Predicting the Members of the NBA Hall of Fame

Josh Miller, Kaizhi Lu, Mingyang Xue

March 16, 2021

Abstract

Our objective is to predict the members of the NBA Hall of Fame. Using basic counting stats in the NBA, we will construct three models and compare their predictive power.

1 Introduction and Overview

Nowadays basketball has become one of the most favorite and popular sports. Among all of the professional leagues in the world, the National Basketball Association (NBA) is the premier men's league. And among all of the NBA awards and honors, being a member of the Hall of Fame—honoring players who have shown exceptional skill at basketball and major contribution to the sport—is the highest honor. What makes a player being eligible for the membership? Can we tell if a player is going to be a member from observing his basic stats, such as points, assists, rebounds, etc.?

This paper is aiming to answer these questions, to predict if an eligible player will make the Hall of Fame given various statistics. Using stats for a players career on a per game basis, we will try to construct a few models that predict if a player is worthy enough for this achievement. In other words, we are going to build models that can accurately predict the likelihood that a player can make the Hall of Fame from basic game stats.

This paper is divided into three main parts: 1. Methods: In this part, we give a detailed description of where we get the data, how we preprocess them, and a brief introduction of the different models we are going to apply to the data. 2. Computational results: This is the main part of our paper where we use R-studio to compare different models and choose the best one through computing model statistics and making corresponding plots. 3. Summary and Conclusions: This is the part where we summarize our findings and make the real-world conclusions.

2 Methods

To answer our question, we first need to acquire the data from the package nbastatR. We do this by installing the package from Alex Bresler's github with the help of devtools [1]. We also install the following packages to R: caret, e1071, brglm2, rpart, and rpart.plot. Then, we access the dictionary of all players in NBA history with the function assign_nba_players(), which is a built in function for nbastatr.

In order to set up our problem, we need to know the names of each player that is in the NBA Hall of Fame. With the names of the players, we construct a vector containing all Hall of Fame members [3]. We then make a variable that gets all player ids, and randomly select 350 ids out of the total to help build our sample.

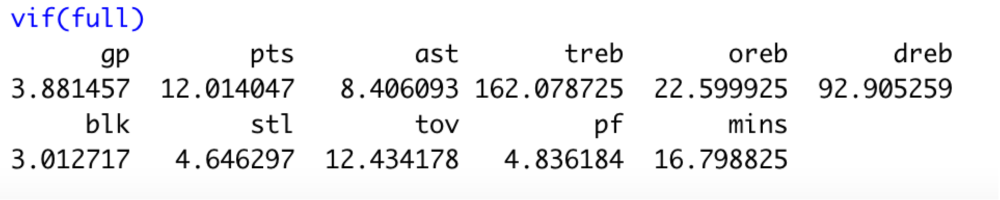
Now, we create our data. We use the player_careers() function, identifying the names of the Hall of Fame players, as well as a few ids due to players having the same name, and the 350 randomly selected player ids. Some of those randomly selected ids may overlap with players we already have from the Hall of Fame vector, and the program recognizes this and does not include duplicates. This led to a sample size of 479. While this is a relatively small sample, we choose this value because the program was unable to run with a large dataset due to the construction of the nbastatR package. We specify the mode as ”PerGame” to get per game stats. This gives us a few data tables that have all of the career statistics for each player that is in our sample. We then create the variables that we want by pulling them from the data table. The variables of interest are: names, games played (gp), points (pts), assists (ast), total rebounds (treb), offensive rebounds (oreb), defensive rebounds (dreb), blocks (blk), steals (stl), turnovers (tov), personal fouls (pf), and minutes (mins), where all predictors except for gp are on a per game basis. We then create a binary response variable, using the ifelse() function to go through all of the names we have and identify which players are in the Hall of Fame and assign them a 1, with all other players getting a 0.

We take all of these variable and build a data frame. Some players have NA values for certain statistics, since stats like blocks and steals weren’t tracked until the 1973-1974 season [2]. So, we use the na.omit function on our data.frame. We also manually removed certain present day players such as LeBron James and Kevin Durant, as we put them in our data to see their percentages out of curiosity, and their results would significantly affect the model as they will be Hall of Fame players when their careers are over. We then build the training and test set, with a 70-30 split, since our dataset is not very large.

