Predicting Concrete Compressive-Strength using Multiple Linear Regression techniques

Carolina Garma Escoffié
Data Engineering
Universidad Politécnica de Yucatán
Mérida, Yucatán
Email: st1809073@upy.edu.mx

***Abstract*—The following paper reports the results of the application of multiple linear regression techniques in order to predict the concrete compressive strength (target) which is a highly nonlinear function of age and ingredients.**

Keywords— Concrete Compressive Strength Data Set, Multiple Linear Regression, Lasso Regression, Ridge Regression, ElasticNet Regression

I. INTRODUCTION

Concrete is a hardened mass that by its very nature is discontinuous and heterogeneous. The properties of any heterogeneous system depend on the physical and chemical characteristics of its constituent materials and the interactions between them.

Concrete is the most important material in civil engineering. Simple compressive strength is the principal mechanical characteristic of concrete. It is defined as the ability to support a load per unit area, and is expressed in terms of stress, usually in kg/cm2, MPa and sometimes in pounds per square inch (psi). Based on the above, concrete compressive strength is a highly nonlinear function of age and ingredients [1].

The dataset here presented was obtained from the UCI Machine Learning Repository. Its original owner and donor, Prof. I-Cheng Yeh donated it on August 3, 2007 [2].

Some relevant information about the dataset is that contains 1030 instances and 9 attributes, all of them are quantitative. In this approach the compressive strength of concrete (in MPa) is a function of the following 8 input features:

1. Cement (kg/m3)
2. Fly Ash (kg/m3)
3. Blast furnace slag (kg/m3)
4. Water (kg/m3)
5. Superplasticizer (kg/m3)
6. Coarse aggregate (kg/m3)
7. Fine aggregate (kg/m3)
8. Age of testing (days)

There is not a specific way of modeling concrete strength, but the proposed regression models in this paper are based on these 8 inputs features.

II. DATA EXPLORATION

A. Attributes Information

#	Column	Non-Null Count	Dtype
0	Cement	1030	Float64
1	Blast furnace slag	1030	Float64
2	Fly Ash	1030	Float64
3	Water	1030	Float64
4	Superplasticizer	1030	Float64
5	Coarse aggregate	1030	Float64
6	Fine aggregate	1030	Float64
7	Age	1030	Int64
8	Compressive strength	1030	Float64

