Text_Analytics

Ryan M. Allen

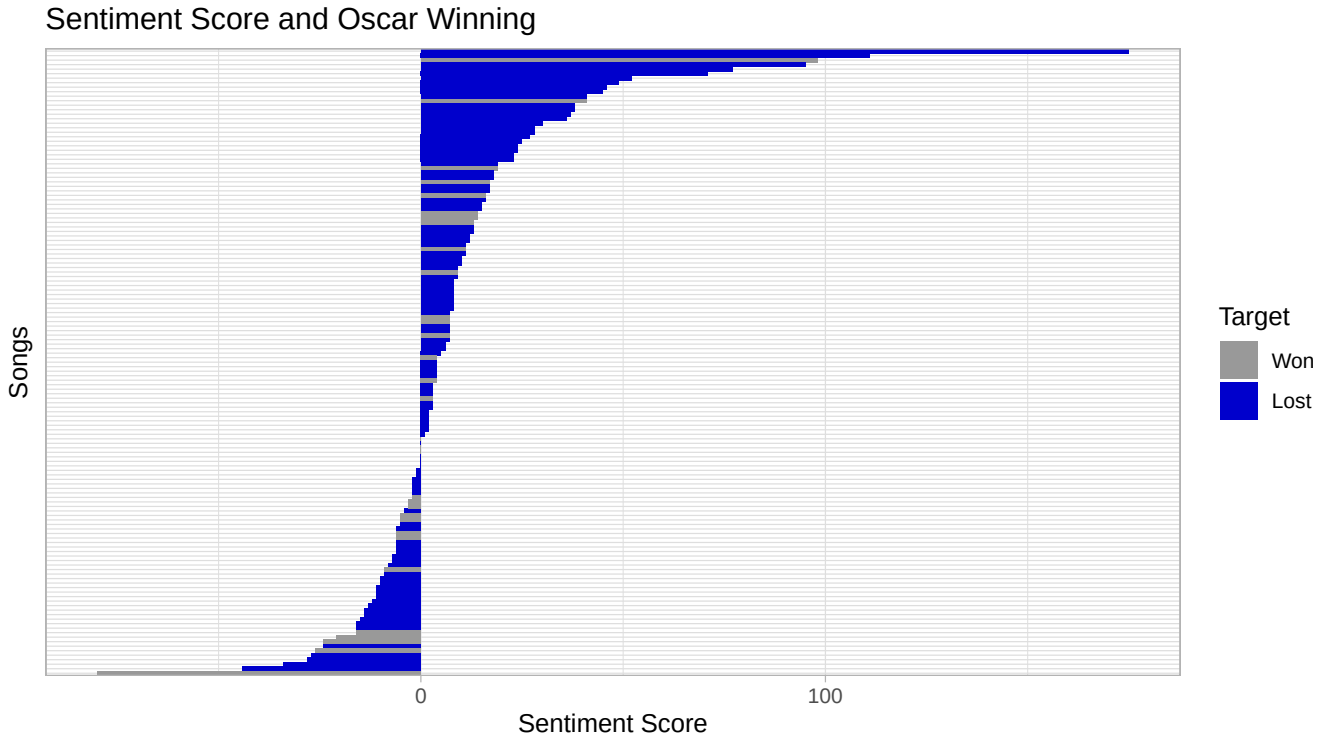
March 17, 2020

Data Collection and Cleansing

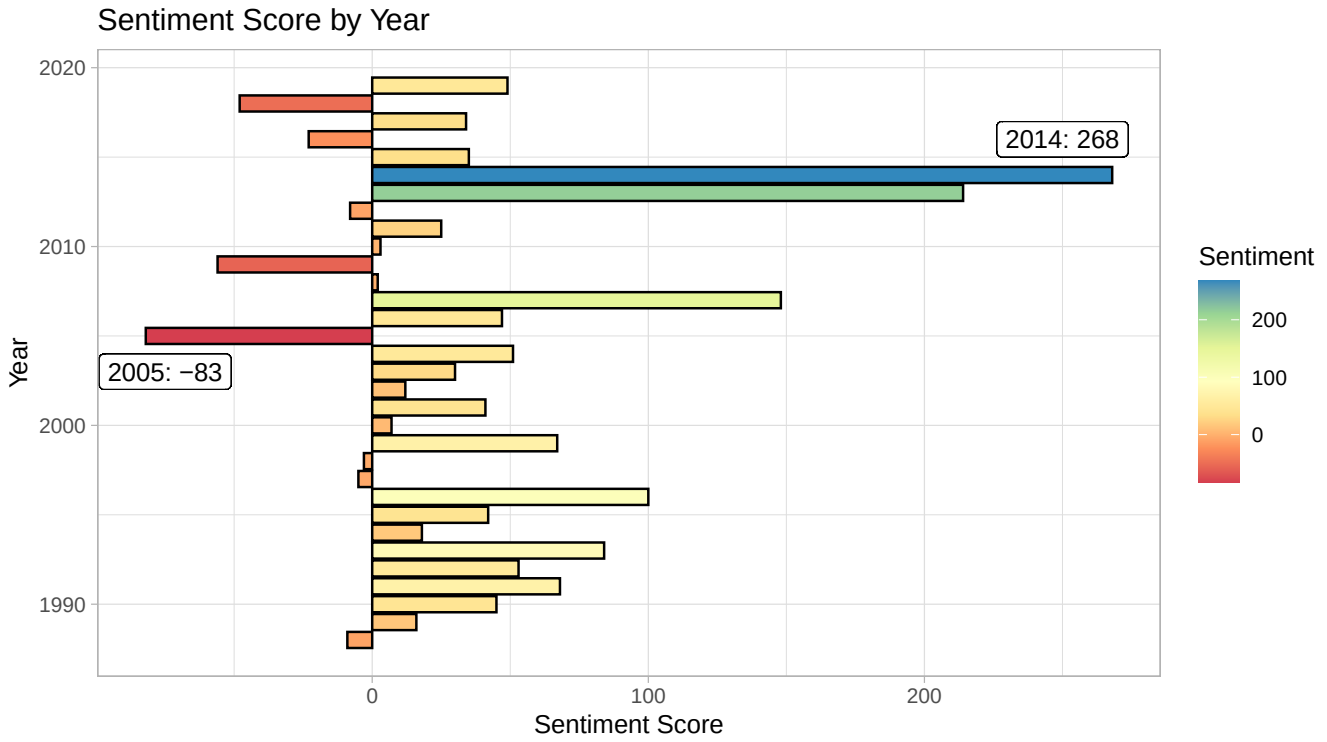
I collected songs that were nominated for the Best Orignal Song. I have a column for the name the song, the year, the Spotify URI (a unique song identifier), the lyrics, and whether or not the song won (1 for won, 0 for nominated but lost). I collected data through 1990. I then used the Spotify API to collect the songs features and then joined that back to the original dataset.

Text Analytics

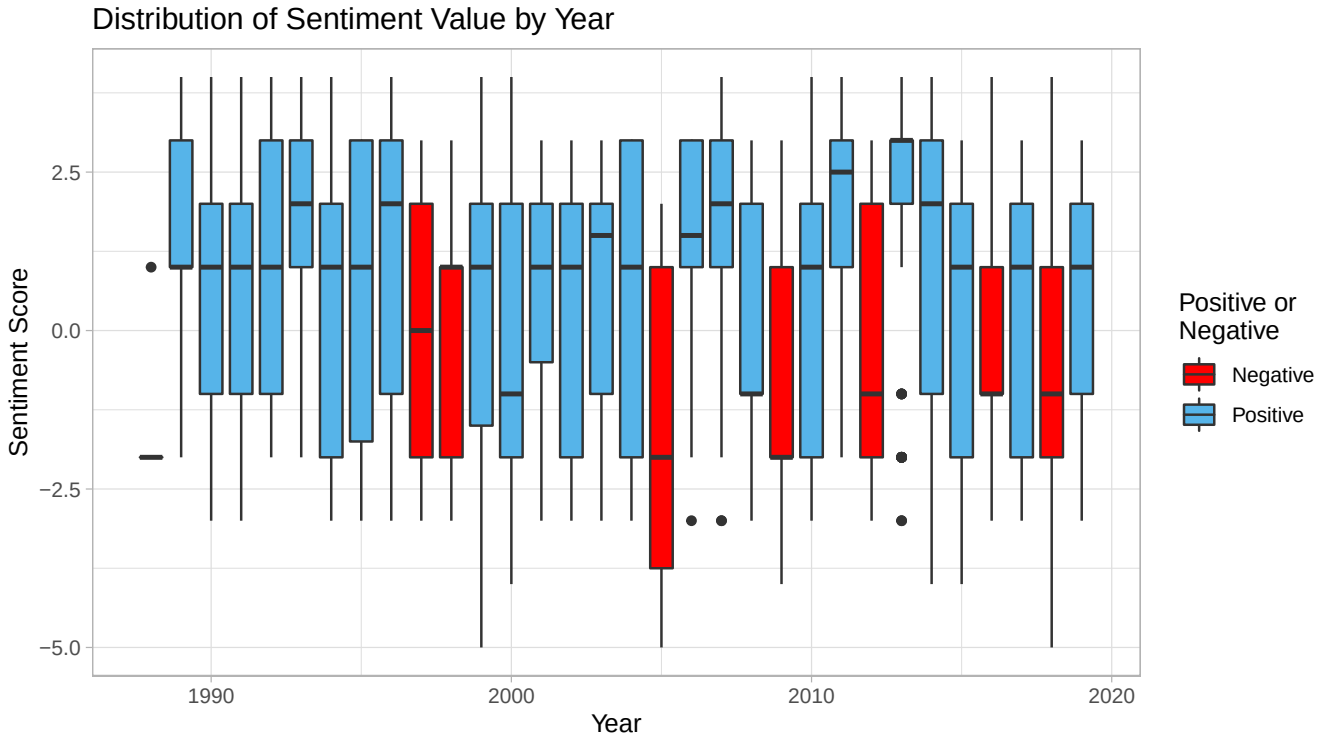
I used two different sentiment libraries to one is the AFINN package from Finn Arup Nielsen and the other is the nrc package from Saif Mohammad and Peter Turney. The AFINN Package assigns a number -4 through 4 to a every word in its dictionary and then for the plots below, I have summed the sentiment scores to get a total sentiment score, the higher the number the more positive the song lyrics are. THe nrc package assigns a feeling/emotion to each word in its library, things like fear, surprise, joy, disgust, anticipation, anger, sadness and trust. I then took the total number of words in each category and divided it by the total number of words (minus stop words) to get a percent anger, or a percent joy.



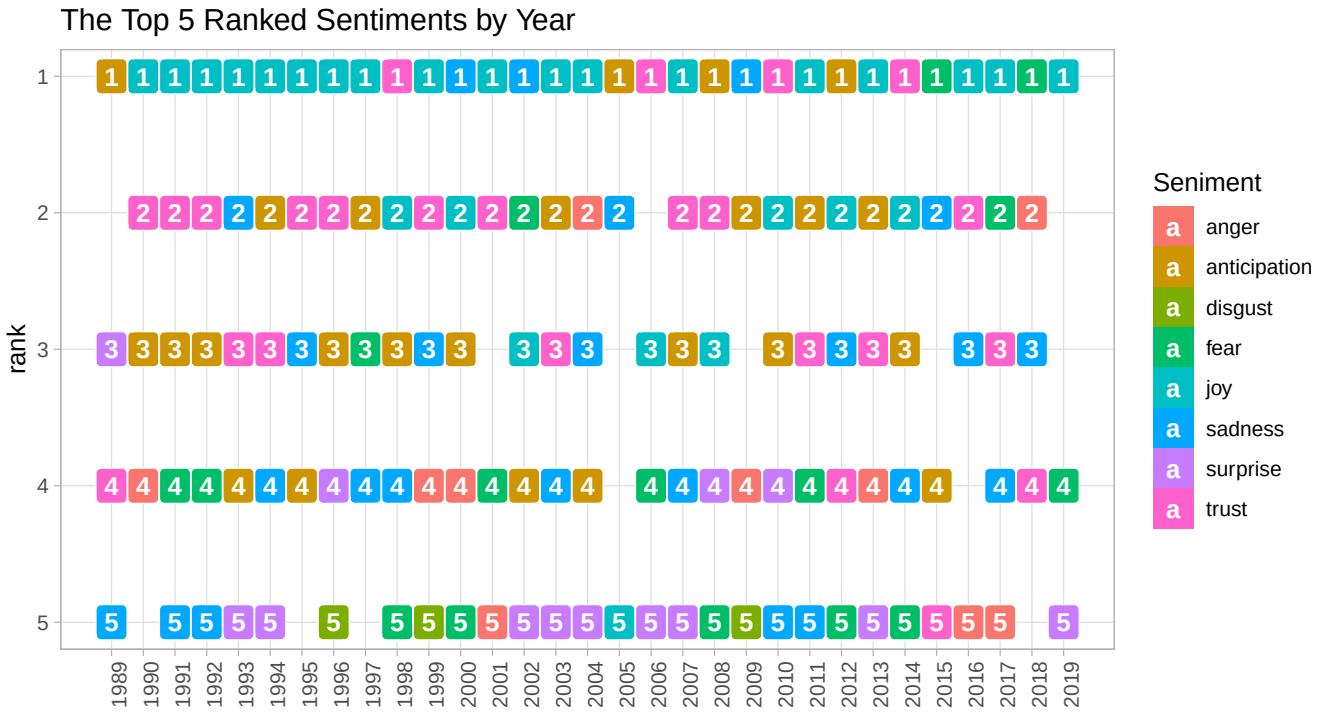
Overall it seems there are more number of songs that have positive sentiments, but there does not seem to be a correlation between winning and sentiments.



There appears to be a more variability in sentiment scores as time moves on. More recent years have wider swings. And years prior to 2005 overall seem to be more positive.



There were 7 years that have an overall average sentiment of negative (less than 0) and we see that 2000 had the lowest median sentiment score of any year.



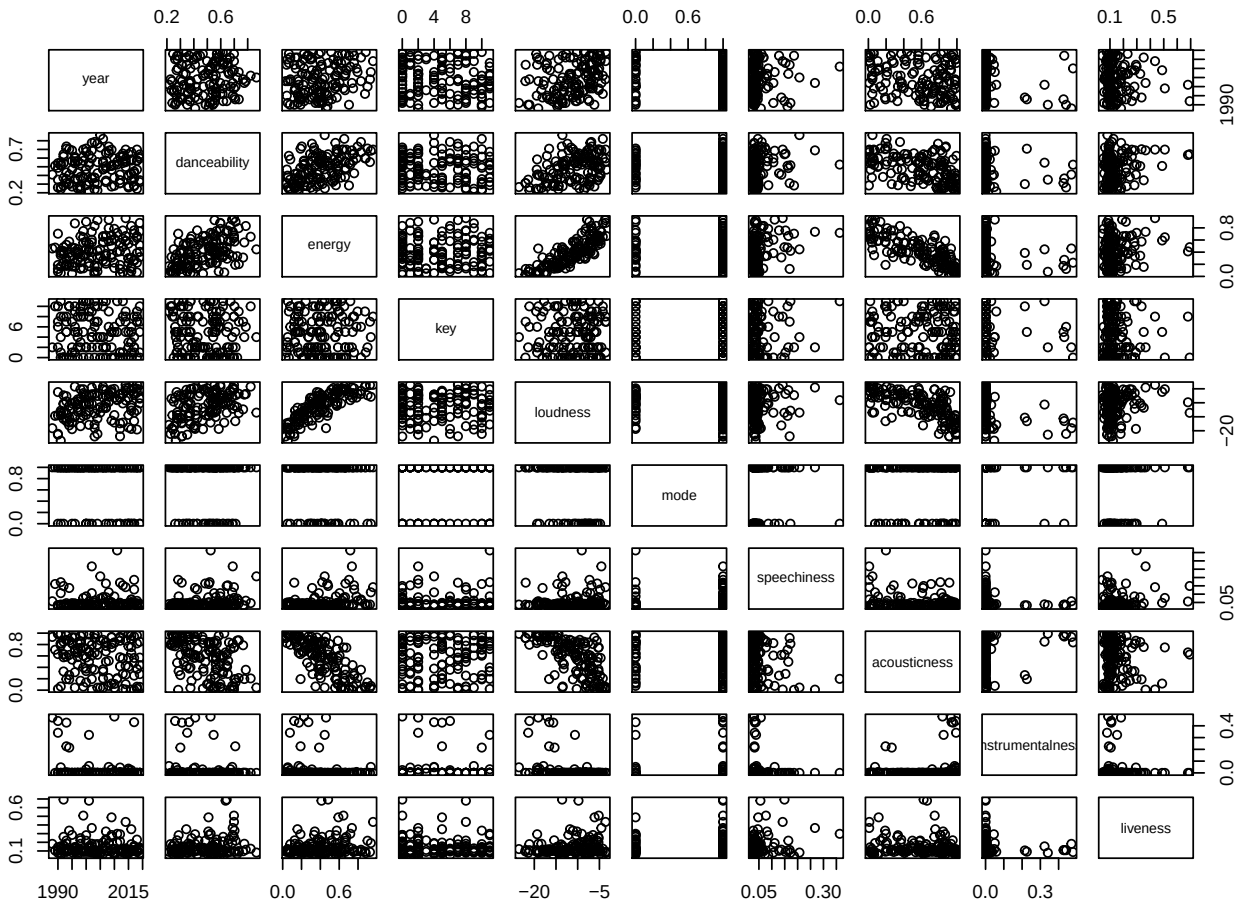
We see that Joy was the prevailing sentiment for much of the 90's and then it becomes less popular. There does not seem to be much of a trend in the top 5 sentiments each year.

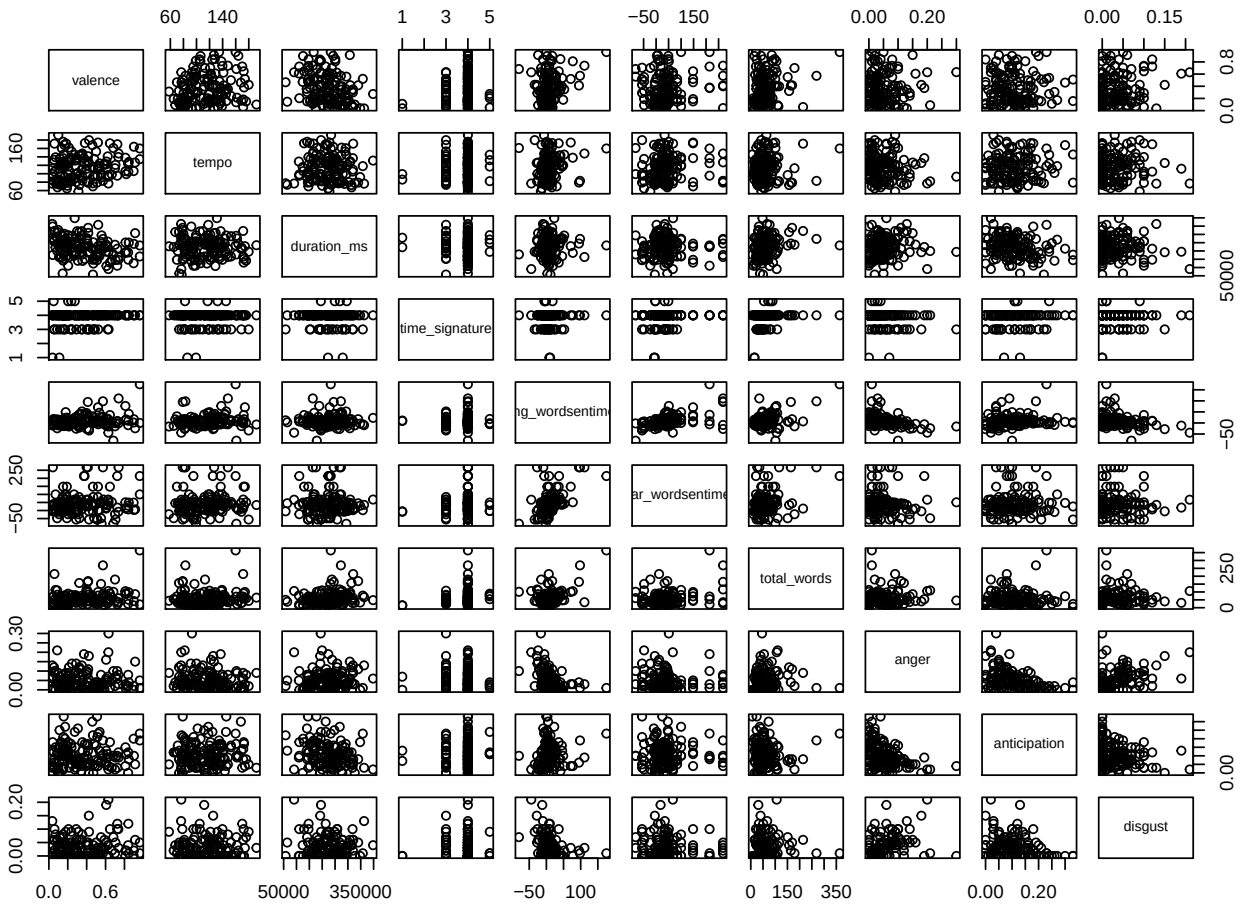
negative

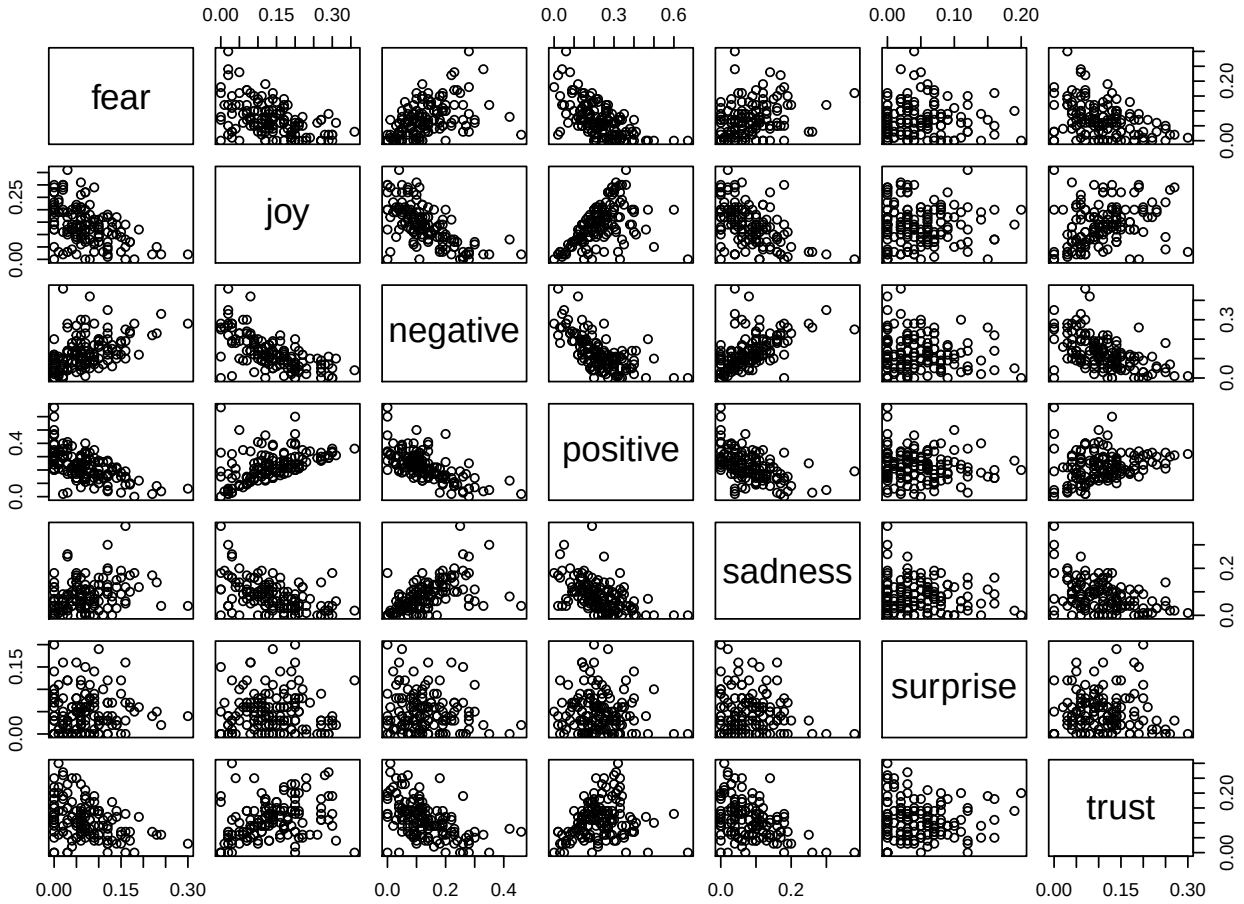
positive

The most popular words split by their positivity. Love is not only the most used positive word, it is the most used word in our lyrics.

