EXPLORING SCHOOL PERFORMANCE FACTORS

Team United

Lasya Naga Sai Lakshmi Paida and Tarun Kumar Bandaru

REPORT

Introduction:

The “Exploring School Performance” dataset, sourced from the NYC student survey, provides an extensive array of metrics aimed at understanding the factors influencing school performance. Our study involves finding the best metric which is improving the school performance. This dataset is particularly exciting due to its real-world relevance and the opportunity it provides to inform educational policies, improve teaching methods, and enhance community involvement.

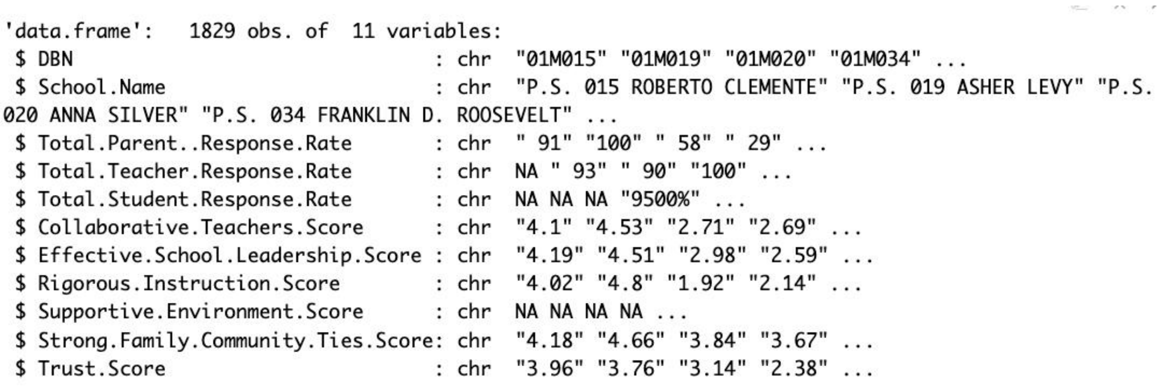
Data Variables:

- DBN: Stands for District Borough Number, a unique identifier for each school in New York City.
- School Name: The name of the school.
- Total Parent Response Rate: The response rate of parents to the school survey. It indicates the percentage of parents who participated in the survey.
- Total Teacher Response Rate: The response rate of teachers to the school survey. It indicates the percentage of teachers who participated in the survey.
- Total Student Response Rate: The response rate of students to the school survey. It indicates the percentage of students who participated in the survey.
- Collaborative Teachers Score: A score representing the level of collaboration among teachers in the school. Higher scores indicate higher levels of collaboration.
- Effective School Leadership Score: A score assessing the effectiveness of school leadership. It reflects the perceived quality of leadership within the school.
- Rigorous Instruction Score: A score evaluating the rigor of instruction in the school. It measures the academic challenge provided to students.
- Supportive Environment Score: A score indicating the level of supportiveness in the school environment. It assesses how well the school supports the needs of its students.

- Community Ties Score: A score reflecting the strength of ties between the school, families, and the community. It measures the level of community engagement and involvement.
- Trust Score: A score assessing the level of trust within the school community. It reflects the trust that students, parents, and teachers have in the school's administration and policies.

Data Preparation:

