MACHINE LEARNING ENGINEER NANODEGREE

June 27, 2018
CAPSTONE PROPOSAL
YESWITHA

Predicting Compressive strength of concrete

DOMAIN BACKGROUND:

• Compressive strength or compression strength is the capacity of a material or structure to withstand loads tending to reduce size, as opposed to tensile strength, which withstands loads tending to elongate.

Compressive strength is one of the most important engineering properties of concrete. It is a standard industrial practice that the concrete is classified based on grades.

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

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

• The major motivation for exploring this project is to develop the skill and understanding how to work out on real time datasets and how Machine-Learning is potentially making the lives of people easier.

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

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.

DATASET AND INPUTS:

I took dataset from kaggle: https://www.kaggle.com/pavanraj159/concrete-compressive-strength-data-set.

There are 9 columns and 1030 data points in dataset.

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)
9) Concrete compressive strength -- quantitative -- MPa

SOLUTION STATEMENT:

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

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

3) I will eventually work out on diverse models such as SVM's, Decision Trees 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.

4) 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.

BENCHMARK MODEL:

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, hence we use r2_score . The coefficient of determination 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 R^2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable.

EVALUATION METRICS:

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

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

PROJECT DESIGN:-

The project will be designed in the following steps:

• 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.