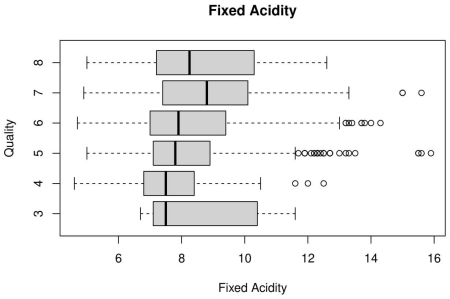
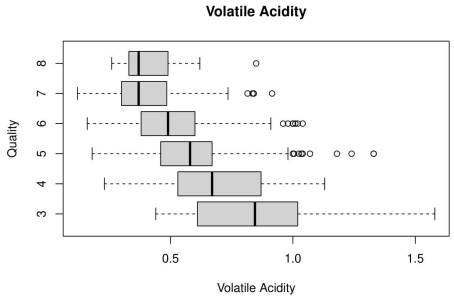
Predicting Wine Preferences

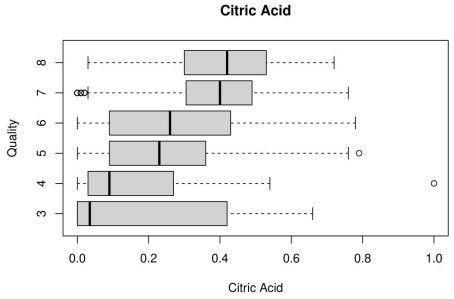
Will Firmin

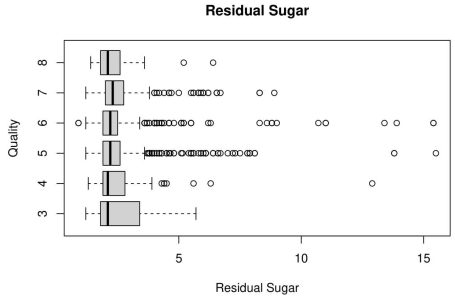
Dataset

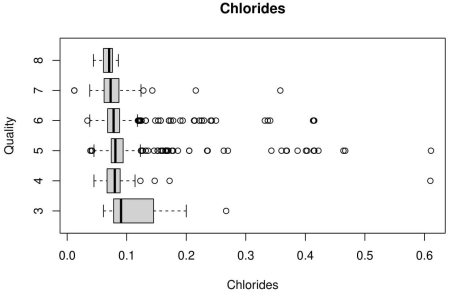
- Wine samples from Northern Portugal
- 1599 observations
- Response variable: quality ratings
- Ratings are on a scale of 1-10
- Data only has ratings from 3-8
- Predictors: physico-chemical characteristics
- 11 numeric predictor variables
    - Alcohol content
    - Density
    - Acidity
    - Etc.

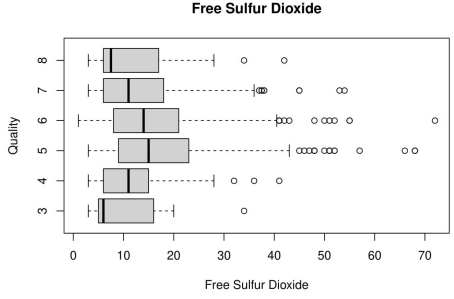

Fixed Acidity
Quality
Fixed Acidity

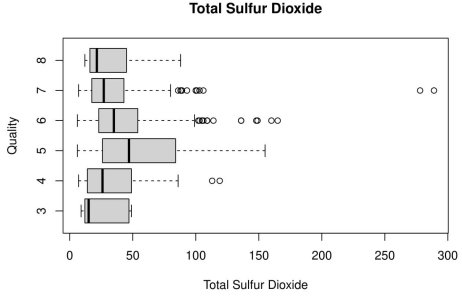

Volatile Acidity
Quality
Volatile Acidity

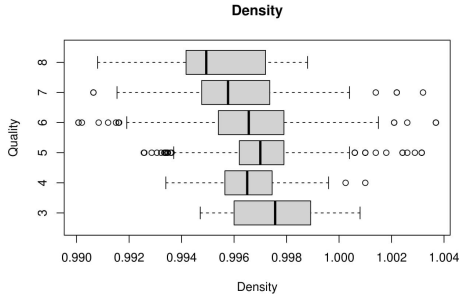

Citric Acid
Quality
Citric Acid

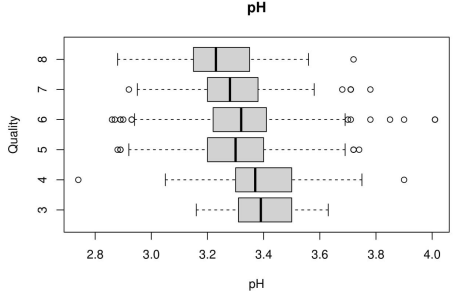

Residual Sugar
Quality
Residual Sugar

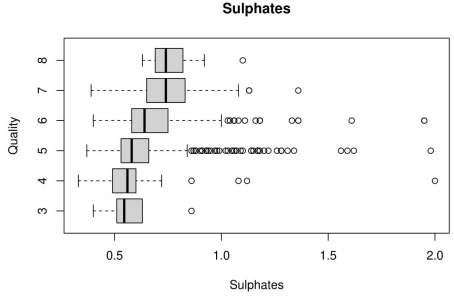

Chlorides
Quality
Chlorides

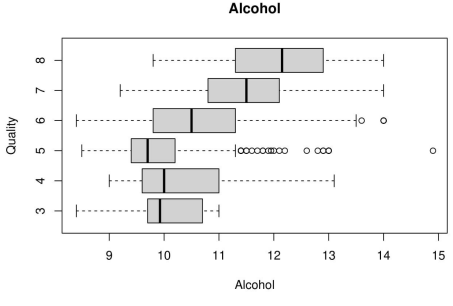

Free Sulfur Dioxide
Quality
Free Sulfur Dioxide


Total Sulfur Dioxide
Quality
Total Sulfur Dioxide

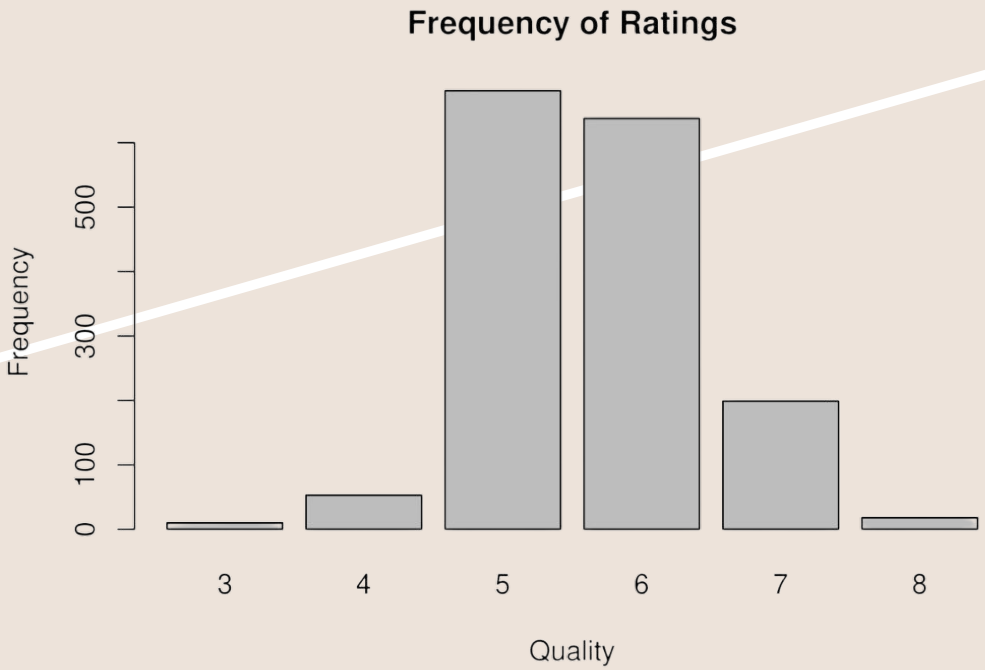

Density
Quality
Density


pH
Quality
pH


Sulphates
Quality
Sulphates


Alcohol
Quality
Alcohol

Binary Classification



Logistic Regression

```
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           57.779068  96.092487   0.601   0.5476
## fixed.acidity          0.141564   0.119470   1.185   0.2360
## volatile.acidity      -3.175509   0.588609  -5.395 6.85e-08 ***
## citric.acid           -1.416295   0.683255  -2.073   0.0382 *
## residual.sugar         0.010385   0.067937   0.153   0.8785
## chlorides             -3.528084   1.813208  -1.946   0.0517 .
## free.sulfur.dioxide    0.015112   0.009999   1.511   0.1307
## total.sulfur.dioxide  -0.015412   0.003556  -4.334 1.47e-05 ***
## density              -64.890175  98.105536  -0.661   0.5083
## pH                    -0.791903   0.859508  -0.921   0.3569
## sulphates              2.505958   0.539572   4.644 3.41e-06 ***
## alcohol                0.935579   0.127084   7.362 1.81e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Accuracy: **74.3%**

Important variables line up with previous observations

Linear Discriminant Analysis

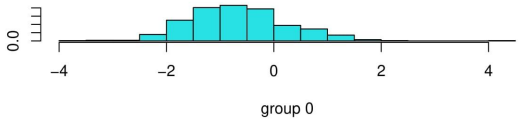
```
## Call:
## lda(quality ~ ., data = wineC, subset = train)
##
## Prior probabilities of groups:
##         0         1
## 0.4557641 0.5442359
##
## Group means:
##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 0      8.117255        0.5857941   0.2369608       2.564314 0.09368627
## 1      8.444499        0.4738916   0.2993596       2.486946 0.08284236
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## 0            16.66078             53.32843 0.9970740 3.313118 0.6220000
## 1            14.98358             38.11987 0.9963625 3.311806 0.6893924
##     alcohol
## 0  9.919706
## 1 10.912972
```

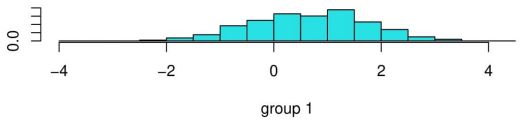
Accuracy: **73.8%**

Less accurate than logistic

Sizable differences in group means

Noticable separation below





Quadratic Discriminant Analysis

