Wine Quality Prediction

Yang Cao
cao9@clemson.edu

I. Introduction

THIS project is aim to predict the quality ratings from various physicochemical properties of red and white wines. Due to the red wine dataset contains less observations, several machine learning methods were mainly applied to model the white wine dataset under regression and classification approaches. In order to get better performance of a data science model, the original dataset was splitted into two parts: training data and test data. It is found that the response (quality) is associated with most of predictors provided in this dataset. The random forest model shows the best performance in regression.

II. Dataset

In order to explore the relationships between various physicochemical parameters and the quality ratings for both Red and White wine. The data sets can be obtained from UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Wine+ Quality, and consists of 6000+ samples data for combined Red and White wine types. [1]

A. About Dataset

In the above reference, two datasets are avaliable, related to red and white vinho verde wine samples, from the north of Portugal. The predictors include physicochemical tests (e.g. PH and alcohol values), and the response is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). The classes are ordered but not balanced. [2] Number of Instances: red wine - 1599; white wine - 4898. Number of Attributes: 11 inputs + 1 output attribute. None missing attribute values exists.

B. Attributes Information

Input variables (based on physicochemical tests): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. Output variable (based on sensory data): quality (score between 0 and 10).

III. Exploratory data analysis

After loading the red and white wine datasets in R, I created a new attribute called ”color” in each file, and then merged these files into one. The new file is named ”wine”, which can be used for exploratory data analysis.

A. Summary of the Data Set

```
> dim(wine)
[1] 6497   13
> names(wine)
 [1] "fixed.acidity"       "volatile.acidity"    "citric.acid"         "residual.sugar"      "chlorides"           "free.sulfur.dioxide"  "total.sulfur.dioxide"
 [8] "density"             "pH"                  "sulphates"           "alcohol"             "quality"             "color"
> head(wine)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality color
1           3.8             0.310       0.02          11.1     0.036                  20                  114 0.99248 3.75      0.44    12.4       6 white
2           3.9             0.225       0.40           4.2     0.030                  29                  118 0.98900 3.57      0.36    12.8       8 white
3           4.2             0.170       0.36           1.8     0.029                  93                  161 0.98999 3.65      0.89    12.0       7 white
4           4.2             0.215       0.23           5.1     0.041                  64                  157 0.99688 3.42      0.44     8.0       3 white
5           4.4             0.320       0.39           4.3     0.030                  31                  127 0.98904 3.46      0.36    12.8       8 white
6           4.4             0.460       0.10           2.8     0.024                  31                  111 0.98816 3.48      0.34    13.1       6 white
> str(wine)
'data.frame':   6497 obs. of  13 variables:
 $ fixed.acidity       : num  3.8 3.9 4.2 4.2 4.4 4.4 4.4 4.5 4.6 4.6 ...
 $ volatile.acidity    : num  0.31 0.225 0.17 0.215 0.32 0.46 0.54 0.19 0.445 0.52 ...
 $ citric.acid         : num  0.02 0.4 0.36 0.23 0.39 0.1 0.09 0.21 0 0.15 ...
 $ residual.sugar      : num  11.1 4.2 1.8 5.1 4.3 2.8 5.1 0.95 1.4 2.1 ...
 $ chlorides           : num  0.036 0.03 0.029 0.041 0.03 0.024 0.038 0.033 0.053 0.054 ...
 $ free.sulfur.dioxide : num  20 29 93 64 31 31 52 89 11 8 ...
 $ total.sulfur.dioxide: num  114 118 161 157 127 111 97 159 178 65 ...
 $ density             : num  0.992 0.989 0.99 0.997 0.989 ...
 $ pH                  : num  3.75 3.57 3.65 3.42 3.46 3.48 3.41 3.34 3.79 3.9 ...
 $ sulphates           : num  0.44 0.36 0.89 0.44 0.36 0.34 0.4 0.42 0.55 0.56 ...
 $ alcohol             : num  12.4 12.8 12 8 12.8 13.1 12.2 8 10.2 13.1 ...
 $ quality             : int  6 8 7 3 8 6 7 5 5 4 ...
 $ color               : chr  "white" "white" "white" "white" "white" ...
> summary(wine)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  1.00      Min.   :  6.0       Min.   :0.9871   Min.   :2.720   Min.   :0.2200
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0       1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00      Median :118.0       Median :0.9949   Median :3.210   Median :0.5100
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53      Mean   :115.7       Mean   :0.9947   Mean   :3.219   Mean   :0.5313
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0       3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00      Max.   :440.0       Max.   :1.0390   Max.   :4.010   Max.   :2.0000
    alcohol         quality         color
 Min.   : 8.00   Min.   :3.000   Length:6497
 1st Qu.: 9.50   1st Qu.:5.000   Class :character
 Median :10.30   Median :6.000   Mode  :character
 Mean   :10.49   Mean   :5.818
 3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :14.90   Max.   :9.000
```

