CAPSTONE PROJECT

Predicting Compressive strength of concrete

PROJECT OVERVIEW:

Concrete has a versatile use in the construction practice for its availability, cheap rate, flexibility of handling and giving shape to any desired form. Designing a concrete structure requires the concrete compressive strength to be used. The design strength of the concrete normally represents its 28th day strength. In case of construction work 28 days is considerable time to wait for the test results of concrete strength, while it also represents the quality control process of concrete mixing, placing, proper curing etc. Hence, a rapid and reliable concrete strength prediction would be of great significance. So, I chosen this project to predict the compressive strength of concrete which is useful for industries to classify them into grades.

Link for the data set in Kaggle: https://www.kaggle.com/pavanraj159/concrete-compressivestrength-data-set.

From the following research papers I understood the importance of problem.
http://dergipark.gov.tr/download/article-file/217736.
http://www.iebconferences.info/haspre.pdf

Using the data effectively we can find the strength of concrete which is important material in industries.

PROBLEM STATEMENT:

- By accurately predicting the strength of concrete.

- By using the dataset, the task is to predict the Concrete compressive strength score that tells the strength of concrete. The model utilizes the important characteristics of the data to develop models that can predict the scores.

- By using machine learning techniques we can predict the strength of concrete. Several steps are involved in the project like Data exploration, Data processing and finally testing various algorithms and techniques.

- By considering the given problem domain I will apply several machine learning supervised models to figure out the concrete compressive strength.

- Initially, I will apply bench mark model for the training data and calculate its performance.

- I will eventually work out on diverse models such as SVM's, Gradient boost and Random Forests etc., and decide which one is the best model used on the performances and best model will be optimized further using GridSearchCV.

- Finally I will compare the performances of the benchmark model and optimized model and come up to a conclusion about the best model that could potentially fit the input data.

METRICS:

Problem is a regression task, since it takes certain features as inputs and figures out a score that determines the concrete compressive strength.

Hence, I decided to use Coefficient of determination(R2 score) as the performance metric that could be used to check the performance of the scores obtained from the Bench Mark Model and the Optimal Model considered.

The Coefficient of Determination(R^2) is the key output of the Regression Analysis. It can be defined as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The value of R^2 == 0, tells that the model is a worst fit to the given data.
- The value of R^2 == 1, tells that the model is the best fit to the given data.

The formula for Coefficient Of Determination(R^2) is given by:

If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured using three sums of squares formulas:

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The regression sum of squares, also called the explained sum of squares:

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Reference:
https://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination

DATA EXPLORATION:

Data exploration is the important part in any kind of analysis tasks. It tells about the Data and its features. These features and description of data can be used for further analysis.

In this phase, I first acquired the data which I collected from Kaggle.

These are few data points:

	Cement	Blast_Furnace_Slag	Fly_Ash	Water	Superplasticizer	Coarse_Aggregate	Fine_Aggregate	Age	Concrete_compressive_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

Later, I found some of useful description of data like median, mean, count etc.

	Cement	Blast_Furnace_Slag	Fly_Ash	Water	Superplasticizer	Coarse_Aggregate	Fine_Aggregate	Age	Concrete_compressive_stre
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.00
mean	281.167864	73.895825	54.188350	181.567282	6.204660	972.918932	773.580485	45.662136	35.81
std	104.506364	86.279342	63.997004	21.354219	5.973841	77.753954	80.175980	63.169912	16.70
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.33
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.71
50%	272.900000	22.000000	0.000000	185.000000	6.400000	968.000000	779.500000	28.000000	34.44
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.13
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.60

Since, the main goal is construct a model that has the capability to predict strength of concrete , we need to divide the data into features and target variable.

Description of features:

Concrete is inert mass which grows from a cementing medium. Concrete is a product of two major components, one is the cement paste and another is the inert mass. In order to form the cementing medium, cement would mix with water. Coarse aggregates and fine aggregates are the part of inert mass.

Concrete compressive strength depends on the following features.

