# Predicting Heart disease regression

*Joshua Smith*

*April 16, 2016*

```
r   knitr::opts_chunk$set(prompt=TRUE, comment="", echo=FALSE)
```

## Model One

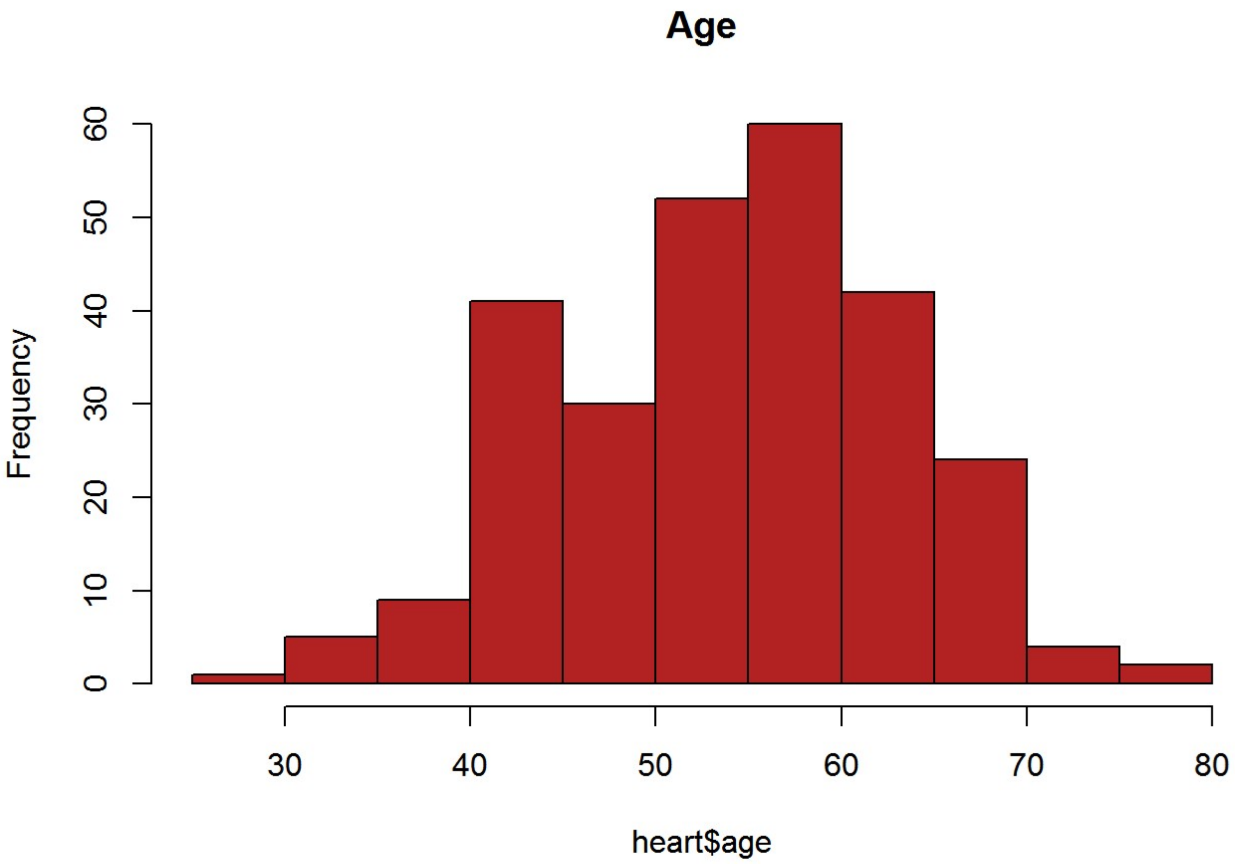
In this Model I just wanted to take a look at some of the data to see the general age of the subjects as well as the ratio on males to females. The first table belows data shows sex as 1’s and 0’s. By doing some google research I found out the the ones are male and the latter female.

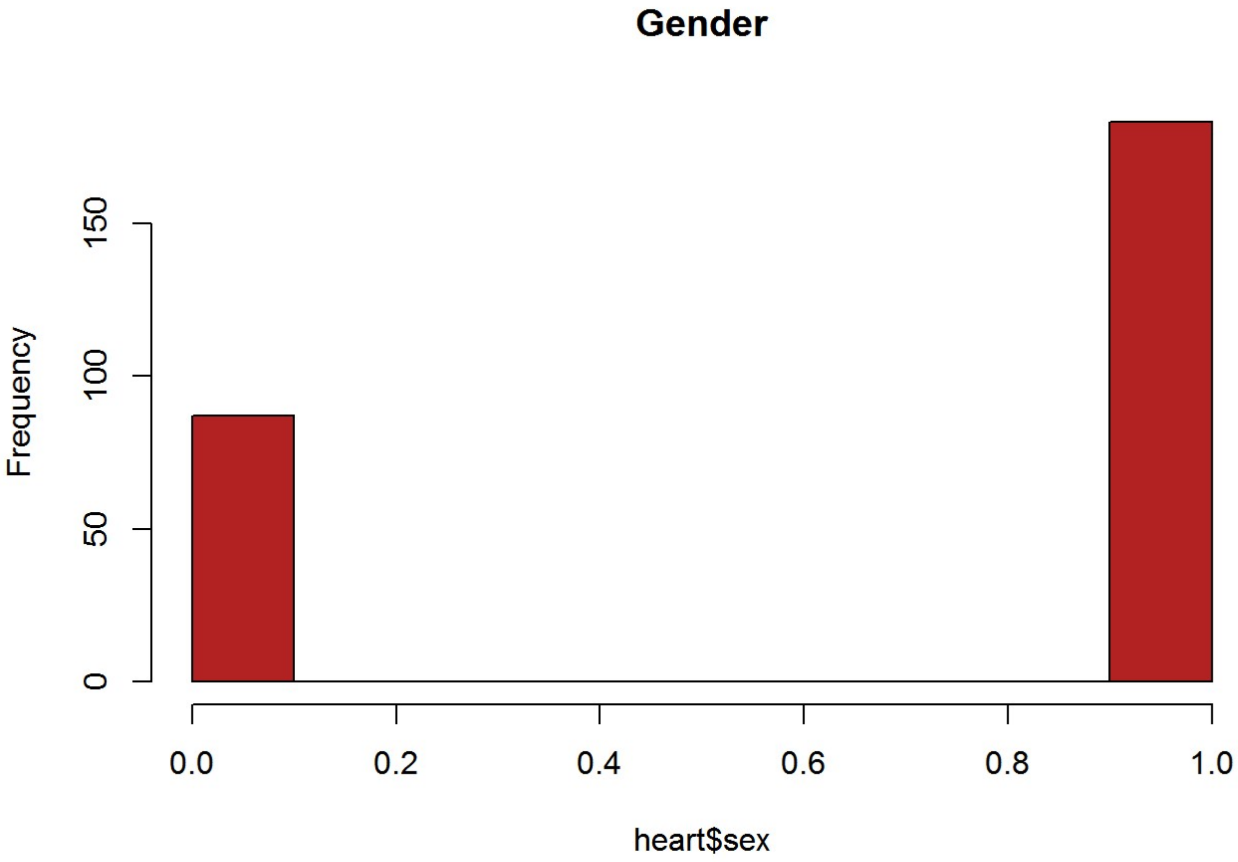
```
'data.frame':   270 obs. of  14 variables:
 $ age      : num  70 67 57 64 74 65 56 59 60 63 ...
 $ sex      : num  1 0 1 1 0 1 1 1 1 0 ...
 $ chestpain: num  4 3 2 4 2 4 3 4 4 4 ...
 $ restbp   : num  130 115 124 128 120 120 130 110 140 150 ...
 $ chol     : num  322 564 261 263 269 177 256 239 293 407 ...
 $ sugar    : num  0 0 0 0 0 0 1 0 0 0 ...
 $ ecg      : num  2 2 0 0 2 0 2 2 2 2 ...
 $ maxhr    : num  109 160 141 105 121 140 142 142 170 154 ...
 $ angina   : num  0 0 0 1 1 0 1 1 0 0 ...
 $ dep      : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4 ...
 $ exercise : num  2 2 1 2 1 1 2 2 2 2 ...
 $ fluor    : num  3 0 0 1 1 0 1 1 2 3 ...
 $ thal     : num  3 7 7 7 3 7 6 7 7 7 ...
 $ output   : num  1 0 1 0 0 0 1 1 1 1 ...
```

Looking at the summary of the data really helped when looking at the features age, sex, and chest pain. The min and max ages are 29 years old and and the max is 77 years old. The mean for sex is 0.6778 which means there are more males that females. The mean chestpain is 3.174 and min in 1 and max is 4. So the majority of the subjects had high chest pain.

```
      age             sex           chestpain        restbp     
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0  
 Median :55.00   Median :1.0000   Median :3.000   Median :130.0  
 Mean   :54.43   Mean   :0.6778   Mean   :3.174   Mean   :131.3  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0  
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0  
      chol           sugar             ecg            maxhr      
 Min.   :126.0   Min.   :0.0000   Min.   :0.000   Min.   : 71.0  
 1st Qu.:213.0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:133.0  
 Median :245.0   Median :0.0000   Median :2.000   Median :153.5  
 Mean   :249.7   Mean   :0.1481   Mean   :1.022   Mean   :149.7  
 3rd Qu.:280.0   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:166.0  
 Max.   :564.0   Max.   :1.0000   Max.   :2.000   Max.   :202.0  
     angina            dep          exercise         fluor       
 Min.   :0.0000   Min.   :0.00   Min.   :1.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :0.0000   Median :0.80   Median :2.000   Median :0.0000  
 Mean   :0.3296   Mean   :1.05   Mean   :1.585   Mean   :0.6704  
 3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :6.20   Max.   :3.000   Max.   :3.0000  
      thal           output      
 Min.   :3.000   Min.   :0.0000  
 1st Qu.:3.000   1st Qu.:0.0000  
 Median :3.000   Median :0.0000  
 Mean   :4.696   Mean   :0.4444  
 3rd Qu.:7.000   3rd Qu.:1.0000  
 Max.   :7.000   Max.   :1.0000  
```

I wanted to take a closer look at the age and sex distrabutions. These Graphs show the amomount of ages of the subject and also shows the amount of males in the data compaired to the amont of females. The ages seem to be mostly from 45 to 65 years old. Also, there looks like there are twice as many males to females in the data.


Age
Frequency
0
10
20
30
40
50
60
30
40
50
60
70
80
heart$age


Gender
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
heart$sex

# The best features selection from our first model

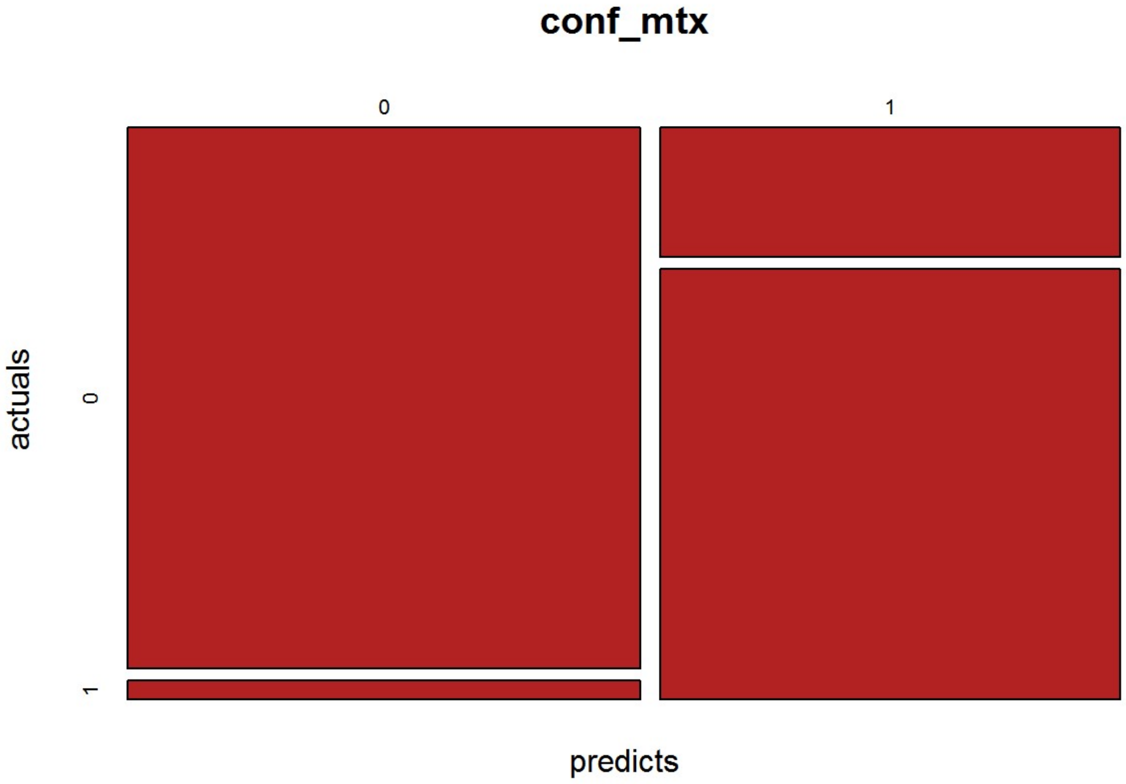
```

Call:
glm(formula = output ~ age + maxhr + sex + chol + chestpain + 
    ecg, family = binomial, data = heart)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3710  -0.7548  -0.2146   0.6667   2.8216  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.557490   2.198782  -1.618   0.1057    
age          0.030279   0.020187   1.500   0.1336    
maxhr       -0.035511   0.008476  -4.190 2.79e-05 ***
sex          2.080494   0.397204   5.238 1.62e-07 ***
chol         0.008009   0.003207   2.497   0.0125 *  
chestpain    0.977577   0.191481   5.105 3.30e-07 ***
ecg          0.306981   0.160980   1.907   0.0565 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.96  on 269  degrees of freedom
Residual deviance: 247.99  on 263  degrees of freedom
AIC: 261.99

Number of Fisher Scoring iterations: 5
```

