Lab 2: Logistic Regression

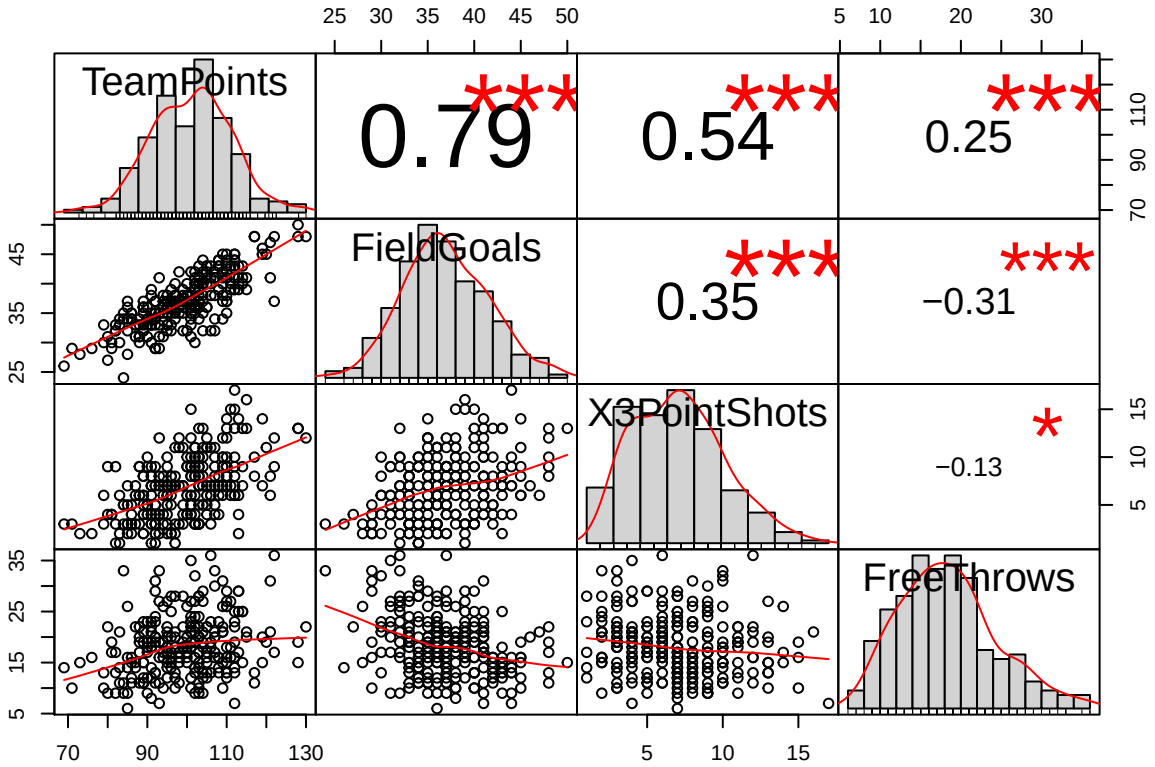
Ashwini Marathe (asm105) & Srishti Saha (ss1078)

