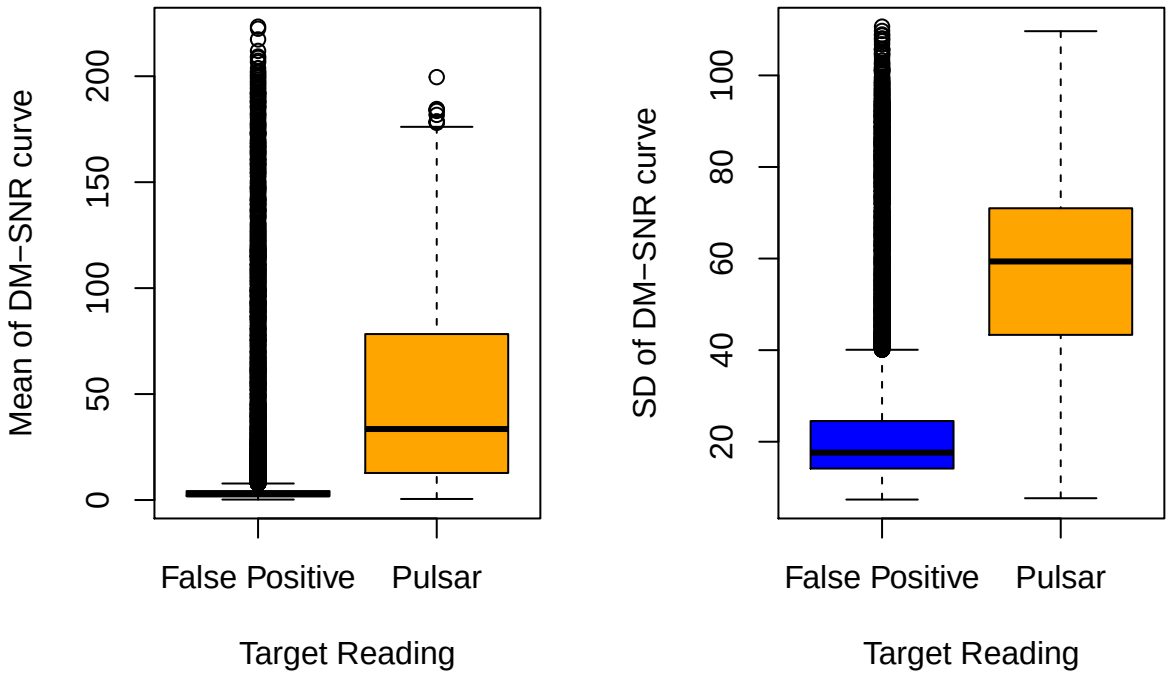
Appendix

Conner McCloney

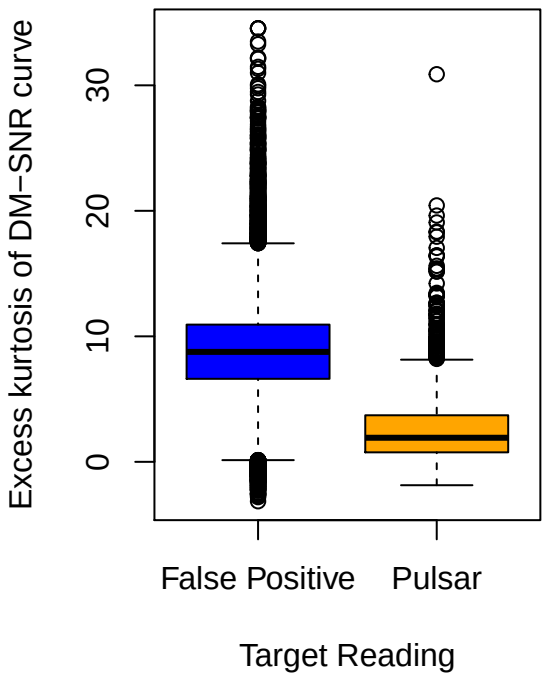
```
pulsar_data <- read.csv("pulsar_stars.csv")
```

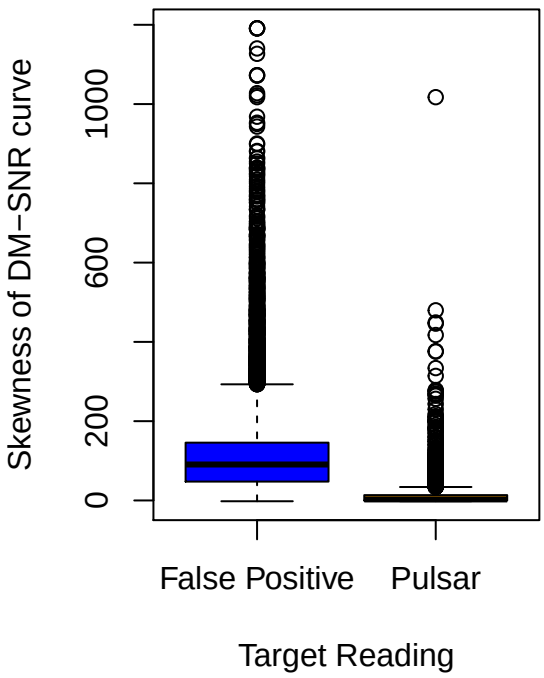
Exploratory Plots

```
pulsar_data$target_class_label[pulsar_data$target_class==1] <- "Pulsar"
pulsar_data$target_class_label[pulsar_data$target_class==0] <- "False Positive"
#DM-SNR Curve
par(mfrow=c(1,2))
boxplot(Mean.of.the.DM.SNR.curve~target_class_label,data=pulsar_data,ylab="Mean of DM-SNR curve", xlab=
boxplot(Standard.deviation.of.the.DM.SNR.curve~target_class_label,data=pulsar_data, ylab="SD of DM-SNR
```