1) Cement (component 1) -- quantitative -- kg in a m3 mixture
2) Blast Furnace Slag (component 2) -- quantitative -- kg in a m3 mixture
3) Fly Ash (component 3) -- quantitative -- kg in a m3 mixture
4) Water (component 4) -- quantitative -- kg in a m3 mixture
5) Superplasticizer (component 5) -- quantitative -- kg in a m3 mixture
6) Coarse Aggregate (component 6) -- quantitative -- kg in a m3 mixture
7) Fine Aggregate (component 7) -- quantitative -- kg in a m3 mixture
8) Age -- quantitative -- Day (1~365)

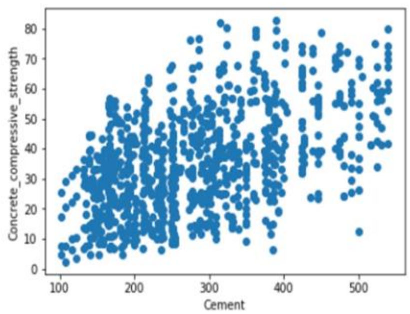
EXPLORATORY VISUALIZATION:

Exploratory Visualization is very important, because by visualizations we can know characteristics of data very clearly.

These visualizations tells about how features are correlated with the target variable.

Plot for Cement:

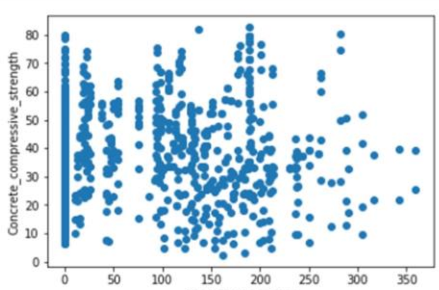
Plot shows high positive correlation between cement and compressive strength



Scatter plot showing that Cement and Concrete compressive strength are highly correlated. So,Cement will be the main feature to predict the strength

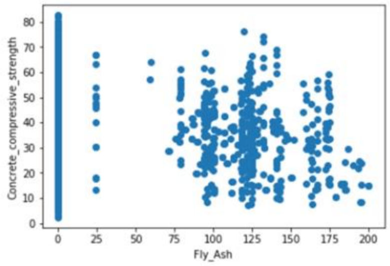
Plot for Blast_Furnace_Slag :

In this plot we cannot say that blast furnace slag and compressive strength highly positive correlated but we can say that there is impact of blast furnace slag on strength of concrete.



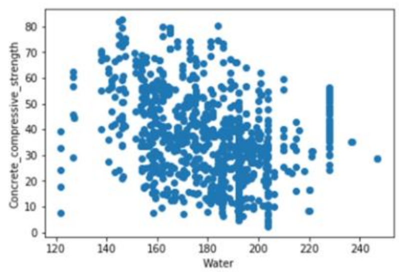
Blast furnace slag is one of the component in the preparation concrete. If we see the scatter plot, range of 100 and 200 and 0 are having more datapoints.

Plot for Fly_Ash:



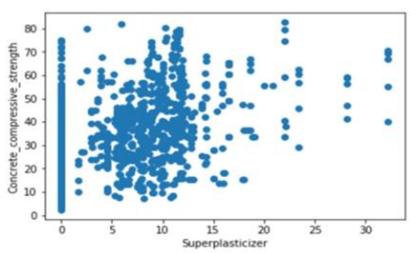
If concrete contains less amount of Fly ash then concrete will be no delay in hardening of concrete.If we see the scatter it also shows more datapoints at 0.

Plot for Water:



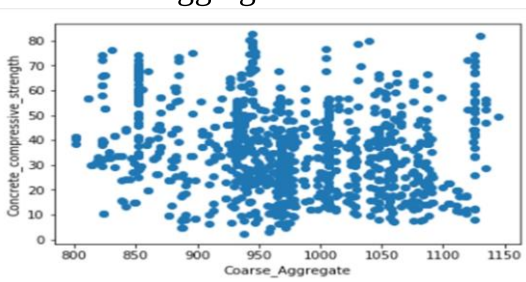
Water is one of the main ingredient in preparation of concrete.Efficient amount of water is needed for good strength of concrete.Scatter plot also tells the same point that water is highly important.

