A Statistical Model for Predicting the Compressive Strength of Concrete

Syed Abdul Hadi

Abstract

In this paper we will attempt to explain the use of statistical models for predicting the compressive strength of concrete. Regression techniques with interactions between variables and ensemble random forests were tested. Random Forests gave the best results having explained 92.3% of the variance.

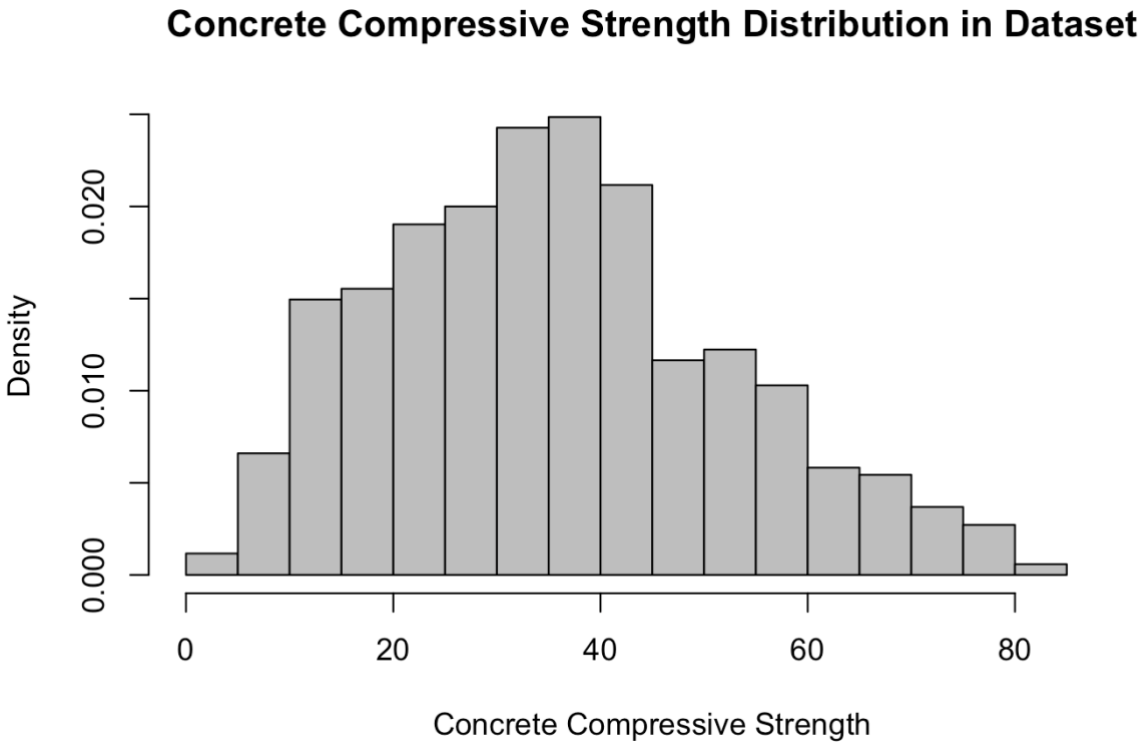
Concrete is characterized by 3 different types of strengths- Tensile, Flexural and Compressive. Tensile strength measures the ability of concrete to resist tensile force whereas flexural strength measures the ability of concrete to withstand bending. Finally, compressive strength measures the ability of concrete to withstand compressive force. Each unique application of concrete defines a minimum required threshold for each type of strength of concrete. Civil engineers and concrete manufacturers are often interested in accurately predicting the strength of concrete before using it in a job. Being able to accurately predict this is critical for the longevity of the final product and for the overall cost effectiveness of the project.

Studies have shown that the compressive strength of concrete is a highly non-linear function of the age and ingredients of the concrete. I will be examining the compressive strength of concrete and building a model to predict the compressive strength of concrete using its age and ingredients as regressors. In this paper I will attempt to answer the following questions 1) To what extent can the compressive strength of concrete be predicted using its ingredients as predictors? 2) Are certain ingredients weighted more than others in determining the compressive strength of concrete? 3) Can regression techniques accurately predict the compressive strength of concrete?

