Final Report

Brian Hallberg

6/23/2018

Introduction

Election to the Baseball Hall of Fame is the highest honor bestowed on any player, it indicates that they have made great contributions to baseball and are legedary players that have made baseball a fan favorite for over 100 years.

With over 100 years of players and their playing statistics, the ability to predict if a player will become a member of the Hall of Fame would be helpful to active players and their agents as well as members of the baseball writers that elect the players and the veterans committe that elects those no longer eligible. Players and agents can use this in salary negotions. Baseball writers can insure that they continue to nominate and elect the elite players in the game. The veterns committee can identify players that were missed in the past but have the statistics to have been elected.

Data

This project makes use of a data set called the History of Baseball. It contains data from professional baseball covering the period from 1871 to 2015. This includes All-Star Game and Hall of Fame selections. Additionally, there is yearly statistics on over 18,000 players. These sets of data will provide the information needed and are described below. There are additional data on teams and franchises that are avaialble, but not used in this project. The data is available on the Kraggle web side and can be obtained at this link:

- https://www.kaggle.com/seanlahman/the-history-of-baseball

Player Table

From the Kraggle site this is called the Master Table and is stored in the file master.csv. It made more sense to me to call this is the player table and so I saved the file as player.csv in the repository and always use the table name of player instead.

The player table contains unique code assigned to each player (player_id) which is used as key into each of the other tables. This has many details about the player, but for this model only the player_id, player's first name (nameFirst), player's last name (nameLast), date the player made their first major league appearance (debut), and date the player made their last major league appearence (finalGame).

All Star Table

The all_star table contains fields related to those players that were selected for the all-star game or games, since some years there were multiple games played. The fields used were player_id, year (YearID), for those selected for the game.

Batting Table

The batting table contains fields related to the hitting statistics for the players by year(yearID) and team(stint). A number of fields were initially reviewed, but the following were the ones that appeared to be the most significiant for the model; player_id, at bats(AB), runs(R), hits(H), home runs(HR), and runs batted in(RBI).

Hall of Fame Table

The hall_of_fame table contains fields related to the Hall of Fame elections for each year(YearID) starting in 1936. The fields player_id, and inducted which is Y if they were elected that year and a N if not.

Pitching Table

The pitching table contains fields related to the pitching statistics for the player by year(YearID) and team(stint). A number of fields were initially reviewed, but the following were the ones that appeared to be the most significiant for the model; playerID, games pitched(G), wins(W), and strikeouts(SO).

Data Manipulation

To keep the tables smaller, some columns were removed from the player table.

```r
player <- player %>%
  select(-starts_with("birth"), -starts_with("death"), -height, -weight) %>%
  select(-bats, -throws, -bbref_id, -retro_id, -name_given)
```

Summarize data

The *batting* and *pitching* tables are broken down beyond the player so that a row of the table represents the team and year because a player could have played for more that one team in a year. I only care about the totals for the players career so summaries of these two tables were created with these statements.

```r
#  Summarize the batting data first
sumhit <- batting %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  select(-year, -stint)
#  Now get the pitching data
sumpitch <- pitching %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  select(-year, -stint, -baopp, -era)
```

The *all_star* table has a row that represents a player, year and which all-star game they were voted to play (yes, in 4 years they played 2 all-star games in the year). I only wanted the number of all-star games each played was voted to play in so another summary was done with this statement.

```r
sumallstar <- all_star %>%
  select(player_id, league_id, year) %>%
  group_by(player_id) %>%
  count(player_id)
```

Finally the *hall_of_fame* table shows each year a player was voted on by the baseball writers including the number of votes they received. I only care if the player was inducted so the following statements were used to reduce all the data to a unique set of players inducted.

```r
sumhof <- hof %>%
  filter(category =="Player") %>%
  select(player_id, inducted) %>%
  filter(inducted == "Y")
```

Build a final data set of all information by player

Since data was spread across different tables, the corresponding data frames needed to be joined in order to pull the data together into a single data frame. The statements below shows the joining of the four tables into one data frame used.

```r
#  Add player and summarized Hall of Fame data
fullplay <- full_join(player, sumhof, by = "player_id")
#  Add the summarized All-Star data
fullplay <- full_join(fullplay, sumallstar, by = "player_id")
#  Add the summarized batting data
fullplay <- full_join(fullplay, sumhit, by = "player_id")
#  And finally add the summarized pitching data
fullplay <- full_join(fullplay, sumpitch, by = "player_id")
```