Plot for superplasticizer:



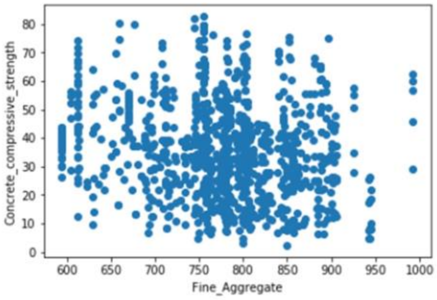
Superplasticizer is nothing but water reducers.They reduce water by 40%. Sactterplot showing correlation so it also an important feature for predicting strength of concrete.

Plot for Coarse Aggregate:



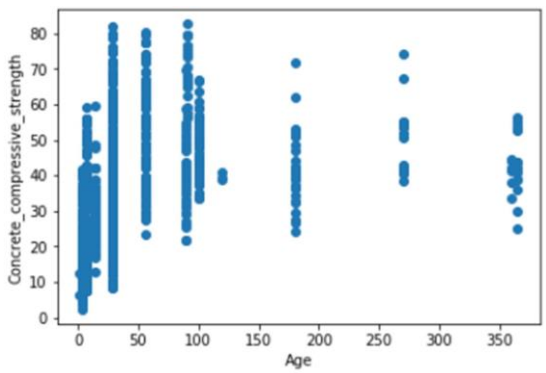
Coarse_Aggregate is highly correlated with compressive strength.So, it is most important feature

Plot for Fine Aggregate:



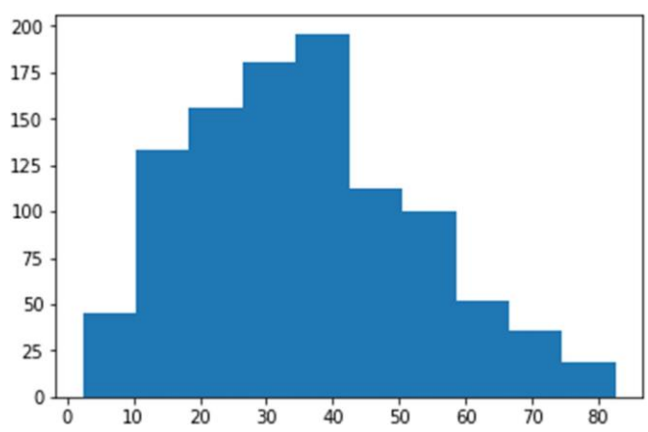
Scatterplot shows high correlation of Fine aggregate with concrete compressive strength.So,it is also one important main feature in predicting strength of concrete

Plot for Age:

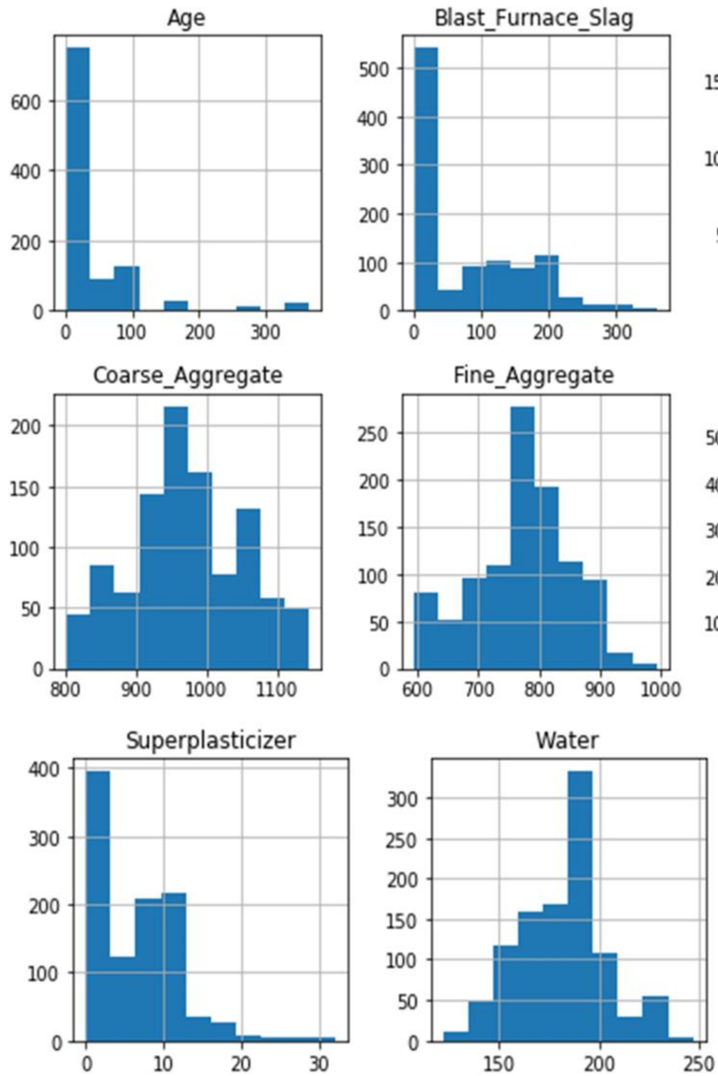


It describes a sheer intuition that the Age shows a striking factor in influencing strength of concrete.

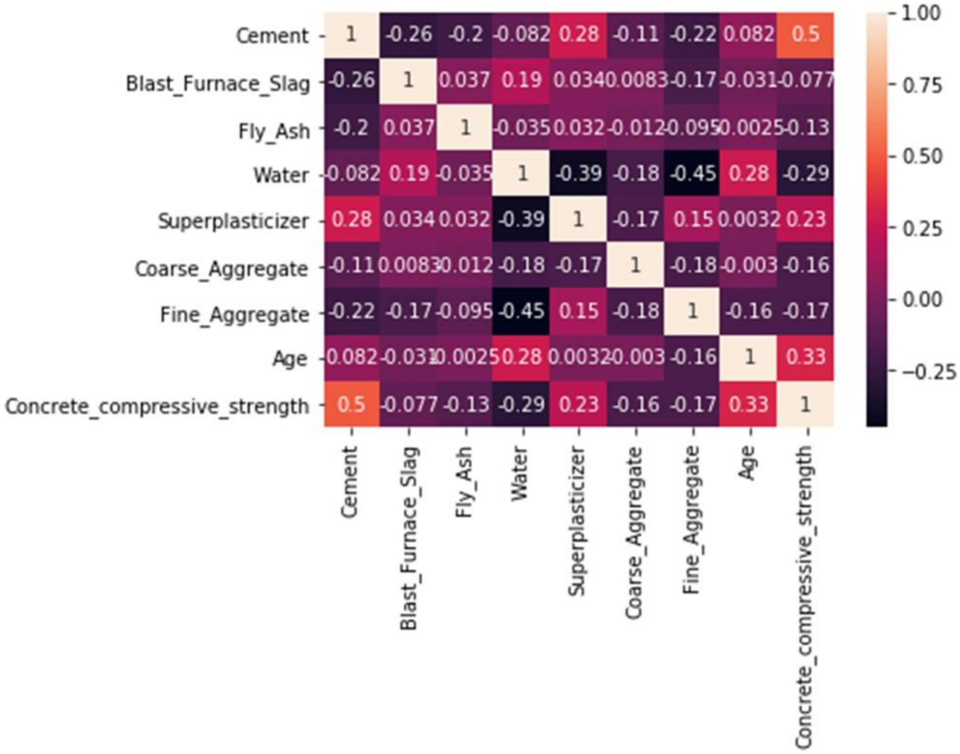
This is the histogram for target label (Concrete compressive strength). I drew this histogram to know spread of values.



These are histograms of model features:



Heat map which shows us the correlation between features.



ALGORITHMS AND TECHNIQUES:

By considering the given problem domain I will apply several machine learning supervised models to figure out the concrete compressive strength.