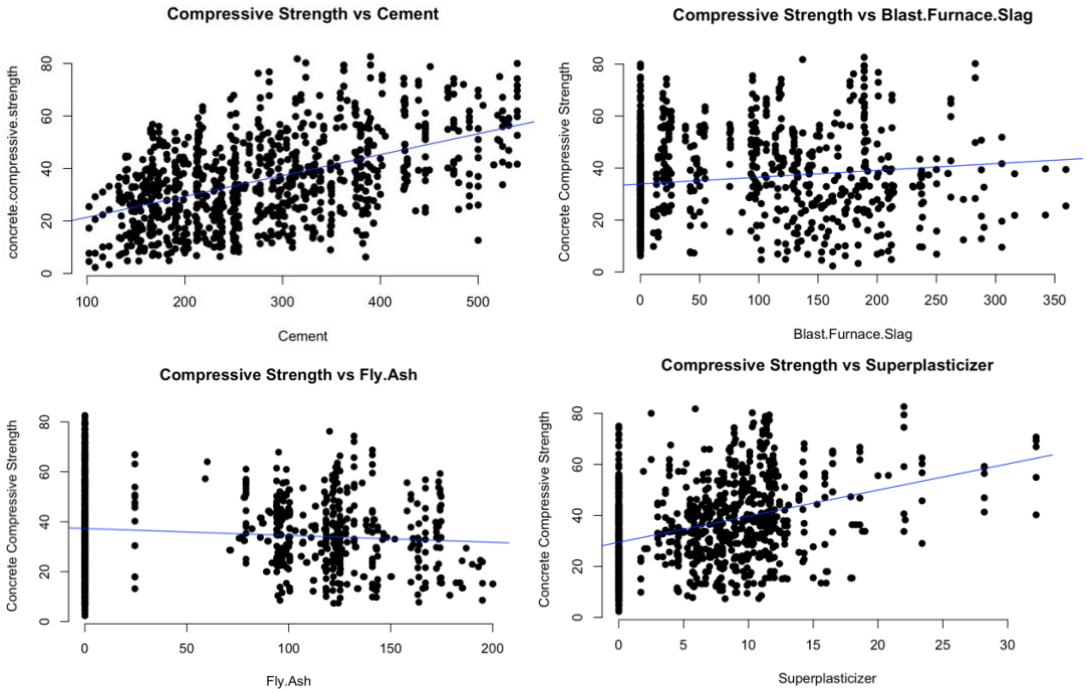
The Data

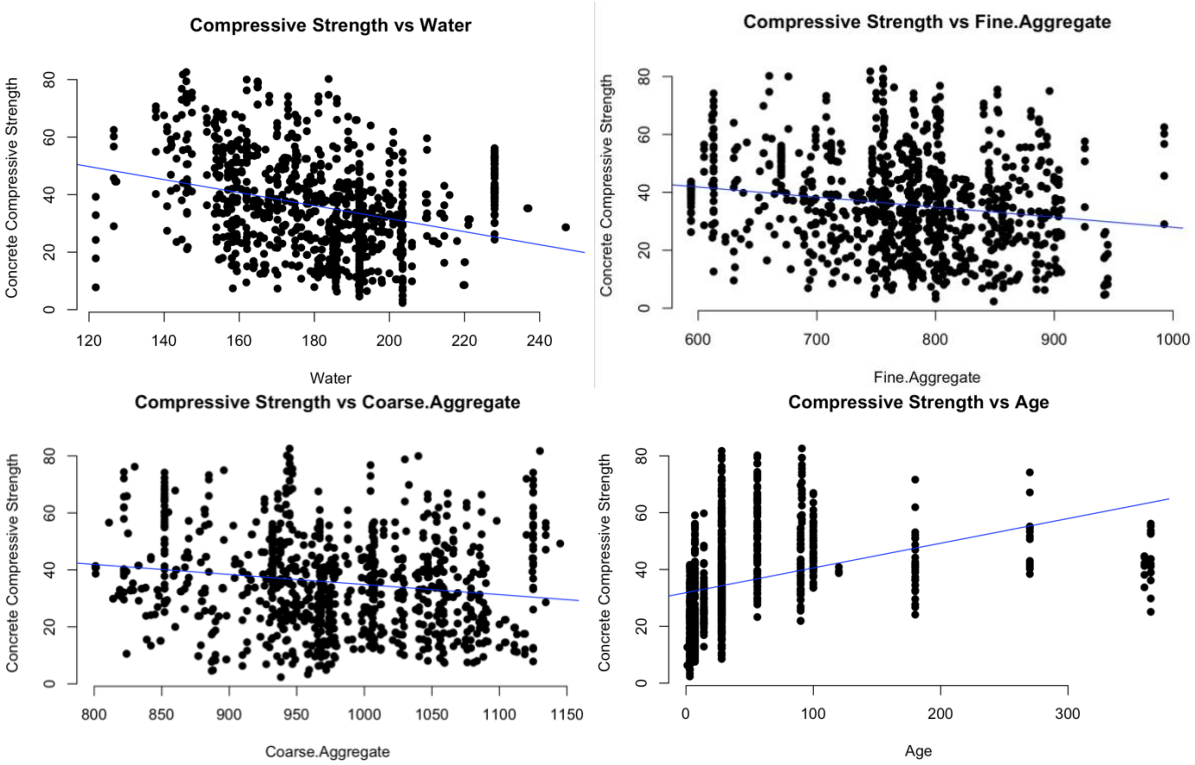
We start off by understanding the dataset. The data has been sourced from the UCI Machine learning repository and was originally created by I-Cheng Yeh[1]. The dataset has 1030 observations and 9 variables (8 explanatory variables and 1 response variable). There are no missing values, and all variables are continuous. The response variable is the compressive strength of concrete. The distribution of the target variable can be seen in the following histogram.

[1] I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998).



We now look at the relationship of the target variable with the individual explanatory variables through scatterplots to get a better sense of how the compressive strength varies with changes in the individual variables.





It can be observed that some variables (such as Age, Fly.Ash and Superplasticizer) affect the compressive strength in a categorical manner as the effects are bundled in the scatterplot. This is an indication that we will have to use complex regression techniques with knots to capture the variance in the response variable.

Simple Linear Regression and Regression with Interaction Terms

We start off by setting a benchmark by using simple linear regression. On RStudio, the lm() function allows for this conveniently. The results of this model are given below:

```
Call:
lm(formula = Concrete.compressive.strength ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-28.654  -6.302   0.703   6.569  34.450

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -23.331214  26.585504  -0.878 0.380372