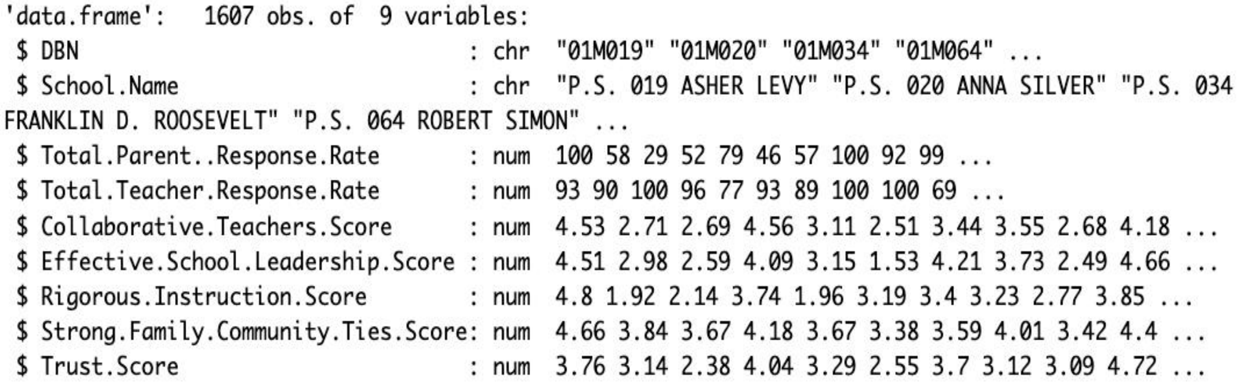
Our Data:



```
'data.frame':	1829 obs. of  11 variables:
 $ DBN                               : chr  "01M015" "01M019" "01M020" "01M034" ...
 $ School.Name                       : chr  "P.S. 015 ROBERTO CLEMENTE" "P.S. 019 ASHER LEVY" "P.S. 
020 ANNA SILVER" "P.S. 034 FRANKLIN D. ROOSEVELT" ...
 $ Total.Parent..Response.Rate       : chr  " 91" "100" " 58" " 29" ...
 $ Total.Teacher.Response.Rate       : chr  NA " 93" " 90" "100" ...
 $ Total.Student.Response.Rate       : chr  NA NA NA "9500%" ...
 $ Collaborative.Teachers.Score      : chr  "4.1" "4.53" "2.71" "2.69" ...
 $ Effective.School.Leadership.Score : chr  "4.19" "4.51" "2.98" "2.59" ...
 $ Rigorous.Instruction.Score        : chr  "4.02" "4.8" "1.92" "2.14" ...
 $ Supportive.Environment.Score      : chr  NA NA NA NA ...
 $ Strong.Family.Community.Ties.Score: chr  "4.18" "4.66" "3.84" "3.67" ...
 $ Trust.Score                       : chr  "3.96" "3.76" "3.14" "2.38" ...
```

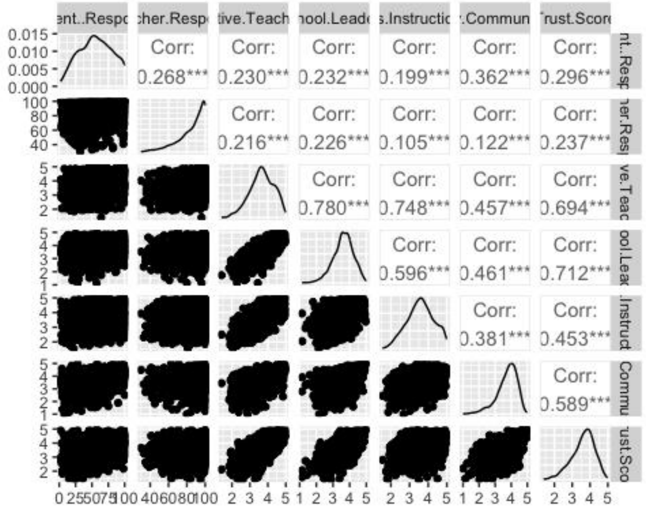
- Converted the data type of every variable to numeric type.
- While trying to find the complete cases, we have noticed that our dataset has two columns (Total.Student.Response.Rate, Supportive.Environment.Score) which have majority of missing values.
- So we have dropped those columns.
- We have also dropped all the other rows which are having missing values.
- Then we are left with 1607 rows.

Overview of the prepared data:



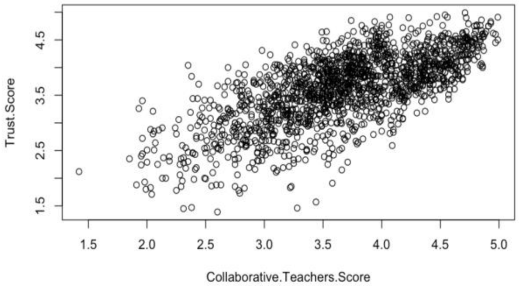
```
'data.frame':	1607 obs. of  9 variables:
 $ DBN                               : chr  "01M019" "01M020" "01M034" "01M064" ...
 $ School.Name                       : chr  "P.S. 019 ASHER LEVY" "P.S. 020 ANNA SILVER" "P.S. 034 
FRANKLIN D. ROOSEVELT" "P.S. 064 ROBERT SIMON" ...
 $ Total.Parent..Response.Rate       : num  100 58 29 52 79 46 57 100 92 99 ...
 $ Total.Teacher.Response.Rate       : num  93 90 100 96 77 93 89 100 100 69 ...
 $ Collaborative.Teachers.Score      : num  4.53 2.71 2.69 4.56 3.11 2.51 3.44 3.55 2.68 4.18 ...
 $ Effective.School.Leadership.Score : num  4.51 2.98 2.59 4.09 3.15 1.53 4.21 3.73 2.49 4.66 ...
 $ Rigorous.Instruction.Score        : num  4.8 1.92 2.14 3.74 1.96 3.19 3.4 3.23 2.77 3.85 ...
 $ Strong.Family.Community.Ties.Score: num  4.66 3.84 3.67 4.18 3.67 3.38 3.59 4.01 3.42 4.4 ...
 $ Trust.Score                       : num  3.76 3.14 2.38 4.04 3.29 2.55 3.7 3.12 3.09 4.72 ...
```

