Predicting Compressive Strength of High Performance Concrete

Sunidhi Taneja

ID-20804148

A project report submitted for the partial fulfilment of course STAT 844, Spring term

University of Waterloo
August 2019

1 Introduction

Concrete is the most widely used construction material on earth, mainly due to its low cost, durability, availability, fire resistance and long service life. High-performance concrete (HPC) is a terminology used for concrete that conforms to a set of standards above those of the conventional concrete, not limited to strength. In addition to the basic ingredients used in conventional concrete, i.e., cement, water, coarse and fine aggregates, HPC incorporates supplementary materials such as fly ash, blast furnace slag and superplasticizer. Even though the recipe for HPC seems fairly simple with its components, tremendous amount of complexity is involved in selecting the exact proportions of the constituents in order to simulate the desired performance.

Usually, compressive strength tests of concrete are conducted 7 to 28 days after pouring the concrete mix. Neglecting these tests might compromise on the quality control but the 28-days waiting period can delay the construction process. Thus, rapid and reliable prediction of concrete compressive strength is really important for quality control as well as predesign. It gives an option to do the essential modifications in the constituent proportions so as to avoid the circumstances where concrete does not attain the required design levels. It enables us to make economic use of raw materials and have fewer construction failures, thus reducing construction costs and loss of lives along with time. The early prediction of concrete compressive strength can also give us an idea about the time for concrete form removal and project scheduling.

The Abrams' water-cement ratio law (proposed in 1918) is considered as the most significant advancement in the field of concrete technology. It states that the strength of concrete is inversely proportional to the ratio of water to cement (w/c). It implies that as long as water to cement ratios of several comparable concrete mixes are same, their strengths will be same independent of the quantities of water and cement and even any other ingredients present in the mixture.

Several studies have independently shown that concrete strength is determined not only by w/c ratio but quantities of other ingredients as well. But the empirical equations used in codes and standards for estimation of concrete compressive strength are based on tests of concrete mixes without supplementary ingredients. Thus, validity of these relationships for concrete mixes with supplementary ingredients is questionable. It is really important to understand the relationship between composition of concrete and its strength to optimize the concrete mixture.

This study aims at using various regression models to accurately predict the compressive strength of HPC at varying ages. Similar works in the past employ traditional linear regression and artificial neural networks [1] and support vector machines (SVM), bagging regression trees (BRT) and multiple additive regression trees (MART)[2]. This study exploits regression techniques, namely multiple additive regression splines (MARS), random forests and gradient boosting.

The study will utilise the dataset in two different ways: one of them will consider water and cement as two independent input variables while the other method will consider water to cement ratio as an input variable and drop the features 'water' and 'cement'. The comparison between models obtained both the ways will be useful to analyse if the quantities of water and cement play a role in determining concrete compressive strength.

2 Dataset

The study uses a dataset on Concrete compressive strength published in UCI Machine Learning Repository [3] by Prof. I-Cheng Yeh [1].

The dataset with 1030 samples made with ordinary Portland cement and cured under normal conditions was derived from 17 different sources and has 9 quantitative attributes, including 1 output variable and 8 input variables. Data is in raw form (not scaled) and there are no missing attribute values. The following table gives a brief description of the attributes:

Name	Measurement unit	Description
Cement	$\frac{kg}{m^3}$	Input Variable
Blast Furnace Slag	$\frac{kg}{m^3}$	Input Variable
Fly Ash	$\frac{kg}{m^3}$	Input Variable
Water	$\frac{kg}{m^3}$	Input Variable
Superplasticizer	$\frac{kg}{m^3}$	Input Variable
Coarse Aggregate	$\frac{kg}{m^3}$	Input Variable
Fine Aggregate	$\frac{kg}{m^3}$	Input Variable
Age	days	Input Variable
Concrete compressive strength	MPa	Output Variable

The attributes have been described in more detail below:

2.1 Target

Concrete compressive strength The specimens were converted into 15-cm cylinders through standard guidelines and tested under the action of compressive loads to determine the compressive strength.

2.2 Features

- **Cement** Cement has adhesive and cohesive properties. It acts as a binding agent in the concrete. It chemically combines with water (hydration) to form a hardened mass.

- **Blast Furnace Slag and Fly Ash** Both improve workability and durability of the concrete mix and reduce permeability. The hardening process takes longer but over time, the strength attained is higher.

- **Water** The water performs hydration with cement and also makes concrete mix fluent and workable.

- **Superplasticizer** It improves concrete workability and reduces the amount of water required in the concrete mix, thus increasing the strength by decreasing water to cement ratio.

- **Coarse Aggregate** Coarse aggregate comprises the strongest part of concrete mix and reduces amount of shrinkage while concrete cures.

- **Fine Aggregate** It works as a space filling agent between coarse aggregates. It adds volume to the concrete (i.e., less air and more strength).

- **Age** Concrete compressive strength increases with age. Usually, it is tested after 28 days but it is less than the long term strength that it can gain with age.