Cleanup some bad column names

The joins of the batting and pitching tables, created some conflicting column names and so they were updated to make some plots more readable and also to help simplify some future code. These are included below

```r
fullplay <- rename(fullplay, allstar = n, rallow = r.y, hrallow = hr.y, bballow = bb.y, hallow = h.y)
fullplay <- rename(fullplay, gpitch = g.y, sopitch = so.y, batterhit = hbp.y)
fullplay <- rename(fullplay, game = g.x, run = r.x, hit = h.x, hr = hr.x, bb = bb.x, so = so.x, hbp = hb
```

Add fields and update some missing data values

Some fields needed to be derived. For example, the batting average is calculated. There are also some missing data that is replaced with values to help with plotting. Examples are included below.

```r
#  Calculate the batting average
fullplay <- fullplay %>% mutate(ba = hit/ab)
#  Replace missing data in specific columns
fullplay <- fullplay %>% mutate_at(vars(inducted), funs(replace(., is.na(.), 'N')))
fullplay <- fullplay %>% mutate_at(vars(allstar), funs(replace(., is.na(.), 0)))
```

Reduce to Hall of Fame Eligible Players

In this section, reduce the data set of players to those eligible for induction into the hall of fame. The rules say that the player must have played major league baseball for 10 years and they must be out of the game for 5 years. Also they must be elected within 10 years of retiring. So we'll reduce the set of players to those that meet those requirements.

```r
#  Add a column that calulates the years a player was in baseball
eligible <- fullplay %>%
  mutate(timein = floor(difftime(final_game, debut, units="days")/365)) %>%
  filter(timein > 10) %>%
  filter(final_game <= "2011-01-01")
```

Additional steps needed for the model builds

The column *inducted* either has the value "Y" or "N". To process the regression we need to create another column that has the value 1 if *inducted* is "Y" or 0 if it is "N".

```
eligible$Indicator <- ifelse(eligible$inducted == "Y", 1, 0)
```

Replace missing values with 0 in the eligible table.

```
eligible[is.na(eligible)] <- 0
```

It seems that it might make sense that the models will work better if the players are split by those that are pitchers and those that are not. Pitchers in general are not elected for their hitting abilities and fielder are not selected based on how they pitch. There are a suprising number of players that have pitched in at least one game. I made the threshold of games pitched, 163 based on Babe Ruth who certainly pitched more games than many, but was not elected for his pitching abilities.

```
#  First get the pitchers
pitchelig <- eligible %>%
  filter(!is.na(gpitch)) %>%
  filter(gpitch>163) %>%
  select(-ab, -ba,-bb, -cs, -double, -hbp, -hit, -hr, -ibb, -rbi) %>%
  select(-run, -sb, -sh, -so, -triple, -game)
# now the non pitchers
nonpitchelig <- eligible %>%
  filter(gpitch<164 | is.na(gpitch)) %>%
  select(-w, -l, -gpitch, -gs, -cg, -sho, -sv, -ipouts, -hallow, -er) %>%
  select(-hrallow, -bballow, -sopitch, -wp, -batterhit, -bk, -bfp) %>%
  select(-gf, -rallow)
```