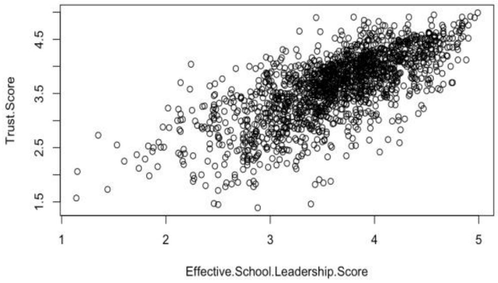
Exploratory Data Analysis (EDA):



From the above, we have also observed that Total Teacher Response Rate is skewed to the right.

We can also find that, Collaborative Teachers Score and Effective School Leadership Score are the most correlated variable with Trust Score (Target variable). Considering correlation, these two are the important predictors.



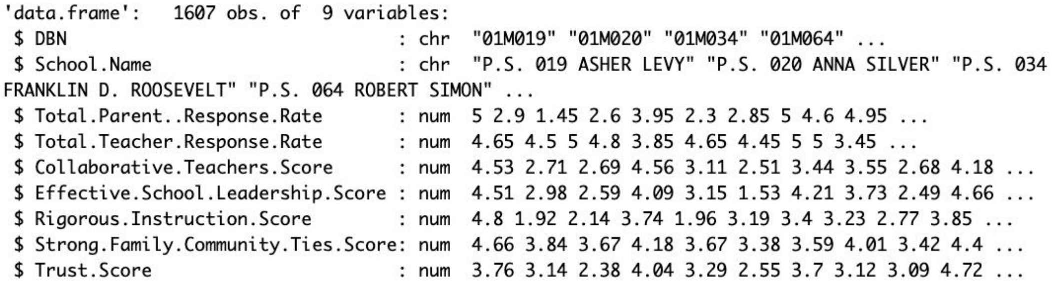


These are the plots of the target variable(Trust Score) with the most important predictors.

Diagnostics:

Total.Parent.Response.Rate and Total.Teacher.Response.Rate are scaled to 100 while other measures are scaled to 5. So we scaled these two variables also to 5.

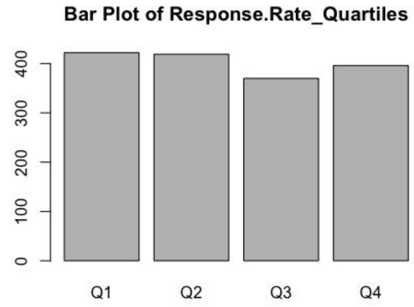
The below data represents the scaled variables.



```
'data.frame':   1607 obs. of  9 variables:
 $ DBN                              : chr  "01M019" "01M020" "01M034" "01M064" ...
 $ School.Name                      : chr  "P.S. 019 ASHER LEVY" "P.S. 020 ANNA SILVER" "P.S. 034
FRANKLIN D. ROOSEVELT" "P.S. 064 ROBERT SIMON" ...
 $ Total.Parent..Response.Rate       : num  5 2.9 1.45 2.6 3.95 2.3 2.85 5 4.6 4.95 ...
 $ Total.Teacher.Response.Rate       : num  4.65 4.5 5 4.8 3.85 4.65 4.45 5 5 3.45 ...
 $ Collaborative.Teachers.Score      : num  4.53 2.71 2.69 4.56 3.11 2.51 3.44 3.55 2.68 4.18 ...
 $ Effective.School.Leadership.Score : num  4.51 2.98 2.59 4.09 3.15 1.53 4.21 3.73 2.49 4.66 ...
 $ Rigorous.Instruction.Score        : num  4.8 1.92 2.14 3.74 1.96 3.19 3.4 3.23 2.77 3.85 ...
 $ Strong.Family.Community.Ties.Score: num  4.66 3.84 3.67 4.18 3.67 3.38 3.59 4.01 3.42 4.4 ...
 $ Trust.Score                      : num  3.76 3.14 2.38 4.04 3.29 2.55 3.7 3.12 3.09 4.72 ...
```

