Wine...

Business Objective

- This wine, Vinho Verde, is very popular in Portugal, especially in the summer.
- From this data, our goal is to predict wine quality (0-10); this wine was rated by certain wine experts.
- Useful for the producers so they can consistently see what factors play into their low/high quality wine

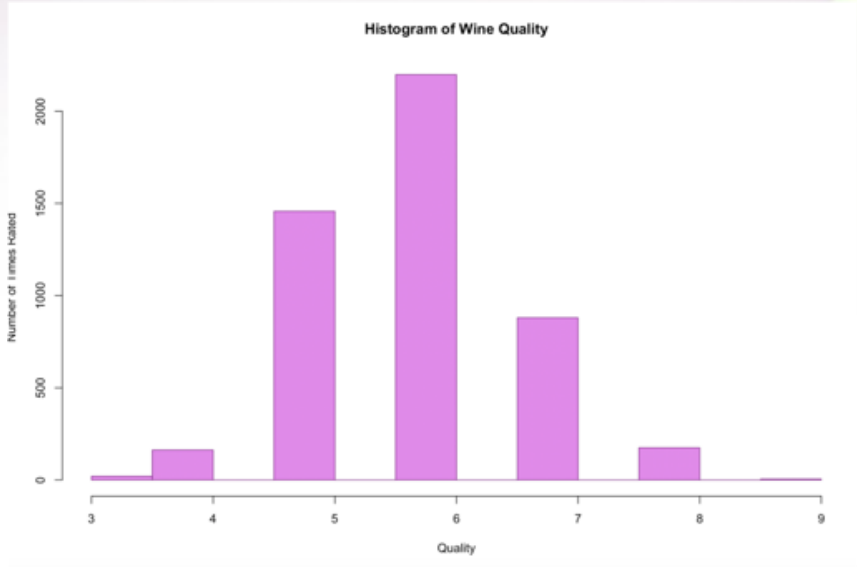
Overview of the Data

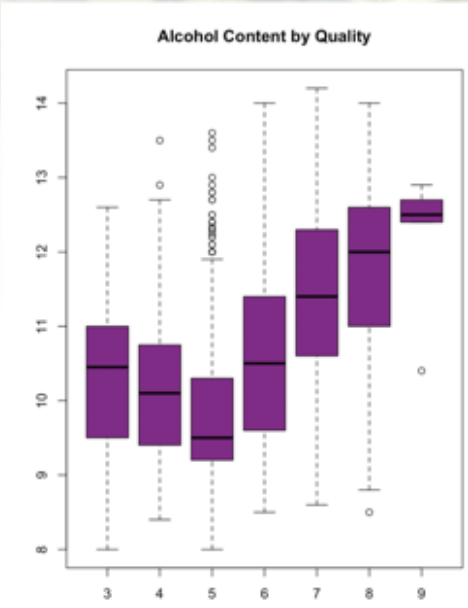
- *White Wine Data*: 4898 observations x 12 variables

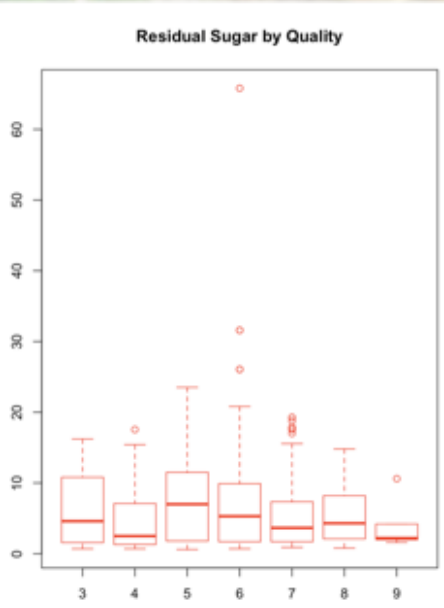
Predictors (not all of them)

quality	Outcome Variable	Scale from 0-10
residual.sugar	Measure of sweetness	*g/dm^3*
citric.acid	Natural preservative to add sour taste	*g/dm^3*
alcohol	Alcohol	*By Volume*
pH	Measurement of strength of acids present	*Any pH less than 7 is acidic, while any pH greater than 7 is basic.*