Cement               0.119804   0.008489  14.113  < 2e-16 ***
Blast.Furnace.Slag   0.103866   0.010136  10.247  < 2e-16 ***
Fly.Ash              0.087934   0.012583   6.988 5.02e-12 ***
Water               -0.149918   0.040177  -3.731 0.000201 ***
Superplasticizer     0.292225   0.093424   3.128 0.001810 **
Coarse.Aggregate     0.018086   0.009392   1.926 0.054425 .
Fine.Aggregate       0.020190   0.010702   1.887 0.059491 .
Age                  0.114222   0.005427  21.046  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 1021 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.6125
F-statistic: 204.3 on 8 and 1021 DF,  p-value: < 2.2e-16
```

While this model is only able to explain 61.2% of the variance in the response variable, these are promising results because we can see that all the variables are statistically significant. At this point, we are in a good position to move forward and test out more advanced techniques. Next we attempt to build a similar model which also explores the interactions between the variables to predict the compressive strength of concrete. The lm() function in R allows for this as well. The results from this model are:

```
Call:
lm(formula = Concrete.compressive.strength ~ (.)^2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-27.0211  -5.5775   0.0581   5.9405  30.8974

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.795e+02  1.417e+02  -1.267 0.205554
Cement                           3.199e-01  1.505e-01   2.125 0.033829 *
Blast.Furnace.Slag               7.063e-02  2.063e-01   0.342 0.732138
```