At this point, we are able to begin our analysis. We build a logistic regression model with all of our predictors, and perform tests to find the best models for a basic model with no interaction, an interaction model, and a tree model. For the interaction model, we fit a brglm model with the help of the brglm2 package. The model selection process includes observing variance inflation factors, stepwise regression, and anova. Once we have the best model, we make predictions, construct a confusion matrix for our test set to see the true negatives/positives and false negatives/positives, and plot a bar graph of some players. This is done for the basic and interaction model. For tree, we plot the tree model and make predictions on the test set.

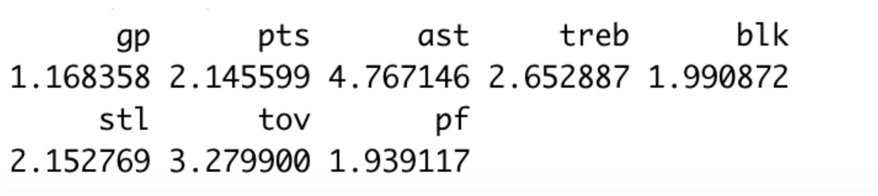
3 Computational Results

First, we need to check the multicollinearity between the variables that we are interested in using. In Figure 1, we see the results the variance inflation factors of each predictor. As we see, the variables treb, oreb, dreb, and minutes have high values, as well as pts and turnovers. Intuitively, since treb is the sum of oreb and dreb, there will be problems, so we chose to remove oreb and dreb. Also, by playing more minutes, a player will accumulate more stats. So, we remove minutes as well. We can see the variance inflation factors for the remaining variables now in Figure 2.



```
vif(full)
       gp       pts       ast       treb      oreb      dreb
 3.881457 12.014047  8.406093 162.078725 22.599925 92.905259
      blk       stl       tov         pf      mins
 3.012717  4.646297 12.434178   4.836184 16.798825
```

Figure 1: Variance Inflation Factors for all predictors.

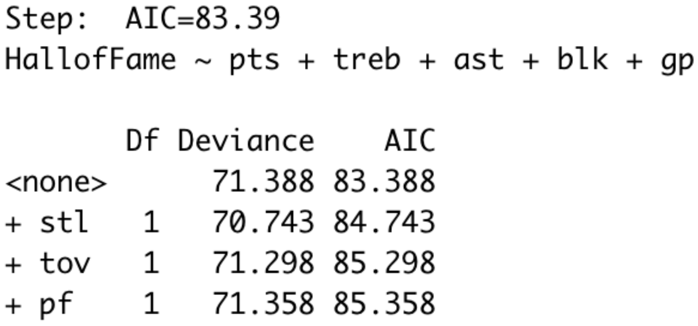


```
      gp      pts      ast     treb      blk
1.168358 2.145599 4.767146 2.652887 1.990872
     stl      tov       pf
2.152769 3.279900 1.939117
```

Figure 2: Variance inflation factors for predictors excluding oreb, dreb, and mins.

These values are all below 5, and thus are good. So, we use these variables in our full model before model selection.

Now, we are ready to select the best model without interaction (basic model) from these variables. We perform stepwise regression in both the backward and forward directions. As it turns out, both the forward and backward method resulted in the same model, and the results are shown in Figure 3. Since the models are the same, we do not need to check the Leave-One-Out Cross-Validation.

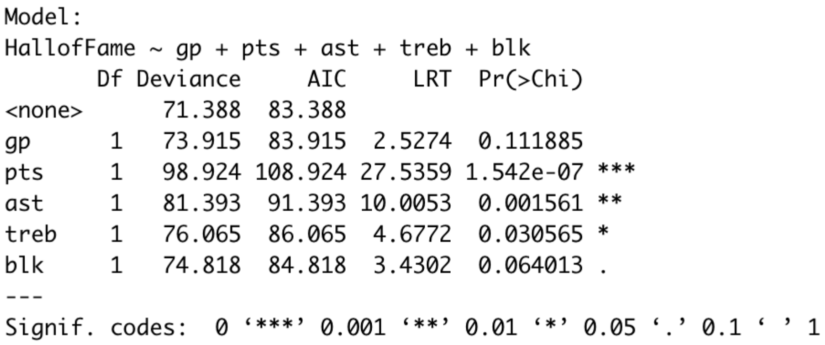


```
Step:  AIC=83.39
HallofFame ~ pts + treb + ast + blk + gp

       Df Deviance    AIC
<none>      71.388 83.388
+ stl   1   70.743 84.743
+ tov   1   71.298 85.298
+ pf    1   71.358 85.358
```

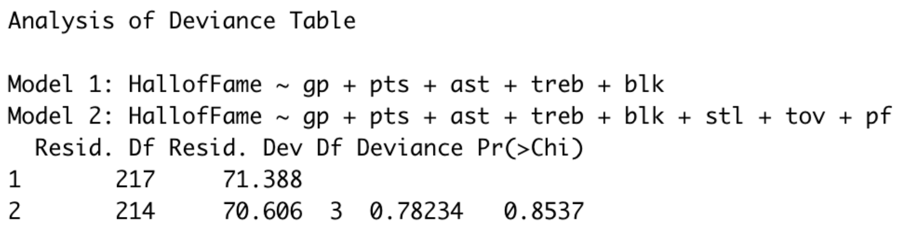
Figure 3: Stepwise Regression Results.

