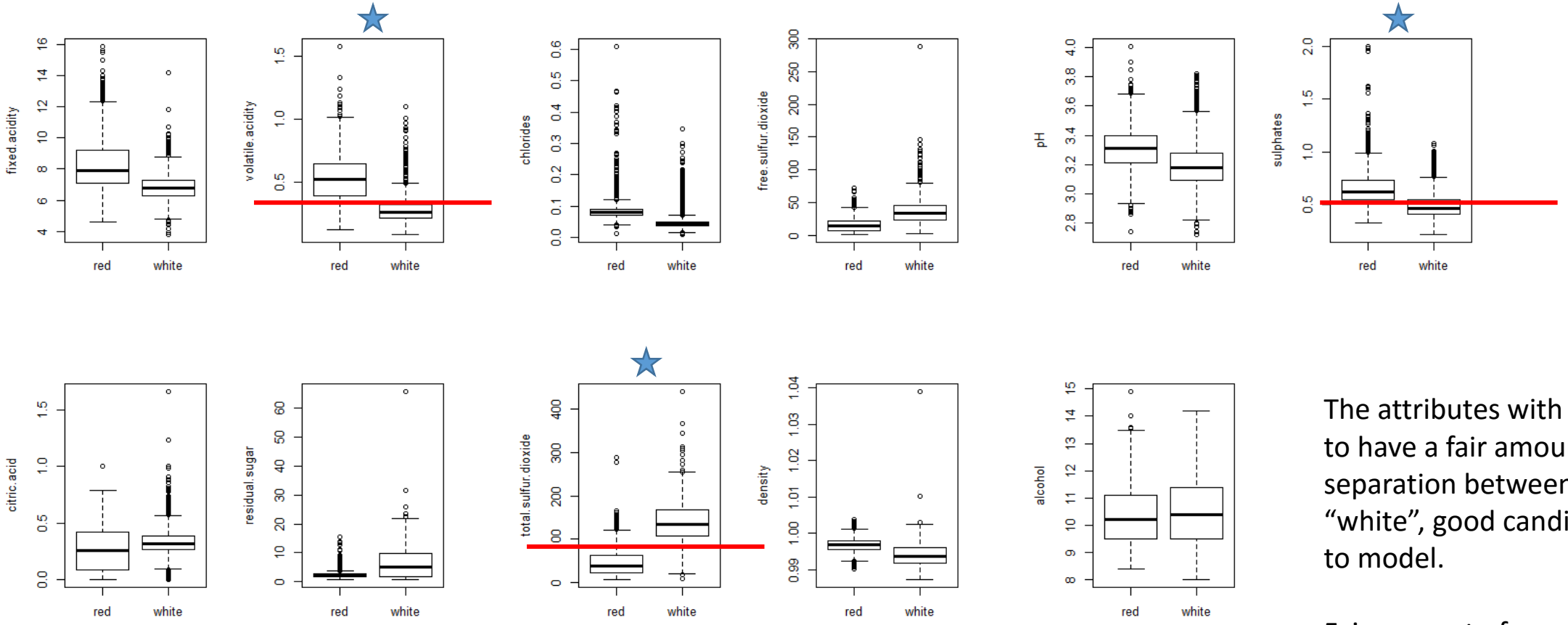
Red or White

Data

- 11 wine attributes
- 1599 red wine
- 4898 white wine

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

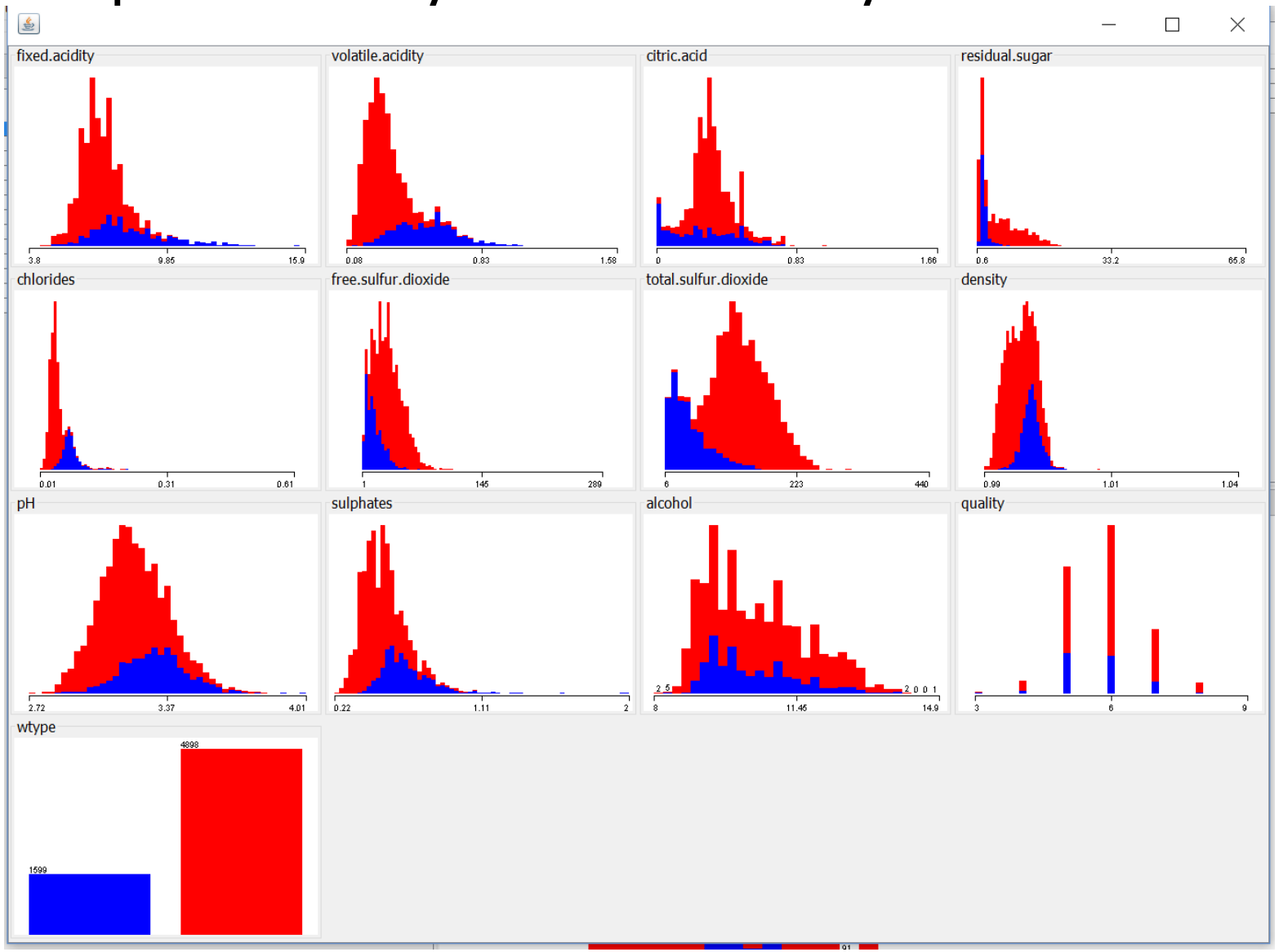
Exploratory Data Analysis



The attributes with a star appears to have a fair amount of separation between “red” and “white”, good candidates as inputs to model.

Fair amount of overlap among the extreme values!!