3 Dataset exploration and preprocessing

The original Concrete_Data.xls file was saved with .csv extension before importing it to R. After loading the data, the column headers were changed to make them more accessible. The duplicated rows in the dataset were removed as the trend of the duplicated rows showed that several concrete samples with same composition of ingredients had exact same strength on 3rd day as well as on 7th, 28th, 56th and 91st day. This is highly improbable considering the complexity and quality controls involved in the formation of concrete and it might have occured due to typographical errors. After removing the duplicated rows, the dataset contained 1005 observations. The first few rows of the data after preprocessing are as follows:

	Cement	Slag	Flyash	Water	Superplasticizer	Coarse	Fine	Age	Strength
1	540.0	0.0	0	162.0	2.5	1040.0	676.0	28	79.99
2	540.0	0.0	0	162.0	2.5	1055.0	676.0	28	61.89
3	332	142.5	0	228.0	0.0	932.0	594.0	270	40.27
4	332	142.5	0	228.0	0.0	932.0	594.0	365	41.05
5	198.6	132.4	0	192	0.0	978.4	825.5	360	44.30
6	266.0	114.0	0	228	0.0	932.0	670.0	90	47.03

The following table presents summary statistics of the dataset used in this study (the attribute water to cement ratio has been included as well):

	min	**median**	**mean**	**max**	**std**
Cement	102.0	265.0	278.6	540.0	104.3443
Blast Furnace Slag	0.00	20.00	72.04	359.40	86.17081
Fly Ash	0.00	0.00	55.54	200.10	64.20797
Water	121.8	185.7	182.1	247.0	21.33933
Superplasticizer	0.000	6.100	6.033	32.200	5.919967
Coarse Aggregate	801.0	968.0	974.4	1145.0	77.57967
Fine Aggregate	594.0	780.0	772.7	992.6	80.34043
Age	1.00	28.00	45.86	365.00	63.73469
Water to Cement ratio	0.2669	0.6753	0.7483	1.8820	0.3140055
Concrete Compressive Strength	2.33	34.45	35.82	82.60	16.70574

From the scatterplot matrix of Concrete Compressive Strength dataset given by Figure 1, there is no high correlation between any two attributes. The same can be verified from the correlation matrix. Figure 2 gives the heatmap of correlation between attributes. Water and Superplasticizer have the highest correlation of around 0.65 (which is negative) which is not a very strong correlation.

Figure 3 gives the scatterplot of concrete compressive strength versus water to cement ratio and depicts the inverse proportionality between strength and w/c as proposed by Abrams' water-cement ratio law. On the other hand, it also suggests that for same w/c, we can have different concrete compressive strength thus condemning the claim that it just depends on w/c.