For further evidence that these variables make up the best model, we use the drop1() command to see which variables are necessary for the model at an $\alpha = 0.1$ significance level. As we can see in Figure 4, all of the variables are significant and will be included in the model. We also perform ANOVA, with the null hypothesis that all of the coefficients in the full model that are not in the reduced model are equal to 0. In other words, H_0: $\beta_{stl} = \beta_{tov} = \beta_{pf} = 0$, and the results are shown in Figure 5. Since the p-value is greater than .1, we fail to reject the null hypothesis.



```
Model:
HallofFame ~ gp + pts + ast + treb + blk
       Df Deviance     AIC     LRT  Pr(>Chi)    
<none>      71.388  83.388                      
gp      1   73.915  83.915  2.5274  0.111885    
pts     1   98.924 108.924 27.5359 1.542e-07 ***
ast     1   81.393  91.393 10.0053  0.001561 ** 
treb    1   76.065  86.065  4.6772  0.030565 *  
blk     1   74.818  84.818  3.4302  0.064013 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: Drop1 Results.

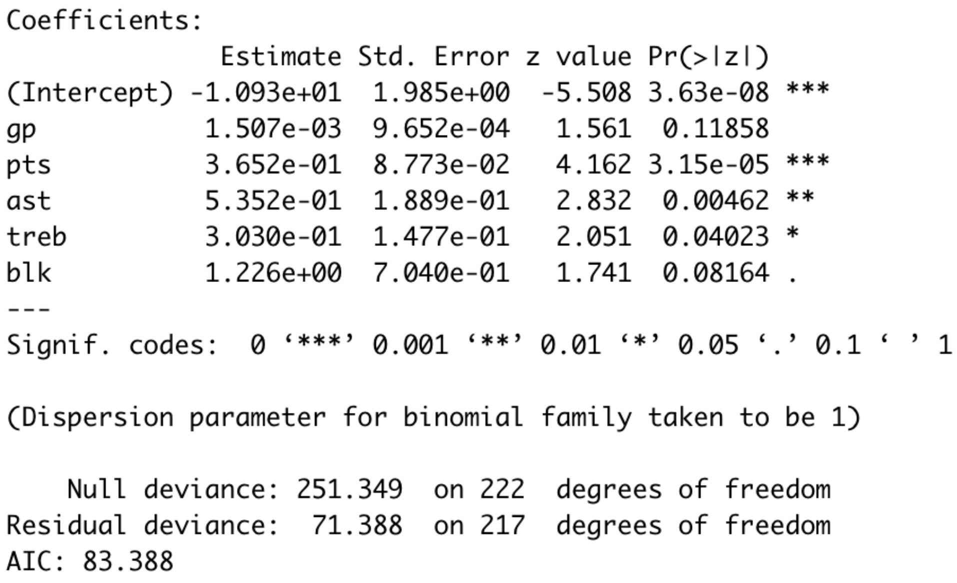


```
Analysis of Deviance Table

Model 1: HallofFame ~ gp + pts + ast + treb + blk
Model 2: HallofFame ~ gp + pts + ast + treb + blk + stl + tov + pf
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       217     71.388
2       214     70.606  3  0.78234   0.8537
```

Figure 5: ANOVA Results.

With these results, we arrive at the conclusion that our best model without interaction contains the variables pts, treb, ast, blk, and gp. Figure 6 shown below shows the summary for the model. We see that all of the predictors except for gp are significant at the $\alpha = .1$ level. The coefficients represent the change in log odds for a player to make the Hall of Fame with a 1 unit increase in the predictor.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.093e+01  1.985e+00  -5.508 3.63e-08 ***
gp           1.507e-03  9.652e-04   1.561  0.11858
pts          3.652e-01  8.773e-02   4.162 3.15e-05 ***
ast          5.352e-01  1.889e-01   2.832  0.00462 **
treb         3.030e-01  1.477e-01   2.051  0.04023 *
blk          1.226e+00  7.040e-01   1.741  0.08164 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.349  on 222  degrees of freedom
Residual deviance:  71.388  on 217  degrees of freedom
AIC: 83.388
```

Figure 6: Summary of Base Model.