They are :-

- Support Vector Machines (SVM)
- Ensemble Methods- Randomboost
- Ensemble Methods-Gradientboost

Support Vector Machines :-

- SVM's are simple, accurate and perform well on smaller and cleaner datasets. It can be more efficient as it uses subset of training points.
- The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. Initially as the output is a real number and continuous it becomes very difficult to predict the information at hand, which has infinite possibilities.
- In the case of regression, the factor margin of tolerance (epsilon) is set in approximation to the SVM.
- The main theme for SVM is always to minimize error, particularize the hyperplane which maximizes the margin, keeping in mind that part of the error is convinced.
- Since For the current regression problem consists of a lot of continuous features the application of SVM's can serve a better purpose.

(Ref: https://data-flair.training/blogs/applications-of-svm/)

Ensemble Methods –Gradientboost:

Gradient Boosting Regresor is a ensemble algorithm for both classification and regression.GBTs build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive. There are typically three parameters - number of trees, depth of trees and learning rate, and the each tree built is generally shallow. Although it may seem GBDTs are better than random forests, GBDTs are prone to overfitting, however there are strategies to overcome same and build more generalized trees using a combination of parameters like learning rate (shrinkage) and depth of tree. Generally the two parameters are kept on the lower side to allow for slow learning and better generalization.

Ensemble Methods - Random Forests:

- Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes in case of the classification tasks or mean prediction for the regression tasks of the individual trees.
- Since Random Forests perform well on almost every machine learning problem and they also show less over fit behaviour when compared to Decision Trees. Since our problem is composed of a lot of continuous features for which Random Forests serve a better choice. (https://en.wikipedia.org/wiki/Random_forest)

Model Tuning:

In this part of the project, I'll apply the GridSearchCV technique to further optimize the best model that was selected from the three supervised learning models stated above.

In this tuning process, hyper parameter I have used is n_estimators. n_estimators tells number of trees in the forest.

GridSearchCV:-

For tuning hyperparameters we use GridSearchCV technique

I will then use the performance metric (r2_score) and compare the three models, model which has the best r2_score will be considered for further analysis.

I'll optimize the selected model by 'GridSearchCV' and evaluate the model by comparing the final r2_score of the optimized model and the benchmark model.

BENCHMARK :

1) Since the given problem expects to predict a continuous output, to determine a metric value(R2_score) that will help us to establish a comparison between the performances of the Bench Mark Model and the Optimal Model .

2) The Bench Mark model is a base model, Since the task is regression I used Linear regressor as Benchmark model, and we use r2_score as performance metric . The coefficient of determination (r2_score) for a model is a useful statistic in regression analysis, as it often describes how "good" that model is at making predictions.

3) The values for R2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable.

R2_Score after applying linear regression is: 0.574.That means 57.4% variance.

DATA PREPROCESSING:

Tasks in data pre-processing:-

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.
- Data integration: using multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.

In our data many of the fields contain missing values they are filled with 0s.

Count of missing values is as follows:

```
Cement                            0
Fly_Ash                         566
Water                             0
Superplasticizer                379
Blast_Furnace_Slag              471
Coarse_Aggregate                  0
Fine_Aggregate                    0
Age                               0
Concrete_compressive_strength     0
```