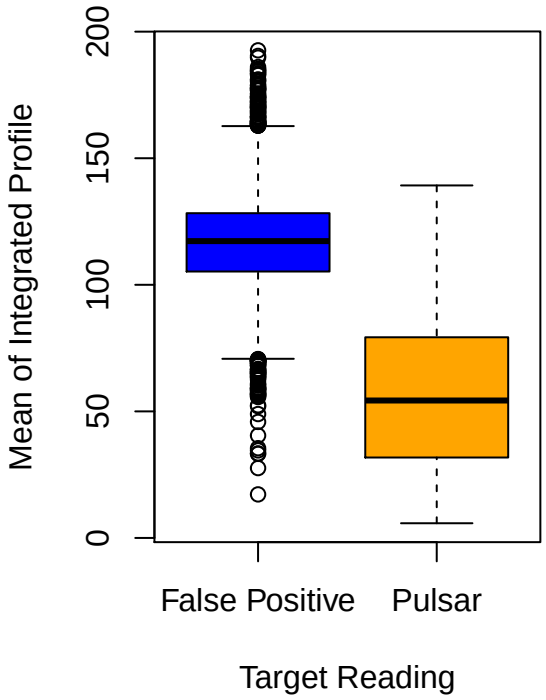


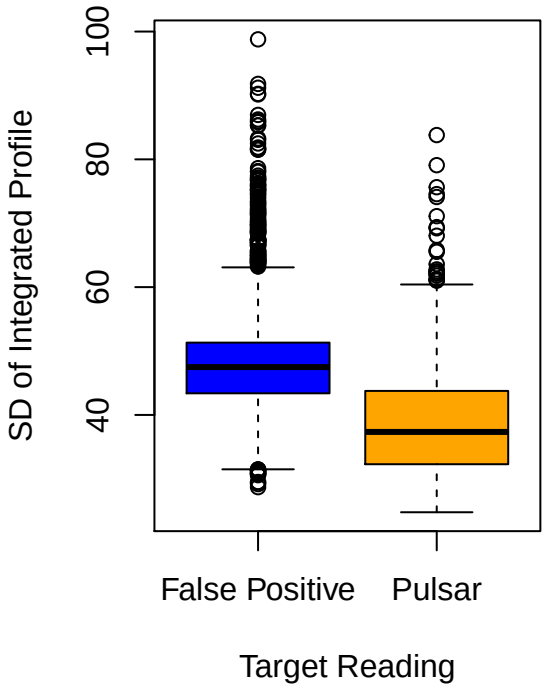

```
par(mfrow=c(1,2))
boxplot(Excess.kurtosis.of.the.DM.SNR.curve~target_class_label,data=pulsar_data, ylab="Excess kurtosis
boxplot(Skewness.of.the.DM.SNR.curve~target_class_label,data=pulsar_data, ylab="Skewness of DM-SNR curv
```



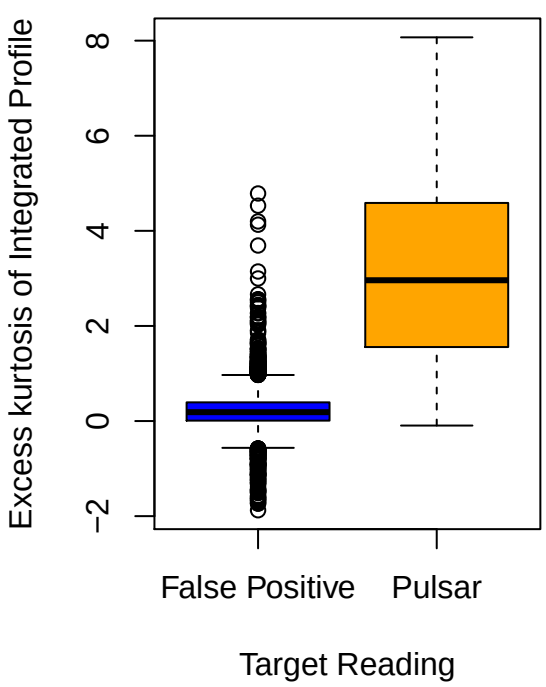


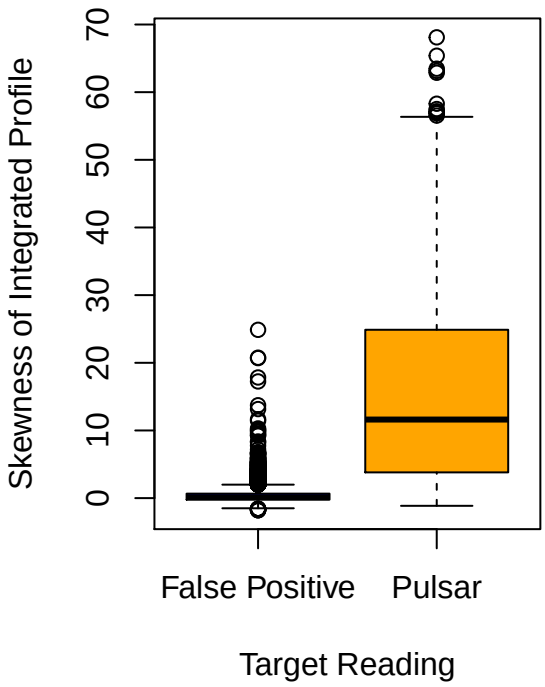

```
par(mfrow=c(1,2))
boxplot(Mean.of.the.integrated.profile~target_class_label,data=pulsar_data,ylab="Mean of Integrated Pro
boxplot(Standard.deviation.of.the.integrated.profile~target_class_label,data=pulsar_data, ylab="SD of I
```






```
par(mfrow=c(1,2))
boxplot(Excess.kurtosis.of.the.integrated.profile~target_class_label,data=pulsar_data, ylab="Excess kur
boxplot(Skewness.of.the.integrated.profile~target_class_label,data=pulsar_data, ylab="Skewness of Integ
```