Using all the feacures seems like a better choice so we are going to build our model using all the feactures.

```
Call:
glm(formula = output ~ ., family = binomial, data = tr_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7784  -0.5171  -0.1682   0.3835   2.3457  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.870830   3.449850  -1.702 0.088800 .  
age         -0.035775   0.030274  -1.182 0.237311    
sex          1.679739   0.624392   2.690 0.007141 ** 
chestpain    0.586304   0.234448   2.501 0.012392 *  
restbp       0.023816   0.012508   1.904 0.056897 .  
chol         0.007595   0.004668   1.627 0.103707    
sugar       -0.780854   0.704344  -1.109 0.267592    
ecg          0.348804   0.224439   1.554 0.120158    
maxhr       -0.028952   0.012336  -2.347 0.018930 *  
angina       0.913179   0.489607   1.865 0.062164 .  
dep          0.106721   0.248217   0.430 0.667231    
exercise     0.396515   0.434664   0.912 0.361646    
fluor        1.345171   0.326295   4.123 3.75e-05 ***
thal         0.406390   0.121295   3.350 0.000807 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.71  on 214  degrees of freedom
Residual deviance: 141.94  on 201  degrees of freedom
AIC: 169.94

Number of Fisher Scoring iterations: 6
```

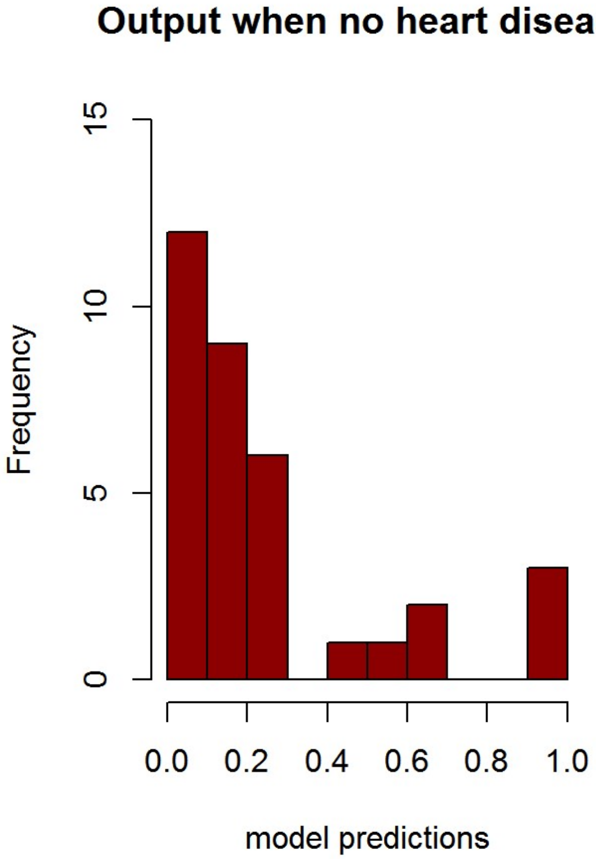
The conf. matrix show that it is a strong predictor. The graph of the matrix shows strong results.

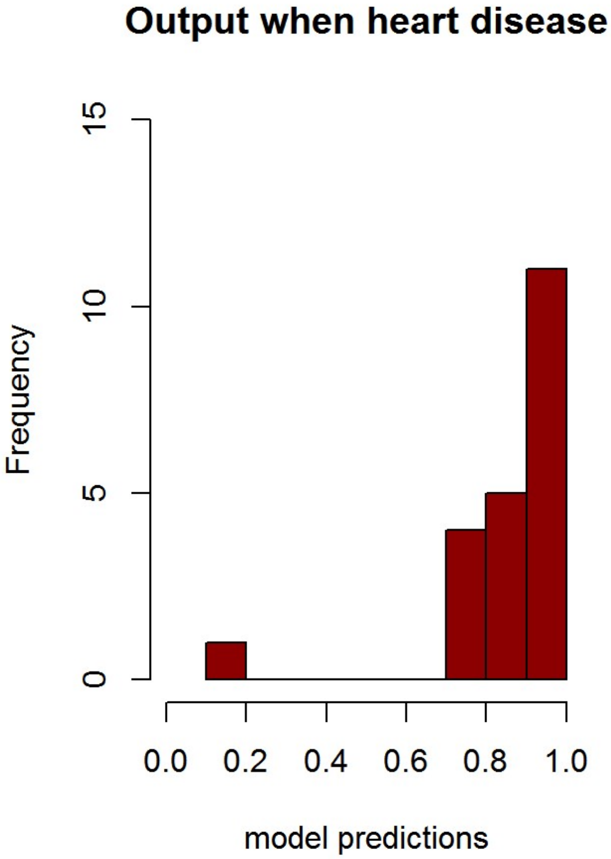
```
        actuals
predicts  0  1
       0 28  1
       1  6 20
```

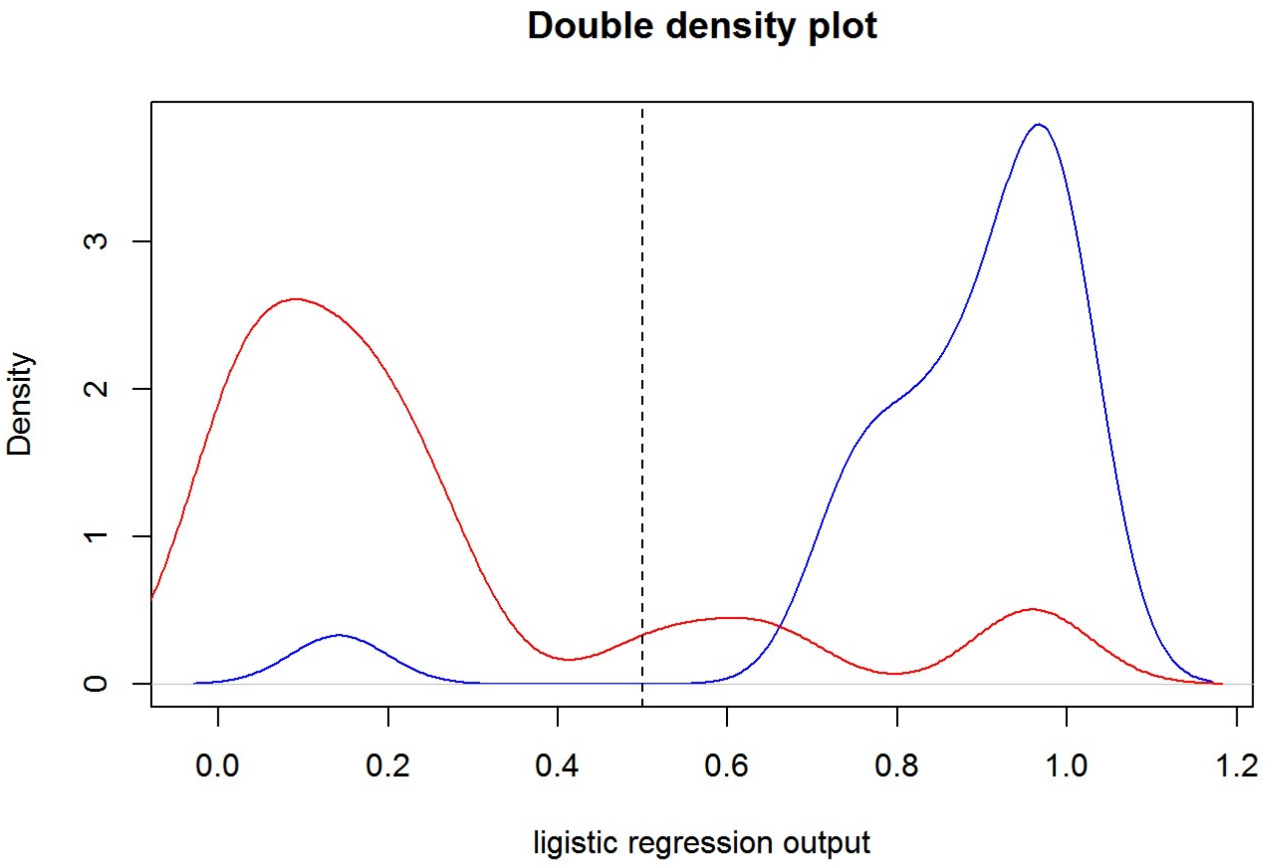


## Assessing the model

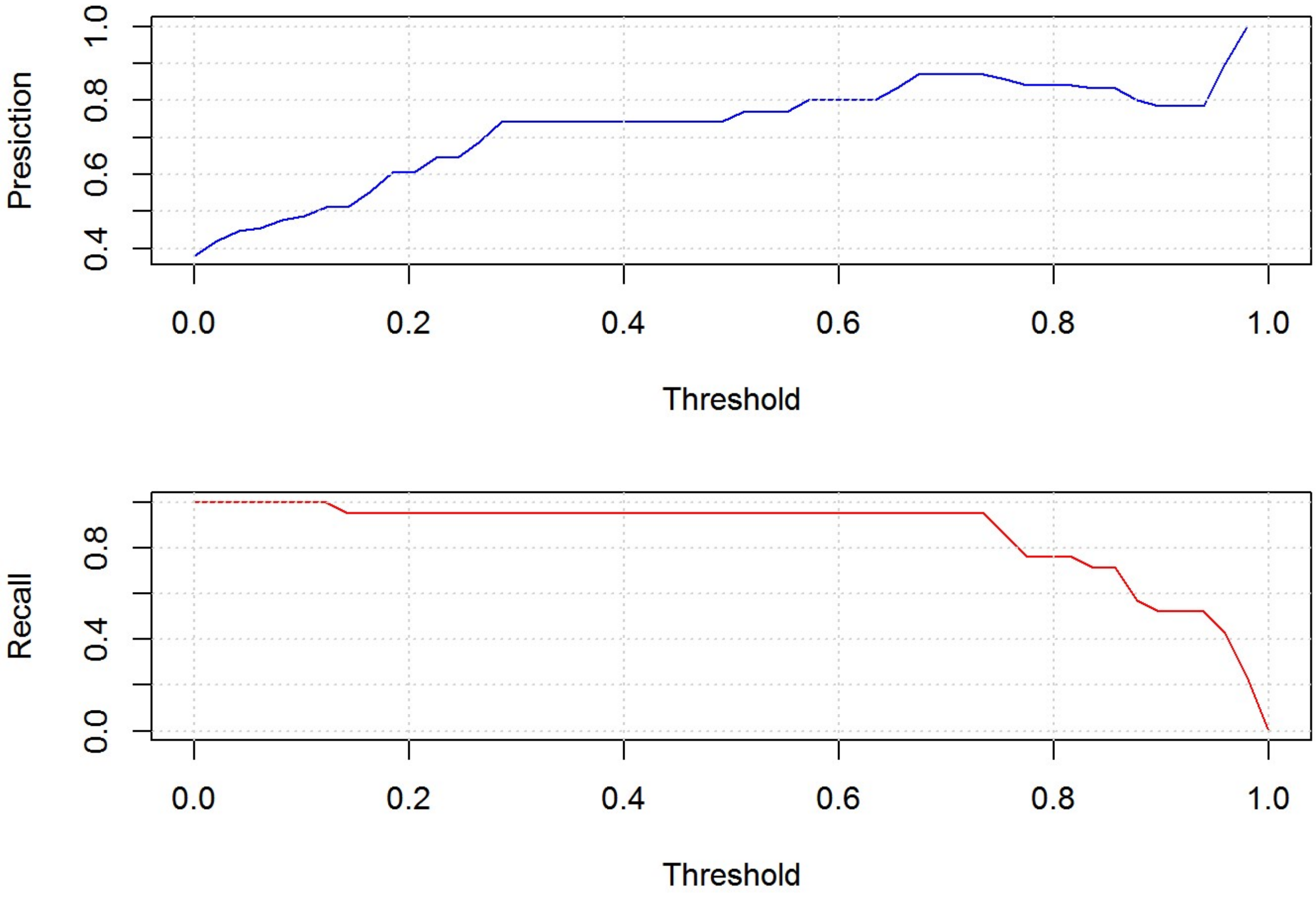
Looking at the output of the model on test cases where heart disease is present and not present we see that the models thrshold is right where we want it.


Output when no heart disease
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
model predictions


Output when heart disease
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
model predictions

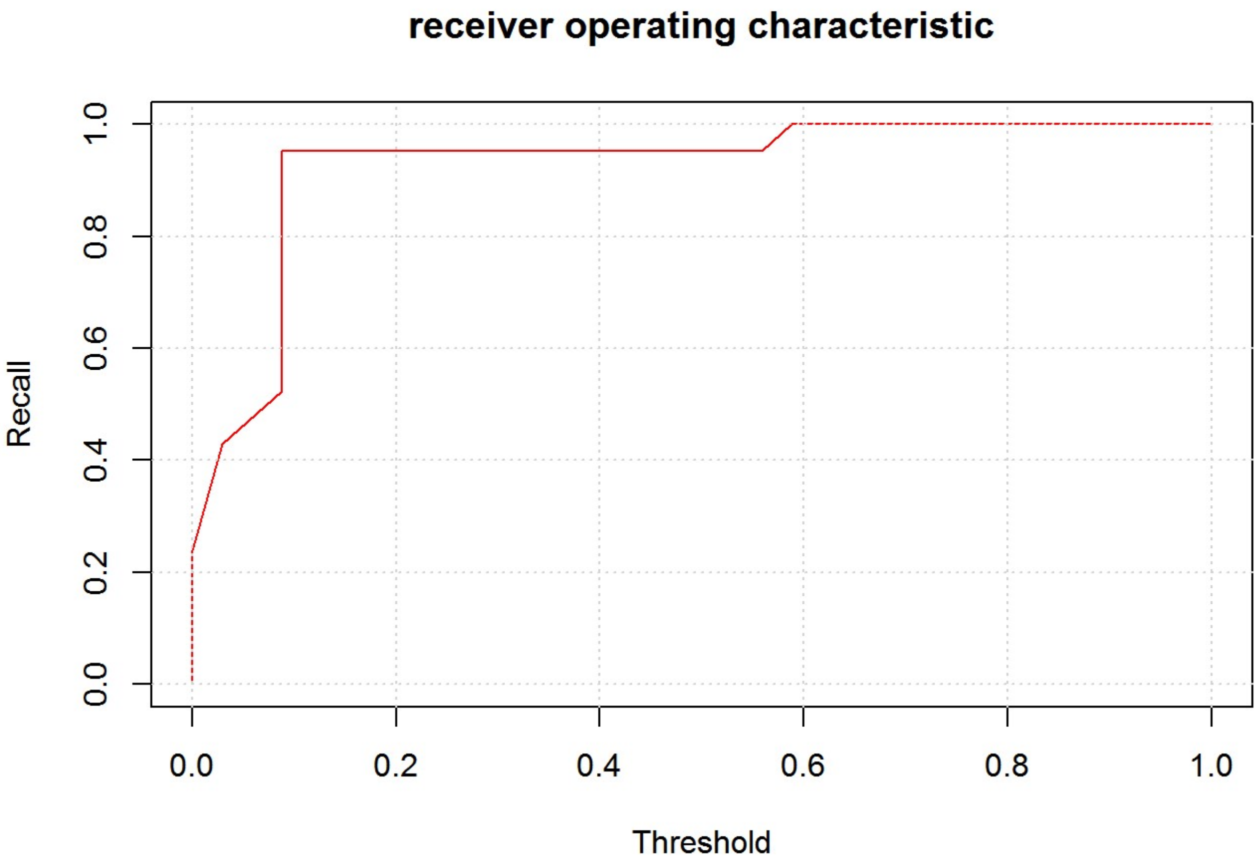


The plots below show the precision and recall. The threshold at 0.5, about 78% of the people diagnosed with heart disease really have it, and about 97% of the people who have it are diagnosed to have it.