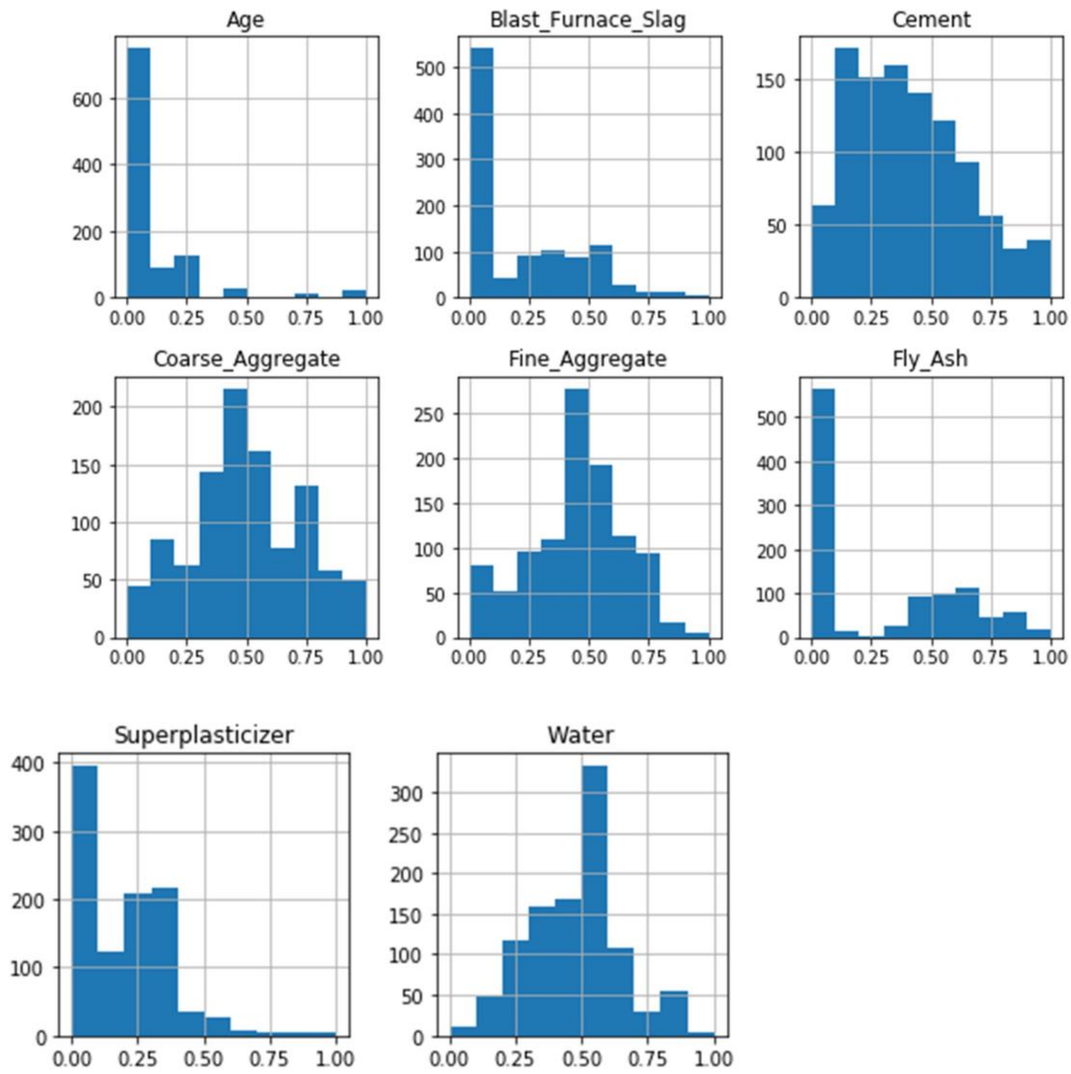
We have to remove those missing values. For that I am replacing 0s with mean value of that column.

This the data after removing missing values.

	Cement	Blast_Furnace_Slag	Fly_Ash	Water	Superplasticizer	Coarse_Aggregate	Fine_Aggregate	Age	Concrete_compressive_strength
0	540.0	136.158676	120.288793	162.0	2.500000	1040.0	676.0	28	79.99
1	540.0	136.158676	120.288793	162.0	2.500000	1055.0	676.0	28	61.89
2	332.5	142.500000	120.288793	228.0	9.816897	932.0	594.0	270	40.27
3	332.5	142.500000	120.288793	228.0	9.816897	932.0	594.0	365	41.05
4	198.6	132.400000	120.288793	192.0	9.816897	978.4	825.5	360	44.30

MinMaxScaler is used to normalize the data. MinMaxScaler Normalize values to the range of 0 and 1.

These are histograms of features after normalization.



Now all the values of features are ranging between 0 and 1.

IMPLEMENTATION:

Select the best model out of 3 models which I mentioned above by calculating r2_score for each model.

Firstly, divide the dataset into training and testing data. After implement Benchmark model and calculate r2_score. Next apply these 3 models on training and testing set. And then calculate the r2_score of each model and then compare each other's score and then pick the model which is having high r2_score.

R2_Score for SVR: 0.294

R2_score for Gradientboosting Regressor: 0.918

R2_Score for Random Forest Regressor: 0.916

Out of three models, Gradientboosting Regressor is having high r2_score. So, Gradientboosting Regressor model is the best model.

