Nomad2018 Predicting Transparent Conductors

Pranjal Vijay

January 17, 2018

Load the libraries

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
```

Load the Data

```
Training_Data<-read.csv("C:/Users/ddddd/Comptitions/Kaggle/Nomad2018 Predicting/train.csv", stri
ngsAsFactors = F)
Test_Data<-read.csv("C:/Users/ddddd/Comptitions/Kaggle/Nomad2018 Predicting/test.csv", stringsAs
Factors = F)
```

Data Exploration

```
names(Training_Data)
```

```
##  [1] "id"                         "spacegroup"
##  [3] "number_of_total_atoms"      "percent_atom_al"
##  [5] "percent_atom_ga"            "percent_atom_in"
##  [7] "lattice_vector_1_ang"       "lattice_vector_2_ang"
##  [9] "lattice_vector_3_ang"       "lattice_angle_alpha_degree"
## [11] "lattice_angle_beta_degree"  "lattice_angle_gamma_degree"
## [13] "formation_energy_ev_natom"  "bandgap_energy_ev"
```

```
str(Training_Data)
```

```
## 'data.frame':    2400 obs. of  14 variables:
##  $ id                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ spacegroup                : int  33 194 227 167 194 227 206 12 206 194 ...
##  $ number_of_total_atoms     : num  80 80 40 30 80 40 80 20 80 80 ...
##  $ percent_atom_al           : num  0.625 0.625 0.812 0.75 0 ...
##  $ percent_atom_ga           : num  0.375 0.375 0.188 0 0.625 ...
##  $ percent_atom_in           : num  0 0 0 0.25 0.375 0 0.875 0.5 0.25 0 ...
##  $ lattice_vector_1_ang      : num  9.95 6.18 9.75 5 6.66 ...
##  $ lattice_vector_2_ang      : num  8.55 6.18 5.66 5 6.66 ...
##  $ lattice_vector_3_ang      : num  9.18 23.63 13.96 13.53 24.58 ...
##  $ lattice_angle_alpha_degree: num  90 90 91 90 90 ...
##  $ lattice_angle_beta_degree : num  90 90 91.1 90 90 ...
##  $ lattice_angle_gamma_degree: num  90 120 30.5 120 120 ...
##  $ formation_energy_ev_natom : num  0.068 0.249 0.1821 0.2172 0.0505 ...
##  $ bandgap_energy_ev         : num  3.44 2.92 2.74 3.35 1.38 ...
```

```
Training_Data$id <- as.numeric(Training_Data$id )
Training_Data$spacegroup <-  as.numeric(Training_Data$spacegroup   )
```

```
str(Training_Data)
```

```
## 'data.frame':    2400 obs. of  14 variables:
##  $ id                        : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ spacegroup                : num  33 194 227 167 194 227 206 12 206 194 ...
##  $ number_of_total_atoms     : num  80 80 40 30 80 40 80 20 80 80 ...
##  $ percent_atom_al           : num  0.625 0.625 0.812 0.75 0 ...
##  $ percent_atom_ga           : num  0.375 0.375 0.188 0 0.625 ...
##  $ percent_atom_in           : num  0 0 0 0.25 0.375 0 0.875 0.5 0.25 0 ...
##  $ lattice_vector_1_ang      : num  9.95 6.18 9.75 5 6.66 ...
##  $ lattice_vector_2_ang      : num  8.55 6.18 5.66 5 6.66 ...
##  $ lattice_vector_3_ang      : num  9.18 23.63 13.96 13.53 24.58 ...
##  $ lattice_angle_alpha_degree: num  90 90 91 90 90 ...
##  $ lattice_angle_beta_degree : num  90 90 91.1 90 90 ...
##  $ lattice_angle_gamma_degree: num  90 120 30.5 120 120 ...
##  $ formation_energy_ev_natom : num  0.068 0.249 0.1821 0.2172 0.0505 ...
##  $ bandgap_energy_ev         : num  3.44 2.92 2.74 3.35 1.38 ...
```

Research question

We are already given a research question by this competition on Kaggle itself.

The prediction of two target properties: the formation energy (which is an indication of the stability of a new material) and the bandgap energy (which is an indication of the potential for transparency over the visible range) to facilitate the discovery of new transparent conductors and allow for advancements in the above-mentioned technologies.

For each id in the test set, we must predict a value for both formation_energy_ev_natom and bandgap_energy_ev.

Solution Starts from here…

Data Preprocessing

```
Training_Data %>%
  filter( id != "NA", spacegroup != "NA", number_of_total_atoms != "NA", percent_atom_al != "NA"
, percent_atom_ga != "NA", percent_atom_in != "NA", lattice_vector_1_ang != "NA", lattice_vector
_2_ang != "NA", lattice_vector_3_ang != "NA", lattice_angle_alpha_degree != "NA", lattice_angle_
beta_degree != "NA", lattice_angle_gamma_degree != "NA", formation_energy_ev_natom != "NA", band
gap_energy_ev != "NA") %>%
  select( id, spacegroup, number_of_total_atoms, percent_atom_al, percent_atom_ga, percent_atom_
in   , lattice_vector_1_ang, lattice_vector_2_ang, lattice_vector_3_ang, lattice_angle_alpha_deg
ree, lattice_angle_beta_degree, lattice_angle_gamma_degree, formation_energy_ev_natom,  bandgap_
energy_ev ) %>%
  str()
```