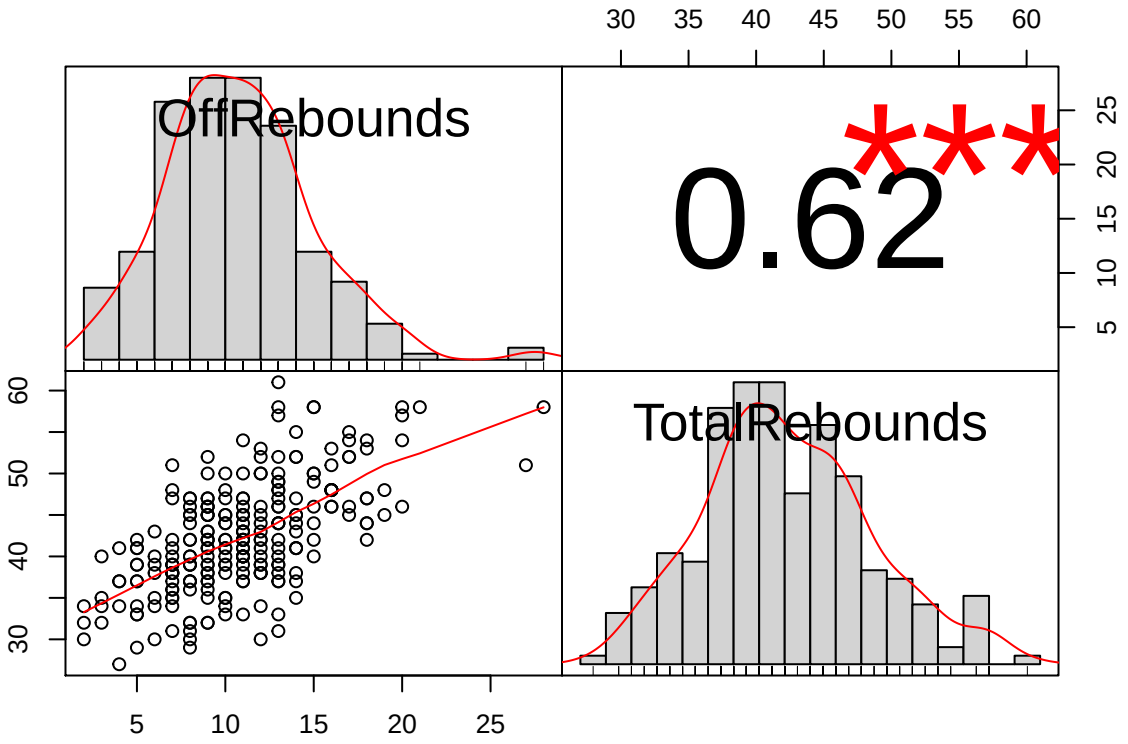
20 September, 2019

Exercise 1: Exploratory plots for Win

- From the histogram plotting **wins versus Home**, we see that the frequency of wins in Home state is greater than that in the matches played Away. We also see that the number of losses at Home are lesser than in matches played Away.
- From the boxplot of **Team points versus Wins and losses**, we see that the team score more points when it won as compared to when it lost.
- The plot shows that the team has **higher number of average field goals in Wins than in losses**. The average percent of goals scored over attempted is over 45% for wins while it is lower for losses.
- Although there is not much of a difference in the **average number of steals across wins and losses**, it is still higher in wins than in losses. The distribution of **number of blocks** follows a very similar pattern. The number of **assists** is also higher for wins than for losses.
- The average number of turnovers is slightly lower in wins than in losses. However, this difference is not very significant.

Exercise 2: Correlations





- We see high correlations between the following pairs:

1. TeamPoints and FieldGoals
2. TeamPoints and X3PointShots
3. TeamPoints and FreeThrows Thus these variables should not be included in the model together.

- We also see a high correlation between OffRebounds and TotalRebounds. Thus, both these variables should also not be included together.
- By definition of the metrics, FieldGoals., FieldGoals and FieldGoalsAttempted also have high correlations. Hence, one of these metrics will be finally selected. A similar case has been observed for X3PointShots and FreeShots and their derived metrics.

Exercise 3: Model

```
##         Home   TeamPoints FieldGoals.      Assists       Steals       Blocks
##     1.170565     1.527458     1.579395     1.319471     1.105540     1.075031
##    Turnovers
##     1.105975
```

According to the VIF values, none of the variable-pairs seem to have a high correlation. Thus, we can safely eliminate the scope of multicollinearity.

Exercise 4: Model Output and Interpretation

```
##
## Call:
## glm(formula = Win ~ Home + TeamPoints + FieldGoals. + Assists +
##     Steals + Blocks + Turnovers, family = binomial, data = nba_reduced_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0389  -0.8190   0.3425   0.7706   2.9590
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.321990   2.334279  -6.564 5.24e-11 ***
## HomeHome      1.021735   0.338800   3.016 0.002563 **
## TeamPoints    0.076893   0.023498   3.272 0.001067 **
## FieldGoals.  15.081376   4.363003   3.457 0.000547 ***
## Assists      -0.003461   0.043652  -0.079 0.936797
## Steals        0.128700   0.055748   2.309 0.020966 *
## Blocks        0.117326   0.074179   1.582 0.113729
## Turnovers    -0.052846   0.043191  -1.224 0.221130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 336.31  on 245  degrees of freedom
## Residual deviance: 245.94  on 238  degrees of freedom
## AIC: 261.94
##
## Number of Fisher Scoring iterations: 5
```

The model results can be interpreted in the following manner:

- The model estimate for Home (level=Home) is 1.02 on the log-scale which converts to 2.77 on the exponential scale. This means that the odds for winning increases 2.77 times when the team plays at Home (as compared to Away matches).
- According to the p-values (ones with low p-values), the significant variables are: Home, TeamPoints, FieldGoals. and Steals.
- The most significant variable is FieldGoals. (field goals scored as a percent of goals attempted) according to the absolute value of its t-value (it is the highest at 3.457)
- The estimate of FieldGoals. on a log-scale is 15.08 which indicates that for an increase in the Field goals ratio by 0.1 (0<=FieldGoals.<=1), there will be a $e^{1.5}$ times increase in the odds of winning.
- For the metric, TeamPoints, the estimate is 0.07 (on the log-scale) which implies that for every 1 point increase in team points, there will be an increase in the odds of winning by $e^{0.07}$ times (i.e. 1.07 times).
- Similarly, for steals, increase in every 1 steal will lead to an increase of odds of winnng by 1.13 times.

Exercise 5: Predictions and Accuracy

Let us look at the confusion matrix for in-sample prediction for the primary basic model with the following variables: Home,TeamPoints,FieldGoals.,Assists,Steals,Blocks,Turnovers

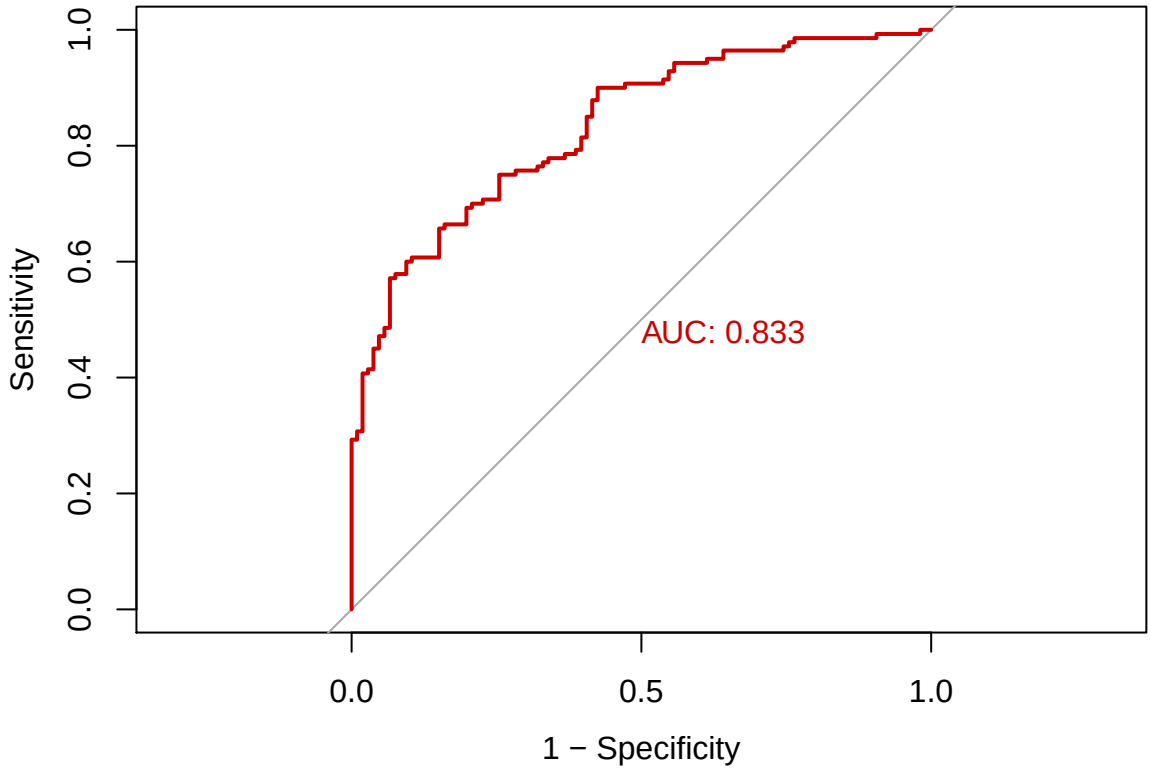
```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   L   W
##          L  67  31
##          W  39 109
##
##                Accuracy : 0.7154
##                  95% CI : (0.6547, 0.771)
##     No Information Rate : 0.5691
##     P-Value [Acc > NIR] : 1.531e-06
##
##                   Kappa : 0.4144
##
##  Mcnemar's Test P-Value : 0.4028
##
##             Sensitivity : 0.7786
##             Specificity : 0.6321
##          Pos Pred Value : 0.7365
##          Neg Pred Value : 0.6837
##              Prevalence : 0.5691
##          Detection Rate : 0.4431
##    Detection Prevalence : 0.6016
##       Balanced Accuracy : 0.7053
##
##        'Positive' Class : W
##
```

The accuracy of this model on the training dataset is 71.5%.

