INNOMATICS RESEARCH LABS

Module: Machine learning project – Regression Problem

Title: Concrete compressive strength

Data Set Information:

Number of instances 1030
Number of Attributes 9
Attribute breakdown 8 quantitative input variables, and 1 quantitative output variable
Missing Attribute Values None

Click here to download the dataset.

Attribute Information:

Given are the variable name, variable type, the measurement unit and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

Name	Data Type	Measurement	Description
Cement (Component 1)	quantitative	kg in a m3 mixture	Input Variable
Blast Furnace Slag (Component 2)	quantitative	kg in a m3 mixture	Input Variable
Fly Ash (Component 3)	quantitative	kg in a m3 mixture	Input Variable
Water (Component 4)	quantitative	kg in a m3 mixture	Input Variable
Superplasticizer (Component 5)	quantitative	kg in a m3 mixture	Input Variable
Coarse Aggregate (Component 6)	quantitative	kg in a m3 mixture	Input Variable
Fine Aggregate (Component 7)	quantitative	kg in a m3 mixture	Input Variable
Age	quantitative	Day (1~365)	Input Variable
Concrete compressive strength	quantitative	MPa	Output Variable

INNOMATICS RESEARCH LABS

Steps to perform in Project:

1. Read the dataset into the notebook
2. Print the shape of the data
3. List out the feature variables and their data-types
4. List out response variable and its data type
5. Perform univariate analysis (be as creative as possible in your analysis)
   - Visualize the shape of the distribution of data. Is every feature variable normally distributed? Why is normal distribution important for data?
   - Check for null values in the feature variables
   - Draw box and whiskers plot of each of the feature variables
   - Check for outliers
   - Is the data distribution skewed? If highly skewed, do you still find outliers which you did not treat?
   - How do the distributions look in terms of variation? Which features are widely spread and which are kind of concentrated towards the mean?
6. Treat outliers. What is your strategy? What other strategies can be used?
7. Perform bi-variate analysis (be as creative as possible)
   - Try creating correlation matrices. See if there are variables which are strongly or weakly related
   - If there are variables showing high correlation, what corrective action is needed? Why is this matter of concern? What if we do not treat the variables showing high degree of correlation?
8. What is the type of machine learning problem at hand? (Supervised or Unsupervised?) Why?
9. What is the category of the machine learning problem at hand? (Classification or Regression?) Why?
10. Perform below algorithms:
    - Linear Regression
    - Lasso Regression
    - Ridge Regression
    - Decision Tree Regressor
    - Random Forest Regressor
    - KNN Regressor
    - SVM Regressor
11. Pick each of the algorithm and perform the below steps:
    - Split your data between train and test steps.
    - Build your model
    - List down the evaluation metrics you would use to evaluate the performance of the model?
    - Evaluate the model on training data
    - Predict the response variables for the test data
    - How are the two scores? Are they significantly different? Are they the same? Is the test score better than the training score?

- Perform hyper parameter tuning and cross validation techniques.
- Evaluate the model on test data.

12. Which algorithm performs better on this dataset and Why?