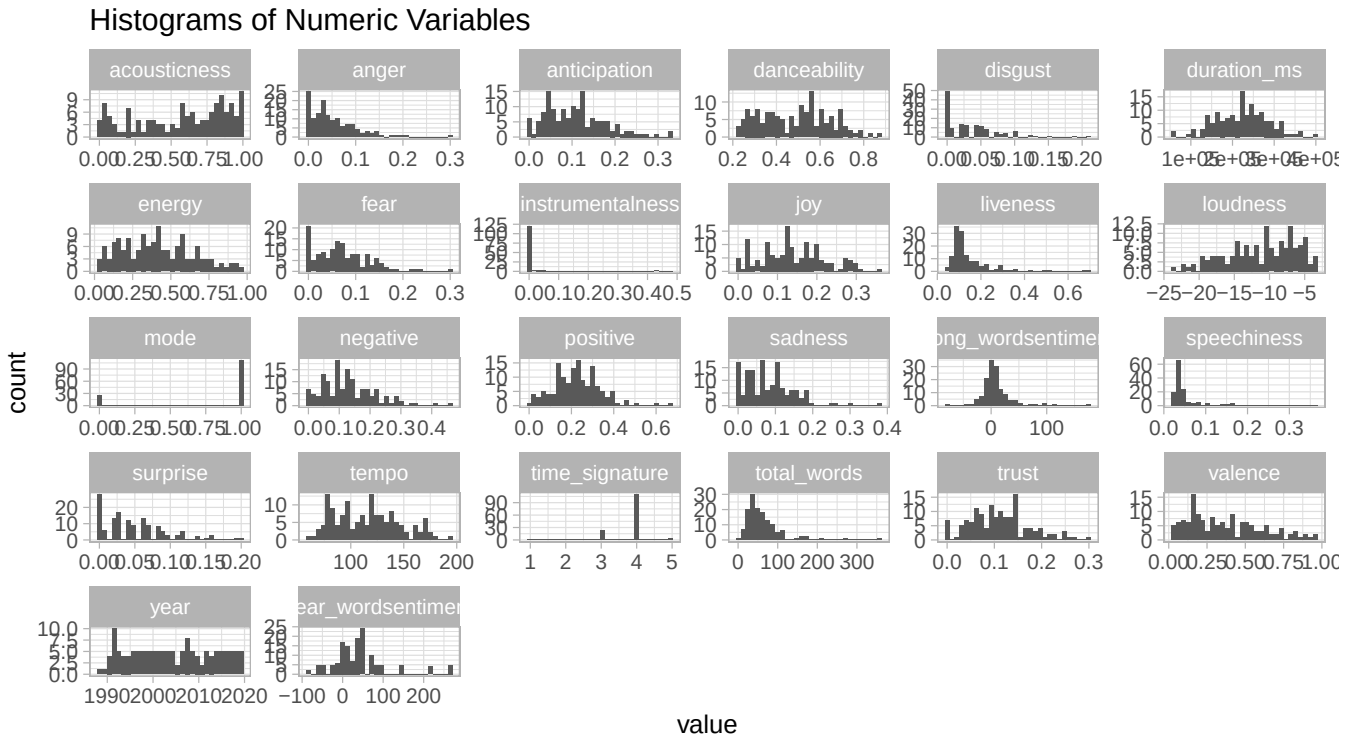
Pairs Plots of Numeric Variables

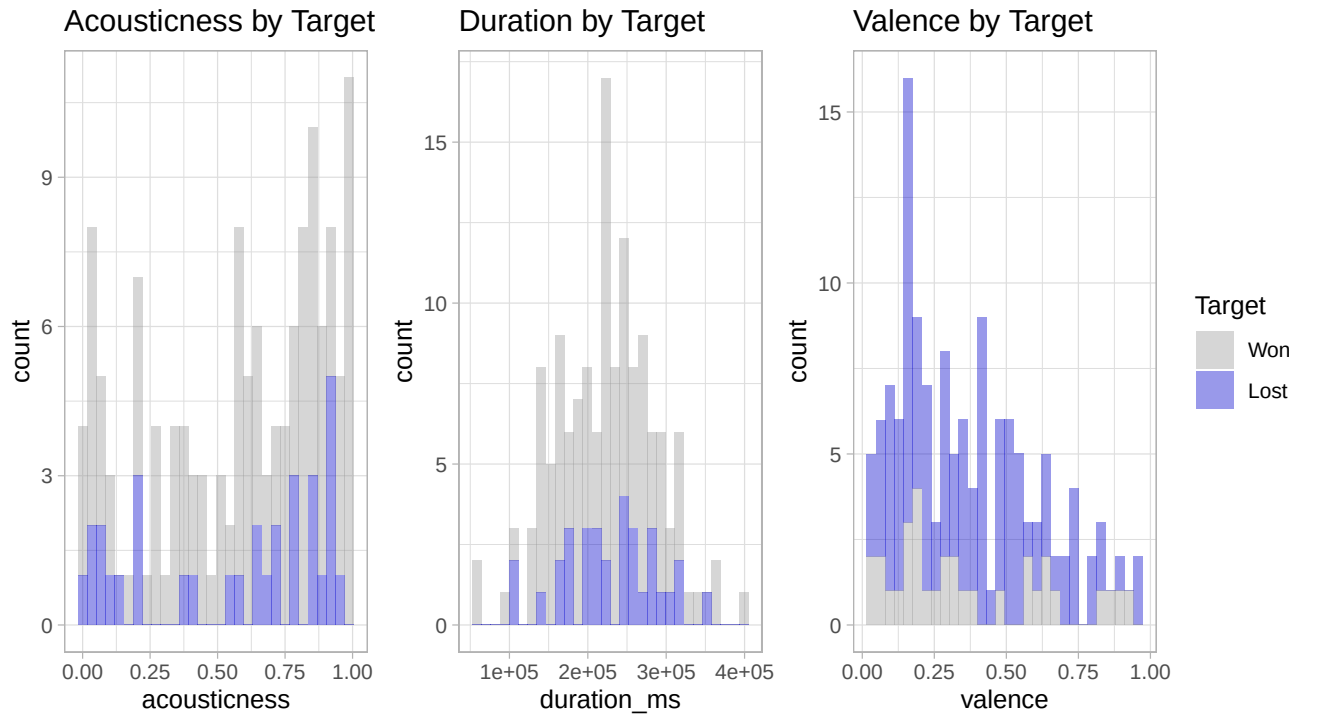

year
danceability
energy
key
loudness
mode
speechiness
acousticness
nstrumentalnes
liveness


valence
tempo
duration_ms
time_signature
total_words
anger
anticipation
disgust



Variables relationships of note: I have a lot of repeat variables that might provide some collinearity, positive is very similar to joy, surprise, trust. Negative also has similar attributes. Energy and loudness are another pair that could prove problematic later on.

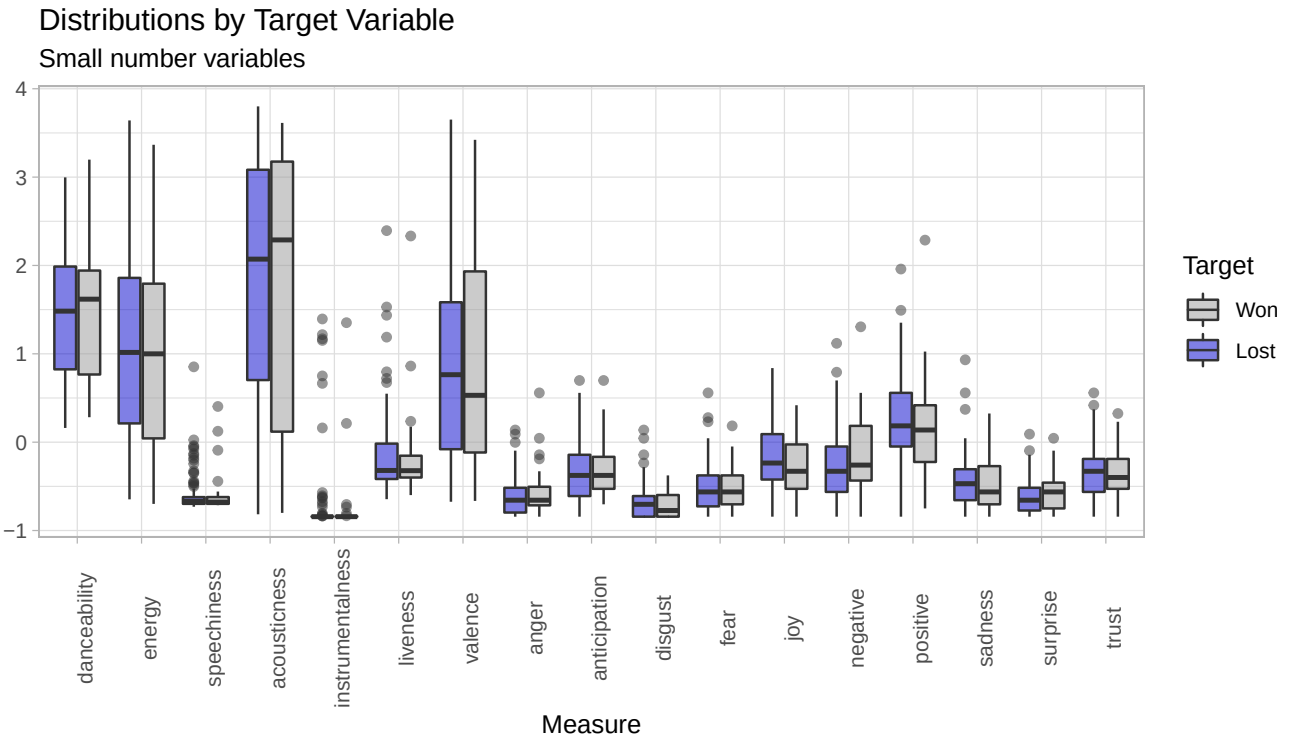


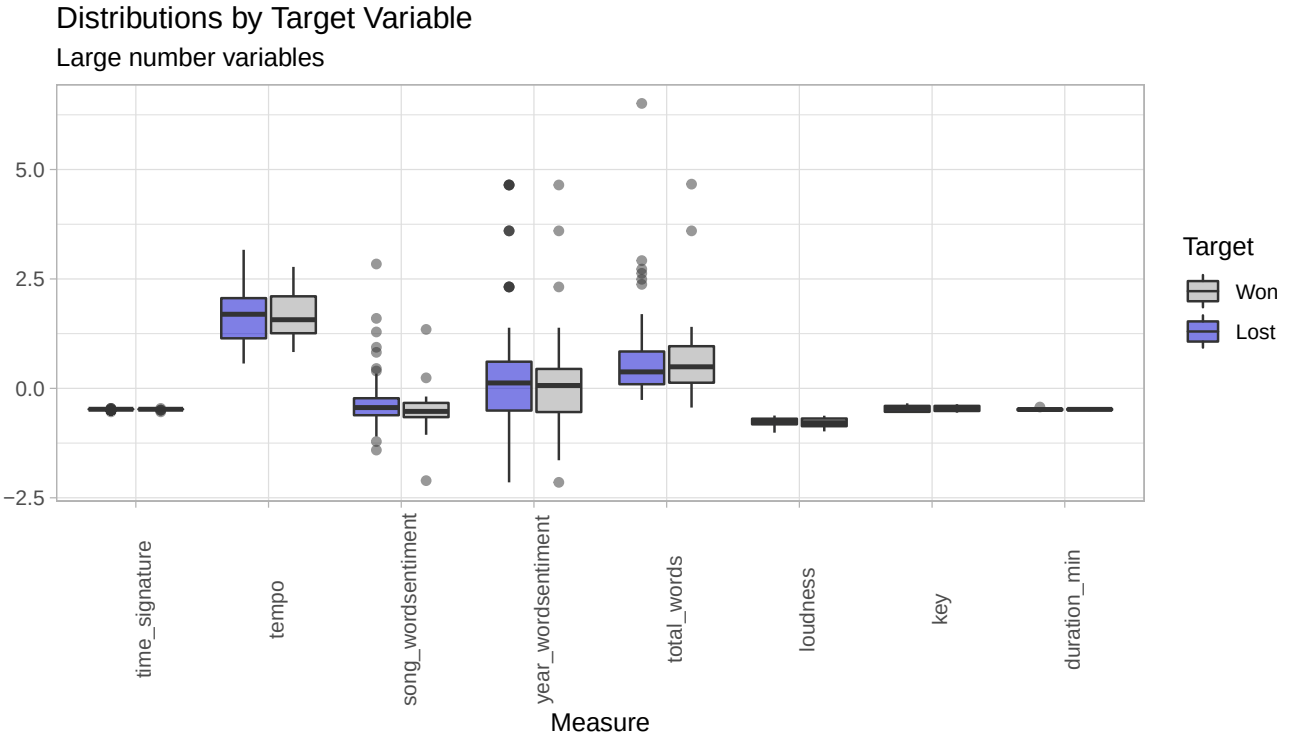


Acousticness has an inverse bell curve, I see a handful of skewed measures; total words, speechiness, disgust, surprise maybe, anger. I only see one that has any left skew which is loudness.

I picked a couple of variables to see if there was a difference in target variable (winning) for any of them. It looks like for duration, valence, and acousticness that there is not much a difference distribution-wise between target variables.

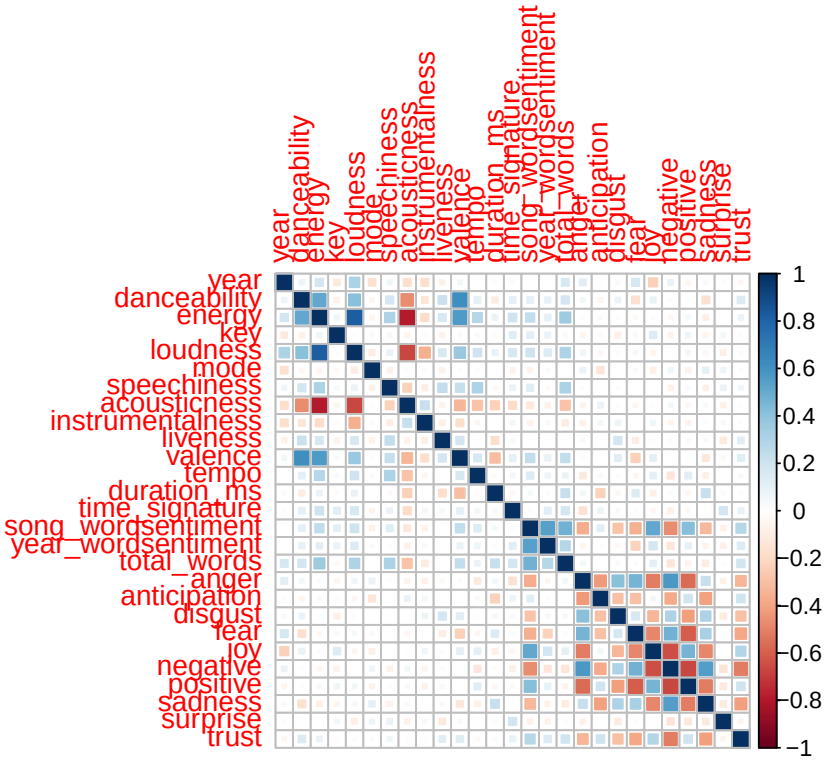
Boxplots by Target Variable





Variables that appear to have a higher median for winning than losing: Danceability, Acousticness, Negative, Surprise, maybe Total Words. Variables that have a visually lower median for winning than losing: Valence, Disgust, Joy, Positive, Sadness, Trust and maybe Tempo.

Corr plot

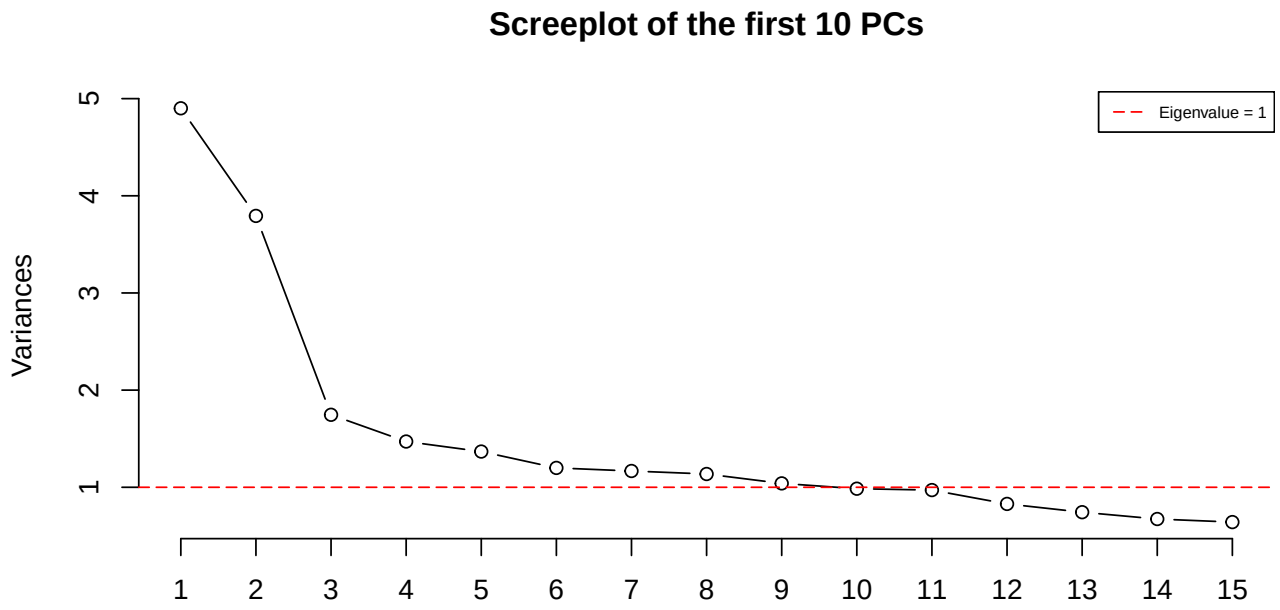


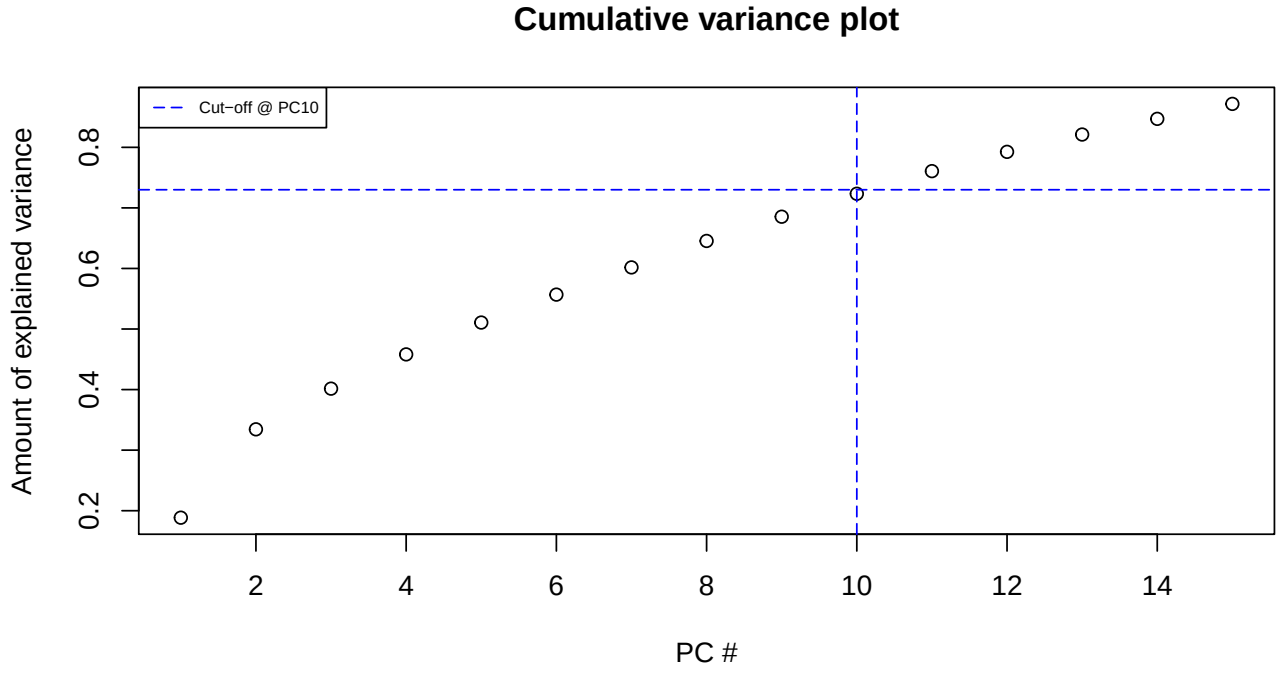
A couple of key observations with correlations, acousticness and energy are very strongly negatively correlated. Most of the lyric sentiment fields are either closely correlated with another field (anger and fear or joy and song sentiment) this could prove problematic with models later on.