As the Total.Teacher.Response.Rate is skewed to the right, we have tried to transform it with log, sqrt, square and log(1+x), but even after trying these it is not symmetric.

So we have converted this into four quartiles and the new variable Response Rate Quartiles is created which is a categorical variable.



Model Selection:

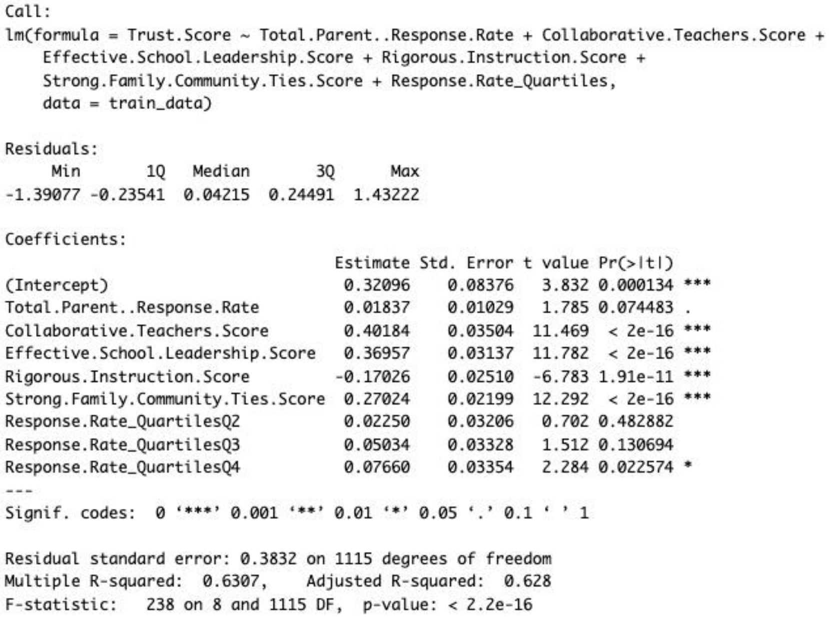
Train Test Split:

The dataset is divided into two parts: 70% as train data and the remaining 30% as test data. We'll build the models on the train data and text its error on the test data.

Full Model:

We hypothesize that the Trust Score can be predicted from all the predictors except DBN and School Name.

Summary:



```
Call:
lm(formula = Trust.Score ~ Total.Parent..Response.Rate + Collaborative.Teachers.Score +
    Effective.School.Leadership.Score + Rigorous.Instruction.Score +
    Strong.Family.Community.Ties.Score + Response.Rate_Quartiles,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39077 -0.23541  0.04215  0.24491  1.43222

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         0.32096    0.08376   3.832 0.000134 ***
Total.Parent..Response.Rate         0.01837    0.01029   1.785 0.074483 .
Collaborative.Teachers.Score        0.40184    0.03504  11.469  < 2e-16 ***
Effective.School.Leadership.Score   0.36957    0.03137  11.782  < 2e-16 ***
Rigorous.Instruction.Score         -0.17026    0.02510  -6.783 1.91e-11 ***
Strong.Family.Community.Ties.Score  0.27024    0.02199  12.292  < 2e-16 ***
Response.Rate_QuartilesQ2           0.02250    0.03206   0.702 0.482882
Response.Rate_QuartilesQ3           0.05034    0.03328   1.512 0.130694
Response.Rate_QuartilesQ4           0.07660    0.03354   2.284 0.022574 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3832 on 1115 degrees of freedom
Multiple R-squared:  0.6307,    Adjusted R-squared:  0.628
F-statistic:   238 on 8 and 1115 DF,  p-value: < 2.2e-16
```

