Investigation of Portuguese Wine Quality Data

Presented by: Amy Gao

Data Description

- A detailed chemical dataset was created in 2009 on 4898 white wines made in Portugal.

- All of the 12 data variables in this dataset are quantitative.

- Data input variables are based on objective physicochemical tests (e.g. Citric Acid, Residual Sugar, PH level)

- Quality is the only output variable, which is based on sensory data graded on a standard scale from 0 (very bad) to 10 (very excellent).

Objective

- build predictive models for the quality of these Portuguese wines and test the predictive power of the models

- we can possibly use the information to create better quality wine by adding or removing some of these factors

Creating Binary variables

Creating binary variable (0,1)for wine quality

Greater than or equal to 6, quality == 1------ Good quality

Less than 6, quality ==0----- Poor quality

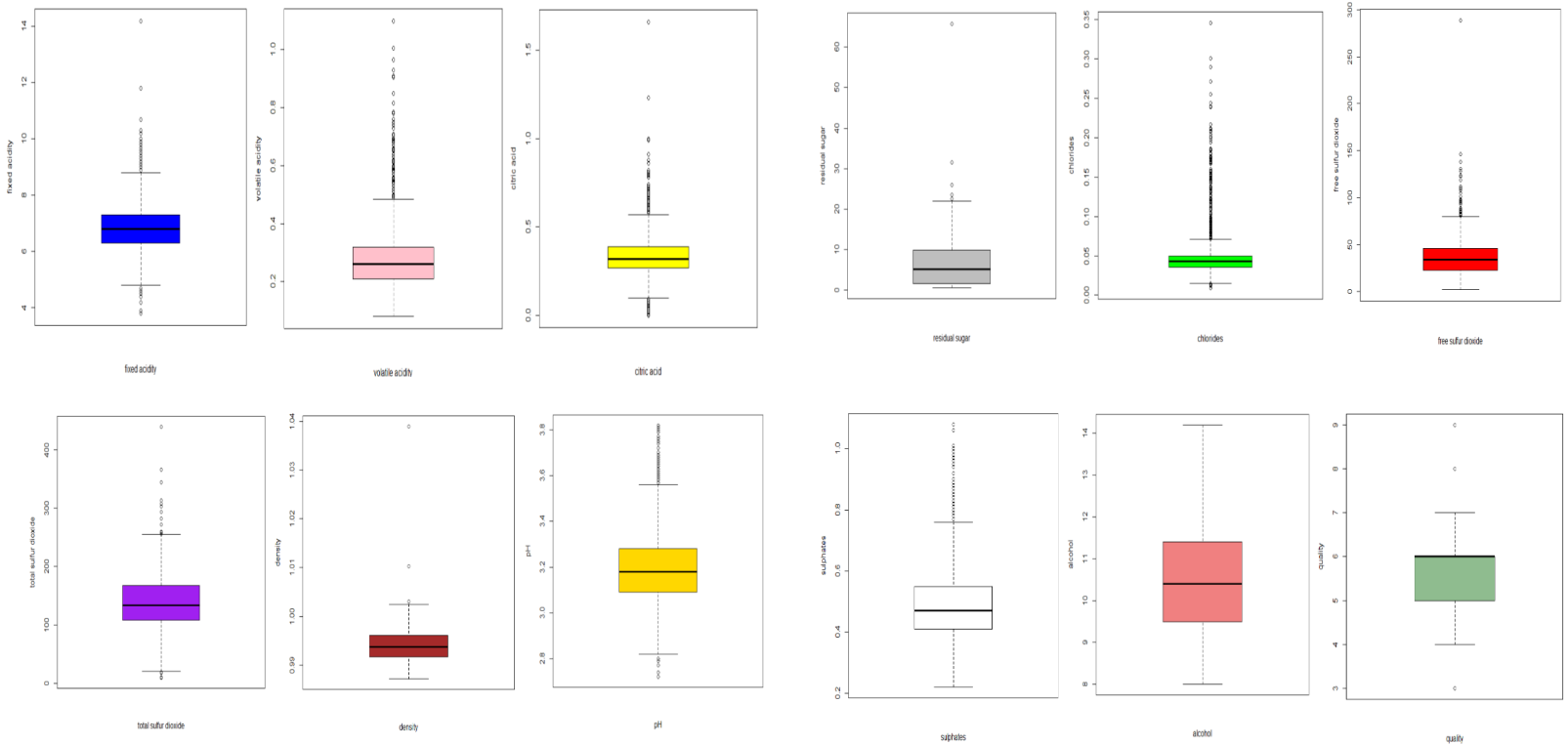
```
> wine$quality<-factor(ifelse(wine$quality>=6,1,0))
> summary(wine$quality)
   0    1
1640 3258
```

