Predicting strength of concrete using linear regression models

Aleksandra Leszczyk

Abstract
Have you considered building a house? You will probably need to use concrete. It consists of three fundamental components: water, cement, and aggregate (e.g. sand, rock, or gravel). This mix can harden into hardy solid-state substance. Before building a house, we should estimate the durability of used materials to avoid a building failure. Thus, we learn from past data how to evaluate the strength of concrete based on the composition and age of the substance.

Keywords
concrete strength — linear regression

Contents

1. The objective

Our goal is to predict the durability of concrete from its composition and age using linear regression model. Thus, we examine different models to pick the one with highest predictive power.

2. Description of the data set

The data set has been downloaded from www.kaggle.com webpage. The data contains 1030 observations of 9 features. The strength of concrete is a target variable. We aim at predicting it when we know the age of concrete and the amount of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate used to prepare concrete. The sample of data set is presented in Table 1 while variables and their types are listed in Table 2. All the variables are continuous numerical except for age which is a discrete numerical variable here.

3. Exploration and cleaning

3.1 Exploratory data analysis

Since the variables are numerical, we focus on distributions of numerical variables. Table 4 summarizes the statistical features of variables, while Figure 2 presents histograms showing data distributions (in diagonal). We see immediately that data is not standardized. Most variables are not normally distributed. Age is right skew as the data set examines mostly young buildings. Some features, that is blast furnace slag, fly ash, and superplasticizer, are right-skew because they have many zero-values. We can conclude that these ingredients are optional as they are not always in many concrete compositions. The exact measures of skewness are in Table **??**.

Furthermore, we examine correlations between variables using pairplot (see Figure 2) and heatplot (see Figure 1). We observe that some independent variables

Table 1. Data sample.

	Cement	Blast furnace slag	Fly ash	Water	Superplasticizer	Coarse aggregate	Fine aggregate	Age	Strength
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30

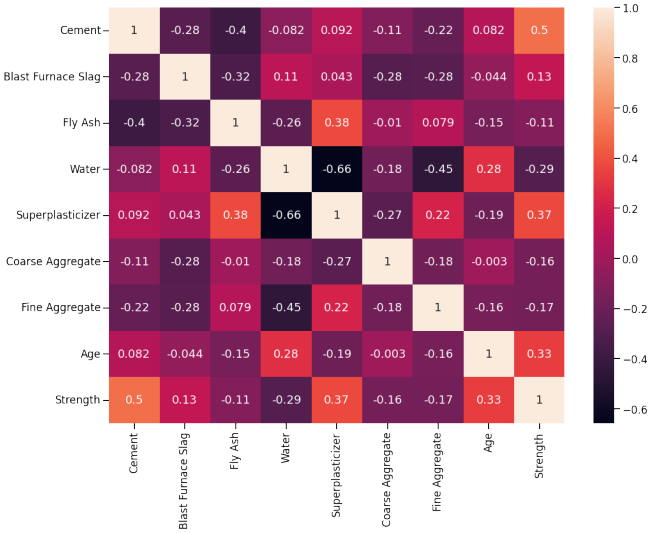
Table 2. Column names, number of values for each column and data types.

No.	Column	Non-null count	data type	
0	Cement	1030 non-null	float64	continuous numerical
1	Blast furnace slag	1030 non-null	float64	continuous numerical
2	Fly ash	1030 non-null	float64	continuous numerical
3	Water	1030 non-null	float64	continuous numerical
4	Superplasticizer	1030 non-null	float64	continuous numerical
5	Coarse aggregate	1030 non-null	float64	continuous numerical
6	Fine aggregate	1030 non-null	float64	continuous numerical
7	Age	1030 non-null	int64	discrete numerical
8	Strength	1030 non-null	float64	continuous numerical

Table 3. Skewness od data distributions.

Column	Skew
Cement	0.509481
Blast furnace slag	0.800717
Fly ash	0.537354
Water	0.074628
Superplasticizer	0.907203
Coarse aggregate	-0.040220
Fine aggregate	-0.253010
Age	3.269177
Strength	0.416977

Figure 1. Correlation between variables.



are negatively correlated, as superplasticizer and water, or fine aggregate and water. However, we do not need to worry about dependencies between independent variables, as adding ingredients to mixture is independent. Thus, we should not get a spurious correlation.

