R Notebook

Dataset – Wine Quality

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Background of data

This dataset is obtained from UCI Machine learning repository to classify the quality of wines

Dataset Characteristics: Multivariate

Attribute characteristics: Real

Date Donated: 2009/ 10/07

Number of instances: 4898

Number of Attributes: 12

Missing values: N/A

Attributes

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

```
#Meenakshi Nagarajan
#nagarajan.12@wright.edu
#Load the data into 'wine_quality'
wine_quality <-
read.csv(file="/Users/meenakshinagarajan/Desktop/Datamining/winequality-
red.csv", header=TRUE,sep=";")
head(wine_quality)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
str(wine_quality)
```

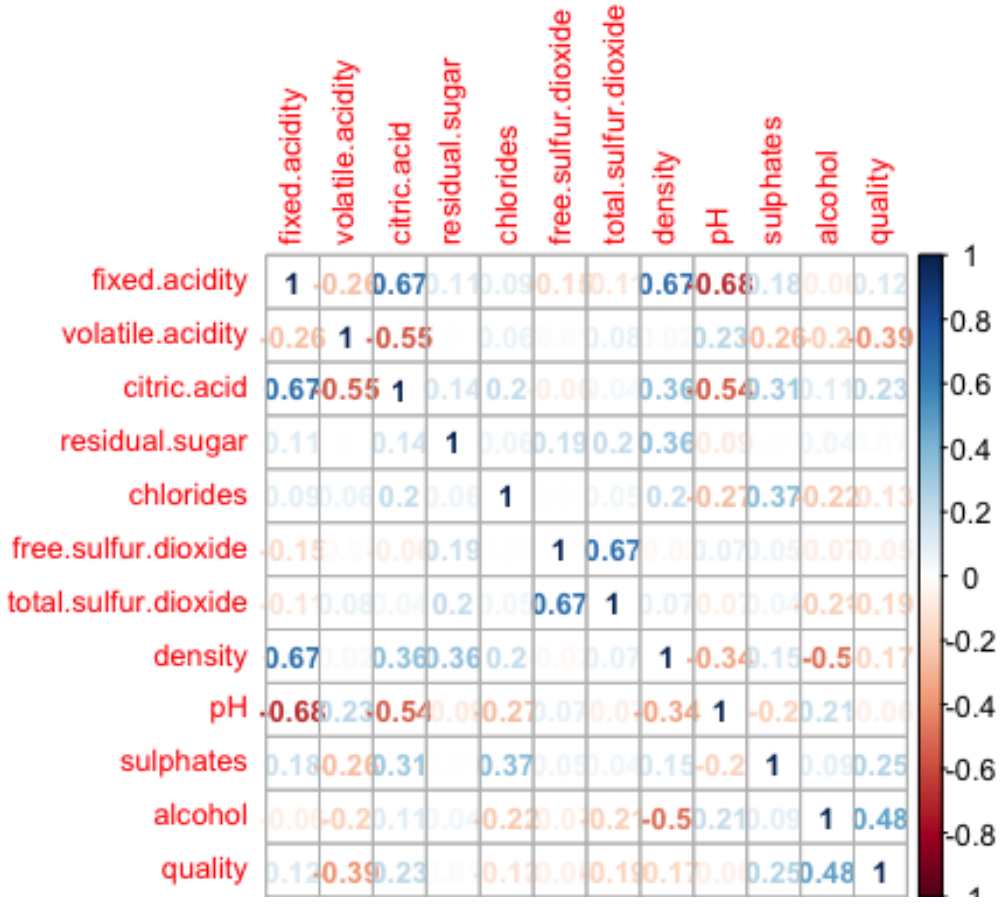
```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36
3.35 ...
```

```
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

We are going to look if we can predict the alcohol content of wine using other variables. Therefore,form a model with training data and use the model to predict alcohol with test data

Finding the correlation between variables

```
library(corrplot)
par(mfrow = c(1,1))
cor.wine_quality <- cor(wine_quality[1:12])
#correlation plot
corrplot(cor.wine_quality, method = 'number',tl.cex = 0.8,number.cex=0.8)
```



Correlation plot shows that alcohol content is strongly correlated with quality and density. All the other variables shows lesser correlation compared to these two variables.

Split between training and testing data

```
## 75% of the sample size
smp_size <- floor(0.75 * nrow(wine_quality))