Exploratory Data Analysis

```
> summary(wine)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700   1st Qu.:0.03600
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200   Median :0.04300
 Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391   Mean   :0.04577
 3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900   3rd Qu.:0.05000
 Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800   Max.   :0.34600
 free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :  2.00      Min.   :  9.0        Min.   :0.9871   Min.   :2.720
 1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090
 Median : 34.00      Median :134.0        Median :0.9937   Median :3.180
 Mean   : 35.31      Mean   :138.4        Mean   :0.9940   Mean   :3.188
 3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280
 Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :3.820
   sulphates         alcohol          quality
 Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
 Median :0.4700   Median :10.40   Median :6.000
 Mean   :0.4898   Mean   :10.51   Mean   :5.878
 3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
 Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Based on the outputs, there are 12 variables in the wine dataset:

fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality .



- No outliers for Alcohol variable.
- Quality, Density and Residual sugar have a few outlier
- Mostly outliers are on the larger side.

Split as training/testing datasets:

```
> subset <- sample(nrow(wine), nrow(wine) * 0.9)
> wine.train = wine[subset, ]
> wine.test=wine[-subset, ]
```

Logistic Regression (Full Model)

```
> wine.glm1<-glm(quality~ .,family=binomial,wine.train)
> summary(wine.glm1)

Call:
glm(formula = quality ~ ., family = binomial, data = wine.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1079  -0.9045   0.4495   0.8051   3.0125

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.536e+02  7.492e+01   3.384 0.000713 ***
fixed.acidity         4.134e-02  7.539e-02   0.548 0.583423
volatile.acidity     -6.559e+00  4.366e-01 -15.022  < 2e-16 ***
citric.acid           2.413e-01  3.196e-01   0.755 0.450307
residual.sugar        1.641e-01  2.851e-02   5.756 8.61e-09 ***
chlorides             9.452e-02  1.774e+00   0.053 0.957521
free.sulfur.dioxide   9.196e-03  2.946e-03   3.121 0.001801 **
total.sulfur.dioxide -1.006e-03  1.277e-03  -0.788 0.430628
density              -2.658e+02  7.593e+01  -3.500 0.000465 ***
pH                    1.001e+00  3.812e-01   2.625 0.008663 **
sulphates             1.887e+00  3.815e-01   4.946 7.57e-07 ***
alcohol               7.243e-01  9.913e-02   7.306 2.75e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5619.3  on 4407  degrees of freedom
Residual deviance: 4466.4  on 4396  degrees of freedom
AIC: 4490.4

Number of Fisher Scoring iterations: 5
```

Stepwise variable selection

```
Step:  AIC=4484.14
quality ~ volatile.acidity + residual.sugar + free.sulfur.dioxide +
    density + pH + sulphates + alcohol

                      Df Deviance    AIC
<none>                     4468.1 4484.1
- pH                   1   4477.4 4491.4
- free.sulfur.dioxide  1   4479.6 4493.6
- sulphates            1   4492.9 4506.9
- density              1   4496.7 4510.7
- residual.sugar       1   4537.9 4551.9
- alcohol              1   4590.9 4604.9
- volatile.acidity     1   4763.8 4777.8
```

7 Variables selected out of 11: volatile.acidity, residual sugar, free.sulfur.dioxide, density, PH, sulphates and alcohol

The Best Model and Model evaluation

```
> wine.glmbest<-glm(quality~ volatile.acidity + residual.sugar + free.sulfur.dioxide +density + pH + sulphates + alcohol, family=binomial, wine.train)
> AIC(wine.glmbest)
[1] 4484.137
> AIC(wine.glm1)
[1] 4490.448
```

AIC(wine.glm1)= 4490.4

AIC(wine.glmbest)=4484.1

```
> BIC(wine.glm1)
[1] 4567.142
> BIC(wine.glmbest)
[1] 4535.267
```

BIC(wine.glm1)=4567.1

BIC(wine.glmbest)=4535.3

In-sample Prediction