3.2 Data cleaning

The data set do not contain missing values as it is shown in Table 2. Furthermore, the Table 4 and Figure **??** suggest that data set do not contain outliers. Thus, no action is required.

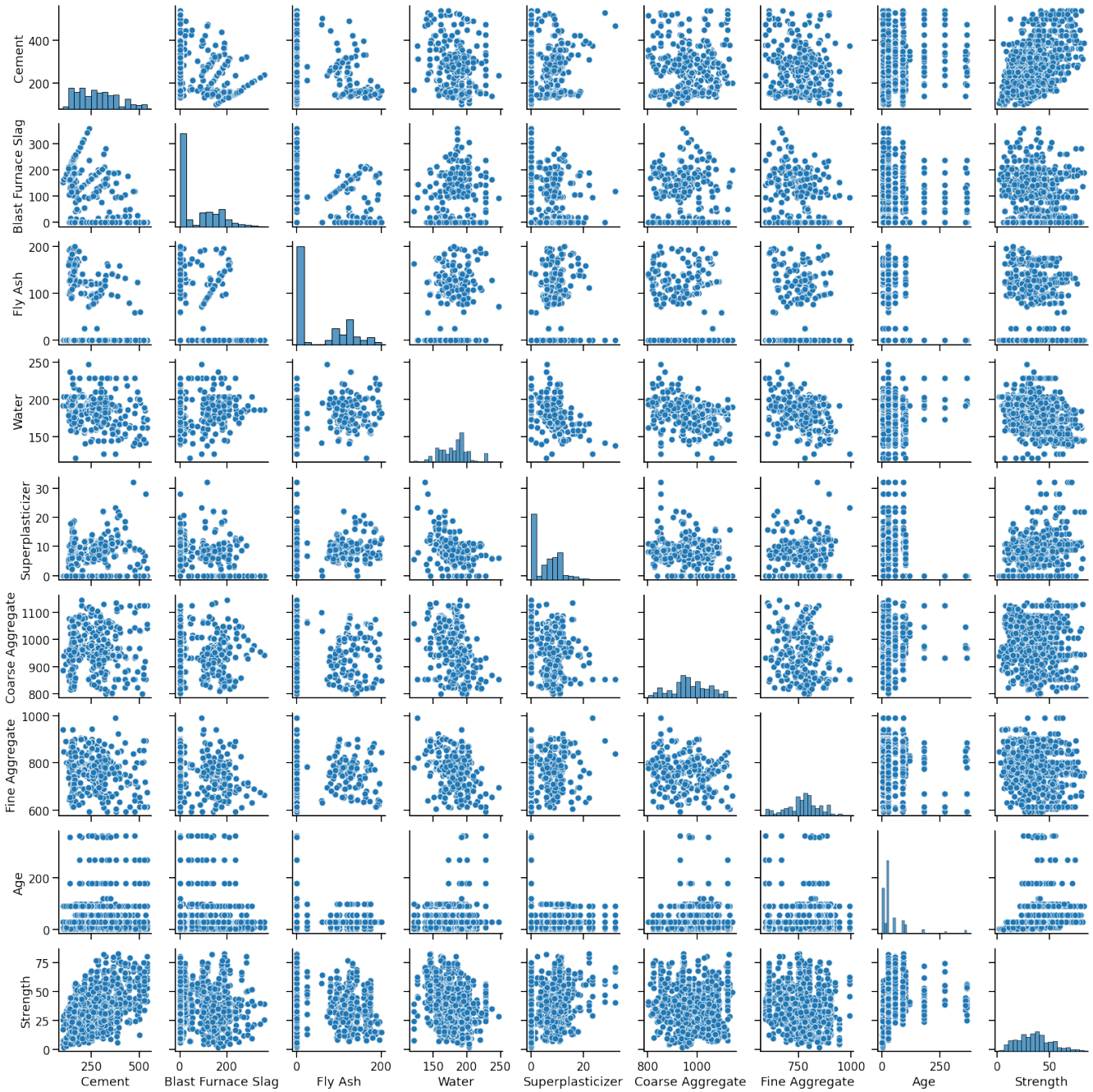
3.3 Feature engineering

In chemistry, we rarely have a situation that a product is a simple sum of substracts. The reactions and interactions between ingredients introduce non-linear effects. Therefore, we add polynomial terms of second and third degree to account for complex nature of material engineering.

Table 4. Summary statistics of data set.