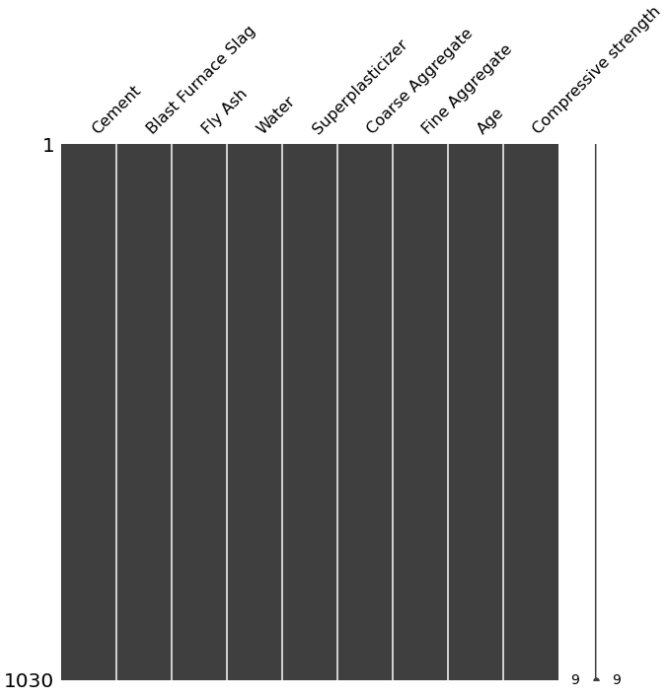


Figure 1. Null-value count

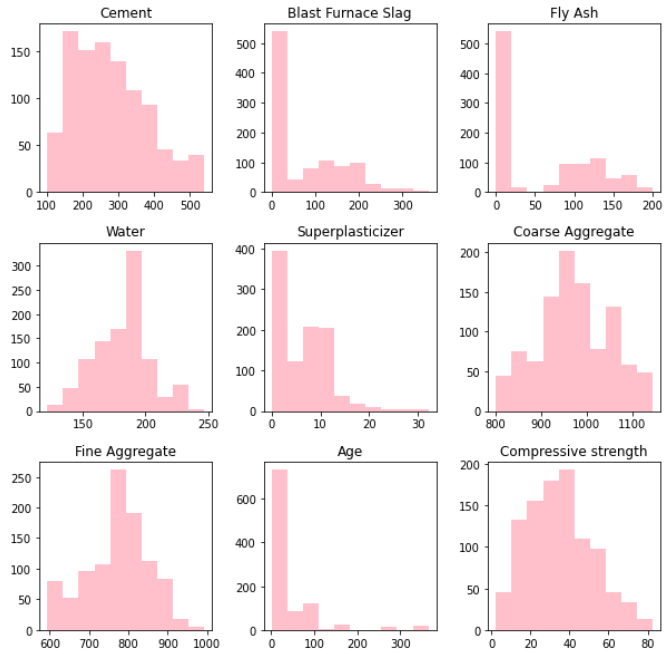


Figure 1. Dataset histogram

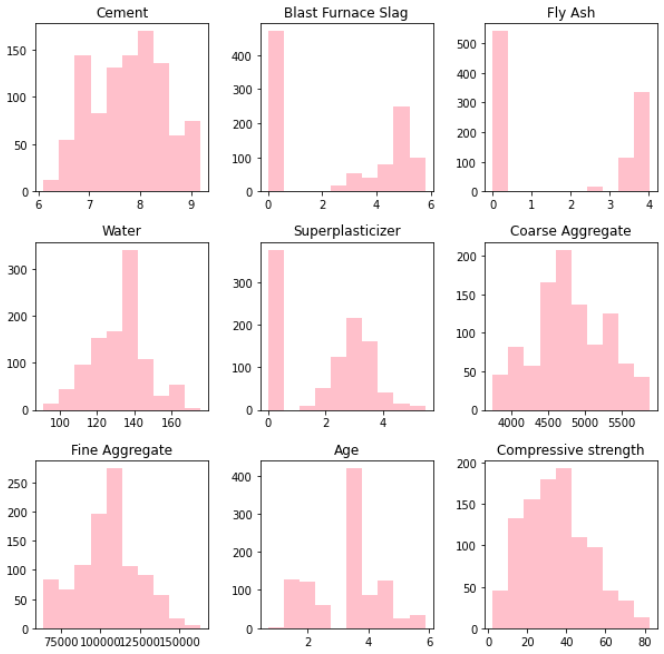


Figure 2. Gaussian-like dataset histogram

The dataset contains 1030 records, 1005 are unique records, 9 attributes and 0 null values. The histogram for each attribute is presented below.

We can see there is a tendency for ASF > 0, Positive Asymmetry. In the case of Fine and Coarse Aggregate we can see ASF < 0, Negative Asymmetry. Some skew transformations need to be implemented.

III. DATA PREPROCESSING

There were 25 duplicate data records that were properly removed from the dataset; thus, the new dataset contains 1005 unique records.

Skewed data may act as an outlier for the statistical model and we know that outliers adversely affect the model's performance especially regression-based models. So, there is a necessity to transform the skewed data to close enough to a Gaussian distribution or Normal distribution [3].

To correct the left skew tendency observed during the exploration. data transformation was conducted using a PowerTransformer() function. After the transformation, the following more Gaussian Like histogram was obtained:

A. Data Description

	count	mean	std	min	25%	50%	75%	max
Cement	981.0	7.768931	0.711013	6.099733	7.189006	7.773297	8.265756	9.189167
Blast Furnace Slag	981.0	2.420814	2.378395	-0.000000	-0.000000	2.970171	4.888185	5.789037
Fly Ash	981.0	1.715529	1.851748	-0.000000	-0.000000	-0.000000	3.725025	4.026326
Water	981.0	132.789648	13.568880	102.274275	122.900043	134.855790	139.720988	163.095705
Superplasticizer	981.0	1.884046	1.562778	0.000000	0.000000	2.474869	3.233100	5.468651
Coarse Aggregate	981.0	4805.302927	480.270930	3744.492311	4534.972509	4757.611977	5146.191833	5883.324934
Fine Aggregate	981.0	104042.732621	19080.398258	63973.081172	91849.653911	105140.246039	116210.375268	149212.968211
Age	981.0	3.228217	1.106798	0.692674	2.075184	3.356142	4.026978	5.868416
Compressive strength	981.0	34.856729	15.865778	2.331808	23.345657	33.718824	44.520450	76.800732

