NBA(1)

Sahba Salarian

2019-02-22

Introduction

```
NBA1 <-read.csv("/Users/sahba/Dropbox/Data Science/NBA longevity/nba_logreg.csv", header=T, stringsAsFa
```

```
str(NBA1)
```

```
## 'data.frame':    1340 obs. of  21 variables:
##  $ Name       : chr  "Brandon Ingram" "Andrew Harrison" "JaKarr Sampson" "Malik Sealy" ...
##  $ GP         : int  36 35 74 58 48 75 62 48 65 42 ...
##  $ MIN        : num  27.4 26.9 15.3 11.6 11.5 11.4 10.9 10.3 9.9 8.5 ...
##  $ PTS        : num  7.4 7.2 5.2 5.7 4.5 3.7 6.6 5.7 2.4 3.7 ...
##  $ FGM        : num  2.6 2 2 2.3 1.6 1.5 2.5 2.3 1 1.4 ...
##  $ FGA        : num  7.6 6.7 4.7 5.5 3 3.5 5.8 5.4 2.4 3.5 ...
##  $ FG.        : num  34.7 29.6 42.2 42.6 52.4 42.3 43.5 41.5 39.2 38.3 ...
##  $ X3P.Made   : num  0.5 0.7 0.4 0.1 0 0.3 0 0.4 0.1 0.1 ...
##  $ X3PA       : num  2.1 2.8 1.7 0.5 0.1 1.1 0.1 1.5 0.5 0.3 ...
##  $ X3P.       : num  25 23.5 24.4 22.6 0 32.5 50 30 23.3 21.4 ...
##  $ FTM        : num  1.6 2.6 0.9 0.9 1.3 0.4 1.5 0.7 0.4 1 ...
##  $ FTA        : num  2.3 3.4 1.3 1.3 1.9 0.5 1.8 0.8 0.5 1.4 ...
##  $ FT.        : num  69.9 76.5 67 68.9 67.4 73.2 81.1 87.5 71.4 67.8 ...
##  $ OREB       : num  0.7 0.5 0.5 1 1 0.2 0.5 0.8 0.2 0.4 ...
##  $ DREB       : num  3.4 2 1.7 0.9 1.5 0.7 1.4 0.9 0.6 0.7 ...
##  $ REB        : num  4.1 2.4 2.2 1.9 2.5 0.8 2 1.7 0.8 1.1 ...
##  $ AST        : num  1.9 3.7 1 0.8 0.3 1.8 0.6 0.2 2.3 0.3 ...
##  $ STL        : num  0.4 1.1 0.5 0.6 0.3 0.4 0.2 0.2 0.3 0.2 ...
##  $ BLK        : num  0.4 0.5 0.3 0.1 0.4 0 0.1 0.1 0 0 ...
##  $ TOV        : num  1.3 1.6 1 1 0.8 0.7 0.7 0.7 1.1 0.7 ...
##  $ TARGET_5Yrs: num  0 0 0 1 1 0 1 1 0 0 ...
```

```
#NBA1$Name <- c(as.factor(NBA1$Name))
#str(NBA1)
```