	Cement	Blast f. s.	Fly ash	Water	Superplasticizer	Coarse aggregate	Fine aggregate	Age	Strength
mean	281	74	54	182	6	973	774	46	36
std	105	86	64	21	6	78	80	63	17
min	102	0	0	122	0	801	594	1	2
25%	192	0	0	165	0	932	731	7	24
50%	273	22	0	185	6	968	780	28	34
75%	350	143	118	192	10	1029	824	56	46
max	540	359	200	247	32	1145	993	365	83

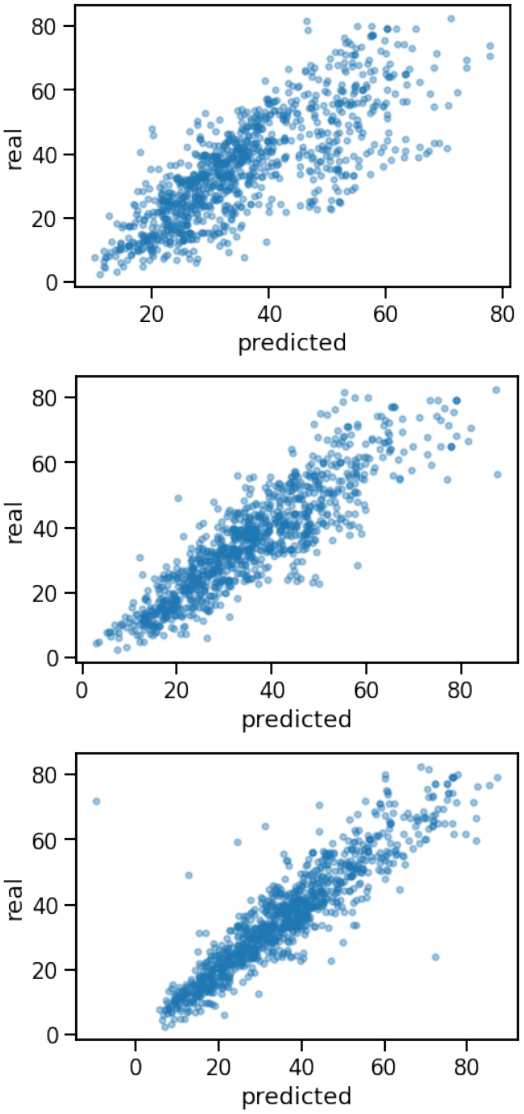
Figure 2. Relationships between variables.



4. Training linear regression models

In this section, we train different linear regression models to select the one that predicts concrete strength with the highest accuracy. To evaluate performance, we use the mean of R^2 values obtained with the 3-fold cross-validation method. We use R^2 score because it measures how much a model can explain the behavior of a target.

Figure 3. Predicted versus real values of concrete strength using basic linear regression model. We see results for model that do not include polynomial features (top), the model which includes second-order polynomial features (middle), and the model that includes third-order polynomial features (bottom).



4.1 Basic linear regression

Let us examine a basic linear regression model. We use only features that have been provided in the data set, without adding non-linear features. The data is scaled using the standard scaling approach, although it is not expected to change the results qualitatively here. With this method, we obtain the coefficient of determination equal to 0.61.

Next, we examine if we can obtain a higher R^2 score if we use polynomial features of the second and third degree. We obtained R^2 equal to 0.79 if we use second-degree polynomial features, and 0.83 if third-degree polynomial features are included. Although R^2 rises, suggesting that we decrease the bias, we have the problem with over-fitting. We observe that single but distant point appears in Figure 6 if third-order polynomial features are introduced. Thus, we need to examine models that introduce a penalty for high-value coefficients obtained from linear regression models.

4.2 LASSO regression

The LASSO regression modifies a cost function by adding a penalty proportional to the absolute value of coefficients. It is tuned by the α parameter. We need to find the best regularization strength using cross-validation. The results are presented in Figure 4.

This model reduces many coefficients to zero, and thus, it can be used to exclude some features. The model becomes easier to interpret, but here we aim at predicting the strength of concrete, and, thus, we need to examine another approach, that is, Ridge regression.

4.3 Ridge regression

The Ridge regression modifies a cost function by adding a penalty proportional to the square of coefficients. In this subsection, we perform a similar regularization procedure as for LASSO regression. The results are presented in Figure 5.

