Featurization and Model Tuning Project

Data Description:

The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory. Data is in raw form (not scaled). The data has 8 quantitative input variables, and 1 quantitative output variable, and 1030 instances (observations).

Domain:

Cement manufacturing

Context:

Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.

Attribute Information:

• Cement	: measured in kg in a m3 mixture
• Blast	: measured in kg in a m3 mixture
• Fly ash	: measured in kg in a m3 mixture
• Water	: measured in kg in a m3 mixture
• Superplasticizer	: measured in kg in a m3 mixture
• Coarse Aggregate	: measured in kg in a m3 mixture
• Fine Aggregate	: measured in kg in a m3 mixture
• Age	: day (1~365)
• Concrete compressive strength	measured in MPa

Learning Outcomes:

- Exploratory Data Analysis
- Building ML models for regression
- Hyper parameter tuning

Objective:

Modeling of strength of high performance concrete using Machine Learning

Steps and tasks:

1. Deliverable -1 (Exploratory data quality report reflecting the following) (20 marks)
    a. Univariate analysis (5 marks)
        i. Univariate analysis – data types and description of the independent attributes which should include (name, meaning, range of values observed, central values (mean and median), standard deviation and quartiles, analysis of the body of distributions / tails, missing values, outliers
    b. Multivariate analysis (5 marks)
        i. Bi-variate analysis between the predictor variables and between the predictor variables and target column. Comment on your findings in terms of their relationship and degree of relation if any. Presence of leverage points. Visualize the analysis using boxplots and pair plots, histograms or density curves. Select the most appropriate attributes
    c. Pick one strategy to address the presence outliers and missing values and perform necessary imputation (10 marks)
2. Deliverable -2 (Feature Engineering techniques) (15 marks)
    a. Identify opportunities (if any) to create a composite feature, drop a feature etc. (5 marks)
    b. Decide on complexity of the model, should it be simple linear model in terms of parameters or would a quadratic or higher degree help (5 marks)
    c. Explore for gaussians. If data is likely to be a mix of gaussians, explore individual clusters and present your findings in terms of the independent attributes and their suitability to predict strength (5 marks)
3. Deliverable -3 (create the model) (15 marks)
    a. Obtain feature importance for the individual features and present your findings
4. Deliverable -4 (Tuning the model) (20 marks)
    a. Algorithms that you think will be suitable for this project (5 marks)

b. Techniques employed to squeeze that extra performance out of the model without making it overfit or underfit (5 marks)
c. Model performance range at 95% confidence level (10 marks)

References:

- Medium article on hyper parameter tuning