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

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



From the ROC curve, the AUC value is 0.833.

Exercise 6: Improved Model

```
##
## Call:
## glm(formula = Win ~ Home + TeamPoints + FieldGoals. + Assists +
##     Steals + Blocks + Turnovers + Opp.FieldGoals., family = binomial,
##     data = nba_reduced_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80388  -0.34533   0.08365   0.42748   2.20197
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.24990    3.39040  -1.254 0.210020
## HomeHome          0.75555    0.44558   1.696 0.089949 .
## TeamPoints        0.15621    0.03691   4.232 2.31e-05 ***
## FieldGoals.      21.57964    6.03051   3.578 0.000346 ***
## Assists           0.01722    0.05645   0.305 0.760284
## Steals            0.21410    0.07650   2.799 0.005131 **
## Blocks           -0.11401    0.10167  -1.121 0.262145
## Turnovers        -0.12066    0.05814  -2.075 0.037950 *
## Opp.FieldGoals. -46.59914    7.06709  -6.594 4.29e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 336.31  on 245  degrees of freedom
## Residual deviance: 152.20  on 237  degrees of freedom
## AIC: 170.2
##
## Number of Fisher Scoring iterations: 6
```

On addition of the variable Opp.FieldGoals., it turns out to be relevant based on its low p-value. The absolute value of its t-stat is 6.59 (highest). Thus, it is the most significant variable.

The estimate of FieldGoals. on a log-scale is -46.60 which indicates that for an increase in the Field goals ratio by 0.1 (0<=FieldGoals.<=1), there will be a $e^{4.6}$ times decrease in the odds of winning. (*This statistic may be up for investigation!*)

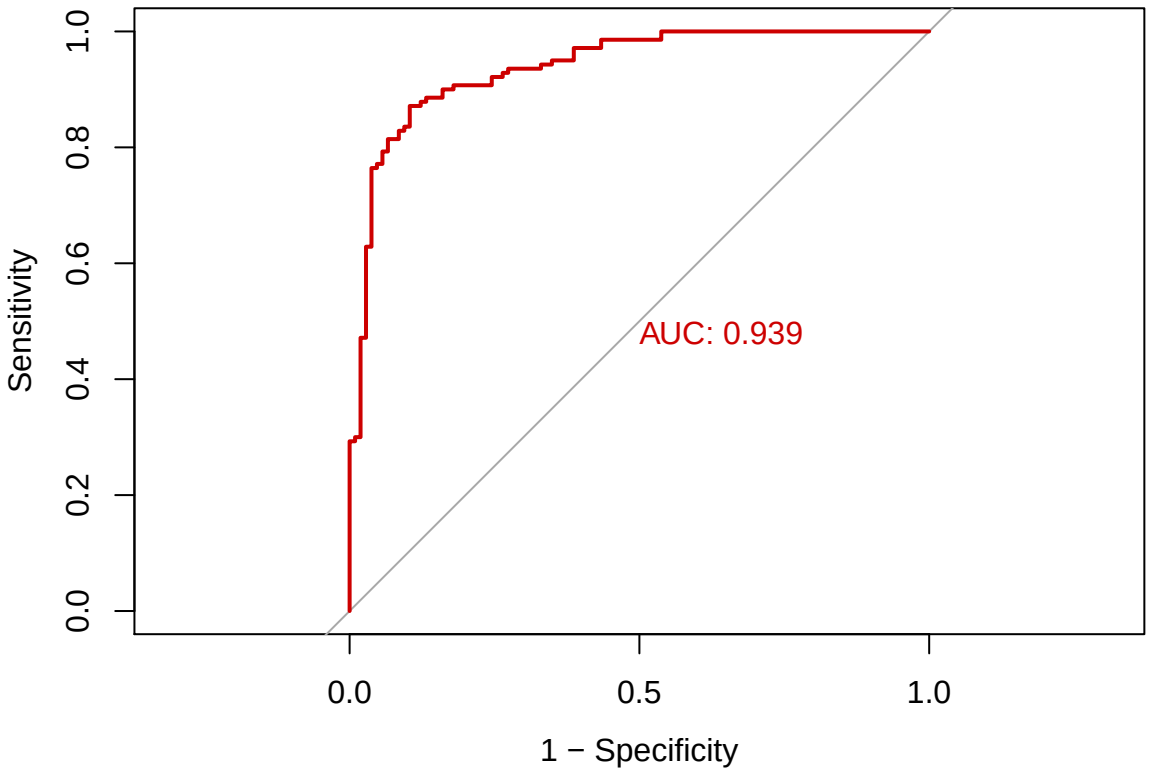
Exercise 7: Confusion Matrix and ROC of improved model