Fig. 1. Summary of the Data Set

Mean residual.sugar level is 5.4 g/l, but there exists a very sweet wine with 65.8 g/l residual.sugar (should be an outlier). Mean free.sulfur.dioxide is 30.5 ppm, while its maximum value is 289, which is much greater than Q3 (41 ppm). PH of wine is within range from 2.7 to 4, mean 3.2, which is relatively stable. Alcohol has the lightest concentration which is 8%, and the strongest is 14.9%. The Minimum quality ratings is 3, the mean is 5.8, and the maximum value is 9.

B. Distribution of Single Variables

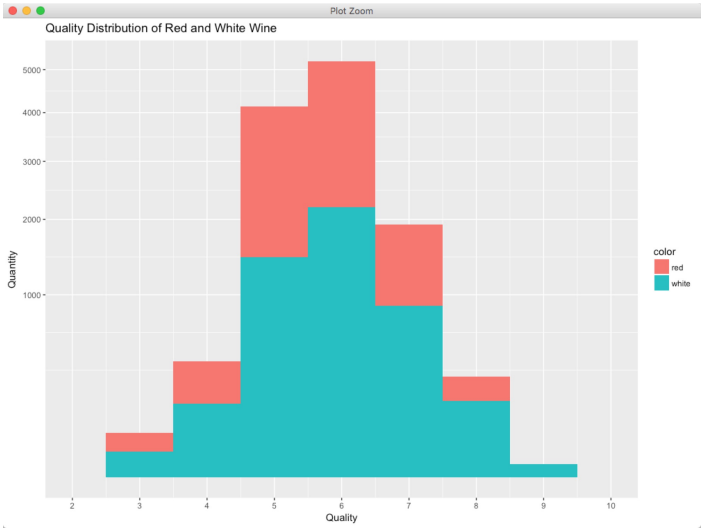

Fig. 2. Quality Distribution of Red and White Wine

Although the number of observations for red and white wine are different in dataset, but still we can see that for both colors, its normal distribution with almost the same picks at 5 and 6 quality ratings.