For the outlier detection, Interquartile Range (IQR) (low=0.25, upper=0.75) was implemented After filtering the dataset, the following boxplot was obtained (a standardized version of the dataset is shown for better visualization).

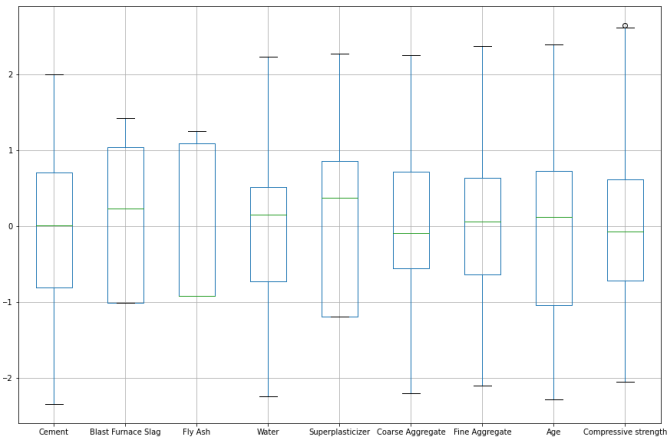


Figure 3. Standarized boxplot after IQR filtering

IV. DATA CORRELATION

A correlation matrix over the dataset was conducted to see which attributes have more impact over the target value. The absolute correlation indexes for the target vary from 0.038 to 0.57. The most correlated attributes (using absolutes) were Age (0.57), Cement (0.47) and Superplasticizer (0.32) and the least were Fly Ash (0.038), Blast Furnace Slag (0.12) and Coarse Aggregate (0.15).