Data engineering

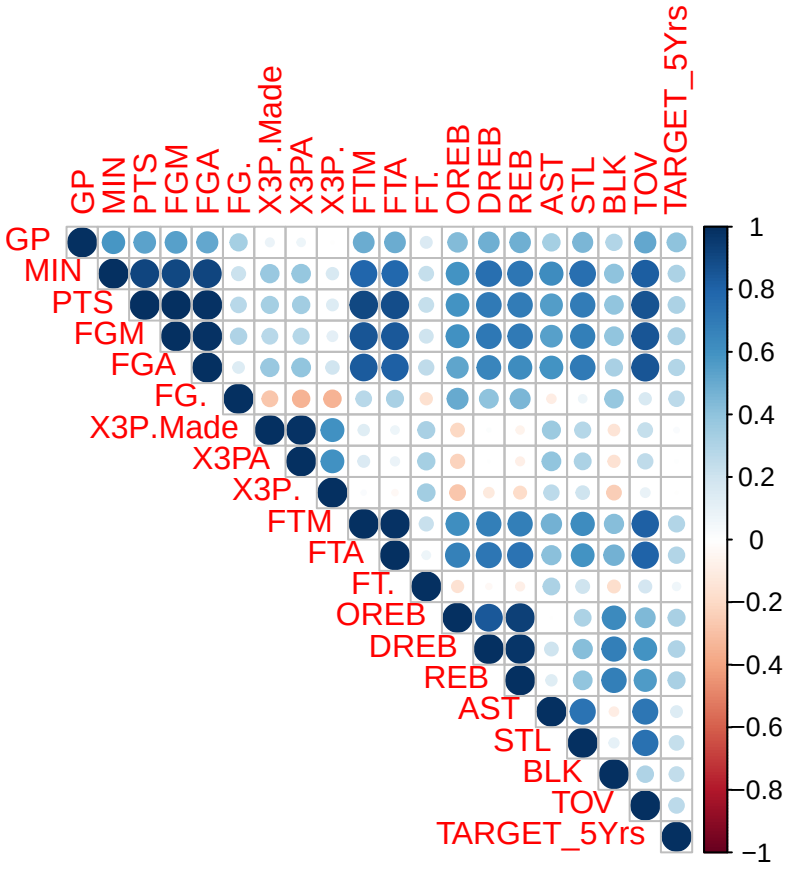
Train/Test Split:

```
str(train)
```

```
## 'data.frame':    793 obs. of  21 variables:
##  $ Name       : chr  "Brandon Ingram" "JaKarr Sampson" "Malik Sealy" "Matt Geiger" ...
##  $ GP         : int  36 74 58 48 48 42 35 40 27 40 ...
##  $ MIN        : num  27.4 15.3 11.6 11.5 10.3 8.5 6.9 6.7 6.6 6.1 ...
##  $ PTS        : num  7.4 5.2 5.7 4.5 5.7 3.7 2.3 3.6 1.3 2.6 ...
##  $ FGM        : num  2.6 2 2.3 1.6 2.3 1.4 0.9 1.2 0.6 0.9 ...
##  $ FGA        : num  7.6 4.7 5.5 3 5.4 3.5 2.4 3 1.3 1.8 ...
```

```
##  $ FG.         : num  34.7 42.2 42.6 52.4 41.5 38.3 36.5 39.8 47.2 51.4 ...
##  $ X3P.Made    : num  0.5 0.4 0.1 0 0.4 0.1 0 0.1 0 0.1 ...
##  $ X3PA        : num  2.1 1.7 0.5 0.1 1.5 0.3 0.1 0.6 0 0.4 ...
##  $ X3P.        : num  25 24.4 22.6 0 30 21.4 33.3 13.6 0 14.3 ...
##  $ FTM         : num  1.6 0.9 0.9 1.3 0.7 1 0.5 1.1 0.1 0.7 ...
##  $ FTA         : num  2.3 1.3 1.3 1.9 0.8 1.4 0.6 1.5 0.3 1 ...
##  $ FT.         : num  69.9 67 68.9 67.4 87.5 67.8 81.8 77.6 28.6 68.4 ...
##  $ OREB        : num  0.7 0.5 1 1 0.8 0.4 0.5 0.5 0.6 0.1 ...
##  $ DREB        : num  3.4 1.7 0.9 1.5 0.9 0.7 0.3 0.8 1.4 0.3 ...
##  $ REB         : num  4.1 2.2 1.9 2.5 1.7 1.1 0.9 1.2 2 0.4 ...
##  $ AST         : num  1.9 1 0.8 0.3 0.2 0.3 0.7 0.4 0.2 1.4 ...
##  $ STL         : num  0.4 0.5 0.6 0.3 0.2 0.2 0.1 0.3 0.2 0.3 ...
##  $ BLK         : num  0.4 0.3 0.1 0.4 0.1 0 0.1 0.1 0.6 0 ...
##  $ TOV         : num  1.3 1 1 0.8 0.7 0.7 0.3 0.6 0.3 0.8 ...
##  $ TARGET_5Yrs: num  0 0 1 1 1 0 0 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int  339 340 341 359 387 398 508 510 511 522 ...
##   ..- attr(*, "names")= chr  "339" "340" "341" "359" ...
```