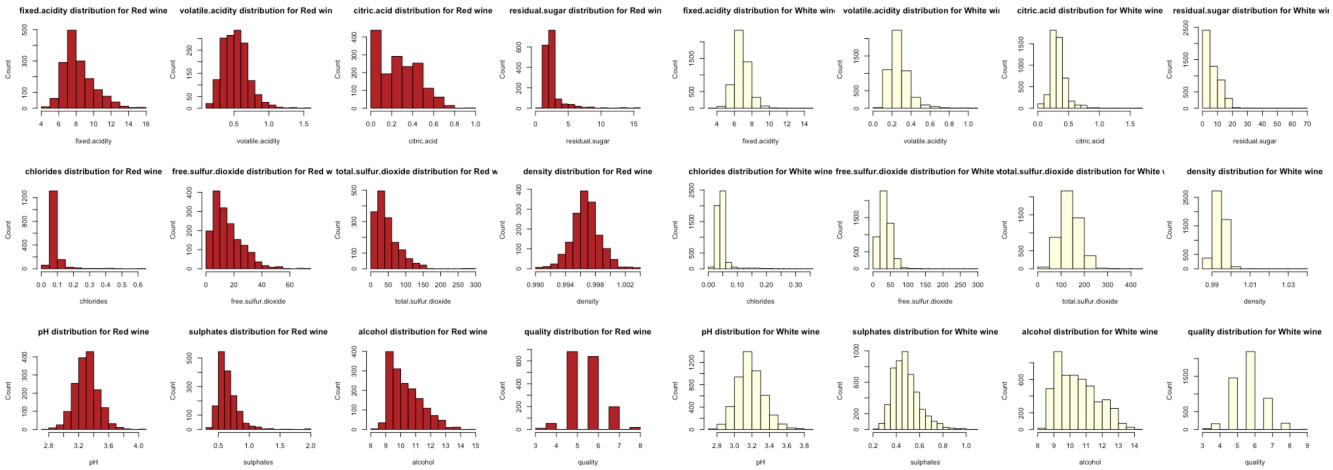

Fig. 3. Distribution of each variable in red wine dataset(left) and white wine dataset(right)

1) Histograms: Histogram is a good tool to show the distribution of variable values. You can find quality attribute has most values concentrated in the categories 5, 6 and 7. Only a small proportion is in the categories [3, 4] and [8, 9] and none in the categories [1, 2] and 10. Residual.sugar has a positively skewed distribution, even after eliminating the outliers, its distribution will remain skewed. Alcohol has an irregular shaped distribution but it does not have pronounced outliers.

2) Boxplots: Boxplot is a great method for each of the variables as another indicator of spread. In the following figures, fixed.acidity, residual.sugar, chlorides and free.sulfur.dioxide have outliers. If those outliers are eliminated, distribution of these variables may be taken to be symmetric. Some of the variables, such as density has a few outliers, but these are very different from the rest. Mostly outliers are on the larger side.

C. Distribution of Two and More Variables

A scatterplot matrix is derived to help us analyze the behavious and correlation of all variables. The scatterplots indicate the relationships between the target and predictors. A simple linear regression line is added to each single plot. In this figure, you can find the residual.sugar and free.sulfur.dioxide have outliers whose values are much greater than other values. I removed these two observations for reducing errors.

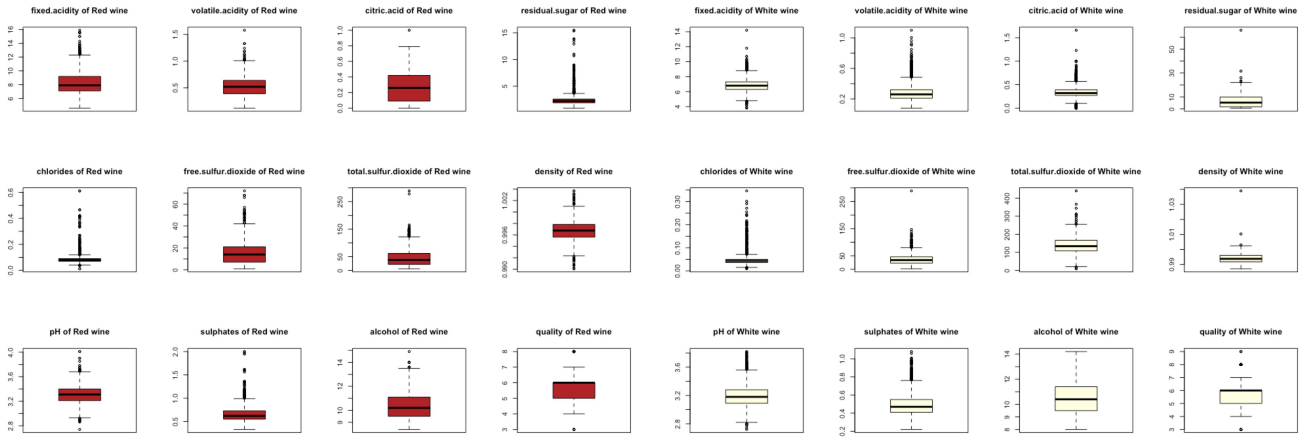

Fig. 4. Boxplots of each variable in red wine dataset(left) and white wine dataset(right)