## set the seed to make your partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(wine_quality)), size = smp_size)

train <- wine_quality[train_ind, ]
test <- wine_quality[-train_ind, ]
```

Forming the model with all variables using training data

```
fit1<- lm(alcohol ~ ., data=train)
summary(fit1)
```

```
##
## Call:
## lm(formula = alcohol ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0910 -0.3716 -0.0531  0.3516  2.3337
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.489e+02  1.519e+01  36.143  < 2e-16 ***
## fixed.acidity         4.711e-01  2.311e-02  20.380  < 2e-16 ***
## volatile.acidity      5.488e-01  1.274e-01   4.308 1.79e-05 ***
## citric.acid           8.260e-01  1.521e-01   5.430 6.82e-08 ***
## residual.sugar        2.753e-01  1.449e-02  18.994  < 2e-16 ***
## chlorides            -8.131e-01  4.643e-01  -1.751   0.0802 .
## free.sulfur.dioxide  -2.891e-03  2.365e-03  -1.222   0.2218
## total.sulfur.dioxide -1.200e-03  8.088e-04  -1.484   0.1380
## density              -5.591e+02  1.554e+01 -35.971  < 2e-16 ***
## pH                    3.580e+00  1.765e-01  20.284  < 2e-16 ***
## sulphates             9.214e-01  1.222e-01   7.541 9.21e-14 ***
## quality               2.503e-01  2.546e-02   9.832  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5941 on 1187 degrees of freedom
## Multiple R-squared:  0.6897, Adjusted R-squared:  0.6868
## F-statistic: 239.8 on 11 and 1187 DF,  p-value: < 2.2e-16
```

removing highly insignificant variable 'free.sulfur.dioxide'

```
fit2<- lm(alcohol ~ .-free.sulfur.dioxide, data=train)
summary(fit2)
```

```
##
## Call:
## lm(formula = alcohol ~ . - free.sulfur.dioxide, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07717 -0.37065 -0.05812  0.34265  2.36597
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.474e+02  1.514e+01  36.155  < 2e-16 ***
## fixed.acidity         4.681e-01  2.299e-02  20.359  < 2e-16 ***
## volatile.acidity      5.724e-01  1.260e-01   4.544 6.07e-06 ***
## citric.acid           8.560e-01  1.502e-01   5.701 1.50e-08 ***
## residual.sugar        2.738e-01  1.445e-02  18.954  < 2e-16 ***
## chlorides            -8.425e-01  4.638e-01  -1.817   0.0695 .
## total.sulfur.dioxide -1.881e-03  5.872e-04  -3.203   0.0014 **
## density              -5.575e+02  1.549e+01 -35.990  < 2e-16 ***
## pH                    3.549e+00  1.747e-01  20.318  < 2e-16 ***
## sulphates             9.163e-01  1.221e-01   7.502 1.23e-13 ***
## quality               2.487e-01  2.543e-02   9.779  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5942 on 1188 degrees of freedom
## Multiple R-squared:  0.6893, Adjusted R-squared:  0.6867
## F-statistic: 263.5 on 10 and 1188 DF,  p-value: < 2.2e-16
```

After removing free.sulfur.dioxide the model has been improved, but there still is a variable which is less significant compared to others.

removing insignificant variable 'chlorides' to improve the model

```
fit3<- lm(alcohol ~ .-free.sulfur.dioxide-chlorides, data=train)
summary(fit3)
```

```
##
## Call:
## lm(formula = alcohol ~ . - free.sulfur.dioxide - chlorides, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10120 -0.36774 -0.05437  0.34227  2.35085
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.527e+02  1.487e+01  37.173  < 2e-16 ***
## fixed.acidity         4.801e-01  2.205e-02  21.771  < 2e-16 ***
## volatile.acidity      5.215e-01  1.229e-01   4.243 2.38e-05 ***
## citric.acid           7.954e-01  1.466e-01   5.428 6.91e-08 ***
```

```
## residual.sugar        2.755e-01  1.443e-02  19.093  < 2e-16 ***
## total.sulfur.dioxide -1.739e-03  5.825e-04  -2.986  0.00289 **
## density              -5.633e+02  1.518e+01 -37.116  < 2e-16 ***
## pH                    3.642e+00  1.672e-01  21.785  < 2e-16 ***
## sulphates             8.473e-01  1.162e-01   7.292 5.56e-13 ***
## quality               2.532e-01  2.533e-02   9.997  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5948 on 1189 degrees of freedom
## Multiple R-squared:  0.6884, Adjusted R-squared:  0.6861
## F-statistic: 291.9 on 9 and 1189 DF,  p-value: < 2.2e-16
```

Comparing the nested models