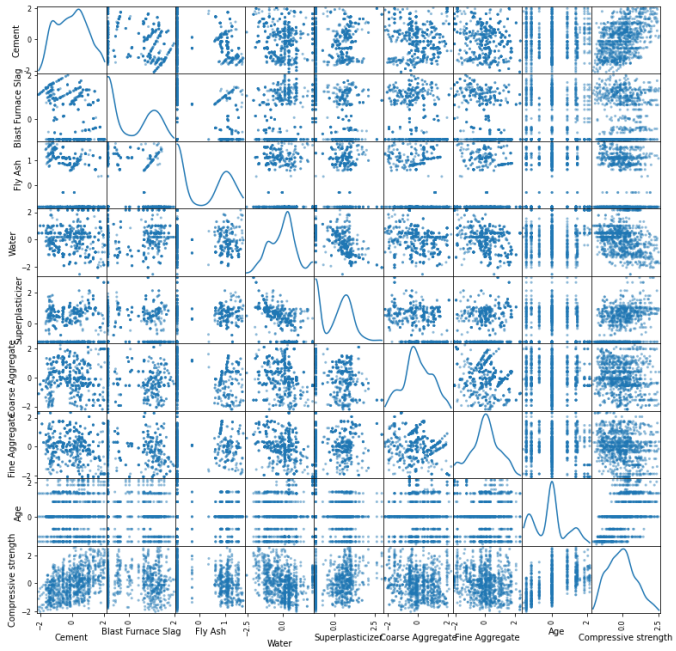


Figure 4. Dataset Scatter Plot

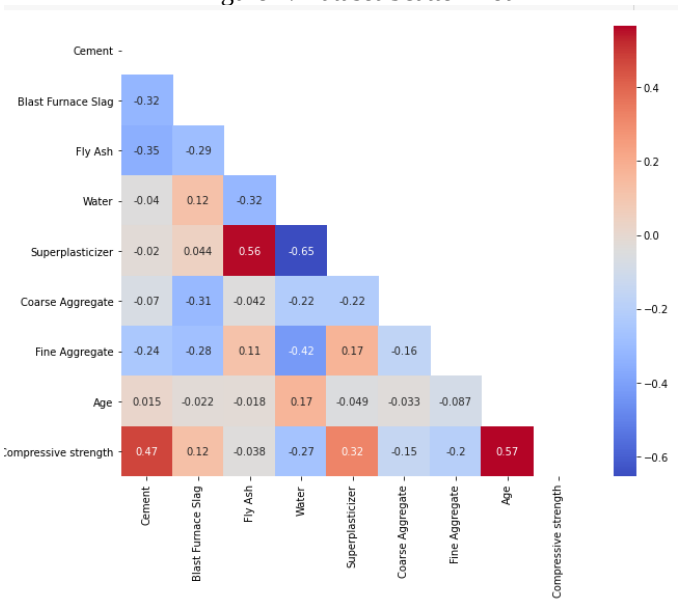


Figure 5. Dataset Heatmap

For the next section, the 8 input attributes will be held to carry out the regression models.

V. PREVIOUS WORK

The first paper released about this dataset was made by its donor, I-C. Yeh, from the department of Civil Engineering of the Chung-Hua University in 1998. This study was aimed at demonstrating the possibilities of adapting neural networks to predict the compressive strength of concrete. A set of trial batches of HPC was produced at the laboratory and showed satisfactory experimental results. However, the method is not applicable to extrapolation beyond the domain of the data accumulated in the past.

A second study by I-C. Yeh was related in the same year (1998), this time evaluating Augment-Neuron Networks. The results showed that the logarithm neurons and exponent neurons in the network provide an enhanced network architecture to improve performance of these networks for modeling concrete strength significantly. A neural network–based concrete mix optimization methodology is proposed and is verified to be a promising tool for mix optimization [4].

Other studies made by I-C. Yeh exploring other features of concrete strength can be found in the references [5] [6] [7] [8].

VI. DATA MODELING

For this section, four different model approaches using 8 input attributes will be carried out throughout the sklearn module:

A. *Multiple Linear Regression (OLS)*

$$J(\beta) = \frac{1}{2N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2$$

OSL Cost function

B. *Lasso Regression or L1 Regularization*

$$J(\beta) = \frac{1}{2N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{d}|\beta_j|^1$$

Lasso Cost function

C. *Ridge Regression or L2 Regularization*

$$J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\beta_j^2$$

Ridge Cost function

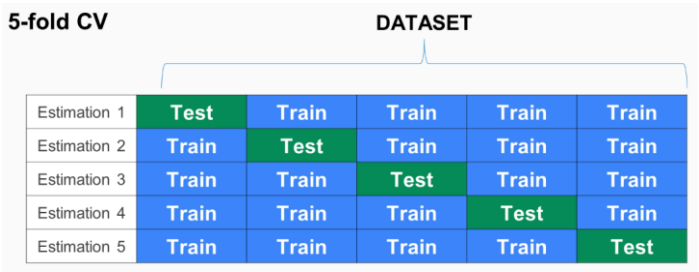
D. *ElasticNet Regression (L1 and L2 Regularization)*

$$J(\beta) = \frac{1}{2N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2 + \alpha\rho\sum_{j=1}^{d}|\beta_j|^1 + \frac{\alpha(1-\rho)}{2}\sum_{j=1}^{d}|\beta_j|^2$$

ElasticNet Cost Function

A. *Test and Training Sets*

For training the data and getting the optimal hyperparameters a 5-folds cross validation was implemented.



Model	Train Set	Test Set	CV
OSL	4/5	1/5	5
Lasso	4/5	1/5	5
Ridge	4/5	1/5	5
ElasticNet	4/5	1/5	5

VII. RESULTS

A. *Lasso Regression Best Hyperparameter: alpha = 0*

The results from the 5-fold cross validation showed that the optimal coefficient for alpha is 1e-6, in other words, very close to zero. When alpha tends to zero, this means that L1 regularization is ignored for the model, thus becoming a simple OLS.