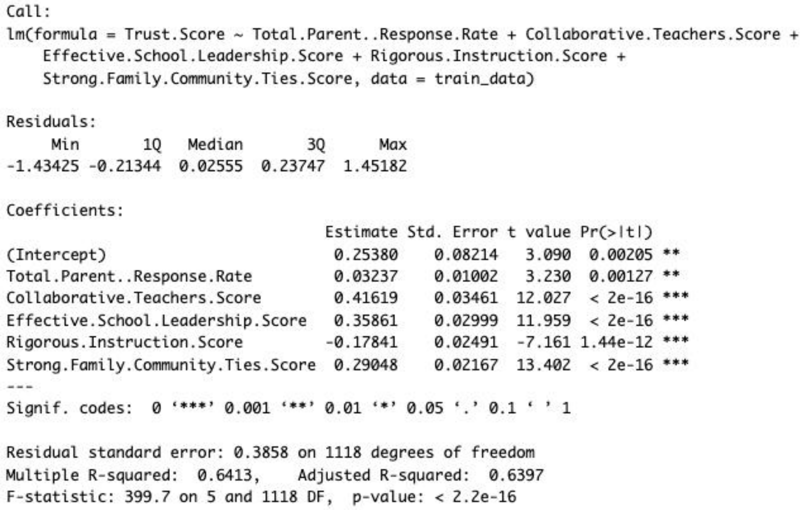
Backwards Selection:

Lets consider backward stepwise selection, to find a model with much more significance.

The model is:

Trust.Score ~ Total.Parent..Response.Rate + Collaborative.Teachers.Score + Effective.School.Leadership.Score + Rigorous.Instruction.Score + Strong.Family.Community.Ties.Score

Summary:



```
Call:
lm(formula = Trust.Score ~ Total.Parent..Response.Rate + Collaborative.Teachers.Score +
    Effective.School.Leadership.Score + Rigorous.Instruction.Score +
    Strong.Family.Community.Ties.Score, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43425 -0.21344  0.02555  0.23747  1.45182

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.25380    0.08214   3.090  0.00205 **
Total.Parent..Response.Rate        0.03237    0.01002   3.230  0.00127 **
Collaborative.Teachers.Score       0.41619    0.03461  12.027  < 2e-16 ***
Effective.School.Leadership.Score  0.35861    0.02999  11.959  < 2e-16 ***
Rigorous.Instruction.Score        -0.17841    0.02491  -7.161 1.44e-12 ***
Strong.Family.Community.Ties.Score  0.29048    0.02167  13.402  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3858 on 1118 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.6397
F-statistic: 399.7 on 5 and 1118 DF,  p-value: < 2.2e-16
```

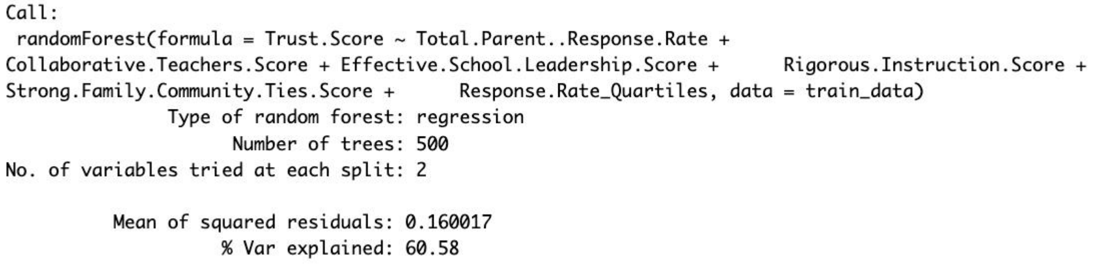
From the above summary, we can say that most of the non-significant predictors are not considered and the model also have a better R^2 value than our full model.

Now, lets further investigate the dataset by modelling it using Random Forest Model.

Random Forest Model:

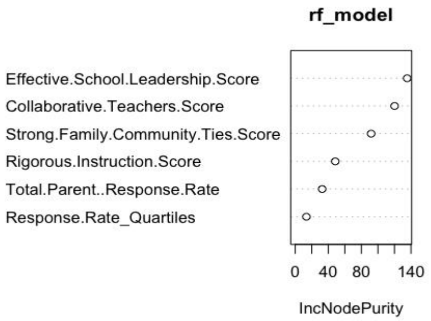
A Random Forest is an ensemble learning method used for both classification and regression tasks. It is a type of ensemble machine learning model that operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

Implemented random forest regression with all the variables and using the tree count 500.