```
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + quality
## Model 2: alcohol ~ (fixed.acidity + volatile.acidity + citric.acid +
residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + quality) - free.sulfur.dioxide
## Model 3: alcohol ~ (fixed.acidity + volatile.acidity + citric.acid +
residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + quality) - free.sulfur.dioxide -
##     chlorides
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1187 418.96
## 2   1188 419.49 -1  -0.52732 1.4940 0.22184
## 3   1189 420.65 -1  -1.16538 3.3018 0.06946 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 1 does not perform significantly better than Model 2 and Model 2 is not significantly better than Model 3. Therefore, Model 3 is used to predict the alcohol content.

Calculating Root-Mean-Square-Error

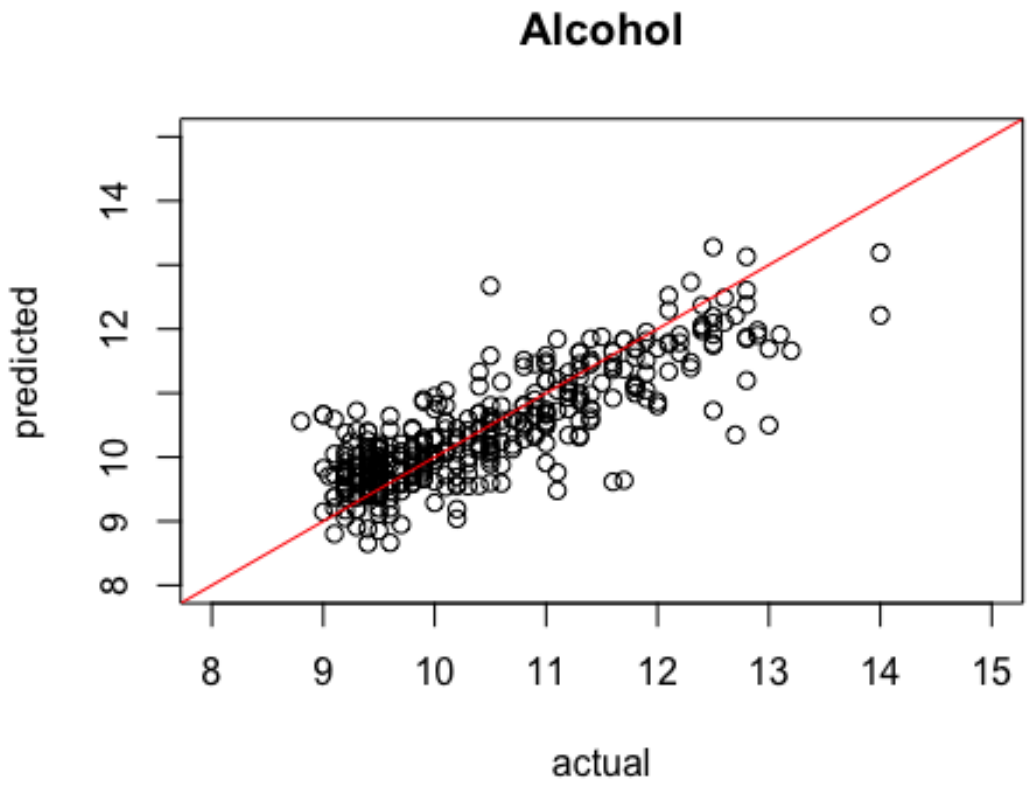
```
RMSE <- function(predicted, true) mean((predicted-true)^2)^.5
RMSE(predict(fit3, test), test$alcohol)
```

```
## [1] 0.6019659
```

The obtained value is not very less or not too high. This suggests finding accuracy of the model by comparing actual vs. predicted

Compare actual vs. predicted values.

```
plot(test[,"alcohol"], predict(fit3, test),
  xlim=c(8,15), ylim=c(8,15), xlab = "actual", ylab = "predicted",
  main = "Alcohol")
abline(0,1, col="red")
```




```
cor(test[,"alcohol"], predict(fit3, test))
```

```
## [1] 0.8298045
```

The actual vs. predicted values graph shows that the predicted values are reasonably close to the regressed dialgonal line. The correlation value between actual and predicted is 0.829 which shows that model's predictions are good.

Thus, this model could be used to get an approximate value of alcohol content