Data Exploration

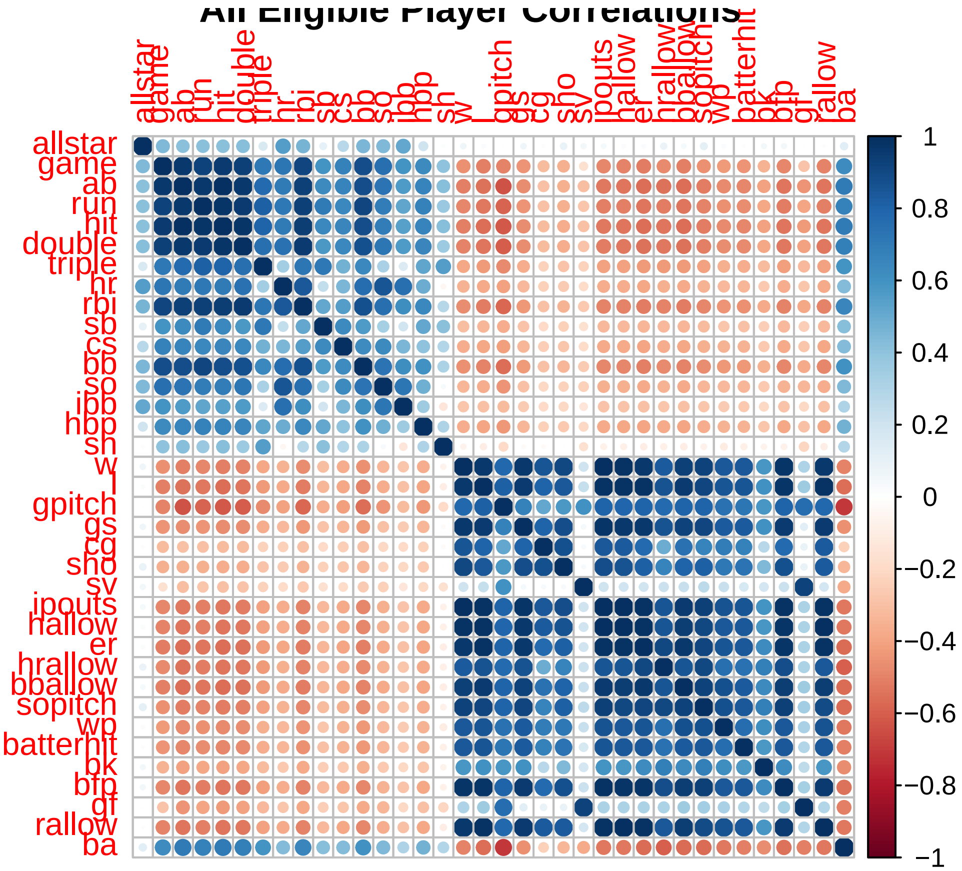
Box plots of some of the data

Scatter plots

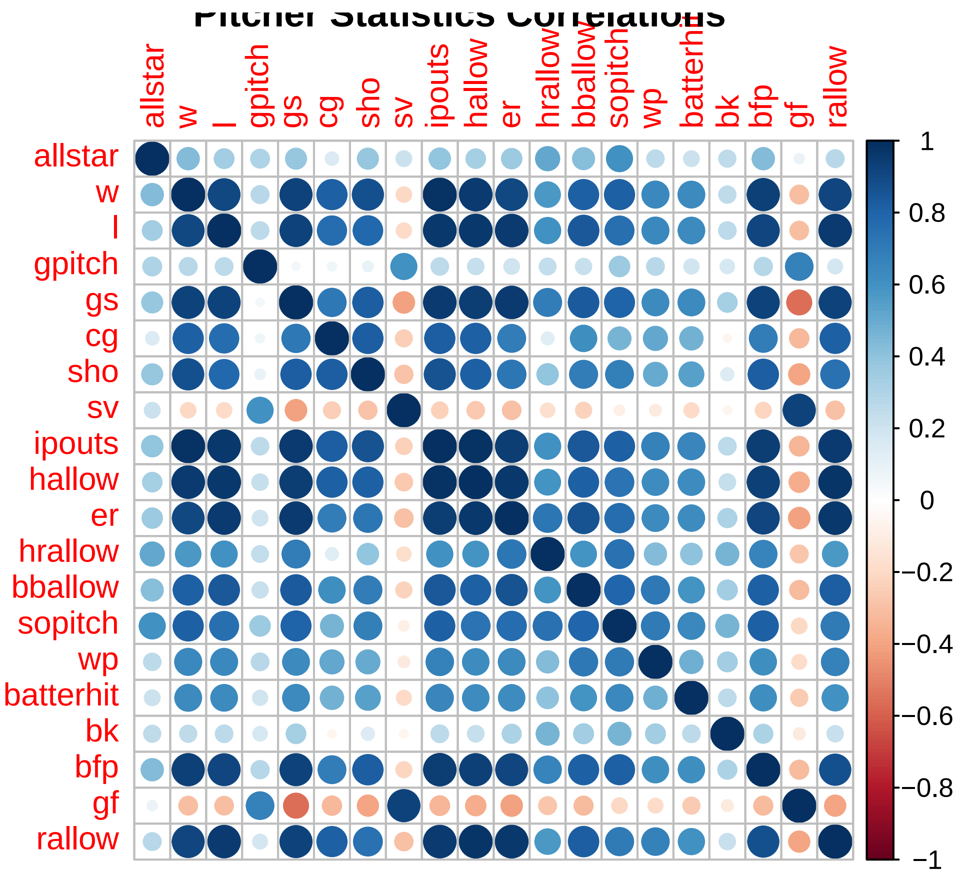
Look at the corelation

Some correlation plots might show us something.