```
## 'data.frame':    2400 obs. of  14 variables:
##  $ id                        : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ spacegroup                : num  33 194 227 167 194 227 206 12 206 194 ...
##  $ number_of_total_atoms     : num  80 80 40 30 80 40 80 20 80 80 ...
##  $ percent_atom_al           : num  0.625 0.625 0.812 0.75 0 ...
##  $ percent_atom_ga           : num  0.375 0.375 0.188 0 0.625 ...
##  $ percent_atom_in           : num  0 0 0 0.25 0.375 0 0.875 0.5 0.25 0 ...
##  $ lattice_vector_1_ang      : num  9.95 6.18 9.75 5 6.66 ...
##  $ lattice_vector_2_ang      : num  8.55 6.18 5.66 5 6.66 ...
##  $ lattice_vector_3_ang      : num  9.18 23.63 13.96 13.53 24.58 ...
##  $ lattice_angle_alpha_degree: num  90 90 91 90 90 ...
##  $ lattice_angle_beta_degree : num  90 90 91.1 90 90 ...
##  $ lattice_angle_gamma_degree: num  90 120 30.5 120 120 ...
##  $ formation_energy_ev_natom : num  0.068 0.249 0.1821 0.2172 0.0505 ...
##  $ bandgap_energy_ev         : num  3.44 2.92 2.74 3.35 1.38 ...
```

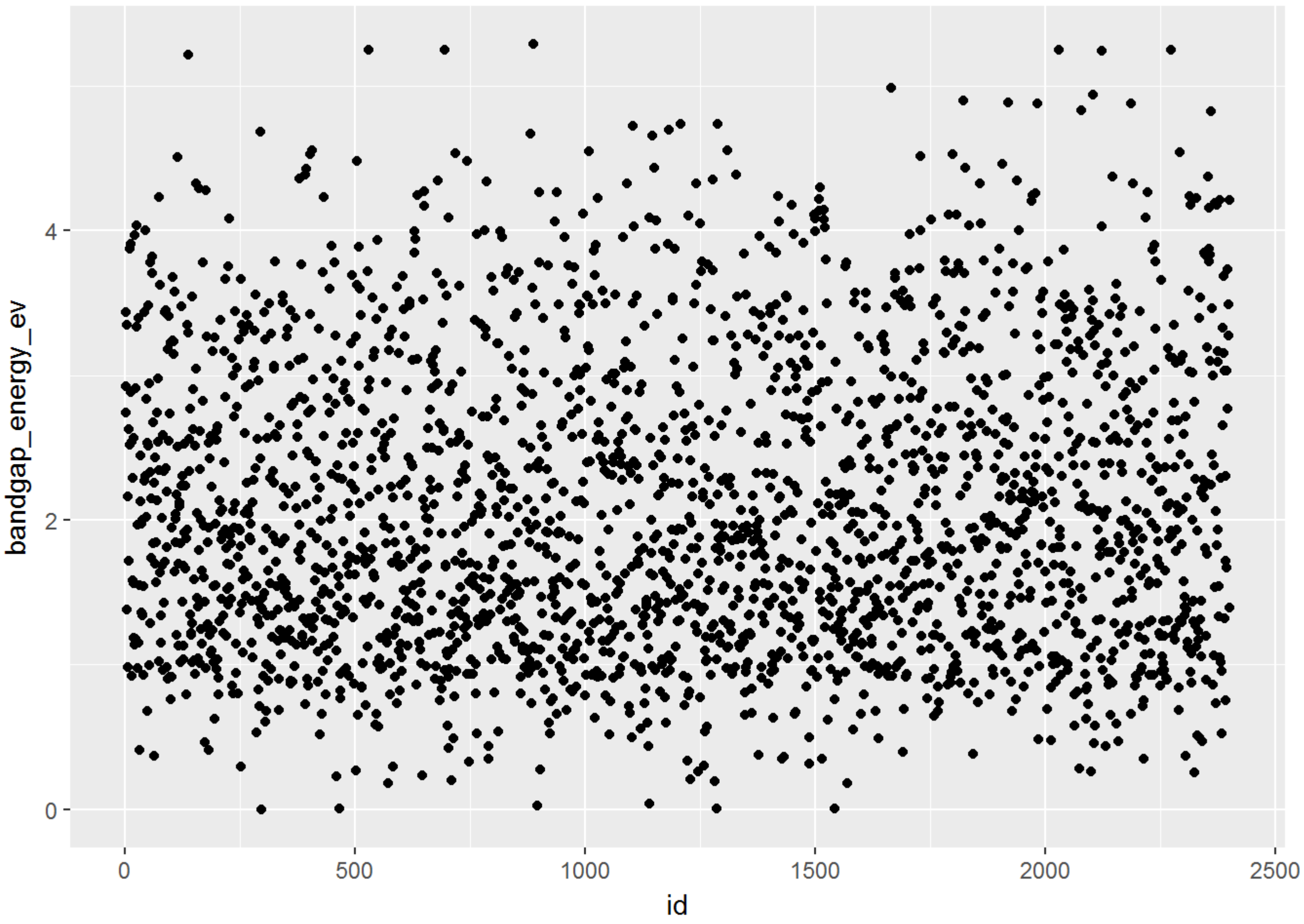
```
Training_Data %>%
  filter( id != "NA", spacegroup != "NA", number_of_total_atoms != "NA", percent_atom_al != "NA"
, percent_atom_ga != "NA", percent_atom_in != "NA", lattice_vector_1_ang != "NA", lattice_vector
_2_ang != "NA", lattice_vector_3_ang != "NA", lattice_angle_alpha_degree != "NA", lattice_angle_
beta_degree != "NA", lattice_angle_gamma_degree != "NA", formation_energy_ev_natom != "NA", band
gap_energy_ev != "NA") %>%
  select( id, spacegroup, number_of_total_atoms, percent_atom_al, percent_atom_ga, percent_atom_
in   , lattice_vector_1_ang, lattice_vector_2_ang, lattice_vector_3_ang, lattice_angle_alpha_deg
ree, lattice_angle_beta_degree, lattice_angle_gamma_degree, formation_energy_ev_natom,  bandgap_
energy_ev ) %>%
  group_by(id, spacegroup, number_of_total_atoms, percent_atom_al, percent_atom_ga, percent_atom