```
## Quadratic Discriminant Analysis
##
## 1599 samples
##   11 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1008, 1007, 1007, 1007, 1007, 1007, ...
## Resampling results:
##
##   Accuracy  Kappa
##   0.717648  0.4188196
```

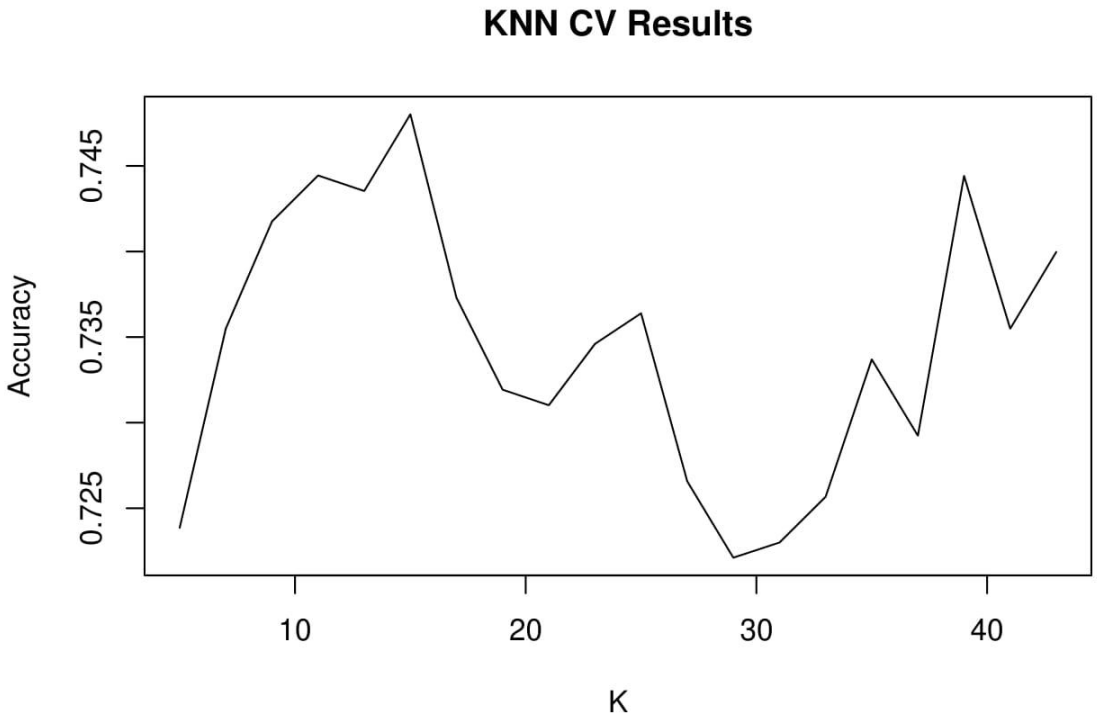
Accuracy: **69.8%**

Worse than both previous

Indicates more linear data

qda.pred	0	1
0	139	50
1	95	196

K-Nearest Neighbors



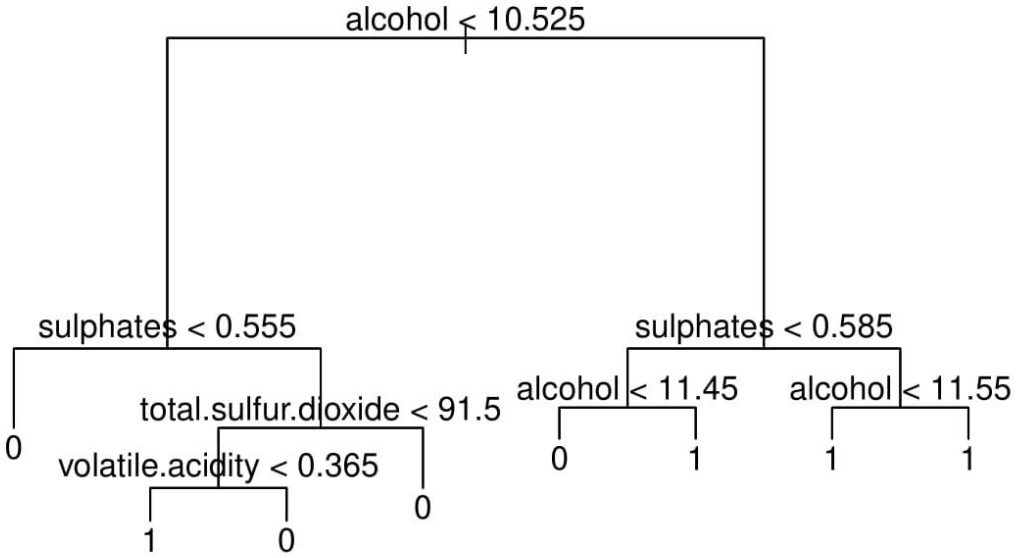
Accuracy: **69.4%**

Best at K = 15

Even worse than QDA

Indicates linear relationship

Classification Tree

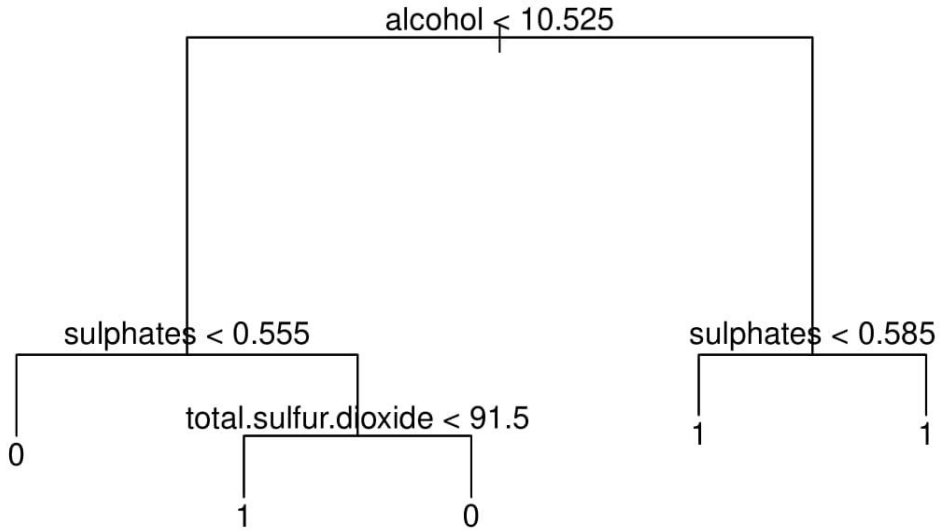


Accuracy: **70.6%**

Important variables line up with previous observations

Not best accuracy, but good

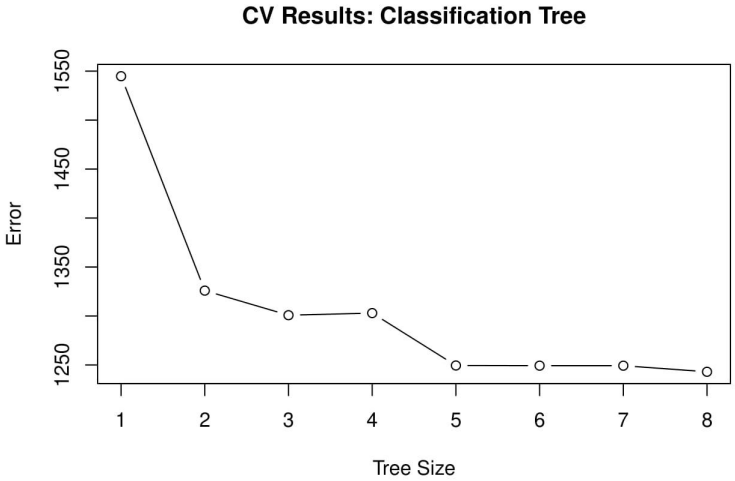
Pruned Classification Tree



Accuracy: **65.6%**

Best at full complexity

5 is comparable



Support Vector Machines

SVC:

- Cost = 0.2976
- Accuracy = **72.9%**

SVM (Radial):

- Cost = 0.8858668
- Gamma = 0.4641589
- Accuracy = **75.6%**

SVM (Polynomial):

- Cost = 1000
- Degree = 2.371374
- Accuracy = **61.0%**

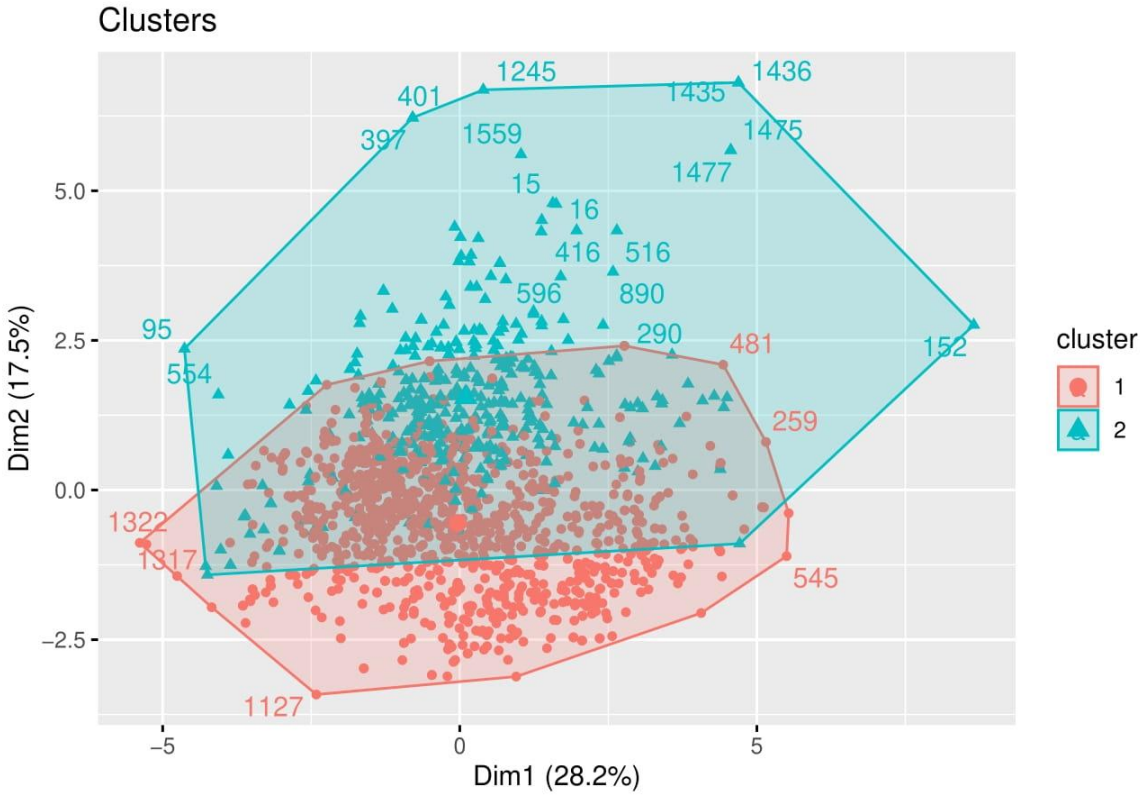
Radial kernel for SVM performed best

Better than all previous models

Polynomial kernel for SVM performed worst

Worse than all previous models

K-Means Clustering



Cluster 1:

- Good: 60.0%
- Bad: 40.0%

Cluster 2:

- Good: 35.2%
- Bad: 64.8%

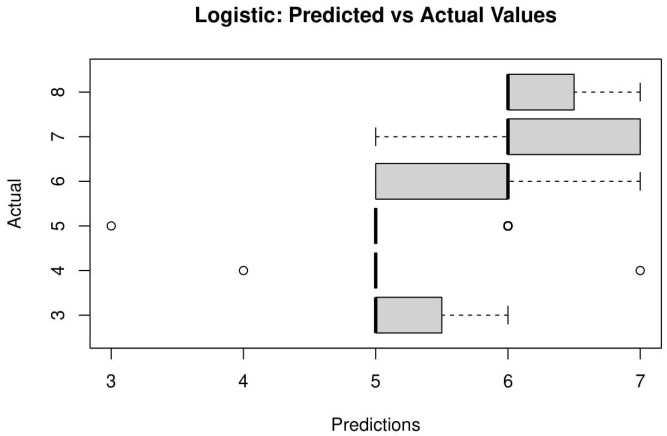
Multiclass Classification

Logistic Regression