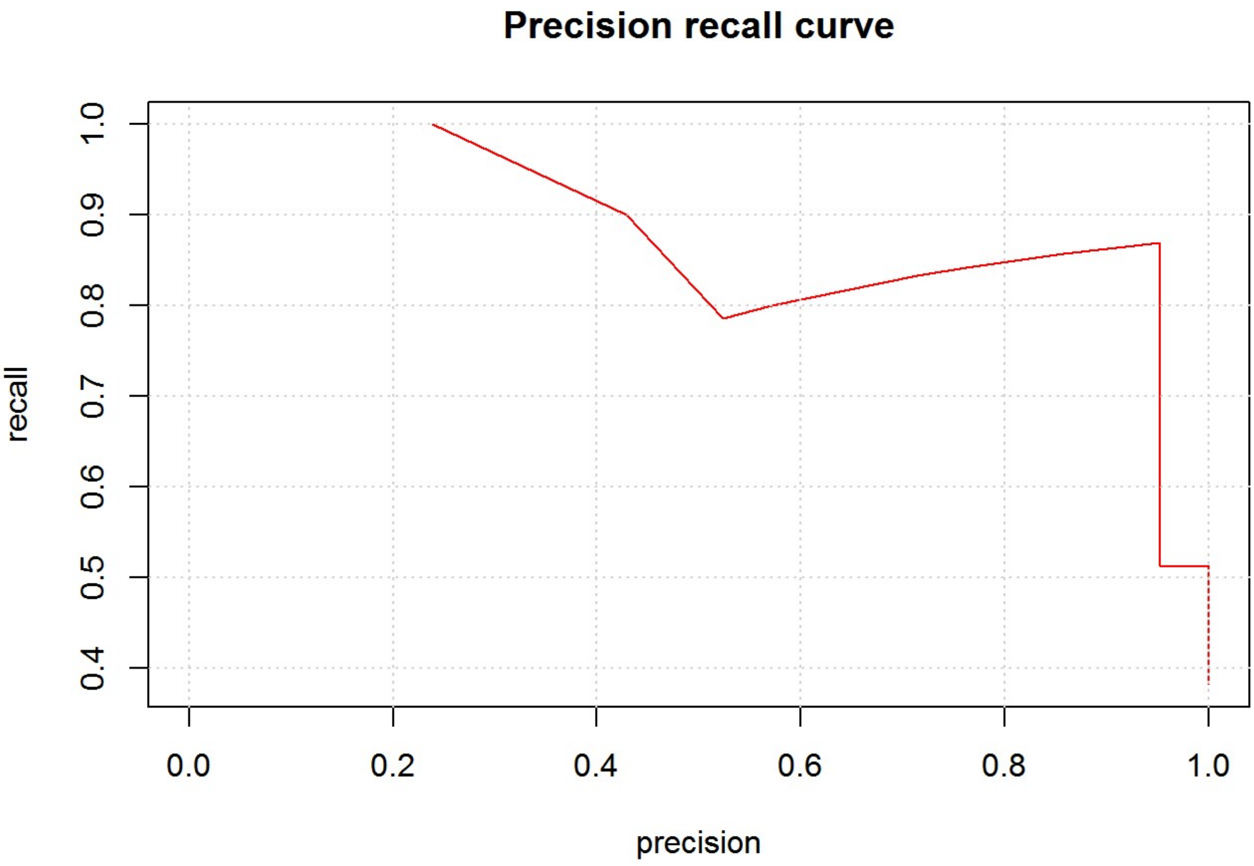


The ROC plot shows that the classifiers are working good.

What we are looking for is the graph like to be snugg up against the upper left corner.



The precision-recall plot shows us the tradeoffs we can get when getting a feel for precision versus recall.



## Model Two

# Female data exploration

The data has been split up into male and female. This model is testing the fmale data, but remember that there are twice as many males than females to we have about half the amount of data to test the females.

Looking at the best female features, chestpain seems to be an okay predictor, and thal, fluor, maxhr, sugar, and restbp seem to barley have an influence. The rest of the features are not correlated.

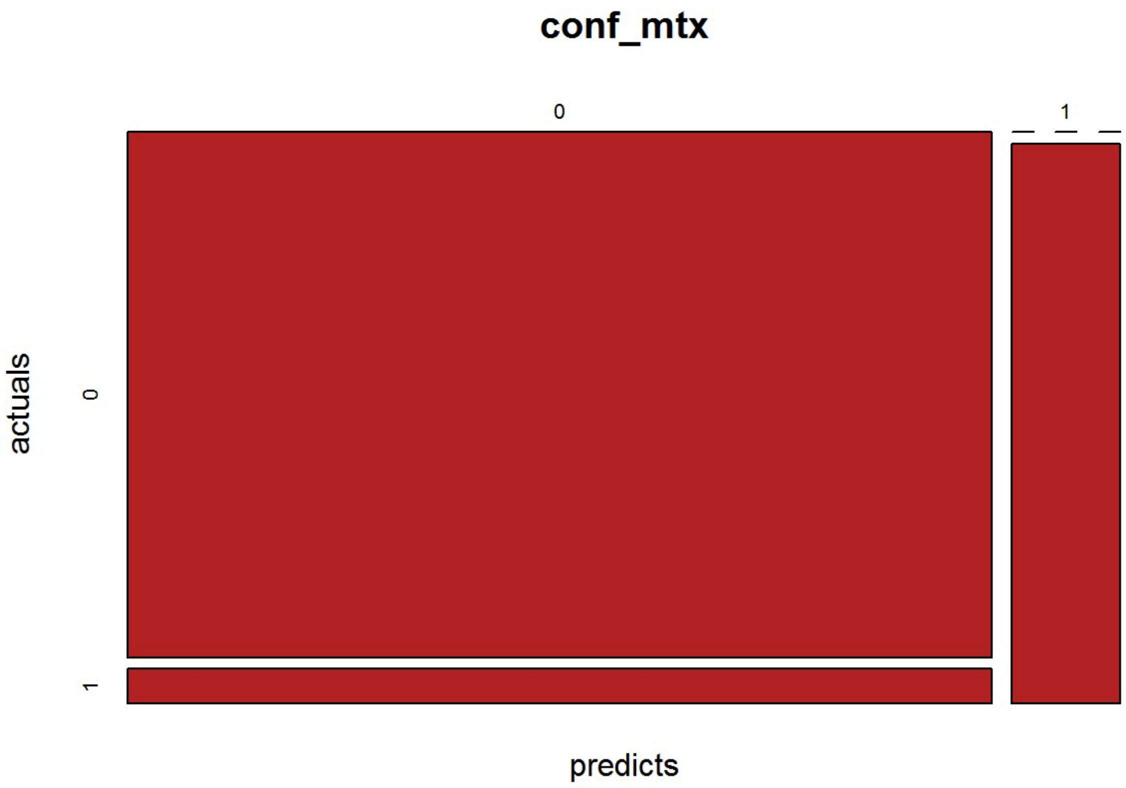
```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

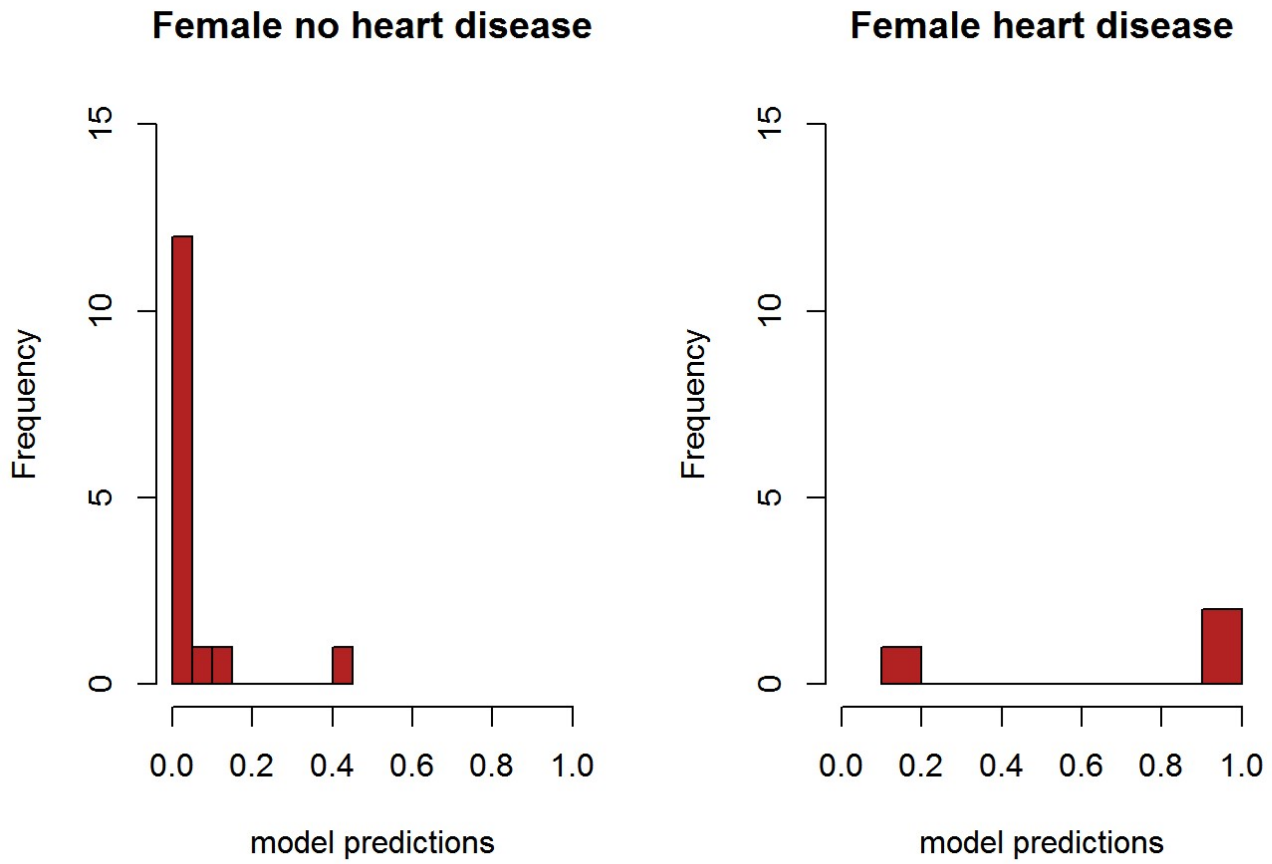
```

Call:
glm(formula = output ~ ., family = binomial, data = tr_dataf)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.48269  -0.18699  -0.00460  -0.00001   2.27620  

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)  
(Intercept) -88.28953   40.91153  -2.158   0.0309 *
age           0.23811    0.16212   1.469   0.1419  
sex                NA         NA      NA       NA  
chestpain     5.67915    2.76992   2.050   0.0403 *
restbp        0.17907    0.09701   1.846   0.0649 .
chol         -0.04156    0.03092  -1.344   0.1788  
sugar         6.63098    3.92960   1.687   0.0915 .
ecg           2.77774    1.76686   1.572   0.1159  
maxhr         0.13770    0.07451   1.848   0.0646 .
angina        0.53094    2.33336   0.228   0.8200  
dep           1.46076    1.08507   1.346   0.1782  
exercise      0.67527    2.04739   0.330   0.7415  
fluor         3.61127    2.15694   1.674   0.0941 .
thal          3.15615    1.75073   1.803   0.0714 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77.048  on 68  degrees of freedom
Residual deviance: 19.709  on 56  degrees of freedom
AIC: 45.709