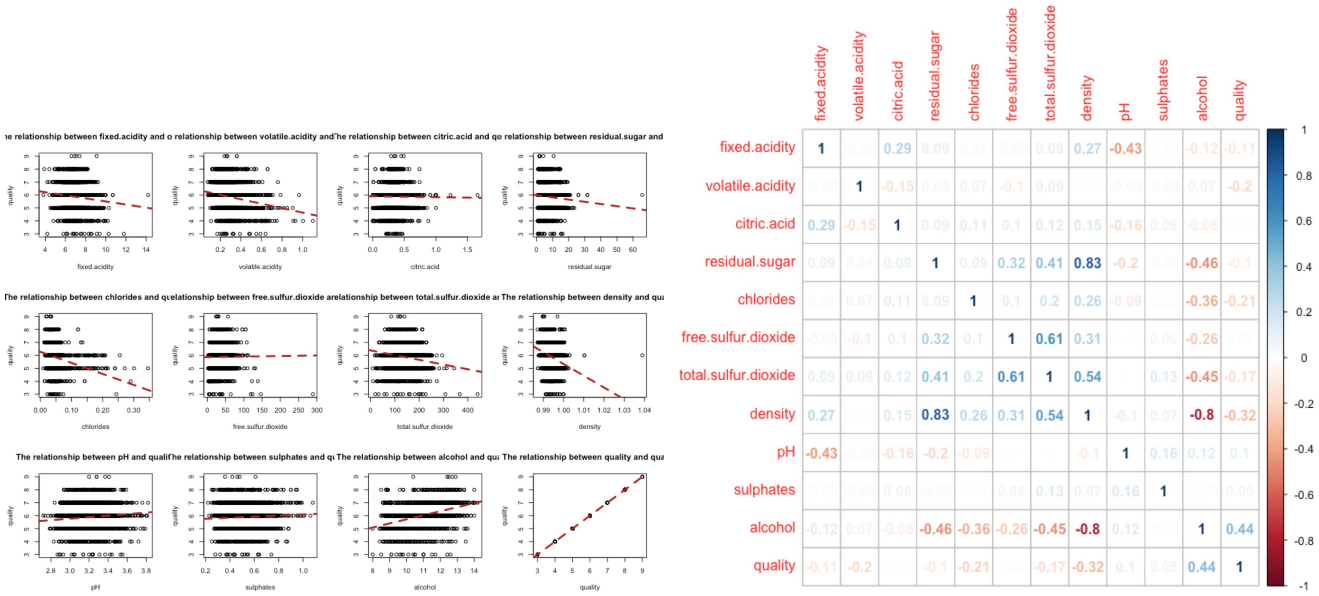

Fig. 5. Scatterplots of quality with each predictor(left) and Correlation between all the variables(right)

The color attribute was removed for correlation computing. We can see the correlation between variables: density and residual.sugar have a correlation coeficient 0.83, free.sulfur.dioxide and total.sulfur.dioxide have a correlation coeficient 0.61, which means a strong correlation between these predictors. The coeficient of quality and citric.acid indicate a poor correlation.

D. Create train and test sets

In order to predict the quality ratings more accurately, I used the white wine dataset which has greater observations. First divide the white wine dataset into train sets and a test sets. After sampleing the 75% observations randomly for training, the samples in whiteTrain dataset is 3672, and the whiteTest dataset has 1224 samples.

E. Feature Selection

Not all predictors are significant. A forward selection method is employed to build a working model. As the figure showing, we got the primary predictors: alcohol, volatile.acidity, residual.sugar, density, pH, sulphates, fixed.acidity, free.sulfur.dioxide.

IV. Model Selection and Validation

A. Multiple Linear Regression

A Multiple Linear Regression model was fitted using the key predictors obtained from the previous step. Look at the summary of this model, The residual standard error (RSE) is 0.75. A value of 0.75 means that actual quality ratings deviate from the true

```
Call:
lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
    density + pH + sulphates + fixed.acidity + free.sulfur.dioxide,
    data = whiteTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4901 -0.4900 -0.0473  0.4625  3.1087

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.000e+02  2.473e+01   8.087 8.20e-16 ***
alcohol              1.177e-01  3.257e-02   3.614 0.000306 ***
volatile.acidity    -1.733e+00  1.257e-01 -13.791  < 2e-16 ***
residual.sugar       9.640e-02  9.487e-03  10.161  < 2e-16 ***
density             -2.005e+02  2.504e+01  -8.006 1.58e-15 ***
pH                   8.196e-01  1.254e-01   6.538 7.11e-11 ***
sulphates            6.290e-01  1.184e-01   5.314 1.14e-07 ***
fixed.acidity        1.094e-01  2.561e-02   4.273 1.98e-05 ***
free.sulfur.dioxide  3.442e-03  8.092e-04   4.254 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7468 on 3663 degrees of freedom
Multiple R-squared:  0.2605,    Adjusted R-squared:  0.2589
F-statistic: 161.3 on 8 and 3663 DF,  p-value: < 2.2e-16
```