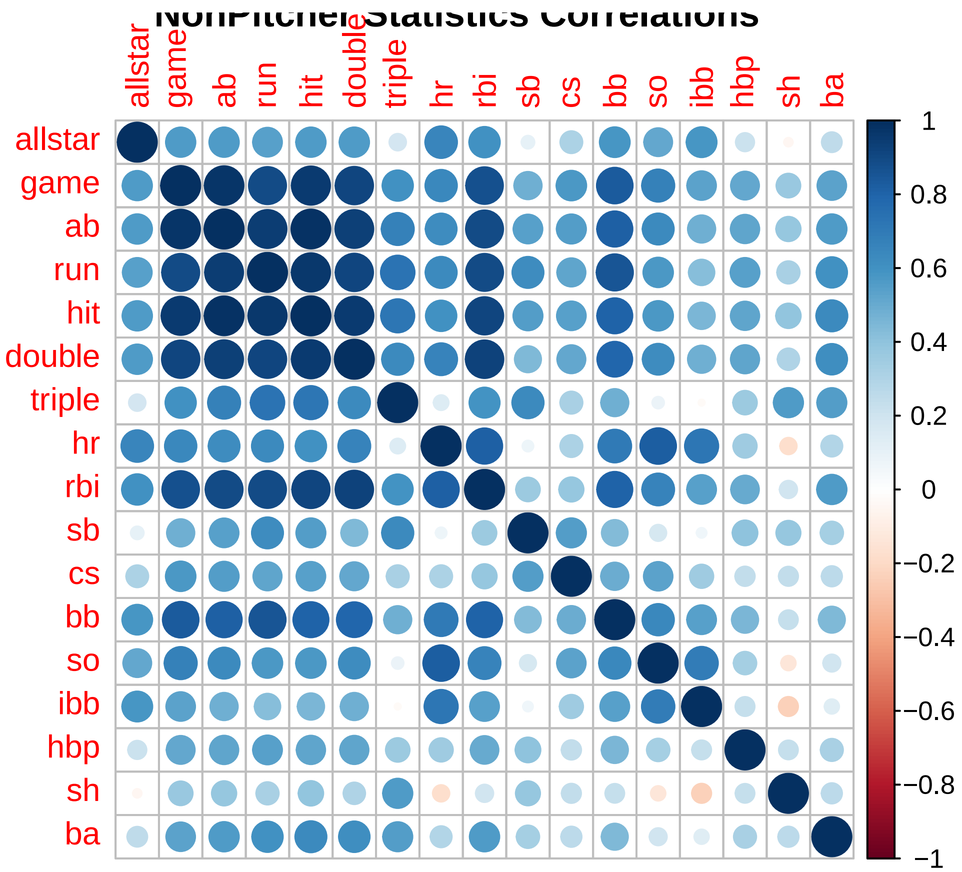
```
theCor <- cor(eligible[,7:42])
corrplot(theCor, title="All Eligible Player Correlations")
```



```
theCor <- cor(pitchelig[,7:26])
corrplot(theCor, title="Pitcher Statistics Correlations")
```



```
theCor <- cor(nonpitchelig[,7:23])
corrplot(theCor, title="NonPitcher Statistics Correlations")
```



Predictive Model

After some experimentation, Logistic Regression was selected as the type of predictive model to be used with the data set. For each model, data was split into training and test data subsets in order to evaluate the accuracy of the model.

The most promising predictors

The initial theory was that All Star game appearances my be a good predictor of a Hall of Fame career. I did also select these additional predictors based on what seems to be popular today as indicators of great players.