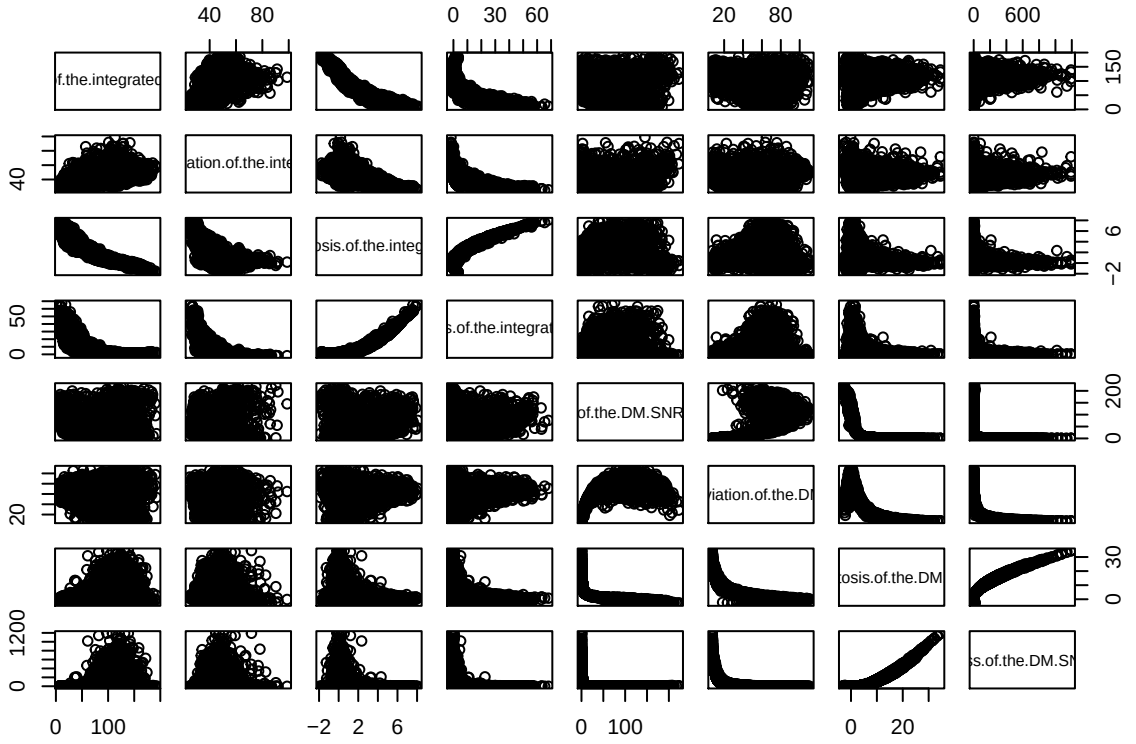
Multicollinearity

```
suppressMessages(library(car))
temp <- subset( pulsar_data, select = -c(target_class_label,target_class))
cor(temp)
```

```
##                                                Mean.of.the.integrated.profile
## Mean.of.the.integrated.profile                                      1.0000000
## Standard.deviation.of.the.integrated.profile                        0.5471369
## Excess.kurtosis.of.the.integrated.profile                          -0.8738984
## Skewness.of.the.integrated.profile                                 -0.7387748
## Mean.of.the.DM.SNR.curve                                           -0.2988408
## Standard.deviation.of.the.DM.SNR.curve                             -0.3070158
## Excess.kurtosis.of.the.DM.SNR.curve                                 0.2343312
## Skewness.of.the.DM.SNR.curve                                        0.1440330
##                                                Standard.deviation.of.the.integrated.profile
## Mean.of.the.integrated.profile                                                    0.547136926
## Standard.deviation.of.the.integrated.profile                                      1.000000000
## Excess.kurtosis.of.the.integrated.profile                                        -0.521435272
## Skewness.of.the.integrated.profile                                               -0.539792970
## Mean.of.the.DM.SNR.curve                                                          0.006868735
## Standard.deviation.of.the.DM.SNR.curve                                           -0.047631587
## Excess.kurtosis.of.the.DM.SNR.curve                                               0.029429387
## Skewness.of.the.DM.SNR.curve                                                      0.027691480
```

```
##                                                 Excess.kurtosis.of.the.integrated.profile
## Mean.of.the.integrated.profile                                                 -0.8738984
## Standard.deviation.of.the.integrated.profile                                   -0.5214353
## Excess.kurtosis.of.the.integrated.profile                                       1.0000000
## Skewness.of.the.integrated.profile                                              0.9457291
## Mean.of.the.DM.SNR.curve                                                        0.4143676
## Standard.deviation.of.the.DM.SNR.curve                                          0.4328802
## Excess.kurtosis.of.the.DM.SNR.curve                                            -0.3412090
## Skewness.of.the.DM.SNR.curve                                                   -0.2144909
##                                                 Skewness.of.the.integrated.profile
## Mean.of.the.integrated.profile                                          -0.7387748
## Standard.deviation.of.the.integrated.profile                            -0.5397930
## Excess.kurtosis.of.the.integrated.profile                                0.9457291
## Skewness.of.the.integrated.profile                                       1.0000000
## Mean.of.the.DM.SNR.curve                                                 0.4120564
## Standard.deviation.of.the.DM.SNR.curve                                   0.4151400
## Excess.kurtosis.of.the.DM.SNR.curve                                     -0.3288433
## Skewness.of.the.DM.SNR.curve                                            -0.2047825
##                                                 Mean.of.the.DM.SNR.curve
## Mean.of.the.integrated.profile                              -0.298840844
## Standard.deviation.of.the.integrated.profile                 0.006868735
## Excess.kurtosis.of.the.integrated.profile                    0.414367611
## Skewness.of.the.integrated.profile                           0.412056437
## Mean.of.the.DM.SNR.curve                                     1.000000000
## Standard.deviation.of.the.DM.SNR.curve                       0.796554844
## Excess.kurtosis.of.the.DM.SNR.curve                         -0.615970831
## Skewness.of.the.DM.SNR.curve                                -0.354269152
##                                                 Standard.deviation.of.the.DM.SNR.curve
## Mean.of.the.integrated.profile                                              -0.30701583
## Standard.deviation.of.the.integrated.profile                                -0.04763159
## Excess.kurtosis.of.the.integrated.profile                                    0.43288016
## Skewness.of.the.integrated.profile                                           0.41513996
## Mean.of.the.DM.SNR.curve                                                     0.79655484
## Standard.deviation.of.the.DM.SNR.curve                                       1.00000000
## Excess.kurtosis.of.the.DM.SNR.curve                                         -0.80978582
## Skewness.of.the.DM.SNR.curve                                                -0.57579983
##                                                 Excess.kurtosis.of.the.DM.SNR.curve
## Mean.of.the.integrated.profile                                           0.23433123
## Standard.deviation.of.the.integrated.profile                             0.02942939
## Excess.kurtosis.of.the.integrated.profile                               -0.34120902
## Skewness.of.the.integrated.profile                                      -0.32884331
## Mean.of.the.DM.SNR.curve                                                -0.61597083
## Standard.deviation.of.the.DM.SNR.curve                                  -0.80978582
## Excess.kurtosis.of.the.DM.SNR.curve                                      1.00000000
## Skewness.of.the.DM.SNR.curve                                             0.92374273
##                                                 Skewness.of.the.DM.SNR.curve
## Mean.of.the.integrated.profile                                    0.14403302
## Standard.deviation.of.the.integrated.profile                      0.02769148
## Excess.kurtosis.of.the.integrated.profile                        -0.21449091
## Skewness.of.the.integrated.profile                               -0.20478247
## Mean.of.the.DM.SNR.curve                                         -0.35426915
## Standard.deviation.of.the.DM.SNR.curve                           -0.57579983
## Excess.kurtosis.of.the.DM.SNR.curve                               0.92374273
## Skewness.of.the.DM.SNR.curve                                      1.00000000
```

```r
pairs(temp)
```



