Exploratory Data Analysis

Exploratory data analysis (EDA) is used to analyze and examine data sets and summarize their key characteristics, often using data visualization methods. It helps determine how best to manipulate data sources in order to get the answers you need, making it much easier to discover and recognize patterns, spot variances, test a hypothesis, or check assumptions. While employing EDA, we checked for missing values, which were none. Furthermore, the distribution of the input and output attributes were checked and most of them were nearly representing a normal distribution, thus not requiring any further modifications. Next, the correlation, in-between the parameters and between the parameters and the output label, was checked. Only the quantity of cement and the Concrete Strength had a slight correlation, but it was deemed ignorable.

Methodology of models

A number of machine learning models were tried and tested for predicting the concrete strength from the provided parameters. A comparative analysis was carried out between the various models employed and the model providing the highest accuracy was further tuned for the given data.

1. **Linear Regression:**

Linear regression is conceivably one of the most eminent and well understood algorithms in statistics and machine learning. The principal idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error is the least. Error is defined as the distance between the points to the regression line. Mathematically,

Y(pred) = b0 + b1*x

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

2. **Lasso Regression:**

Lasso Regression is an extension or variant of linear regression. The main difference between these two algorithms is that Linear Regression gives the regression coefficients directly from the dataset whereas Lasso Regression uses a method called “shrinkage” that compresses the parameters towards zero. This method in a way regularizes the coefficients.It is advantageous in two ways. It prevents overfitting and it helps to apply this on different datasets.

Mathematical Equation of Lasso Regression:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

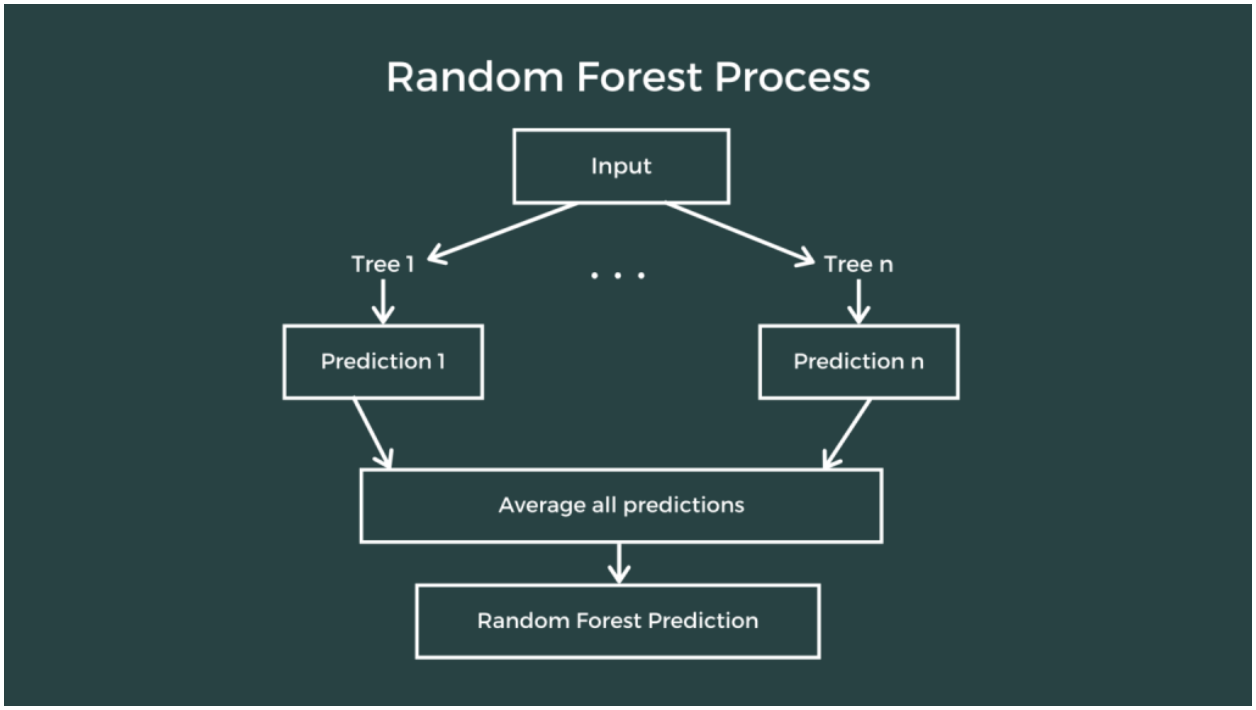
λ is the shrinkage parameter.

3. **Ridge Regression:**

   Ridge Regression is also a variant of Linear Regression. The singular part of this method is that it is used mainly to deal with problems of multicollinearity. This method also penalizes the coefficients i.e regularization in order to prevent overfitting and get more accurate results.

4. **Random Forest Regressor:**

   The Random Forest Regressor model uses the concept of ensemble learning. Ensemble learning means the model combines the eclectic predictions from various Machine Learning models to get a more accurate prediction. As shown in the figure below, the tree is fed with input then it is split into n number of trees at each layer .In the final layer, the average of all predictions gathered from various models is taken in account as the final prediction.



5. **<u>Decision Tree Regressor:</u>**

   A decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model gets confident enough to make a single prediction. The order of the question as well as their content are being determined by the model. In addition, the questions asked are all in a True/False form.

6. **<u>Gradient Boost Regressor:</u>**

   Gradient boosting is one of the variants of ensemble methods where you create multiple weak models and combine them to get better performance as a whole.

7. **<u>Bagging Regressor:</u>**

   A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

8. **AdaBoost Regressor:**

   AdaBoost Regressor is a method that works by combining multiple weak classifiers into one strong and reliable classifier. The uniqueness of this method is that it assigns more weight on instances which are difficult to predict and less weights on instances that are tough to predict.

9. **KNN Regressor:**

   The KNN algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. In the final step the prediction is the closest data point in the set. The ‘n’ in KNN stands for the number of neighbors. In this case the final prediction is the average of ‘n’ neighbors.

10. **Extra Tree Regressor:**

    The extra tree regressor fits a number of randomized decision trees i.e. extra trees on various sub samples of the dataset. The metric to find the prediction used here is averaging. This improves accuracy significantly and reduces the chances of overfitting. Similar to the Random Forest approach, the Extra Trees algorithm will randomly sample the features at every split point of a decision tree. But, where it differs from Random Forest algorithm is that Random Forest algorithm employs a greedy algorithm to select an optimal split point, whereas the Extra Trees algorithm selects a split point at random.