Next comes the prediction. We first plotted the cutoffs for deciding which player is a Hall of Fame member, and chose the one with the highest accuracy on the training data. This plot can be seen in Figure 7, where a 0.65 cutoff resulted in a 94.17% accuracy in our training set. Then, we perform our prediction on our test data, and plot the confusion matrix to show the accuracy. The confusion matrix is shown in Figure 8, with the green being true positives and true negatives, while the red is false positives and false negatives. The prediction on the test set had 96.88% accuracy.

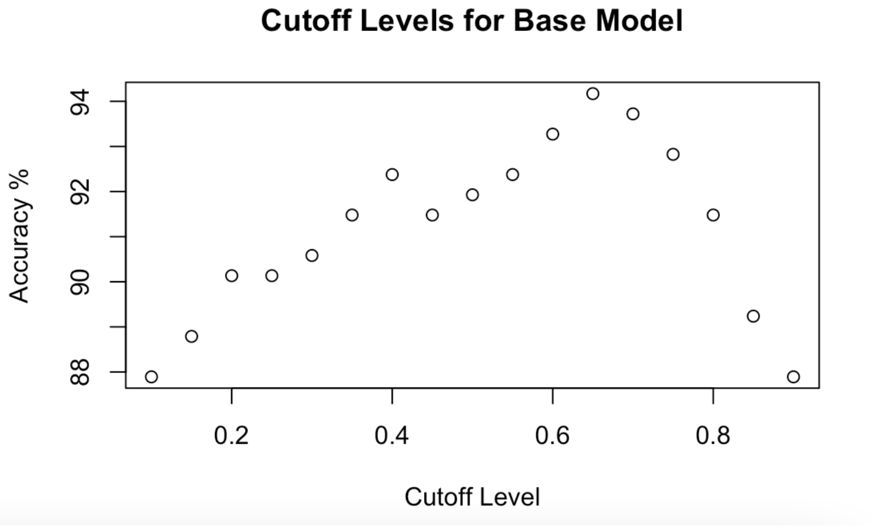


Figure 7: Accuracy levels for each cutoff to decide if Hall of Fame or not.

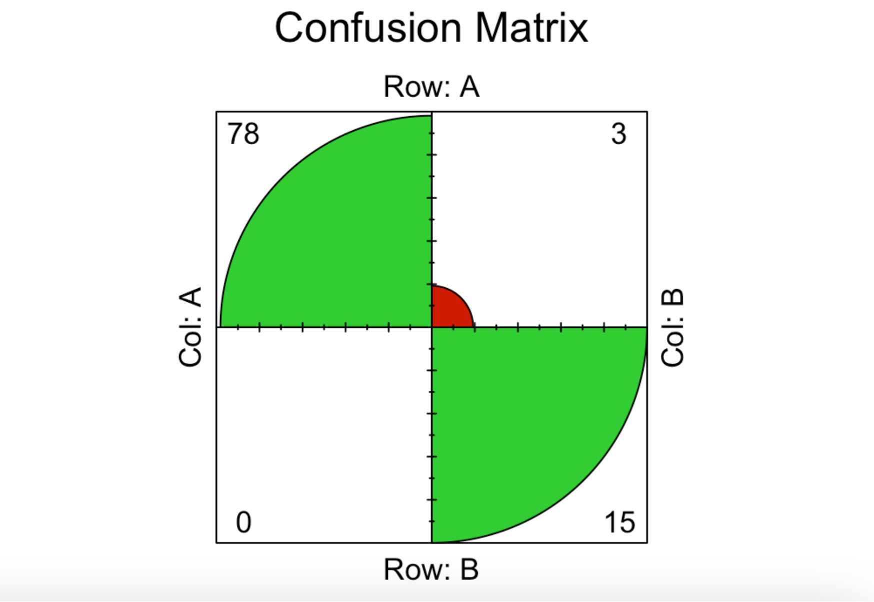
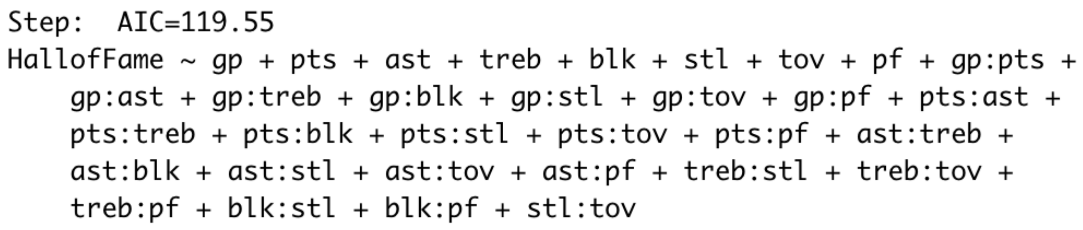