Models

From exploratory plots, all seem useful in detecting difference between False Positive and true Pulsar reading.

```r
full.model <- glm(target_class~Mean.of.the.integrated.profile+Standard.deviation.of.the.integrated.profi
                  Excess.kurtosis.of.the.integrated.profile+Skewness.of.the.integrated.profile+
                  Mean.of.the.DM.SNR.curve+Standard.deviation.of.the.DM.SNR.curve+
                  Excess.kurtosis.of.the.DM.SNR.curve+Skewness.of.the.DM.SNR.curve,data=pulsar_data ,family
summary(full.model)
```

```
##
## Call:
## glm(formula = target_class ~ Mean.of.the.integrated.profile +
##     Standard.deviation.of.the.integrated.profile + Excess.kurtosis.of.the.integrated.profile +
##     Skewness.of.the.integrated.profile + Mean.of.the.DM.SNR.curve +
##     Standard.deviation.of.the.DM.SNR.curve + Excess.kurtosis.of.the.DM.SNR.curve +
##     Skewness.of.the.DM.SNR.curve, family = binomial(link = "logit"),
##     data = pulsar_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3813  -0.1644  -0.1000  -0.0565   3.6178
##
```

```
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -9.019954   0.977017  -9.232
## Mean.of.the.integrated.profile               0.030260   0.005925   5.107
## Standard.deviation.of.the.integrated.profile -0.035430   0.010386  -3.411
## Excess.kurtosis.of.the.integrated.profile    6.577063   0.300110  21.916
## Skewness.of.the.integrated.profile          -0.616243   0.039709 -15.519
## Mean.of.the.DM.SNR.curve                    -0.028585   0.003268  -8.747
## Standard.deviation.of.the.DM.SNR.curve       0.053166   0.007364   7.220
## Excess.kurtosis.of.the.DM.SNR.curve          0.047749   0.085758   0.557
## Skewness.of.the.DM.SNR.curve                -0.004750   0.003051  -1.557
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## Mean.of.the.integrated.profile              3.27e-07 ***
## Standard.deviation.of.the.integrated.profile 0.000647 ***
## Excess.kurtosis.of.the.integrated.profile    < 2e-16 ***
## Skewness.of.the.integrated.profile           < 2e-16 ***
## Mean.of.the.DM.SNR.curve                     < 2e-16 ***
## Standard.deviation.of.the.DM.SNR.curve      5.21e-13 ***
## Excess.kurtosis.of.the.DM.SNR.curve         0.577676
## Skewness.of.the.DM.SNR.curve                0.119474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10959.5  on 17897  degrees of freedom
## Residual deviance:  2615.8  on 17889  degrees of freedom
## AIC: 2633.8
##
## Number of Fisher Scoring iterations: 8
```

```
vif(full.model)
```

```
##               Mean.of.the.integrated.profile
##                                     4.207534
## Standard.deviation.of.the.integrated.profile
##                                     1.726942
##    Excess.kurtosis.of.the.integrated.profile
##                                    12.353817
##           Skewness.of.the.integrated.profile
##                                     6.980475
##                      Mean.of.the.DM.SNR.curve
##                                     3.963631
##       Standard.deviation.of.the.DM.SNR.curve
##                                     9.612362
##          Excess.kurtosis.of.the.DM.SNR.curve
##                                    37.028818
##                 Skewness.of.the.DM.SNR.curve
##                                    15.877252
```

```
#p-value based reduction
reduced.model1 <- glm(target_class~Mean.of.the.integrated.profile+Standard.deviation.of.the.integrated.
               Excess.kurtosis.of.the.integrated.profile+Skewness.of.the.integrated.profile+
               Mean.of.the.DM.SNR.curve+Standard.deviation.of.the.DM.SNR.curve,data=pulsar_data ,family
```

```r
summary(reduced.model1)
```

```
##
## Call:
## glm(formula = target_class ~ Mean.of.the.integrated.profile +
##     Standard.deviation.of.the.integrated.profile + Excess.kurtosis.of.the.integrated.profile +
##     Skewness.of.the.integrated.profile + Mean.of.the.DM.SNR.curve +
##     Standard.deviation.of.the.DM.SNR.curve, family = binomial(link = "logit"),
##     data = pulsar_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4113  -0.1643  -0.1021  -0.0593   3.6389
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                -9.122269   0.742402 -12.288
## Mean.of.the.integrated.profile              0.030509   0.005931   5.144
## Standard.deviation.of.the.integrated.profile -0.037829   0.010328  -3.663
## Excess.kurtosis.of.the.integrated.profile   6.575552   0.300899  21.853
## Skewness.of.the.integrated.profile         -0.613499   0.040346 -15.206
## Mean.of.the.DM.SNR.curve                   -0.031502   0.002968 -10.616
## Standard.deviation.of.the.DM.SNR.curve      0.059558   0.004004  14.874
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## Mean.of.the.integrated.profile             2.69e-07 ***
