2011147018 배진훈

1) your dataset

- Brief

Title: Concrete Compressive Strength Data Set

Link: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Description: Concrete is the core material in **civil engineering**. The concrete compressive strength(y) is a mostly non-linear function of ingredients and age(X).

- Detail

Input(X): Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age

Output(y): Concrete compressive strength (MPa, megapascals)

Input	**Data Type**	**Measurement**
Cement (component 1)	Quantitative	Kg in a m3 mixture
Blast Furnace Slag (component 2)		
Fly Ash (component 3)		
Water (component 4)		
Superplasticizer (component 5)		
Coarse Aggregate (component 6)		
Fine Aggregate (component 7)		
Age		Day (1~365)
Output	**Data Type**	**Measurement**
Concrete compressive strength	Quantitative	MPa

Describe: 1030 rows

2) your overall approach

- Metric

According to KCS (KOREA CONSTRUCTION STARNDARDS), the design standard compressive strength of high-strength concrete is generally 40MPa or more, and high-strength lightweight aggregate concrete is 27 MPa or more. [1]That is, the absolute error and ratio of the error to actual value are important and it is easy to intuitively judge using **MAE** among various metrics used for

[1] https://www.kcsc.re.kr/StandardCode/Viewer/517

regression. Then, I'll use mean error ratio(**MER**) calculated by dividing the absolute error by the actual target value

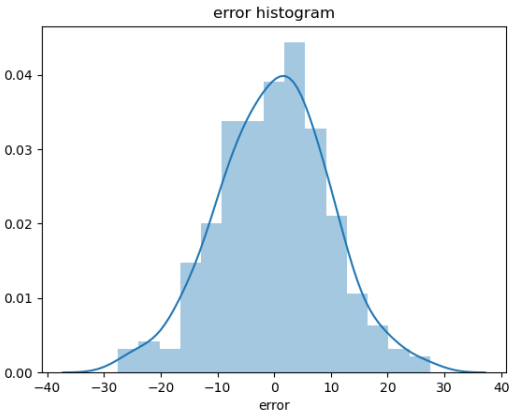
- **Baseline**

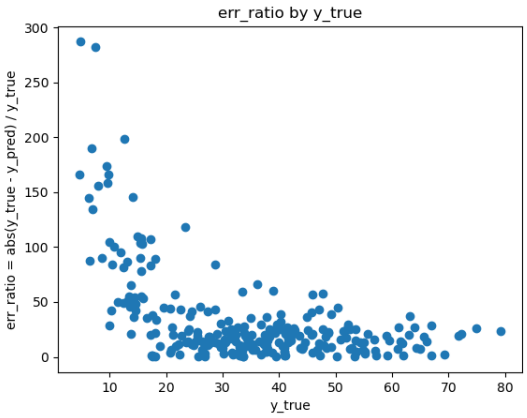
Intro

Linear Model – Elastic Net
Not feature extraction

Result (MAE: 7.7MPa, MER: 33%)

```
2020-06-24 04:14:10:pid_45692:parent:<INFO> mean of err_ratio: 32.924543776131735
2020-06-24 04:14:10:pid_45692:elastic_net:<INFO> tuned params: {'alpha': 10, 'l1_ratio': 0.0, 'max_iter': 10}
2020-06-24 04:14:10:pid_45692:elastic_net:<INFO> beta_coef:
            column       beta
0           Cement   0.102615
1            Blast   0.087667
2              Fly   0.070557
3            Water  -0.237636
4  Superplasticizer   0.174586
5           Coarse   0.001949
6             Fine   0.000539
7              Age   0.111609
8        intercept  31.332524
2020-06-24 04:14:10:pid_45692:parent:<INFO> metric is 7.716627039419666
```





Analysis

It is not bad that MAE is 7-8. However, considering that MER is 33, it has a large error compared to the actual y value. It should be developed so that MER is at least 10%

iii) your featuring engineering ideas

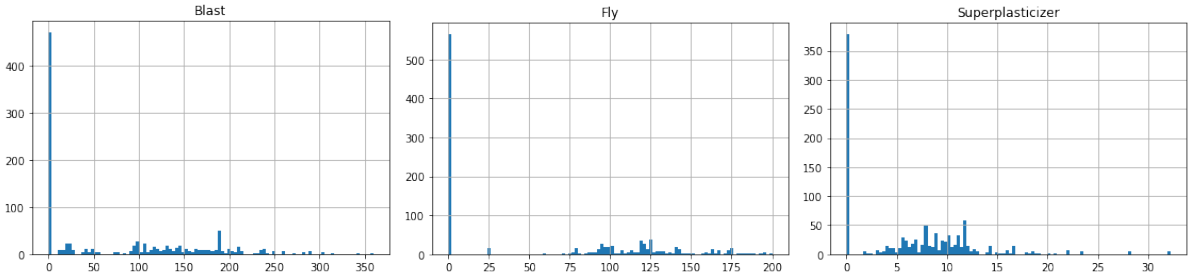
- **Processed**

Intro

Linear Model – Elastic Net

+ Feature extraction

data_analysis.ipynb



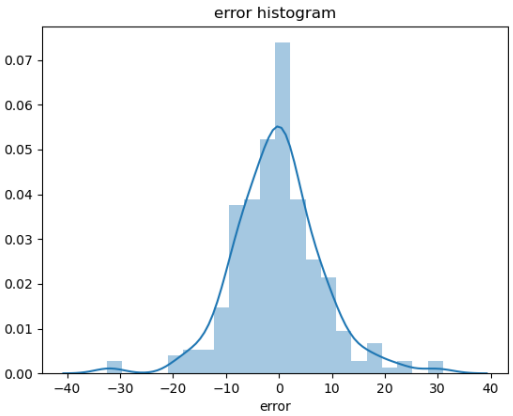

```
zero = df.applymap(lambda x: x == 0.0)
zero.sum().to_frame().T
```