The TARGET_5Yrs should be analyzed versus other information for each athlete: GP, MIN, PTS, FGM, FGA, FG.,X3P.Made, X3PA, X3P.,FTM, FTA,FT.,OREB, DREB, REB, AST, STL, BLK, TOV



Use glm() to fit the model

```
LogFit1 <- glm(TARGET_5Yrs ~  GP+ MIN+ PTS+ FGM+ FGA+ FG.+X3P.Made+ X3PA+ X3P.+FTM+ FTA+ FT.+OREB +DREB
stargazer (LogFit1, type="latex", title="ANOVA test of the logistic Regresssion ", header = FALSE)
```

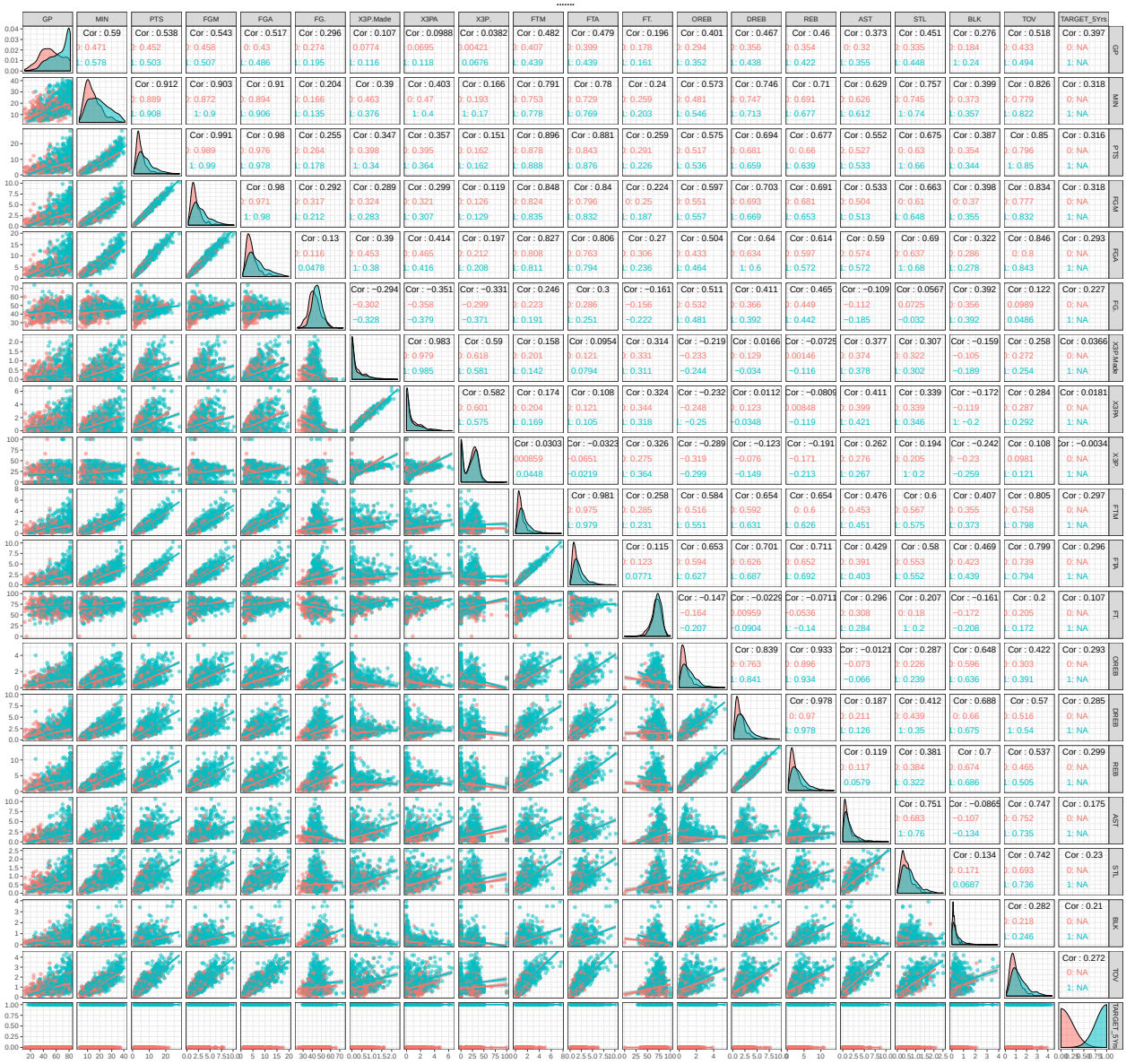

GP
MIN
PTS
FGM
FGA
FG.
X3P.Made
X3PA
X3P.
FTM
FTA
FT.
OREB
DREB
REB
AST
STL
BLK
TOV
TARGET_5Yrs

Figure 1: „„„

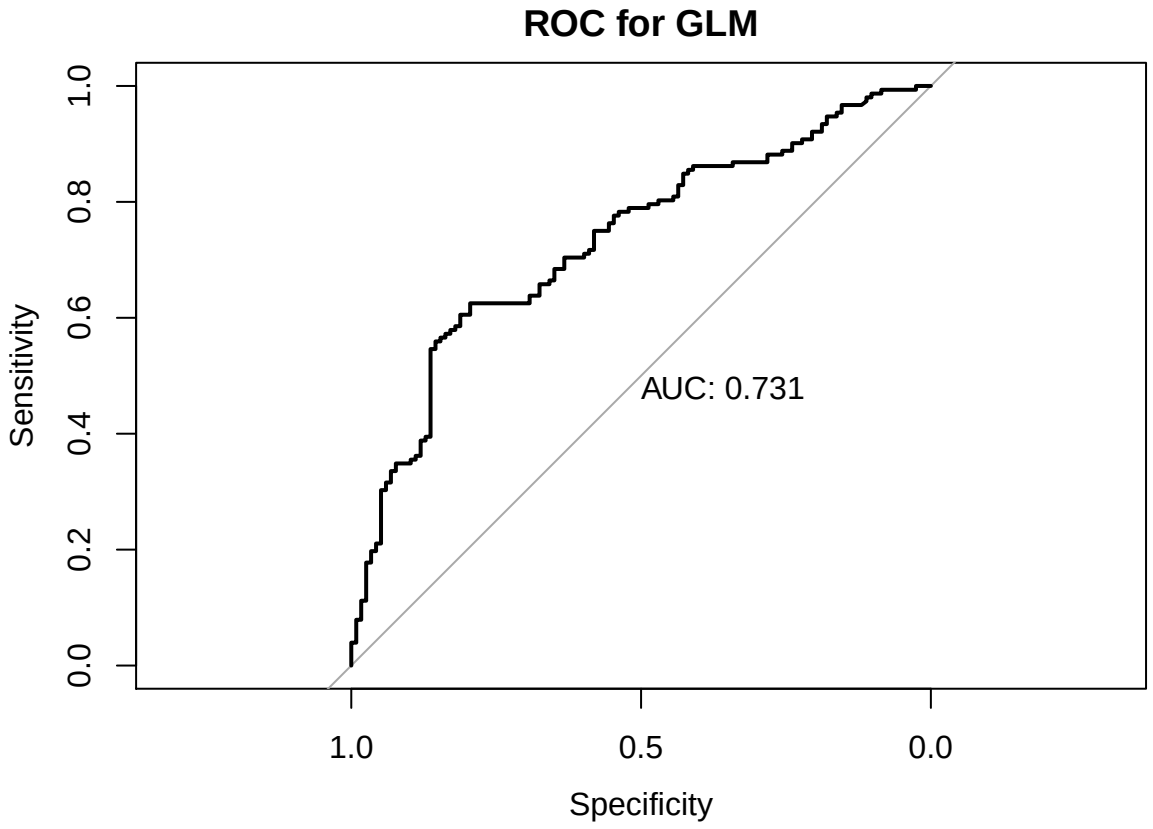
```
##
## \begin{table}[!htbp] \centering
##   \caption{ANOVA test of the logistic Regresssion }
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & TARGET\_5Yrs \\
## \hline \\[-1.8ex]
##  GP & 0.037$^{***}$ \\
##   & (0.006) \\
##   & \\
##  MIN & $-$0.093$^{**}$ \\
##   & (0.043) \\
##   & \\
##  PTS & 0.051 \\
##   & (1.158) \\
##   & \\
##  FGM & $-$0.940 \\
##   & (2.331) \\
##   & \\
##  FGA & 0.521$^{*}$ \\
##   & (0.301) \\
##   & \\
##  FG. & 0.059$^{**}$ \\
##   & (0.029) \\
##   & \\
##  X3P.Made & 2.865$^{*}$ \\
##   & (1.709) \\
##   & \\
##  X3PA & $-$1.054$^{**}$ \\
##   & (0.526) \\