```
## Coefficients:
##   (Intercept) fixed.acidity volatile.acidity citric.acid residual.sugar
## 4   12.268849    -0.3811757        -4.721874   -1.921828     -0.1547333
## 5   -2.818620    -0.2705864        -7.284544   -1.592367     -0.3050784
## 6    0.430362    -0.1574736        -9.915406   -3.143702     -0.3385994
## 7   12.841774    -0.2619752       -11.679394   -2.544609     -0.1568228
## 8   10.752679    -0.6367343        -7.344396    0.349823     -0.3870262
##    chlorides free.sulfur.dioxide total.sulfur.dioxide   density        pH
## 4  -8.620334        -0.046405129          0.034284551 -5.151393  -3.646719
## 5 -11.680105        -0.007591673          0.044732170 21.836353  -5.366425
## 6 -13.582024         0.005062219          0.030420956 11.283595  -5.891750
## 7 -23.495671         0.012391775          0.014603423 -4.315503  -7.123874
## 8 -49.188135         0.011868527          0.007273219  7.980098 -13.088652
##   sulphates   alcohol
## 4  3.518983 1.2671826
## 5  2.977010 0.9748738
## 6  4.951917 1.8788797
## 7  7.685613 2.4666715
## 8  9.192534 3.2262319
```

Accuracy: **63.1%**

RMSE: **0.6630**



Linear Discriminant Analysis

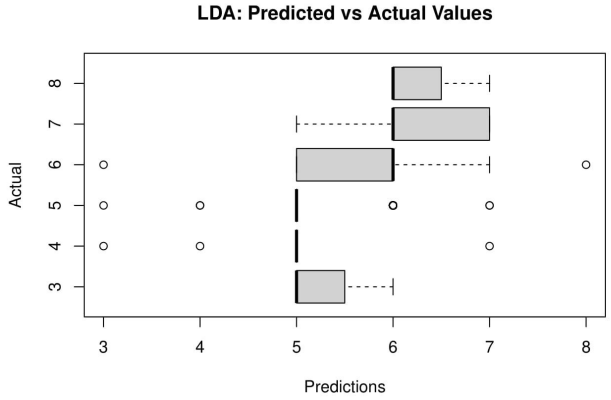
```
## Prior probabilities of groups:
##           3           4           5           6           7           8
## 0.006255585 0.034852547 0.414655943 0.395889187 0.134941912 0.013404826
##
## Group means:
##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 3      7.785714        0.9678571   0.1114286       2.764286 0.11371429