Number of Fisher Scoring iterations: 10
```

The predicted no heart disease looks right on point with woman that didn't have heart disease, but the recall was bad. Only 50% of women told had heart disease actually had it.

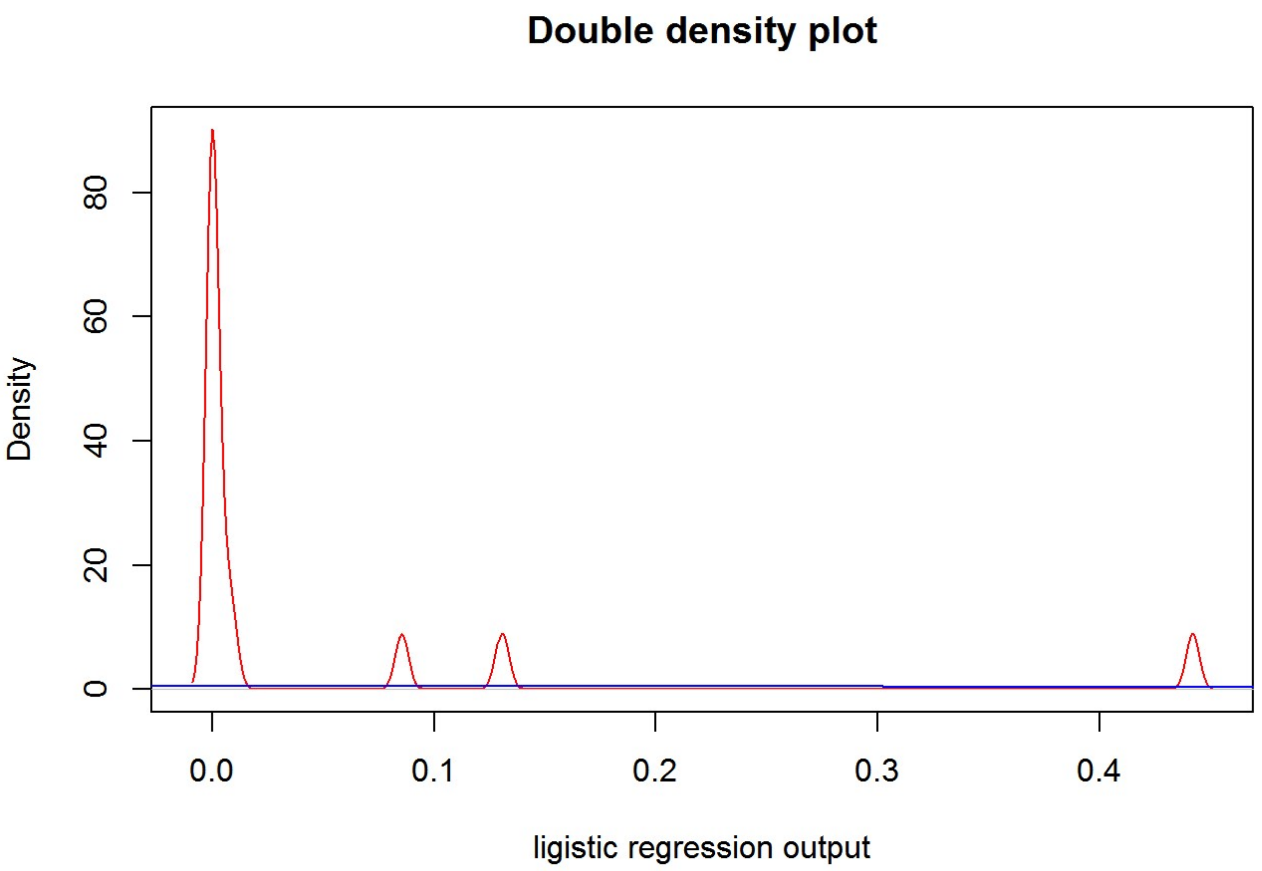
```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
        actuals
predicts  0  1
       0 15  1
       1  0  2
```

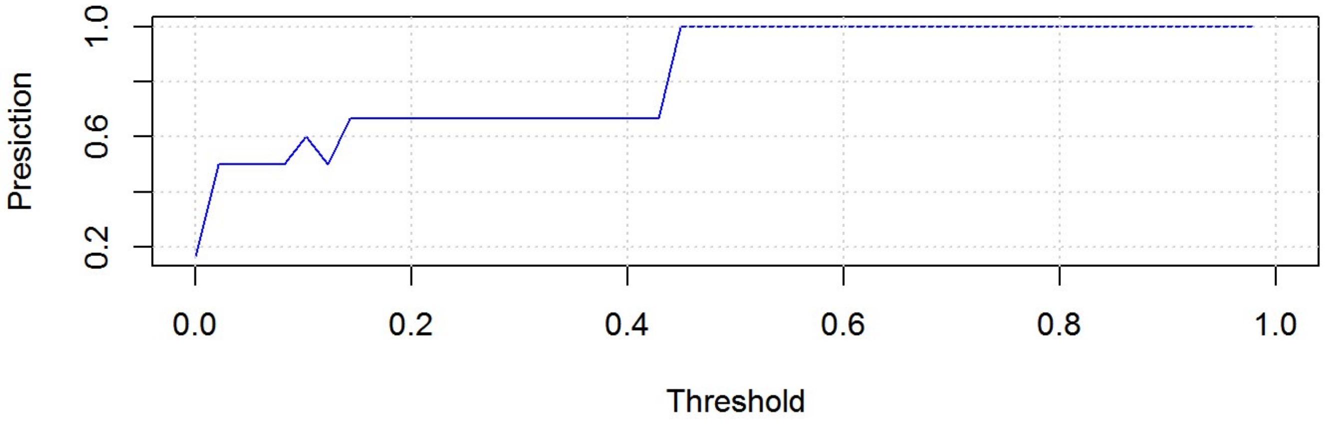

conf_mtx
0
1
actuals
0
1
predicts

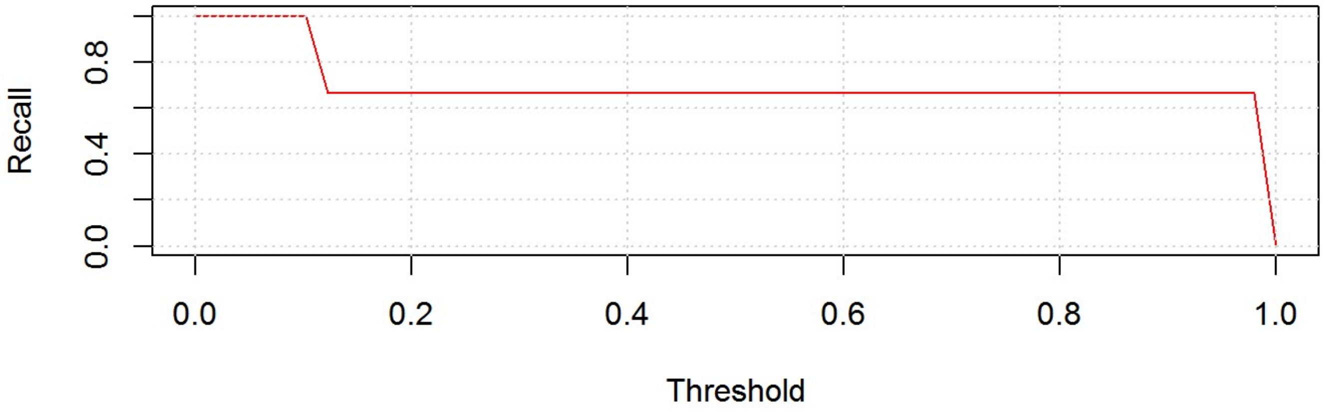


Double density plot for this model:



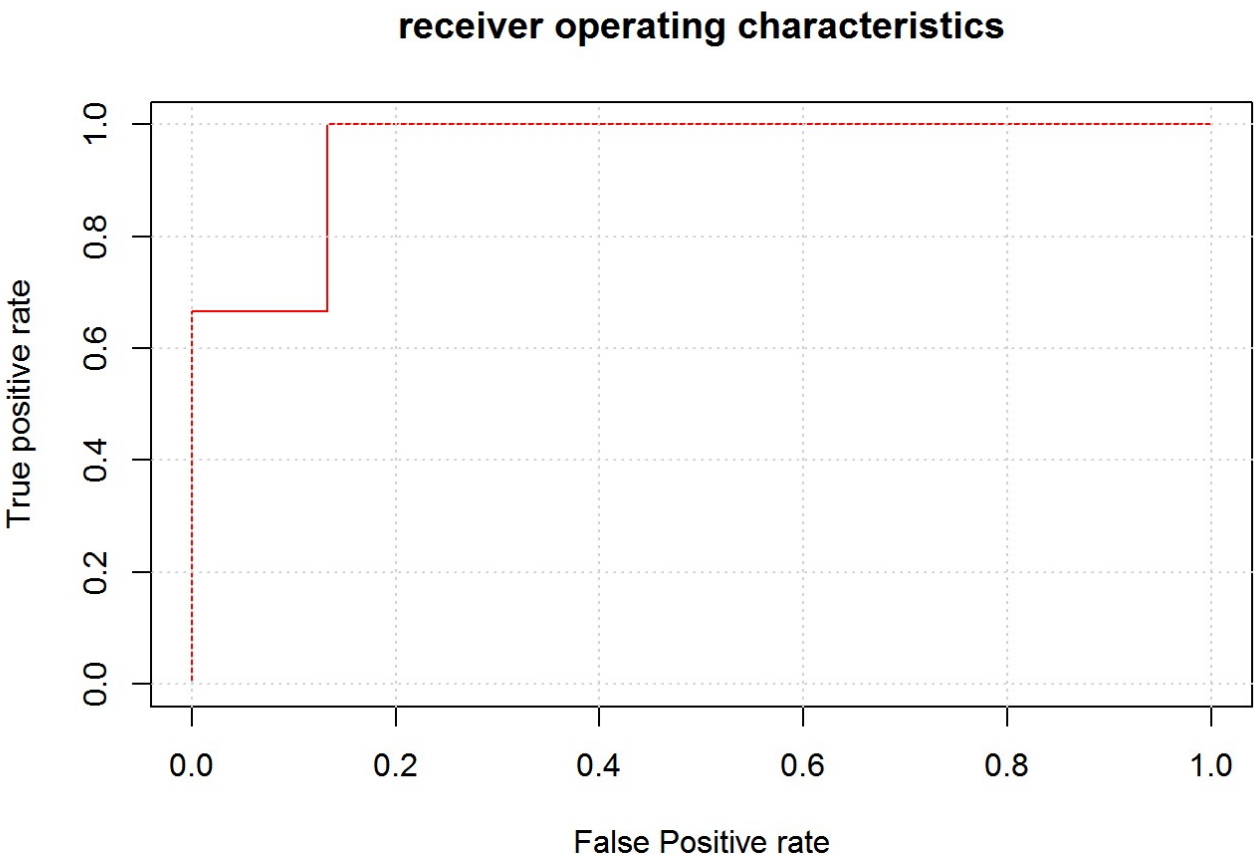
These plots show precision and recall by threshold value. With the threshold at 0.5, about 99% of the women diagnosed with heart disease really have it, and about 65% of the women who have it are diagnosed to have it.





The ROC plot looks like the classifier is working well.

The receiver operating characteristic (ROC) plot gives an overall idea of how well the classifier is working. The curve for a perfect classifier would hug the left and top edges of the plot. The curve for a classifier that makes random decisions would be a diagonal line from the lower-left to the upper-righ



This time we are going to use all the features but only for the males in the data.

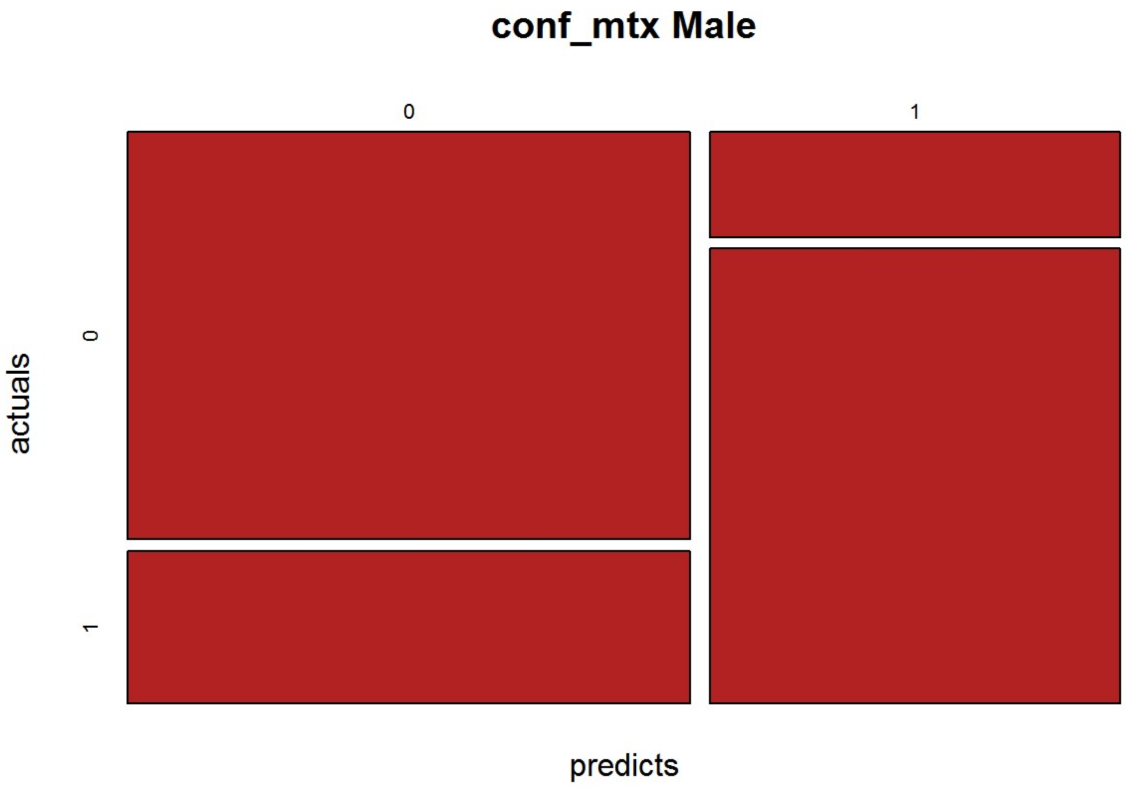
Looking at the best male features, chestpain and thal seems to be a good predictor, fluor is the highest rated, maxhr and dep seem to barley have an influence. The rest of the features are not correlated.

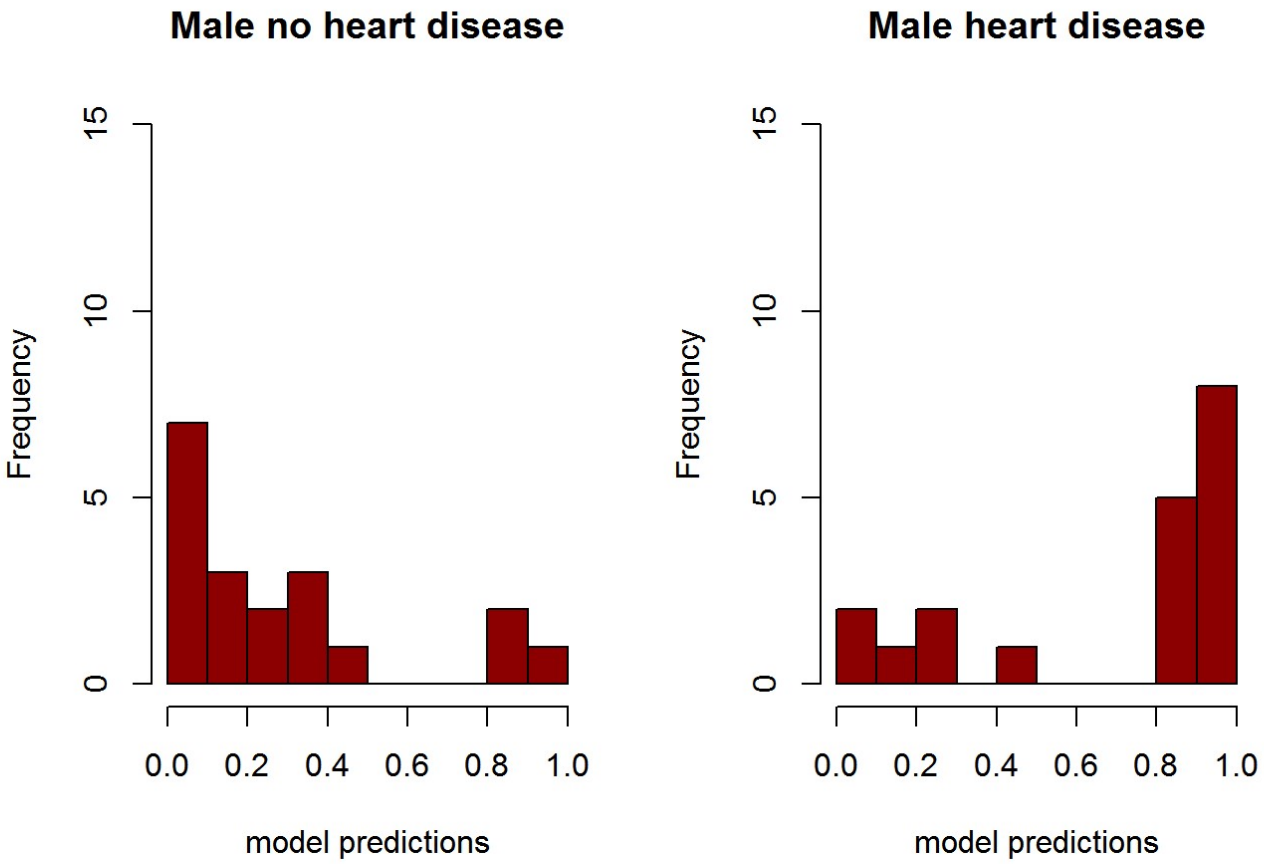
```
Call:
glm(formula = output ~ ., family = binomial, data = tr_dataM)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.77796  -0.50656   0.08837   0.43492   2.39932  