Let us look at the confusion matrix for in-sample prediction for the improved model with the following variables: Home,TeamPoints,FieldGoals.,Assists,Steals,Blocks,Turnovers,Opp.FieldGoals.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   L   W
##          L  87  14
##          W  19 126
##
##                Accuracy : 0.8659
##                  95% CI : (0.8168, 0.9058)
##     No Information Rate : 0.5691
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7249
##
##  Mcnemar's Test P-Value : 0.4862
##
##             Sensitivity : 0.9000
##             Specificity : 0.8208
##          Pos Pred Value : 0.8690
##          Neg Pred Value : 0.8614
##              Prevalence : 0.5691
##          Detection Rate : 0.5122
##    Detection Prevalence : 0.5894
##       Balanced Accuracy : 0.8604
##
##        'Positive' Class : W
##
```

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

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



The accuracy of the improved model is 86.6%. The true-positive rate (sensitivity) of the model is 90% which indicates that of all the wins, the model can predict 90% of the cases. For the previous model, the sensitivity was 77.9%. Thus, the new improved model is predicting the odds of winning better.

The AUC of the improved model is 0.939.

Exercise 8: Suggestions of Coach

We have the following recommendations:

1. Improve the percentage of field goals scored as a percent of the attempts. This would mean that the accuracy of the goals scored should increase.
2. The average team scores (team points) should increase. This variable is highly significant while improving the odds of winning.
3. Improve the defense such that the number of steals increase thus increasing the odds of winning.

Exercise 9: Out-of-sample predictions

Let us look at the confusion matrix for out-of-sample prediction for the improved model with the following variables: Home,TeamPoints,FieldGoals.,Assists,Steals,Blocks,Turnovers,Opp.FieldGoals.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 44  2
##          1 16 20
##
##                Accuracy : 0.7805
##                  95% CI : (0.6754, 0.8644)
##     No Information Rate : 0.7317
##     P-Value [Acc > NIR] : 0.192781
##
##                   Kappa : 0.5347
##
##  Mcnemar's Test P-Value : 0.002183
##
##             Sensitivity : 0.9091
##             Specificity : 0.7333
##          Pos Pred Value : 0.5556
##          Neg Pred Value : 0.9565
##              Prevalence : 0.2683
##          Detection Rate : 0.2439
##    Detection Prevalence : 0.4390
##       Balanced Accuracy : 0.8212
##
##        'Positive' Class : 1
##
```