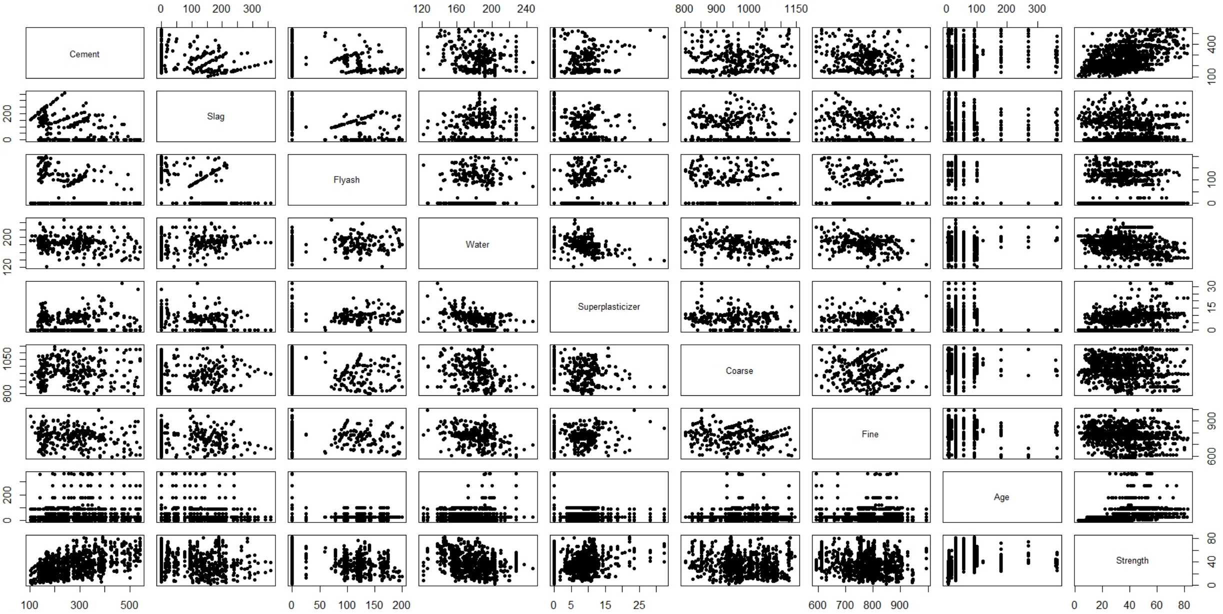

Figure 1: Scatterplot matrix of Concrete compressive strength data

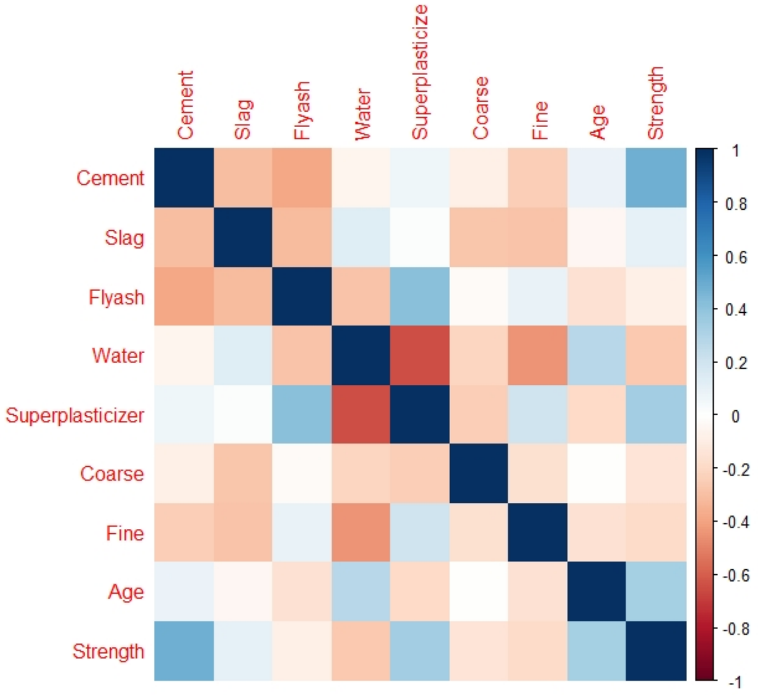

Figure 2: Heatmap of correlation matrix of Concrete compressive strength data

Before using various regression techniques to build models on the dataset, the dataset was divided into training and test dataset and the split used was 75% and 25% respectively. Further, to analyse the models with water to cement ratio, features 'water' and 'cement' columns were removed in training and test datasets and another feature 'wcratio' (water to cement ratio) was added. For convenience, we will be denoting the two kinds of models as models fit to dataset 1 and dataset 2 respectively.

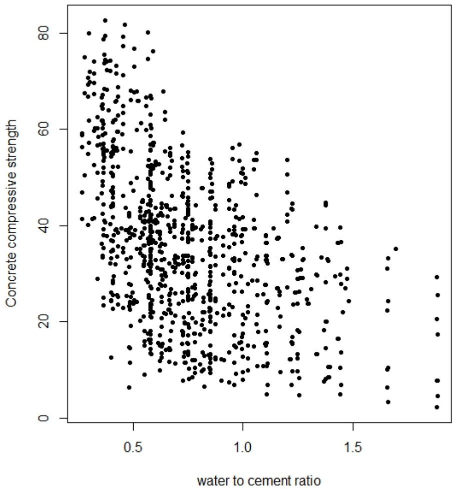

Figure 3: Scatterplot of Concrete compressive strength vs water to cement ratio

4 Regression techniques

Smoothing methods like Smoothing spline and Kernel smoothing are affected by 'curse of dimensionality'. One of the issue is the data requirement, which increases exponentially with increase in the number of predictors. The dataset on Concrete Compressive Strength has 8 predictors but not a very large number of observations. Thus, it is not recommended to use smoothing methods on this dataset. But there are better methods which work well in multiple dimensions. Some of them have been used in this study.

4.1 Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) is a non-parametric regression algorithm, which incorporates non-linearities and interactions between variables. It gives a piecewise linear model and is an improvement over splines as well as trees simultaneously. There are two important tuning parameters associated with MARS model: maximum degree of interaction (degree) and the maximum number of terms to be kept in the final model (nprune).

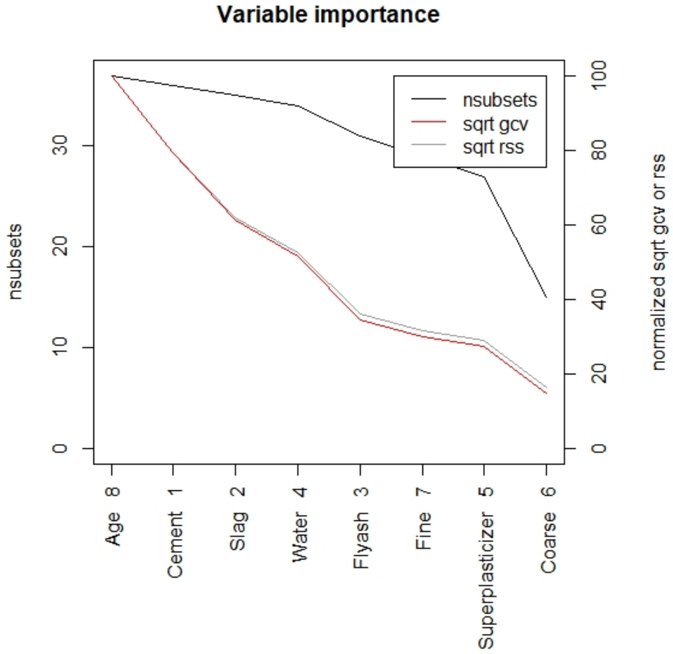

Figure 4: Variable importance plot for MARS model fit to training dataset 1

To fit MARS model to the training datasets, the parameters were tuned using **train()** function from **caret** package. A CV grid search was performed to identify the optimal hyperpameter mix. 5-fold CV was stored as the resampling method in trainControl() to train the tuning parameters and quantitative measure RMSE was used to select the optimal model using the smallest value. Thus, cross validated RMSE was used as the estimate of prediction error for tuning the parameters. The model was pruned using backward elimination and default value of penalty was used since penalty