## Standard.deviation.of.the.integrated.profile 0.000249 ***
## Excess.kurtosis.of.the.integrated.profile   < 2e-16 ***
## Skewness.of.the.integrated.profile          < 2e-16 ***
## Mean.of.the.DM.SNR.curve                    < 2e-16 ***
## Standard.deviation.of.the.DM.SNR.curve      < 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: 10959.5  on 17897  degrees of freedom
## Residual deviance:  2626.7  on 17891  degrees of freedom
## AIC: 2640.7
##
## Number of Fisher Scoring iterations: 8
```

```r
vif(reduced.model1)
```

```
##               Mean.of.the.integrated.profile
##                                     4.235012
## Standard.deviation.of.the.integrated.profile
##                                     1.719797
##    Excess.kurtosis.of.the.integrated.profile
##                                    12.330768
##           Skewness.of.the.integrated.profile
##                                     7.006026
##                     Mean.of.the.DM.SNR.curve
##                                     3.309988
```

```
##       Standard.deviation.of.the.DM.SNR.curve
##                                     2.943439
```

```
#vif-based reduction
reduced.model2 <- glm(target_class~Mean.of.the.integrated.profile+Standard.deviation.of.the.integrated.p
              Mean.of.the.DM.SNR.curve+Standard.deviation.of.the.DM.SNR.curve,data=pulsar_data ,family
summary(reduced.model2)
```

```
##
## Call:
## glm(formula = target_class ~ Mean.of.the.integrated.profile +
##     Standard.deviation.of.the.integrated.profile + Mean.of.the.DM.SNR.curve +
##     Standard.deviation.of.the.DM.SNR.curve, family = binomial(link = "logit"),
##     data = pulsar_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6606  -0.1758  -0.0901  -0.0444   3.9198
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  4.706789   0.304829  15.441
## Mean.of.the.integrated.profile              -0.101844   0.003059 -33.299
## Standard.deviation.of.the.integrated.profile  0.007772   0.007876   0.987
## Mean.of.the.DM.SNR.curve                    -0.027967   0.002367 -11.816
## Standard.deviation.of.the.DM.SNR.curve       0.079145   0.003290  24.059
##                                             Pr(>|z|)
## (Intercept)                                   <2e-16 ***
## Mean.of.the.integrated.profile                <2e-16 ***
## Standard.deviation.of.the.integrated.profile   0.324
## Mean.of.the.DM.SNR.curve                      <2e-16 ***
## Standard.deviation.of.the.DM.SNR.curve        <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: 10959.5  on 17897  degrees of freedom
## Residual deviance:  3367.5  on 17893  degrees of freedom
## AIC: 3377.5
##
## Number of Fisher Scoring iterations: 8
```

```
vif(reduced.model2)
```

```
##               Mean.of.the.integrated.profile
##                                     1.465249
## Standard.deviation.of.the.integrated.profile
##                                     1.543699
##                      Mean.of.the.DM.SNR.curve
##                                     3.188099
##       Standard.deviation.of.the.DM.SNR.curve
##                                     2.941211
```

```
#p-value and vif-based reduction
reduced.model3 <- glm(target_class~Mean.of.the.integrated.profile+
```

```
              Mean.of.the.DM.SNR.curve+Standard.deviation.of.the.DM.SNR.curve,data=pulsar_data ,family
summary(reduced.model3)
```

```
##
## Call:
## glm(formula = target_class ~ Mean.of.the.integrated.profile +
##     Mean.of.the.DM.SNR.curve + Standard.deviation.of.the.DM.SNR.curve,
##     family = binomial(link = "logit"), data = pulsar_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6136  -0.1754  -0.0898  -0.0446   3.9207
##
## Coefficients:
##                                       Estimate Std. Error z value
## (Intercept)                           4.893816   0.239811   20.41
## Mean.of.the.integrated.profile       -0.100314   0.002615  -38.36
## Mean.of.the.DM.SNR.curve             -0.027346   0.002281  -11.99
## Standard.deviation.of.the.DM.SNR.curve  0.079113   0.003292   24.03
##                                      Pr(>|z|)
## (Intercept)                            <2e-16 ***
## Mean.of.the.integrated.profile         <2e-16 ***
## Mean.of.the.DM.SNR.curve               <2e-16 ***
## Standard.deviation.of.the.DM.SNR.curve   <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: 10959.5  on 17897  degrees of freedom
## Residual deviance:  3368.5  on 17894  degrees of freedom
## AIC: 3376.5
##
## Number of Fisher Scoring iterations: 8
```

```
vif(reduced.model3)
```

```
##         Mean.of.the.integrated.profile
##                                1.072640
##               Mean.of.the.DM.SNR.curve
##                                2.967376
## Standard.deviation.of.the.DM.SNR.curve
##                                2.956700
```