```
> prob.glmbest.insample <- predict(wine.glmbest, type = "response")
> predicted.glmbest.insample <- prob.glmbest.insample >0.5
> predicted.glmbest.insample <- as.numeric(predicted.glmbest.insample)
> table(wine.train$quality, predicted.glmbest.insample, dnn = c("Truth", "Predicted")
+ )
     Predicted
Truth    0    1
    0  722  753
    1  347 2586

> mean(ifelse(wine.train$quality != predicted.glmbest.insample, 1, 0))
[1] 0.2495463
```

Set cut-off probability as 0.5

Misclassification rate: 24.95%

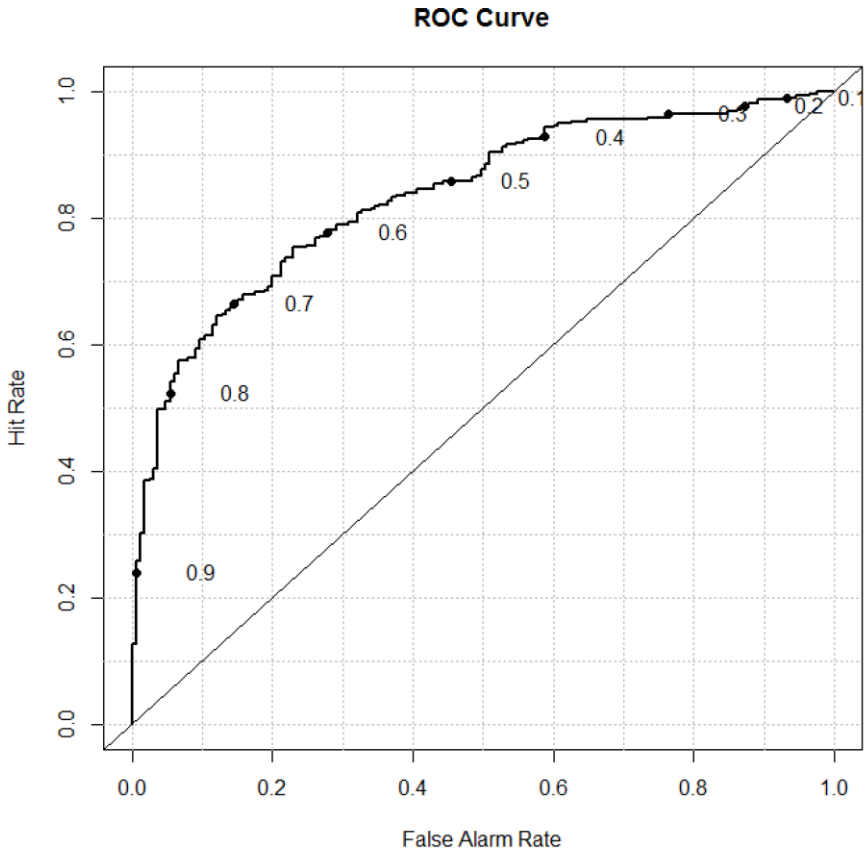
Out-of-sample Prediction

```
> prob.glmbest.outsample <- predict(wine.glmbest, wine.test, type= "response")
> predicted.glmbest.outsample <- prob.glmbest.outsample >0.5
> predicted.glmbest.outsample <- as.numeric(predicted.glmbest.outsample)
> table(wine.test$quality, predicted.glmbest.outsample, dnn = c("Truth", "Predicted"))
     Predicted
Truth   0   1
    0  89  76
    1  46 279
> mean(ifelse(wine.test$quality != predicted.glmbest.outsample, 1, 0))
[1] 0.2489796
```

Out-of-sample misclassification rate: 24.90%

In-sample misclassification rate: 24.95%

Out-of-sample ROC Curve




```
> roc.plot(wine.test$quality == "1", prob.glmbest.outsample)$roc.vol
    Model      Area      p.value binorm.area
1 Model  1 0.8353753 3.203871e-34          NA
```

Area under curve: 0.835

Cross validation and cost function