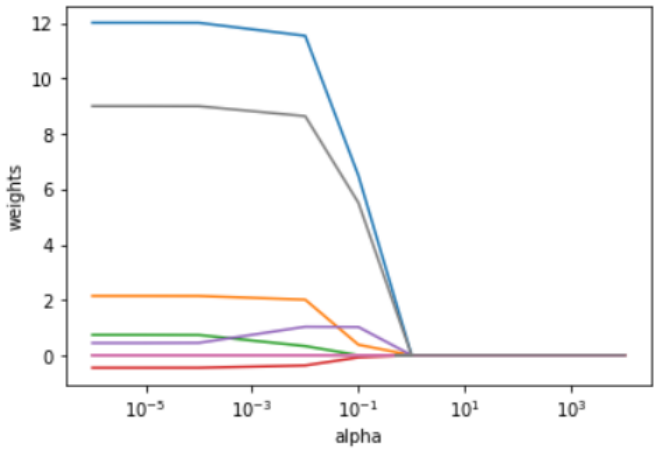


Figure. Lasso Coefficients as a function of regularization

B. *Ridge Regression Best Hyperparameter: alpha = 0*

The results from the 5-fold cross validation showed that the optimal coefficient for alpha is 1e-6, in other words, very close to zero. When alpha tends to zero, this means that L2 regularization is ignored for the model, thus becoming a simple OLS.

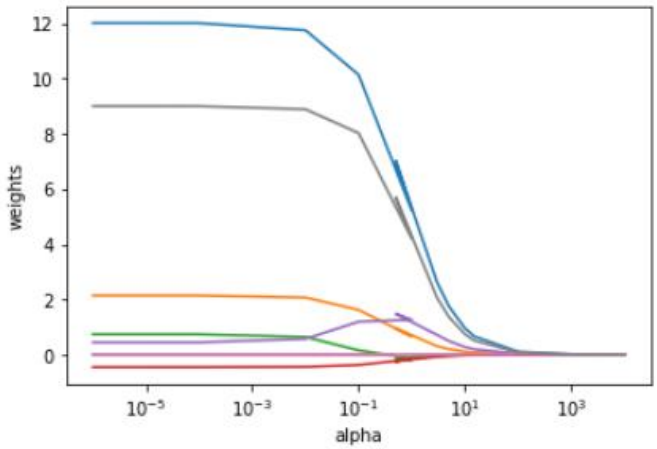


Figure. Ridge Coefficients as a function of regularization

C. *ElasticNet Best Hyperparameters: alpha=0, l1_ratio=1*

The results from the 5-fold cross validation showed that the optimal coefficient for alpha is 1e-6, in other words, very close to zero. For the penalization, we obtained 1 .0 which means a L1 penalization, i.e., Lasso Regression, but given that the alpha is almost zero, this means that both regularizations are ignored for the model, thus becoming a simple OLS.

D. *Table of regression coefficients*

Attribute	OLS	Lasso	Ridge	Elastic Net
Cement	11.7279	11.7279	11.7279	11.7279
Blast furnace slag	2.0719	2.0719	2.0719	2.0719
Fly Ash	0.6695	0.6695	0.6695	0.6695
Water	-0.4579	-0.4579	-0.4579	-0.4579
Superplasticizer	0.5107	0.5107	0.5107	0.5107
Coarse aggregate	-0.0026	-0.0026	-0.0026	-0.0026
Fine aggregate	-0.0001	-0.0001	-0.0001	-0.0001
Age	8.9972	8.9972	8.9972	8.9972

Here we can see that the coefficients of coarse and fine aggregate should be eliminated given that they are almost zero and will not affect the prediction.