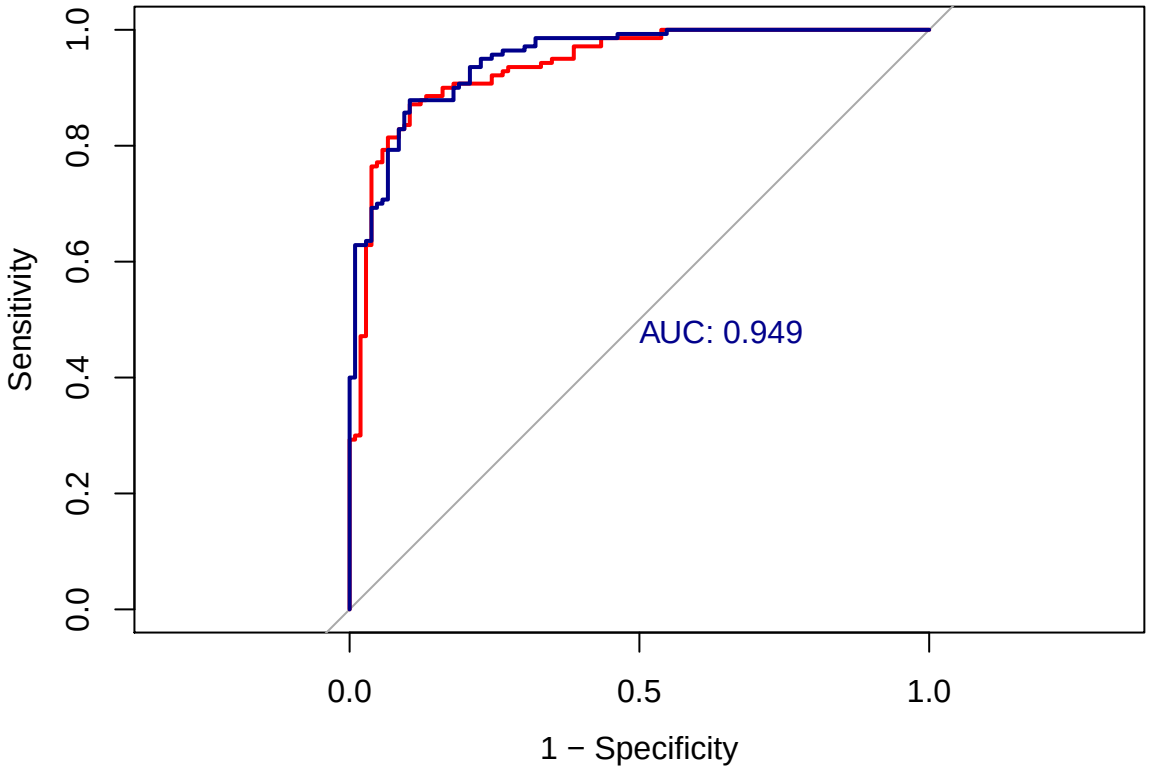
The out-of-sample accuracy of this model is 78.05%. The model seems to perform well even for out-of-sample data (2017-2018).

Exercise 10: Change in Deviance Test

Let us look at the confusion matrix for in-sample prediction for the improved model with the following variables: Home,TeamPoints,FieldGoals.,Assists,Steals,Blocks,Turnovers,Opp.FieldGoals,Opp.Assists,Opp.Blocks

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   L   W
##          L  87  15
##          W  19 125
##
##                Accuracy : 0.8618
##                  95% CI : (0.8123, 0.9023)
##     No Information Rate : 0.5691
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7169
##
##  Mcnemar's Test P-Value : 0.6069
##
##             Sensitivity : 0.8929
```

```
##             Specificity : 0.8208
##          Pos Pred Value : 0.8681
##          Neg Pred Value : 0.8529
##              Prevalence : 0.5691
##          Detection Rate : 0.5081
##    Detection Prevalence : 0.5854
##       Balanced Accuracy : 0.8568
##
##        'Positive' Class : W
##


## Analysis of Deviance Table
##
## Model 1: Win ~ Home + TeamPoints + FieldGoals. + Assists + Steals + Blocks +
##     Turnovers + Opp.FieldGoals.
## Model 2: Win ~ Home + TeamPoints + FieldGoals. + Assists + Steals + Blocks +
##     Turnovers + Opp.FieldGoals. + Opp.Assists + Opp.Blocks
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       237     152.20
## 2       235     140.09  2   12.101 0.002356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

## Setting direction: controls < cases

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

## Setting direction: controls < cases
```



According to the results of the change in deviance test, the p-value is very low. Given that the null-hypothesis is that the new model is equivalent to the previous model, we can reject the the same. Thus, the new model is better than the previous one.

Addition of a new variable:

We thought of adding the variable Opp.Turnovers to our existing model. This is because by definituion of the metric, if the number of times the ball was won back from the opponent (when they had the possession), this should increase the odds of winning.

Let us look at the confusion matrix for in-sample prediction for the improved model with the following variables: Home, TeamPoints, FieldGoals., Assists, Steals, Blocks, Turnovers, Opp.FieldGoals, Opp.Assists, Opp.Blocks, Opp.Turnovers

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   L   W
##          L  90  15
##          W  16 125
##
##                Accuracy : 0.874
##                  95% CI : (0.8259, 0.9127)
##     No Information Rate : 0.5691
##     P-Value [Acc > NIR] : <2e-16
##
```

```
##                   Kappa : 0.7428
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8929
##             Specificity : 0.8491
##          Pos Pred Value : 0.8865
##          Neg Pred Value : 0.8571
##              Prevalence : 0.5691
##          Detection Rate : 0.5081
##    Detection Prevalence : 0.5732
##       Balanced Accuracy : 0.8710
##
##        'Positive' Class : W
##
```