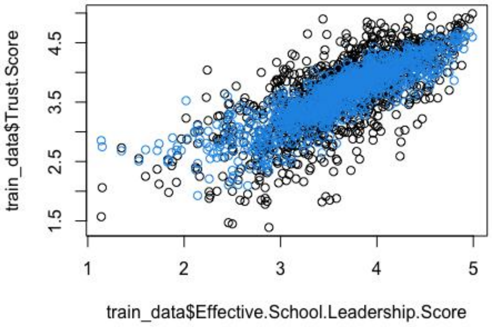


```
Call:
 randomForest(formula = Trust.Score ~ Total.Parent..Response.Rate +
Collaborative.Teachers.Score + Effective.School.Leadership.Score +      Rigorous.Instruction.Score +
Strong.Family.Community.Ties.Score +      Response.Rate_Quartiles, data = train_data)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.160017
                    % Var explained: 60.58
```

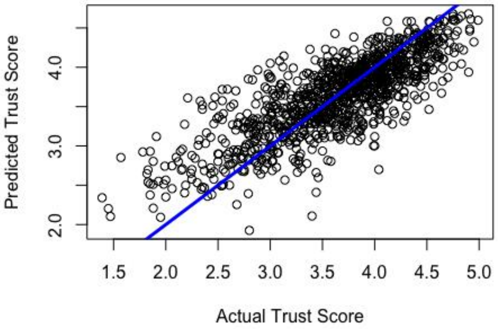
VarImp Plot:



From the above varImp plot, our most important variable is: Effective School Leadership Score.



This figure is the plot of Trust Score with the most important variable: Effective School Leadership Score and it is combined with the predicted Trust Score with Random Forest.



This figure is the plot of the Actual Trust Score vs Predicted Trust Score along with the regression line.

From these plots we can notice that the model worked well and there are not much deviations in the predicted values from the actual values.

Prediction Error (Generalization Error):

We have Calculated the generalization error for backward selection and random forest models.

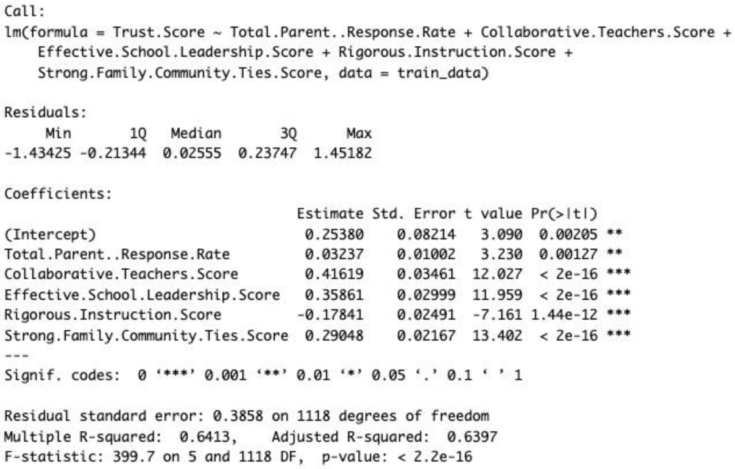
Backward Selection Model	Random Forest Model
0.1395026	0.1500384

Based on this, our backward selection model is a better model due to its less generalization error.

Final Model:

Summary:

Lets look at the summary of our final model:

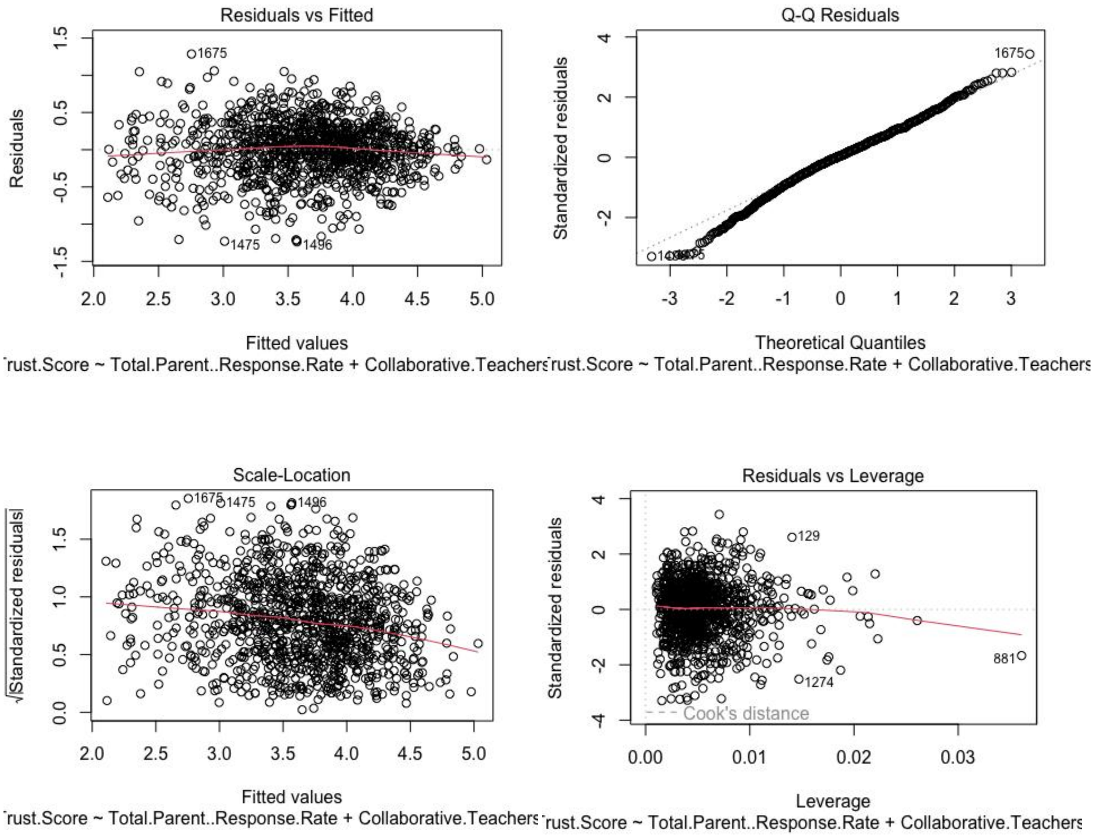


```
Call:
lm(formula = Trust.Score ~ Total.Parent..Response.Rate + Collaborative.Teachers.Score +
    Effective.School.Leadership.Score + Rigorous.Instruction.Score +
    Strong.Family.Community.Ties.Score, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43425 -0.21344  0.02555  0.23747  1.45182

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         0.25380    0.08214   3.090  0.00205 **
Total.Parent..Response.Rate         0.03237    0.01002   3.230  0.00127 **
Collaborative.Teachers.Score        0.41619    0.03461  12.027  < 2e-16 ***
Effective.School.Leadership.Score   0.35861    0.02999  11.959  < 2e-16 ***
Rigorous.Instruction.Score         -0.17841    0.02491  -7.161 1.44e-12 ***
Strong.Family.Community.Ties.Score  0.29048    0.02167  13.402  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3858 on 1118 degrees of freedom
Multiple R-squared:  0.6413,	Adjusted R-squared:  0.6397
F-statistic: 399.7 on 5 and 1118 DF,  p-value: < 2.2e-16
```

From the above summary, we can say that the best model is considered with the most important predictors.

Assessment Plots:

Lets look at the assessment plots of our final model:



- In Residual vs Fitted Plot, the linearity is satisfied. There is no definite pattern and no heteroscedasticity.
- In Q-Q Residual plot, initially the points: Outliers(1475 and 41) are away from the line but gradually the points are on the straight line. But, at the end a few points: Outliers(1675) go away from the line.
- In Scale-Location plot, the points are scattered randomly around the line and here isno funnel shaped structure in the points which indicates homoscedasticity.
- In Residuals-Leverage plot, 41, 131 and 1675 are the residuals with more leverage.

Overall, The model seems to be linear.

Conclusion:

Collaborative Teachers Score is the best factor that influences the School Performance, followed by the Effective School Leadership Score.

This means that, more the level of collaboration among teachers in the school and more the effectiveness of school leadership, the performance of the school is more.

So if the school improves the Student-Teacher collaboration and Leadership, the school will be the best performed school comparatively.