Exploratory Data Analysis

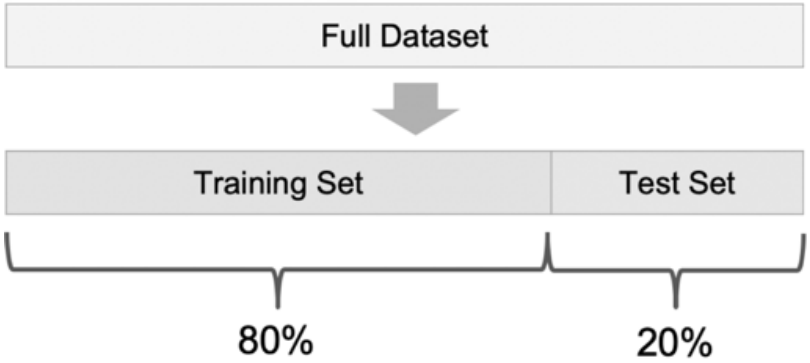






Model Building Process

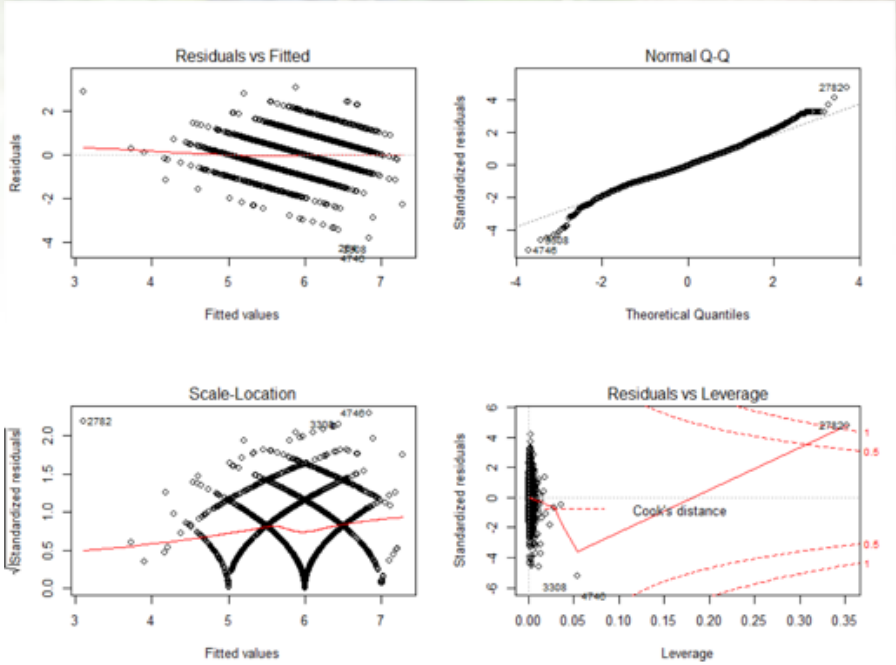
- **Linear Regression**
  - Feature Selection using Forward Selection
  - Validation done by Hold Out Method
- **KNN**
  - Validation done by K-Fold Cross Validation with K = 5
- **Logistic Regression**
- **Random Forest**



Linear Regression - Testing Assumptions

Test for:

- Linearity of predictor-response relationship
- Independence of errors (lack of multicollinearity)
- Normal distribution of errors
- Equal variance of errors (homoscedasticity)
    - Standardized residuals plot does not indicate homoscedasticity