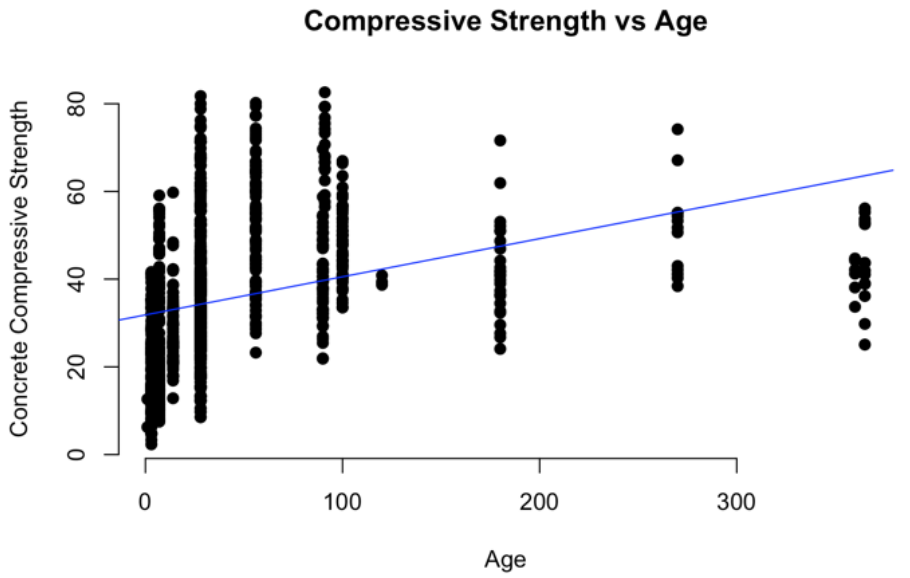
```
Fly.Ash                                -1.862e-01  3.015e-01  -0.618 0.537009
Water                                   1.931e+00  4.641e-01   4.161 3.45e-05 ***
Superplasticizer                        5.957e-03  4.875e+00   0.001 0.999025
Coarse.Aggregate                        6.025e-02  9.894e-02   0.609 0.542710
Fine.Aggregate                         -1.288e-01  9.953e-02  -1.294 0.196040
Age                                    -4.438e-01  4.848e-01  -0.915 0.360186
Cement:Blast.Furnace.Slag               1.312e-04  5.743e-05   2.285 0.022535 *
Cement:Fly.Ash                          2.935e-04  8.091e-05   3.628 0.000300 ***
Cement:Water                           -1.800e-03  3.790e-04  -4.750 2.33e-06 ***
Cement:Superplasticizer                -4.092e-03  1.887e-03  -2.169 0.030347 *
Cement:Coarse.Aggregate                 3.007e-05  6.668e-05   0.451 0.652126
Cement:Fine.Aggregate                   1.023e-04  5.918e-05   1.729 0.084157 .
Cement:Age                              5.477e-04  1.788e-04   3.064 0.002244 **
Blast.Furnace.Slag:Fly.Ash              4.713e-04  1.165e-04   4.044 5.65e-05 ***
Blast.Furnace.Slag:Water               -1.304e-03  5.267e-04  -2.476 0.013435 *
Blast.Furnace.Slag:Superplasticizer    -2.334e-03  2.265e-03  -1.031 0.302845
Blast.Furnace.Slag:Coarse.Aggregate    -4.179e-05  8.869e-05  -0.471 0.637642
Blast.Furnace.Slag:Fine.Aggregate       3.060e-04  7.318e-05   4.181 3.15e-05 ***
Blast.Furnace.Slag:Age                  8.268e-04  1.806e-04   4.577 5.31e-06 ***
Fly.Ash:Water                          -1.992e-03  6.505e-04  -3.063 0.002251 **
Fly.Ash:Superplasticizer               -7.336e-03  2.935e-03  -2.500 0.012588 *
Fly.Ash:Coarse.Aggregate                1.163e-04  1.255e-04   0.927 0.354327
Fly.Ash:Fine.Aggregate                  4.928e-04  1.337e-04   3.685 0.000241 ***
Fly.Ash:Age                             1.715e-03  2.980e-04   5.753 1.17e-08 ***
Water:Superplasticizer                  7.853e-03  6.165e-03   1.274 0.203040
Water:Coarse.Aggregate                 -1.018e-03  2.731e-04  -3.728 0.000204 ***
Water:Fine.Aggregate                   -4.752e-04  2.731e-04  -1.740 0.082137 .
Water:Age                              -4.614e-04  8.274e-04  -0.558 0.577225
Superplasticizer:Coarse.Aggregate       1.411e-03  1.827e-03   0.772 0.440083
Superplasticizer:Fine.Aggregate        -9.478e-04  2.169e-03  -0.437 0.662185
Superplasticizer:Age                    5.887e-03  2.325e-03   2.532 0.011502 *
Coarse.Aggregate:Fine.Aggregate         1.436e-04  6.959e-05   2.064 0.039262 *
Coarse.Aggregate:Age                    2.316e-05  1.495e-04   0.155 0.876931
Fine.Aggregate:Age                      5.076e-04  2.072e-04   2.449 0.014483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.388 on 993 degrees of freedom
Multiple R-squared:  0.7567,    Adjusted R-squared:  0.7479
F-statistic: 85.81 on 36 and 993 DF,  p-value: < 2.2e-16
```