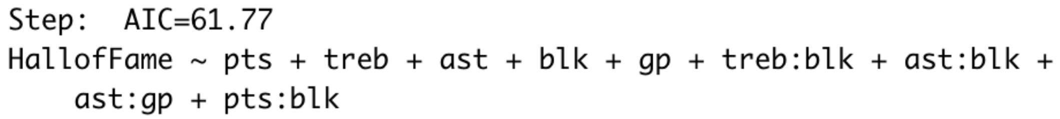
Figure 8: Confusion Matrix results for our Test Data.

Now, we build our interaction model. We use the same variables that we determined are safe from our variance inflation factors, and we add all 2 way interactions. We then perform stepwise regression on the full interaction model to determine a reduced model. The results for both forward and backward stepwise regression are shown below in Figure 9 and Figure 10.



```
Step:  AIC=119.55
HallofFame ~ gp + pts + ast + treb + blk + stl + tov + pf + gp:pts +
    gp:ast + gp:treb + gp:blk + gp:stl + gp:tov + gp:pf + pts:ast +
    pts:treb + pts:blk + pts:stl + pts:tov + pts:pf + ast:treb +
    ast:blk + ast:stl + ast:tov + ast:pf + treb:stl + treb:tov +
    treb:pf + blk:stl + blk:pf + stl:tov
```

Figure 9: Backward Stepwise Regression on Interaction Model.

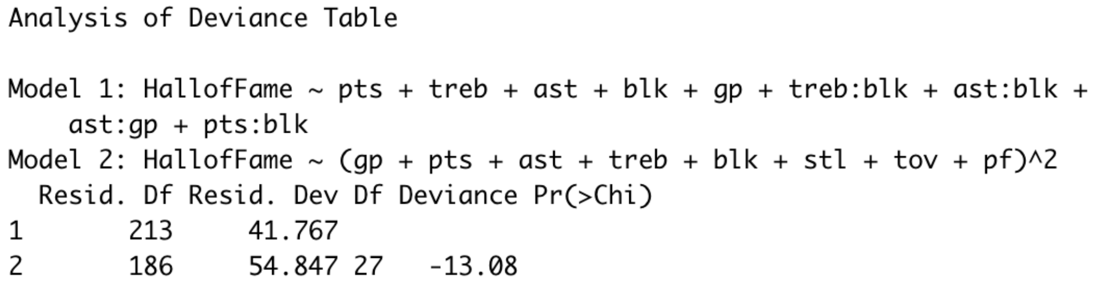


```
Step:  AIC=61.77
HallofFame ~ pts + treb + ast + blk + gp + treb:blk + ast:blk +
    ast:gp + pts:blk
```

Figure 10: Forward Stepwise Regression on Interaction Model.

As we see, these two models differ. So, we need to look at the Leave One Out Cross Validation value, and choose the model that yields the smaller value. The results are 1.289 for the backward regression and 0.379 for the forward regression. Thus, we proceed with the forward selection model.