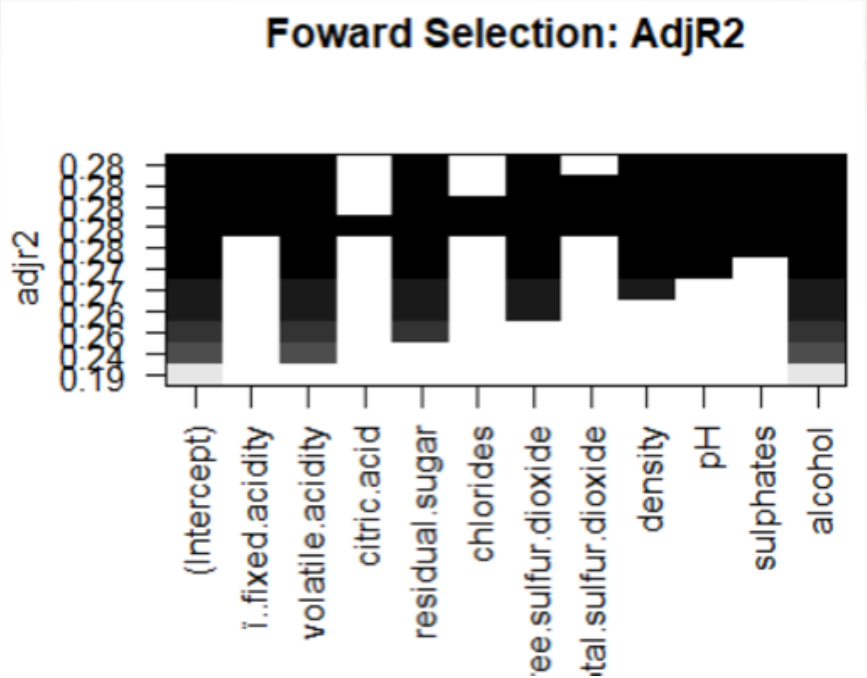


Linear Regression - Initial Model Comparison

```
> qpred = lm(quality~ï..fixed.acidity+volatile.acidity+citric.acid+r
esidual.sugar-chlorides+free.sulfur.dioxide+total.sulfur.dioxide+den
sity+pH+sulphates+alcohol,data=whitewine)
> summary(qpred)

Call:
lm(formula = quality ~ ï..fixed.acidity + volatile.acidity +
    citric.acid + residual.sugar - chlorides + free.sulfur.dioxide +

    total.sulfur.dioxide + density + pH + sulphates + alcohol,
    data = whitewine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8321 -0.4917 -0.0384  0.4670  3.1151

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.515e+02  1.856e+01   8.164 4.08e-16 ***
ï..fixed.acidity      6.713e-02  2.056e-02   3.265   0.0011 **
volatile.acidity     -1.868e+00  1.132e-01 -16.501  < 2e-16 ***
citric.acid           1.746e-02  9.521e-02   0.183   0.8545
residual.sugar        8.215e-02  7.379e-03  11.134  < 2e-16 ***
free.sulfur.dioxide   3.722e-03  8.437e-04   4.411 1.05e-05 ***
total.sulfur.dioxide -2.874e-04  3.780e-04  -0.760   0.4471
density              -1.517e+02  1.882e+01  -8.062 9.33e-16 ***
pH                    6.942e-01  1.039e-01   6.678 2.69e-11 ***
sulphates             6.332e-01  1.003e-01   6.312 2.99e-10 ***
alcohol               1.936e-01  2.422e-02   7.993 1.63e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7513 on 4887 degrees of freedom
Multiple R-squared:  0.2818,    Adjusted R-squared:  0.2804
F-statistic: 191.8 on 10 and 4887 DF,  p-value: < 2.2e-16
```

```
> confint(qpred)
                          2.5 %        97.5 %
(Intercept)          1.151569e+02  1.879390e+02
ï..fixed.acidity     2.681768e-02  1.074504e-01
volatile.acidity    -2.090247e+00 -1.646305e+00
citric.acid         -1.692025e-01  2.041181e-01
residual.sugar       6.768826e-02  9.662036e-02
free.sulfur.dioxide  2.067510e-03  5.375644e-03
total.sulfur.dioxide -1.028458e-03  4.536919e-04
density             -1.885843e+02 -1.148109e+02
pH                   4.903916e-01  8.979404e-01
sulphates            4.365255e-01  8.298206e-01
alcohol              1.460950e-01  2.410529e-01
> par(mfrow=c(2,2))
> plot(qpred)
> vif(qpred)
    ï..fixed.acidity     volatile.acidity          citric.acid
            2.612817             1.129962             1.151897
      residual.sugar  free.sulfur.dioxide total.sulfur.dioxide
           12.152487             1.786346             2.239028
             density                   pH            sulphates
           27.475237             2.137241             1.136952
             alcohol
            7.706337
```