## 4      7.538462        0.6855128   0.1538462       2.579487 0.09412821
## 5      8.170905        0.5716487   0.2458405       2.560022 0.09334698
## 6      8.392551        0.4939052   0.2788262       2.416027 0.08578555
## 7      8.586755        0.4188079   0.3512583       2.687748 0.07574834
## 8      8.546667        0.4373333   0.3833333       2.560000 0.06733333
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## 3            9.285714             24.85714 0.9968571 3.422857 0.5314286
## 4           12.153846             36.25641 0.9962192 3.389231 0.6107692
## 5           17.150862             55.19289 0.9971491 3.305065 0.6243103
## 6           15.393905             39.67720 0.9965642 3.316117 0.6724831
## 7           13.930464             33.86093 0.9959053 3.303179 0.7305960
## 8           13.466667             35.00000 0.9950087 3.271333 0.7740000
##     alcohol
## 3 10.035714
## 4 10.237179
## 5  9.891272
## 6 10.685290
## 7 11.456402
## 8 12.166667
```

Accuracy: **61.7%**

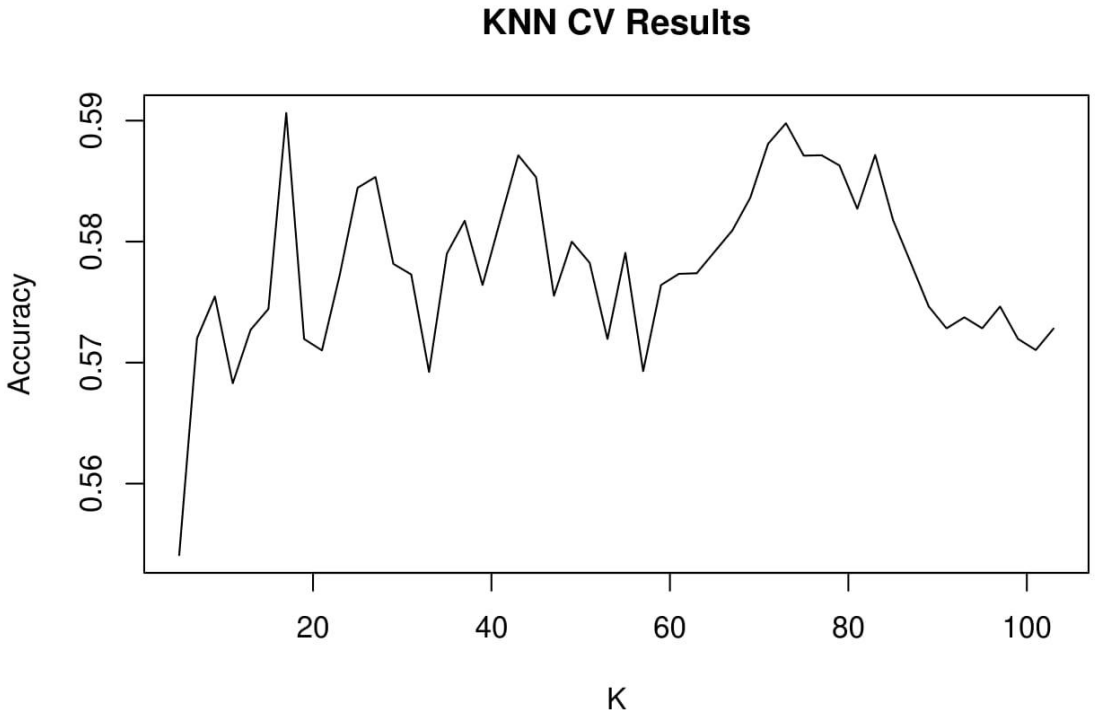
RMSE: **0.6997**

Worse than logistic

Similar plots of predictions



K-Nearest Neighbors

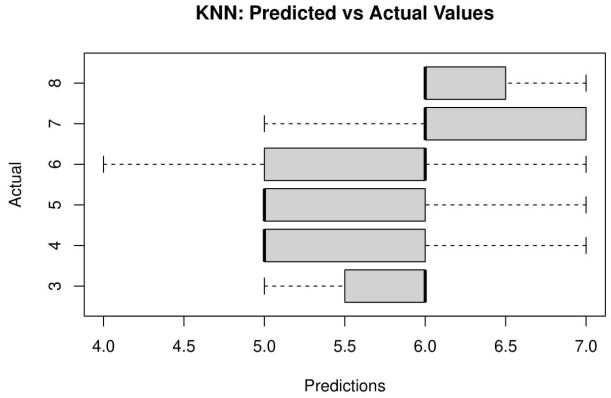


Accuracy: **58.3%**

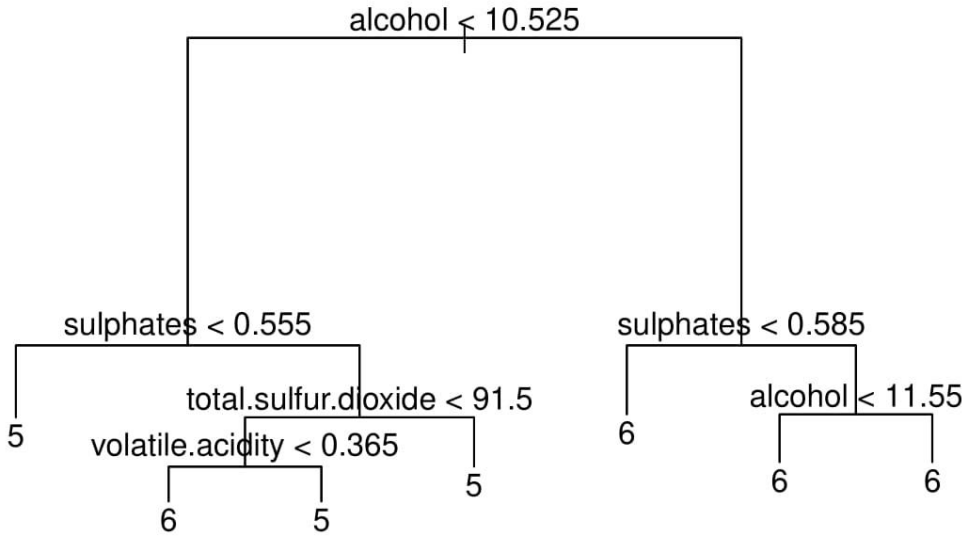
RMSE: **0.7486**

Worse both others

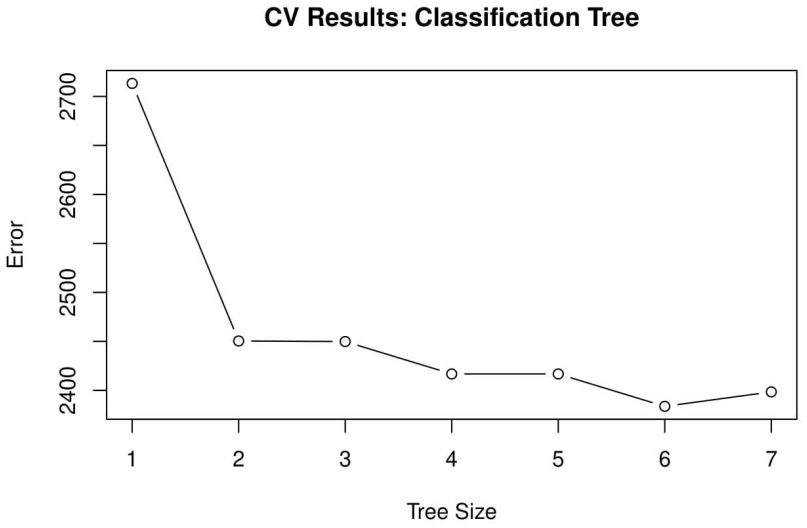
Further evidence of linearity



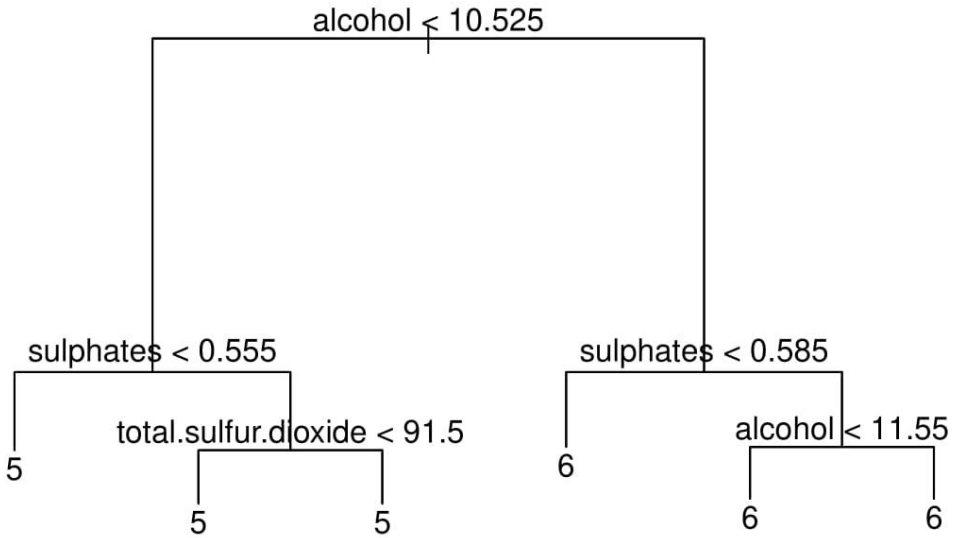