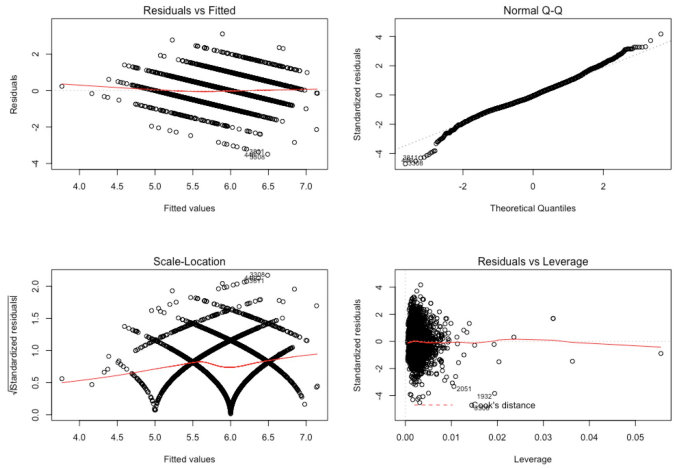

Fig. 6. Multiple Linear Regression Summary(left) and Plots(right)

regression line by approximately 0.75 units, on average. The R-squared is 26%. There is a relationship between the predictors and the response by testing the null hypothesis of whether all the regression coefficients are zero. The F-statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis. Looking at the p-values associated with each predictor's t-statistic, we see that all of these selected predictors have a statistically significant relationship with quality. Application of this model on test data gives the MSE is 0.56.

B. Polynomial Regression

```
Call:
lm(formula = quality ~ poly(alcohol, 2) + poly(volatile.acidity,
    2) + poly(residual.sugar, 4) + poly(free.sulfur.dioxide,
    5) + poly(fixed.acidity, 2) + sulphates + poly(density, 3) +
    poly(pH, 2), data = whiteTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2528 -0.4997 -0.0250  0.4463  3.2133

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.58214    0.05755  96.992  < 2e-16 ***
poly(alcohol, 2)1              5.06921    2.45665   2.063 0.039139 *
poly(alcohol, 2)2              5.10831    0.94426   5.410 6.71e-08 ***
poly(volatile.acidity, 2)1   -10.47305    0.76601 -13.672  < 2e-16 ***
poly(volatile.acidity, 2)2     2.33394    0.72789   3.206 0.001355 **
poly(residual.sugar, 4)1      26.83339    2.93148   9.154  < 2e-16 ***
poly(residual.sugar, 4)2     -10.37822    1.50855  -6.880 7.03e-12 ***
poly(residual.sugar, 4)3      -7.38820    1.68582  -4.383 1.21e-05 ***
poly(residual.sugar, 4)4      -2.90075    0.99358  -2.919 0.003527 **
poly(free.sulfur.dioxide, 5)1  3.22075    0.79113   4.071 4.78e-05 ***
poly(free.sulfur.dioxide, 5)2 -7.88995    0.73262 -10.769  < 2e-16 ***
poly(free.sulfur.dioxide, 5)3  3.53689    0.72857   4.855 1.26e-06 ***
poly(free.sulfur.dioxide, 5)4 -4.97615    0.72125  -6.899 6.13e-12 ***
poly(free.sulfur.dioxide, 5)5  1.37136    0.72075   1.903 0.057161 .
poly(fixed.acidity, 2)1        6.35929    1.30761   4.863 1.20e-06 ***
poly(fixed.acidity, 2)2       -2.47506    0.74877  -3.305 0.000957 ***
sulphates                      0.58034    0.11559   5.021 5.39e-07 ***
poly(density, 3)1            -37.10767    4.47273  -8.296  < 2e-16 ***
poly(density, 3)2              8.91594    1.63750   5.445 5.53e-08 ***
poly(density, 3)3             11.13194    1.83663   6.061 1.49e-09 ***
poly(pH, 2)1                   7.88063    1.15305   6.835 9.60e-12 ***
poly(pH, 2)2                   1.56682    0.73880   2.121 0.034008 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7176 on 3650 degrees of freedom
Multiple R-squared:  0.3198,    Adjusted R-squared:  0.3159
F-statistic: 81.72 on 21 and 3650 DF,  p-value: < 2.2e-16
```