Linear Regression - Feature Selection

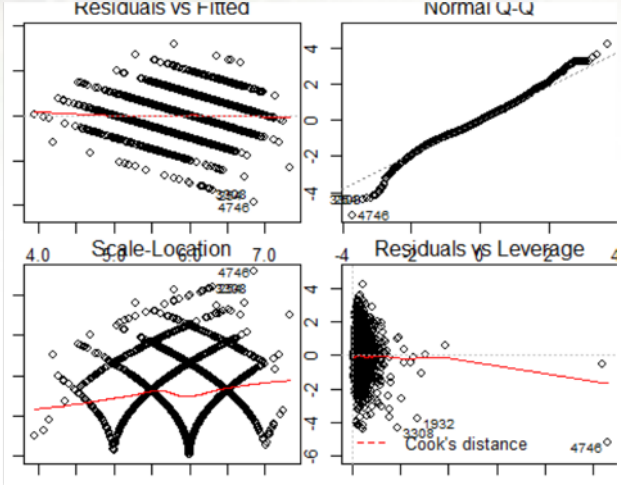



```
> plot(model_fwd, scale="adjr2", main="Foward Selection: AdjR2")
> model_fwd_summary = summary(model_fwd)
> which.max(model_fwd_summary$adjr2)
[1] 8
> summary(model_fwd)$which[8,]
          (Intercept)      ï..fixed.acidity      volatile.acidity
                 TRUE                  TRUE                  TRUE
          citric.acid        residual.sugar             chlorides
                FALSE                  TRUE                 FALSE
 free.sulfur.dioxide total.sulfur.dioxide               density
                TRUE                 FALSE                  TRUE
                  pH             sulphates               alcohol
                TRUE                  TRUE                  TRUE
```

```
> vif(qpred)
     ï..fixed.acidity      volatile.acidity           citric.acid
             2.612817              1.129962              1.151897
       residual.sugar  free.sulfur.dioxide total.sulfur.dioxide
            12.152487              1.786346              2.239028
             density                    pH             sulphates
            27.475237              2.137241              1.136952
             alcohol
             7.706337
```

Linear Regression - New Model

```
> qpredadj2 = lm(quality~ï..fixed.acidity+volatile.acidity+residual.
sugar+free.sulfur.dioxide+pH+sulphates+alcohol, data=whitewine)
> vif(qpredadj2)
   ï..fixed.acidity    volatile.acidity      residual.sugar
           1.242182            1.030666            1.375945
free.sulfur.dioxide                  pH           sulphates
           1.147581            1.295197            1.033553
             alcohol
           1.303358
```



Linear Regression - Performance

Test MSE

```
> mse
[1] 0.5689397
```

```
Call:
lm(formula = quality ~ ï..fixed.acidity + volatile.acidity +
    residual.sugar + free.sulfur.dioxide + pH + sulphates + alcohol

    data = whitewine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8605 -0.4982 -0.0372  0.4599  3.2028

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.892072   0.337214   5.611 2.12e-08 ***
ï..fixed.acidity     -0.055632   0.014278  -3.896 9.90e-05 ***
volatile.acidity     -2.034356   0.108888 -18.683  < 2e-16 ***
residual.sugar        0.025225   0.002500  10.089  < 2e-16 ***
free.sulfur.dioxide   0.003551   0.000681   5.215 1.92e-07 ***
pH                    0.153290   0.081479   1.881     0.06 .
sulphates             0.383213   0.096303   3.979 7.01e-05 ***
alcohol               0.377600   0.010029  37.650  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7565 on 4890 degrees of freedom
Multiple R-squared:  0.2714,    Adjusted R-squared:  0.2703
F-statistic: 260.2 on 7 and 4890 DF,  p-value: < 2.2e-16

>
```

K-Nearest Neighbor (KNN)