Training the Models

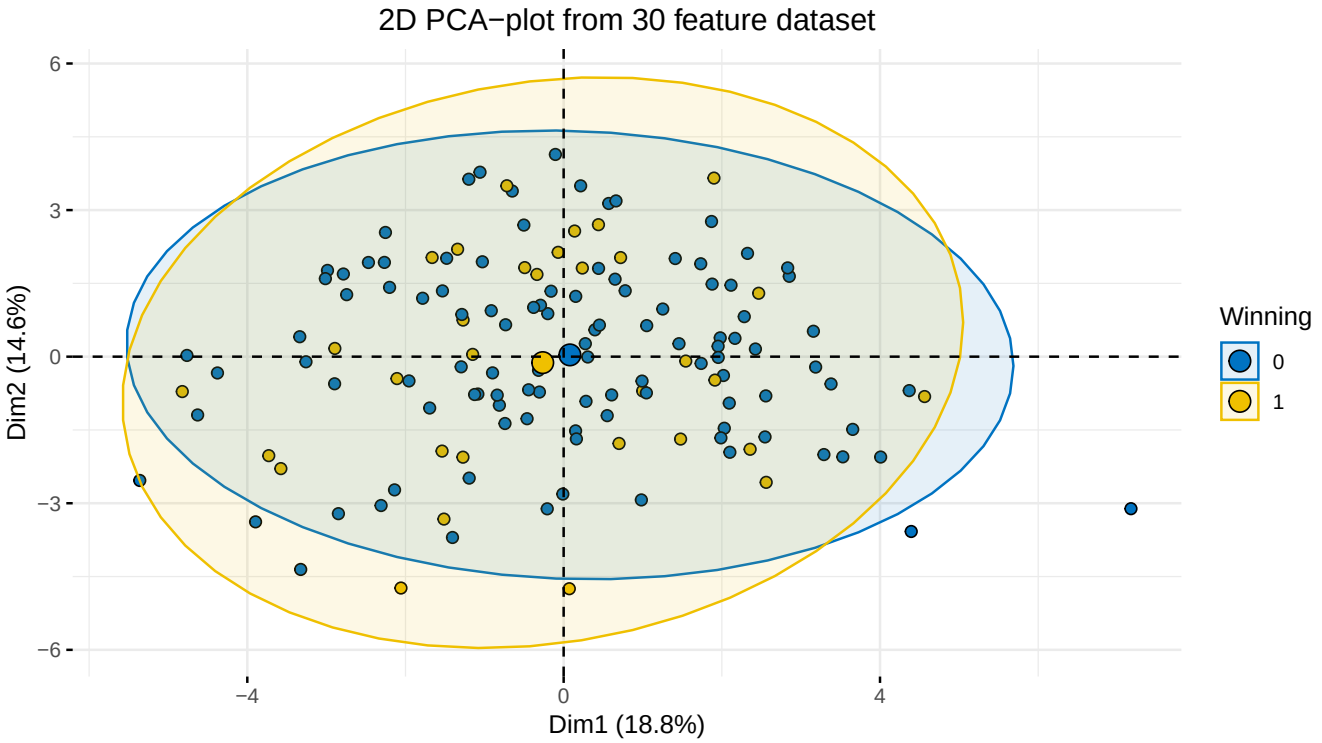
Support Vector Machine

Principal Component Analysis






```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

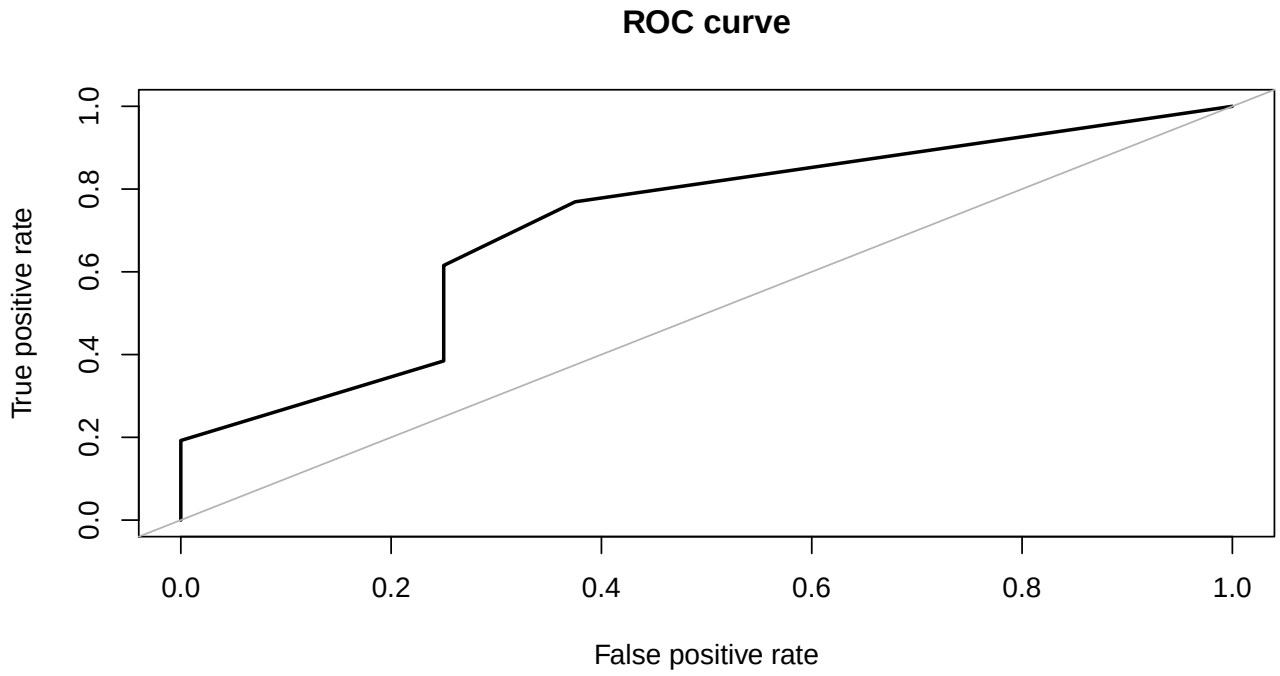


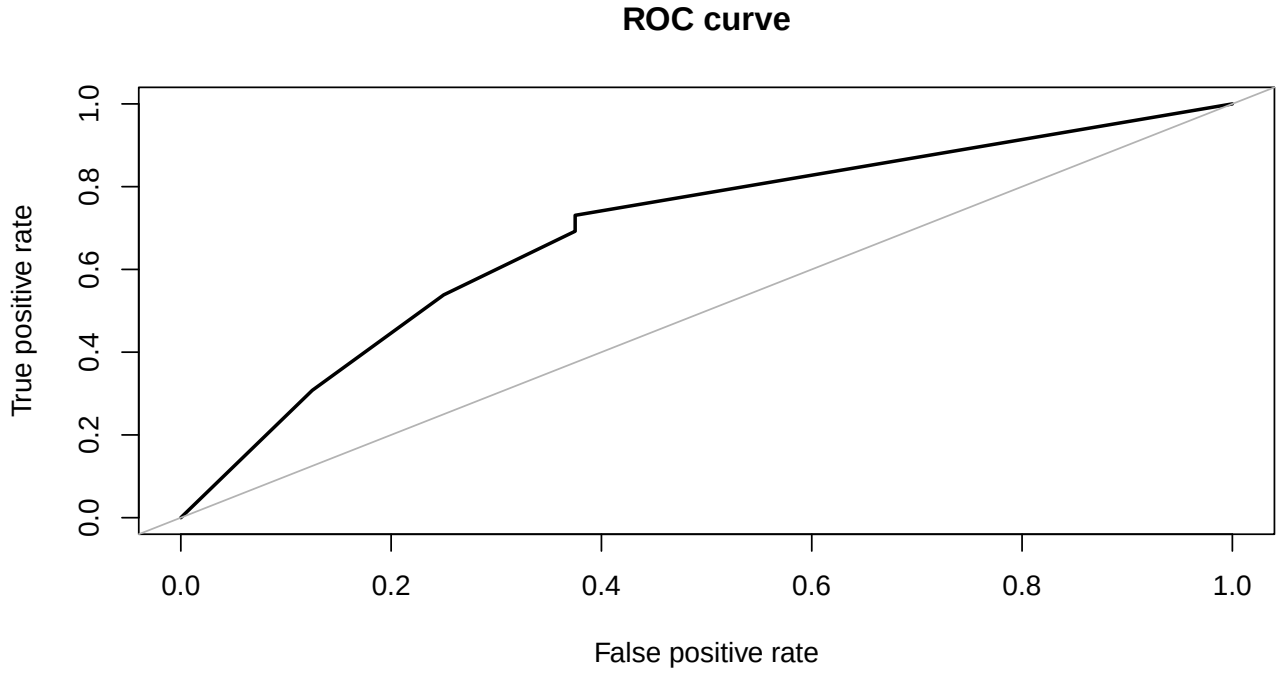
From our plot we see that there is not much separation between our winning and losing variable (1, 0) respectively.

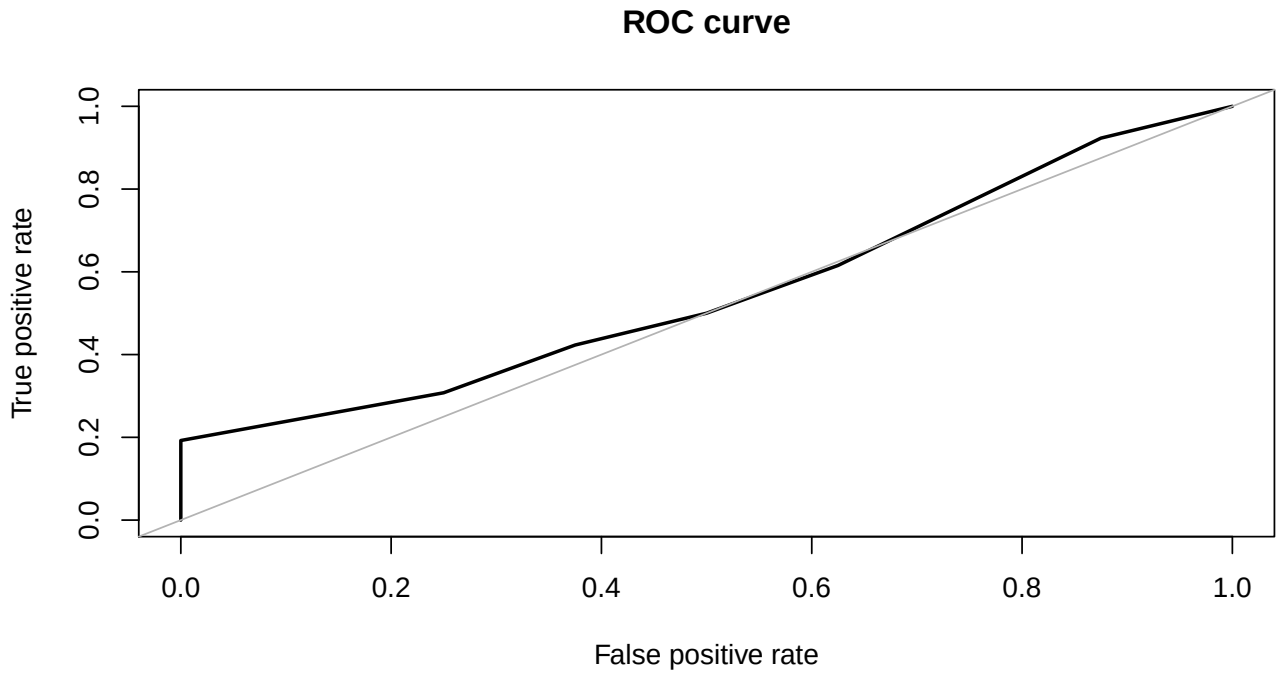
Models

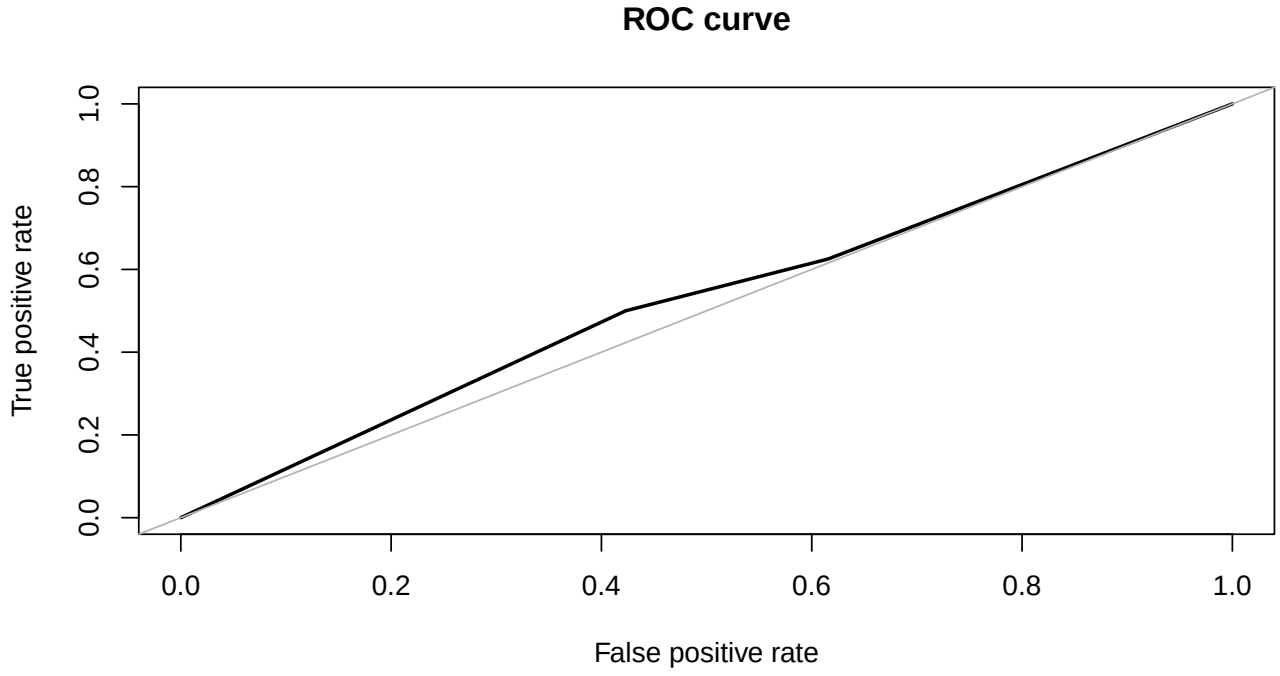
SVM and Regression Trees

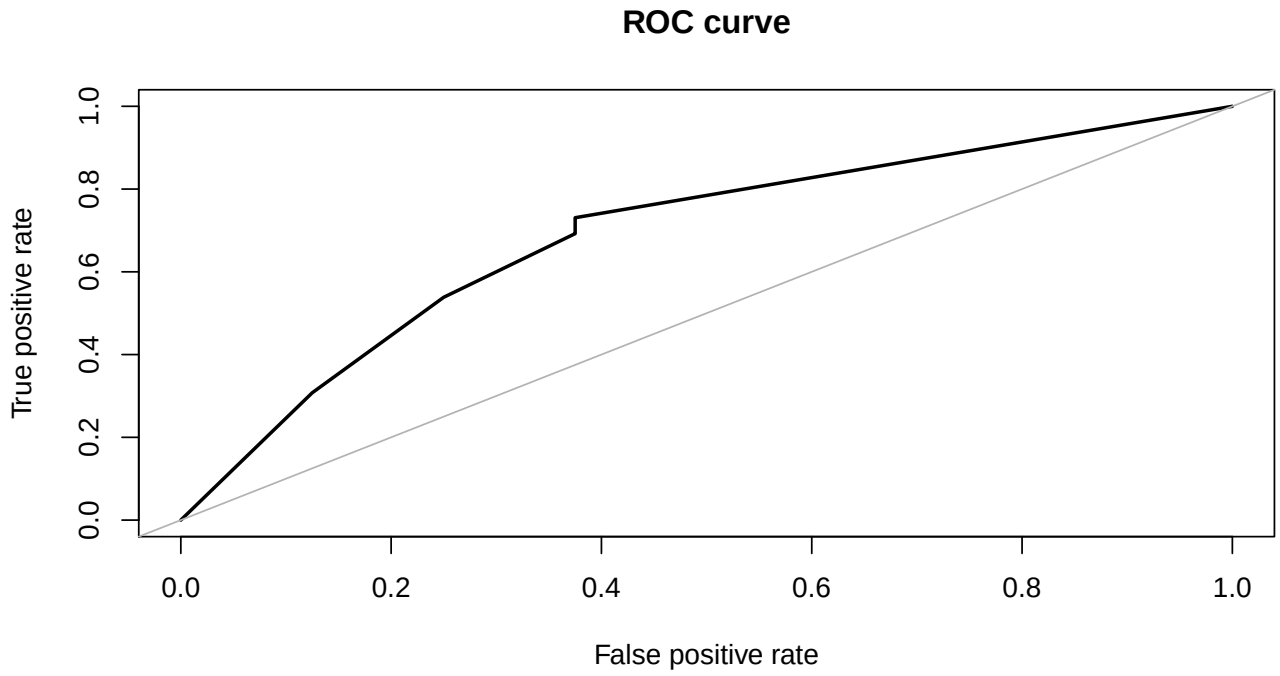
These models were attempted on three different datasets: raw, scaled, and the principal component analysis data. I also use each of the datasets as is and then perform some sampling methods to handle target imbalance. I use over sampling of the minority samples, under sampling of the majority samples, or both. The rose variation sets the probability of the minority sample.


ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

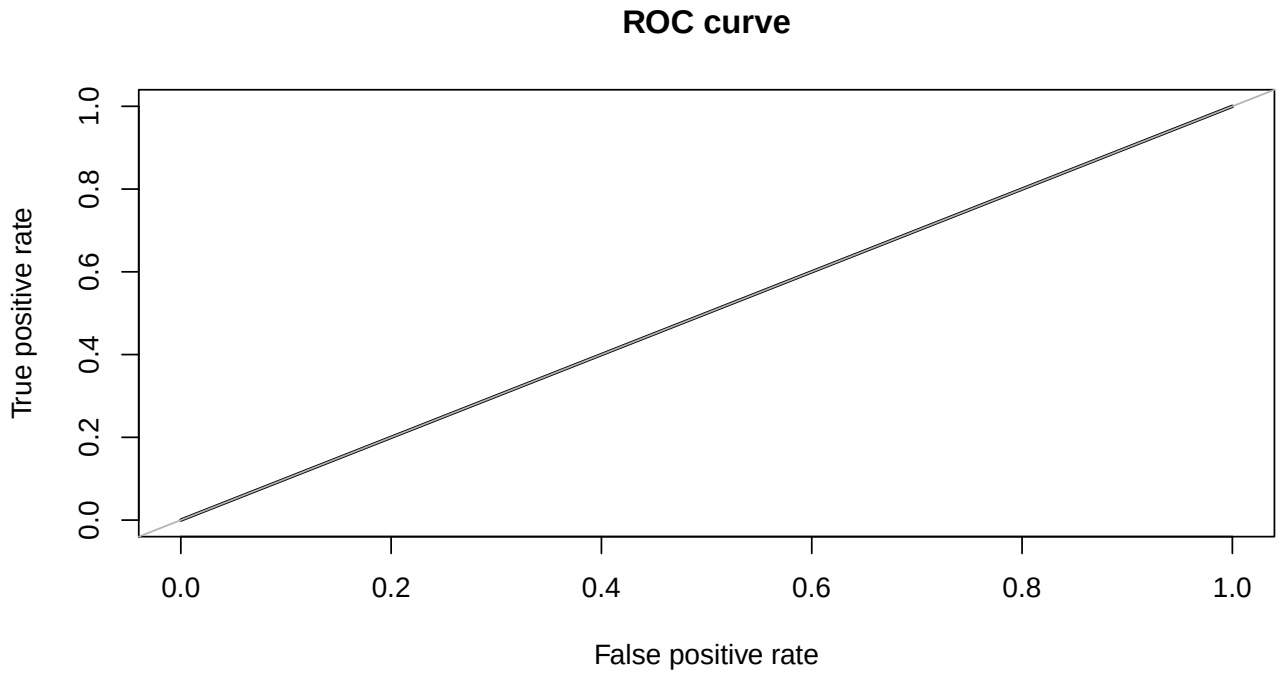

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

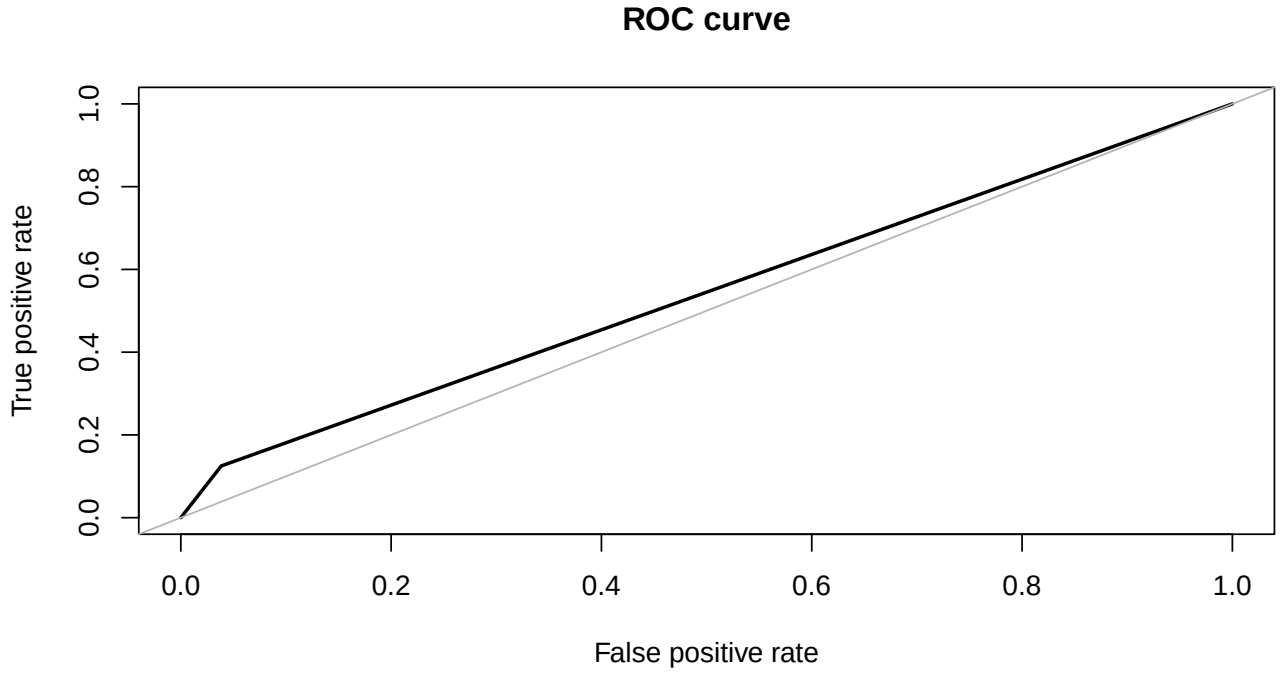

ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

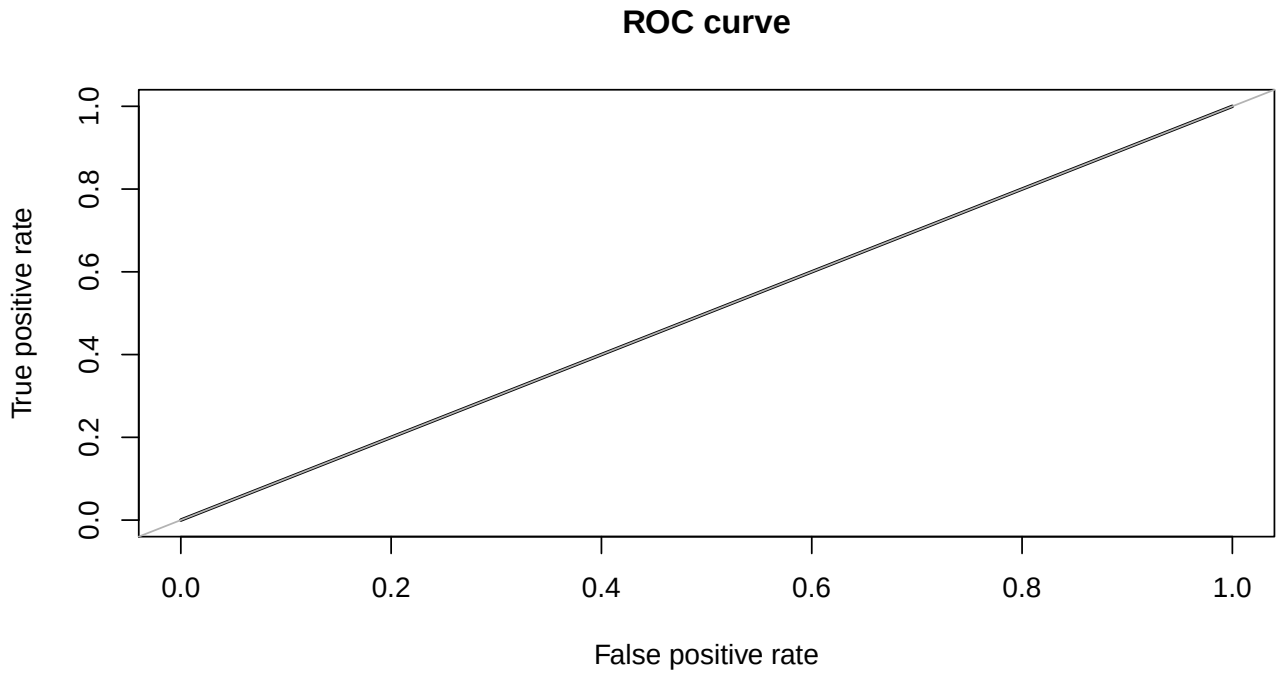

ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

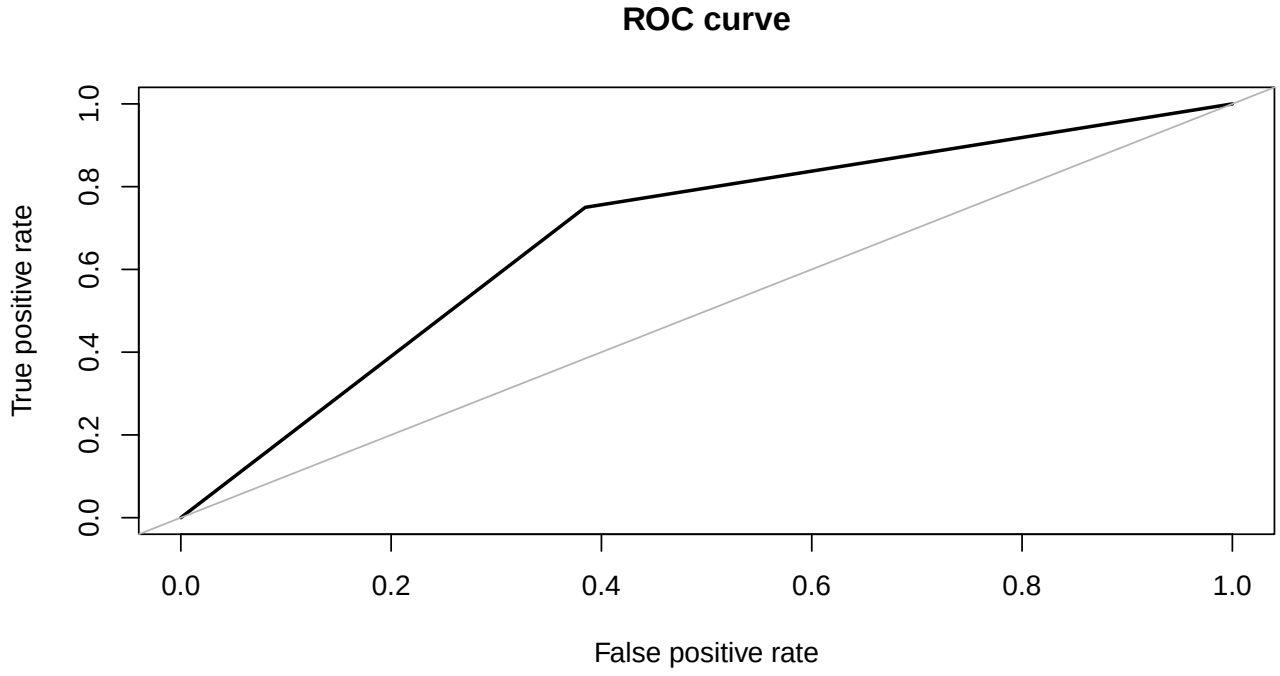


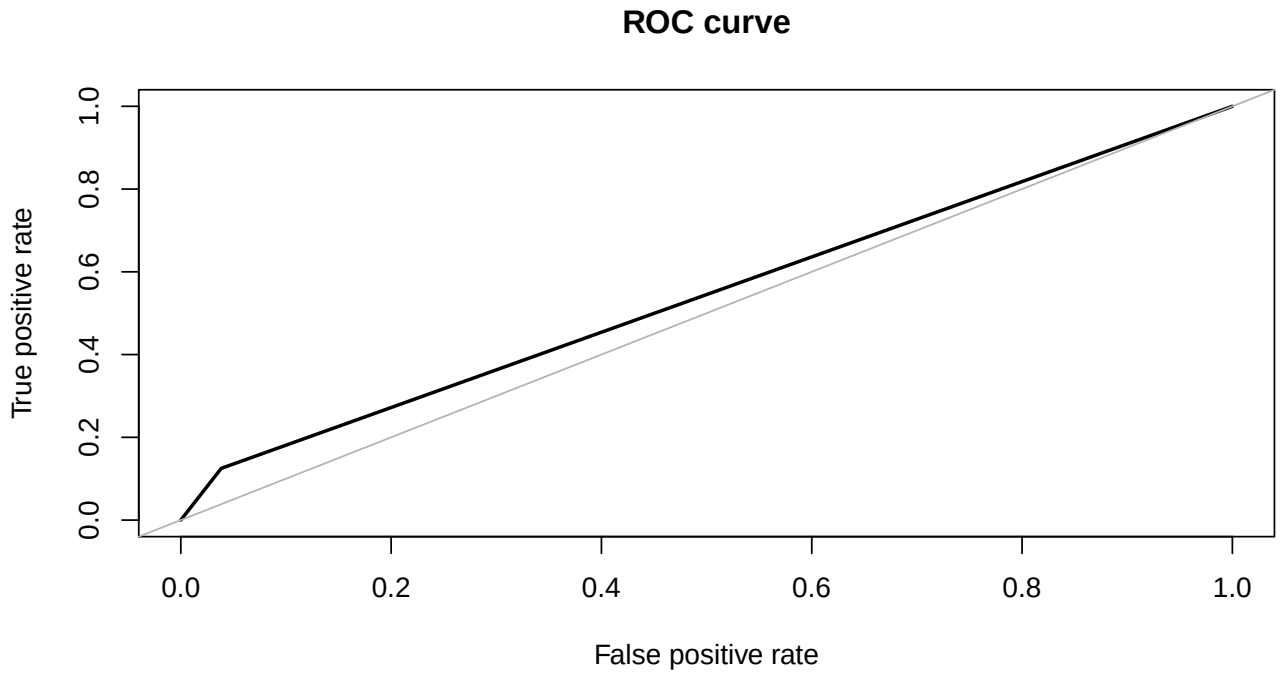