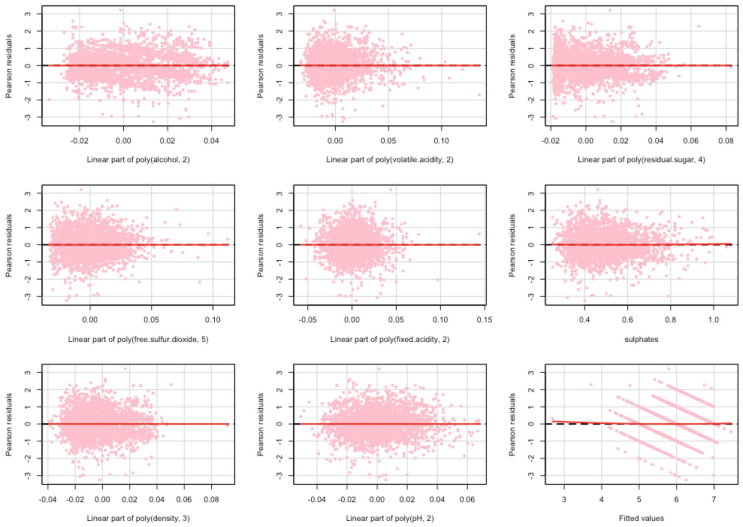

Fig. 7. Polynomial Regression Summary(left) and Plots(right)

In order to investigate whether a polynomial relationship fits the model better, an alternative model with polynomial terms of the significant variables is tried, which improves R2 value to 32%. Application of this model on test data gives the MSE is 0.54.