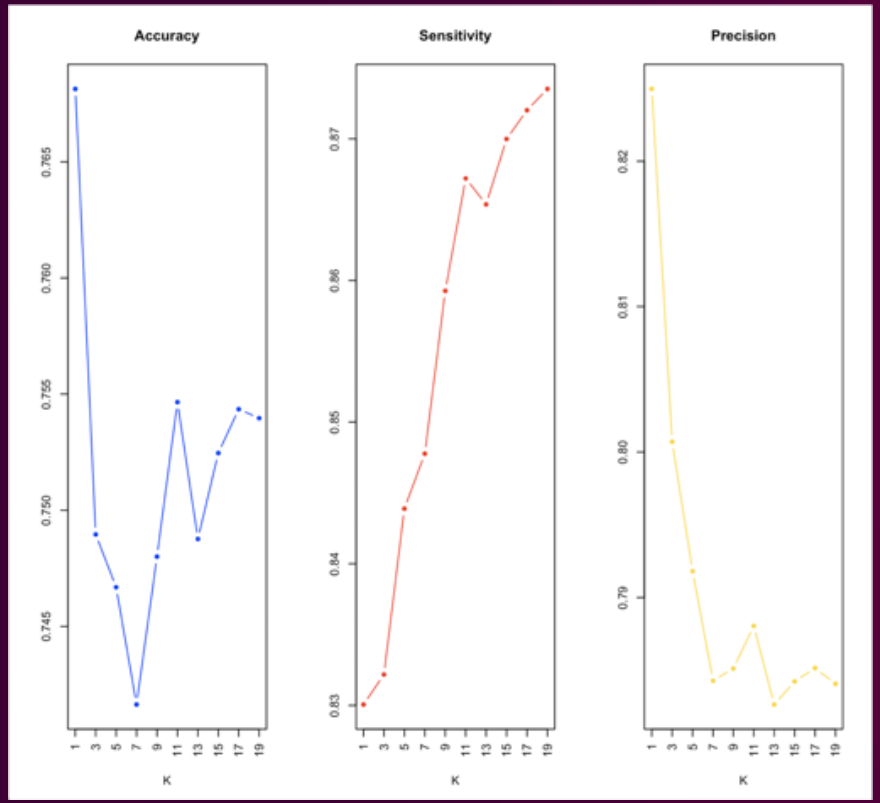
- Data splitting: training and testing
- Normalize continuous variables
- Classify `quality` as *high* and *low*

```
    quality
Min.   :3.000
1st Qu.:5.000
Median :6.000
Mean   :5.878
3rd Qu.:6.000
Max.   :9.000
```

```
high  low
3258 1640
```

```
     high       low
0.6651695 0.3348305
```

- 5-fold cross validation
  - With parameter tuning to avoid overfitting



	K	accuracy	sensitivity	precision
1	1	0.7681403	0.8300638	0.8249668
2	3	0.7489585	0.8321766	0.8007035
3	5	0.7466856	0.8438826	0.7918093
4	7	0.7416392	0.8477614	0.7842883
5	9	0.7480035	0.8592628	0.7851236
6	11	0.7546537	0.8672012	0.7880528
7	13	0.7487621	0.8653649	0.7826546
8	15	0.7524584	0.8699866	0.7842388
9	17	0.7543504	0.8720169	0.7851576
10	19	0.7539609	0.8735198	0.7840778

- K = 3
  - Confusion matrix:

```
      pred.class
       high low
high    537  97
low     139 207
```

  - Accuracy, sensitivity & precision (for *high*):