REFINEMENT:

In this section of the project, the model (Gradientboosting Regressor Model) is optimized by the applying GridSearchCV technique for fine tuning the parameters

In this tuning process, hyper parameter I have used is n_estimators. n_estimators tells number of trees in the forest.

Gradientboosting Regressor Model showed a best improvement upon Model Tuning using the GridSearchCV technique.

It produces r2_score of 0.93.

Table shows the results before and after tuning.

Metric	Gradientboosting Regressor (before tuning)	Gradientboosting Regressor(after tuning)
R2_score	0.918	0.93

MODEL EVALUATION AND VALIDATION:

Table shows the results of benchmark model and optimized model.

Metric	Linear regressor	Gradientboosting Regressor
R2_score	0.574	0.93

It shows that random forest regression is best optimized model for predicting the concrete compressive strength.

JUSTIFICATION:

By observing the validation results above, it is quite evident that the model is performing well on the given data.

Outliers can be bad for boosting because boosting builds each tree on previous trees' residuals/errors. Outliers will have much larger residuals than non-outliers, so gradient boosting will focus a disproportionate amount of its attention on those points. So, before performing boosting data must be free from outliers. For, this dataset there are no outliers, we can know that by seeing plots. Each and every feature is mostly correlated with the label. So, I feel that this model will work fine.

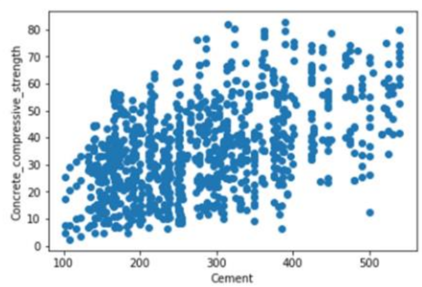
When compared to the BenchMark Model, the Optimal Model (Gradientboosting Regressor Model') shows a performance as shown above.

- The performance of the BenchMark Model is 57 % and that of the Optimal Model is 93%.

Since, we got very good R2_score when we apply it to a real world application we get almost accurate result.

FREE-FORM VISUALIZATION:

All features are important to predict the strength of concrete .Out of all Cement is having high quality.
If we see the below scatter plot, as the values of cement increases Concrete compressive strength(target) increases.



Scatter plot showing that Cement and Concrete compressive strength are highly correlated. So,Cement will be the main feature to predict the strength

• Data Acquisition: First step of the project is acquiring data which I collected from the Kaggle.
• Data Exploration and Visualization: Exploring and visualizing the dataset, cleaning and observing the every feature.
• Data Pre-Processing: Data Pre-processing is the key factor in Machine-Learning, in this project I used this process to clean and remove any irrelevant data.
• Model Evaluation and Validation: Model is tested for the performance in its Benchmark State and the Optimized State respectively.
• Optimization: The model is optimized by the application of GRIDSEARCHCV which helps in tuning the parameters optimal for the model

REFLECTION:

These are the steps I followed:

1) I collected data from kaggle (link provided in the overview phase).
2) I explored the data and I found some values of features are filled with 0s and visualized the relation between each feature to the target variable.
3) In the next pre-processing step I replaced all missing values with the mean of that column and then I normalized the data.
4) Now, my data is cleaned. Next I split the data into training and testing test.
5) Since the problem is regression problem, I used linear regression as bench mark and then calculated r2_score.
6) Next I chosen 3 supervised algorithms (SVR, GradientBoost, RANDOM FOREST) for predicting strength and calculated r2_score. At last I opted model which is having high r2_score that is Gradientboosting Regressor
7) I tuned the best model using GridSearchCV technique. Finally r2_score is calculated after tuning.
8) Lastly, I compared optimized model to the bench mark.
9) Results are as I expected optimized model highly accurate than benchmark model.

IMPROVEMENT:

Not only using regressors we can also use multi-layer feed-forward neural networks (MFNNs) for predicting compressive strength of concrete. A MFNN works well for the complex nonlinear relationship between the inputs (many factors that influence concrete strength) and the output (concrete strength). The neural network (NN) models give high prediction accuracy.