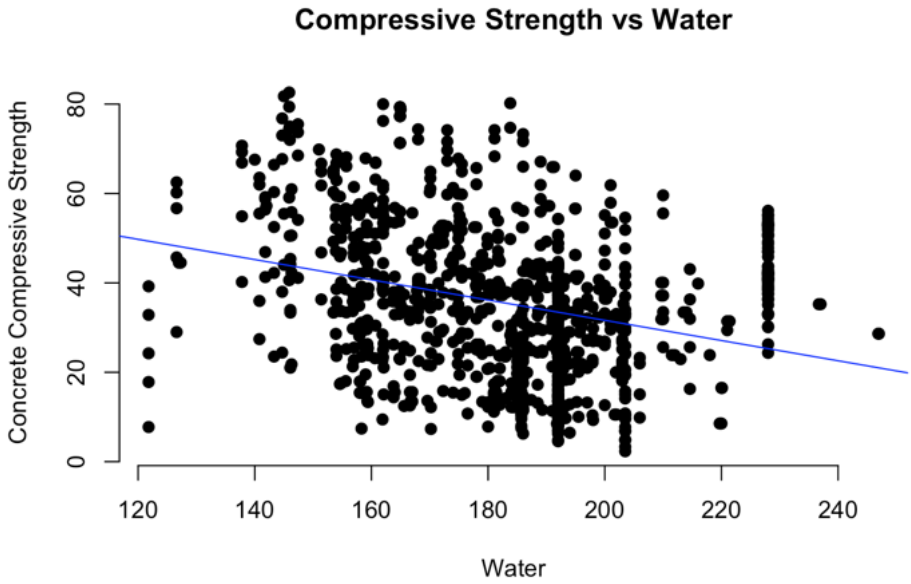
It can be observed that some of the interaction terms are more significant in explaining the variance in the response variable than the main variables and this model can explain 74.8% of the variance in the compressive strength of concrete. Although this is an improvement from the previous model, it does not meet industry standards for a model that can accurately predict the compressive strength of concrete using its age and ingredients. Moreover, it is

our understanding that the compressive strength of concrete is a non-linear function of its age and ingredients therefore the reason we included this model in this paper is to demonstrate the need for a more complex model to do the job.

Multivariate Adaptive Regression Spline

Next, we make use of a Multivariate Adaptive Regression Spline (MARS) model to predict the compressive strength of concrete. The motivation is to model the breaks and jumps in the response variable that we can observe in the scatterplot. Below is one example of the non-linearity and jumps in the response variable as the predictor changes.