```
> pcut = 0.5
> cost1 <- function(r, pi) {mean(((r == 0) & (pi > pcut)) | ((r == 1) & (pi < pcut))) }
> cost2 <- function(r, pi) {
+ weight1 = 2
+ weight0 = 1
+ c1= (r == 1) & (pi < pcut)
+ c0= (r == 0) & (pi > pcut)
+ return(mean(weight1 * c1 + weight0 * c0))
+ }
> library(boot)
> wine.glmcross<-glm(quality~ volatile.acidity + residual.sugar + free.sulfur.dioxide +density + pH + sulphates + alcohol, family=binomial, wine)
> cv.result = cv.glm(wine, wine.glmcross, cost1, 10)
> cv.result$delta
[1] 0.2503062 0.2508992
> cv.result = cv.glm(wine, wine.glmcross, cost2, 10)
> cv.result$delta
[1] 0.3307472 0.3320951
```

10-fold cross validation with symmetric cost function: 0.25

10-fold cross validation with asymmetric cost function: 0.33 (double penalizing getting 1 wrong)

Out-of sample performance: 0.249

Conclusion for Logistic Regression

7 predictors: volatile.acidity, residual sugar, free.sulfur.dioxide, density, PH, sulphates and alcohol

Misclassification rate: 25%

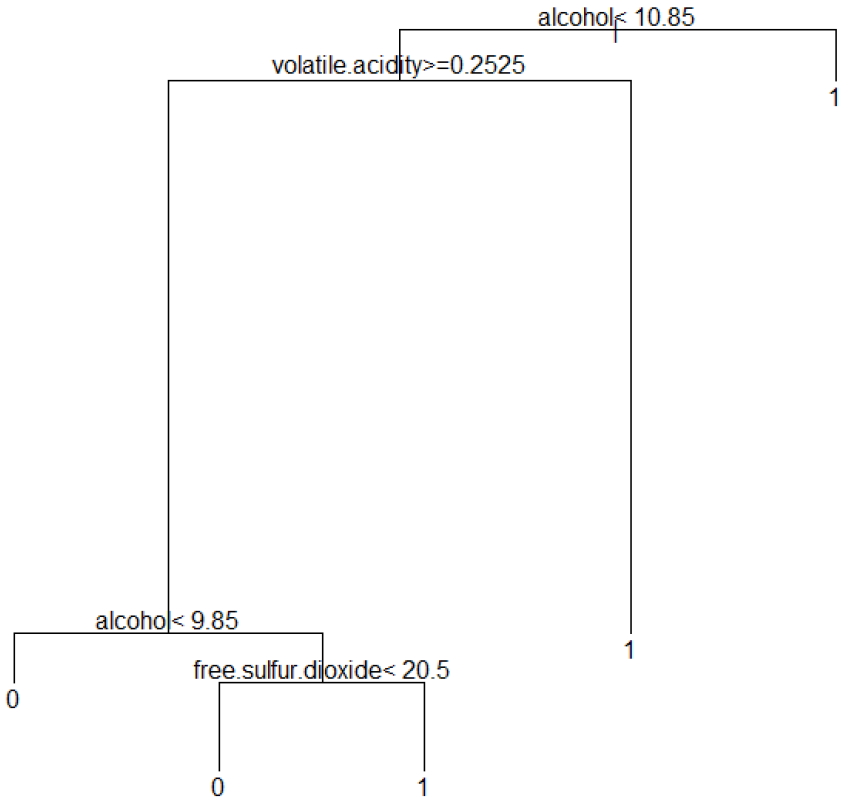
AUC=0.835
$0.8 \leq AUC < 0.9$ Excellent classification
Logistic regression model provides excellent classification.

Classification Tree

```
> wine.rpart <- rpart(formula = quality~ ., data = wine.train, method = "class")
> wine.rpart
n= 4408

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 4408 1475 1 (0.3346189 0.6653811)
   2) alcohol< 10.85 2770 1272 1 (0.4592058 0.5407942)
     4) volatile.acidity>=0.2525 1434  548 0 (0.6178522 0.3821478)
       8) alcohol< 9.85 925  287 0 (0.6897297 0.3102703) *
       9) alcohol>=9.85 509  248 1 (0.4872299 0.5127701)
        18) free.sulfur.dioxide< 20.5 141   44 0 (0.6879433 0.3120567) *
        19) free.sulfur.dioxide>=20.5 368  151 1 (0.4103261 0.5896739) *
     5) volatile.acidity< 0.2525 1336  386 1 (0.2889222 0.7110778) *
   3) alcohol>=10.85 1638  203 1 (0.1239316 0.8760684) *
> plot(wine.rpart)
> text(wine.rpart, pretty = TRUE)
```



Classification Tree

3 Predictors:

Alcohol

Volatile. Acidity

Free. Sulfur dioxide

In-sample prediction