Exploratory Data Analysis

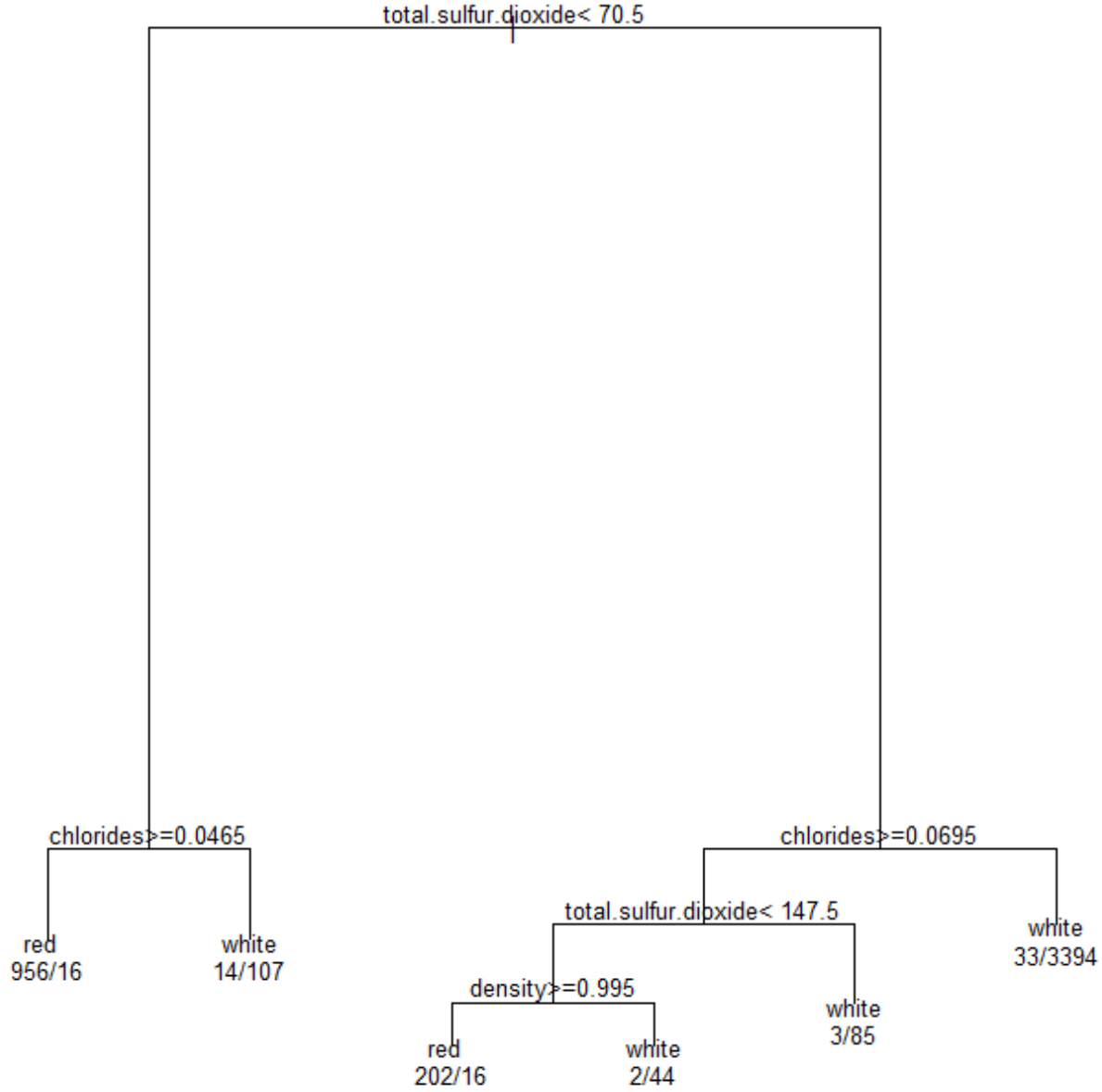


Use rpart()

```
library( rpart )
idx <- sample( c(1: dim(wine)[1] ), 0.75* dim(wine)[1] )
wine.rpart <- rpart( wtype ~ fixed.acidity + volatile.acidity + citric.acid +
                 residual.sugar + chlorides + free.sulfur.dioxide +
                total.sulfur.dioxide + density + pH +
                sulphates + alcohol, + data = wine[ idx, ] )
wine.rpart
```

n= 4872 node), split, n, loss, yval, (yprob) * denotes terminal node

1) root 4872 1210 white (0.248357964 0.751642036)
2) total.sulfur.dioxide< 70.5 1093 123 red (0.887465691 0.112534309)
4) chlorides>=0.0465 972 16 red (0.983539095 0.016460905) *
5) chlorides< 0.0465 121 14 white (0.115702479 0.884297521) *
3) total.sulfur.dioxide>=70.5 3779 240 white (0.063508865 0.936491135)
6) chlorides>=0.0695 352 145 red (0.588068182 0.411931818)
12) total.sulfur.dioxide< 147.5 264 60 red (0.772727273 0.227272727)
24) density>=0.99498 218 16 red (0.926605505 0.073394495) *
25) density< 0.99498 46 2 white (0.043478261 0.956521739) *
13) total.sulfur.dioxide>=147.5 88 3 white (0.034090909 0.965909091) *
7) chlorides< 0.0695 3427 33 white (0.009629413 0.990370587) *