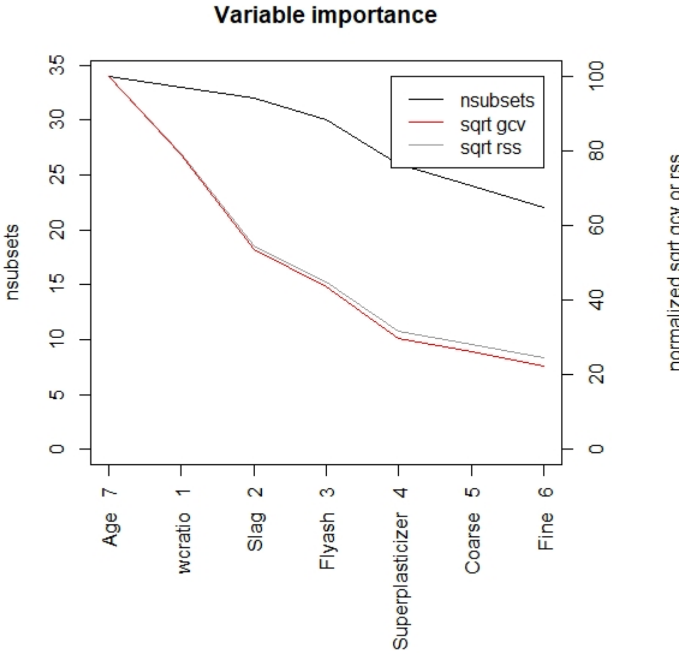

Figure 5: Variable importance plot for MARS model fit to training dataset 2 (with features water and cement replaced by water-cement ratio)

tuning is not generally recommended.

The default value of nk (i.e., the maximum number of terms created by the forward pass) is 21 and that acts as the stopping criterion for the forward pass. The argument nk was set to a higher value so as to prevent it from stopping early.

Figure 4 gives variable importance based on three criteria: nsubsets, square root of GCV and sqrt of RSS. All the three criteria give the same trend for variable importance of features. Age is the most important variable. Cement is the most important ingredient in predicting the strength of HPC followed by slag and water and coarse aggregate is the least important variable.

Figure 5 gives variable importance plot when water to cement ratio (w/c) is used instead of using water and cement independently. w/c is the most important variable. Slag is still one of the important variables and coarse aggregate is still one of the less important variables.

4.2 Random Forests

Random forests are a modification of bootstrap aggregation and create an ensemble of 'de-correlated' trees. This de-correlation of trees is achieved by injecting some randomness in the tree growing process. Each time a split is performed, the search for splitting variable is limited to a random subsample of size m of the p explanatory variables. Thus, number of variables to randomly sample as candidates at each split (mtry) is an important tuning parameter in random forests. It is also suggested to tune the minimum size of terminal nodes (nodesize).

To fit random forest models to the training datasets, a full grid search was performed using **ranger** package to tune these parameters. Out-of-bag (OOB) error was used as the estimate of prediction error to select the optimal model using smallest value.