Classification Tree



Best tree at size 6



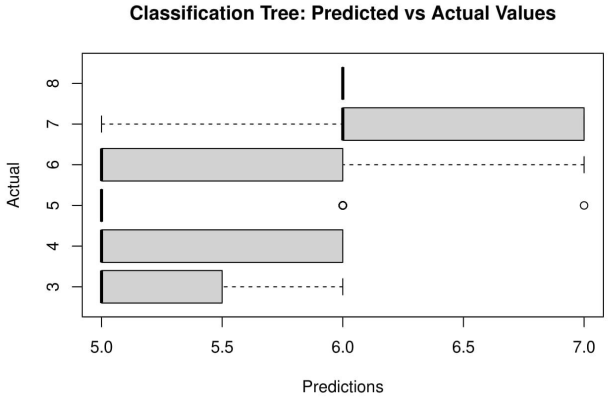
Pruned Classification Tree



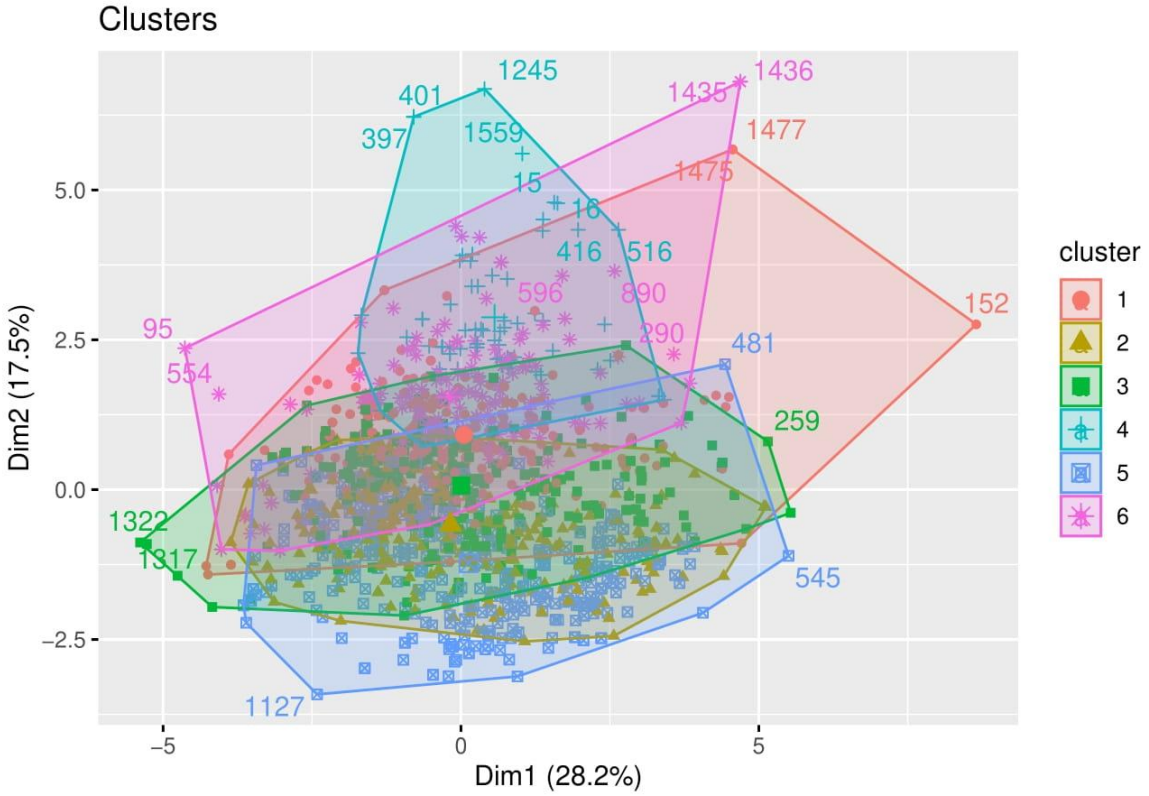
Accuracy: **56.0%**

RMSE: **0.7624**

Worst so far



K-Means Clustering



Cluster 1: overall representative

Cluster 2: slightly heavy on high ratings

Cluster 3: slightly heavy on high ratings

Cluster 4: dominated by 5's (87%)

Cluster 5: more evened out (20% 7's)

Cluster 6: mostly 5's (64%)

Regression

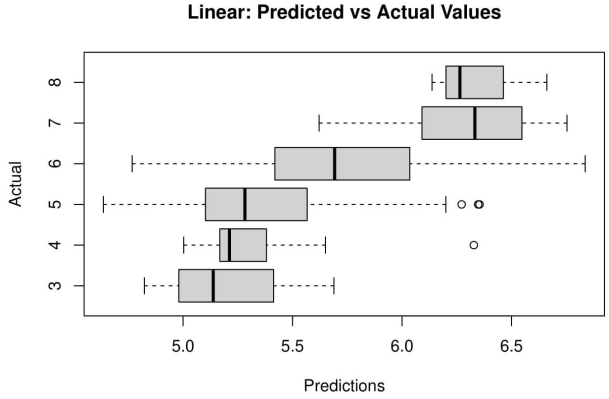
Linear Regression

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

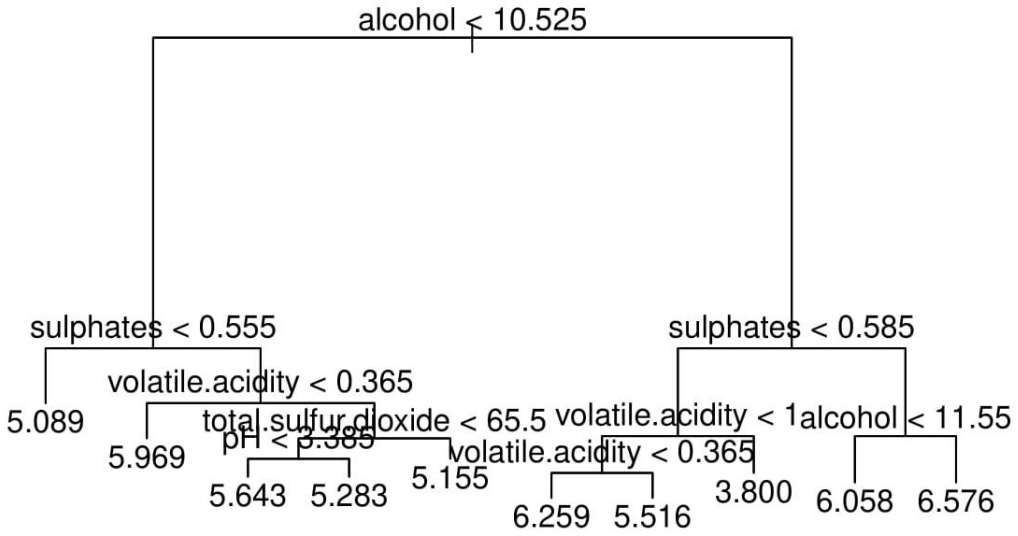
RMSE: **0.6063**

Very good, the previous best is at 0.6630 (logistic)

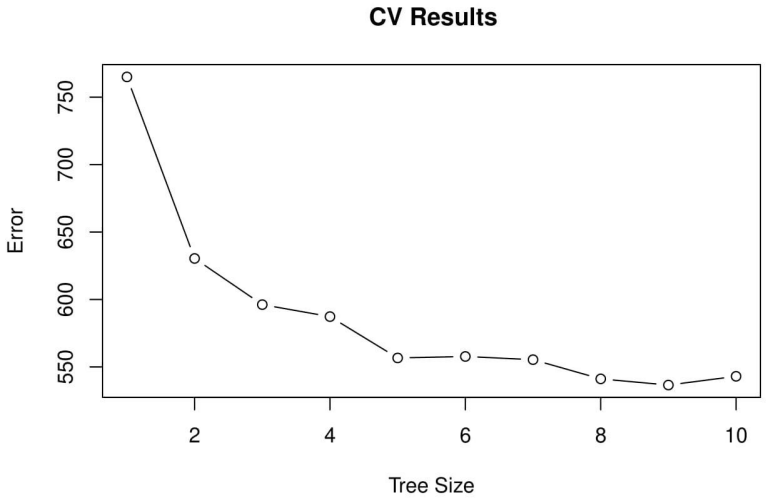
Positive trend in below plot



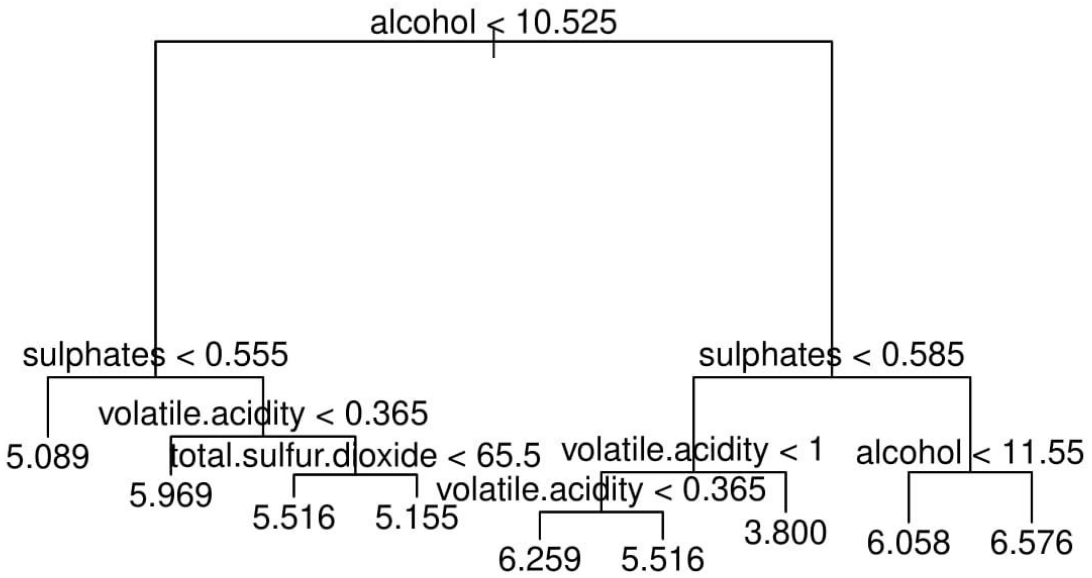
Regression Tree



Best tree at size 9



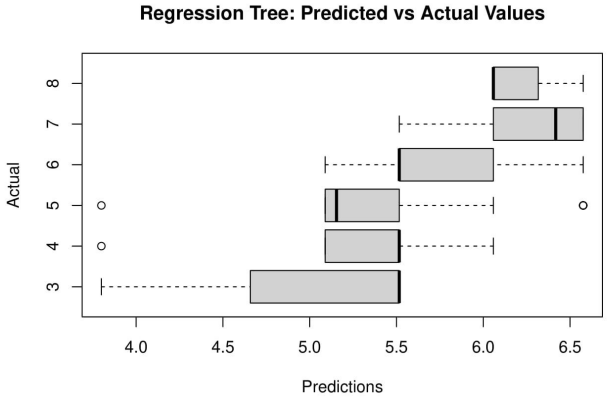
Pruned Regression Tree



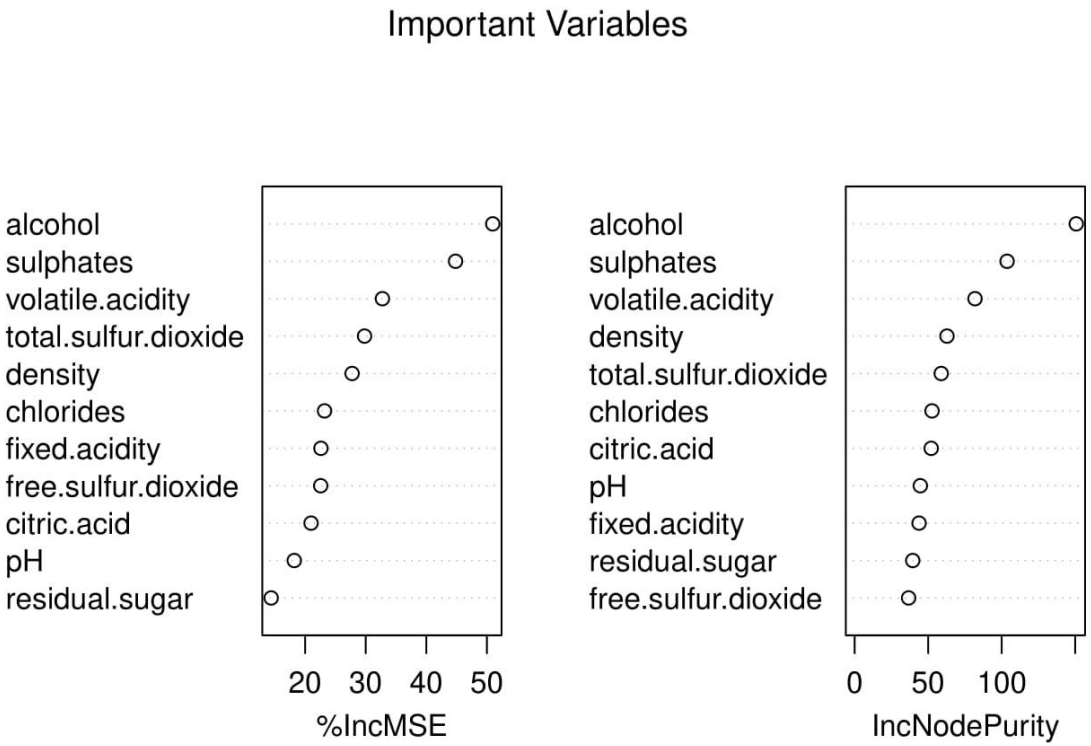
RMSE: **0.6342**

Not as good as linear regression, but still better than all others

Some trend below



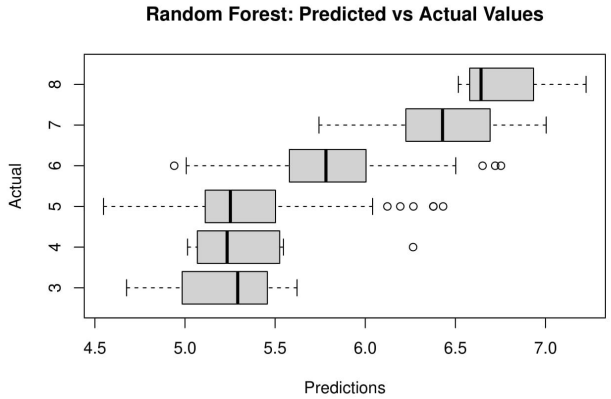
Random Forest



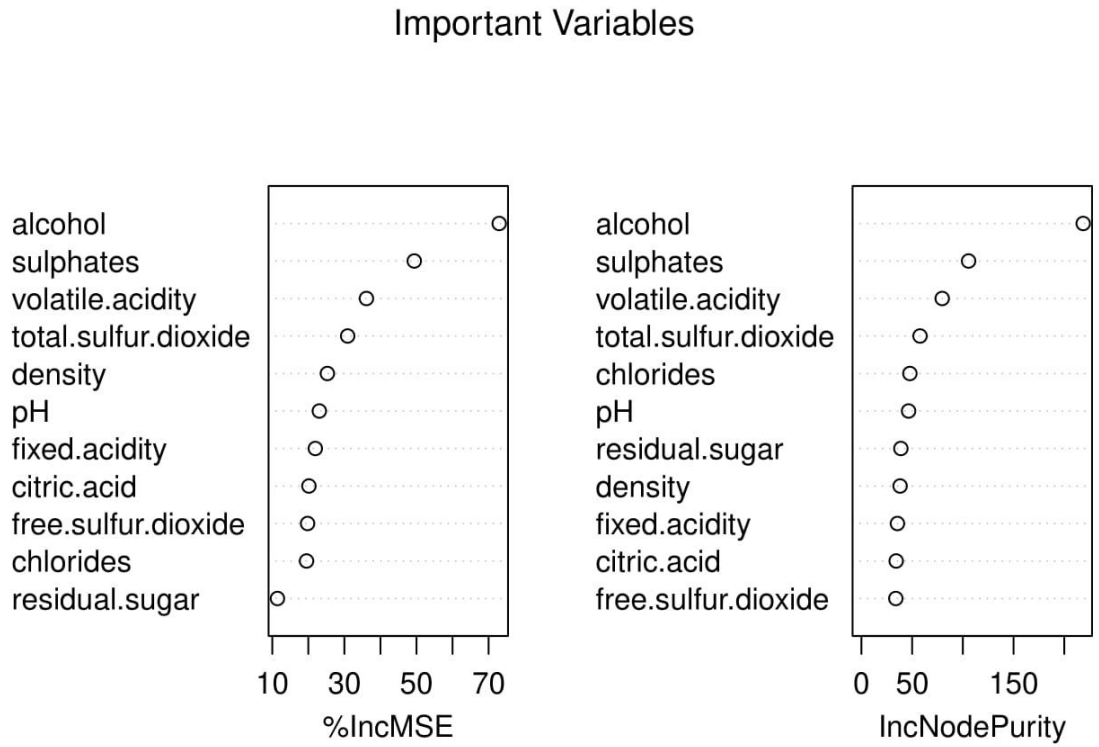
RMSE: **0.5510**

Great improvement, previous best at 0.6063 (linear)

Good trend in below plot



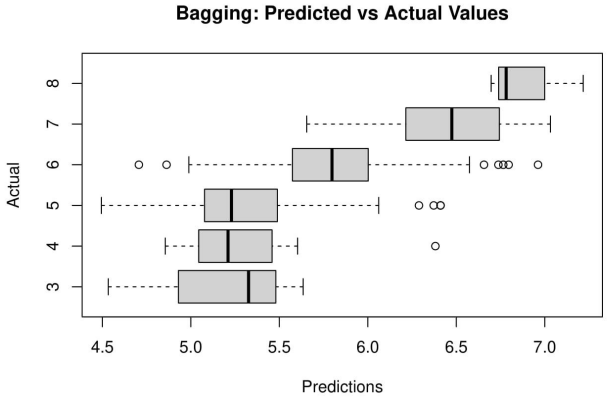
Bagging



RMSE: **0.5521**

Very similar to random forest

Same top variables



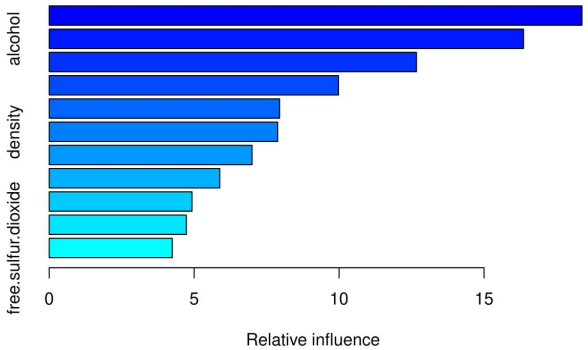
Boosting