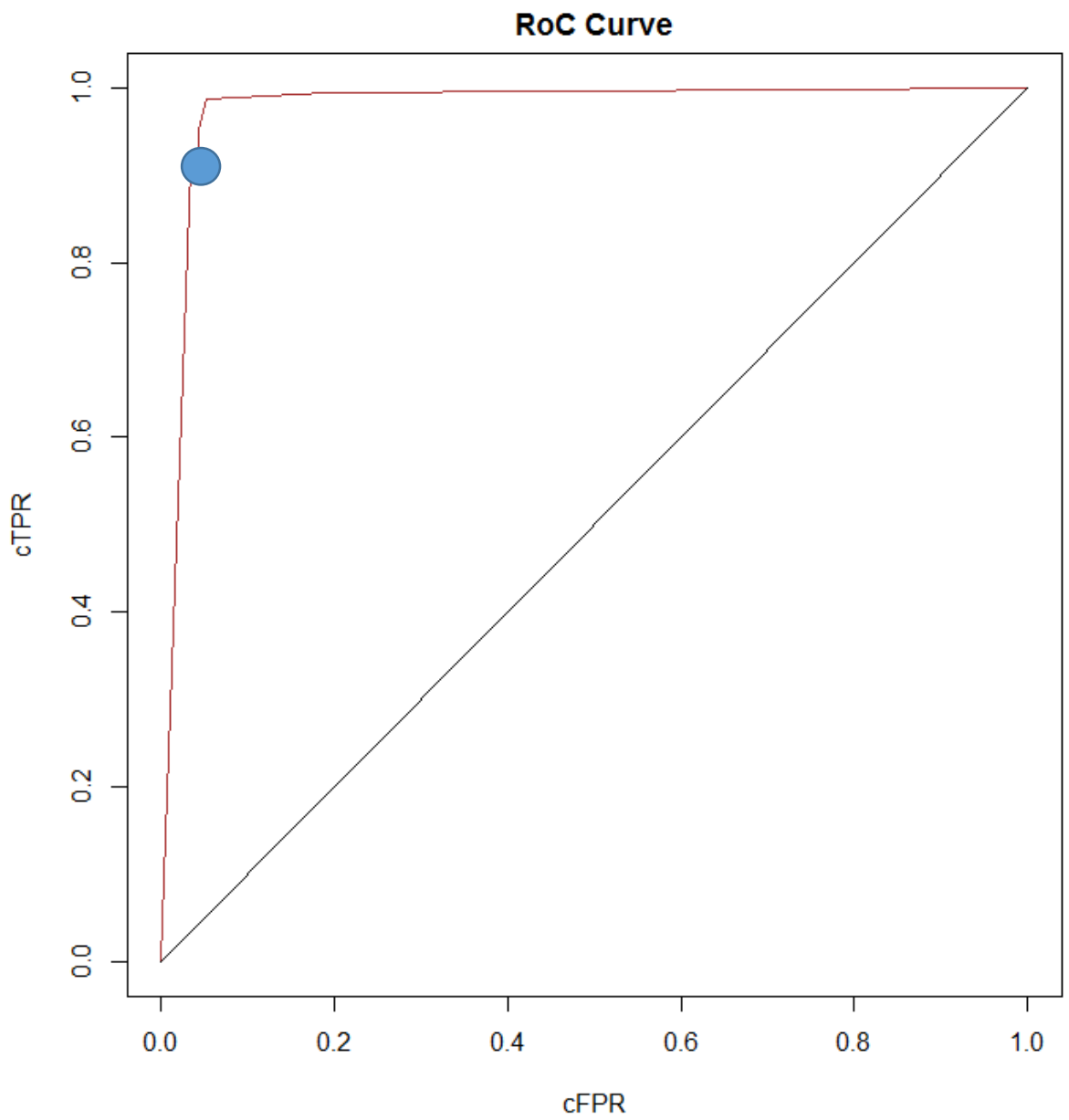


Evaluate model

```
table( wine$wtype[-idx], wine.pred[,2] > 0.5 )
        FALSE TRUE
red       369   20
white      16 1220

prop.table( table( wine$wtype[-idx], wine.pred[,2] > 0.5 ) )
          FALSE       TRUE
red     0.227076923 0.012307692
white   0.009846154 0.750769231