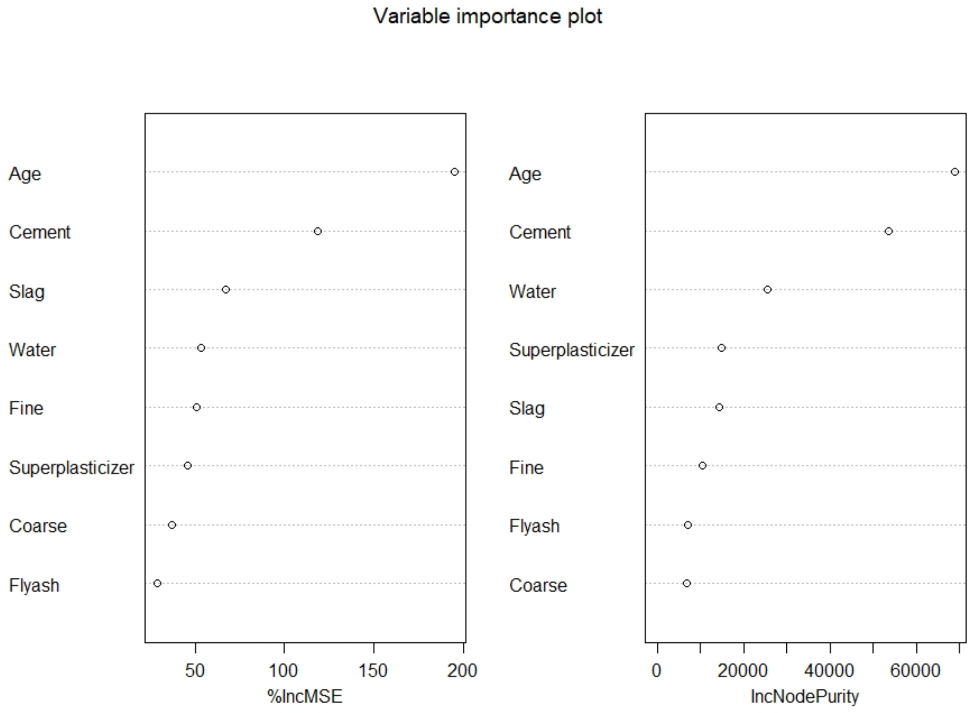

Figure 6: Variable importance plot for Random Forests model fit to training dataset 1

Figure 6 and 7 give the variable importance plots for random forest models fit to the original training dataset and training dataset with water and cement variables replaced by water to cement ratio respectively. The variable importance is based on two criteria: Mean decrease in accuracy (%IncMSE) and mean decrease in RSS or Increase in Node Purity (IncNodePurity). For model fit to dataset 1, age is the most important variable, cement is the most important out of all the ingredients in predicting strength and coarse aggregate and flyash are least important ingredients based upon both the criteria. For model fit to dataset 2, age and wcratio are most important followed by slag and coarse aggregate and flyash are again the least important.

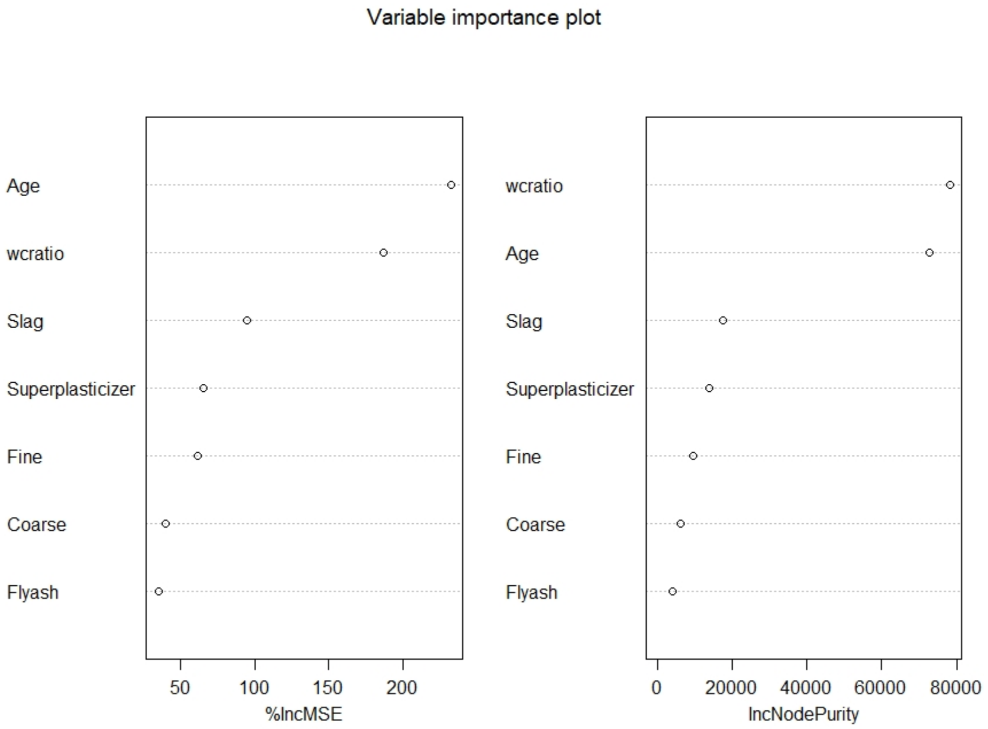

Figure 7: Relative variable importance for Random Forests model fit to training dataset 1

4.3 Boosting methods

While random forests build an ensemble of independent and deep trees, gradient boosting models build an ensemble of shallow and weak trees sequentially with each successive tree learning and improving on the whole previous ensemble. When the outputs of these many“weak” trees is combined, they produce a hard to beat powerful “committee”.

4.3.1 Stochastic Gradient Boosting

With Stochastic gradient boosting, at each iteration, a fraction of the training set observations is randomly selected and the next tree is grown using that subset of training observations. This is the only point which differentiates it from the general gradient boosting algorithm.