- allstar - All Star appearances
- hit - Career Hits
- hr - Career Home Runs
- rbi - Career Runs Batted In
- run - Career Runs Scored
- ba - Career Batting Average
- w - Career Wins for Pitchers
- gpitch - Career Games Pitched
- sopitch - Career Strikeouts thrown by a Pitcher
- er - Career Earned Runs allowed

After running a number of regression trials, based on correlation and relavance from those tests, the following were determine to be the best predictors.

- allstar - All Star appearances

- hr - Career Home Runs
- rbi - Career Runs Batted In
- run - Career Runs Scored
- ba - Career Batting Average
- w - Career Wins for Pitchers
- er - Career Earned Runs allowed

Looking at all players and their All-Star game appearances only

Using the full data set of eligible players. Let's look at all-star appearances only in regression and see what we get.

Make the trial and test sets of data

```
split <- sample.split(eligible$Indicator, SplitRatio = 0.7)
Train <- subset(eligible, split == TRUE)
Test <- subset(eligible, split == FALSE)
fit.allstar <- glm(Indicator ~ allstar, data = Train, na.action = na.pass, family = "binomial")
summary(fit.allstar)
```

```
##
## Call:
## glm(formula = Indicator ~ allstar, family = "binomial", data = Train,
##     na.action = na.pass)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4689  -0.3208  -0.2675  -0.2675   2.5877
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.31219    0.13578  -24.39   <2e-16 ***
## allstar      0.37125    0.02758   13.46   <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: 999.79  on 1669  degrees of freedom
## Residual deviance: 762.60  on 1668  degrees of freedom
## AIC: 766.6
##
## Number of Fisher Scoring iterations: 5
```

Now we'll fit the test data with the model and compute accuracy

```
predTest <- predict(fit.allstar, newdata = Test)
table(Test$Indicator, predTest > 0.5)
```

```
##
##     FALSE TRUE
##   0   649    3
##   1    53   10
```

```r
tmp <- myconfusion(Test$Indicator, predTest > 0.5)
```

```
##
##  Accuracy= 0.9216783
##  Sensitivity= 0.1587302
##  Specificity= 0.9953988
```

The accuracy is 92%. However it does not do a good job of predicting the true hall of fame players from the test set. Only correct on 15% of the cases.

If we change the threshold to 0.01, the true hall of famers from all star appearances only gets to 28% with a 93% accuracy.

Try with the best fit based on trial fits with all data

```r
fit.all <- glm(Indicator ~ allstar + hr + rbi + w, data = Train, family = "binomial")
summary(fit.all)
```

```
##
## Call:
## glm(formula = Indicator ~ allstar + hr + rbi + w, family = "binomial",
##     data = Train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.13530  -0.21583  -0.10497  -0.05456   2.94053
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.7560828  0.4732398 -16.389  < 2e-16 ***
## allstar      0.2918904  0.0388005   7.523 5.36e-14 ***
## hr          -0.0136736  0.0017742  -7.707 1.29e-14 ***
## rbi          0.0075262  0.0006118  12.301  < 2e-16 ***
## w            0.0242583  0.0021454  11.307  < 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: 999.79  on 1669  degrees of freedom
## Residual deviance: 463.29  on 1665  degrees of freedom
## AIC: 473.29
##
## Number of Fisher Scoring iterations: 7
```

Now we'll fit the test data with the model and compute accuracy

```r
predTest <- predict(fit.allstar, newdata = Test)
table(Test$Indicator, predTest > 0.5)
```

```
##
##     FALSE TRUE
##   0   649    3
##   1    53   10
```

```r
tmp <- myconfusion(Test$Indicator, predTest > 0.5)
```

```
##
##  Accuracy= 0.9216783
##  Sensitivity= 0.1587302
##  Specificity= 0.9953988
```

This did not really make any difference in the results.