_in   , lattice_vector_1_ang, lattice_vector_2_ang, lattice_vector_3_ang, lattice_angle_alpha_de
gree, lattice_angle_beta_degree, lattice_angle_gamma_degree, formation_energy_ev_natom,  bandgap
_energy_ev)
```

```
## # A tibble: 2,400 x 14
## # Groups:   id, spacegroup, number_of_total_atoms, percent_atom_al,
## #   percent_atom_ga, percent_atom_in, lattice_vector_1_ang,
## #   lattice_vector_2_ang, lattice_vector_3_ang,
## #   lattice_angle_alpha_degree, lattice_angle_beta_degree,
## #   lattice_angle_gamma_degree, formation_energy_ev_natom,
## #   bandgap_energy_ev [2,400]
##       id spacegroup number_of_total_atoms percent_atom_al percent_atom_ga
##    <dbl>      <dbl>                 <dbl>           <dbl>           <dbl>
##  1     1         33                    80          0.6250          0.3750
##  2     2        194                    80          0.6250          0.3750
##  3     3        227                    40          0.8125          0.1875
##  4     4        167                    30          0.7500          0.0000
##  5     5        194                    80          0.0000          0.6250
##  6     6        227                    40          0.5625          0.4375
##  7     7        206                    80          0.0312          0.0938
##  8     8         12                    20          0.5000          0.0000
##  9     9        206                    80          0.5312          0.2188
## 10    10        194                    80          0.4062          0.5938
## # ... with 2,390 more rows, and 9 more variables: percent_atom_in <dbl>,
## #   lattice_vector_1_ang <dbl>, lattice_vector_2_ang <dbl>,
## #   lattice_vector_3_ang <dbl>, lattice_angle_alpha_degree <dbl>,
## #   lattice_angle_beta_degree <dbl>, lattice_angle_gamma_degree <dbl>,
## #   formation_energy_ev_natom <dbl>, bandgap_energy_ev <dbl>
```

Scatter plots to see the correlation between variable of x axis and y axis

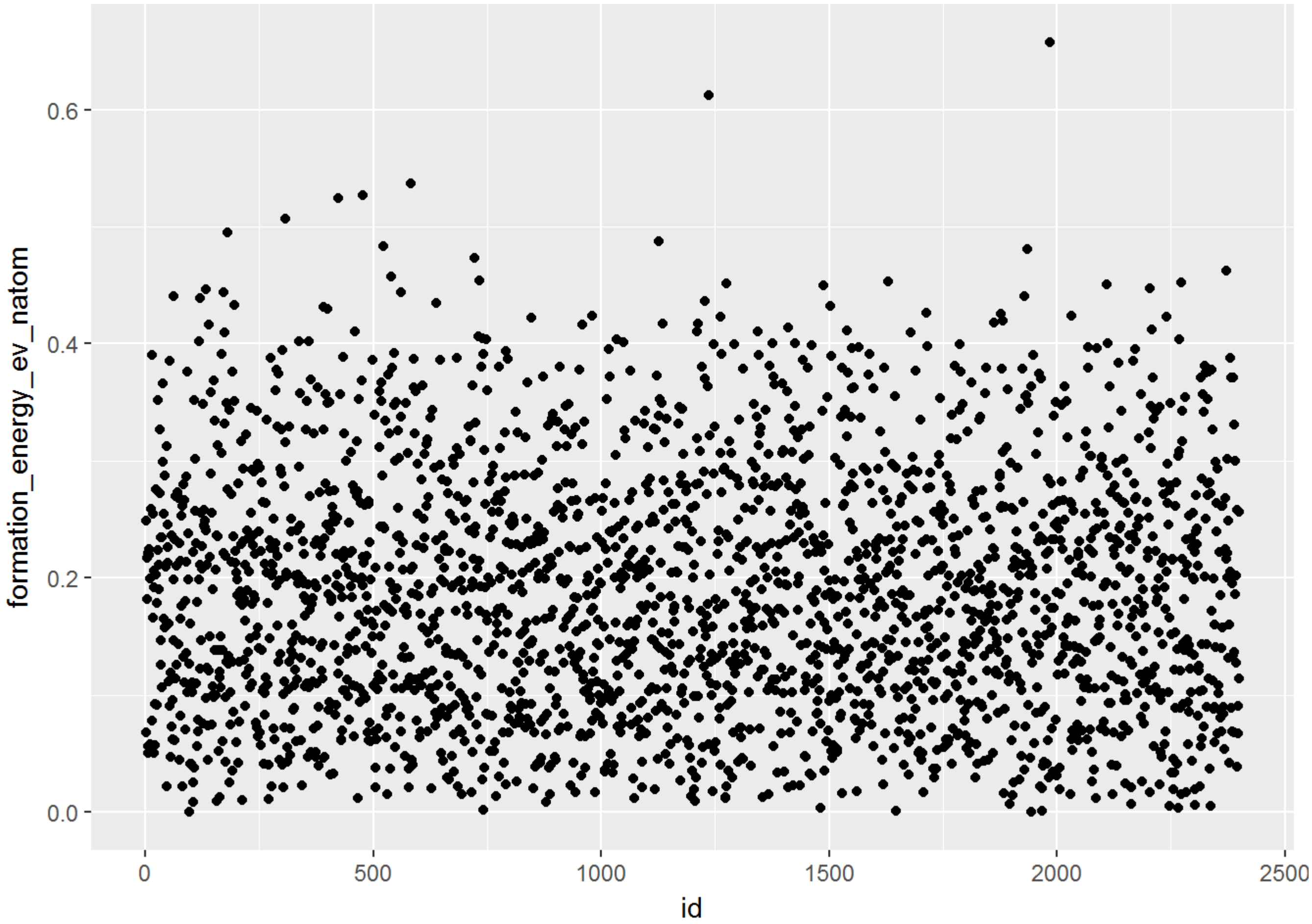
Plot for bandgap energy versus id