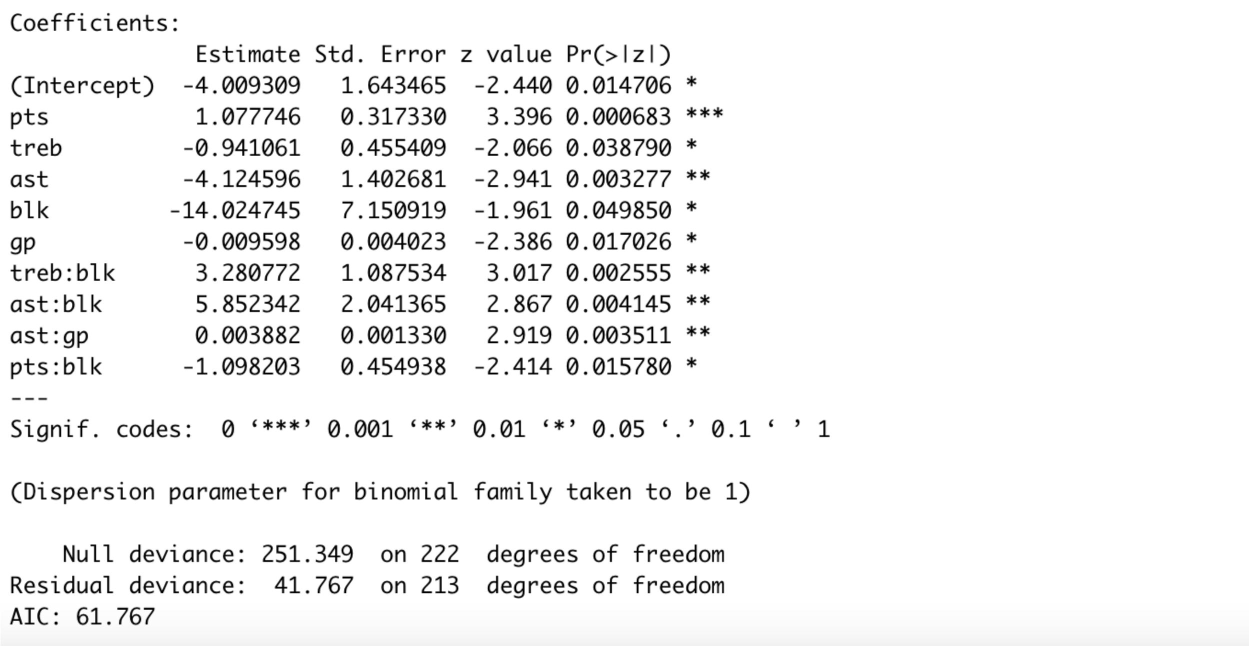
We need to confirm that the reduced model from the forward selection process is better than the full interaction model. This is done with ANOVA. In Figure 11, we see that we get a negative difference for the deviance. Checking the p-value with the built in command pchisq(), we find that the p-value is 1. Thus, we reject the null hypothesis that the full model is better, and conclude that our reduced model is the one we should use. Figure 12 below shows the summary of the reduced interaction model we use, and we can see that all of the predictors are significant at the $\alpha = .05$ significance level, and thus significant at our $\alpha = .1$ level.



```
Analysis of Deviance Table

Model 1: HallofFame ~ pts + treb + ast + blk + gp + treb:blk + ast:blk +
    ast:gp + pts:blk
Model 2: HallofFame ~ (gp + pts + ast + treb + blk + stl + tov + pf)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       213     41.767
2       186     54.847 27   -13.08
```