##   & \\
##  X3P. & 0.011 \\
##   & (0.007) \\
##   & \\
##  FTM & $-$0.395 \\
##   & (1.319) \\
##   & \\
##  FTA & 0.260 \\
##   & (0.607) \\
##   & \\
##  FT. & 0.019 \\
##   & (0.013) \\
##   & \\
##  OREB & $-$0.352 \\
##   & (1.747) \\
##   & \\
##  DREB & $-$1.543 \\
##   & (1.744) \\
##   & \\
##  REB & 1.431 \\
```

```
##   & (1.740) \\
##   & \\
##  AST & 0.307$^{**}$ \\
##   & (0.143) \\
##   & \\
##  STL & 0.278 \\
##   & (0.432) \\
##   & \\
##  BLK & 0.862$^{**}$ \\
##   & (0.366) \\
##   & \\
##  TOV & $-$0.323 \\
##   & (0.352) \\
##   & \\
##  Constant & $-$6.336$^{***}$ \\
##   & (1.606) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 793 \\
## Log Likelihood & $-$422.615 \\
## Akaike Inf. Crit. & 885.231 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
summary(LogFit1)
```

```
##
## Call:
## glm(formula = TARGET_5Yrs ~ GP + MIN + PTS + FGM + FGA + FG. +
##     X3P.Made + X3PA + X3P. + FTM + FTA + FT. + OREB + DREB +
##     REB + AST + STL + BLK + TOV, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9575  -0.9445   0.4726   0.8353   2.3490
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.335548   1.605904  -3.945 7.97e-05 ***