```
qplot(id, bandgap_energy_ev,data=Training_Data)
```



Plot for Formation Energy versus id

```
qplot(id, formation_energy_ev_natom,data=Training_Data)
```



Modelling

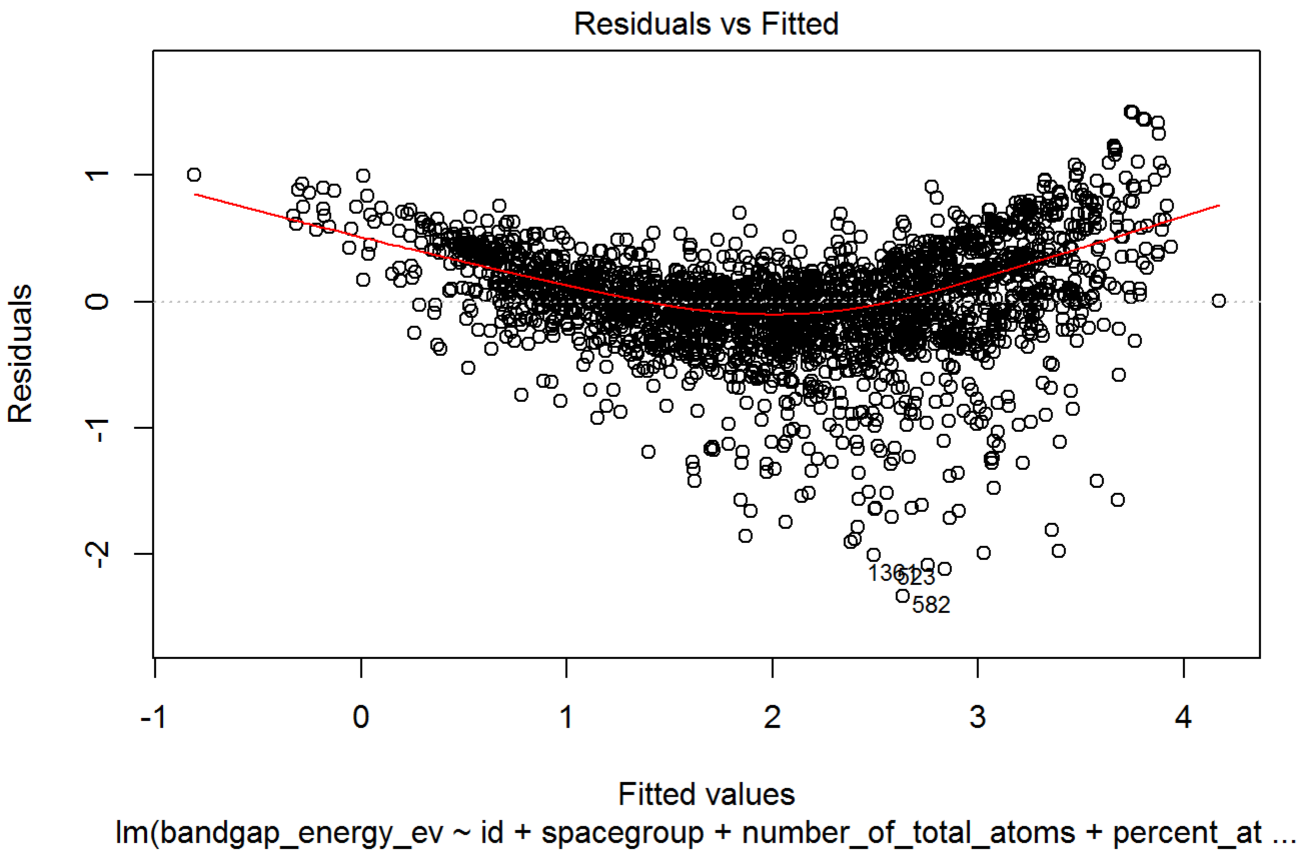
This is for prediction of two target properties :the formation energy (which is an indication of the stability of a new material) and the bandgap energy (which is an indication of the potential for transparency over the visible range)

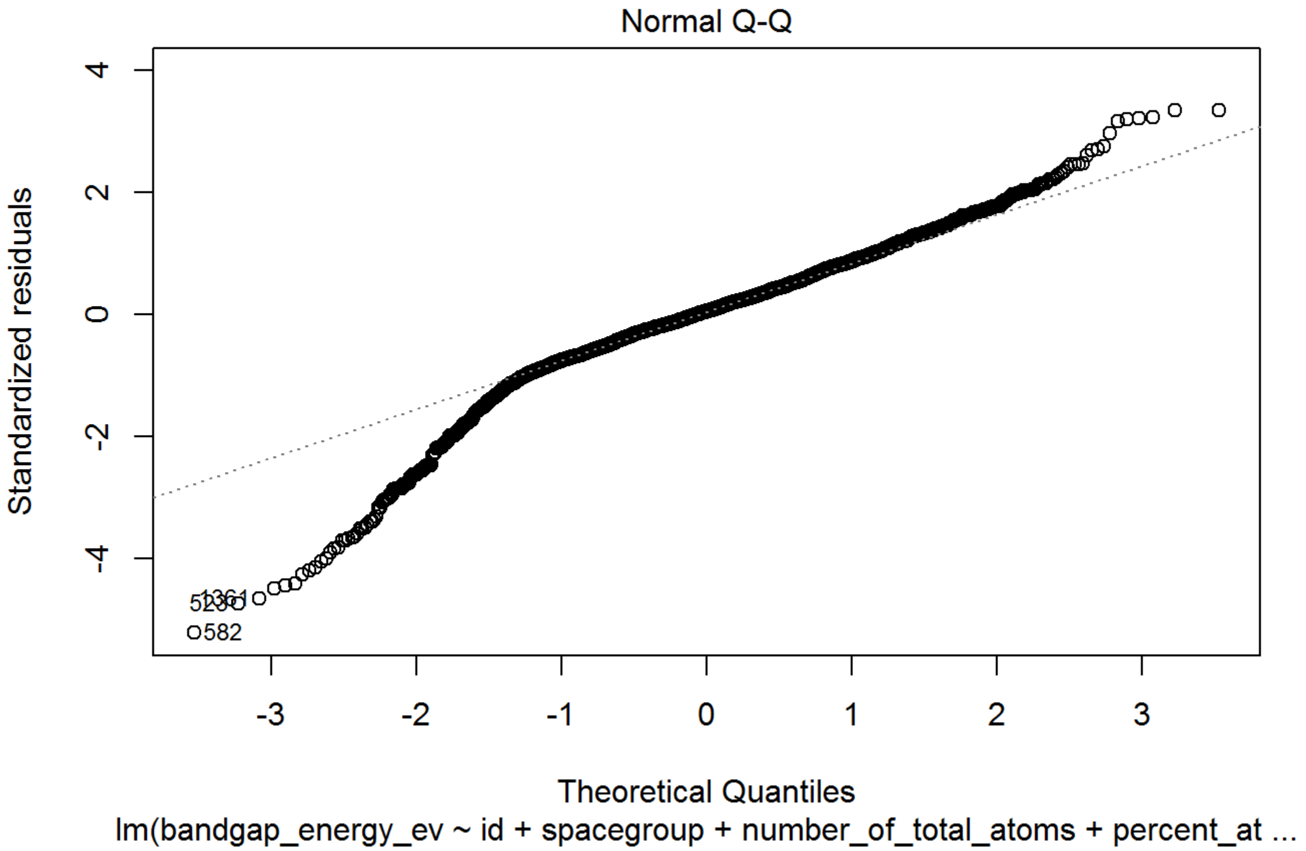
Model for bandgape energy

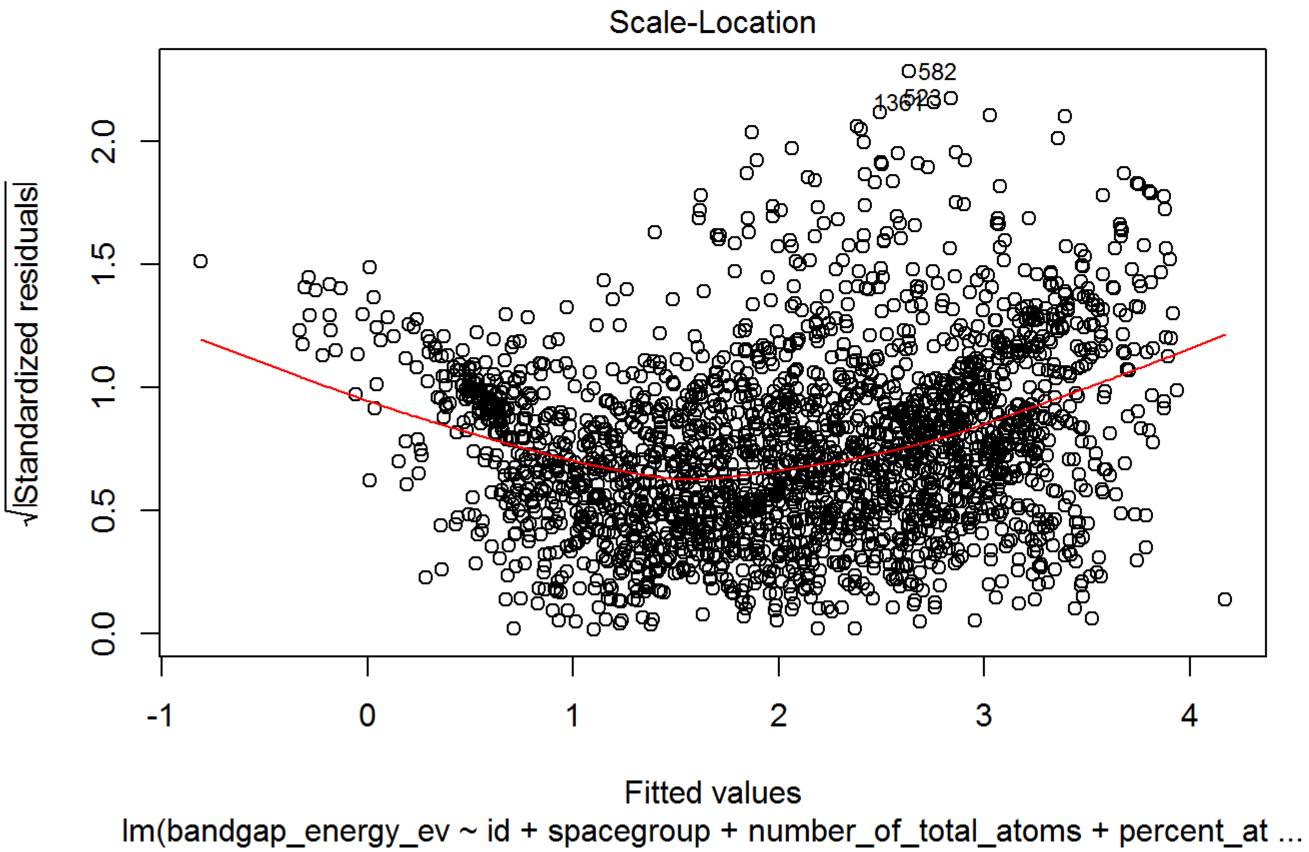
```
bandgap_energy_Model = lm(bandgap_energy_ev ~   id+spacegroup+number_of_total_atoms+percent_atom
_al +           percent_atom_ga+percent_atom_in+lattice_vector_1_ang+lattice_vector_2_ang +
lattice_vector_3_ang+lattice_angle_alpha_degree + lattice_angle_beta_degree+ lattice_angle_gamma
_degree    ,data=Training_Data)
summary(bandgap_energy_Model)
```

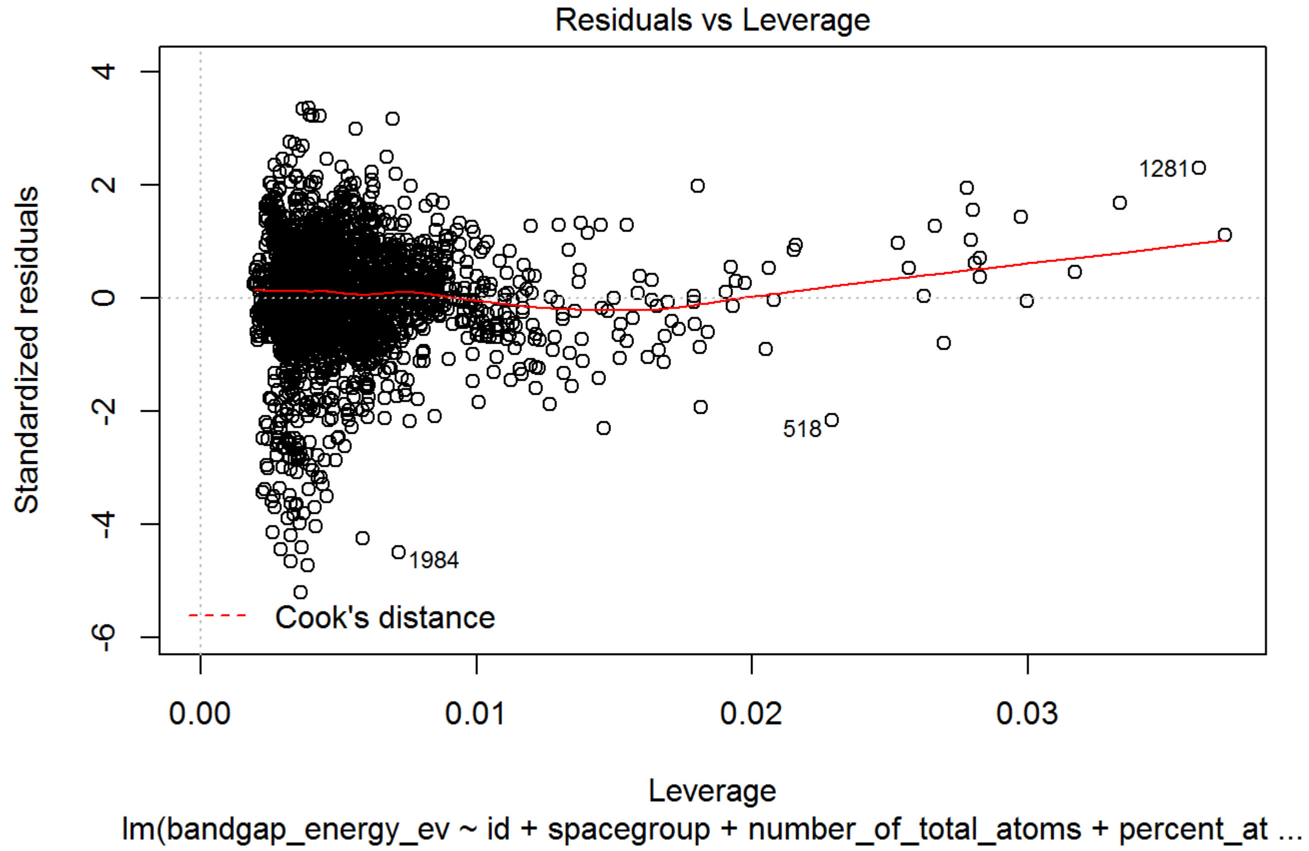
```
##
## Call:
## lm(formula = bandgap_energy_ev ~ id + spacegroup + number_of_total_atoms +
##     percent_atom_al + percent_atom_ga + percent_atom_in + lattice_vector_1_ang +
##     lattice_vector_2_ang + lattice_vector_3_ang + lattice_angle_alpha_degree +
##     lattice_angle_beta_degree + lattice_angle_gamma_degree, data = Training_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33608 -0.22049  0.02426  0.26132  1.50190
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.484e+02  2.227e+02   2.462  0.01388 *
## id                          3.762e-05  1.324e-05   2.841  0.00454 **
## spacegroup                 -3.123e-04  1.601e-04  -1.951  0.05119 .
## number_of_total_atoms       1.280e-02  1.582e-03   8.091 9.30e-16 ***
## percent_atom_al            -5.466e+02  2.227e+02  -2.454  0.01419 *
## percent_atom_ga            -5.481e+02  2.227e+02  -2.461  0.01392 *
## percent_atom_in            -5.499e+02  2.227e+02  -2.469  0.01362 *
## lattice_vector_1_ang       -5.580e-02  7.279e-03  -7.666 2.57e-14 ***
## lattice_vector_2_ang       -1.379e-01  1.897e-02  -7.266 4.98e-13 ***
## lattice_vector_3_ang       -9.701e-02  4.841e-03 -20.041  < 2e-16 ***
## lattice_angle_alpha_degree  5.855e-02  9.736e-03   6.014 2.09e-09 ***
## lattice_angle_beta_degree  -2.374e-02  5.799e-03  -4.094 4.39e-05 ***
## lattice_angle_gamma_degree  6.942e-03  5.944e-04  11.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4486 on 2387 degrees of freedom
## Multiple R-squared:  0.8025, Adjusted R-squared:  0.8015
## F-statistic: 808.1 on 12 and 2387 DF,  p-value: < 2.2e-16
```