Figure 11: ANOVA results for the interaction model.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.009309   1.643465  -2.440 0.014706 *
pts           1.077746   0.317330   3.396 0.000683 ***
treb         -0.941061   0.455409  -2.066 0.038790 *
ast          -4.124596   1.402681  -2.941 0.003277 **
blk         -14.024745   7.150919  -1.961 0.049850 *
gp           -0.009598   0.004023  -2.386 0.017026 *
treb:blk      3.280772   1.087534   3.017 0.002555 **
ast:blk       5.852342   2.041365   2.867 0.004145 **
ast:gp        0.003882   0.001330   2.919 0.003511 **
pts:blk      -1.098203   0.454938  -2.414 0.015780 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.349  on 222  degrees of freedom
Residual deviance:  41.767  on 213  degrees of freedom
AIC: 61.767
```

Figure 12: Summary for the interaction model.

For the prediction portion of our interaction model analysis, we proceed exactly as we did for the base model. We found a cutoff point that resulted in the highest accuracy for our training set, which turned out to be a cutoff at 0.5 with an accuracy of 96.41%. Figure 13 shows the confusion matrix of our predictions for the test set, with 88 of the 96 players correctly classified. The resulting accuracy on our test data was 91.67%.

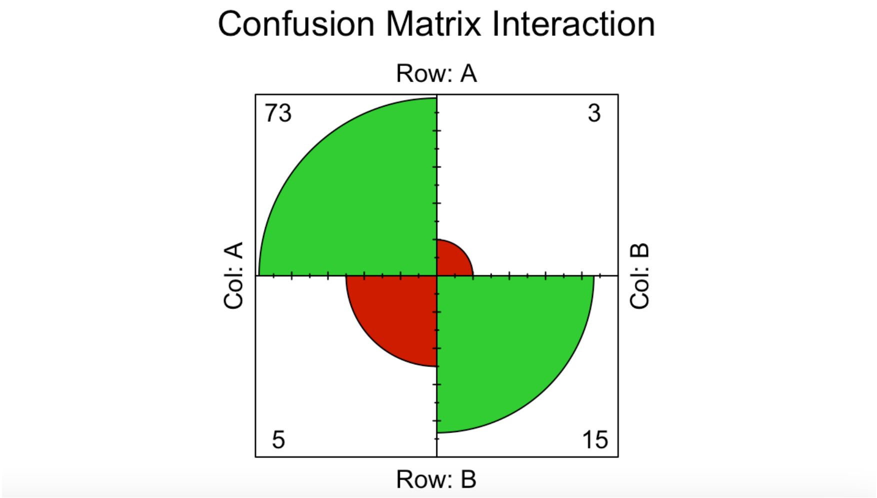


Figure 13: Confusion Matrix for the interaction model on the test set.

Notice that the training accuracy was higher in the interaction model than the base model, but the test accuracy was higher for the base model. This could be a result of over-fitting in the interaction model, but since the test set is not that much worse, it is most likely not the case. While the prediction power of both the base model and the interaction model are above 90%, the base model is about 5% more accurate, so it is technically the better model.

Now, we will check the results for 4 Hall of Fame members, and see how our model performed with these individual players. As we see from Figure 14 and Figure 15, both the base model and interaction model gave similar results, with slightly differing odds (seen on the y axis). However, we see that some players, namely Dennis Rodman and Sarunas Marciulionis, had very low odds to get into the Hall of Fame. This shows that individual stats may not be the only key to getting into the Hall of Fame. Dennis Rodman was known for his defense which cannot be easily tracked by stats and Sarunas Marciulionis was a pioneer for international players coming to the NBA [4].

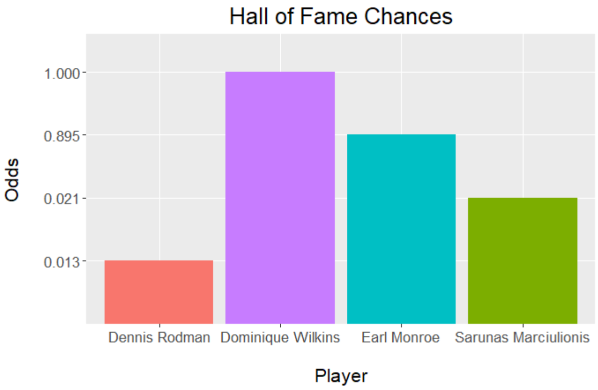


Figure 14: Bar Graph for base model.

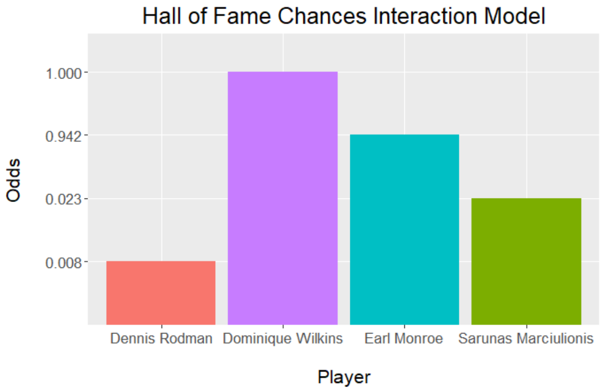


Figure 15: Bar graph for interaction model.

Lastly, we fit a tree model with the 8 predictors that we determined were not correlated with each other. We fit our tree and then pruned the tree. Pruning was decided by finding the smallest model whose error is below the limit, which can be seen in Figure 16. Figure 17 shows the tree model after pruning, with two splits decided by points and treb.

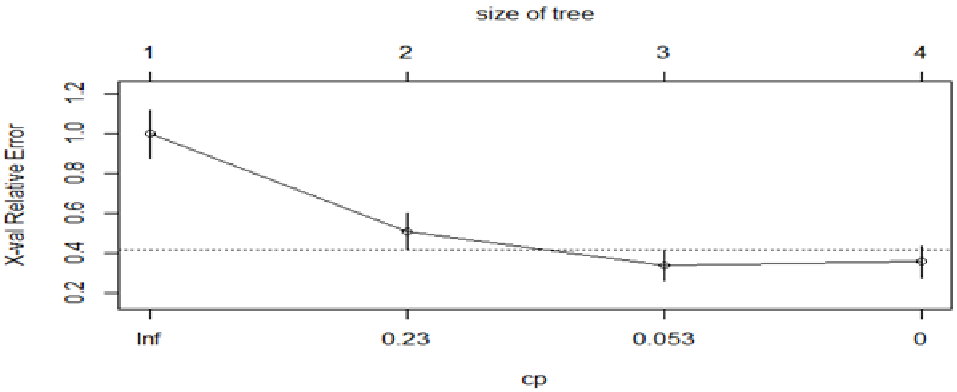