roc.curve(yy=( wine$wtype[-idx]=='white')*1.0, s=0.5,
yhat=wine.pred[,2], print=TRUE)
          Predicted
Data        0    1
0         369   20
1          16 1220
FPR       TPR     Recall    F1Score
0.05141388 0.98705502 0.98387097 0.98546042
```

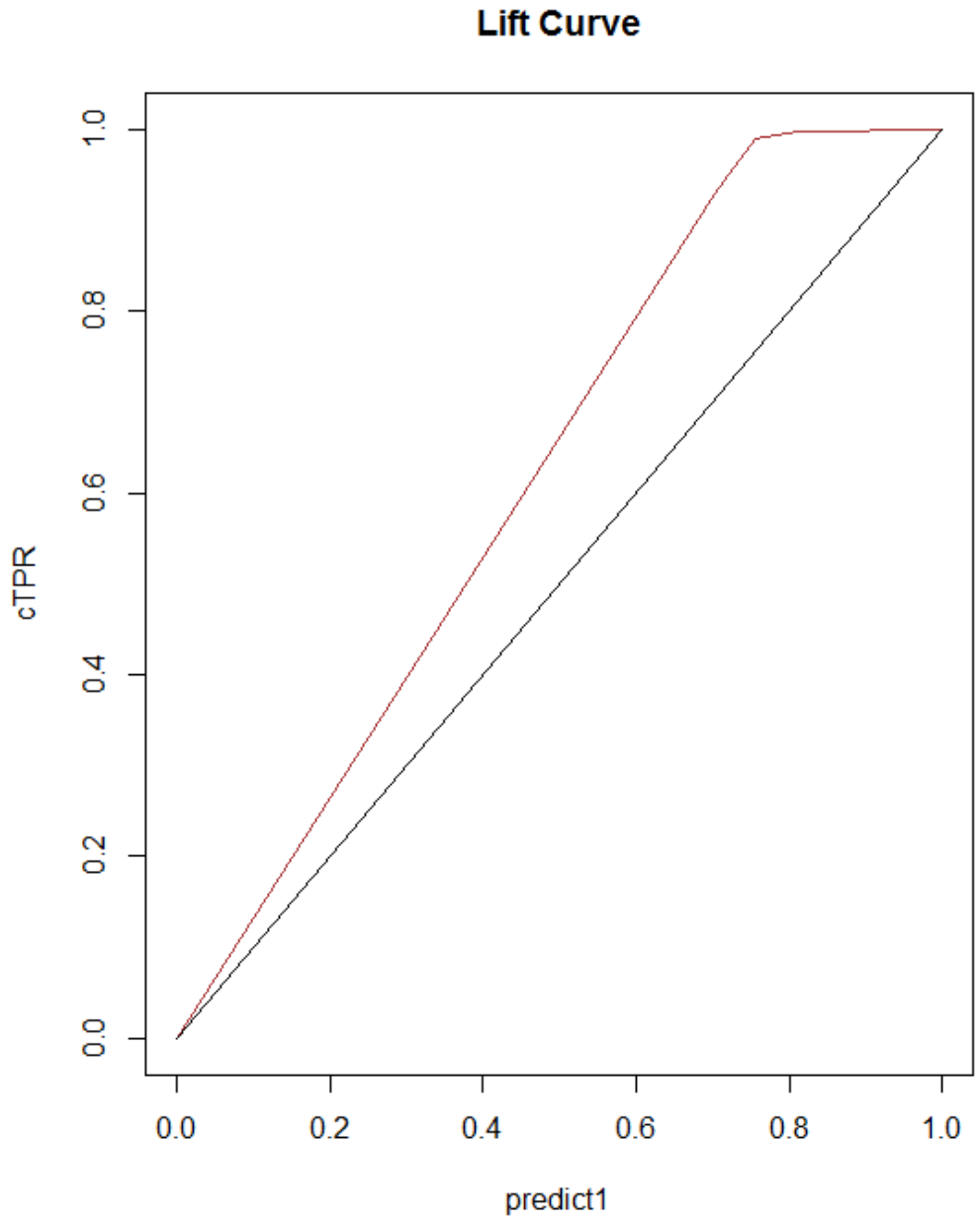


Lift chart

Lift chart:

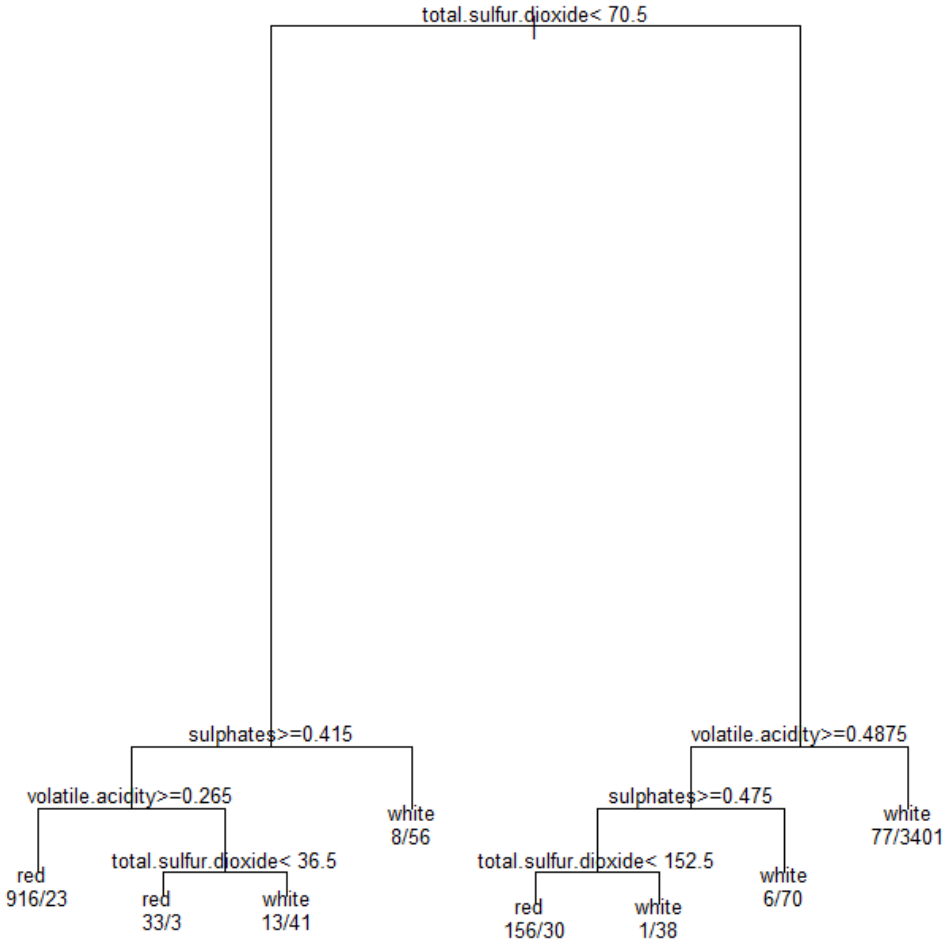
Y-axis: TPR
X-axis: (TP+FP)/observed

Ref: Vuk and Curk (2006)