## GP           0.037087   0.006235   5.948 2.72e-09 ***
## MIN         -0.093299   0.042810  -2.179   0.0293 *
## PTS          0.050503   1.158169   0.044   0.9652
## FGM         -0.940060   2.330690  -0.403   0.6867
## FGA          0.521196   0.300832   1.733   0.0832 .
## FG.          0.058884   0.028795   2.045   0.0409 *
## X3P.Made     2.865413   1.709017   1.677   0.0936 .
## X3PA        -1.054156   0.526354  -2.003   0.0452 *
## X3P.         0.010725   0.007493   1.431   0.1523
## FTM         -0.394939   1.319184  -0.299   0.7646
## FTA          0.259550   0.607014   0.428   0.6690
## FT.          0.019345   0.013087   1.478   0.1394
```

```
## OREB        -0.351588   1.746831  -0.201   0.8405
## DREB        -1.543372   1.743850  -0.885   0.3761
## REB          1.430777   1.740041   0.822   0.4109
## AST          0.306758   0.142630   2.151   0.0315 *
## STL          0.278445   0.432034   0.644   0.5193
## BLK          0.862308   0.365947   2.356   0.0185 *
## TOV         -0.322579   0.351823  -0.917   0.3592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1044.67  on 792  degrees of freedom
## Residual deviance:  845.23  on 773  degrees of freedom
## AIC: 885.23
##
## Number of Fisher Scoring iterations: 5
```