Look at the results if we split into pitchers and non-pitchers

fit again, make train and test sets for both

```r
split = sample.split(pitchelig$Indicator, SplitRatio = 0.7)
pTrain = subset(pitchelig, split == TRUE)
pTest = subset(pitchelig, split == FALSE)

split = sample.split(nonpitchelig$Indicator, SplitRatio = 0.7)
npTrain = subset(nonpitchelig, split == TRUE)
npTest = subset(nonpitchelig, split == FALSE)

pfit.all <- glm(Indicator ~ allstar + w + er, data = pTrain, family = "binomial")
ppredTest <- predict(pfit.all, newdata=pTest)

npfit.all <- glm(Indicator ~ allstar + hr + rbi + run + ba, data = npTrain, family = "binomial")
nppredTest <- predict(npfit.all, newdata=npTest)
```

Prediction on the Pitchers

```r
summary(pfit.all)
```

```
##
## Call:
## glm(formula = Indicator ~ allstar + w + er, family = "binomial",
##     data = pTrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.22040  -0.11356  -0.04305  -0.02031   2.89081
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.226480   1.271146  -7.258 3.92e-13 ***
## allstar      0.396179   0.092230   4.296 1.74e-05 ***
## w            0.071891   0.012078   5.952 2.65e-09 ***
## er          -0.006714   0.001663  -4.037 5.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 326.80  on 582  degrees of freedom
## Residual deviance: 103.15  on 579  degrees of freedom
## AIC: 111.15
```

```
##
## Number of Fisher Scoring iterations: 8
```

```
table(pTest$Indicator, ppredTest > 0.3)
```

```
##
##     FALSE TRUE
##   0   226    3
##   1     5   15
```

```
tmp <- myconfusion(pTest$Indicator, ppredTest > 0.3)
```

```
##
##  Accuracy= 0.9678715
##  Sensitivity= 0.75
##  Specificity= 0.9868996
```

This shows an 97% accuracty rate and it does identify 75% of the actual Hall of Fame pitchers in the Test set. Much better than the combined results.

Prediction on the Non Pitchers

```
summary(npfit.all)
```

```
##
## Call:
## glm(formula = Indicator ~ allstar + hr + rbi + run + ba, family = "binomial",
##     data = npTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1895  -0.1697  -0.0650  -0.0219   3.2976
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.043e+01  2.743e+00  -7.447 9.52e-14 ***
## allstar      3.071e-01  5.243e-02   5.857 4.70e-09 ***
## hr          -1.060e-02  2.322e-03  -4.565 4.99e-06 ***
## rbi          4.256e-03  1.033e-03   4.119 3.81e-05 ***
## run          2.080e-03  6.955e-04   2.991  0.00278 **
## ba           4.491e+01  9.542e+00   4.706 2.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 672.27  on 1086  degrees of freedom
## Residual deviance: 254.30  on 1081  degrees of freedom
## AIC: 266.3
##
## Number of Fisher Scoring iterations: 8
```

```
table(npTest$Indicator, nppredTest > 0.3)
```

```
##
##     FALSE TRUE
##   0   421    2
##   1    24   19
```

```
tmp <- myconfusion(npTest$Indicator, nppredTest > 0.3)
```

```
##
##  Accuracy= 0.944206
##  Sensitivity= 0.4418605
##  Specificity= 0.9952719
```

For the non pitchers the accuracy is 94% and it predicts 44% of the actual hall of fame players in the test set.

I'm going to see about another set of data, for those not eligible in the data set but active or not yet elected to see if any of them have since been elected. A little help on this one would be appreciated.

Conclusions

All Star appearances alone is not a good predictor of Hall of Fame induction. It does seem that splitting the players into those that pitch and those that do not give better results. All Star appearances do seem to be a factor for both groups. In addition, for a pitcher the number of wins they have and the number of earned runs they allow are important factors. For hitters their home runs, batting average, runs batted in and runs scored are factors. This makes sense, as an indicator of their offense abilities. This model may be usefull to the veterns committee to review those player not in the hall of fame, that the model indicates should be. This number is not very large, but maybe after time some of the reasons they were not elected are not that critical in todays world.

The model is still much better at predicting those that will not make the Hall of Fame, but in just pure numbers that have played the game it makes sense. There are about 220 players in the hall of fame in this data set and approximately 18,800 players. That is a very small percentage (~1.2%).

Based on this project, it would be recommended that the model be evaluated to include some fielding data which does appear to be avaialble. The defensive ability of players may have an impact on induction as well. My quick look at the data did not seem to show many fields that would be helpful, but once the data is added maybe it would. It may also be worth looking at this from the standpoint of who should be voted on as a possible Hall of Fame player. That may be a better prediction since many things can influence the voters.