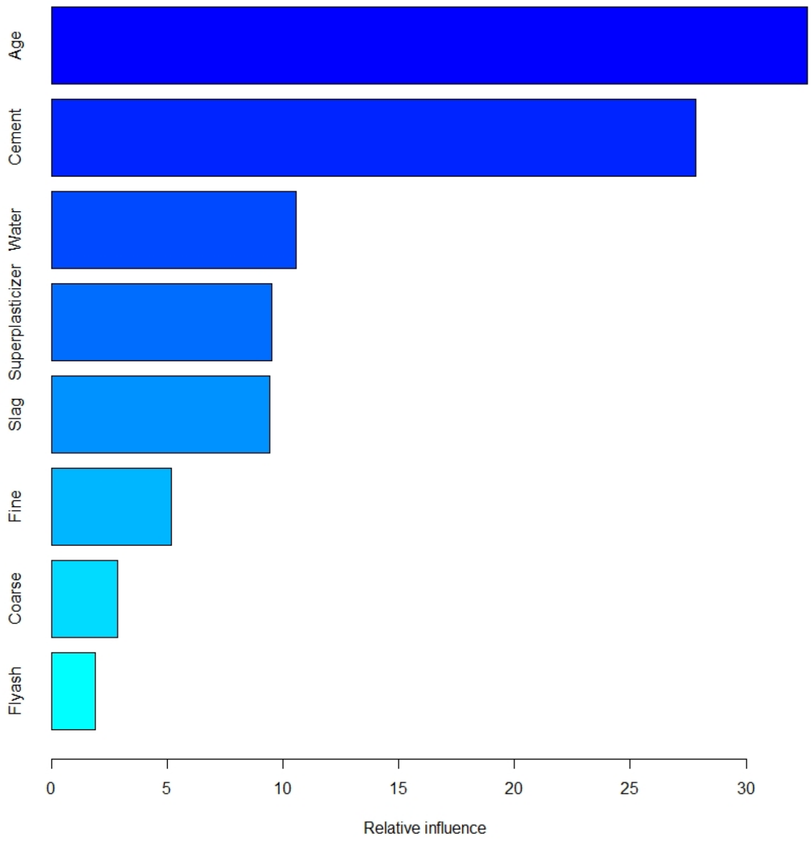

Figure 8: Relative variable importance for Stochastic Gradient model fit to training dataset 1

There are four important tuning parameters associated with Stochastic Gradient Boosting:

- Total number of trees to fit (n.trees)
- Maximum depth of each tree (i.e., the highest level of variable interactions allowed) (interaction.depth)

- Learning rate (shrinkage)
- The fraction of the training set observations randomly selected without replacement to propose the next tree (bag.fraction)

A manual grid search was performed to tune interaction.depth, shrinkage and bag.fraction. An experimental grid of hyperparameter combinations was constructed and 5-fold CV was used as the resampling method. Cross-validation error was used as the estimate of prediction error to select the optimal model using smallest value. The number of trees were set to 2500 and internal CV was allowed to choose the optimal number of trees for given set of tuning parameters. **gbm()** was used to fit the models.

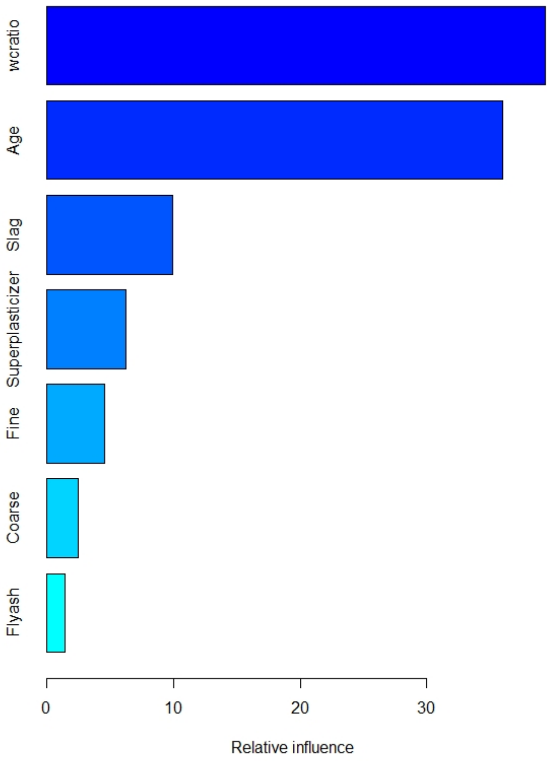

Figure 9: Variable importance for Stochastic Gradient model fit to training dataset 2

Figure 8 and 9 give the relative variable importance plots for stochastic gradient boosting models fit to the original training dataset and training dataset with water and cement variables replaced by water to cement ratio respectively. In Figure 8, age is the most important variable and cement is the most important ingredient in predicting the strength of concrete followed by water. Figure 9 shows that w/c is most important. In both the cases, flyash is the least important variable.

4.3.2 eXtreme gradient boosting

eXtreme gradient boosting is an implementation of the gradient boosting but uses a more regularized model formalization to control over-fitting, which gives it better performance.

The parameters tuned here are same as those in stochastic gradient boosting but are denoted by nrounds, max_depth, eta and subsample respectively. The same procedure was followed as in stochastic gradient boosting. The number of trees were set to 1000. **xgb.cv()** was used to train the models during grid search and **xgboost()** was used to fit the final model with optimal values of parameters.