The AIC-Backard method features: - X3P.
- FGA
- X3PA
- FT.
- FG.
- X3P.Made
- MIN
- BLK
- AST
- DREB

- REB
- GP

The BIC-Backward method decreased the features to 3, GP , DREB , REB but with AIC=901.29.

The BestGLM method decreased the features but with AIC= 924 and 952, which are larger than the first backward_AIC model applied. So the backward_AIC is the one applied.

new GLm based on the model developed from backward_AIC:

```
LogFit2 <- glm(TARGET_5Yrs ~ GP + MIN + FGA + FG. + X3P.Made + X3PA + X3P. +
    FT. + DREB + REB + AST + BLK , data=train[,2:21], family=binomial(link='logit'))
stargazer (LogFit2, type="latex", title="ANOVA test of the logistic Regresssion ", header = FALSE)
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{ANOVA test of the logistic Regresssion }
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & TARGET\_5Yrs \\
## \hline \\[-1.8ex]
##  GP & 0.036$^{***}$ \\
##   & (0.006) \\
##   & \\
##  MIN & $-$0.080$^{**}$ \\
##   & (0.039) \\
##   & \\
##  FGA & 0.101 \\
##   & (0.066) \\
##   & \\
##  FG. & 0.030$^{*}$ \\
##   & (0.018) \\
##   & \\
##  X3P.Made & 2.351$^{*}$ \\
##   & (1.320) \\
##   & \\
##  X3PA & $-$0.773 \\
##   & (0.488) \\
##   & \\
##  X3P. & 0.011 \\
##   & (0.007) \\
##   & \\
##  FT. & 0.015$^{*}$ \\
##   & (0.009) \\
##   & \\
##  DREB & $-$1.254$^{***}$ \\
##   & (0.399) \\
##   & \\
##  REB & 1.095$^{***}$ \\
```