```
## Area under the curve (AUC): 0.690

## Area under the curve (AUC): 0.548

## Area under the curve (AUC): 0.526

## Area under the curve (AUC): 0.690

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  8
##          1  5  0
##
##                Accuracy : 0.6176
##                  95% CI : (0.4356, 0.7783)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9831
##
##                   Kappa : -0.221
##
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.8077
##             Specificity : 0.0000
##          Pos Pred Value : 0.7241
##          Neg Pred Value : 0.0000
##              Prevalence : 0.7647
##          Detection Rate : 0.6176
##    Detection Prevalence : 0.8529
##       Balanced Accuracy : 0.4038
##
##        'Positive' Class : 0
##
```


ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

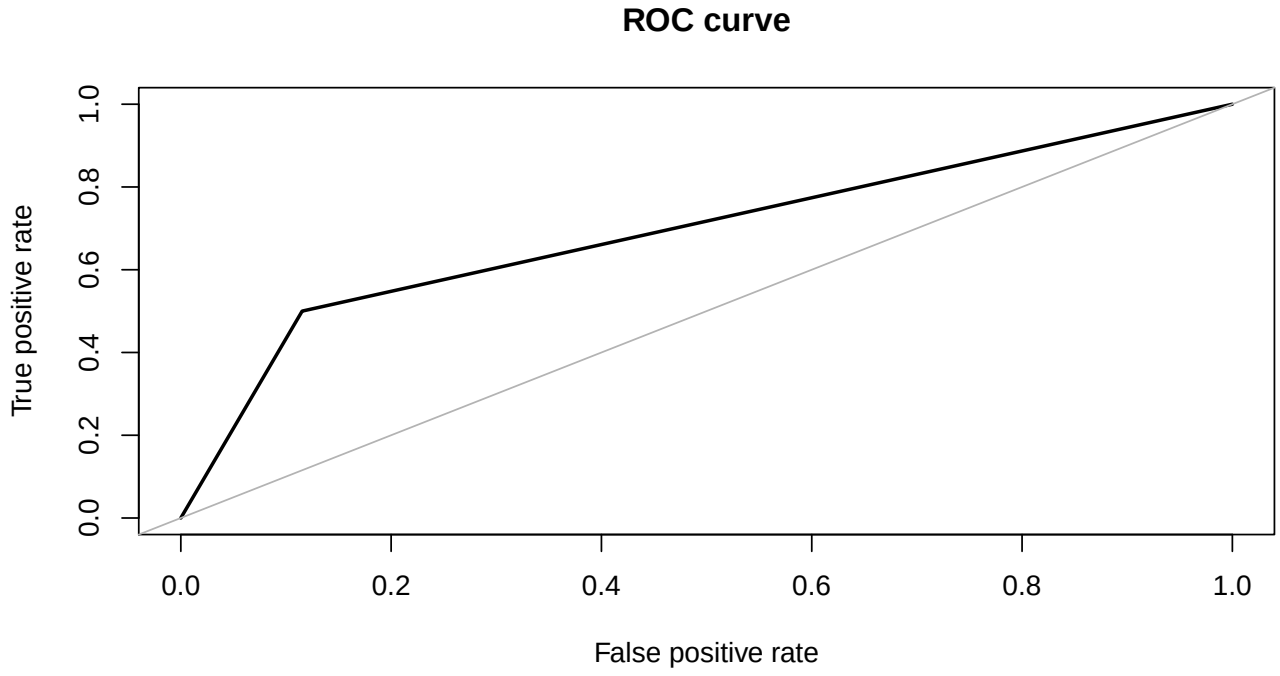

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

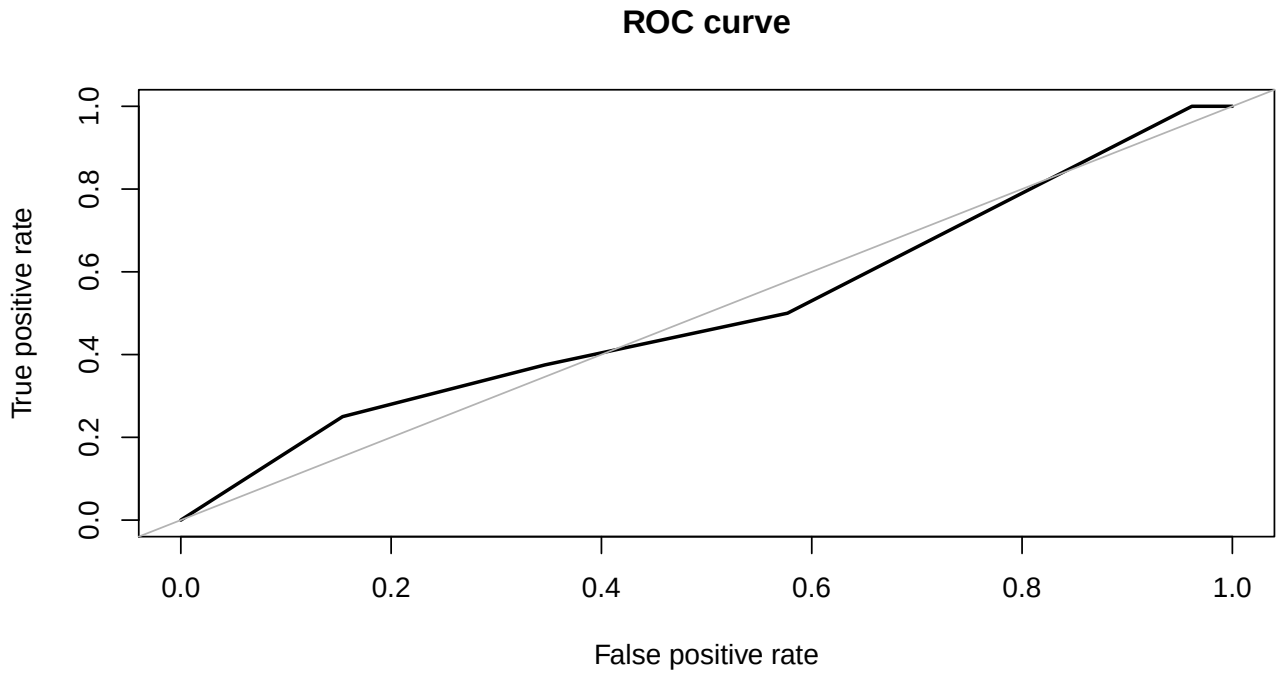

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

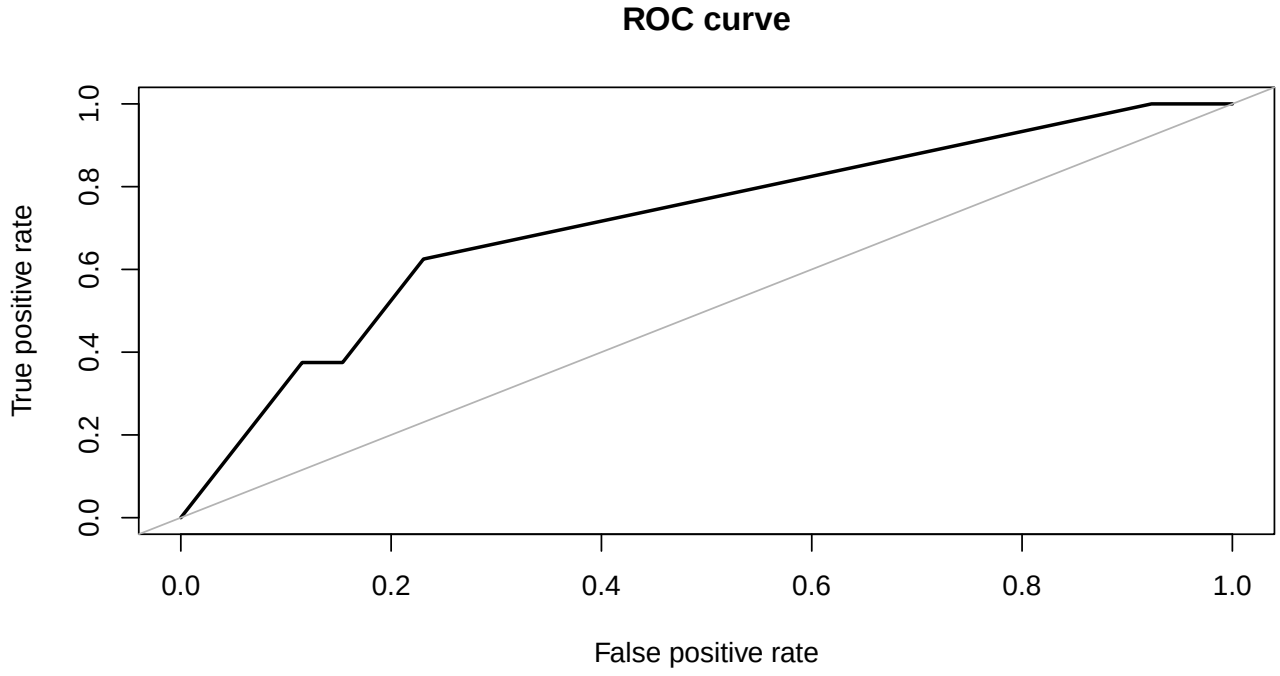

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

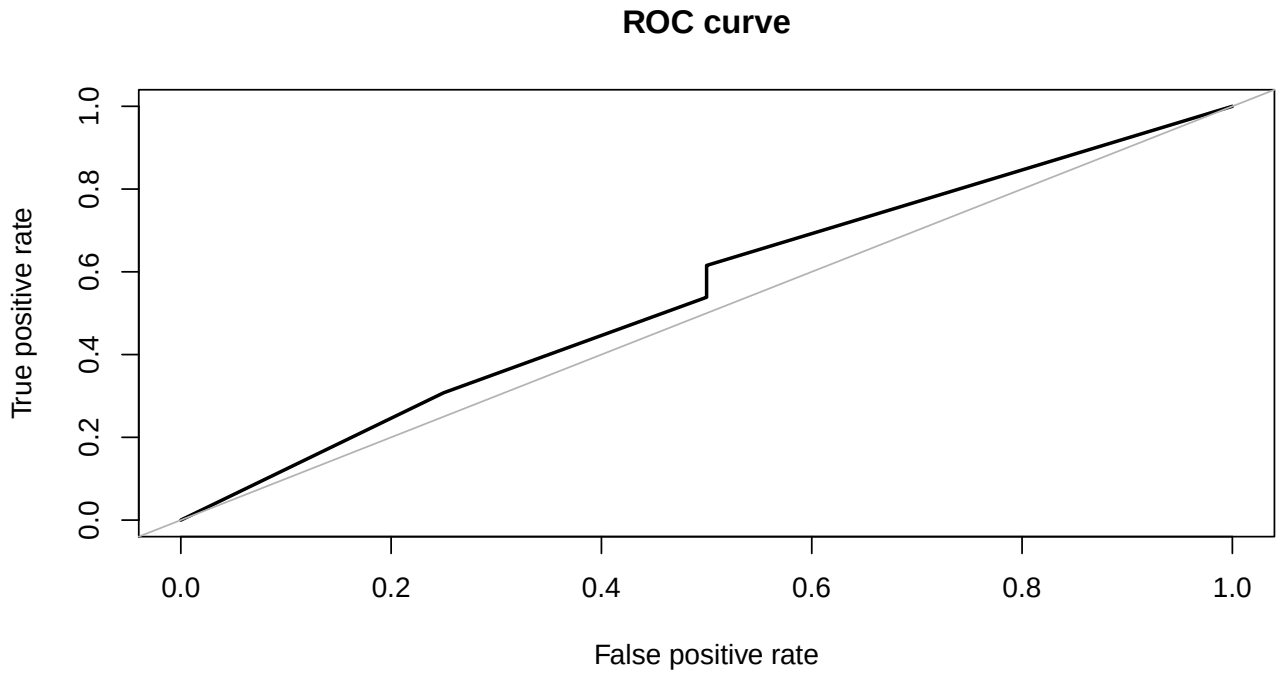


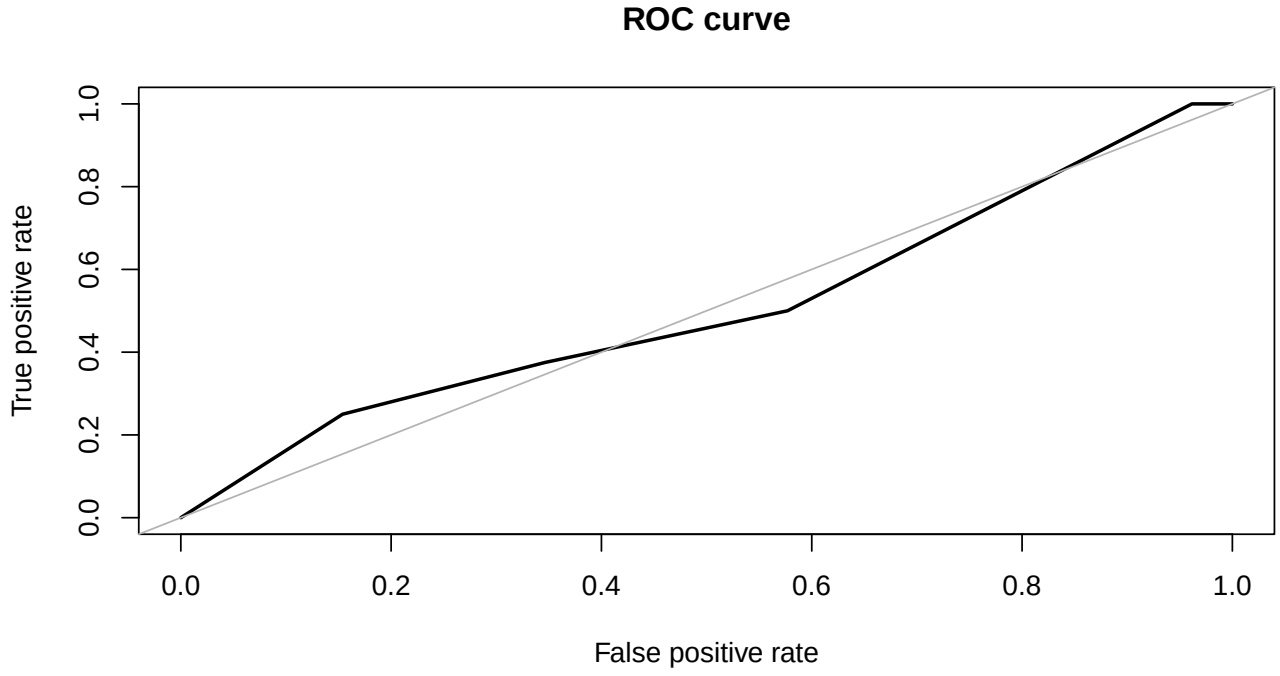