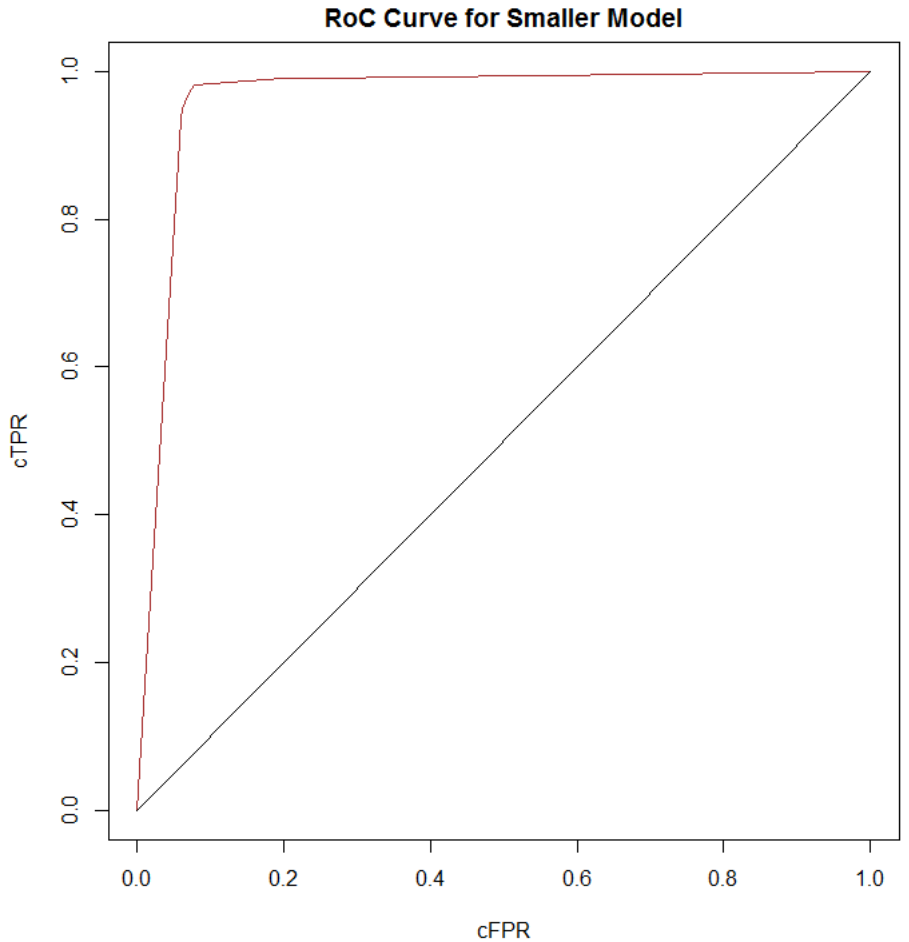
Smaller model

```
wine.rpart2 <- rpart( wtype ~ volatile.acidity +
        total.sulfur.dioxide +sulphates , data = wine[ idx, ] )
```




```
> table( wine$wtype[-idx], wine.pred2[,2] > 0.5 )
              FALSE TRUE
red             359   30
white            22 1214

> prop.table( table( wine$wtype[-idx], wine.pred2[,2] > 0.5 ) )
              FALSE      TRUE
red   0.22092308 0.01846154
white 0.01353846 0.74707692
```



Change the cutoff

```
> roc.curve(yy=( wine$wtype[-idx]=='white')*1.0,    s= 4898/(4898+1599),
                        yhat=wine.pred2[,2], print=TRUE)

              Predicted
Data          0    1
0             359   30
1             22 1214
-------------------------------
FPR       TPR     Recall    F1Score
0.07712082 0.98220065 0.97588424 0.97903226
```

R: library(ROCR)