This model does not reduce coefficients to zero. We observe that adding polynomial features of the third degree requires a high value of α parameter. It suggests that third-level polynomial features introduce overfitting, and thus, the tuning parameter must be high to suppress this effect.

Figure 4. R^2 score from the 3-fold cross-validation evaluation using LASSO regression model. We see results for model that do not include polynomial features (top), the model which includes second-order polynomial features (middle), and the model that includes third-order polynomial features (bottom).

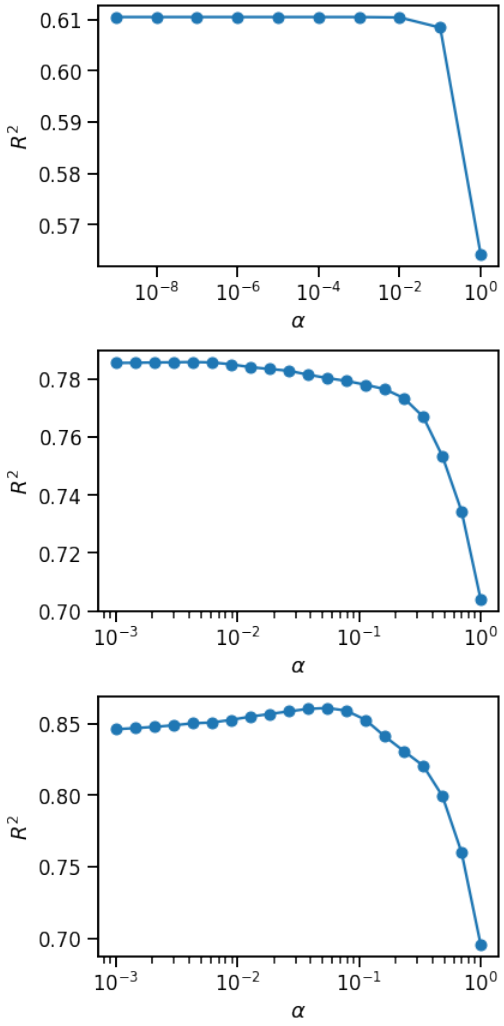
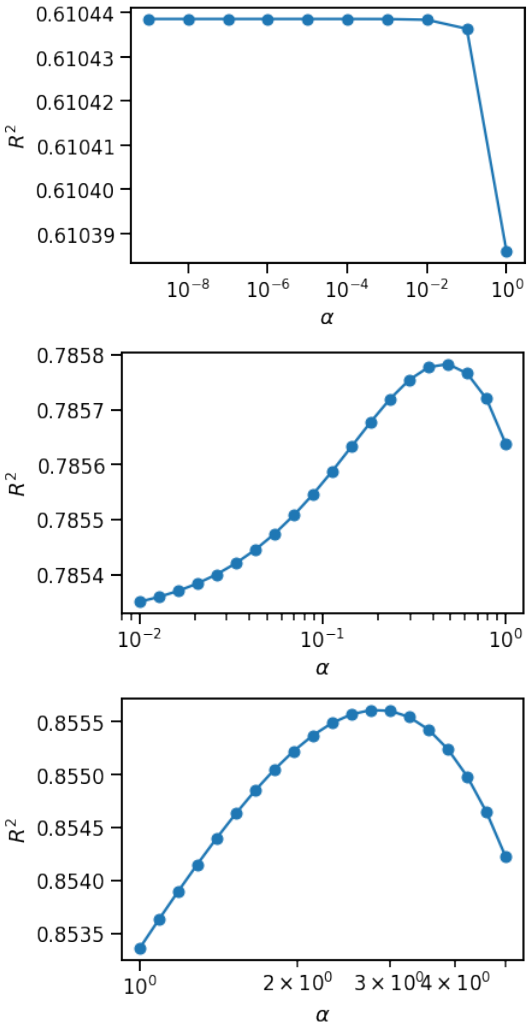


Figure 5. R^2 score from the 3-fold cross-validation evaluation using Ridge regression model. We see results for model that do not include polynomial features (top), the model which includes second-order polynomial features (middle), and the model that includes third-order polynomial features (bottom).