Figure 16: CP of tree to determine where to prune.

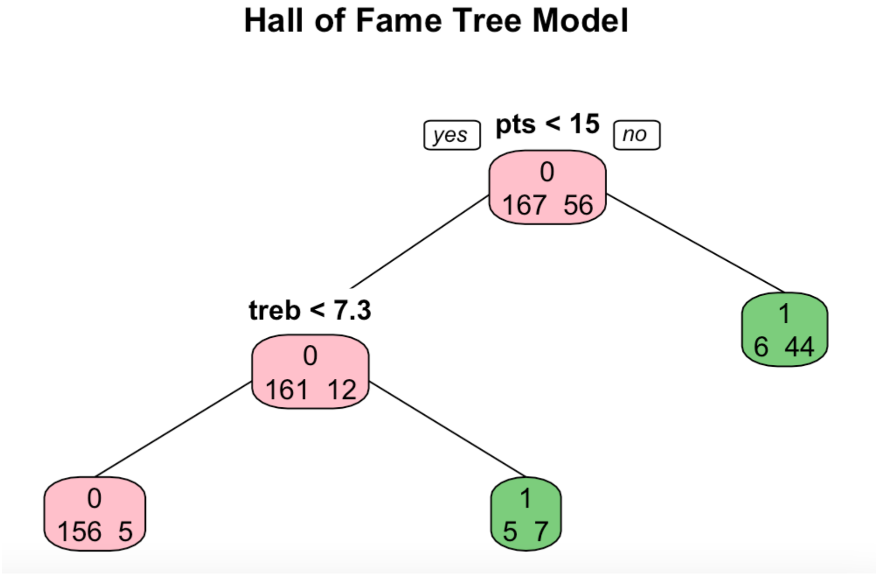


Figure 17: Tree model Diagram.

Using this pruned tree, we see that pts and treb were the only two predictors that contributed to classification. We made predictions on the training and test set, and the tree model on the training set was 92.74% accurate and was 95.05% accurate on the test set. This test accuracy was higher than that for the interaction model, but lower than that of the basic model.

4 Summary and Conclusions

As we described in the introduction, our goal is to predict the members of the NBA Hall of Fame by plugging some basic game statistics of players into our model. In the methods and computational results part, we designed three models (i.e. two logistic regression models and a tree model) and checked multiple statistical criteria for these models such as the variance inflation factors, AIC, ANOVA test. We choose logistic regression model and tree model because they work well for classification problems. The model that performs the best in our test is the logistic regression model with five predictors (i.e. game played, points, assists, total rebounds, blocks) and no interaction effect. It turns out the chosen model shows an accuracy of prediction up to 96.88% as we use a portion of data randomly selected from the original set of data as the test set.

On the other hand, we are aware that there may be some other factors to make certain players get into the Hall of Fame. This is perfectly reasonable as every player is unique, so basic stats obviously can not tell the whole story, and this is also the reason why our model is not 100% accurate when we applied it to our test set.

Another possible improvement could be made on the sample size. The sample size of our data is 479, which is not terrifically large, and we decided to select 70% as our training set and 30% as our testing set. Although we consider the size to be large enough for building and training our selected model, we still feel like the size of test set is not large enough. In general, a larger size of dataset will lead to a more accurate model selection and prediction. Therefore, if more data is available, we could probably further improve our model and raise the predictive accuracy.

Overall, given analysis we performed based on the data we currently have, our model is fairly accurate, and capable to be seriously considered as a reliable predictive model for testing whether a NBA player could get into the Hall of Fame after retirement as long as we have the data of his games.

References

[1] Alex Bresler. *nbastatR*. URL: `http://asbcllc.com/nbastatR/`. Accessed: 03.10.2021.

[2] Brian Martin. *Season Total Tiers: Rebounds, Steals and Blocks*. URL: `https://www.nba.com/stats/articles/season-total-tiers-rebounds-steals-and-blocks`. Accessed: 03.11.2021.

[3] *NBA Players*. URL: `https://en.hispanosnba.com/players/hall-of-fame/index`. Accessed: 03.10.2021.

[4] *Sarunas Marciulionis*. URL: `http://www.hoophall.com/hall-of-famers/sarunas-marciulionis`. Accessed: 03.11.2021.