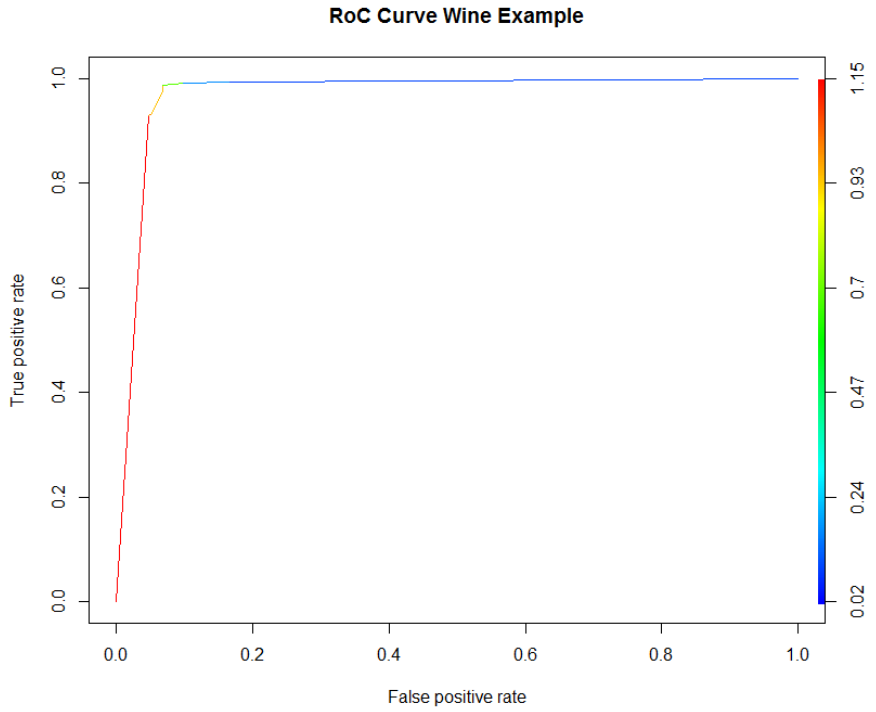
RoC curve:
Y-axis: TPR
X-axis: FPR

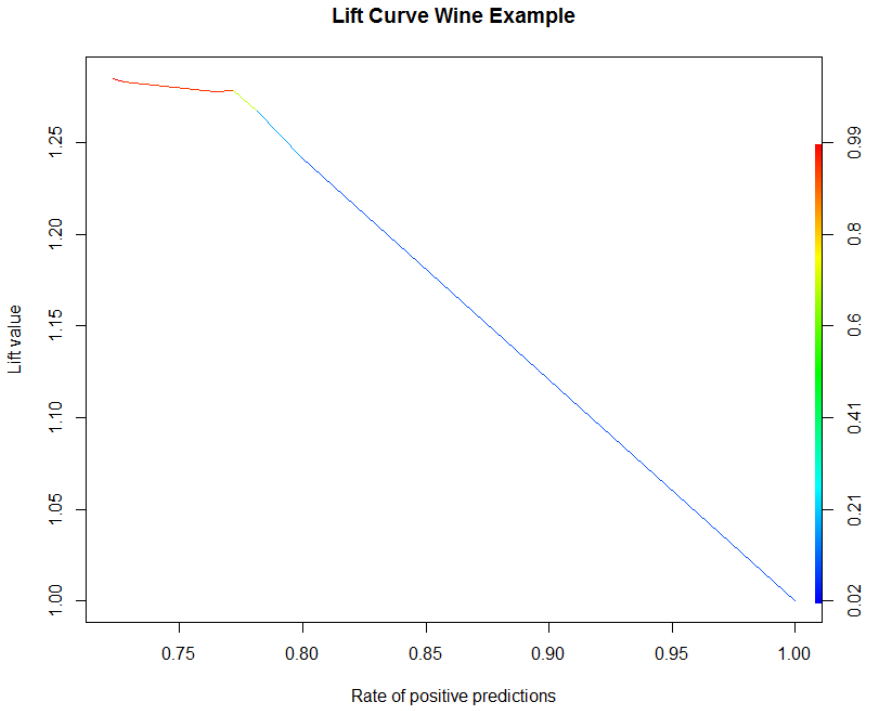
Lift chart:
Y-axis: TPR/((TP+FP)/total)
X-axis: (TP+FP)/(total)

```
library(ROCR)

wine.yhy <- prediction( wine.pred[,2], wine$wtype[-idx] )
wine.roc <- performance( wine.yhy, 'tpr', 'fpr' )
plot( wine.roc, main='RoC Curve Wine Example', colorize=T )

wine.lft <- performance( wine.yhy, 'lift', 'rpp')
plot( wine.lft, main='Lift Curve Wine Example', colorize=T )
```





R:

PR curve:
Y-axis: Precision, Recall
X-axis: Cutoff

The wine example has slightly class imbalance situation. White wine composed of about 75% of both train and test data.

The Precision-Recall curve allow us to examine a cutoff that leads to high precision and high recall. There is a region of possibility. Cutoff between 0.70 to 0.93 yield desirable precision and recall.