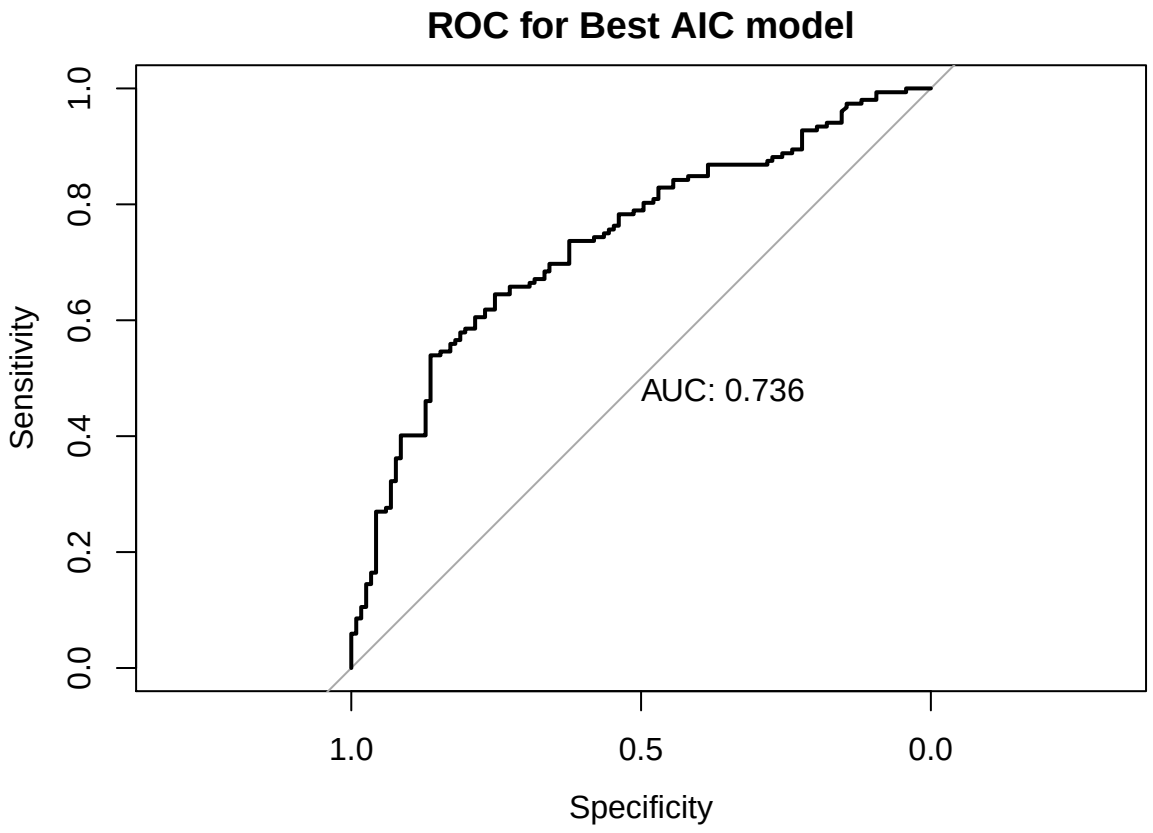
```
##   & (0.292) \\
##   & \\
##  AST & 0.258$^{**}$ \\
##   & (0.110) \\
##   & \\
##  BLK & 0.824$^{**}$ \\
##   & (0.366) \\
##   & \\
##  Constant & $-$4.817$^{***}$ \\
##   & (0.999) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 793 \\
## Log Likelihood & $-$424.265 \\
## Akaike Inf. Crit. & 874.531 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
summary(LogFit2)
```

```
##
## Call:
## glm(formula = TARGET_5Yrs ~ GP + MIN + FGA + FG. + X3P.Made +
##     X3PA + X3P. + FT. + DREB + REB + AST + BLK, family = binomial(link = "logit"),
##     data = train[, 2:21])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9872  -0.9573   0.4732   0.8619   2.2479
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.817183   0.999090  -4.822 1.42e-06 ***
## GP           0.036463   0.006160   5.919 3.24e-09 ***
## MIN         -0.079850   0.039396  -2.027 0.042678 *
## FGA          0.101427   0.066325   1.529 0.126203
## FG.          0.030484   0.018069   1.687 0.091594 .
## X3P.Made     2.350754   1.319644   1.781 0.074854 .
## X3PA        -0.773226   0.488313  -1.583 0.113316
## X3P.         0.010549   0.007357   1.434 0.151614
## FT.          0.015374   0.009237   1.664 0.096050 .
## DREB        -1.254388   0.399174  -3.142 0.001675 **
## REB          1.094629   0.292179   3.746 0.000179 ***
## AST          0.258388   0.110035   2.348 0.018863 *
## BLK          0.824465   0.365807   2.254 0.024207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1044.67  on 792  degrees of freedom
## Residual deviance:  848.53  on 780  degrees of freedom
```

```
## AIC: 874.53
##
## Number of Fisher Scoring iterations: 5
```

ROC for best backward_AIC model:



SVMfit

```
#install.package("e1071")
library(e1071)
svmfit <- svm (TARGET_5Yrs ~ GP+ MIN+ PTS+ FGM+ FGA+ FG.+X3P.Made+ X3PA+ X3P.+FTM+ FTA+ FT.+OREB +DREB +
print (svmfit)
```

