Concrete strength prediction

Zoltán Gyurkó
zoligyurko@gmail.com
Supervisor: András Zempléni

***Abstract*—The used dataset consists the strength of concrete and the components of concrete mix. The aim with this data could be to estimate the strength of concrete based on its components. Or, based on its strength, estimate some of the components' amount.**

I. Data

DATA is available here on Kaggle. The feature set includes the following:

- Cement [kg/m^3; *float64*]
- Blast Furnace Slag [kg/m^3; *float64*]
- Fly Ash [kg/m^3; *float64*]
- Water [kg/m^3; *float64*]
- Super-plasticizer [kg/m^3; *float64*]
- Coarse Aggregate [kg/m^3; *float64*]
- Fine Aggregate [kg/m^3; *float64*]
- Age [day (1 365); *int64*]

The target set is *Strength of the concrete* [N/mm^2;*float64*]. The dataset includes 1030 entries.

II. Data analysis

To analyze the data I have calculated the min, max, mean, and standard deviation values for each column. It seems that there are some very high values in some of the columns that are probably outliers. To test that, I have created a boxplot, which showed that there are outliers in:

- Blast furnace slag
- Water
- Superplasticizer
- and fine aggregate.

These values most probably have to be filtered out. In the case of *Age*, I would not call a higher value an outlier, as it just shows when the measurement was done.

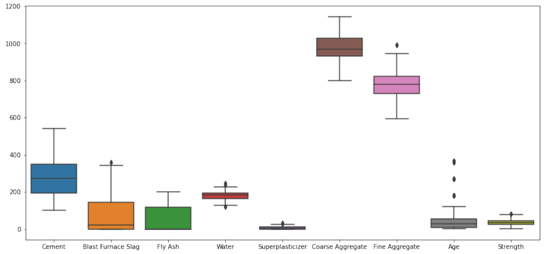

Figure 1. Boxplot.

Pairplots were also created to understand the relationship of each component. It was found that fly ash vs. cement and blast furnace slag vs. cement seems to have some relationship, which is not surprising as usually these materials are added to the mix as a percentage of cement. In fly ash, blast furnace slag, and superplasticizer, it can be seen that sometimes their values are zeros. This is not an error; these materials are optional components of the concrete mix.

A. Feature correlation

I have analyzed on a heatmap the correlation of the features of the data set. Strength has a strong correlation with cement, superplasticizer, age, and negative correlation with water. The other factors have some effect as well, but significantly smaller. Superplasticizer has a strong correlation with water, which is, again, not a surprise, as superplasticizer is mainly added to the mix if the amount of water is low. I have visualized the most promising correlations (e.g. cement vs. strength).

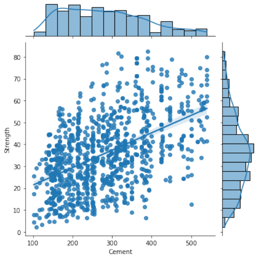

Figure 2. Jointplot of cement amount vs. concrete strength.

III. Literature review

I have performed a brief literature review to identify what is known and what we should expect. It was interesting that most research focused on the effect of one or two components but not more and rarely investigated the cross effects. I have summarized the result in a separate document. Shortly:

- With the increase of superplasticizer the strength is decreasing
- with age the strength is increasing (but always less and less)
- with the fine aggregate the strength is increasing
- with the increase of the water/cement ratio the strength is decreasing
- with the increase of cement amount the strength is increasing until a point when the concrete becomes oversaturated - above that the strength starts to decrease
- slag has a similar effect to cement

This document was created for the Mathematical Modelling Practice course of ELTE's Mathematics Expert in Data Analytics and Machine Learning postgraduate specialization program.

- for fly ash I have found contradictory results

Many researchers point out that it is worth introducing a new variable: the water-to-cement ratio because it has a high correlation to concrete strength. Similarly to aggregate-to-cement ratio and fineness (the ratio of the fine and coarse aggregates). Another interesting finding was that concrete behaves differently in the very low strength (below 8 N/mm^2) area and in the very high strength area (above 65 N/mm^2). It might be worth handling these cases separately. It was also identified that the data set is not complete as some crucial parameters (like the strength of the cement or the type of aggregate) are missing. These parameters would highly influence the strength of concrete.

IV. Further plans

1) **Data cleaning:** removing outliers, physically impossible values, NaNs. Besides that it shall be somehow decided if we want to separate the very low and high strength values and maybe handle them is a separate model. As well as it shall be decided how to handle the valid zero values in fly ash, slag, and superplasticizer.
2) **Feature engineering:** introduction of new variables (composite features); they will be the water-to-cement ratio, aggregate-to-cement ratio, and fineness.
3) **Model creation:** the problem itself is a regression problem, however, it has some classification features (e.g. slag content zero or not) so I would test regression models like random forest regression. Besides that maybe a good idea could be to test a mixture of linear models (e.g. a linear model of all components). Or not linear when we know from the literature or from the data that the relationship is not linear. Maybe it would be worth building a pipeline around that. First, do a kind of classification and then the regression.
4) **Evaluation of the results:** calculate the most important measures (e.g. errors) to evaluate the models.
5) **Add-on 1:** if there is time I would introduce strength classes and would try to estimate only them, not the exact values because in practice this is what is needed.