```
##                                   var   rel.inf
## volatile.acidity       volatile.acidity 18.379366
## alcohol                         alcohol 16.361812
## sulphates                     sulphates 12.668108
## chlorides                     chlorides  9.978477
## total.sulfur.dioxide total.sulfur.dioxide  7.948213
## density                         density  7.882166
## citric.acid                 citric.acid  6.996979
## fixed.acidity             fixed.acidity  5.884200
## pH                                   pH  4.925827
## residual.sugar           residual.sugar  4.731237
## free.sulfur.dioxide   free.sulfur.dioxide  4.243616
```

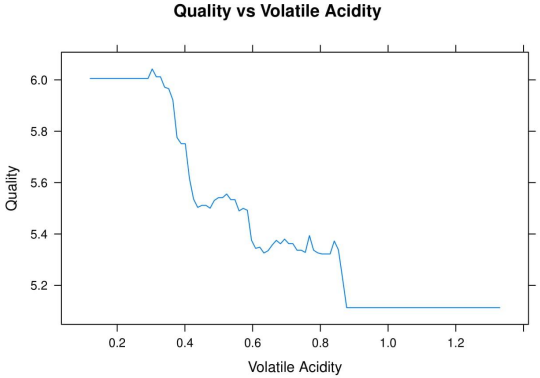
RMSE: **0.4942**

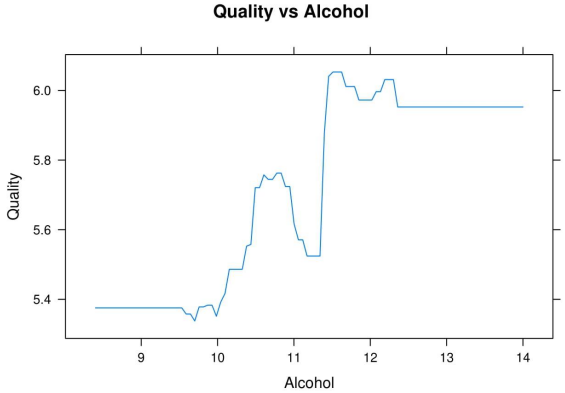
Another great improvement from 0.5510 (random forest)

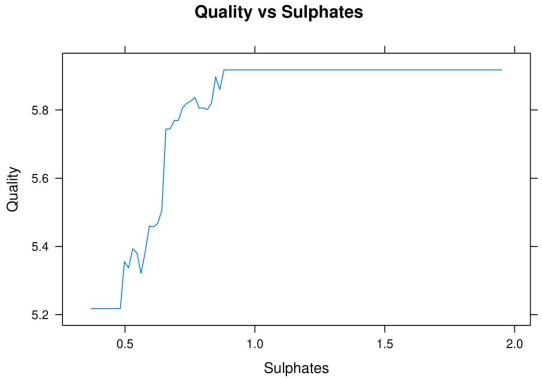
Similar important values

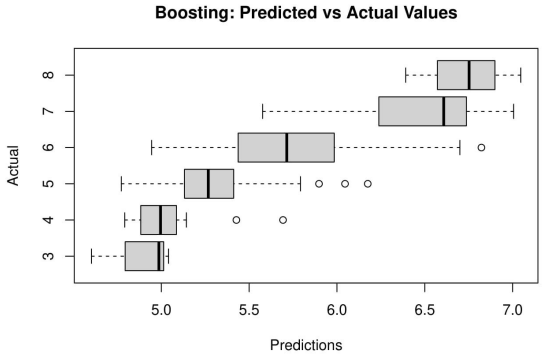


Boosting Cont'd

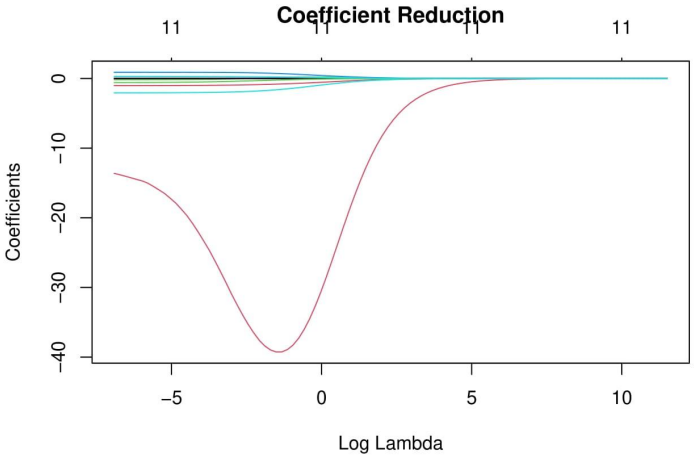


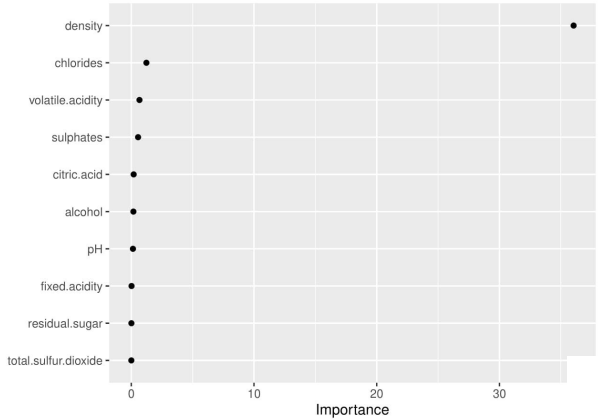






Ridge



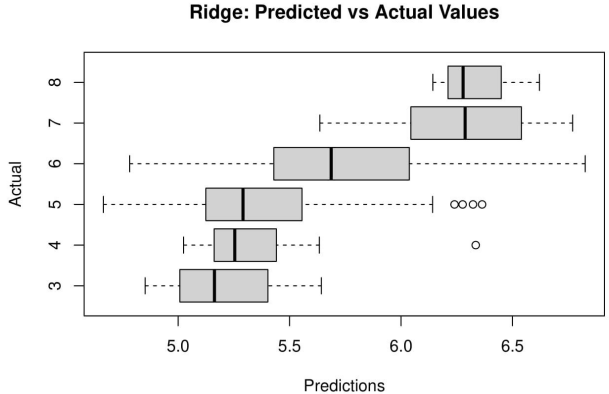


RMSE: **0.6099**

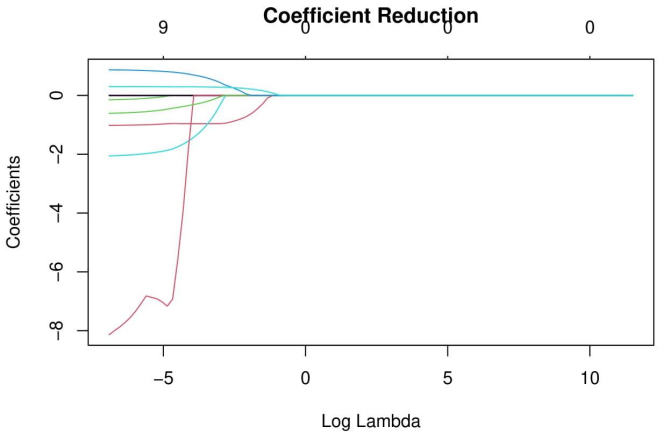
Lambda = 0.0413

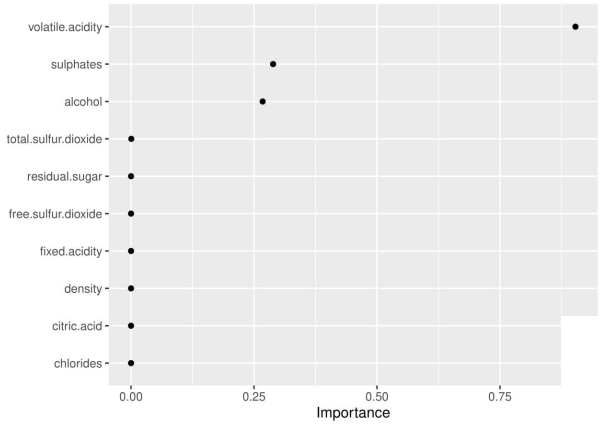
Very similar to linear model

New top variable (density)



Lasso



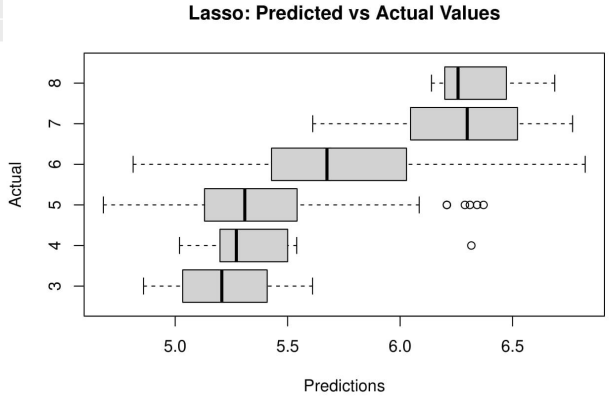


RMSE: **0.6109**

Lambda = 0.0093

Very similar to ridge/linear

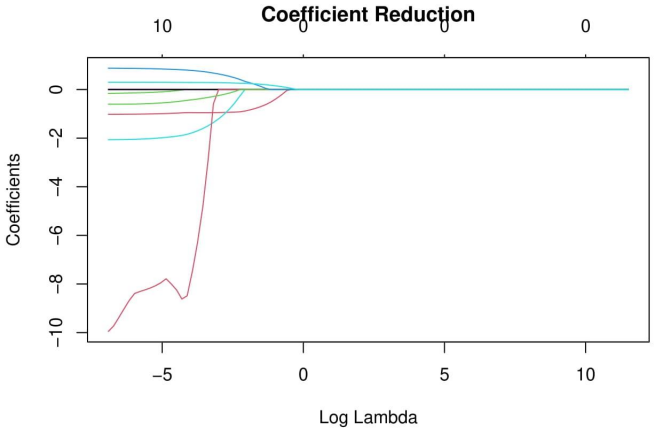
More consistent importance

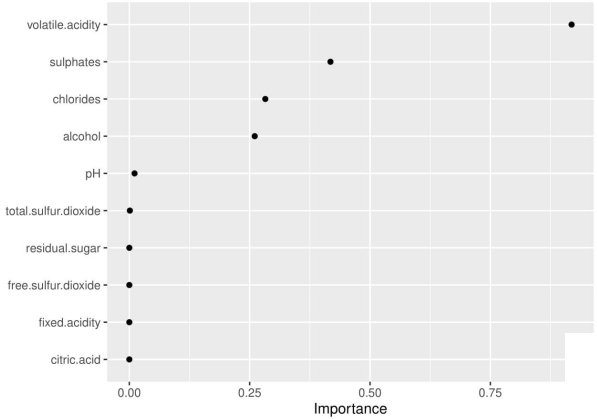


Variables with coefficients reduced to zero:

Fixed acidity, citric acid, residual sugar

Elastic Net



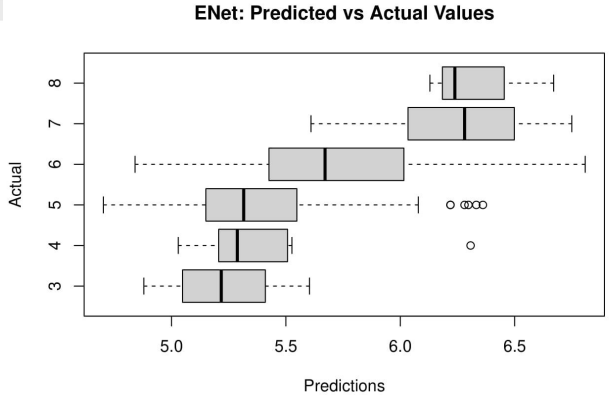


RMSE: **0.6113**

Lambda = 0.0236

Very similar to others

Same variables removed



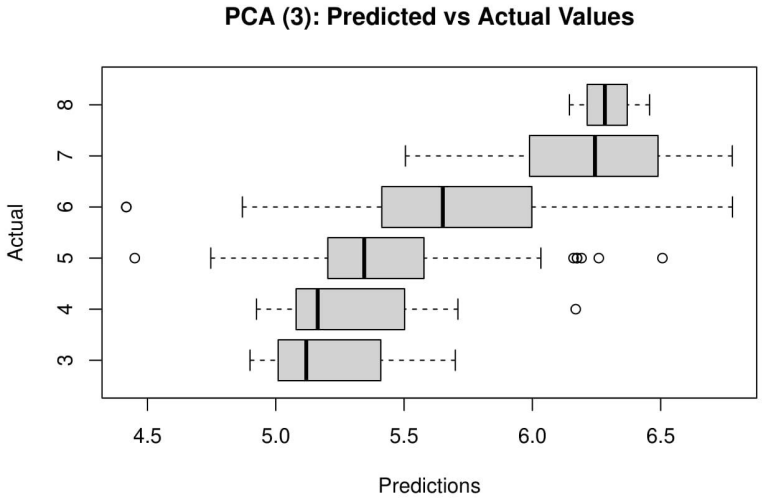
Partial Components Analysis