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.852848   4.342740  -1.117 0.263797    
age         -0.046800   0.037723  -1.241 0.214754    
sex                NA         NA      NA       NA    
chestpain    0.810419   0.285083   2.843 0.004473 ** 
restbp       0.012156   0.015968   0.761 0.446487    
chol         0.010957   0.007251   1.511 0.130784    
sugar       -0.464753   0.836087  -0.556 0.578302    
ecg          0.327679   0.271814   1.206 0.228000    
maxhr       -0.024790   0.014970  -1.656 0.097721 .  
angina       0.508314   0.610358   0.833 0.404950    
dep          0.542457   0.292928   1.852 0.064048 .  
exercise     0.374688   0.511044   0.733 0.463447    
fluor        1.366376   0.385088   3.548 0.000388 ***
thal         0.406192   0.143667   2.827 0.004694 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 199.015  on 144  degrees of freedom
Residual deviance:  97.711  on 132  degrees of freedom
AIC: 123.71

Number of Fisher Scoring iterations: 6
```

# compute confusion matrix for men

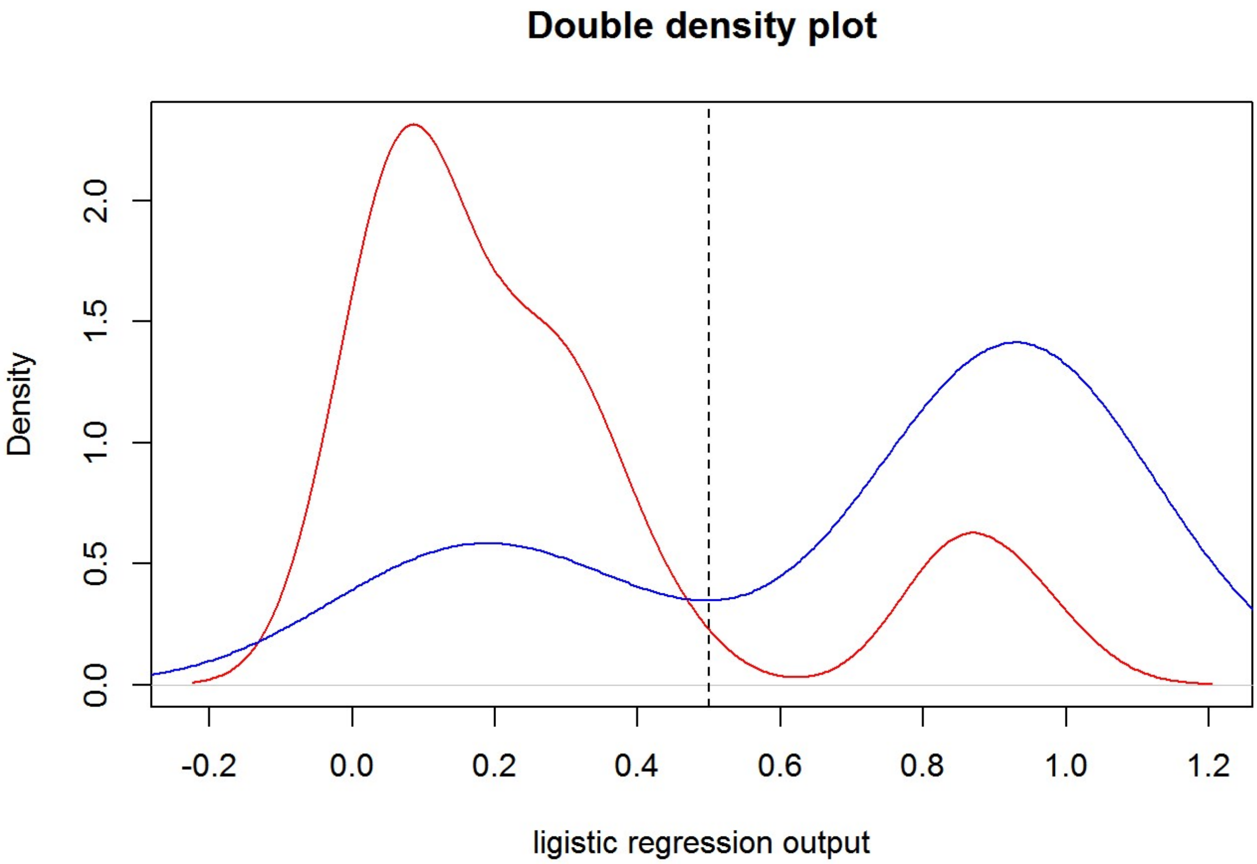
```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
        actuals
predicts  0  1
       0 16  6
       1  3 13
```

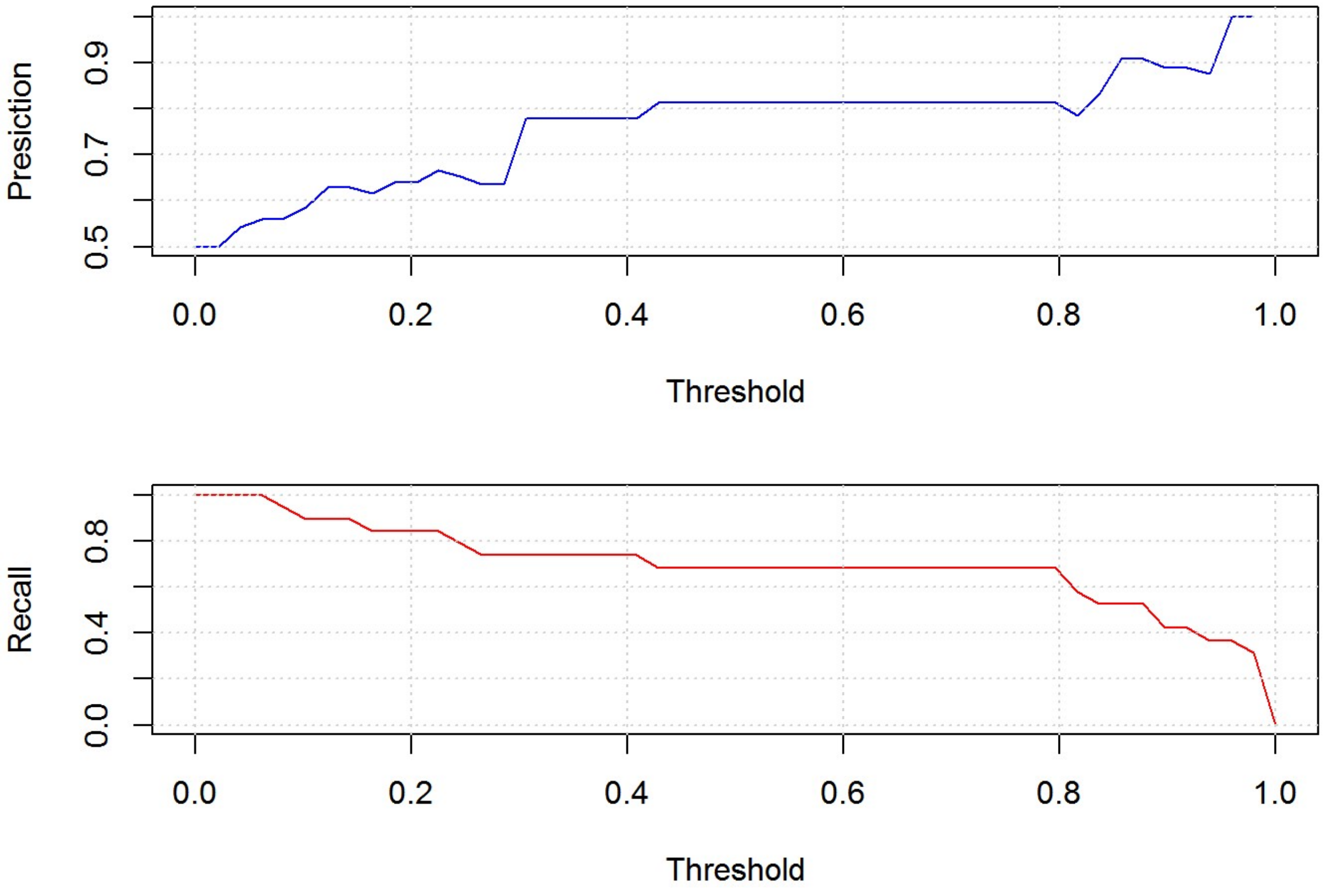

conf_mtx Male
0
1
actuals
0
1
predicts



Double density plot for men:

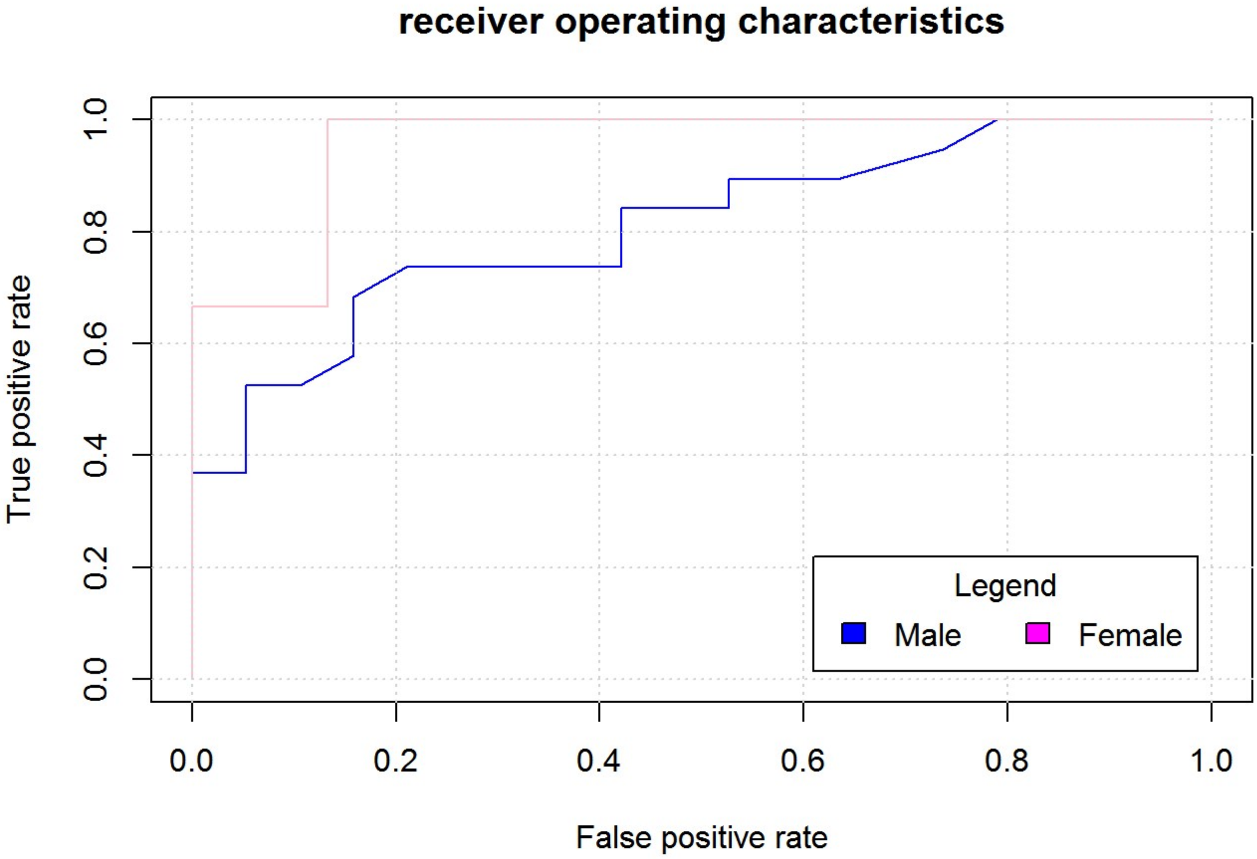


These plots show precision and recall by threshold value. With the threshold at 0.5, about 80% of the men diagnosed with heart disease really have it, and about 85% of the men who have it are diagnosed to have it.



The ROC plot suggests that the classifier is working reasonably well.

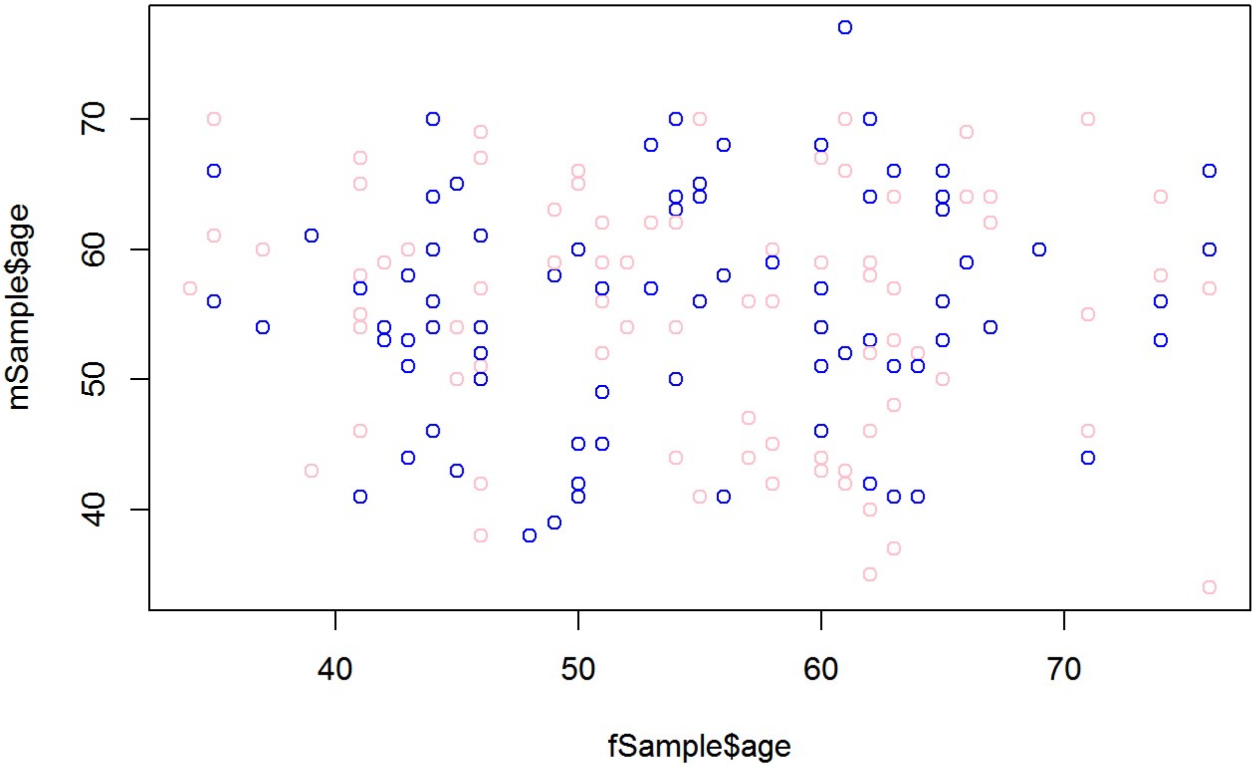
It looks like the female Model is out performing the male model because the pink line is hugging the top left corner more than the blue line.



## model 3

In this model I'm going to try to make the Data Less Bias between Gender. Female and male data is going to be sampled to match the origanal sample data of both genders which is 170.

Just wanted to see how the ages were spread out in terms of male and female from the sampling. They look pretty even with some outliers on the outer limits.



# female data exploration from the sampling

All the features are have “***" I think something could be wrong with the sampling…

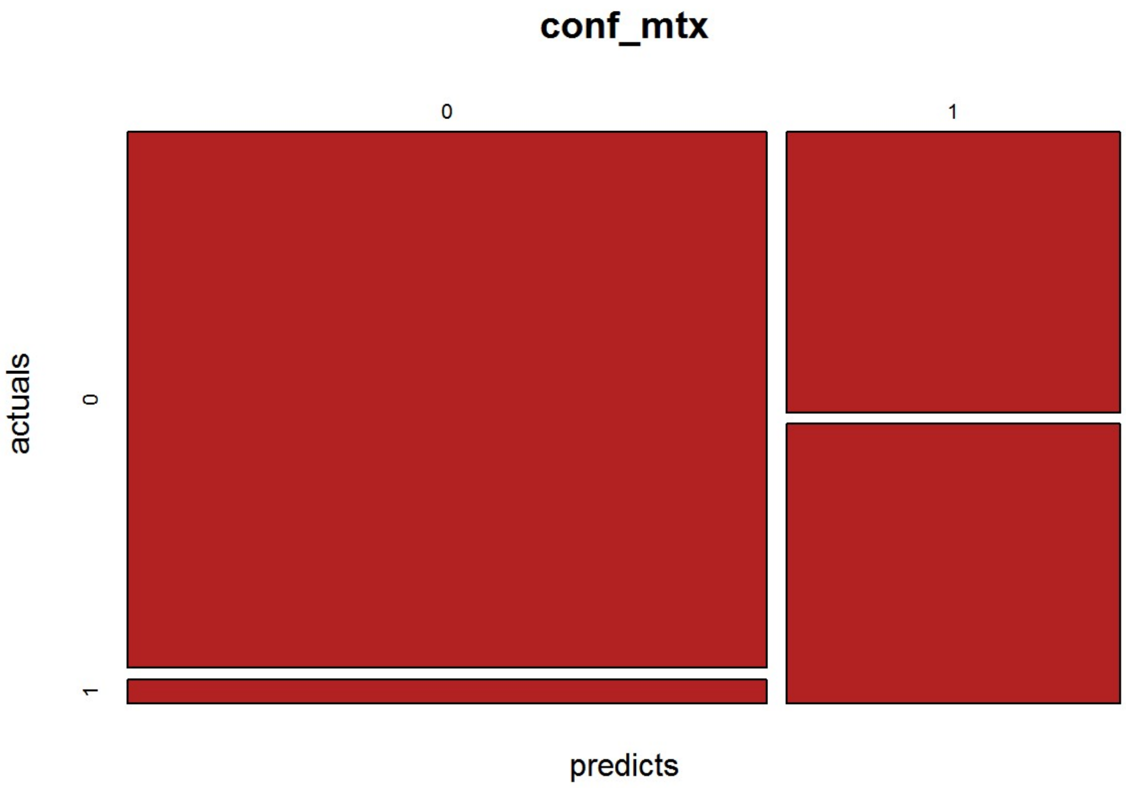
```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```