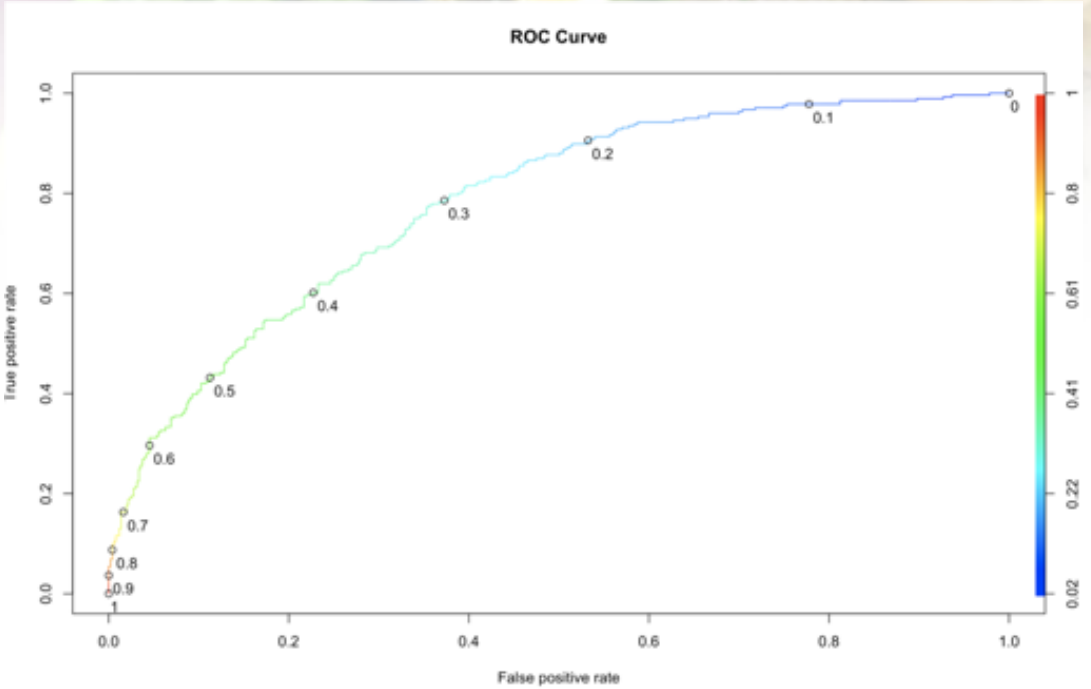
```
         acc sens.high prec.high
1 0.7591837 0.8470032 0.7943787
```

→ Better!

Logistic Regression

```
Call:
glm(formula = as.factor(quality) ~ volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6546  -0.8134  -0.4819   0.9462   2.8910

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.740e+02  2.926e+01 -23.032  < 2e-16 ***
volatile.acidity     6.279e+00  4.349e-01  14.440  < 2e-16 ***
residual.sugar      -2.964e-01  1.604e-02 -18.478  < 2e-16 ***
free.sulfur.dioxide -7.552e-03  2.485e-03  -3.039  0.00237 **
density              6.843e+02  2.967e+01  23.063  < 2e-16 ***
pH                  -1.777e+00  2.700e-01  -6.580 4.71e-11 ***
sulphates           -2.137e+00  3.816e-01  -5.601 2.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4970.9  on 3917  degrees of freedom
Residual deviance: 4029.1  on 3911  degrees of freedom
AIC: 4043.1

Number of Fisher Scoring iterations: 5
```

Reduced Model



Logistic Regression - Performance

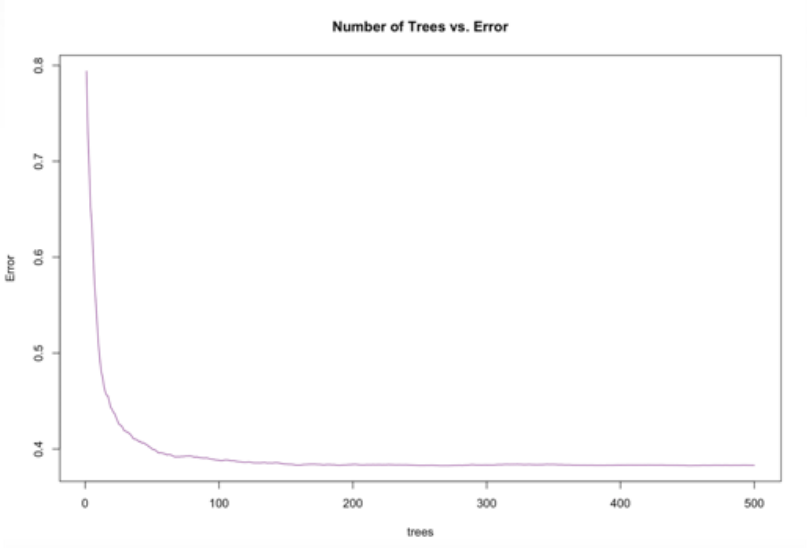
	acc.log.val	sens.log.high.val	prec.log.high.val
1	0.7215837	0.2934783	0.7788462

Validation Results

	acc.log.test	sens.log.high.test	prec.log.high.test
1	0.7183673	0.3150289	0.7364865