Model Testing

```
train=(pulsar_data[0:1250,])
test=pulsar_data[1251:17898,]

glm.probs.full=predict(full.model,test,type="response")
glm.pred.full=rep("0",16648)
glm.pred.full[glm.probs.full>.5]="1"
```

```
table(glm.pred.full,test$target_class,dnn=c("True Label","Predicted Label - Full Model"))
```

```
##           Predicted Label - Full Model
## True Label     0     1
##          0 15013   267
##          1    80  1288
```

```
glm.probs.r1=predict(reduced.model1,test,type="response")
glm.pred.r1=rep("0",16648)
glm.pred.r1[glm.probs.r1>.5]="1"
table(glm.pred.r1,test$target_class,dnn=c("True Label","Predicted Label - Reduced Model #1"))
```

```
##           Predicted Label - Reduced Model #1
## True Label     0     1
##          0 15014   270
##          1    79  1285
```

```
glm.probs.r2=predict(reduced.model2,test,type="response")
glm.pred.r2=rep("0",16648)
glm.pred.r2[glm.probs.r2>.5]="1"
table(glm.pred.r2,test$target_class,dnn=c("True Label","Predicted Label - Reduced Model #2"))
```

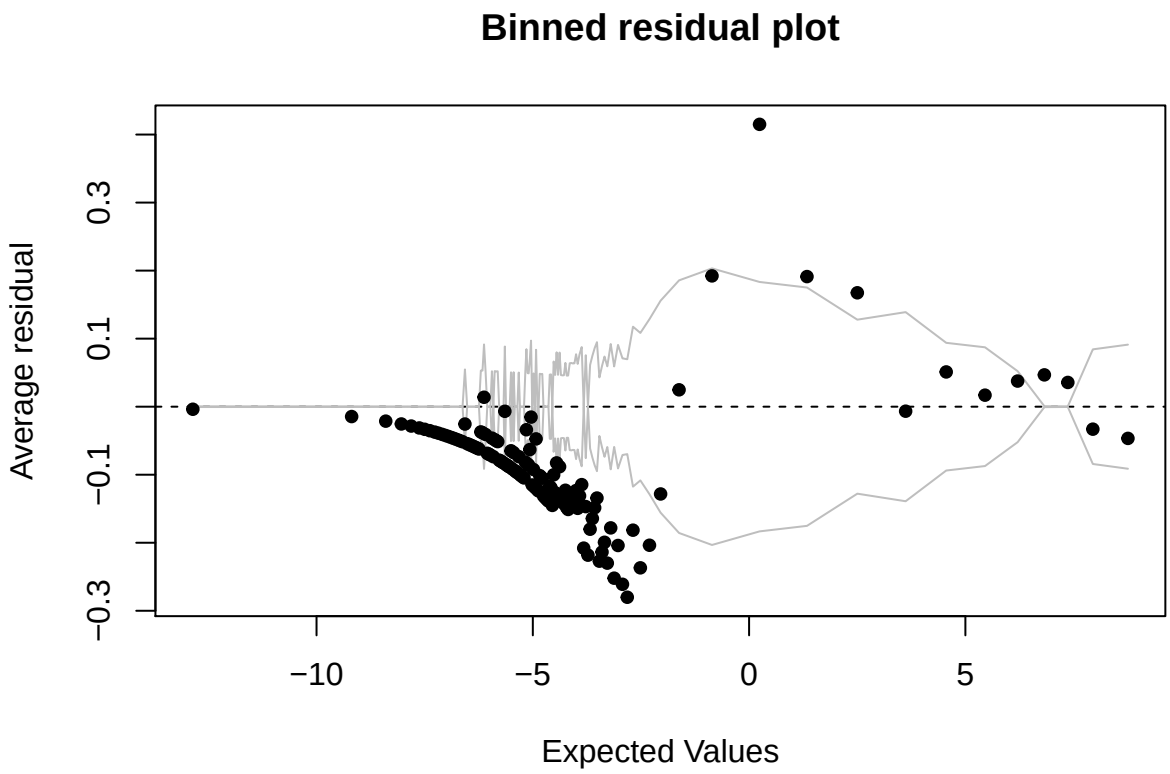
```
##           Predicted Label - Reduced Model #2
## True Label     0     1
##          0 14990   370
##          1   103  1185
```

```
glm.probs.r3=predict(reduced.model3,test,type="response")
glm.pred.r3=rep("0",16648)
glm.pred.r3[glm.probs.r3>.5]="1"
table(glm.pred.r3,test$target_class,dnn=c("True Label","Predicted Label - Reduced Model #3"))
```