Call:
glm(formula = output ~ ., family = binomial, data = tr_dataf)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
 -8.49    0.00    0.00    0.00    8.49  

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error   z value Pr(>|z|)    
(Intercept) -1.038e+16  1.046e+08 -99188643   <2e-16 ***
age          2.057e+13  8.004e+05  25697298   <2e-16 ***
sex                 NA         NA        NA       NA    
chestpain    7.777e+14  8.019e+06  96981709   <2e-16 ***
restbp       2.823e+13  3.971e+05  71089167   <2e-16 ***
chol        -3.954e+12  1.182e+05 -33454040   <2e-16 ***
sugar       -1.257e+14  2.261e+07  -5557357   <2e-16 ***
ecg          4.838e+14  6.506e+06  74368666   <2e-16 ***
maxhr        3.732e+12  3.936e+05   9483485   <2e-16 ***
angina       5.722e+14  1.640e+07  34891134   <2e-16 ***
dep          2.260e+14  8.000e+06  28245440   <2e-16 ***
exercise     2.340e+14  1.467e+07  15950520   <2e-16 ***
fluor        9.536e+14  9.386e+06 101590328   <2e-16 ***
thal         3.109e+14  5.140e+06  60473704   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 145.48  on 134  degrees of freedom
Residual deviance: 576.70  on 122  degrees of freedom
AIC: 602.7

Number of Fisher Scoring iterations: 24
```

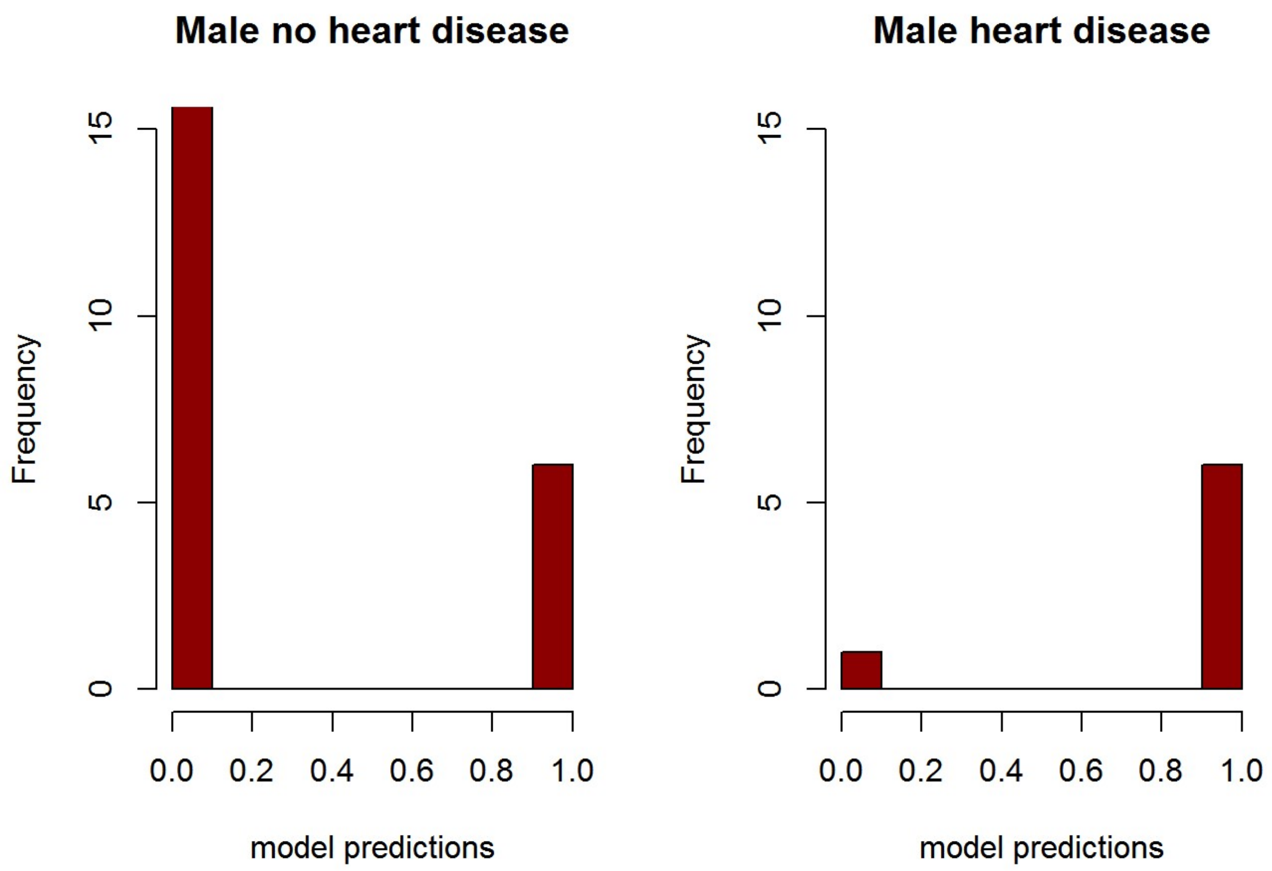
The predictions in the matrix look good!