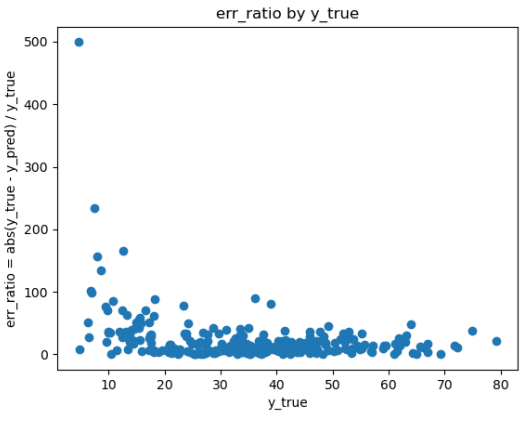
Cement	Blast	Fly	Water	Superplasticizer	Coarse	Fine	Age	Concrete
0	466	566	0	379	0	0	0	0

As a result of checking the histogram of the features, some features have a lot of zero values as sparse data. Especially in the case of 'Fly Ash', zero accounts for about 50%. Therefore, 3 features were engineered into new ones by changing values greater than 0 to 1. And, other numeric features were normalized to close to the normal distribution

Result (MAE: 6.1MPa, MER: 23%)

```
2020-06-24 02:43:52:pid_32472:parent:<INFO> mean of err_ratio: 23.77682995908908
2020-06-24 02:43:52:pid_32472:elastic_net:<INFO> tuned params: {'alpha': 0, 'l1_ratio': 0.0, 'max_iter': 5}
2020-06-24 02:43:52:pid_32472:elastic_net:<INFO> beta_coef:
            column       beta
0           Cement   7.388395
1            Blast  -3.143222
2              Fly  -0.507584
3            Water  -0.029602
4  Superplasticizer   8.363843
5           Coarse   8.763669
6             Fine  -4.320352
7              Age  11.338167
8        intercept  25.468510
2020-06-24 02:43:52:pid_32472:parent:<INFO> metric is 6.108020124279915
```





Analysis

After feature extraction, both MAE and MER have been reduced significantly and will be further reduced by using advanced model.

v) execution results

Method of execution is at the end & The result of 'baseline' & 'processed' is above.

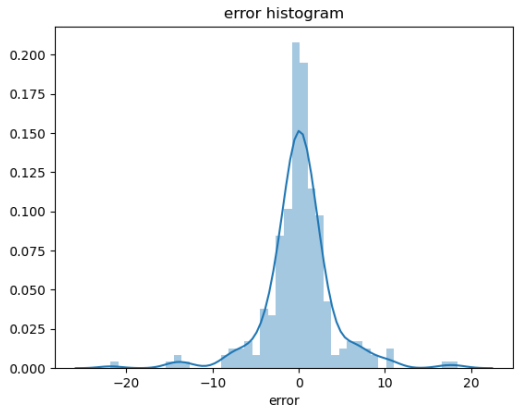
- Advanced

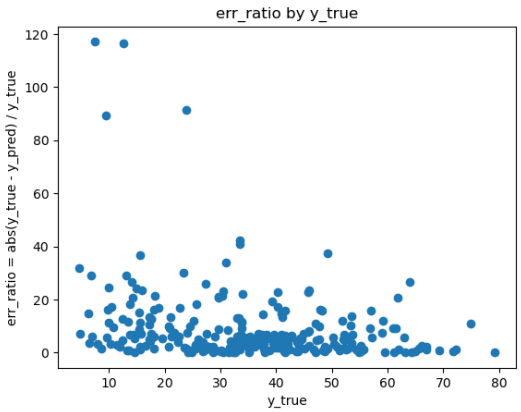
Intro

Ensemble Model – Gradient Boosting
+ Feature extraction

Result (MAE: 2.5MPa, MER: 9%)

```
2020-06-24 04:12:03:pid_41104:parent:<INFO> mean of err_ratio: 9.111544788042194
2020-06-24 04:12:03:pid_41104:gradient_boost:<INFO> tuned params: {'max_depth': 10, 'max_features': 0.32, 'n_estimators': 200}
2020-06-24 04:12:03:pid_41104:parent:<INFO> metric is 2.5386993744579174
```





Analysis

Finally, a reasonable result was obtained as lowering mean of error ratio to less than 10%. The variance of the error histogram decreased, and the error ratio also decreased overall. The error ratio less than 10% means that the error is within 3 when the concrete compressive strength is 30.

Furthermore

- Model advancement (ex. Deep learning)
- Vast amounts of data
- Sophisticated hyperparameter tuning