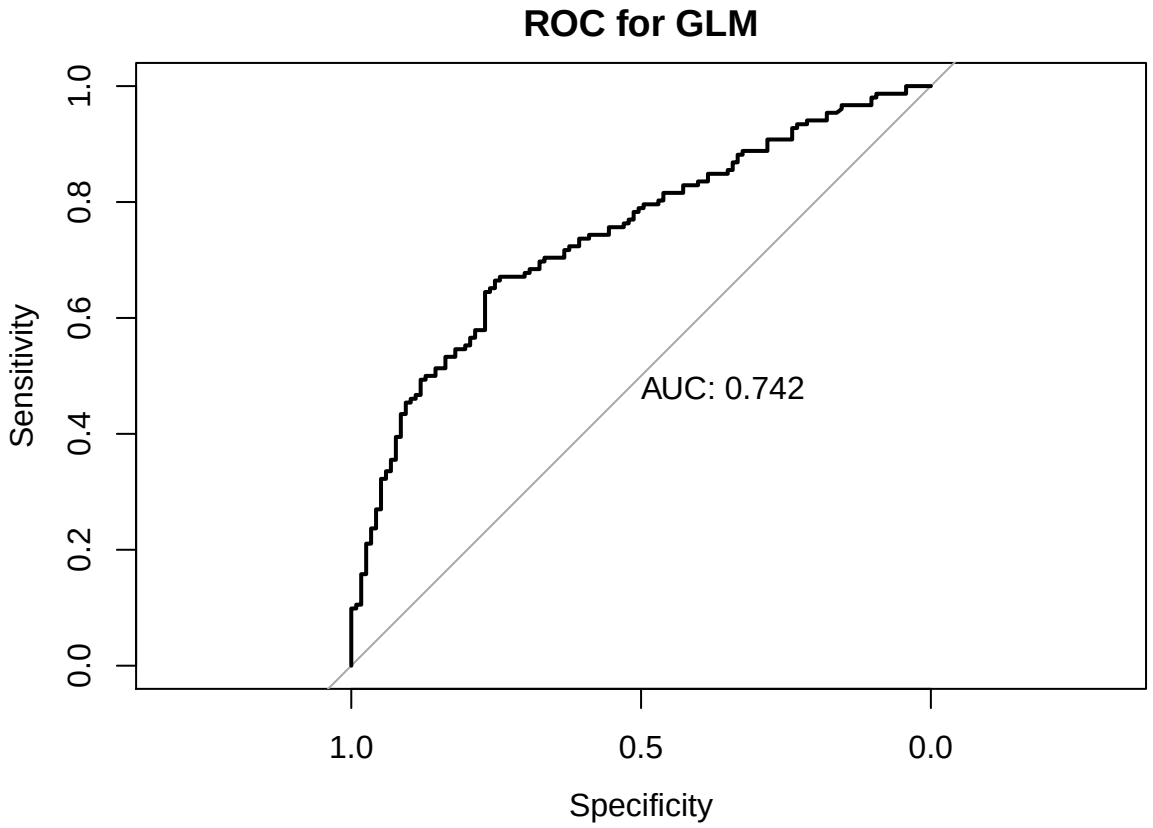
```
##
## Call:
## svm(formula = TARGET_5Yrs ~ GP + MIN + PTS + FGM + FGA + FG. +
##     X3P.Made + X3PA + X3P. + FTM + FTA + FT. + OREB + DREB +
##     REB + AST + STL + BLK + TOV, data = train, kernel = "linear",
##     gamma = 0.05, cost = 10, probability = TRUE, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  linear
```

```
##        cost:  10
##       gamma:  0.05
##     epsilon:  0.1
##
## Sigma:  0.711947
##
##
## Number of Support Vectors:  663
```

```
prob <- predict(svmfit, cv, probability = TRUE)
plot(svmfit, train, TARGET_5Yrs ~ GP)
```

```
svmfit$type
```

```
## [1] 3
```



tunning SVM

cross validation ba caret:

Confusion Matrix

Decision Boundary Logistic & others (probably cannot be drawn dur to large number of features)