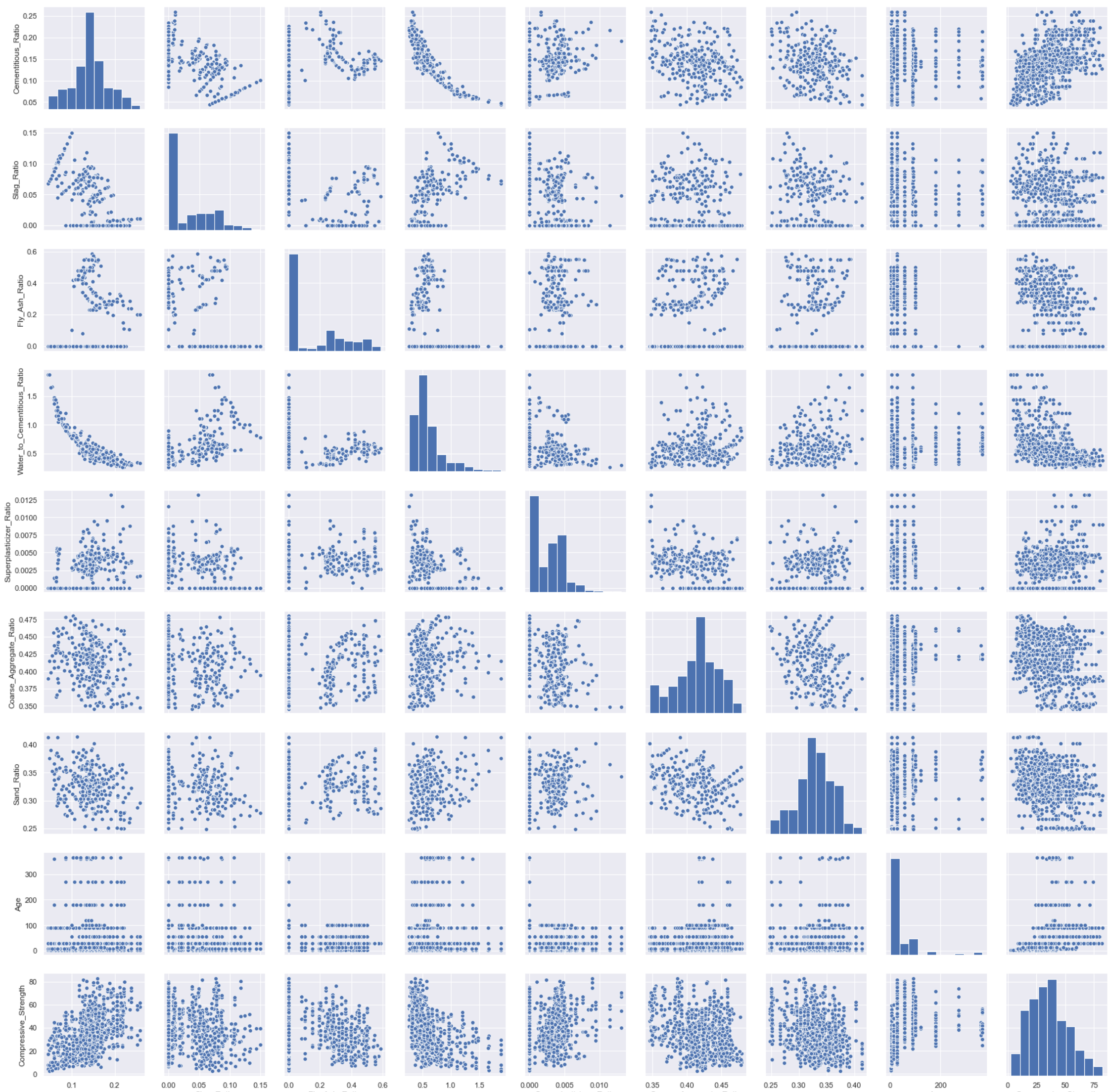
[9]:
```python
ratio_data.describe()
```

[9]:
```
       Cementitious_Ratio   Slag_Ratio  Fly_Ash_Ratio  \
count         1030.000000  1030.000000    1030.000000
mean             0.142726     0.031643       0.155263
std              0.040513     0.036961       0.187884
min              0.044815     0.000000       0.000000
25%              0.124002     0.000000       0.000000
50%              0.143272     0.009455       0.000000
75%              0.162794     0.061972       0.319960
max              0.259517     0.150339       0.588415

       Water_to_Cementitious_Ratio  Superplasticizer_Ratio  \
count                  1030.000000             1030.000000
mean                      0.611796                0.002620
std                       0.278319                0.002494
min                       0.265918                0.000000
25%                       0.447540                0.000000
50%                       0.547837                0.002727
75%                       0.666639                0.004338
max                       1.882353                0.013149

       Coarse_Aggregate_Ratio   Sand_Ratio          Age  Compressive_Strength
count             1030.000000  1030.000000  1030.000000           1030.000000
mean                 0.415166     0.330117    45.662136             35.817836
std                  0.031021     0.033245    63.169912             16.705679
```

min	0.345890	0.247971	1.000000	2.331808
25%	0.392294	0.311208	7.000000	23.707115
50%	0.420464	0.330543	28.000000	34.442774
75%	0.437623	0.354096	56.000000	46.136287
max	0.479846	0.414147	365.000000	82.599225

[10]:
```
sns.pairplot(ratio_data)
```

[10]: <seaborn.axisgrid.PairGrid at 0x1acdb440f40>



1.5 Initial Visual Analysis

It is clear from these pair plots that the ratios in the transformed data are much more strongly correlated with compressive strength than the raw values in the original data.

Below are the observed relationships with compressive strength: * Cementitious_Ratio - A clear positive linear relationship. This was assumed from domain knowledge. * Slag_Ratio - Unclear. But it is unlikely that this substance would contribute anything to the engineering properties of the concrete mix. * Fly_Ash_Ratio - A clear inverse linear relationship. While we assumed from domain knowledge that a high fly ash ratio would reduce compressive strength performance, the clear linearity of the relationship is surprising and should be further studied. * Water_to_Cementitious_Ratio - A clear non-linear inverse relationship. We assumed an inverse relationship from domain knowledge, but the nonlinearity of it is surprising. It appears as if there is a steep reduction in compressive strength up to around 1.0 (equal parts water and cement), then it declines less rapidly the more water is added. This should be further studied. * Superplasticizer_Ratio - Unclear, possibly a positive relationship, should be further studied. * Coarse_Aggregate_Ratio - Unclear. * Sand_Ratio - Unclear. * Age - An assumed positive logarithmic relationship from domain knowledge.

1.6 Conclusions & Future Modeling

We will use the original data during modeling in order to avoid the multicollinearity of the transformed data. Cement, fly ash, and water should be kept as separate quantities during modeling. The w/c and other engineering ratios can be analyzed post-modeling (optional). All input values should be scaled and run through an artificial neural network (ANN) in a train-test split. The performance of the ANN model should be compared with the performance of linear models trained on cement vs. compressive strength, fly ash ratio vs. compressive strength, etc. to determine the best model.