```
## Area under the curve (AUC): 0.500
## Area under the curve (AUC): 0.543
## Area under the curve (AUC): 0.500
## Area under the curve (AUC): 0.683
## Area under the curve (AUC): 0.543
```



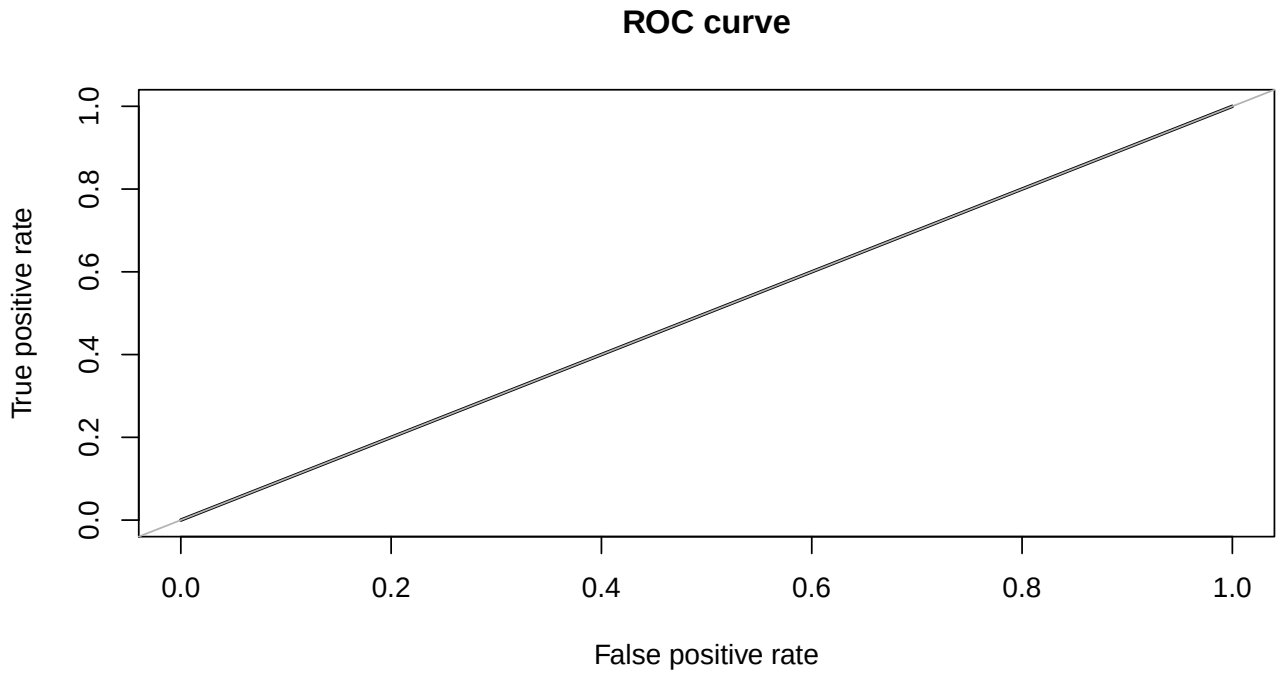

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

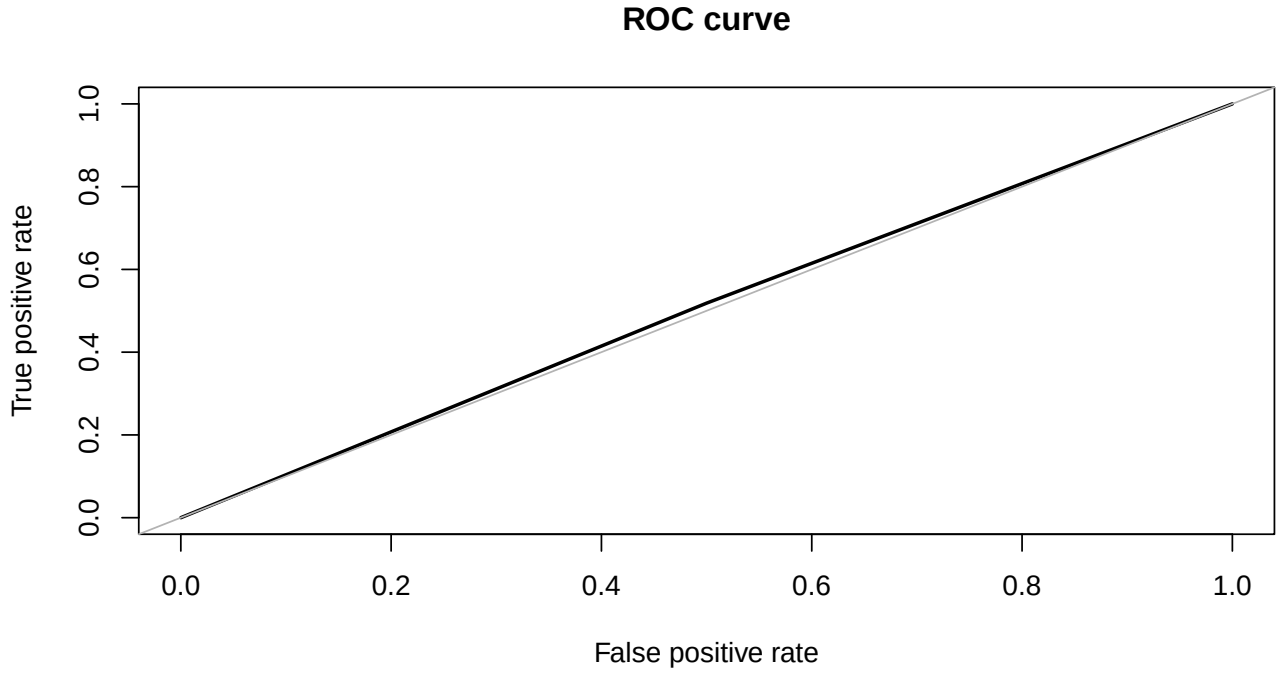

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

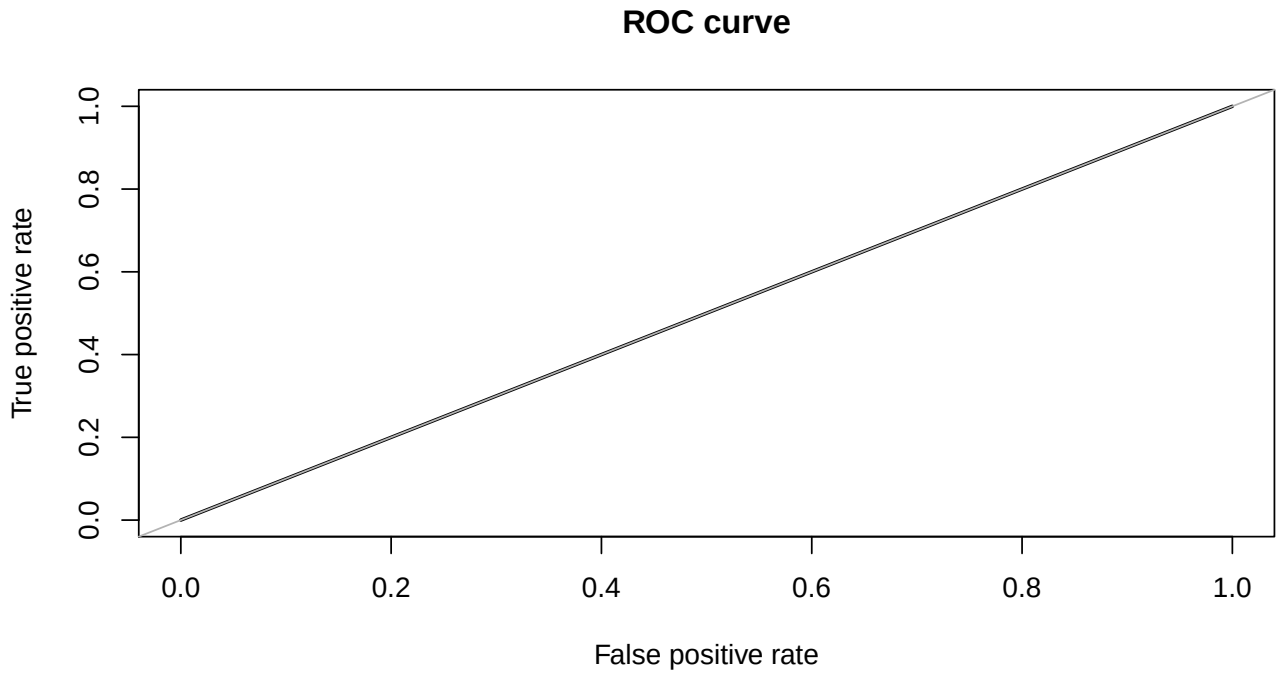


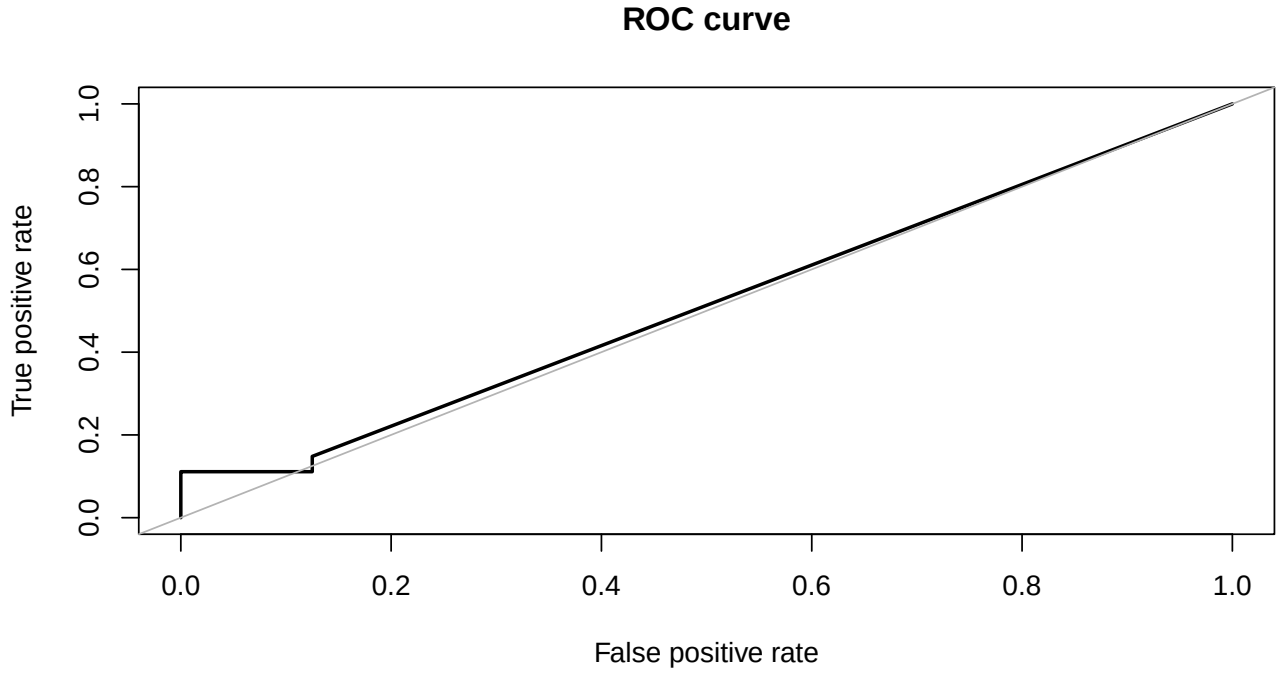


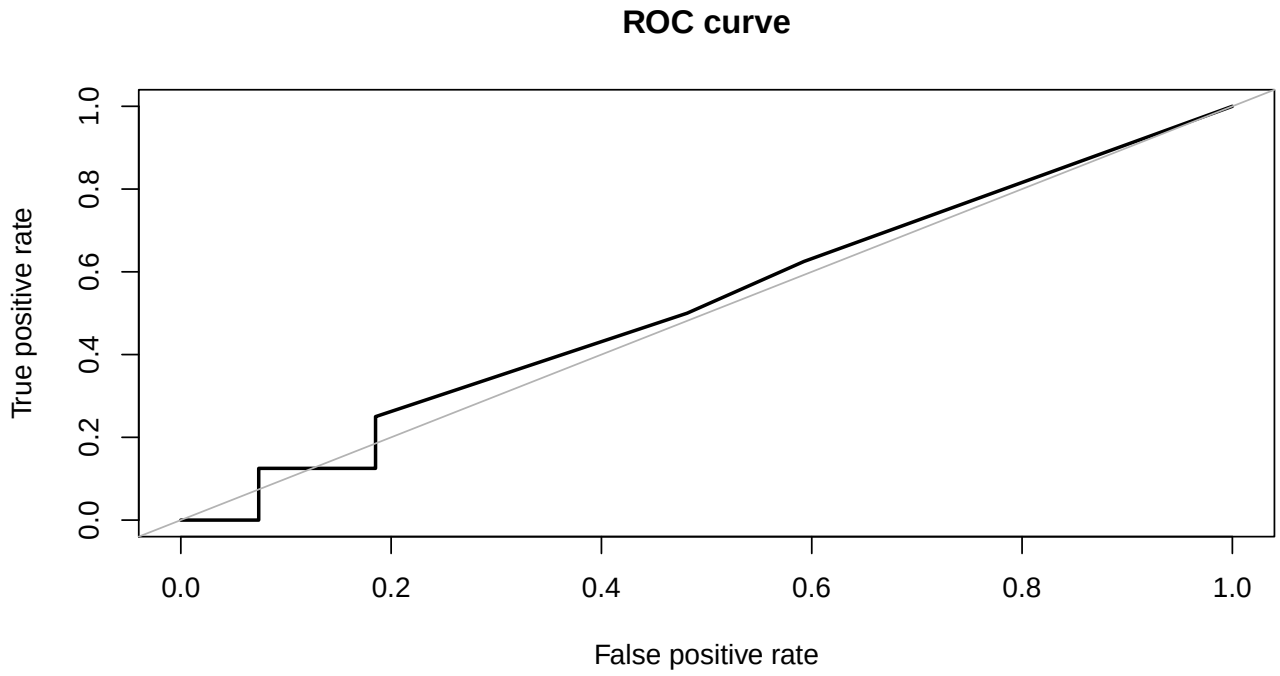

```
## Area under the curve (AUC): 0.692
## Area under the curve (AUC): 0.507
## Area under the curve (AUC): 0.714
## Area under the curve (AUC): 0.548
## Area under the curve (AUC): 0.507
```

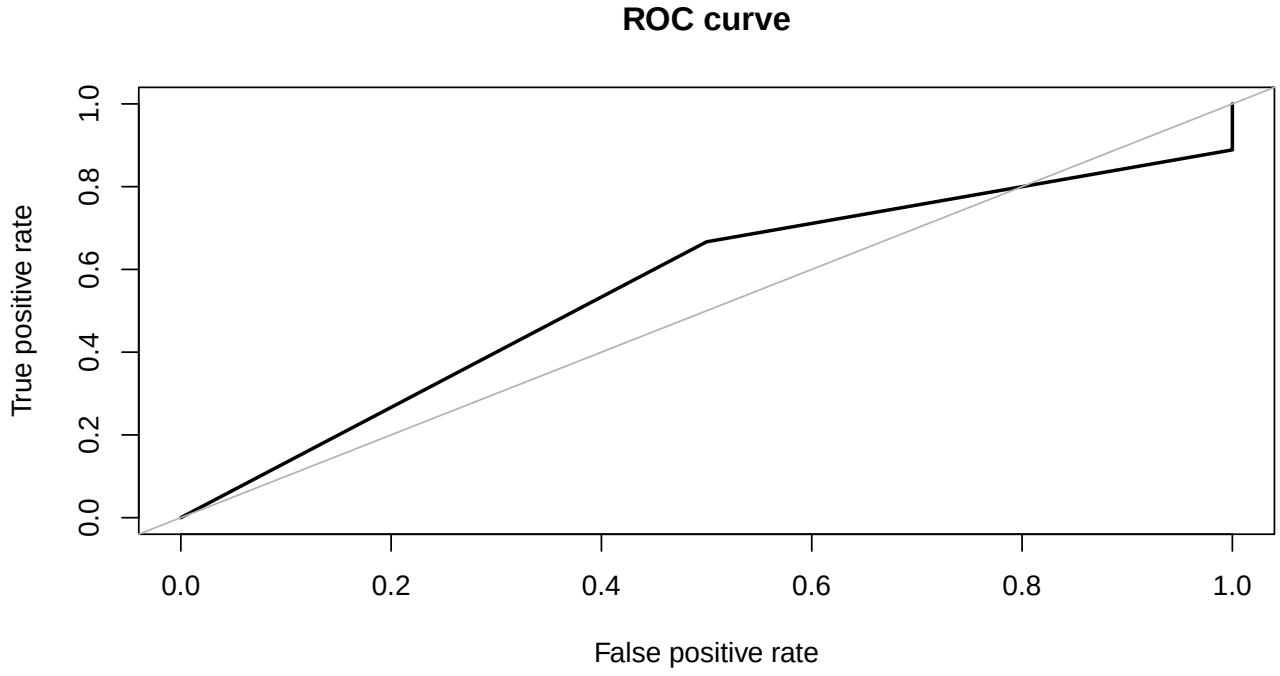

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

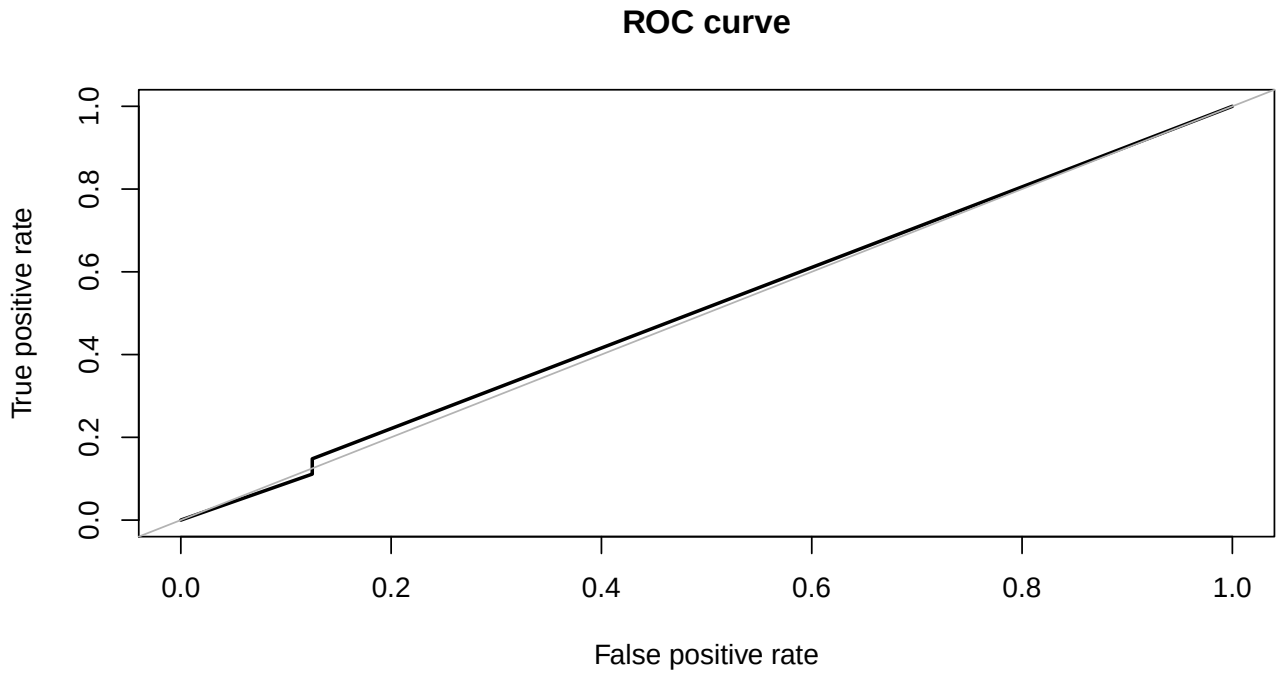

ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0