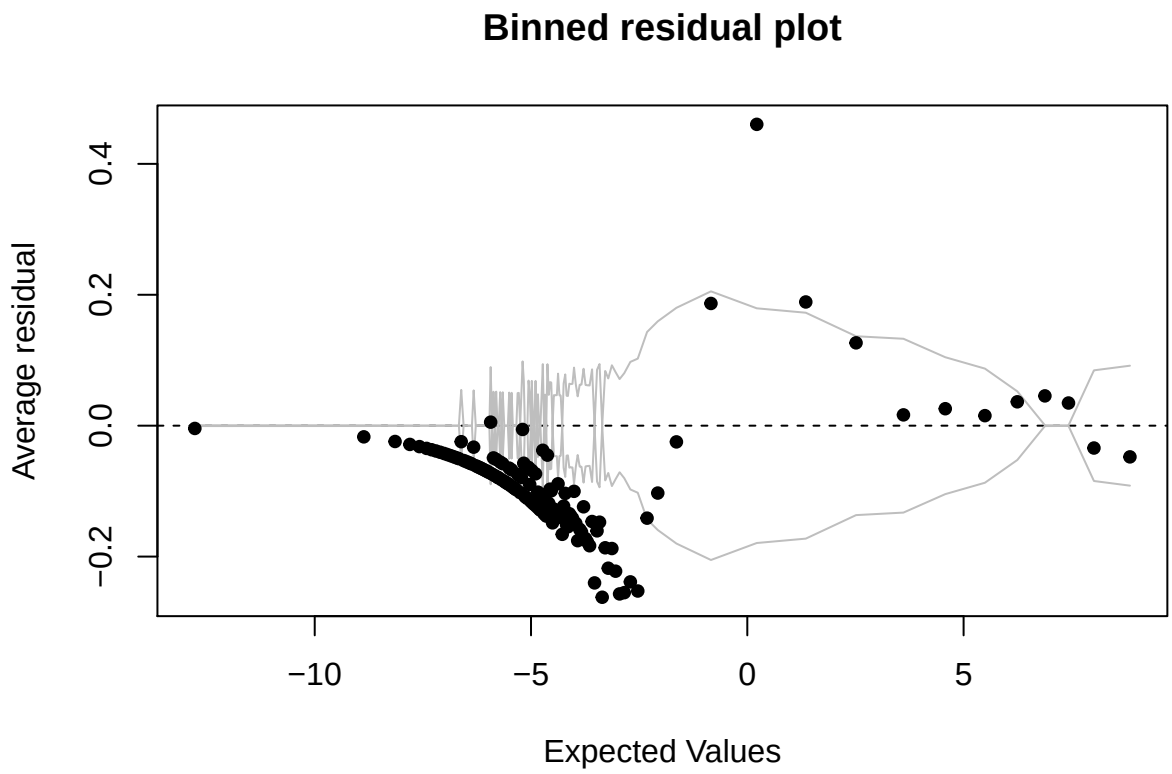
```
##           Predicted Label - Reduced Model #3
## True Label     0     1
##          0 14987   373
##          1   106  1182
```

Diagnostics

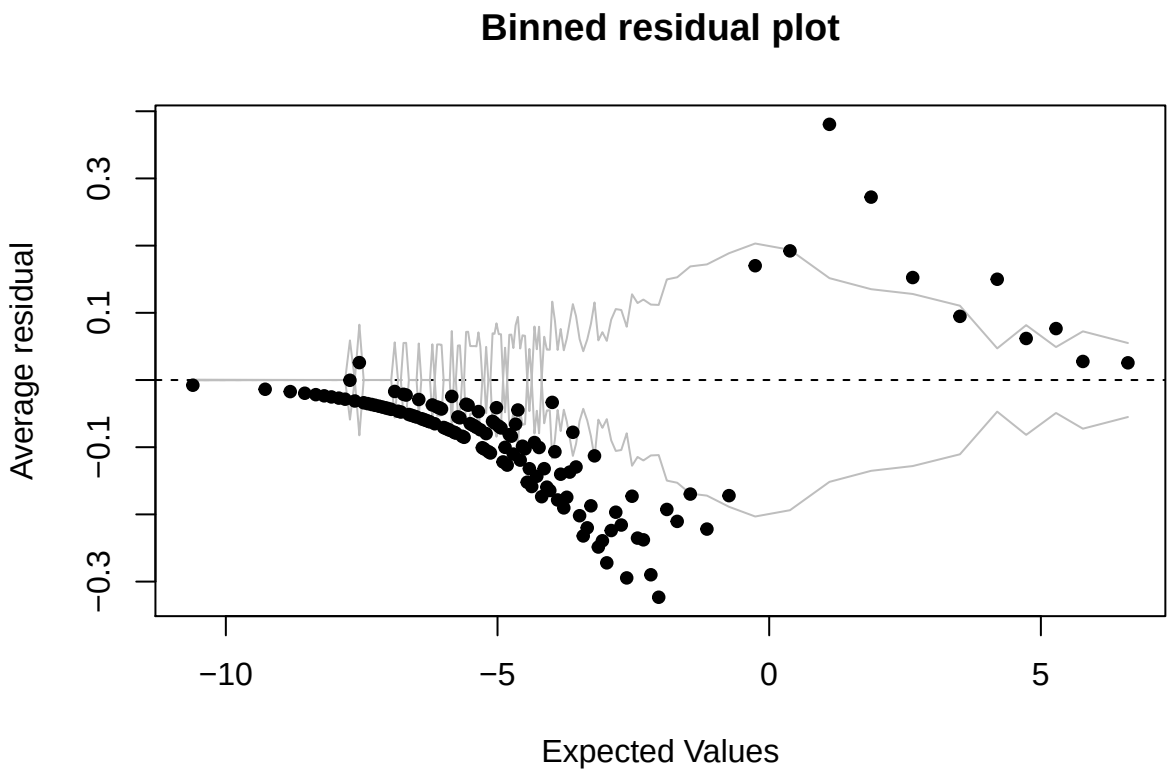
```
suppressMessages(library(arm))
x <- predict(full.model)
y <- resid(full.model)
binnedplot(x,y)
```