5. Selecting the best model

We trained three types of linear regression models – the basic model, LASSO model, and Ridge model. The results are presented in Table 5. We used plain data with polynomial features of second and third-degree, and no polynomial features as well. The data used in all models was standardized. Our purpose is to select the best model for predicting the strength of concrete. Thus, we aim at obtaining the best accuracy. We find that adding third-order polynomial features provides us a better model for interactions between concrete subtracts. However, it also introduces over-fitting. In the 3-fold cross-validation procedure, we found out that two models – LASSO model with $\alpha = 0.05$ and Ridge model with $\alpha = 3.1$ – provide accurate predictions. We found out that the Ridge model with $\alpha = 3.1$ trained with all data including third-order polynomial features provides the highest R^2 score which is equal to 0.856. The LASSO model with $\alpha = 0.05$ trained with all data including third-order polynomial features provides the highest R^2 score which is equal to 0.861. Also, LASSO model allows us to reduce a number of variables to prevent over-fitting that was a result of adding third-order polynomial features. Thus, this is a model we recommend for predicting a concrete

strength.

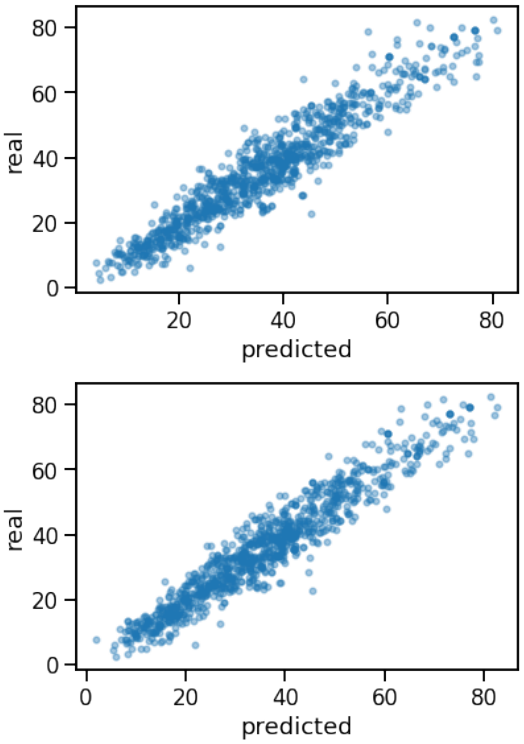
Table 5. R^2 score from the 3-fold cross-validation routine for various linear regression models trained on linear, 2-order, and 3-order polynomial features.

	Only linear	2-order	3-order
Basic	0.610	0.785	0.830
LASSO	0.610	0.786	0.861
Ridge	0.610	0.786	0.856

6. Key findings and insights

No coefficient is reduced to zero by a regularization process when only second-degree polynomial features are used. This observation suggests that a strength of concrete depends on all subtracts and interactions between them. Therefore, we expanded a model with polynomial features of third order.

Figure 6. Predicted versus real values of concrete strength using regression with LASSO ($\alpha = 0.05$, top) and Ridge ($\alpha = 3.1$, bottom) regularization model. The model includes third-order polynomial features.



Better insight into the reactions that occur in concrete is provided by LASSO model, which reduced a number of variables from 164 to 91. The coefficients with highest absolute values are listed in Table 6. We observe strong non-linear relationship between a strength of concrete and its age. The young concrete gets harder and harder when time passes, but this trend reverses after maturing. Blast furnace slag and cement are key factors for a strength of concrete, while adding too much water can severely damage the strength of concrete. With this model, the 0.861 of variance in the strength of concrete can be explained by its composition and age.

Table 6. Coefficients obtained from LASSO model with $\alpha = 0.05$ trained on data including third-order polynomial features.

Feature	Coefficient
Age x Age	-15.159481
Water	-6.763004
Superplasticizer x Age x Age	-5.189929
	...
Blast furnace slag	5.209482
Cement	6.788437
Age	20.530491

7. Next steps

Our best model has the R^2 score equal to 0.861, which suggests that there are other factors that affect a strength of concrete. This agree well with the general knowledge on chemistry, as we know that chemical reactions can be influenced by details of process and external conditions. To obtain better accuracy, we need to know the temperature, air pressure, air humidity, and the order of operations performed while mixing the subtracts. Also, the aging of concrete is affected by external conditions. The information on general climate features, that include information on minimal and maximal temperatures, and number of sunny and rainy days in a year can help us with predicting a strength of concrete with high accuracy. Another factor that introduces bias might be variables that are not normally distributed. Transforming the variables for which it is possible might slightly increase the accuracy of the model.