```
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           0.8268   0.8240   0.7565   0.6850   0.6827   0.6795   0.6801
## adjCV        0.8268   0.8239   0.7559   0.6847   0.6825   0.6793   0.6800
##        7 comps  8 comps  9 comps  10 comps  11 comps
## CV      0.6755   0.6729   0.6684    0.6686    0.6690
## adjCV   0.6753   0.6727   0.6681    0.6683    0.6686
##
## TRAINING: % variance explained
##          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X        28.3425    45.63    59.74    71.23    80.12    85.89    91.10    94.86
## quality   0.8885    16.59    31.74    32.33    33.02    33.08    34.01    34.67
##          9 comps  10 comps  11 comps
## X          97.90     99.48    100.00
## quality    35.63     35.86     35.98
```

Partial Components Analysis Cont'd

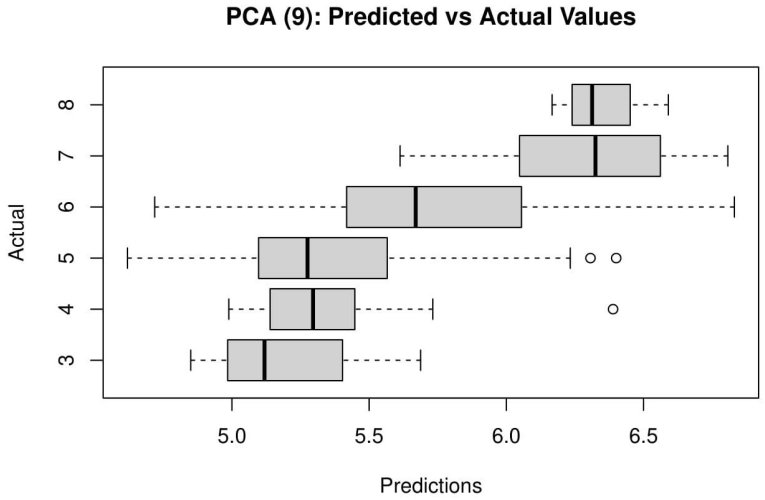
Three Components:

RMSE: **0.6327**

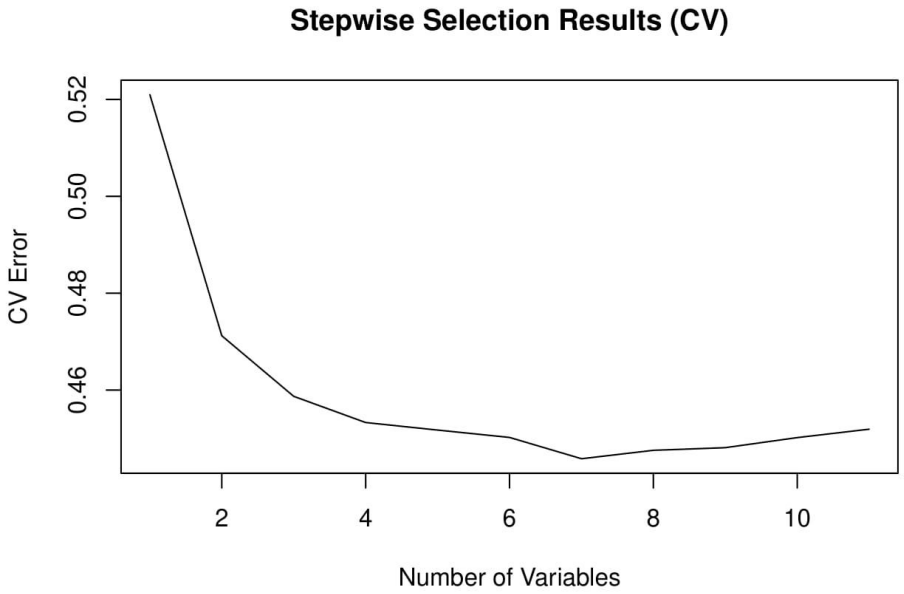


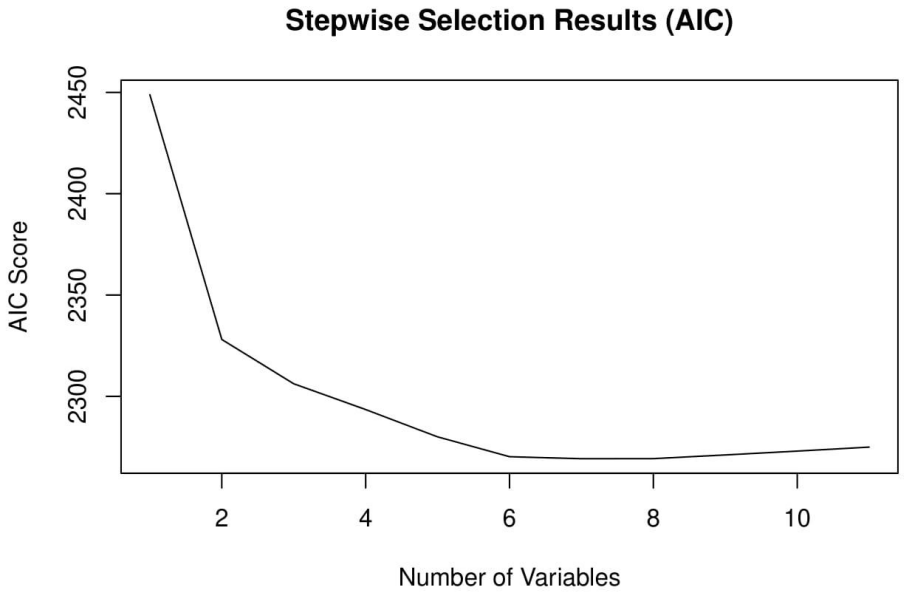
Nine Components:

RMSE: **0.6117**



Forward Stepwise Selection





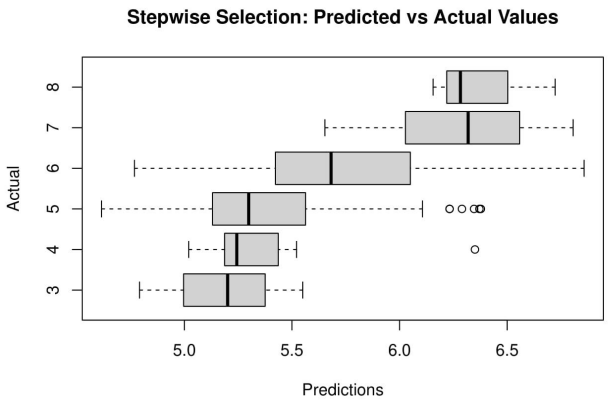
Forward Stepwise Selection Cont'd

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.7834613  0.5546792   8.624  < 2e-16 ***
## alcohol              0.3093595  0.0207331  14.921  < 2e-16 ***
## volatile.acidity    -1.0816790  0.1414013  -7.650 4.35e-14 ***
## sulphates            0.8777227  0.1369913   6.407 2.19e-10 ***
## total.sulfur.dioxide -0.0026454  0.0006513  -4.062 5.21e-05 ***
## chlorides            -2.0273731  0.4760480  -4.259 2.23e-05 ***
## pH                   -0.6080616  0.1602919  -3.793 0.000157 ***
## residual.sugar        0.0024361  0.0150240   0.162 0.871217
## citric.acid          -0.2437766  0.1494258  -1.631 0.103085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

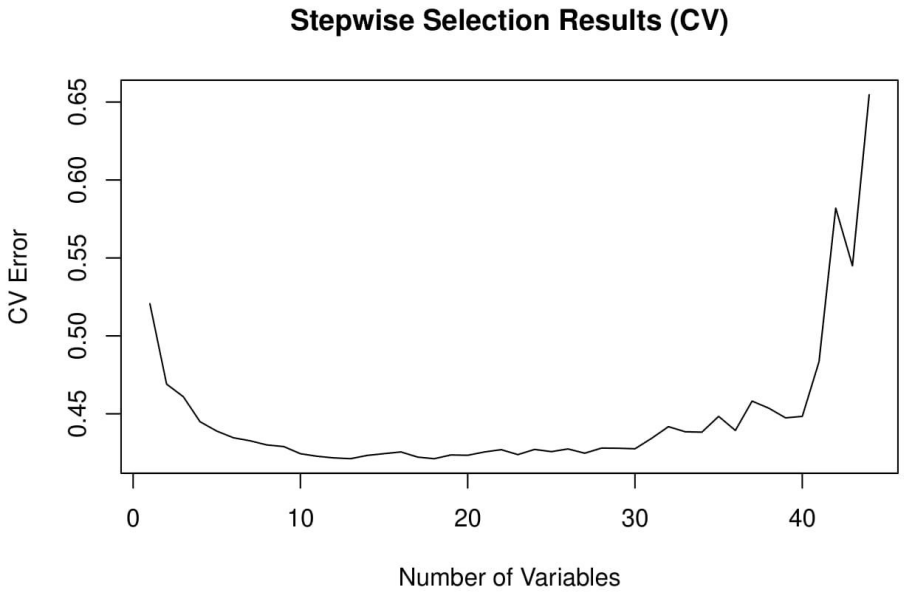
RMSE: **0.6123**

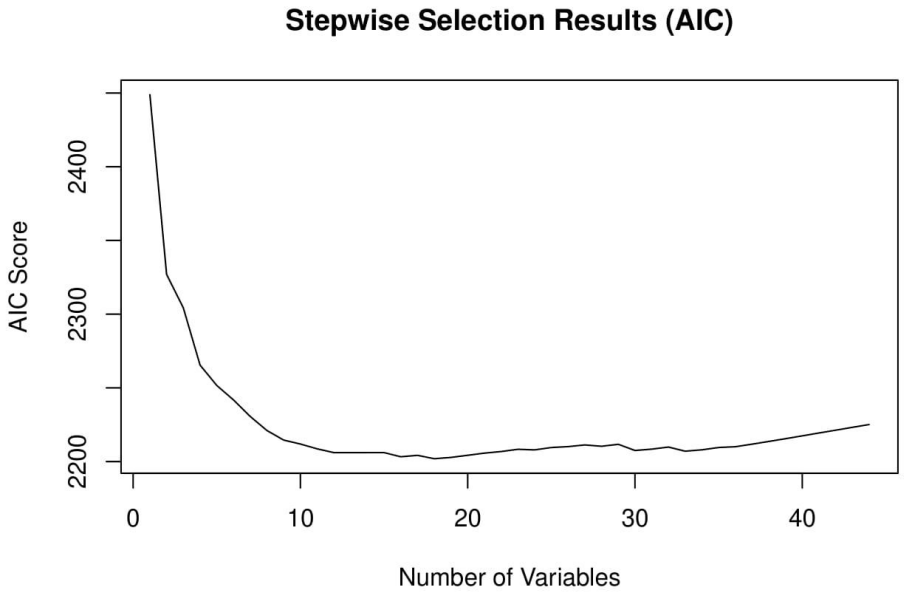
Similar, but only seven vars

Similar plot below