```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
ifelse(type == : prediction from a rank-deficient fit may be misleading
```

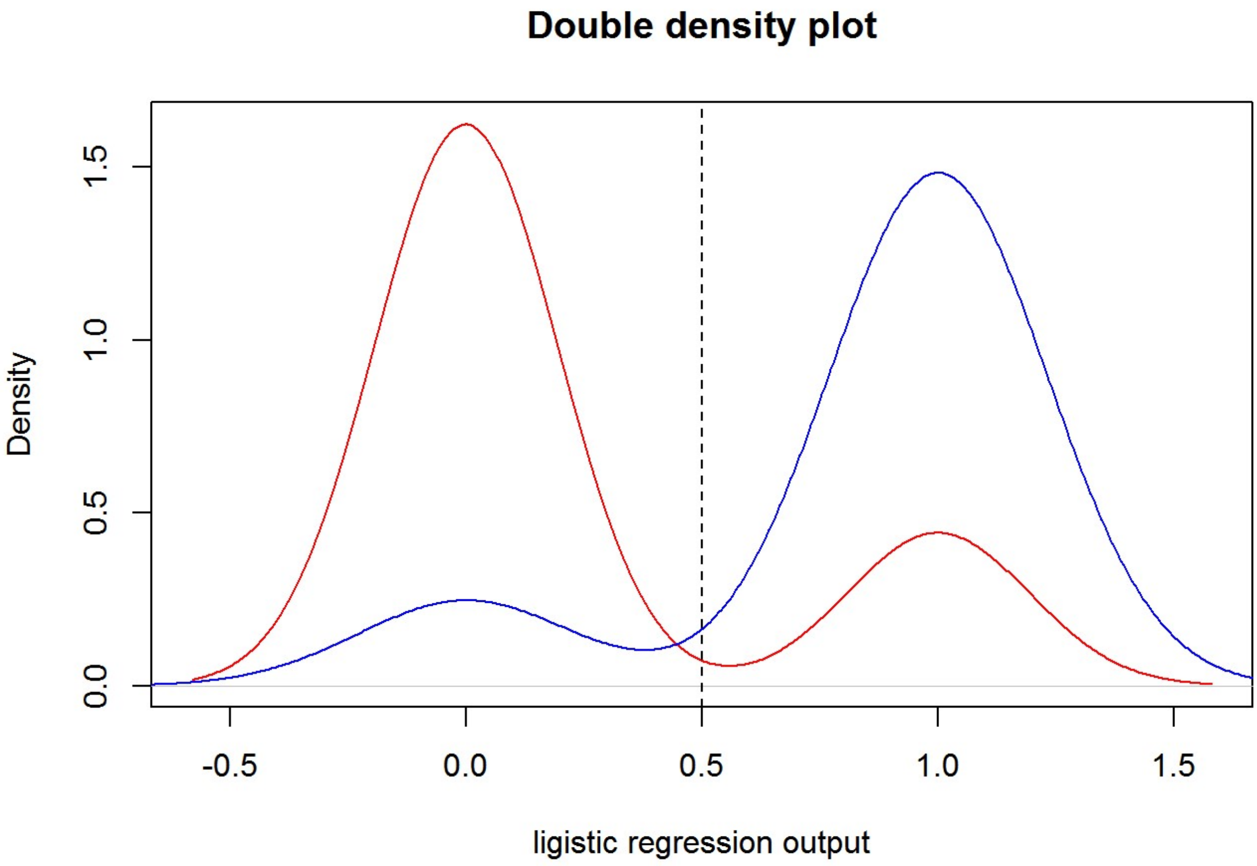
```
        actuals
predicts  0  1
       0 22  1
       1  6  6
```



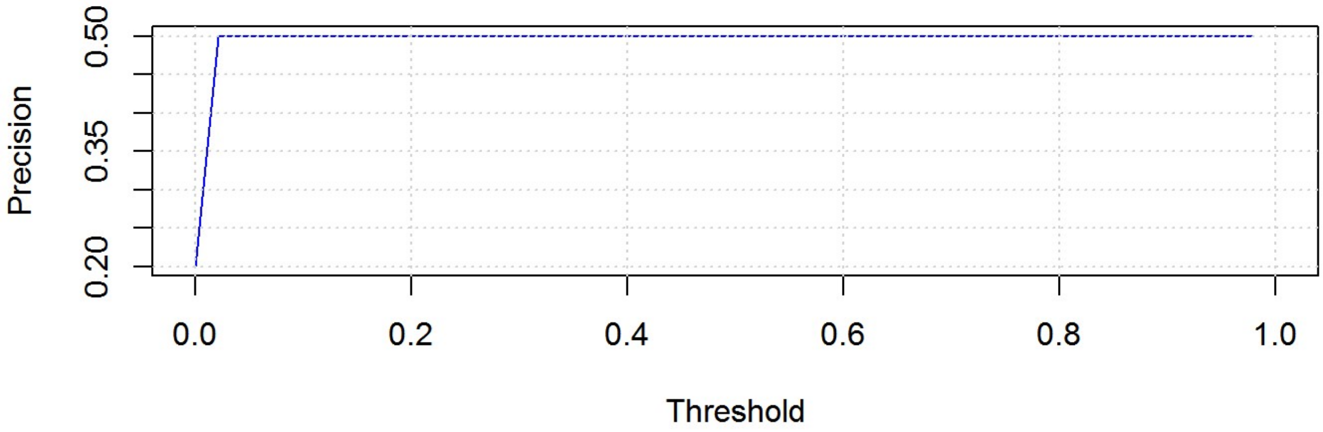
The histigrams don’t look as stable as the last ones. From the looks of the left has some values that are closer to one and the right has values in the 0 range which could mean lower recall and precistion.

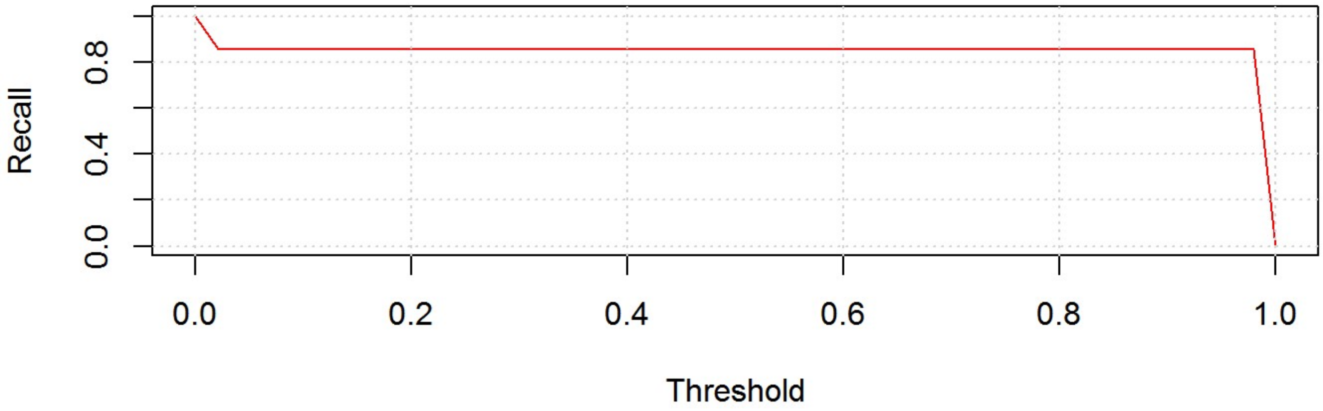


The double density plots for males sampled:

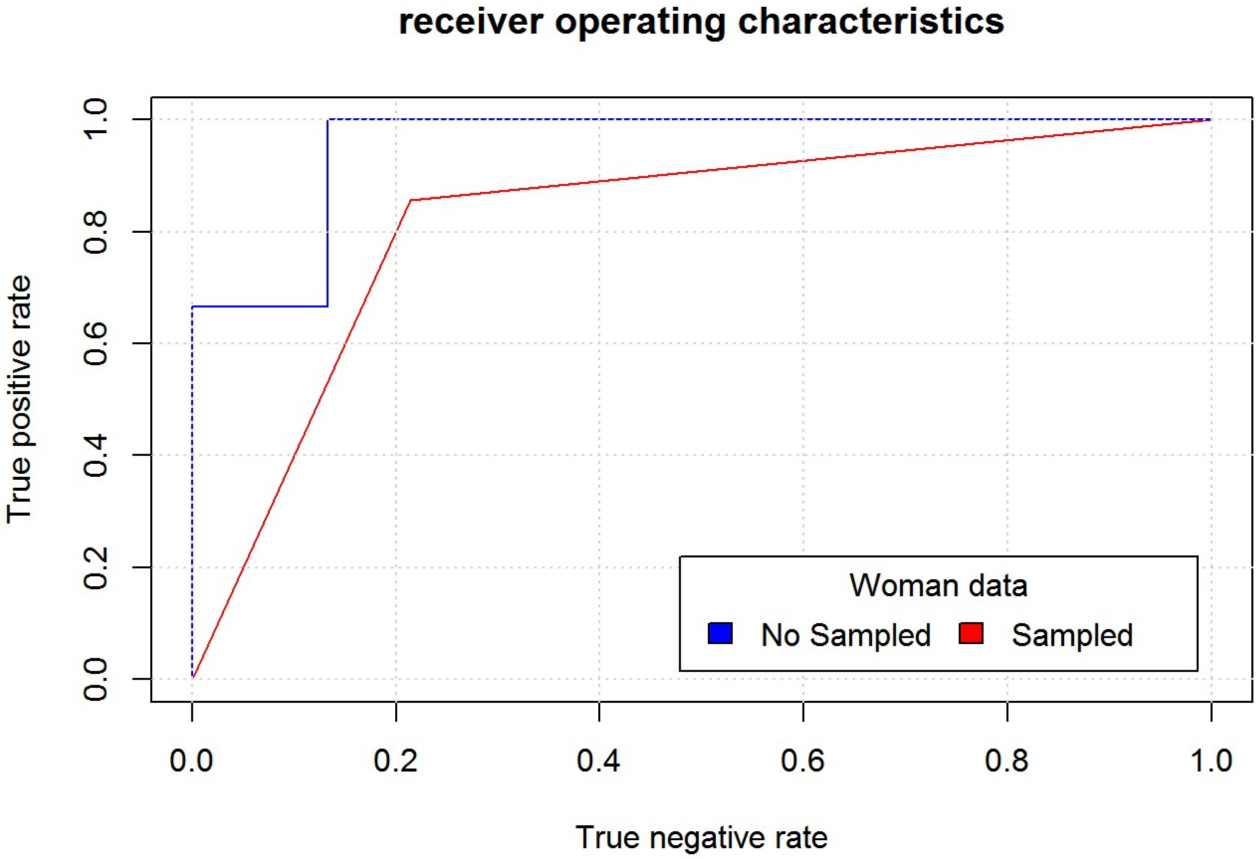


The presistion doesn’t look as good as the last model with the success rate sitting at 50%, the recall is about 82% at a 0.5 threshold with is pretty good.





The data shows that sampling the data from the female data set hurt the models predictions noted by the red line being inside of the blue line.



## Male data exploration sampled

This time we are going to use all the features but only for the male’s sampled data.

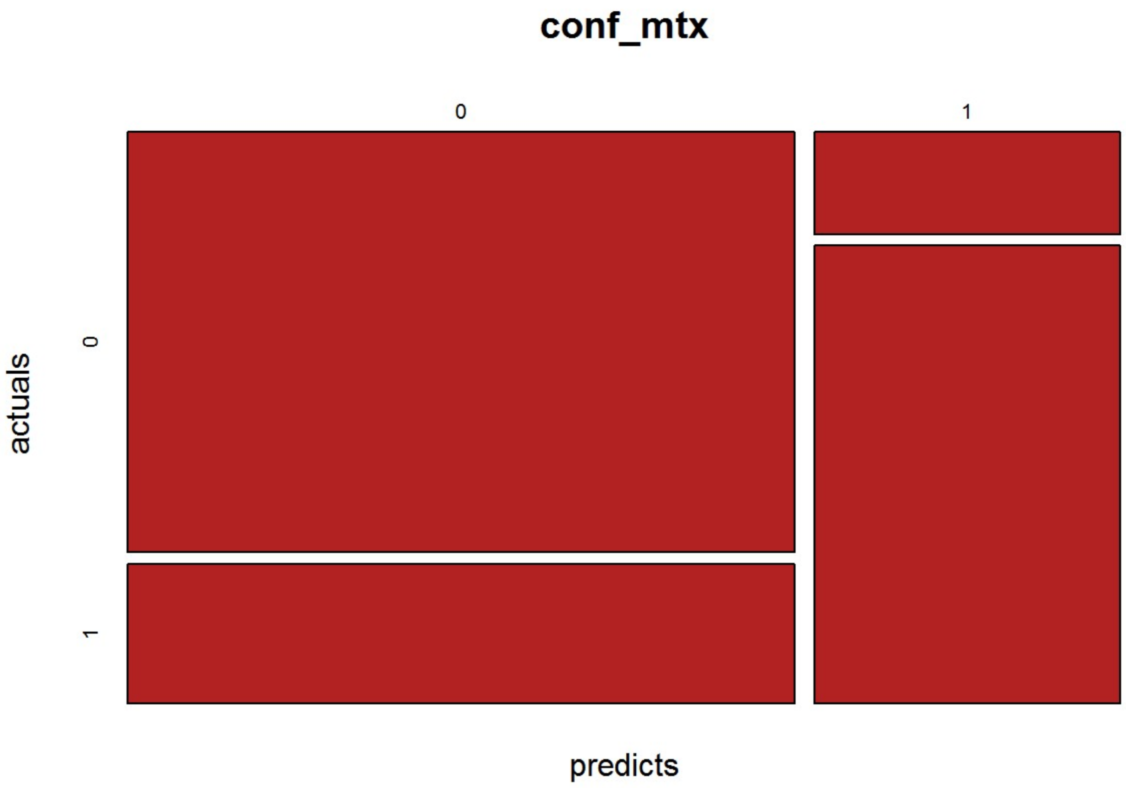
Looking at the best male features. chestpain is good, as well as dep. fluor and thal are almost not correlated, which is different from out last model where they had a strong correlation The rest of the features are not correlated.

```
Call:
glm(formula = output ~ ., family = binomial, data = tr_dataM)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9066  -0.4406   0.1294   0.5688   2.4673  

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)   
(Intercept) -4.166638   4.566195  -0.912  0.36151   
age         -0.035689   0.034214  -1.043  0.29690   
sex                NA         NA      NA       NA   
chestpain    0.764913   0.283841   2.695  0.00704 **
restbp      -0.004365   0.018444  -0.237  0.81292   
chol         0.015644   0.007681   2.037  0.04169 * 
sugar       -0.112078   0.831838  -0.135  0.89282   
ecg          0.478152   0.280945   1.702  0.08877 . 
maxhr       -0.020764   0.016484  -1.260  0.20780   
angina       1.205488   0.576326   2.092  0.03647 * 
dep          0.950422   0.321874   2.953  0.00315 **
exercise    -0.306453   0.560870  -0.546  0.58480   
fluor        0.806936   0.445273   1.812  0.06995 . 
thal         0.295193   0.150748   1.958  0.05021 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 186.787  on 134  degrees of freedom
Residual deviance:  96.174  on 122  degrees of freedom
AIC: 122.17