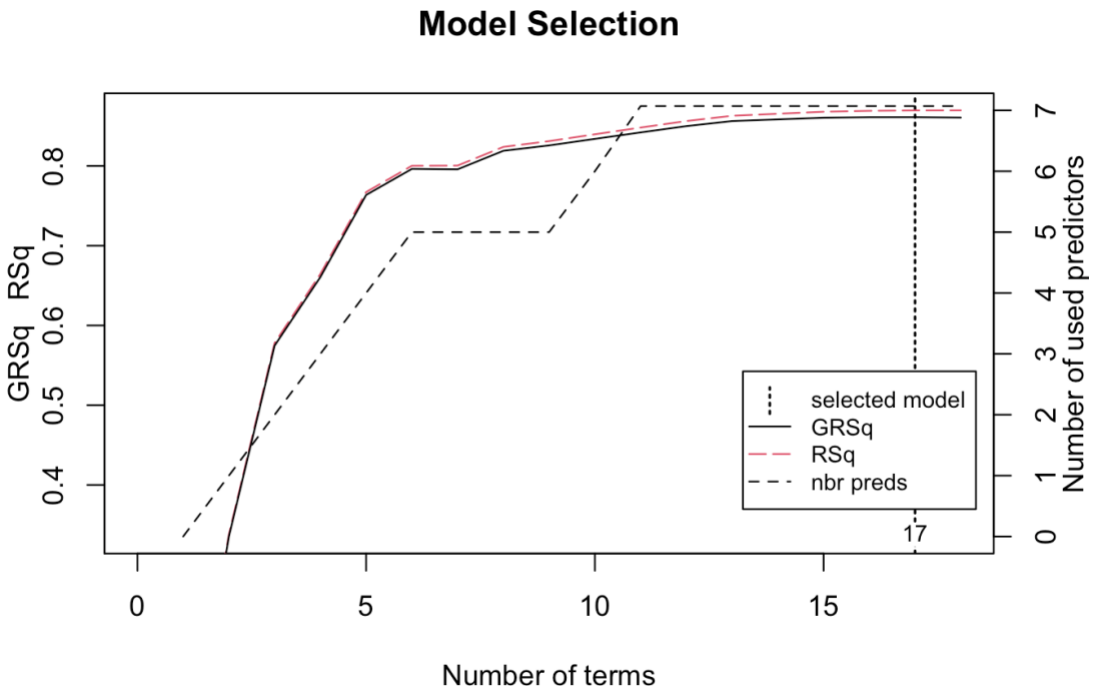




The ‘earth’ library in R allows to conveniently build a MARS model. Below are the results from the model trained over the entire dataset.

```
Call: earth(formula=Concrete.compressive.strength~., data=df)

                            coefficients
(Intercept)                   440.98287
h(525-Cement)                  -0.12020
h(Cement-525)                   0.62283
h(Blast.Furnace.Slag-24)        0.49907
h(53.8-Blast.Furnace.Slag)      0.12153
h(Blast.Furnace.Slag-53.8)     -0.43294
h(174.2-Fly.Ash)               -0.06158
h(Fly.Ash-174.2)               -0.42698
h(Water-127)                   -3.73970
h(218-Water)                   -3.47188
h(Water-218)                    4.31745
h(11.2-Superplasticizer)       -0.15633
h(Superplasticizer-11.2)       -0.84372
h(655-Fine.Aggregate)          -0.15047
h(Age-14)                      -0.90232
h(56-Age)                      -1.26478
h(Age-56)                       0.91171

Selected 17 of 18 terms, and 7 of 8 predictors
Termination condition: Reached nk 21
Importance: Age, Cement, Water, Blast.Furnace.Slag, Fly.Ash, ...
Number of terms at each degree of interaction: 1 16 (additive model)
GCV 38.81178    RSS 37455.59    GRSq 0.8610655    RSq 0.8695723
```

In this model, 7 out of the 8 predictors are used and the model makes use of 17 out of the 18 terms (with knots) to give the final output. MARS evaluates various models by changing the number of predictors and it can be observed that the model with 17 terms gives the best results without redundant complexity.



Age is considered as the most important variable, followed by Cement, Water, Blast Furnace Slag and Fly Ash. The model is able to explain 86.9% of the variance in the response variable which is a significant jump from the benchmark.