```
## Area under the curve (AUC): 0.516
```

```
## Area under the curve (AUC): 0.516
```

```
## Area under the curve (AUC): 0.519
```

```
## Area under the curve (AUC): 0.556
```

```
## Area under the curve (AUC): 0.509
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 24  7
##          1  3  1
##
##                Accuracy : 0.7143
##                  95% CI : (0.537, 0.8536)
##     No Information Rate : 0.7714
##     P-Value [Acc > NIR] : 0.8433
##
##                   Kappa : 0.0169
##
##  Mcnemar's Test P-Value : 0.3428
##
##             Sensitivity : 0.8889
##             Specificity : 0.1250
##          Pos Pred Value : 0.7742
##          Neg Pred Value : 0.2500
##              Prevalence : 0.7714
##          Detection Rate : 0.6857
##    Detection Prevalence : 0.8857
##       Balanced Accuracy : 0.5069
##
```

```
##        'Positive' Class : 0
##
```

The best models so far are the regression trees using the regular data with the pca dataset. It also tied with the rose pca followed by the regression tree with both sampling. I need to test their confusion matrices next.

Logistic Regressions

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = trainScale_songs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5846  -0.6585  -0.3328  -0.1796   2.5908
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.71409    0.35651  -4.808 1.52e-06 ***
## danceability        0.05881    0.43870   0.134   0.8934
## energy              1.49352    0.86979   1.717   0.0860 .
## key                 0.12239    0.31545   0.388   0.6980
## loudness           -1.22178    0.61403  -1.990   0.0466 *
## mode               -0.05439    0.31790  -0.171   0.8642
## speechiness        -0.67668    0.41268  -1.640   0.1011
## acousticness        0.53942    0.57580   0.937   0.3489
## instrumentalness   -0.28455    0.38425  -0.741   0.4590
## liveness            0.78798    0.40616   1.940   0.0524 .
## valence            -0.38673    0.52444  -0.737   0.4609
## tempo               0.31298    0.33854   0.925   0.3552
## duration_ms         0.01909    0.36359   0.053   0.9581
## time_signature      0.06937    0.31500   0.220   0.8257
## song_wordsentiment -0.37169    0.56478  -0.658   0.5105
## year_wordsentiment -0.23017    0.42329  -0.544   0.5866
## total_words         0.76401    0.42314   1.806   0.0710 .
## anger              -1.08265    1.66693  -0.649   0.5160
## anticipation       -2.11372    2.27999  -0.927   0.3539
## disgust            -2.33139    1.39373  -1.673   0.0944 .
## fear               -2.31071    1.89003  -1.223   0.2215
## joy                -3.09942    2.62108  -1.182   0.2370
## negative           -2.90323    2.80649  -1.034   0.3009
## positive           -3.69312    3.61050  -1.023   0.3064
## sadness            -2.15933    2.15112  -1.004   0.3155
## surprise           -0.90402    1.41813  -0.637   0.5238
## trust              -2.35061    2.10555  -1.116   0.2643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 112.88  on 104  degrees of freedom
## Residual deviance:  84.82  on  78  degrees of freedom
## AIC: 138.82
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                  104    112.884
## danceability         1   0.0321       103    112.852 0.857868
## energy               1   0.0173       102    112.835 0.895476
## key                  1   0.0483       101    112.787 0.826120
## loudness             1   2.4822       100    110.304 0.115138
## mode                 1   0.1271        99    110.177 0.721415
## speechiness          1   0.1404        98    110.037 0.707901
## acousticness         1   0.0618        97    109.975 0.803714
## instrumentalness     1   1.1240        96    108.851 0.289066
## liveness             1   0.4929        95    108.358 0.482623
## valence              1   2.7375        94    105.621 0.098018 .
## tempo                1   0.3406        93    105.280 0.559510
## duration_ms          1   0.0028        92    105.277 0.957978
## time_signature       1   0.5260        91    104.751 0.468277
## song_wordsentiment   1   1.6354        90    103.116 0.200954
## year_wordsentiment   1   0.0000        89    103.116 0.999044
## total_words          1   1.4098        88    101.706 0.235095
## anger                1   1.2290        87    100.477 0.267604
## anticipation         1   1.7596        86     98.718 0.184674
## disgust              1   7.3594        85     91.358 0.006671 **
## fear                 1   0.9233        84     90.435 0.336602
## joy                  1   1.5534        83     88.882 0.212634
## negative             1   0.0223        82     88.859 0.881359
## positive             1   0.2154        81     88.644 0.642541
## sadness              1   0.1859        80     88.458 0.666310
## surprise             1   2.3627        79     86.095 0.124264
## trust                1   1.2751        78     84.820 0.258821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Accuracy 0.617647058823529"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  8
##          1  5  0
##
##                Accuracy : 0.6176
##                  95% CI : (0.4356, 0.7783)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9831
##
##                   Kappa : -0.221
##
##  Mcnemar's Test P-Value : 0.5791
##
```

```
##             Sensitivity : 0.8077
##             Specificity : 0.0000
##          Pos Pred Value : 0.7241
##          Neg Pred Value : 0.0000
##              Prevalence : 0.7647
##          Detection Rate : 0.6176
##    Detection Prevalence : 0.8529
##       Balanced Accuracy : 0.4038
##
##        'Positive' Class : 0
##
```

```
## [1] "Accuracy 0.5"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 12  3
##          1 14  5
##
##                Accuracy : 0.5
##                  95% CI : (0.3243, 0.6757)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.99980
##
##                   Kappa : 0.0586
##
##  Mcnemar's Test P-Value : 0.01529
##
##             Sensitivity : 0.4615
##             Specificity : 0.6250
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.2632
##              Prevalence : 0.7647
##          Detection Rate : 0.3529
##    Detection Prevalence : 0.4412
##       Balanced Accuracy : 0.5433
##
##        'Positive' Class : 0
##
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = both_ScaleTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5528  -0.1261   0.0000   0.5896   2.7907
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -11.7914     5.3211  -2.216   0.0267 *
## danceability         3.3796     2.1620   1.563   0.1180
## energy              18.4382     7.3972   2.493   0.0127 *
## key                  0.3542     0.7409   0.478   0.6325
## loudness           -13.7289     6.1209  -2.243   0.0249 *
```

```
## mode                 0.1806     0.5252   0.344   0.7309
## speechiness         -3.6990     1.6460  -2.247   0.0246 *
## acousticness         4.6801     2.1096   2.218   0.0265 *
## instrumentalness   -35.1667    15.6341  -2.249   0.0245 *
## liveness             4.0023     1.8711   2.139   0.0324 *
## valence             -7.2911     3.1029  -2.350   0.0188 *
## tempo                1.8967     1.4786   1.283   0.1996
## duration_ms         -4.5332     2.3621  -1.919   0.0550 .
## time_signature      -3.3969     1.6089  -2.111   0.0347 *
## song_wordsentiment  -5.7378     2.9011  -1.978   0.0480 *
## year_wordsentiment  -0.3870     1.4171  -0.273   0.7848
## total_words          5.9702     2.6464   2.256   0.0241 *
## anger               -1.9480     4.4957  -0.433   0.6648
## anticipation        -8.4494     7.0227  -1.203   0.2289
## disgust            -12.2820     6.1033  -2.012   0.0442 *
## fear                -2.6169     5.0589  -0.517   0.6050
## joy                -14.0411     9.0176  -1.557   0.1195
## negative           -13.1376     9.1776  -1.431   0.1523
## positive            -8.6231     9.7876  -0.881   0.3783
## sadness             -2.3901     5.6118  -0.426   0.6702
## surprise            -1.8475     3.9927  -0.463   0.6436
## trust               -1.1288     4.7868  -0.236   0.8136
## ---
## 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 104  degrees of freedom
## Residual deviance:  57.34  on  78  degrees of freedom
## AIC: 111.34
##
## Number of Fisher Scoring iterations: 11

## [1] "Accuracy 0.558823529411765"

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 16  5
##          1 10  3
##
##                Accuracy : 0.5588
##                  95% CI : (0.3789, 0.7281)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9977
##
##                   Kappa : -0.0079
##
##  Mcnemar's Test P-Value : 0.3017
##
##             Sensitivity : 0.6154
##             Specificity : 0.3750
##          Pos Pred Value : 0.7619
##          Neg Pred Value : 0.2308
##              Prevalence : 0.7647
##          Detection Rate : 0.4706
##    Detection Prevalence : 0.6176
```

```
##       Balanced Accuracy : 0.4952
##
##        'Positive' Class : 0
##

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = rose_ScaleTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5528  -0.1261   0.0000   0.5896   2.7907
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -11.7914     5.3211  -2.216   0.0267 *
## danceability         3.3796     2.1620   1.563   0.1180
## energy              18.4382     7.3972   2.493   0.0127 *
## key                  0.3542     0.7409   0.478   0.6325
## loudness           -13.7289     6.1209  -2.243   0.0249 *
## mode                 0.1806     0.5252   0.344   0.7309
## speechiness         -3.6990     1.6460  -2.247   0.0246 *
## acousticness         4.6801     2.1096   2.218   0.0265 *
## instrumentalness   -35.1667    15.6341  -2.249   0.0245 *
## liveness             4.0023     1.8711   2.139   0.0324 *
## valence             -7.2911     3.1029  -2.350   0.0188 *
## tempo                1.8967     1.4786   1.283   0.1996
## duration_ms         -4.5332     2.3621  -1.919   0.0550 .
## time_signature      -3.3969     1.6089  -2.111   0.0347 *
## song_wordsentiment  -5.7378     2.9011  -1.978   0.0480 *
## year_wordsentiment  -0.3870     1.4171  -0.273   0.7848
## total_words          5.9702     2.6464   2.256   0.0241 *
## anger               -1.9480     4.4957  -0.433   0.6648
## anticipation        -8.4494     7.0227  -1.203   0.2289
## disgust            -12.2820     6.1033  -2.012   0.0442 *
## fear                -2.6169     5.0589  -0.517   0.6050
## joy                -14.0411     9.0176  -1.557   0.1195
## negative           -13.1376     9.1776  -1.431   0.1523
## positive            -8.6231     9.7876  -0.881   0.3783
## sadness             -2.3901     5.6118  -0.426   0.6702
## surprise            -1.8475     3.9927  -0.463   0.6436
## trust               -1.1288     4.7868  -0.236   0.8136
## ---
## 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 104  degrees of freedom
## Residual deviance:  57.34  on  78  degrees of freedom
## AIC: 111.34
##
## Number of Fisher Scoring iterations: 11

## [1] "Accuracy 0.558823529411765"

## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction  0  1
##          0 16  5
##          1 10  3
##
##                Accuracy : 0.5588
##                  95% CI : (0.3789, 0.7281)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9977
##
##                   Kappa : -0.0079
##
##  Mcnemar's Test P-Value : 0.3017
##
##             Sensitivity : 0.6154
##             Specificity : 0.3750
##          Pos Pred Value : 0.7619
##          Neg Pred Value : 0.2308
##              Prevalence : 0.7647
##          Detection Rate : 0.4706
##    Detection Prevalence : 0.6176
##       Balanced Accuracy : 0.4952
##
##        'Positive' Class : 0
##

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = trainPCA_songs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3873  -0.7184  -0.5868  -0.3983   2.3187
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.44513    0.26828  -5.387 7.18e-08 ***
## PC1         -0.02200    0.12113  -0.182    0.856
## PC2         -0.05832    0.13330  -0.437    0.662
## PC3          0.01787    0.19991   0.089    0.929
## PC4          0.22737    0.20734   1.097    0.273
## PC5         -0.37026    0.25412  -1.457    0.145
## PC6          0.18766    0.24917   0.753    0.451
## PC7         -0.29189    0.26038  -1.121    0.262
## PC8          0.09667    0.22572   0.428    0.668
## PC9         -0.09469    0.23823  -0.397    0.691
## PC10         0.14957    0.28758   0.520    0.603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.33  on 103  degrees of freedom
## Residual deviance: 100.32  on  93  degrees of freedom
## AIC: 122.32
##
```

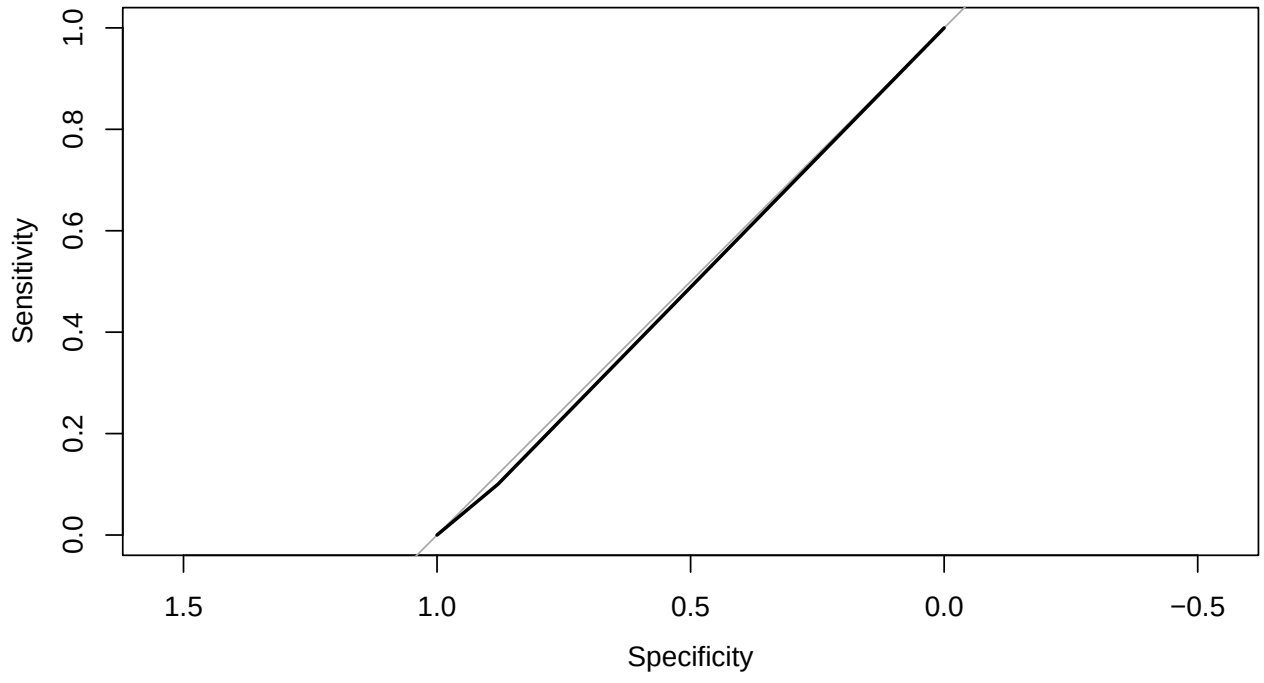
```
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   103     107.33
## PC1   1  0.11598       102     107.21   0.7334
## PC2   1  0.06198       101     107.15   0.8034
## PC3   1  0.00198       100     107.15   0.9645
## PC4   1  1.43735        99     105.71   0.2306
## PC5   1  2.50962        98     103.20   0.1132
## PC6   1  0.73177        97     102.47   0.3923
## PC7   1  1.54936        96     100.92   0.2132
## PC8   1  0.13814        95     100.78   0.7101
## PC9   1  0.19278        94     100.59   0.6606
## PC10  1  0.27131        93     100.31   0.6025
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.49
```




```
## [1] "Accuracy 0.657142857142857"
```

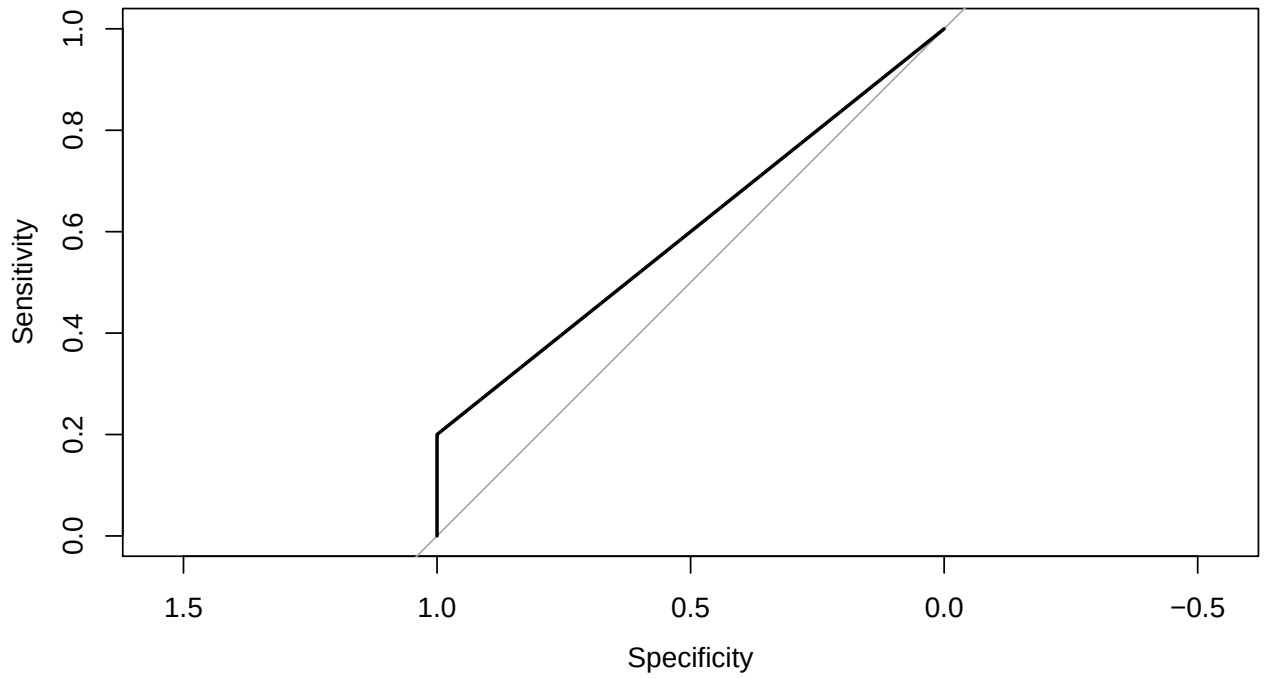
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
```

```
##          0 22  9
##          1  3  1
##
##                Accuracy : 0.6571
##                  95% CI : (0.4779, 0.8087)
##     No Information Rate : 0.7143
##     P-Value [Acc > NIR] : 0.8262
##
##                   Kappa : -0.0244
##
##  Mcnemar's Test P-Value : 0.1489
##
##             Sensitivity : 0.8800
##             Specificity : 0.1000
##          Pos Pred Value : 0.7097
##          Neg Pred Value : 0.2500
##              Prevalence : 0.7143
##          Detection Rate : 0.6286
##    Detection Prevalence : 0.8857
##       Balanced Accuracy : 0.4900
##
##        'Positive' Class : 0
##
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = over_PCATrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.19388  -1.04170  -0.01615   1.06047   1.68281
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.1813986  0.1744346  -1.040   0.2984
## PC1         -0.0419152  0.0727795  -0.576   0.5647
## PC2         -0.0004303  0.0908919  -0.005   0.9962
## PC3         -0.0444348  0.1287723  -0.345   0.7300
## PC4          0.3236013  0.1409261   2.296   0.0217 *
## PC5         -0.1059725  0.1357413  -0.781   0.4350
## PC6          0.0333261  0.1615217   0.206   0.8365
## PC7          0.1888289  0.1597133   1.182   0.2371
## PC8          0.0186202  0.1622029   0.115   0.9086
## PC9         -0.2854233  0.1704022  -1.675   0.0939 .
## PC10        -0.3374949  0.1770795  -1.906   0.0567 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221.81  on 159  degrees of freedom
## Residual deviance: 204.42  on 149  degrees of freedom
## AIC: 226.42
##
## Number of Fisher Scoring iterations: 4
```

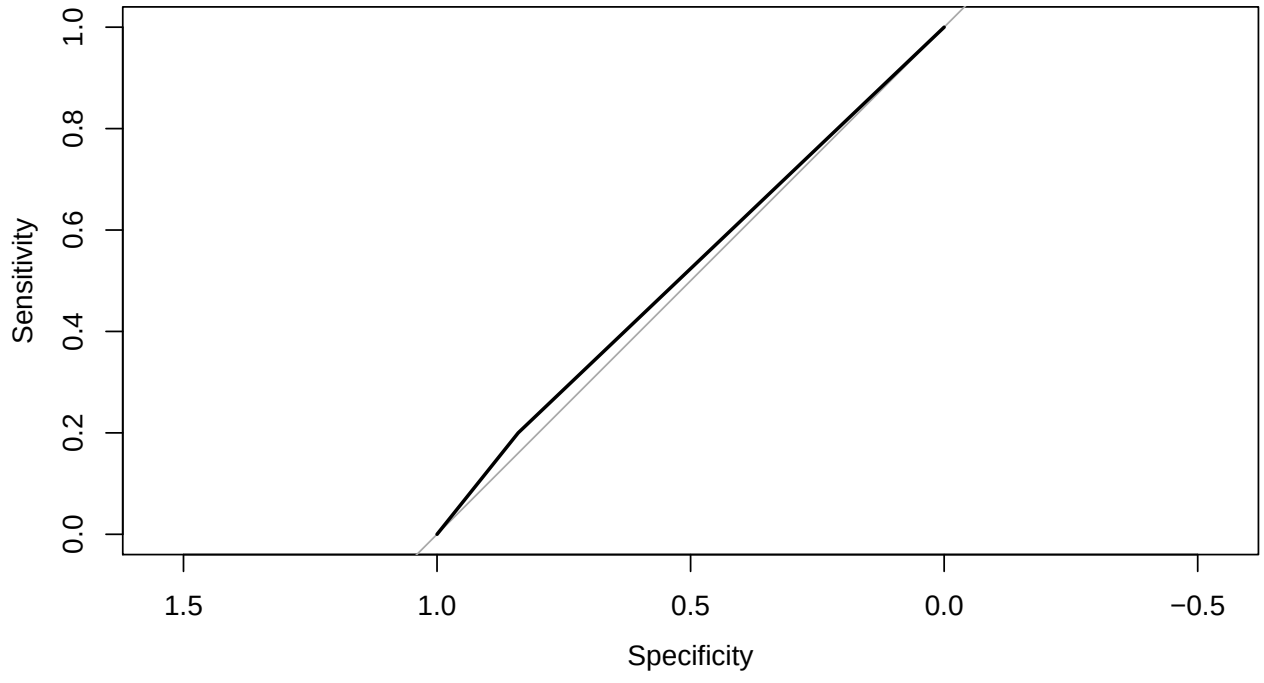
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
## Area under the curve: 0.6
```




```
## [1] "Accuracy 0.771428571428571"
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 25  8
##          1  0  2
##
##                Accuracy : 0.7714
##                  95% CI : (0.5986, 0.8958)
##     No Information Rate : 0.7143
##     P-Value [Acc > NIR] : 0.29413
##
##                   Kappa : 0.2632
##
##  Mcnemar's Test P-Value : 0.01333
##
##             Sensitivity : 1.0000
##             Specificity : 0.2000
##          Pos Pred Value : 0.7576
##          Neg Pred Value : 1.0000
##              Prevalence : 0.7143
##          Detection Rate : 0.7143
##    Detection Prevalence : 0.9429
##       Balanced Accuracy : 0.6000
##
##        'Positive' Class : 0
##
##
```

```
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = under_PCATrain)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.694  -1.068  -0.651   1.047   1.900
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.07491    0.33613   0.223    0.824
## PC1         -0.12135    0.14051  -0.864    0.388
## PC2         -0.01274    0.17873  -0.071    0.943
## PC3         -0.03320    0.23622  -0.141    0.888
## PC4          0.18328    0.25026   0.732    0.464
## PC5         -0.22087    0.26782  -0.825    0.410
## PC6          0.37607    0.36487   1.031    0.303
## PC7          0.18018    0.31420   0.573    0.566
## PC8         -0.51275    0.34342  -1.493    0.135
## PC9         -0.06642    0.36298  -0.183    0.855
## PC10        -0.15533    0.33186  -0.468    0.640
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.235  on 49  degrees of freedom
## Residual deviance: 63.313  on 39  degrees of freedom
## AIC: 85.313
##
## Number of Fisher Scoring iterations: 4

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

## Area under the curve: 0.52
```



```
## [1] "Accuracy 0.657142857142857"
```