C. Generalized Additive Models

```
Call:
lm(formula = white_wine$quality ~ ns(alcohol, 2) + ns(volatile.acidity,
    5) + ns(residual.sugar, 4) + ns(free.sulfur.dioxide, 6) +
    ns(fixed.acidity, 2) + ns(sulphates, 1) + ns(density, 3) +
    ns(pH, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-3.2817 -0.4919 -0.0278  0.4623  3.3184

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.98578    0.22688  21.975  < 2e-16 ***
ns(alcohol, 2)1            0.12826    0.20961   0.612   0.5406
ns(alcohol, 2)2            0.69909    0.15907   4.395 1.13e-05 ***
ns(volatile.acidity, 5)1  -0.48156    0.10754  -4.478 7.72e-06 ***
ns(volatile.acidity, 5)2  -0.64433    0.12184  -5.288 1.29e-07 ***
ns(volatile.acidity, 5)3  -0.56059    0.11275  -4.972 6.86e-07 ***
ns(volatile.acidity, 5)4  -1.73106    0.29169  -5.935 3.15e-09 ***
ns(volatile.acidity, 5)5  -2.29934    0.31987  -7.188 7.55e-13 ***
ns(residual.sugar, 4)1     0.65100    0.06641   9.802  < 2e-16 ***
ns(residual.sugar, 4)2     1.91053    0.14084  13.566  < 2e-16 ***
ns(residual.sugar, 4)3     2.01906    0.27686   7.293 3.53e-13 ***
ns(residual.sugar, 4)4     0.52042    0.44457   1.171   0.2418
ns(free.sulfur.dioxide, 6)1  0.99316    0.08282  11.992  < 2e-16 ***
ns(free.sulfur.dioxide, 6)2  1.16042    0.10329  11.234  < 2e-16 ***
ns(free.sulfur.dioxide, 6)3  1.03973    0.09234  11.260  < 2e-16 ***
ns(free.sulfur.dioxide, 6)4  0.67810    0.11079   6.121 1.00e-09 ***
ns(free.sulfur.dioxide, 6)5  1.32668    0.24081   5.509 3.79e-08 ***
ns(free.sulfur.dioxide, 6)6 -0.59957    0.29828  -2.010   0.0445 *
ns(fixed.acidity, 2)1      1.34472    0.23823   5.645 1.75e-08 ***
ns(fixed.acidity, 2)2     -0.14897    0.33814  -0.441   0.6596
ns(sulphates, 1)           0.63167    0.10490   6.022 1.85e-09 ***
ns(density, 3)1           -2.78345    0.23798 -11.696  < 2e-16 ***
ns(density, 3)2           -2.65912    0.56210  -4.731 2.30e-06 ***
ns(density, 3)3           -1.39404    0.62967  -2.214   0.0269 *
ns(pH, 3)1                 0.36896    0.07437   4.961 7.24e-07 ***
ns(pH, 3)2                 0.29813    0.26945   1.106   0.2686
ns(pH, 3)3                 0.67414    0.15987   4.217 2.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7166 on 4869 degrees of freedom
Multiple R-squared:  0.3476,    Adjusted R-squared:  0.3441
F-statistic: 99.76 on 26 and 4869 DF,  p-value: < 2.2e-16
```

```
Call: gam(formula = white_wine$quality ~ s(alcohol, 2) + s(volatile.acidity,
    5) + s(residual.sugar, 4) + s(free.sulfur.dioxide, 6) + s(fixed.acidity,
    2) + s(sulphates, 1) + s(density, 3) + s(pH, 3))
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.1879 -0.4920 -0.0349  0.4564  3.2266

(Dispersion Parameter for gaussian family taken to be 0.5126)

    Null Deviance: 3832.691 on 4895 degrees of freedom
Residual Deviance: 2495.864 on 4869 degrees of freedom
AIC: 10651.4

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
s(alcohol, 2)              1  664.39  664.39 1296.102 < 2.2e-16 ***
s(volatile.acidity, 5)     1  189.27  189.27  369.237 < 2.2e-16 ***
s(residual.sugar, 4)       1   36.73   36.73   71.658 < 2.2e-16 ***
s(free.sulfur.dioxide, 6)  1   28.51   28.51   55.611 1.040e-13 ***
s(fixed.acidity, 2)        1   12.28   12.28   23.956 1.017e-06 ***
s(sulphates, 1)            1    8.22    8.22   16.029 6.332e-05 ***
s(density, 3)              1   21.64   21.64   42.210 9.018e-11 ***
s(pH, 3)                   1   32.34   32.34   63.088 2.437e-15 ***
Residuals               4869 2495.86    0.51
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects
                          Npar Df Npar F     Pr(F)
(Intercept)
s(alcohol, 2)                1 24.306 8.487e-07 ***
s(volatile.acidity, 5)       4 13.246 9.799e-11 ***
s(residual.sugar, 4)         3 20.719 2.461e-13 ***
s(free.sulfur.dioxide, 6)    5 50.607 < 2.2e-16 ***
s(fixed.acidity, 2)          1 17.678 2.661e-05 ***
s(sulphates, 1)              0  0.126 4.403e-05 ***
s(density, 3)                2 17.944 1.719e-08 ***
s(pH, 3)                     2 11.144 1.483e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 8. Generalized Additive Model1 and 2

Simple GAMs can be estimated with the lm() function using an appropriate choice of basis functions. For example, we can fit a GAM to predict quality ratings using natural spline functions of these primary predictors, such as the Generalized Additive Model1. The R-squared has been improved to 35%. Application of this model on test data shows the MSE is 0.53.

In order to fit more general sorts of GAMs, using smoothing splines or other components that cannot be expressed in terms of basis functions and then fit using least squares regression, we will need to use the gam library. Use the gam() function to fit a model with smoothing splines for these primary predictors, such as the Generalized Additive Model2. Application of this model on test data shows the MSE is 0.52.

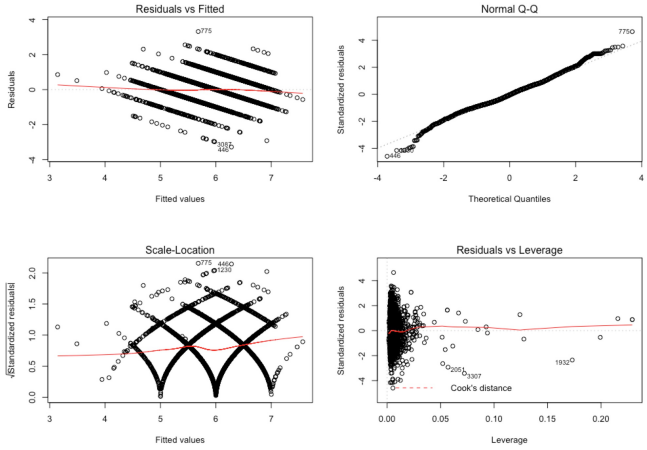

Fig. 9. Generalized Additive Model1 Plots

D. Regression Tree Method

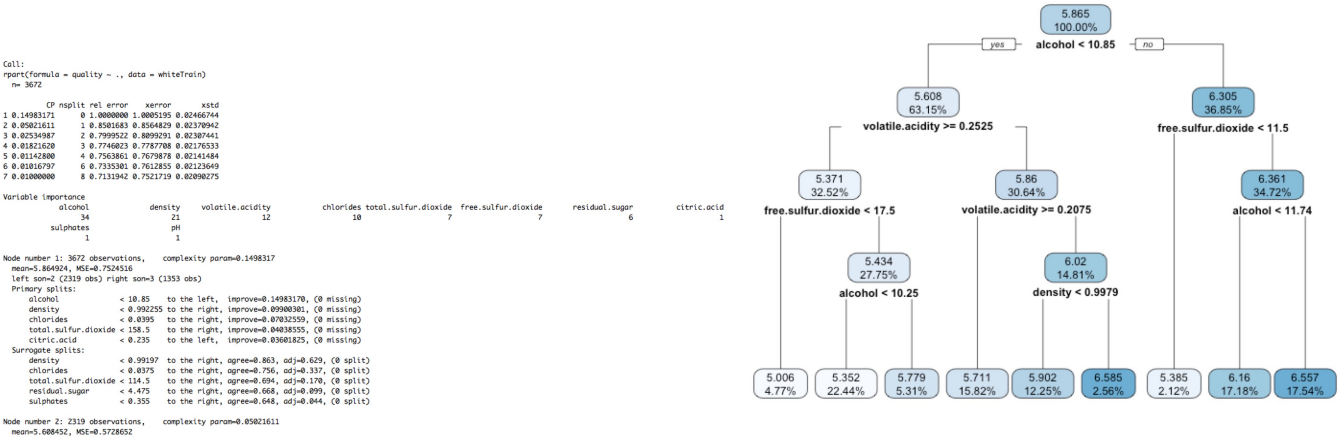

Fig. 10. Regression Tree Method Summary(left) and Tree(right)

A regression tree method is used with default parameters. Using the R formula interface, quality is regard as the response and include all other variables as predictors using the dot notation in training data. A variable importance matrix is given in the above figure, which indicates the alcohol, density and chlorides play important roles in this regression. Application of this model on test data shows the MSE is 0.59.

E. Random Forest