Moreover, we will also use this opportunity to discuss a MARS model which incorporates the interactions between the predictors. The same library is used to build this model and the results are given below. In the results below, it can be observed that the model includes the following interaction terms:

1. Cement and Water
2. Water and Age
3. Fine Aggregate and Age

A total of 19 terms are included in this model and 7 out of 8 predictors are included. One component (Coarse Aggregate) was excluded from both the MARS models we tested and this is an indication of its lack of ability to explain the variance in the compressive strength of concrete.

```
Call: earth(formula=Concrete.compressive.strength~., data=df, degree=2)

                                    coefficients
(Intercept)                           105.302576
h(525-Cement)                          -0.094943
h(Cement-525)                           0.639019
h(53.8-Blast.Furnace.Slag)             -0.138654
h(Blast.Furnace.Slag-53.8)              0.062252
h(174.2-Fly.Ash)                       -0.044284
h(Fly.Ash-174.2)                       -0.454233
h(218-Water)                            0.268492
h(Water-218)                            0.304660
h(11.2-Superplasticizer)               -0.313417
h(Superplasticizer-11.2)               -0.641153
h(Age-14)                              -0.883043
h(56-Age)                              -1.164658
h(Age-56)                               0.965055
h(213.7-Cement) * h(218-Water)         -0.002254
h(218-Water) * h(Age-28)                0.000827
h(218-Water) * h(28-Age)               -0.003040
h(692.6-Fine.Aggregate) * h(Age-56)    -0.001116
h(Fine.Aggregate-692.6) * h(Age-56)    -0.000668

Selected 19 of 20 terms, and 7 of 8 predictors
Termination condition: Reached nk 21
Importance: Age, Cement, Water, Blast.Furnace.Slag, Fly.Ash, ...
Number of terms at each degree of interaction: 1 13 5
GCV 38.35217    RSS 36053.13    GRSq 0.8627107    RSq 0.874456
```

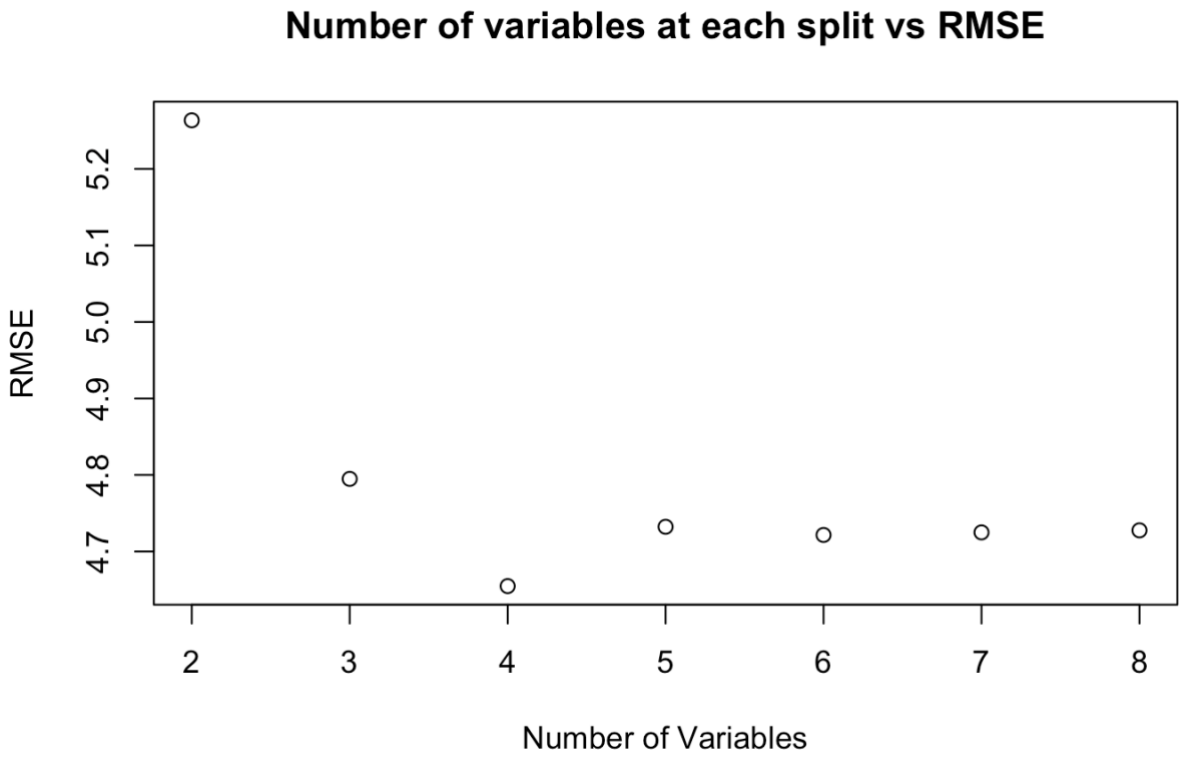
The MARS model with interaction terms included gives an R-squared of 87.4% which is not a significant improvement from the MARS model without interactions (86.9%).