E. *R square train and test results*

Model	R2 Train	R2 test	RMSE	MAE
OLS	0.8167	0.7091	7.7659	6.2568
Lasso	0.8167	0.7091	7.7659	6.2568
Ridge	0.8167	0.7091	7.7659	6.2568
ElasticNet	0.8167	0.7091	7.7659	6.2568

F. Plot of data and a linear regression model fit.

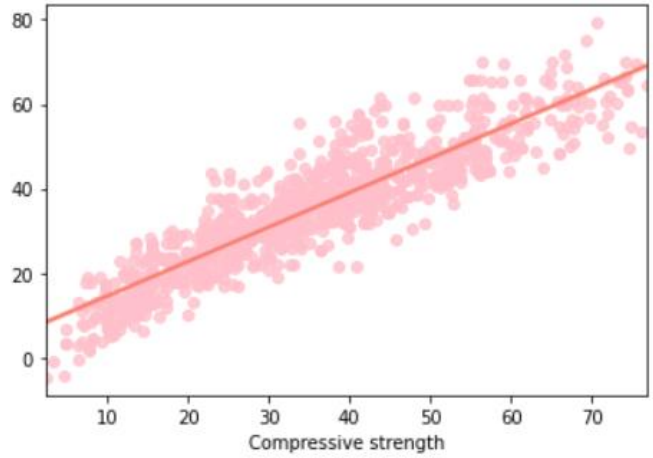


Figure. Regplot

As seen above, the predictions vary from up and down the real data. The MAE and RMSE give us an insight of how much these predictions differ from the data, approximately from 6.2 to 7.7 units., which is not so bad given the standard deviation of the target variable (std=15.86), this means that the predictions are inside the first sigma interval of the normal distribution of values.

VIII. CONCLUSIONS

Based on both tables, coefficients and train and test results, we can see that no model was the best because all of them got

almost the same values. What we can infer from this premise and accordingly to the best hyperparameters for lasso, ridge and elastic net regressions is that no regularization is needed to improve this model because all the mentioned models ended up cancelling their penalizations given their low alphas.

Therefore, the model that better fits this dataset distribution is the ordinary least square regression. How good is that fitting? According to the coefficient of determination, its accuracy is of 70% approximately, which is not the best but neither the worst. So, the predictions conducted using this model for further analysis will not be so far from the real values.

What does that mean regarding the dataset context? Interestingly, this means that all of the 8 input variables are important when determining the Concrete Compressive-Strengt, especially the Age, Cement and Superplasticizer quantities per cubic meter. Something curious is that if only these 3 variables are used for the modeling, the accuracy gets drastically worst (between a 50% and 60%) so, it is better to employ all the variables when the prediction is needed. The only exception would be with coarse and fine aggregates whose coefficients are almost zero and therefore they will be eliminated.

IX. FUTURE WORK

What would be interesting to do next is to make this same analysis over different concrete datasets containing different components and check its relation and how the prediction works and compare it with the previous done.

Also, further analysis regarding the performance differences in other types of modeling (non-linear) to identify what kind of behavior is the one that better fits the concrete distribution data, because in the current paper the accuracy results were acceptable but not optimal.

REFERENCES

[1] 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).

[2] I-Cheng Yeh, “Concrete Compressive Strength Data Set”, UCI Machine Learning Repository, (1998).

[3] Rajat Sharma, “Skewed Data: A problem to your statistical model”, Towards Data Science, (2019).

[4] I-Cheng Yeh, "Modeling Concrete Strength with Augment-Neuron Networks," J. of Materials in Civil Engineering, ASCE, Vol. 10, No. 4, pp. 263-268 (1998).

[5] I-Cheng Yeh, "Design of High Performance Concrete Mixture Using Neural Networks," J. of Computing in Civil Engineering, ASCE, Vol. 13, No. 1, pp. 36-42 (1999).

[6] I-Cheng Yeh, "Prediction of Strength of Fly Ash and Slag Concrete By The Use of Artificial Neural Networks," Journal of the Chinese Institute of Civil and Hydraulic Engineering, Vol. 15, No. 4, pp. 659-663 (2003).

[7] I-Cheng Yeh, "A mix Proportioning Methodology for Fly Ash and Slag Concrete Using Artificial Neural Networks," Chung Hua Journal of Science and Engineering, Vol. 1, No. 1, pp. 77-84 (2003).

[8] Yeh, I-Cheng, "Analysis of strength of concrete using design of experiments and neural networks,": Journal of Materials in Civil Engineering, ASCE, Vol.18, No.4, pp.597-604 ?2006?.

[9] Carolina Garma, “Concrete-Compressive-Strength-Data-Set”, GitHub Repository, (2021).