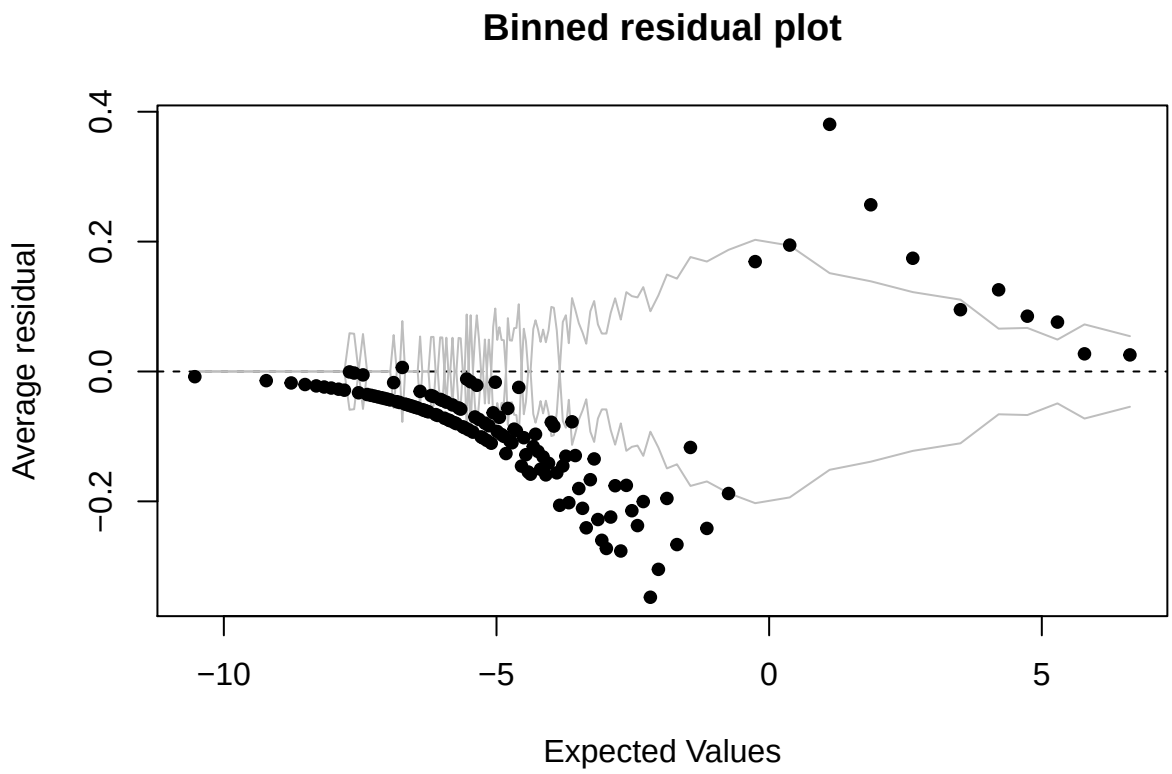

```
x <- predict(reduced.model1)
y <- resid(reduced.model1)
binnedplot(x,y)
```




```
x <- predict(reduced.model2)
y <- resid(reduced.model2)
binnedplot(x,y)
```




```
x <- predict(reduced.model3)
y <- resid(reduced.model3)
binnedplot(x,y)
```




```
#All binned plots bad, interactions don't help
#Binned plots for transformations that seemed to help
```

K-fold CV

```
suppressMessages(library(boot))
pulsar_data <- subset(pulsar_data, select=-target_class_label)
cv.err.full <- cv.glm(pulsar_data,full.model,K=10)
cv.err.full$delta
```

```
## [1] 0.01724807 0.01724266
```

```
cv.err.r1 <- cv.glm(pulsar_data,reduced.model1,K=10)
cv.err.r1$delta
```

```
## [1] 0.01737770 0.01737227
```

```
cv.err.r2 <- cv.glm(pulsar_data,reduced.model2,K=10)
cv.err.r2$delta
```

```
## [1] 0.02337348 0.02336987
```

```
cv.err.r3 <- cv.glm(pulsar_data,reduced.model3,K=10)
cv.err.r3$delta
```

```
## [1] 0.02334646 0.02334459
```