Ensemble Random Forests

Finally, we also test out an ensemble random forest model to predict the compressive strength of concrete. The motivation to test a random forest regression tree is to model the non-linear relationship between the response variable and the predictors.

R offers the 'randomForest' library to build predictive models. By default, the package tests 2 variables at each split. The default model explains 90.86% of the variance in the response variable. While this is a significant improvement from the benchmark and the MARS model, we will not be settling with it. We

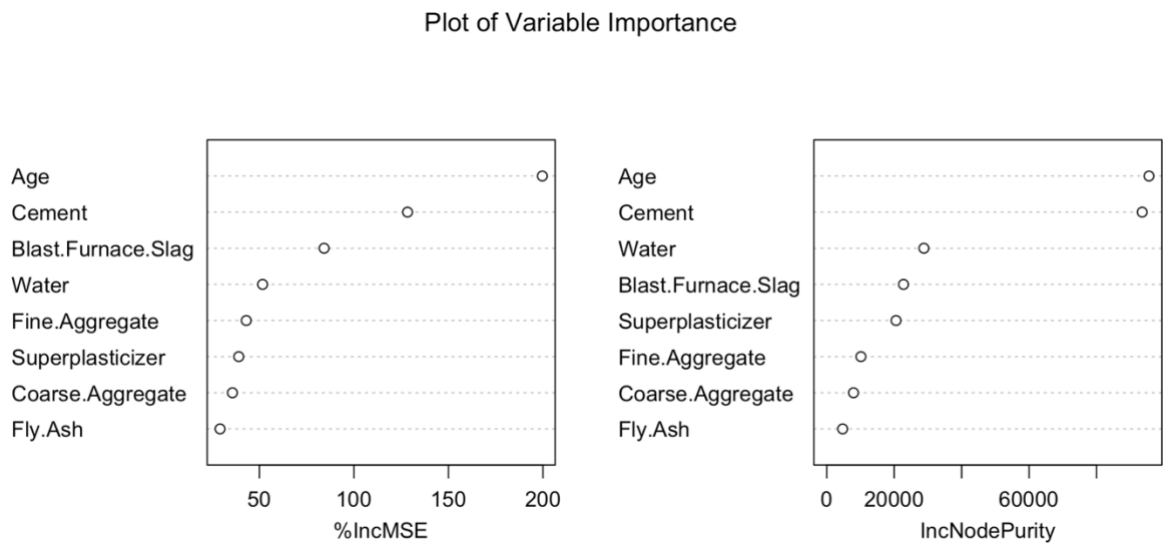
build a number of models by varying the number of variables tested at each split between 2 and 8.



It can be observed that the lowest RMSE is achieved when 4 variables are tested at each split. We will proceed to build a random forest regression tree with the mtry parameter set to 4. The results of this model are given below.

```
Call:
 randomForest(formula = Concrete.compressive.strength ~ ., data = df,      importance = TRUE, mtry = 4)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 21.09512
                    % Var explained: 92.43
```

The model explains 92.43% of the variance in the response variable. We also plot the importance of each predictor in the model. We observed that the same predictors are given the highest ranks in importance as were given in the previous (MARS) model.



Results

Our study with the goal to model the compressive strength of concrete concludes with the results that RandomForests perform better than regression techniques given the non-linearity of the dataset. We can explain 92.43% of the variance in the response variable. Moreover, the variable importance (in decreasing order) is Age, Cement, Water, Blast Furnace Slag, Superplasticizer, Fine Aggregate, Coarse Aggregate and Fly Age. Finally, the highest variance we could explain using a regression model is 86.9% which is significantly lower than what was achieved using an ensemble random forest model.