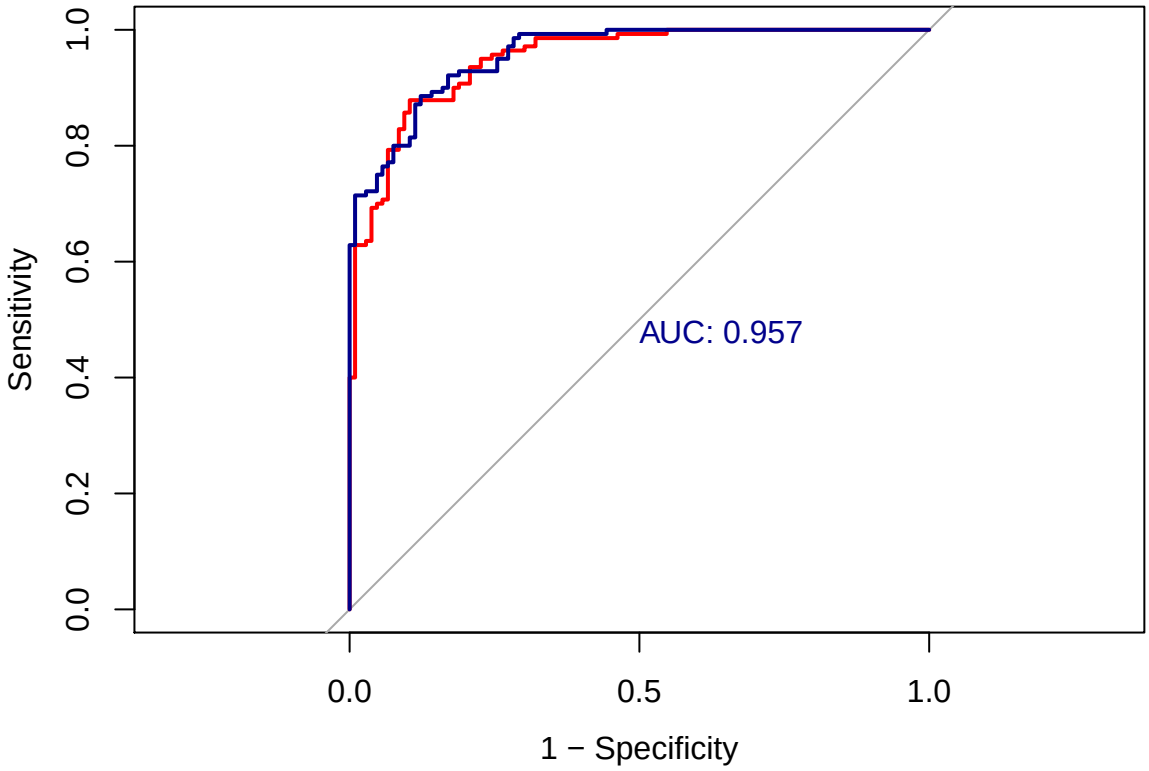
```
## Analysis of Deviance Table
##
## Model 1: Win ~ Home + TeamPoints + FieldGoals. + Assists + Steals + Blocks +
##     Turnovers + Opp.FieldGoals. + Opp.Assists + Opp.Blocks
## Model 2: Win ~ Home + TeamPoints + FieldGoals. + Assists + Steals + Blocks +
##     Turnovers + Opp.FieldGoals. + Opp.Assists + Opp.Blocks +
##     Opp.Turnovers
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       235     140.09
## 2       234     130.82  1   9.2711 0.002328 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

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

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



With the improved model with the variable Opp.Turnovers, we see that this variable is significant owing to its low p-value and high t-statistic.

On comparing the in-sample accuracy of this model with the previous one, we see that the accuracy increased from 86.2% to 87.4% which is an improvement.

Moreover, on comparing the 2 models with a change in deviance test, we see that the p-value is low which implies that the addition of the new variable has improved the model.