```
plot(bandgap_energy_Model)
```


Residuals vs Fitted
Residuals
1
0
-1
-2
1361
523
582
-1
0
1
2
3
4
Fitted values
lm(bandgap_energy_ev ~ id + spacegroup + number_of_total_atoms + percent_at ...


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
1361
523
582
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(bandgap_energy_ev ~ id + spacegroup + number_of_total_atoms + percent_at ...



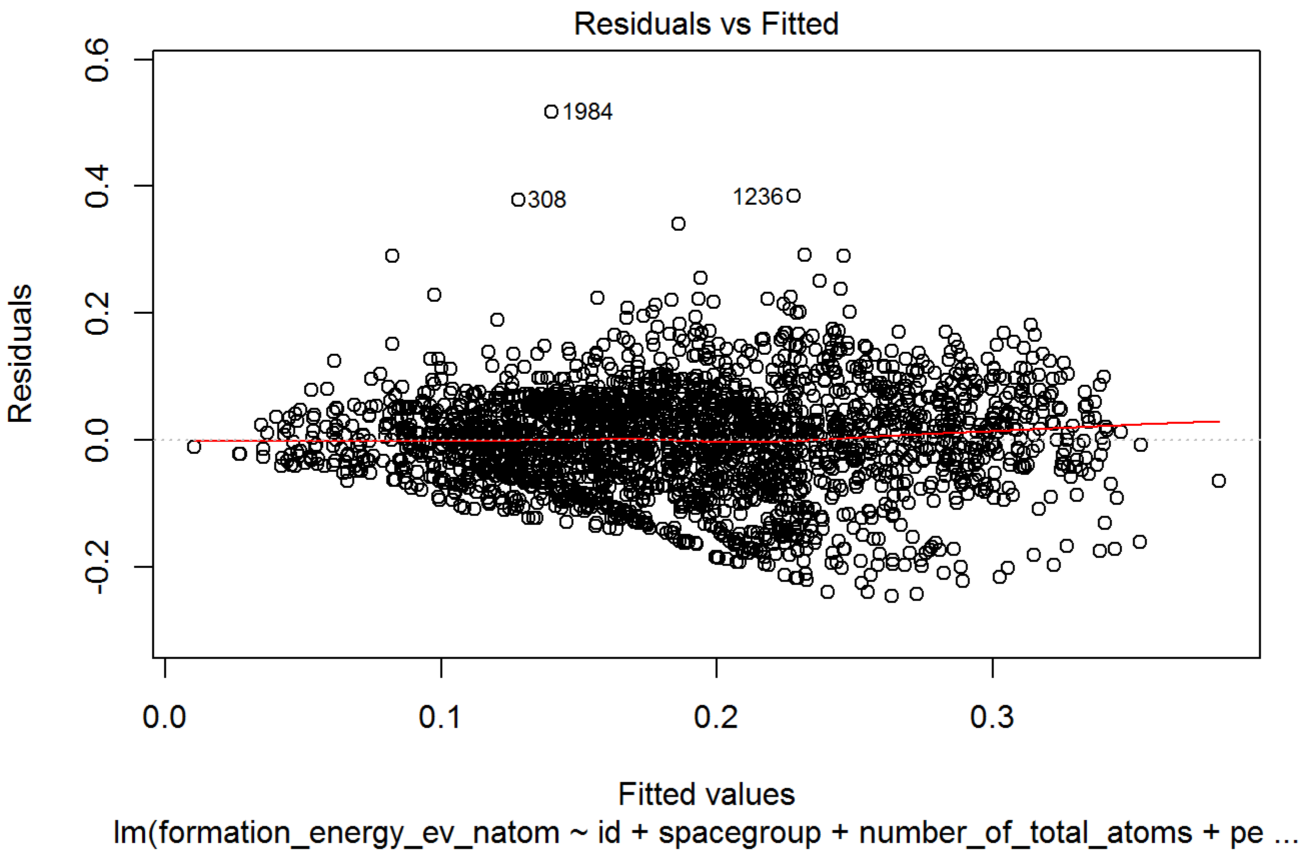


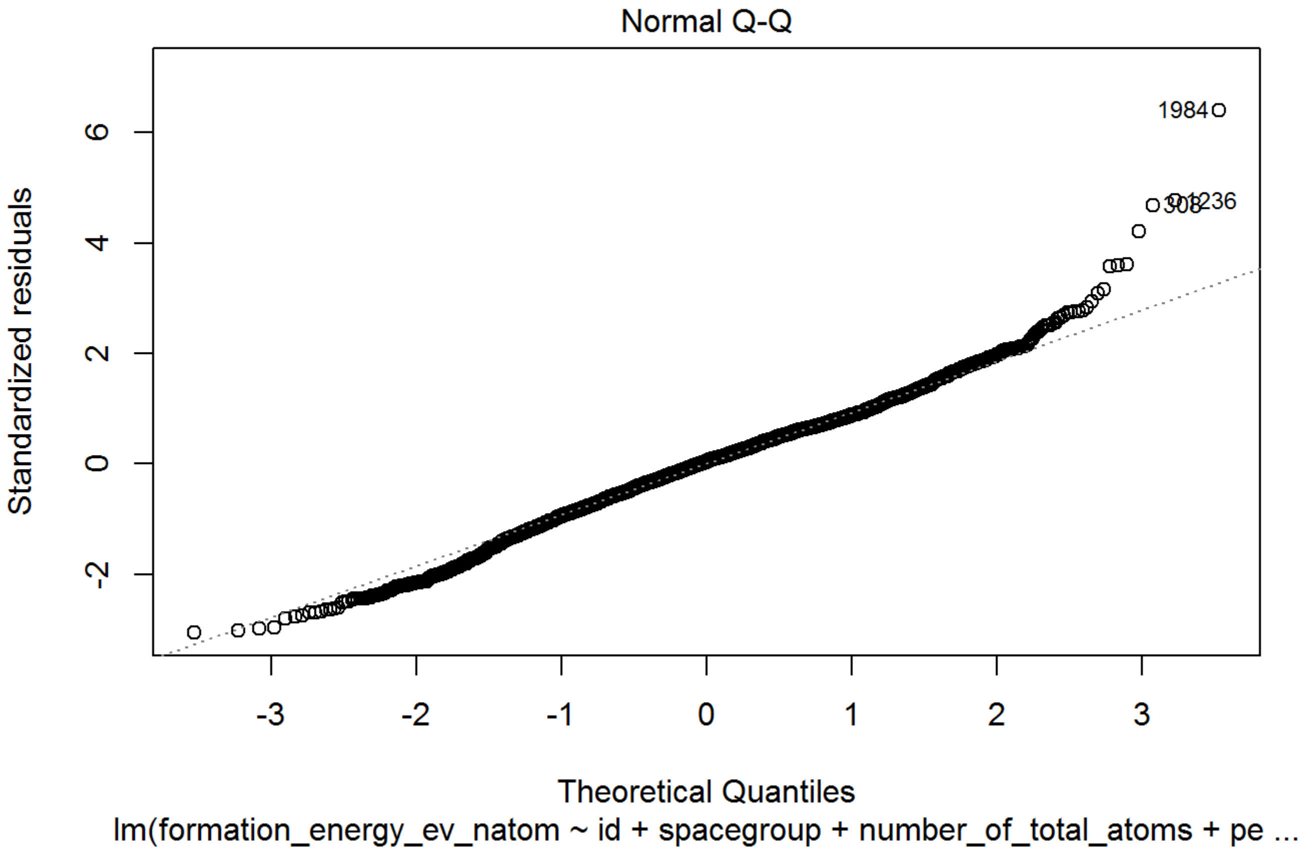
Model for formation energy

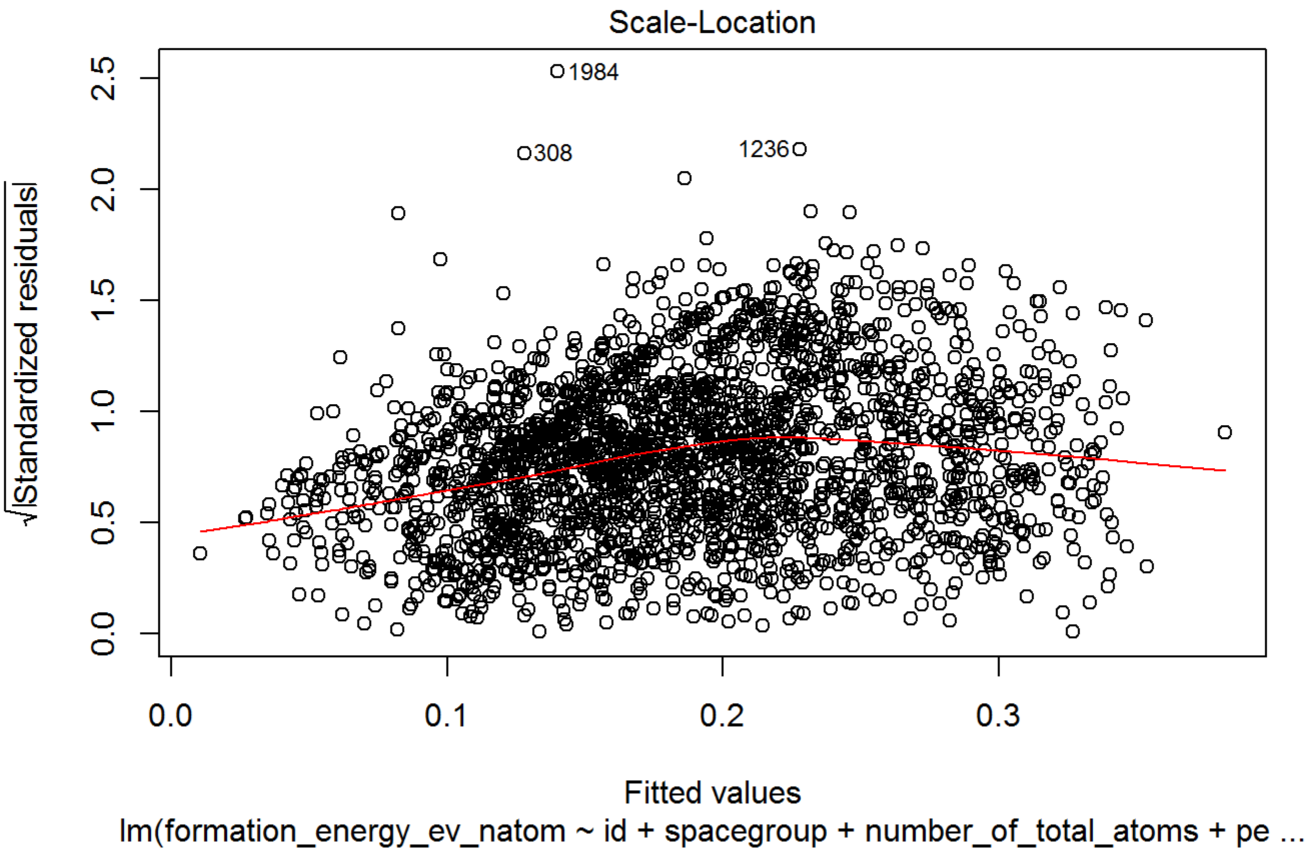
```
formation_energy_Model = lm(formation_energy_ev_natom ~   id+spacegroup+number_of_total_atoms+pe
rcent_atom_al +           percent_atom_ga+percent_atom_in+lattice_vector_1_ang+lattice_vector_2_
ang +      lattice_vector_3_ang+lattice_angle_alpha_degree + lattice_angle_beta_degree+ lattice_
angle_gamma_degree    ,data=Training_Data)
summary(formation_energy_Model)
```

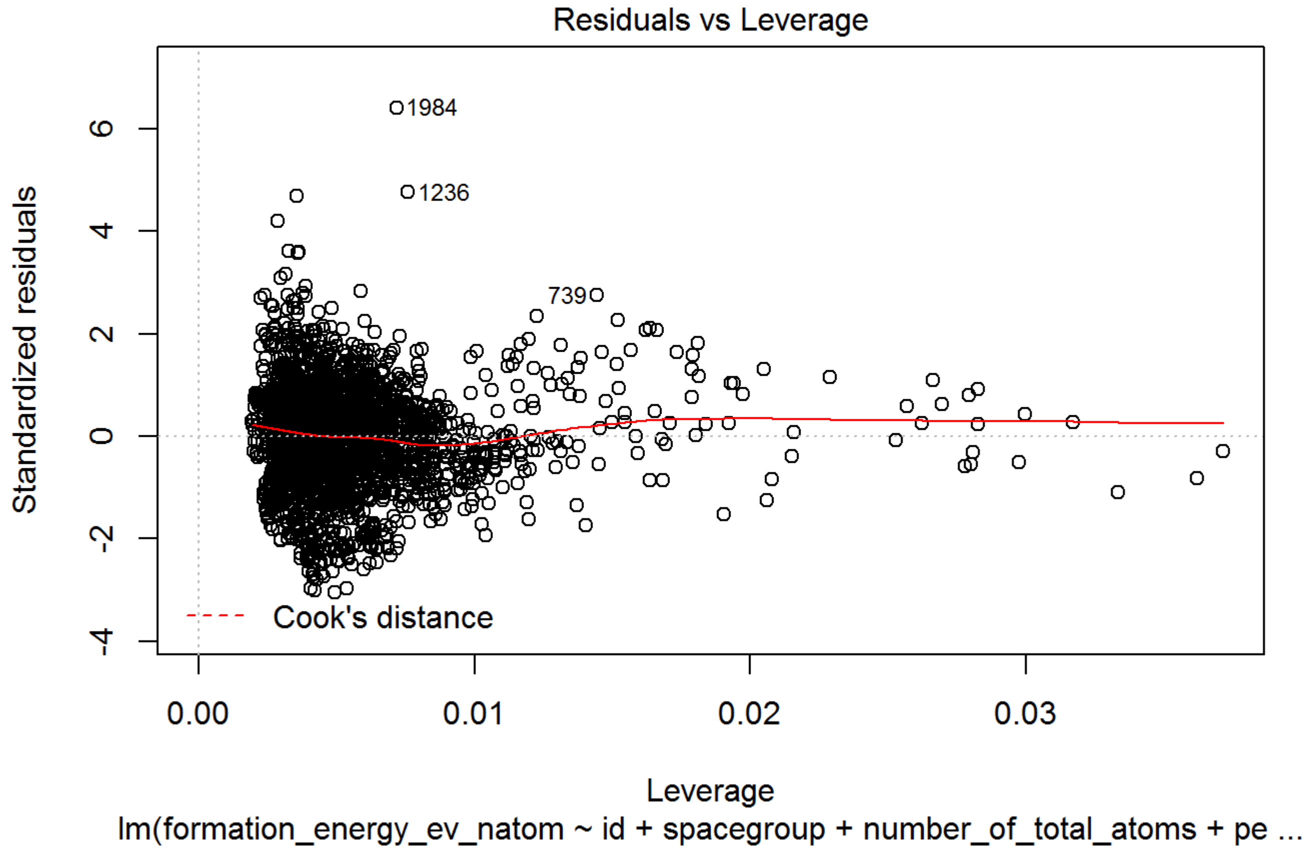
```
## 
## Call:
## lm(formula = formation_energy_ev_natom ~ id + spacegroup + number_of_total_atoms + 
##     percent_atom_al + percent_atom_ga + percent_atom_in + lattice_vector_1_ang + 
##     lattice_vector_2_ang + lattice_vector_3_ang + lattice_angle_alpha_degree + 
##     lattice_angle_beta_degree + lattice_angle_gamma_degree, data = Training_Data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.24675 -0.04983  0.00377  0.05110  0.51706 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -1.560e+02  4.020e+01  -3.879 0.000108 ***
## id                         -4.912e-06  2.391e-06  -2.055 0.040033 *  
## spacegroup                  6.404e-05  2.890e-05   2.216 0.026780 *  
## number_of_total_atoms      -7.824e-04  2.857e-04  -2.739 0.006208 ** 
## percent_atom_al             1.565e+02  4.020e+01   3.892 0.000102 ***
## percent_atom_ga             1.563e+02  4.020e+01   3.889 0.000104 ***
## percent_atom_in             1.565e+02  4.020e+01   3.893 0.000102 ***
## lattice_vector_1_ang        4.266e-03  1.314e-03   3.246 0.001186 ** 
## lattice_vector_2_ang        2.498e-03  3.425e-03   0.729 0.465858    
## lattice_vector_3_ang        1.126e-02  8.739e-04  12.882  < 2e-16 ***
## lattice_angle_alpha_degree -3.040e-03  1.758e-03  -1.730 0.083791 .  
## lattice_angle_beta_degree  -1.200e-03  1.047e-03  -1.147 0.251592    
## lattice_angle_gamma_degree -6.706e-04  1.073e-04  -6.249 4.87e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.08099 on 2387 degrees of freedom
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.3948 
## F-statistic: 131.4 on 12 and 2387 DF,  p-value: < 2.2e-16
```