Forward Stepwise Selection (Nonlinear)



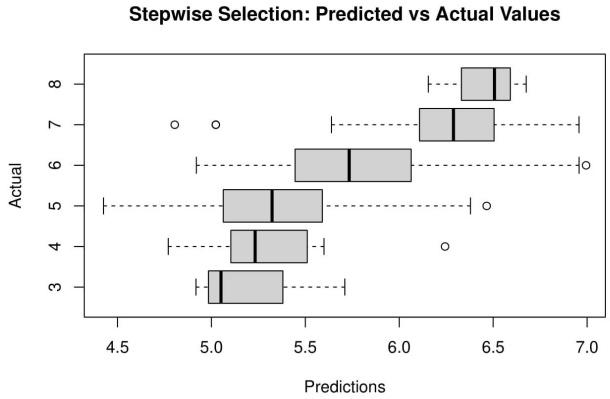


Forward Stepwise Selection (Nonlinear) Cont'd

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.876e+03  2.726e+03   2.523 0.011789 *
## alcohol                 2.598e-01  2.991e-02   8.687  < 2e-16 ***
## volatile.acidity        1.996e-01  4.750e-01   0.420 0.674401
## sulphates               1.020e+01  1.863e+00   5.475 5.41e-08 ***
## sulphates2             -8.345e+00  1.933e+00  -4.318 1.72e-05 ***
## pH4                     6.581e-03  8.655e-03   0.760 0.447221
## fixed.acidity4         -1.521e-03  4.986e-04  -3.050 0.002345 **
## sulphates3              2.027e+00  5.926e-01   3.421 0.000647 ***
## density                -1.034e+04  4.102e+03  -2.520 0.011883 *
## chlorides              -1.510e+00  5.171e-01  -2.921 0.003564 **
## total.sulfur.dioxide   -3.272e-03  1.968e-03  -1.662 0.096771 .
## density3                3.450e+03  1.375e+03   2.509 0.012264 *
## fixed.acidity2         -8.786e-01  2.914e-01  -3.015 0.002632 **
## citric.acid3            1.163e-01  3.609e-01   0.322 0.747280
## fixed.acidity           5.658e+00  1.848e+00   3.062 0.002251 **
## volatile.acidity2      -7.455e-01  3.720e-01  -2.004 0.045301 *
## fixed.acidity3          5.996e-02  1.994e-02   3.006 0.002705 **
## pH                     -1.652e+00  1.317e+00  -1.254 0.210154
## free.sulfur.dioxide     2.503e-03  2.690e-03   0.930 0.352388
## total.sulfur.dioxide2   3.027e-06  1.108e-05   0.273 0.784682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

RMSE: **0.6247**

18 variables



All Models

Boosting: **0.4942**

Random Forest: **0.5510**

Bagging: **0.5521**

Linear: **0.6063**

Ridge: **0.6099**

Lasso: **0.6109**

ENet: **0.6113**

PCA(9): **0.6117**

Stepwise: **0.6123**

Nonlinear Stepwise: **0.6247**

PCA(3): **0.6327**

Regression Tree: **0.6432**

Logistic: **0.6630**

LDA: **0.6997**

KNN: **0.7486**

Classification Tree: **0.7624**