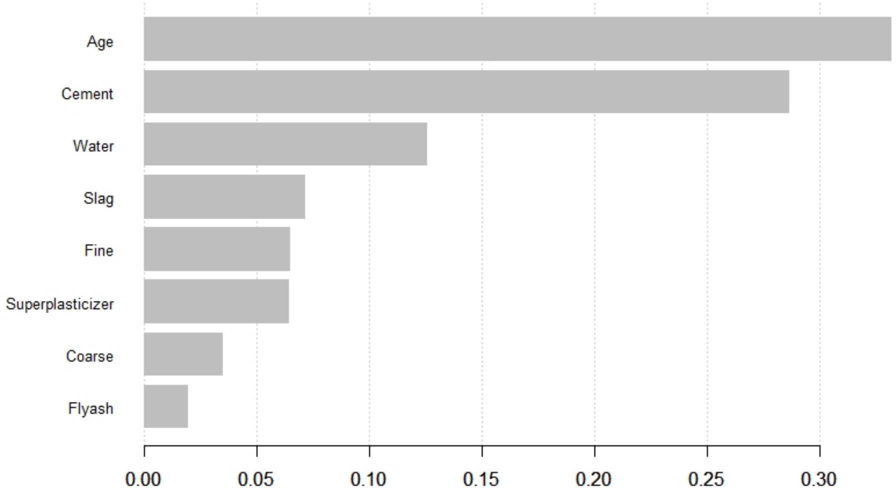

Figure 10: Variable importance plot for eXtreme gradient boosting model fit to training dataset 1, x-axis here gives the relative variable importance measure.

Figure 10 and 11 give the relative variable importance plots for eXtreme gradient boosting models fit to the original training dataset and training

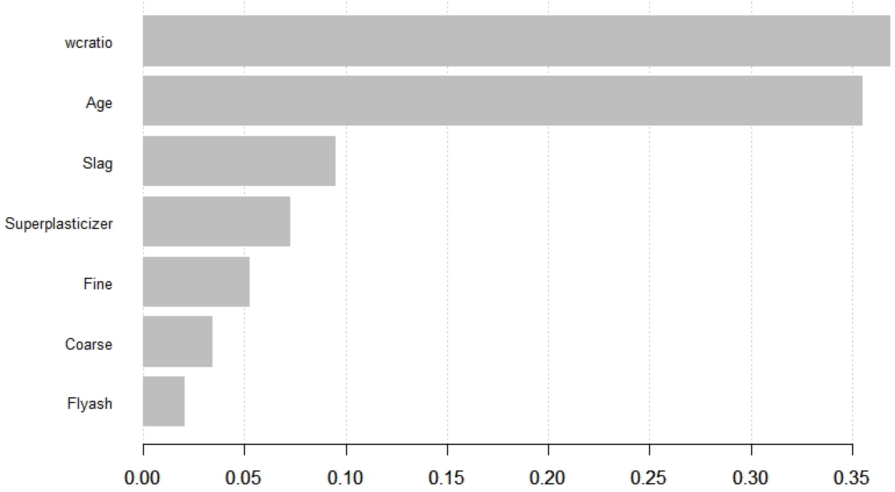

Figure 11: Variable importance plot for eXtreme gradient boosting model fit to training dataset 2, x-axis here gives the relative variable importance measure.

dataset with water and cement variables replaced by water to cement ratio respectively. Variable importance here follows almost the same trend as in the respective cases for stochastic gradient boosting.

5 Previous related works

Yeh [1](1998) modeled HPC strength as a function of cement, blast furnace slag, fly ash, fine aggregate, coarse aggregate, age of testing, superplasticizer and water by using artificial neural networks and obtained promising results ($R^2 = 0.914$ (average)). Chou et. al [2](2011) modeled HPC strength on the same dataset as used in this study using various regression techniques, namely atificial neural network, multiple regression, support vector machine, multiple additive regression trees (MART) and bagging regression trees and the best predictive performance was exhibited by MART ($R^2 = 0.9108$). Another approach used by Yeah and Lien [1](2009) was a genetic operation tree (GOT), which combines a genetic algorithm and an operation tree to produce self-organized formulas automatically to predict the compressive strength of high performance concrete but comparisons showed that GOT ($R^2 = 0.8669$)

was not as accurate as neural networks ($R^2 = 0.9338$).

6 Results and discussion

Mean squared prediction error (MSPE) and coefficient of determination (R^2) were used as the performance measures to evaluate the regression models.

The following tables give the performance of the regression models fit to the training dataset 1 (with no features added or dropped) and training dataset 2 (with features 'Water' and 'Cement' dropped and water to cement ratio added as a feature):

Performance measures for dataset 1		
Regression method	R^2	**MSPE**
MARS	0.8795182	30.78711
Random forests	0.9046908	24.35467
Stochastic gradient boosting	0.9287536	18.20584
eXtreme gradient boosting	0.9313569	17.5406

Performance measures for dataset 2		
Regression method	R^2	**MSPE**
MARS	0.8972387	26.25894
Random forests	0.9033362	24.70083
Stochastic gradient boosting	0.927976	18.40454
eXtreme gradient boosting	0.9281539	18.35906

From the values of MSPE and R^2 in the tables above, it can be seen that the best prediction results in both the cases were obtained by eXtreme gradient boosting followed by stochastic gradient boosting and MARS showcased worst prediction capabilities out of the four models. Figure 12 gives plot of actual vs predicted HPC strength for eXtreme gradient boosting model (which gives the best predictive performance).