```
plot(formation_energy_Model)
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
-0.2
1984
308
1236
0.0
0.1
0.2
0.3
Fitted values
lm(formation_energy_ev_natom ~ id + spacegroup + number_of_total_atoms + pe ...

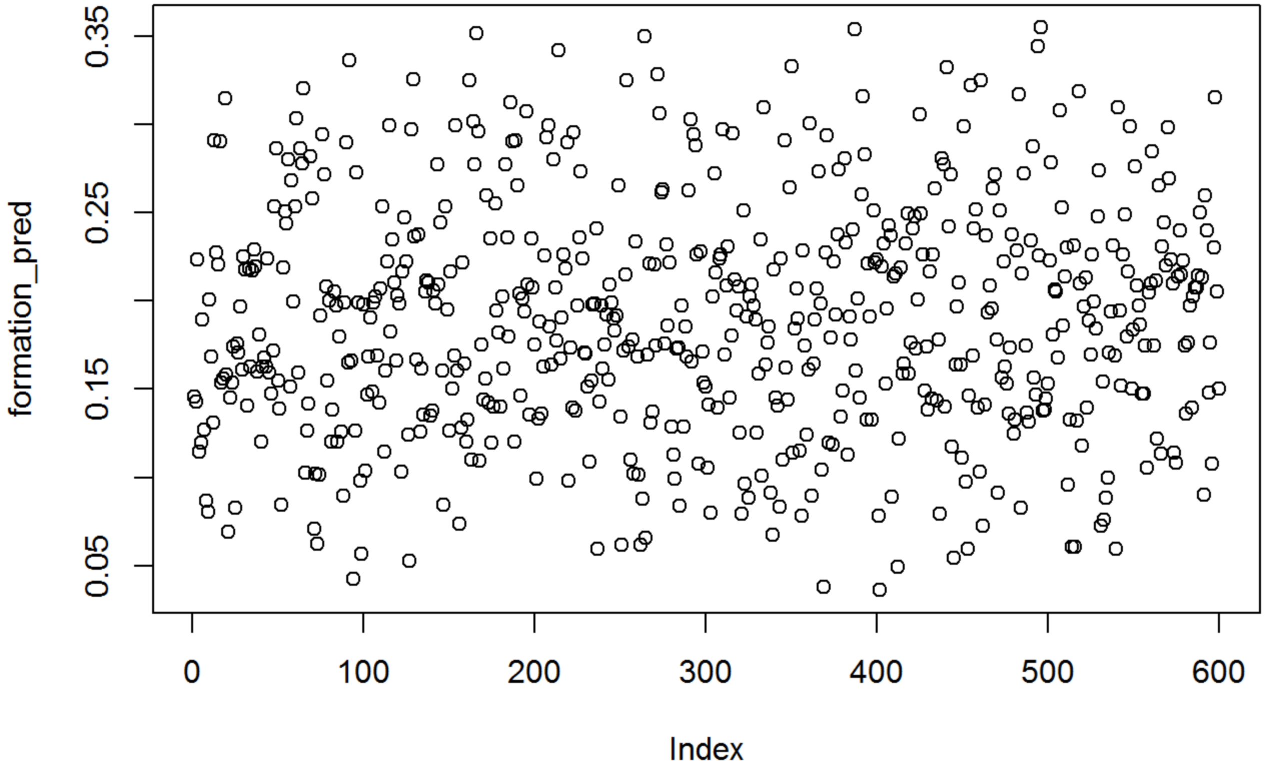

Normal Q-Q
Standardized residuals
6
4
2
0
-2
1984
308
1236
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(formation_energy_ev_natom ~ id + spacegroup + number_of_total_atoms + pe ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.1
0.2
0.3
1984
308
1236
Fitted values
lm(formation_energy_ev_natom ~ id + spacegroup + number_of_total_atoms + pe ...

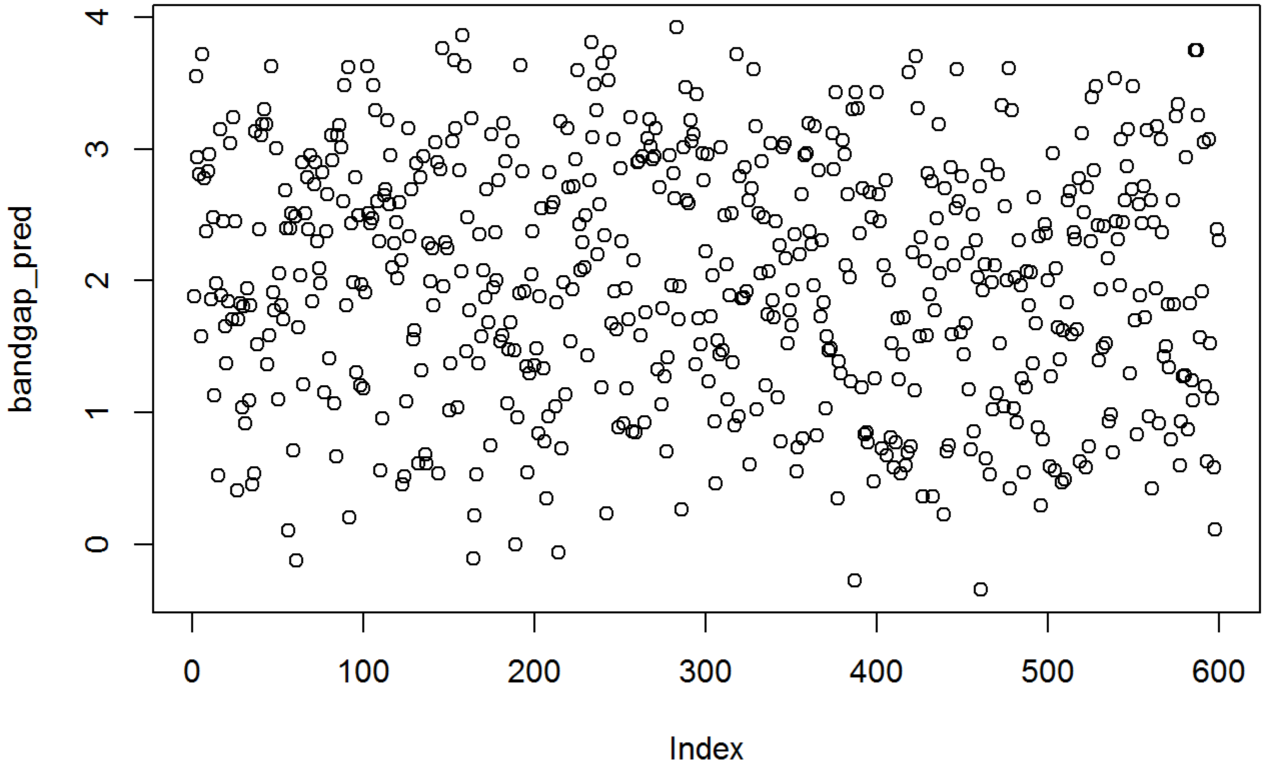

Residuals vs Leverage
Standardized residuals
-4
-2
0
2
4
6
0.00
0.01
0.02
0.03
1984
1236
739
Cook's distance
Leverage
lm(formation_energy_ev_natom ~ id + spacegroup + number_of_total_atoms + pe ...

Predictions on Test Data

```
formation_pred <- predict(formation_energy_Model, Test_Data)
plot(formation_pred)
```



```
bandgap_pred <- predict(bandgap_energy_Model, Test_Data)
plot(bandgap_pred)
```



Data for submission

```
Predicted_Outcome <- data.frame(id = 1:600, formation_energy_ev_natom = formation_pred, bandgap_energy_ev = bandgap_pred)
```

```
colnames(Predicted_Outcome) <- c("id","formation_energy_ev_natom","bandgap_energy_ev")
write.csv(Predicted_Outcome,"Predicted_Outcome.csv",row.names = FALSE)
```

This Prediction helps to avoid costly and inefficient trial-and-error of synthetic methods.