Number of Fisher Scoring iterations: 5
```

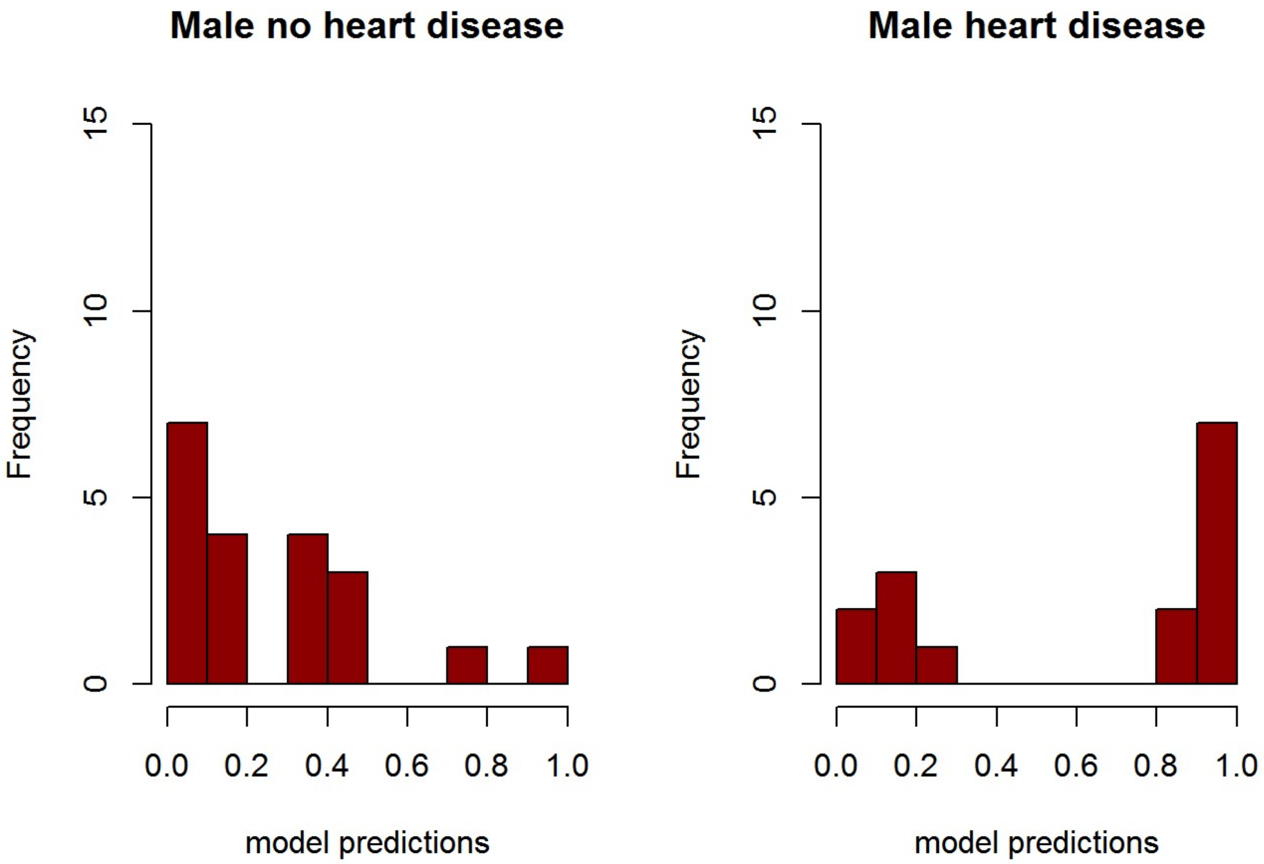
The maxtrix data still looks good. The predicted and actuals decent.

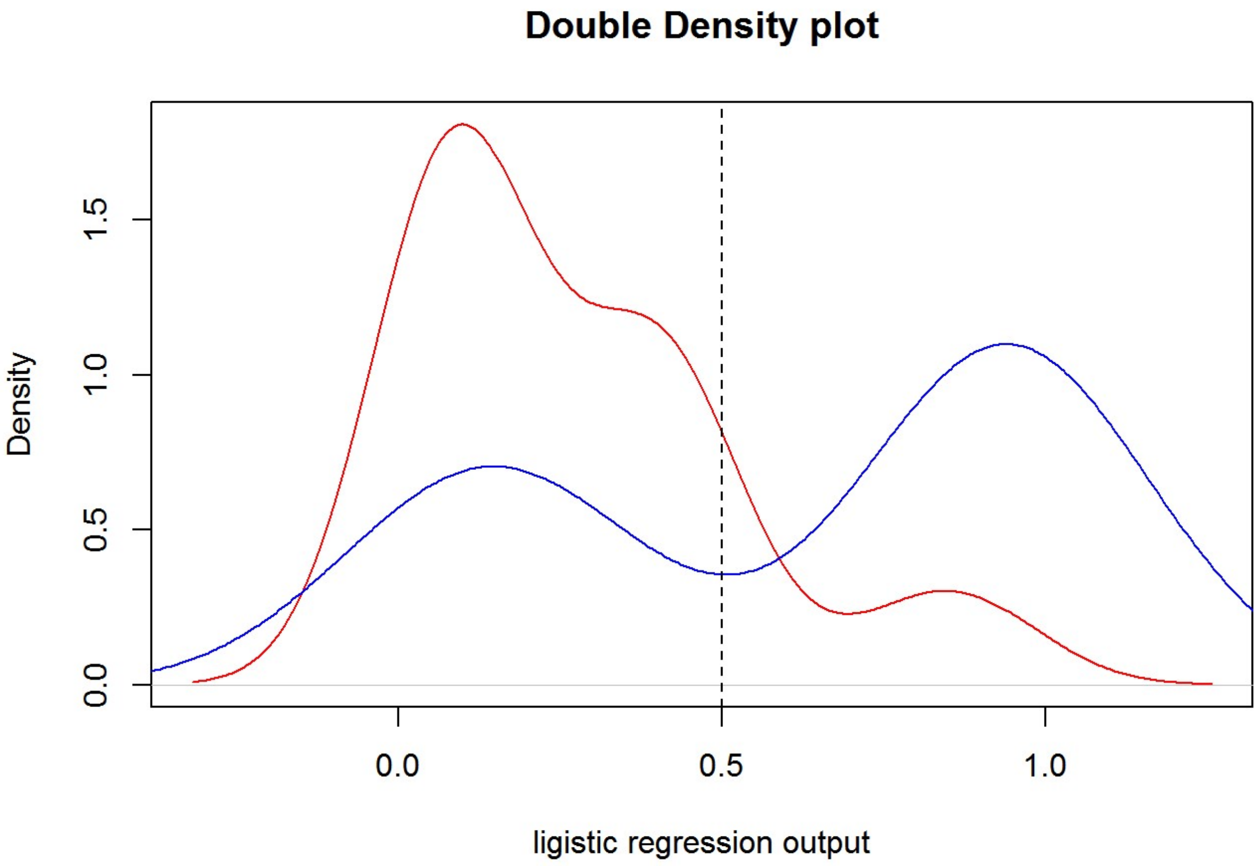
```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
ifelse(type == : prediction from a rank-deficient fit may be misleading
```

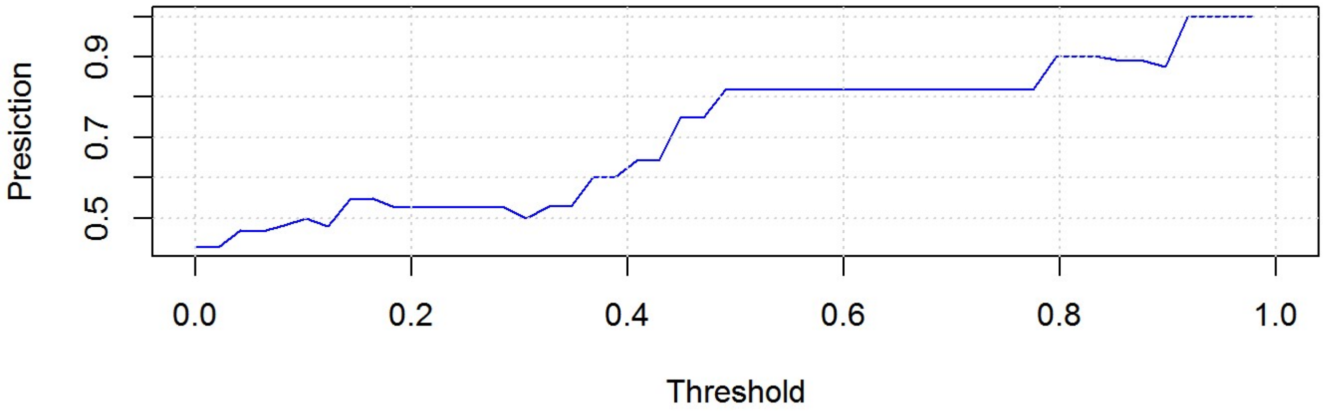
```
        actuals
predicts  0  1
       0 18  6
       1  2  9
```

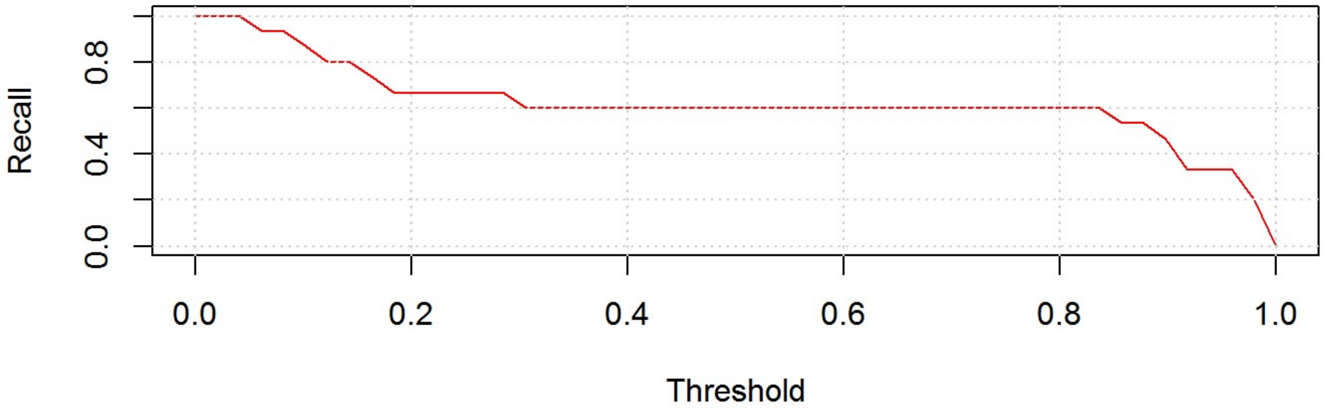


Density plots don’t look as good as either of the other models. The numbers seem to be all over the place in both plots.

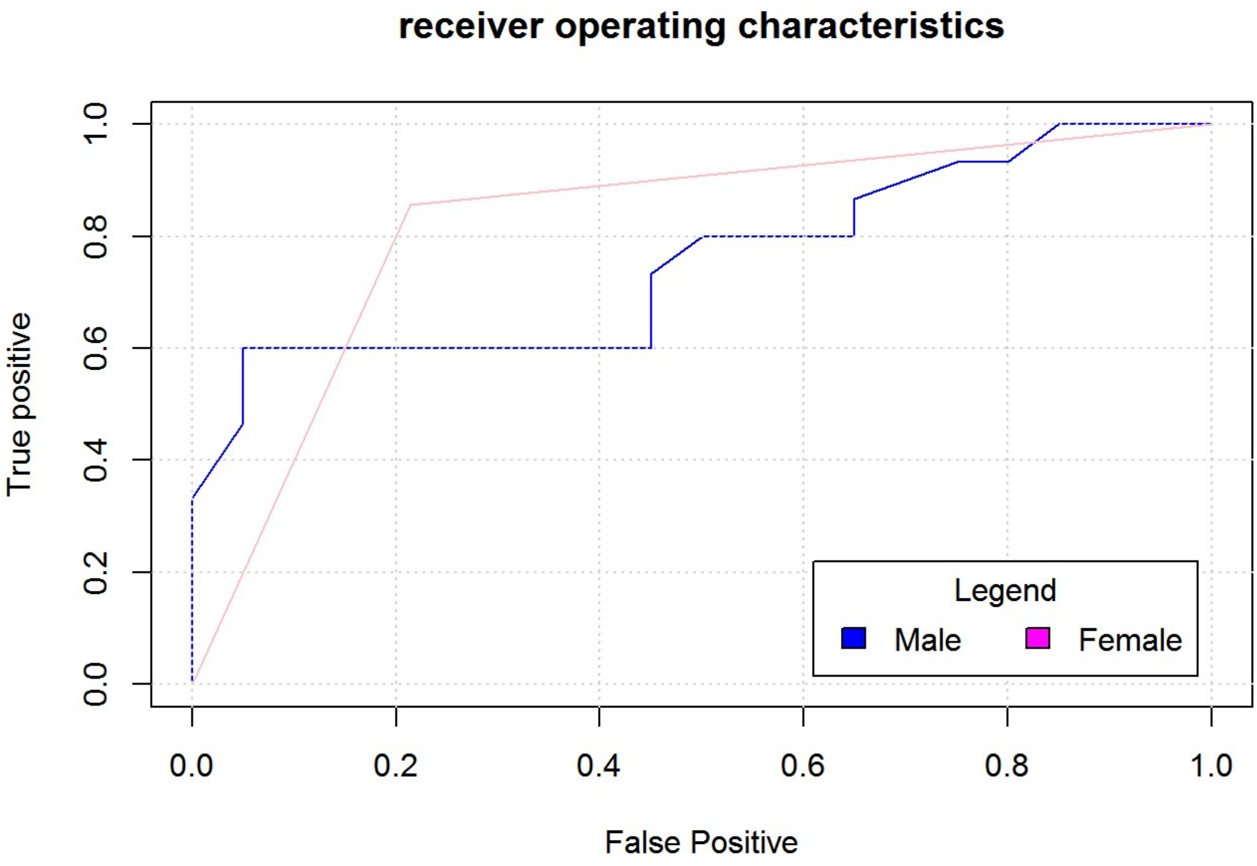

Male no heart disease
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
model predictions
Male heart disease
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
model predictions


Double Density plot
Density
0.0
0.5
1.0
1.5
0.0
0.5
1.0
ligistic regression output

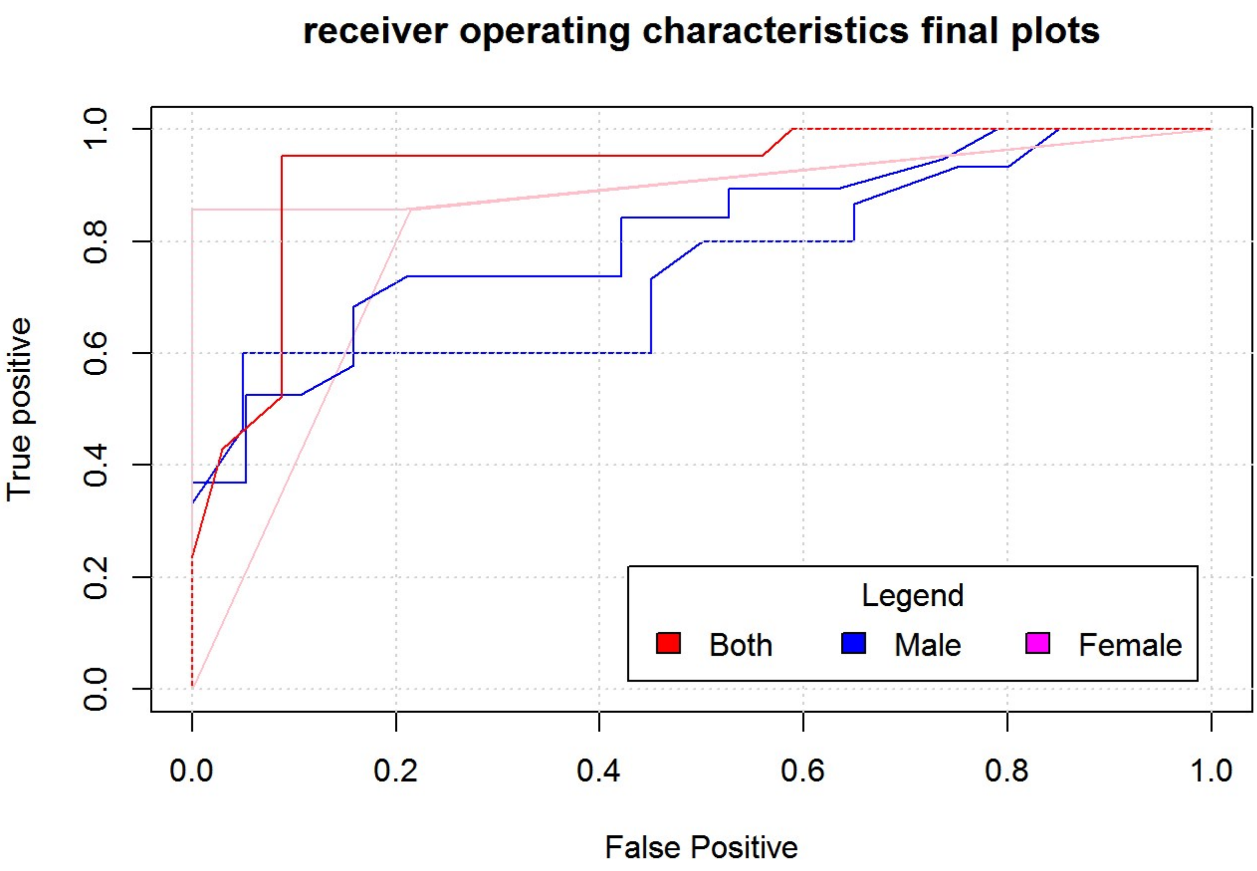




It’s apparent that the female data preformed better than the male data in both situations. Althougth with the sampled data it was closer together.



Over all, from the graph below, the data that preformed the best was the mixed gender data with all the feactures. The data with just females preformed good too! Better than the males. Sampling the male and female data just hurt the model. As you can see below if preformed the lowest with the tures.


receiver operating characteristics final plots
True positive
0.0
0.2
0.4
0.6
0.8
1.0
False Positive
0.0
0.2
0.4
0.6
0.8
1.0
Legend
🟥 Both
🟦 Male
🟪 Female