Time of computation was more in boosting models (around 19 mins for eXtreme gradient boosting and around 6 mins for stochastic gradient boosting while MARS and random forests took around 16 sec and 8 sec respectively for each of the datasets) and random forests performed best in the training time needed to learn the the model but performed bad in accuracy. Even though boosting models performed worse in training time, there was a significant improvement in prediction accuracies as compared to MARS and Random forests. Considering several factors like predictive accuracy, speed,

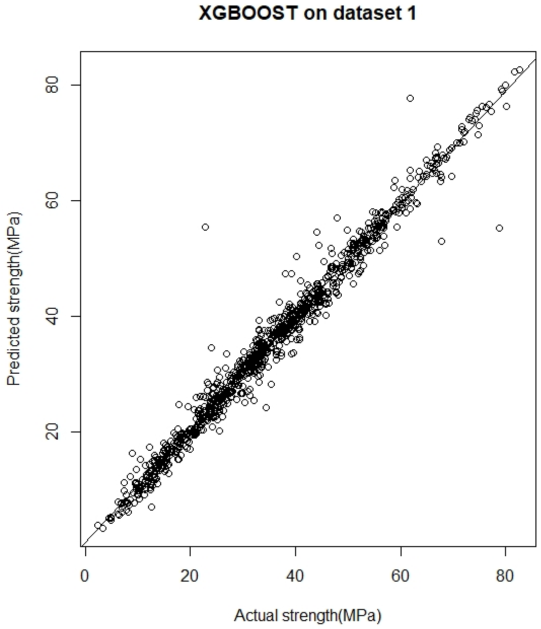

Figure 12: Actual vs predicted HPC strength for eXtreme gradient boosting model fit to dataset with no features added or dropped.

ease of use and interpretability, stochastic gradient boosting is the most reliable predictive model for high performance concrete compressive strength.

When the performance measures in the two tables are compared, it is observed that except MARS, each model fit to training dataset 1 (when no features are added or dropped) outperforms the respective model fit to training dataset 2 (when features 'Water' and 'Cement' are dropped and water to cement ratio is added). In fact, the best model (w.r.t. prediction accuracy) is given by eXtreme gradient boosting fit to dataset 1. Though 'water-cement' ratio plays an important role in concrete technology, the results suggest that we get the best model when 'Water' and 'Cement' are considered as two independent variables instead of being replaced by a single input variable given by water to cement ratio. Thus, as opposed to Abrams' law, only the ratio of water to cement does not matter but their quantities play a role in predicting the strength of high performance concrete as well.

Variable importance plots suggest that age is the most important variable in predicting the compressive strength of HPC followed by cement and water and coarse aggregates and flyash are the least important variables.

7 Conclusions

This study uses regression techniques namely, multiple additive regression splines (MARS), random forests, stochastic gradient boosting and eXtreme gradient boosting to build models to predict compressive strength of high performance concrete. Stochastic gradient boosting is found to work well for this dataset considering several factors like predictive accuracy, speed, ease of use and interpretability. Also, comparisons to primary previous works imply that eXtreme gradient boosting achieves good predictive results ($R^2 = 0.9313569$), which approximate those reported for artificial neural networks ($R^2 = 0.9338$) and based upon variable importance plot, age, cement and water play an important role in predicting the concrete compressive strength of high performance concrete while coarse aggregate and flyash are less important variables.

Based upon comparisons of prediction measures (MSPE and R^2) of models fit to training dataset 1 (when no features were added or dropped) and the models fit to training dataset 2 (when features 'Water' and 'Cement' were dropped and water to cement ratio was added), the study establishes that only the ratio of water to cement does not matter but their quantities play a role in predicting the strength of high performance concrete as well (as opposed to implication of Abram's law).

8 Drawbacks and future work

The dataset does not give any information about the quality controls used during production of concrete like curing conditions, quality of ingredients etc. Superplasticizers were from different manufacturers and had different chemical compositions and there are no details available about them [1]. These conditions and quality of ingredients play an important role in determining the compressive strength of HPC. There is a need of more elaborate dataset to get models with better predictive accuracy. Also, I wish to apply artificial neural networks to the dataset in the future. The training time for my boosting models is higher than expected but I have limited knowledge about the improvements I should make without compromising on the prediction accuracy. I would like to make them more time efficient.

References

[1] Yeh, I.-C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12), 1797-1808.

[2] Chou, J.-S., Chiu, C.-K., Farfoura, M. and Altaharwa, I. (2011). Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data Mining Techniques, *Journal of Computing in Civil Engineering*, 25(3), 242-253.

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[4] Yeh, I.-C and Lien L.C.(2009).Knowledge discovery of concrete material using Genetic Operation Trees . *Expert Systems with Applications*, 36(3), 5807-5812.