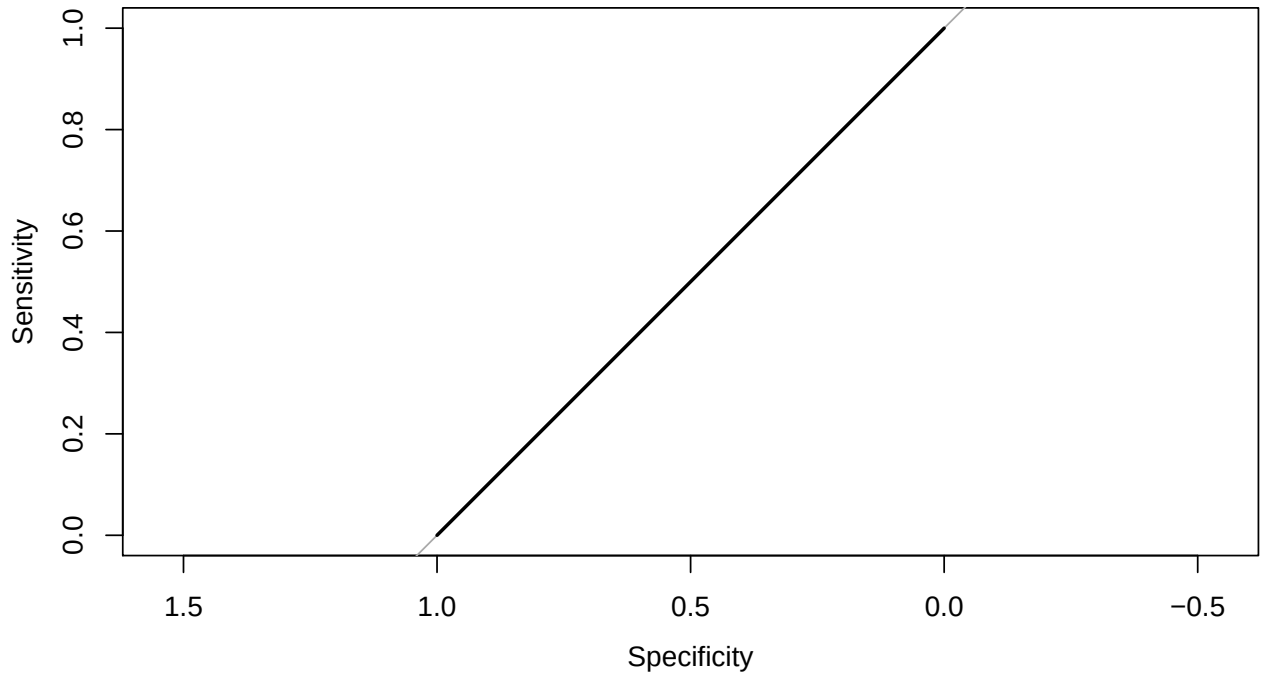
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  8
##          1  4  2
##
##                Accuracy : 0.6571
##                  95% CI : (0.4779, 0.8087)
##     No Information Rate : 0.7143
##     P-Value [Acc > NIR] : 0.8262
##
##                   Kappa : 0.0455
##
##  Mcnemar's Test P-Value : 0.3865
##
##             Sensitivity : 0.8400
##             Specificity : 0.2000
##          Pos Pred Value : 0.7241
##          Neg Pred Value : 0.3333
##              Prevalence : 0.7143
##          Detection Rate : 0.6000
##    Detection Prevalence : 0.8286
##       Balanced Accuracy : 0.5200
##
##        'Positive' Class : 0
##
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = both_PCATrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8145  -1.0274  -0.4638   0.9598   2.6655
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.44478    0.25252  -1.761   0.0782 .
## PC1          0.11341    0.10599   1.070   0.2846
## PC2          0.11208    0.12410   0.903   0.3664
## PC3         -0.30974    0.22317  -1.388   0.1652
## PC4          0.51897    0.20925   2.480   0.0131 *
## PC5         -0.14899    0.16603  -0.897   0.3695
## PC6          0.03106    0.20869   0.149   0.8817
## PC7         -0.21488    0.22988  -0.935   0.3499
## PC8          0.32573    0.23533   1.384   0.1663
## PC9         -0.58399    0.23339  -2.502   0.0123 *
## PC10        -0.05874    0.20790  -0.283   0.7775
## ---
## 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 104  degrees of freedom
```

```
## Residual deviance: 127.36  on  94  degrees of freedom
## AIC: 149.36
##
## Number of Fisher Scoring iterations: 4
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

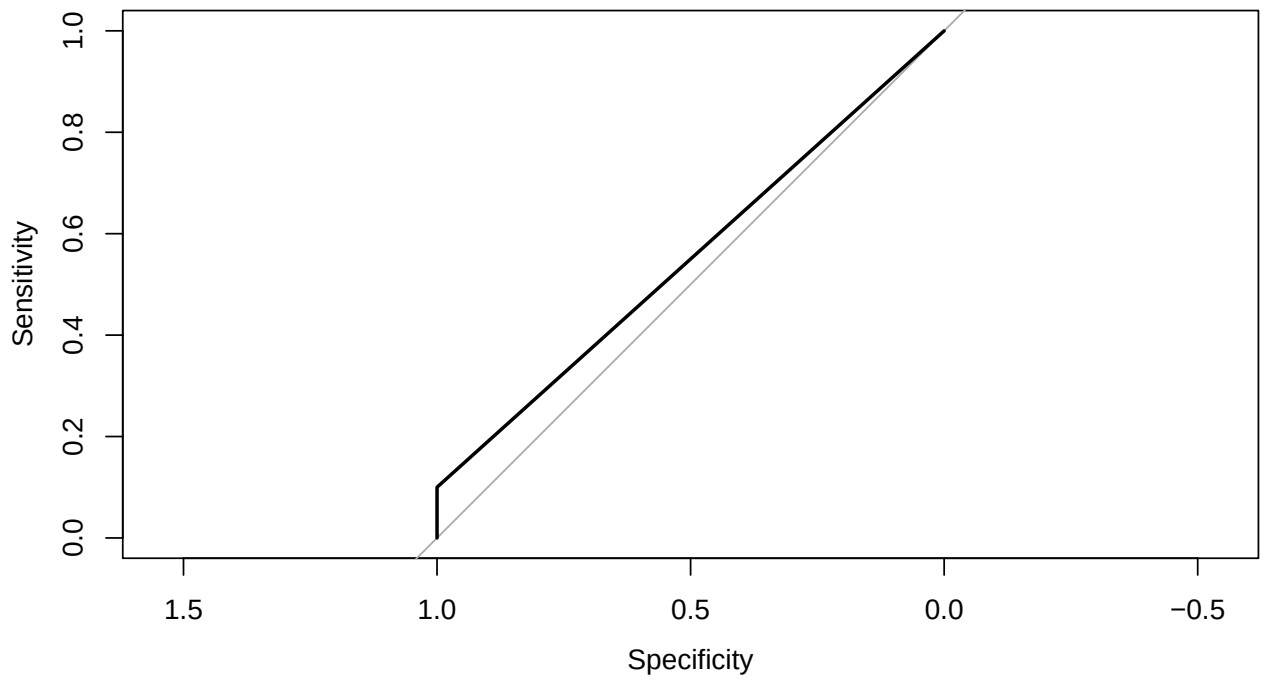
```
## Area under the curve: 0.5
```




```
## [1] "Accuracy 0.714285714285714"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 25 10
##          1  0  0
##
##                Accuracy : 0.7143
##                  95% CI : (0.537, 0.8536)
##     No Information Rate : 0.7143
##     P-Value [Acc > NIR] : 0.584223
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.004427
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.7143
##          Neg Pred Value :    NaN
##              Prevalence : 0.7143
##          Detection Rate : 0.7143
##    Detection Prevalence : 1.0000
```

```
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = rose_PCATrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8784  -1.0132  -0.4639   0.8845   2.7218
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.46491    0.25719  -1.808  0.07066 .
## PC1          0.11995    0.10761   1.115  0.26500
## PC2          0.12857    0.12613   1.019  0.30805
## PC3         -0.25060    0.22736  -1.102  0.27037
## PC4          0.52927    0.21172   2.500  0.01242 *
## PC5         -0.14367    0.16864  -0.852  0.39427
## PC6         -0.01642    0.21417  -0.077  0.93887
## PC7         -0.29018    0.23400  -1.240  0.21495
## PC8          0.39364    0.24520   1.605  0.10841
## PC9         -0.69702    0.25335  -2.751  0.00594 **
## PC10        -0.04091    0.20953  -0.195  0.84521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.02  on 103  degrees of freedom
## Residual deviance: 123.84  on  93  degrees of freedom
## AIC: 145.84
##
## Number of Fisher Scoring iterations: 4

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

## Area under the curve: 0.55
```




```
## [1] "Accuracy 0.742857142857143"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 25  9
##          1  0  1
##
##                Accuracy : 0.7429
##                  95% CI : (0.5674, 0.8751)
##     No Information Rate : 0.7143
##     P-Value [Acc > NIR] : 0.436332
##
##                   Kappa : 0.137
##
##  Mcnemar's Test P-Value : 0.007661
##
##             Sensitivity : 1.0000
##             Specificity : 0.1000
##          Pos Pred Value : 0.7353
##          Neg Pred Value : 1.0000
##              Prevalence : 0.7143
##          Detection Rate : 0.7143
##    Detection Prevalence : 0.9714
##       Balanced Accuracy : 0.5500
##
##        'Positive' Class : 0
##
```

The best model in here is the Rose_PCA model which uses the Rose package and the PCA dataset. It has an accuracy of 71% (sensitivity of .82 and specificity of .29). This is not much better than just saying that all the songs lose. Because there are 4-5 songs nominated each year, you would be correct with around 75-80%

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 23  8
##          1  3  0
##
##                Accuracy : 0.6765
##                  95% CI : (0.4947, 0.8261)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9174
##
##                   Kappa : -0.1472
##
##  Mcnemar's Test P-Value : 0.2278
##
##             Sensitivity : 0.8846
##             Specificity : 0.0000
##          Pos Pred Value : 0.7419
##          Neg Pred Value : 0.0000
##              Prevalence : 0.7647
##          Detection Rate : 0.6765
##    Detection Prevalence : 0.9118
##       Balanced Accuracy : 0.4423
##
##        'Positive' Class : 0
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 23  8
##          1  3  0
##
##                Accuracy : 0.6765
##                  95% CI : (0.4947, 0.8261)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9174
##
##                   Kappa : -0.1472
##
##  Mcnemar's Test P-Value : 0.2278
##
##             Sensitivity : 0.8846
##             Specificity : 0.0000
##          Pos Pred Value : 0.7419
##          Neg Pred Value : 0.0000
##              Prevalence : 0.7647
##          Detection Rate : 0.6765
##    Detection Prevalence : 0.9118
##       Balanced Accuracy : 0.4423
##
##        'Positive' Class : 0
##
```

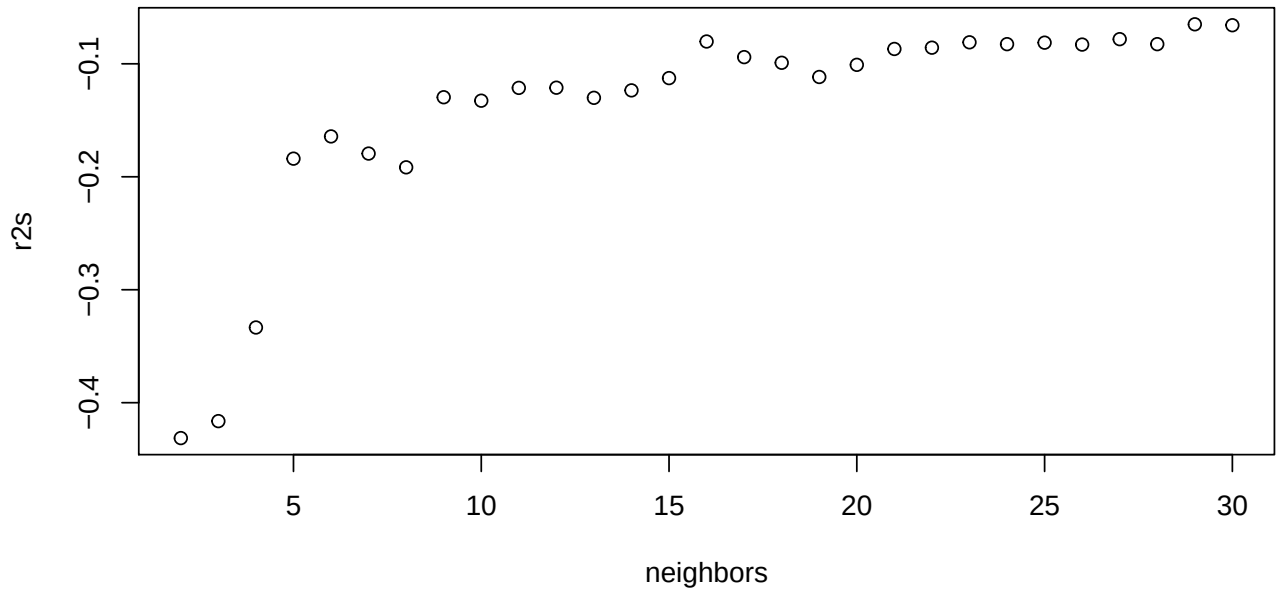
The mlr learning on the Naive Bayes approach has the highest specificity (correctly predicting winning songs) even though the model had an accuracy of 68%.

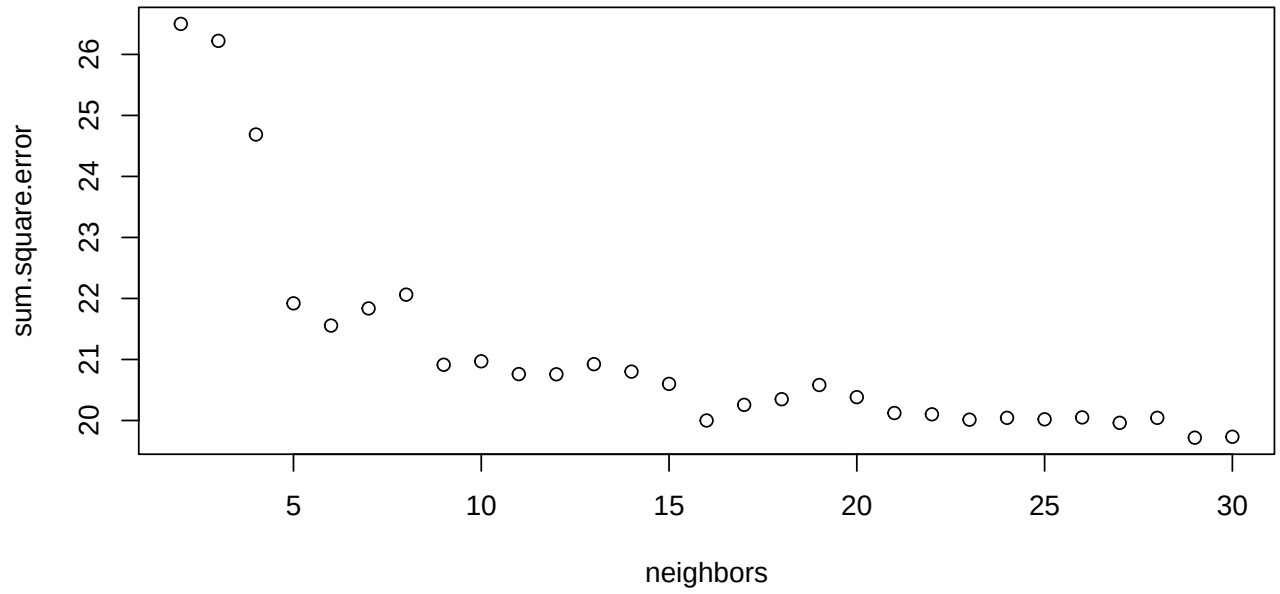
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 16  6
##          1 11  2
##
##                Accuracy : 0.5143
##                  95% CI : (0.3399, 0.6862)
##     No Information Rate : 0.7714
##     P-Value [Acc > NIR] : 0.9998
##
##                   Kappa : -0.129
##
##  Mcnemar's Test P-Value : 0.3320
##
##             Sensitivity : 0.5926
##             Specificity : 0.2500
##          Pos Pred Value : 0.7273
##          Neg Pred Value : 0.1538
##              Prevalence : 0.7714
##          Detection Rate : 0.4571
##    Detection Prevalence : 0.6286
##       Balanced Accuracy : 0.4213
##
##        'Positive' Class : 0
##
```

KNN

PCA Datasets

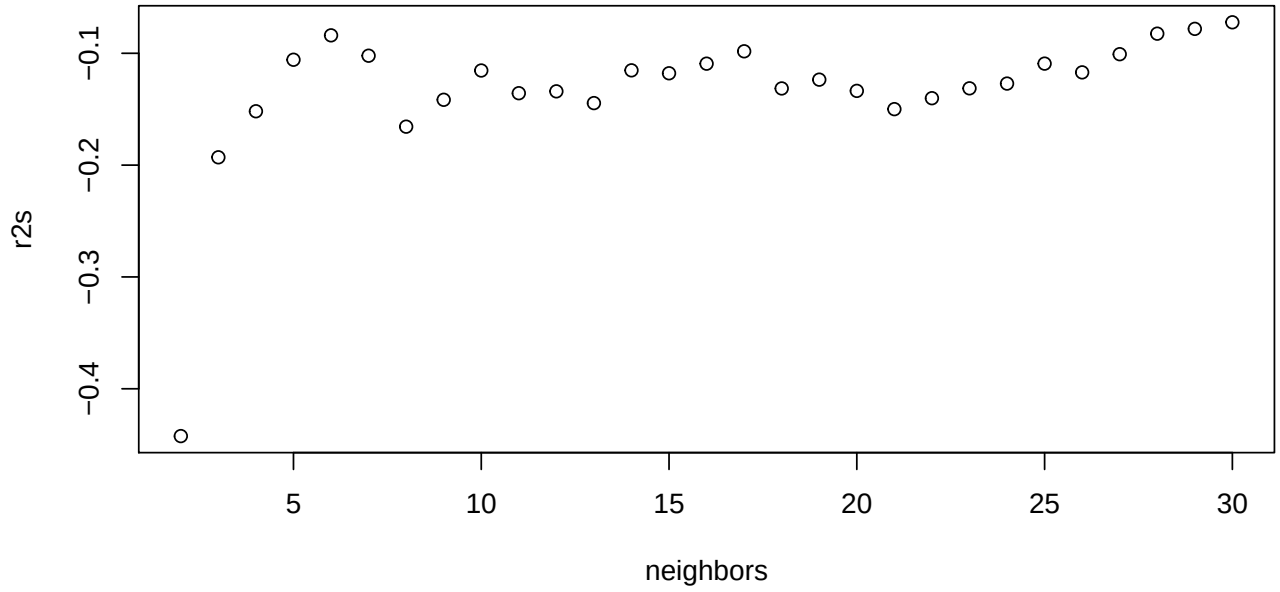
Scale Datasets

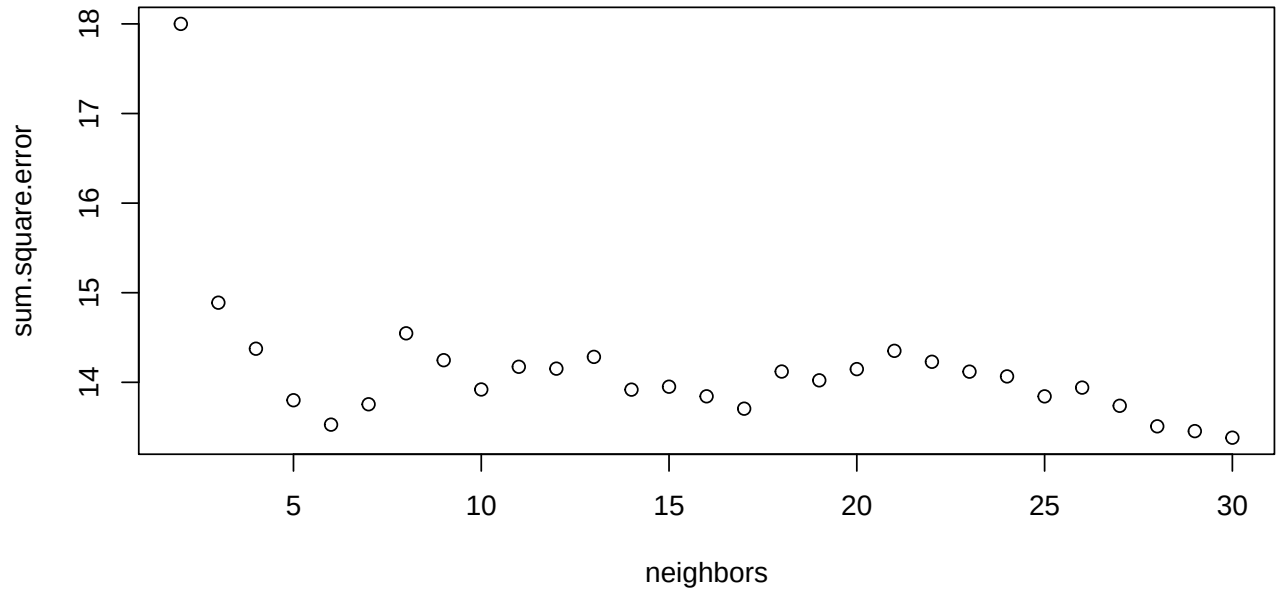





```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 23  7
##          1  3  1
##
##                Accuracy : 0.7059
```