```
> model.rf <- randomForest(white_wine$quality[train] ~ . - quality, data = whiteTrain, importance=TRUE, do.trace=100)
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 100 |   0.3886    51.64 |
 200 |   0.3825    50.83 |
 300 |   0.3814    50.69 |
 400 |    0.379    50.37 |
 500 |   0.3789    50.36 |
> model.rf

Call:
 randomForest(formula = white_wine$quality[train] ~ . - quality,      data = whiteTrain, importance = TRUE, do.trace = 100)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

         Mean of squared residuals: 0.3789302
                   % Var explained: 49.64
```

Fig. 11. Random Forest Method

A Random Forest method is applied for regression in the training data. Application of this model on test data shows the MSE is 0.38, which indicates the best regression preformance in this project.

V. Conclusion

Several regression methods were applied in this project, such as Multiple Linear Regression, Polynomial Regression, Generalized Additive Models, Regression Tree and Random Forest. The evaluation for the performance of these models is based on R-squared and mse values. The Random Forest method shows the best performance compared with the other ones, because it gives the least MSE value, and explains the impact of predictors on the response by 49.6%. Some models do not perform well, a main reason is the samples may be not enough. In addition, feature selection and resampling methods are also used to test the performance of a data science model.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009. ISSN: 0167-9236.

[2] Paulo Cortez, University of Minho, Guimares, Portugal, http://www3.dsi.uminho.pt/pcortez. A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009