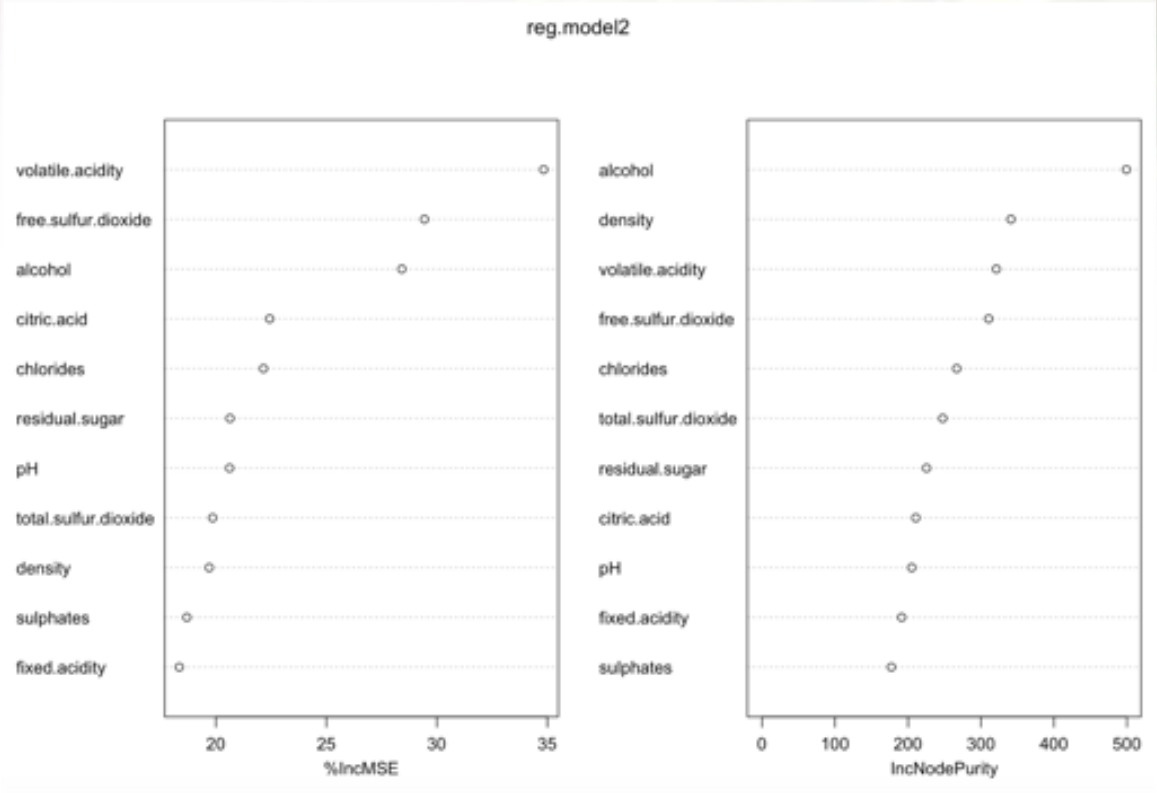
Test Results

Random Forest with Regression Trees




```
Call:
 randomForest(formula = quality ~ ., data = train, ntree = 100,
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 3

          Mean of squared residuals: 0.3941869
                    % Var explained: 50.89
```

Variable Importance Plots



Random Forest - Performance

Validation MSE

```
[1] 0.07319144
```

Test MSE

```
[1] 0.3157703
```

- The validation error was less than the test error
- Not that awful in the grand scheme, Wine Quality would predicted as ± .3

Performance Evaluation/Conclusion

- Classification models (KNN vs. Logistic Regression):
  - KNN returns a better model

```
        acc sens.high prec.high
1 0.7591837 0.8470032 0.7943787
```

- Quantitative models (Linear Regression vs. Random Forest):
  - Random Forest is better in terms of MSE

```
[1] 0.3157703
```

What's next to come...

Feature Engineering

- Also have access to Red Wine Data, combine with White Wine Data and create new variable `color`
  - Determine if new variable `color` is a viable factor when determining `quality`
- Create new variable `flavor` which would be based `citric.acid` and `residual.sugar`
  - Determine if new variable `flavor` is a viable factor when determining `quality`

Questions?