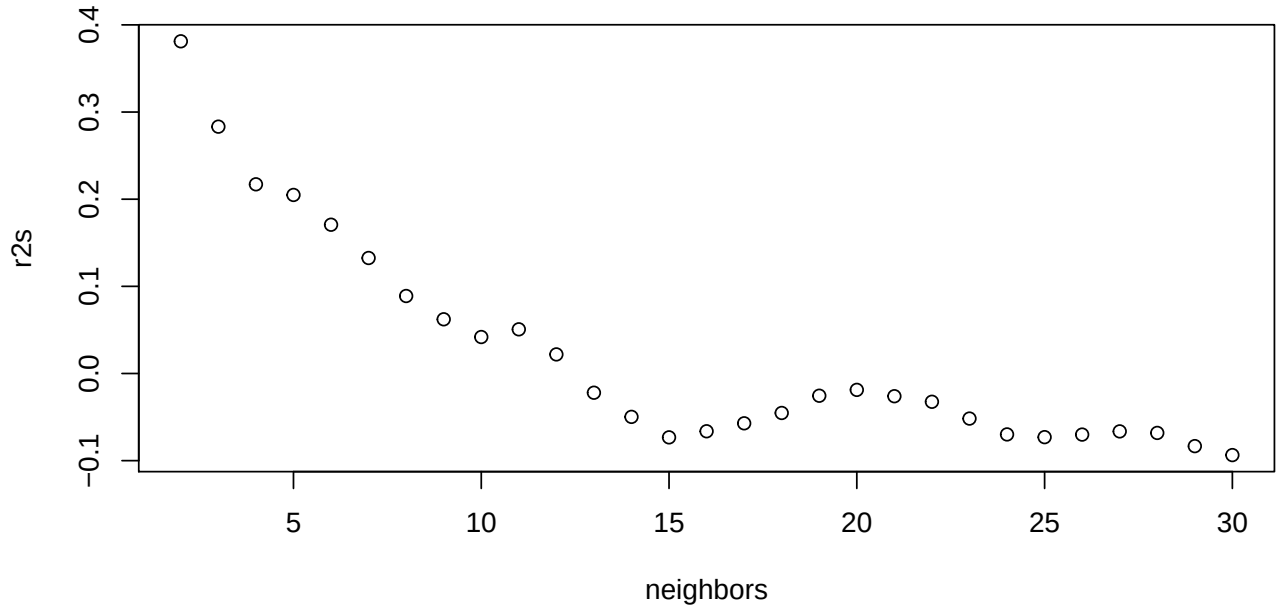
```
##                 95% CI : (0.5252, 0.849)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.8442
##
##                   Kappa : 0.0116
##
##  Mcnemar's Test P-Value : 0.3428
##
##             Sensitivity : 0.8846
##             Specificity : 0.1250
##          Pos Pred Value : 0.7667
##          Neg Pred Value : 0.2500
##              Prevalence : 0.7647
##          Detection Rate : 0.6765
##    Detection Prevalence : 0.8824
##       Balanced Accuracy : 0.5048
##
##        'Positive' Class : 0
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 26  8
##          1  0  0
##
##                Accuracy : 0.7647
##                  95% CI : (0.5883, 0.8925)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.59339
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.01333
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.7647
##          Neg Pred Value :    NaN
##              Prevalence : 0.7647
##          Detection Rate : 0.7647
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

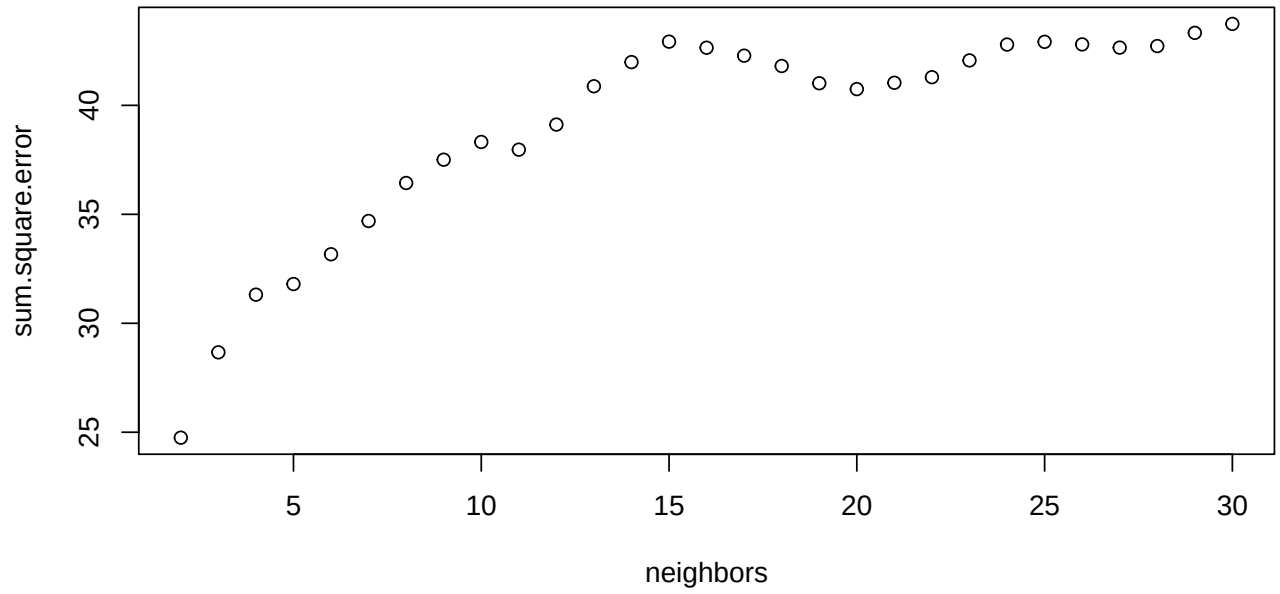





```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 18  5
##          1  8  3
##
##                Accuracy : 0.6176
```

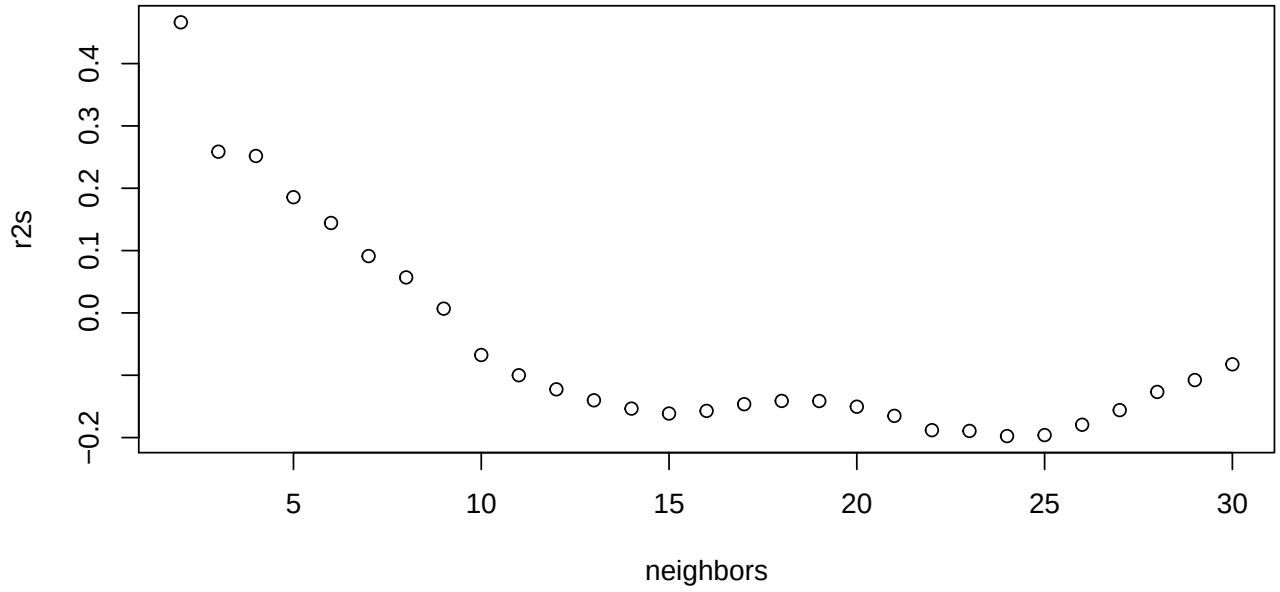
```
##                 95% CI : (0.4356, 0.7783)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9831
##
##                   Kappa : 0.0596
##
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.6923
##             Specificity : 0.3750
##          Pos Pred Value : 0.7826
##          Neg Pred Value : 0.2727
##              Prevalence : 0.7647
##          Detection Rate : 0.5294
##    Detection Prevalence : 0.6765
##       Balanced Accuracy : 0.5337
##
##        'Positive' Class : 0
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 18  5
##          1  8  3
##
##                Accuracy : 0.6176
##                  95% CI : (0.4356, 0.7783)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9831
##
##                   Kappa : 0.0596
##
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.6923
##             Specificity : 0.3750
##          Pos Pred Value : 0.7826
##          Neg Pred Value : 0.2727
##              Prevalence : 0.7647
##          Detection Rate : 0.5294
##    Detection Prevalence : 0.6765
##       Balanced Accuracy : 0.5337
##
##        'Positive' Class : 0
##
```

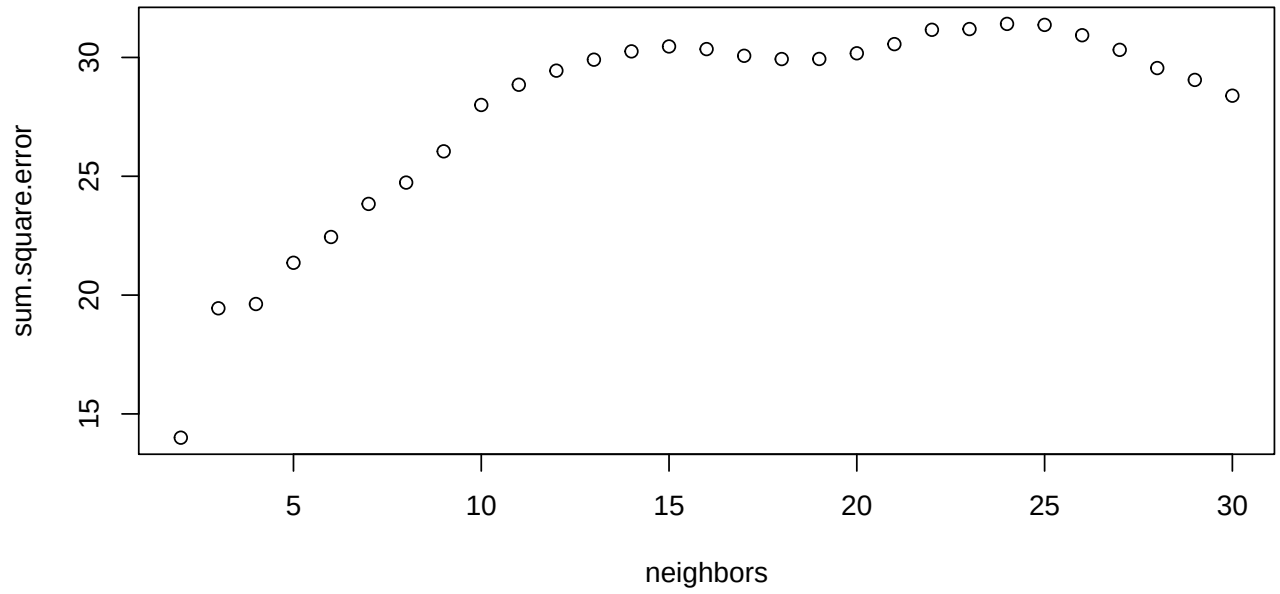





```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  4
##          1  7  4
##
##                Accuracy : 0.6765
```

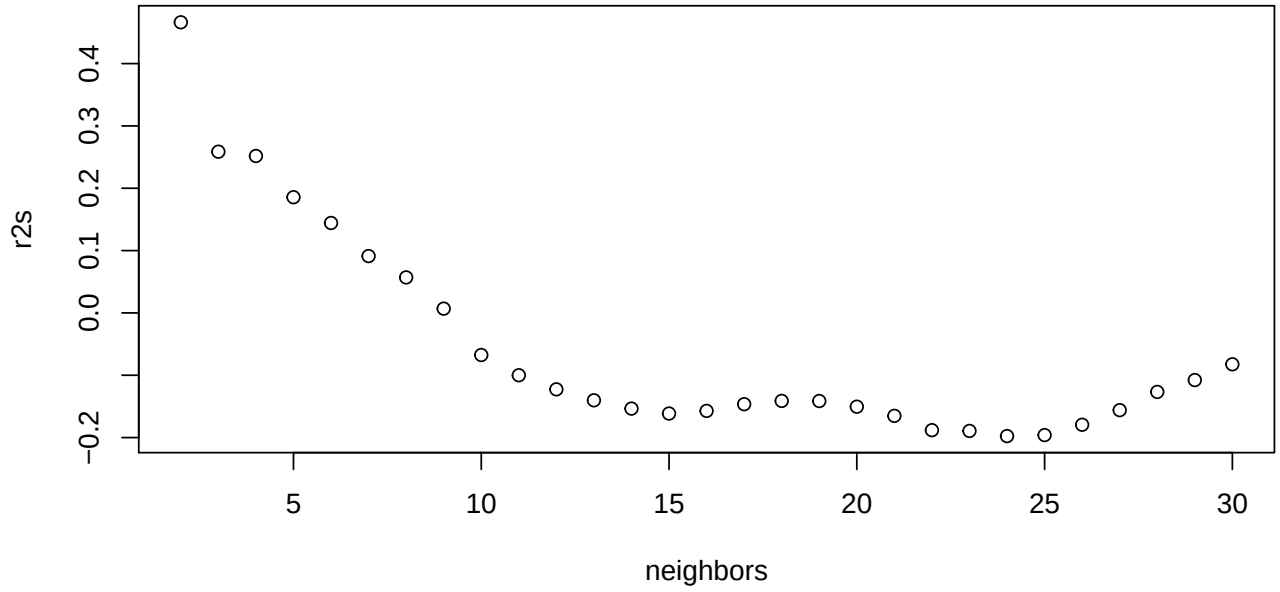
```
##                 95% CI : (0.4947, 0.8261)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9174
##
##                   Kappa : 0.2043
##
##  Mcnemar's Test P-Value : 0.5465
##
##             Sensitivity : 0.7308
##             Specificity : 0.5000
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.3636
##              Prevalence : 0.7647
##          Detection Rate : 0.5588
##    Detection Prevalence : 0.6765
##       Balanced Accuracy : 0.6154
##
##        'Positive' Class : 0
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  4
##          1  7  4
##
##                Accuracy : 0.6765
##                  95% CI : (0.4947, 0.8261)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9174
##
##                   Kappa : 0.2043
##
##  Mcnemar's Test P-Value : 0.5465
##
##             Sensitivity : 0.7308
##             Specificity : 0.5000
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.3636
##              Prevalence : 0.7647
##          Detection Rate : 0.5588
##    Detection Prevalence : 0.6765
##       Balanced Accuracy : 0.6154
##
##        'Positive' Class : 0
##
```

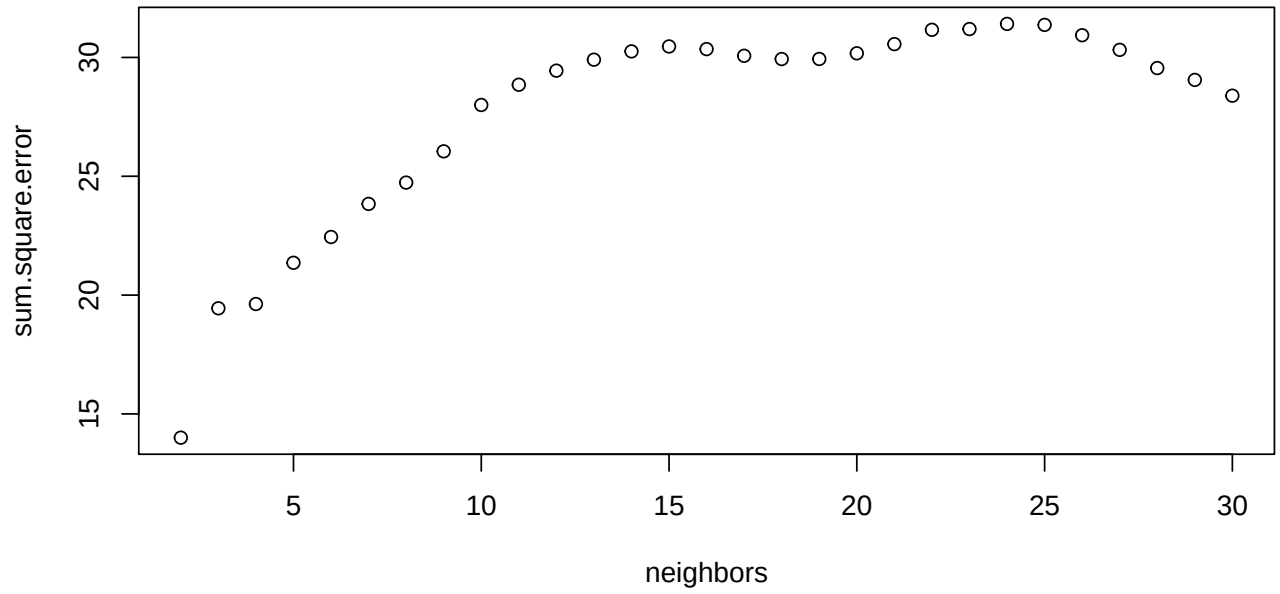





```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 15  4
##          1 11  4
##
##                Accuracy : 0.5588
```

```
##                 95% CI : (0.3789, 0.7281)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9977
##
##                   Kappa : 0.059
##
##  Mcnemar's Test P-Value : 0.1213
##
##             Sensitivity : 0.5769
##             Specificity : 0.5000
##          Pos Pred Value : 0.7895
##          Neg Pred Value : 0.2667
##              Prevalence : 0.7647
##          Detection Rate : 0.4412
##    Detection Prevalence : 0.5588
##       Balanced Accuracy : 0.5385
##
##        'Positive' Class : 0
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 15  4
##          1 11  4
##
##                Accuracy : 0.5588
##                  95% CI : (0.3789, 0.7281)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9977
##
##                   Kappa : 0.059
##
##  Mcnemar's Test P-Value : 0.1213
##
##             Sensitivity : 0.5769
##             Specificity : 0.5000
##          Pos Pred Value : 0.7895
##          Neg Pred Value : 0.2667
##              Prevalence : 0.7647
##          Detection Rate : 0.4412
##    Detection Prevalence : 0.5588
##       Balanced Accuracy : 0.5385
##
##        'Positive' Class : 0
##
```






```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 15  4
##          1 11  4
##
##                Accuracy : 0.5588
```

```
##                 95% CI : (0.3789, 0.7281)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.9977
##
##                   Kappa : 0.059
##
##  Mcnemar's Test P-Value : 0.1213
##
##             Sensitivity : 0.5769
##             Specificity : 0.5000
##          Pos Pred Value : 0.7895
##          Neg Pred Value : 0.2667
##              Prevalence : 0.7647
##          Detection Rate : 0.4412
##    Detection Prevalence : 0.5588
##       Balanced Accuracy : 0.5385
##
##        'Positive' Class : 0
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 11  4
##          1 15  4
##
##                Accuracy : 0.4412
##                  95% CI : (0.2719, 0.6211)
##     No Information Rate : 0.7647
##     P-Value [Acc > NIR] : 0.99999
##
##                   Kappa : -0.0521
##
##  Mcnemar's Test P-Value : 0.02178
##
##             Sensitivity : 0.4231
##             Specificity : 0.5000
##          Pos Pred Value : 0.7333
##          Neg Pred Value : 0.2105
##              Prevalence : 0.7647
##          Detection Rate : 0.3235
##    Detection Prevalence : 0.4412
##       Balanced Accuracy : 0.4615
##
##        'Positive' Class : 0
##
```