```
> wine.train.pred.tree1 = predict(wine.rpart, wine.train, type = "class")
> table(wine.train$quality, wine.train.pred.tree1, dnn = c("Truth", "Predicted"))
     Predicted
Truth    0    1
    0  735  740
    1  331 2602
```

```
> mean(ifelse(wine.train$quality !=  wine.train.pred.tree1, 1, 0))
[1] 0.2429673
```

In-sample misclassification rate: 24.30%

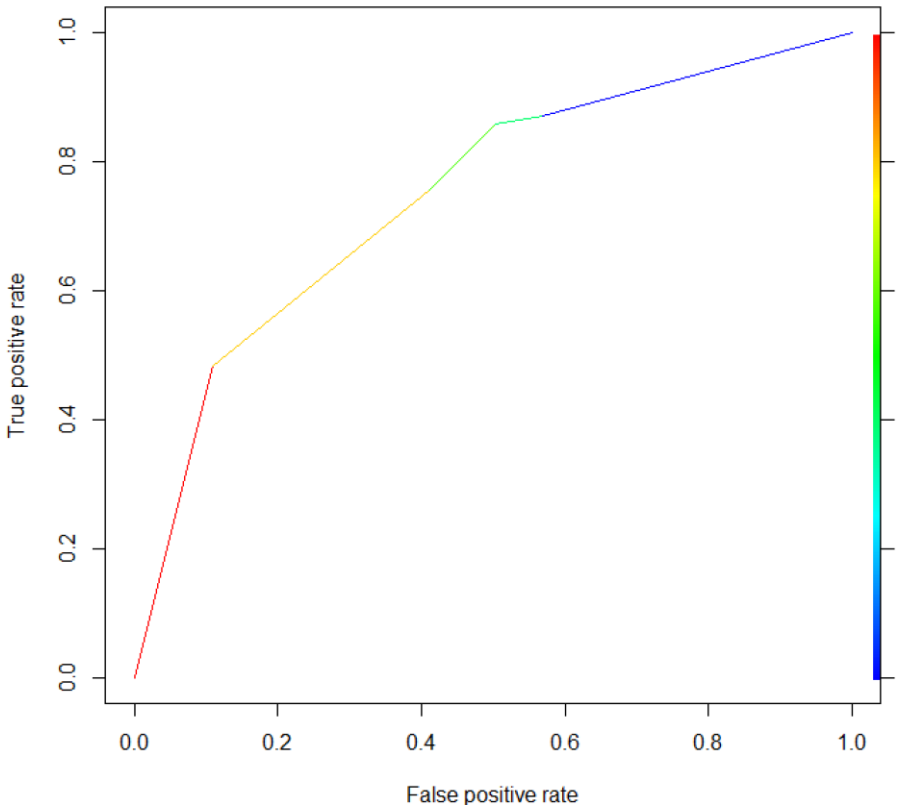
Out-of-sample prediction

```
> wine.test.pred.tree2 = predict(wine.rpart, wine.test, type = "class")
> table(wine.test$quality, wine.test.pred.tree2, dnn = c("Truth", "Predicted"))
     Predicted
Truth   0   1
    0  82  83
    1  46 279
```

```
> mean(ifelse(wine.test$quality !=  wine.test.pred.tree2, 1, 0))
[1] 0.2632653
```

Out-of-sample misclassification rate: 26.33%

Out-of-sample ROC Curve




```
> pred = prediction(wine.test.prob.rpart2[, 2], wine.test$quality)
> perf = performance(pred, "tpr", "fpr")
> plot(perf, colorize = TRUE)
```

```
> slot(performance(pred, "auc"), "y.values")[[1]]
[1] 0.7477949
```

Area under curve= 0.748

Conclusion for classification tree

3 predictors: Alcohol, volatile. acidity, free. Sulfur dioxide

Misclassification rate: 26.33%

AUC=0.748

$0.7 \leq AUC < 0.8$ Acceptable classification

Classification tree model provides acceptable classification.

Comparison and Conclusion

Logistic regression:

7 predictors

Misclassification rate: 24.90%

AUC=0.835, Excellent classification

Classification tree:

3 predictors: Alcohol, Volatile. acidity, free. Sulfur dioxide

Out-of-sample Misclassification rate: 26.33%

AUC= 0.748, Acceptable classification

Logistic regression model is more accurate in predicting wine quality

Classification tree is simpler