ANALYSIS OF BASEBALL HALL OF FAME VOTING PATTERN USING DECISION TREES CLASSIFICATION METHOD

Prepared for UMGC DATA 630: Machine Learning

JULY 7, 2020

VANESSA FOTSO

Introduction

Baseball is one of the most popular sports in the United States that enthusiasts both participants and spectators. Major League Baseball is the highest level of baseball division in the U.S. and every year, a select few, yet greatest players, are honorably enshrined in the National Baseball Hall of Fame (HOF). The selection process of players into the National Hall of Fame has been a subject of debate among many members in the baseball community, including spectators and media, as it relies on subjective votes from a committee of Baseball Writers' Association of America (BBWAA) and the Veteran's Committee (VC).

As per BBWAA criteria, to be considered on voting ballots for election into HOF, a player must have been competed in at least ten seasons at the professional level and have been retired for at least five seasons. To be elected by the BBWAA committee, a player must obtain at least 75% of all ballots cast in an election. If a player meet all the criteria, but is not elected into HOF, he will maintain his consideration for a total of 20 years assuming he has received a "yes" vote on at least 5% of the ballots in each election. However, a player is no longer eligible on the BBWAA ballot if he has not been elected within 20 years of retirement, and his only chance for induction into Hall of Fame will be through the Veterans Committee (Baseball Reference, 2009).

The significant controversy over the election into HOF provides a great platform for machine learning predictive analysis. The subjective election of a player into Hall of Fame makes this prediction not trivial to a model. However, past research investigating

BBWAA decision patterns revealed satisfactory results on the consistency of the committee voting behavior based on standard baseball statistical criteria.

This paper aim to investigate which performance characteristics are the most significant to the voting committee in electing a Hall of Fame member. Decision trees will be used to build a model that can predict the status of Hall of Fame membership for an eligible player, factoring in all attributes that may impact the decision. Decision trees is a supervised machine learning classification technique which successively partition into smaller subsets of record, beginning at the root node and branching out into internal node or leaf node. Each branch represents a data variable that falls within a certain range of values. The leaf node represents the decision or class label, and the path from the root to the leaf node represents the set of rules leading to the decision of classification being made. This technique is suitable for the problem at hand as it is adaptable, measurable and can determine and apply rules in order of importance.

Data Description and Preprocessing

The data was obtained from the Journal of Statistics Education (JSE) Data archive website. It entails the career records for 1340 modern position players eligible for the Major League Baseball Hall of Fame (MLBHOF). The data consists of various attributes (27 variables) for a given eligible baseball player, including the number of seasons played, games played, first, second and third base, home runs, batting averages, Hall of Fame status etc. The Hall of Fame Membership is the variable of interest in this study and has three possible values (0, 1, 2), with 0 meaning the player was not admitted to the

HOF, 1&2 representing the two different ways of election (BBWAA or VC). A summary of the attributes in the dataset, as well as the attributes' types, is provided in figure 2 and 3. Most attributes are numerical, with the exception of the player's name and primary position played. The HOF membership being the responsive variable was transformed to factor variable as the output must be categorical for decision trees method. The summary in figure 2 also shows that the variables strikeouts, stolen bases, caught stealing and stolen base runs contain missing values which were replaced by their respective mean values. Handling missing values here is necessary to optimize the model. Figure 4 presents the data summary after transformation. Out of 1340 eligible players, only 124 were admitted into the Hall of Fame.

Classification Using Decision Tree

In Machine Learning Classification technique, the model is trained with a set of data from the whole dataset called a training set which consist of records having multiple variables. each record in the training set has a labeled class. The goal is to use the training data to build a model that can predict each class given the attributes. The model accuracy is then tested on a test set, which has records with unknown class.

The analysis used R language to implement the conditional inference trees model. Ctree() is a function from R 'party' package used to create decision tree model. The main arguments for this algorithm are the formula and data. The formula is the decision model or equation used to make predictions. The formula is in the form of y ~ x, where y is the response variable the model is trying to predict, and x represents a set of factors for the decision nodes. The data argument provides the ctree function with the dataset to pull the variables given in the formula.

The data was randomly divided into 70% training set and 30% test set. The training set was used with the ctree function to create a conditional inference tree. The formula used for the prediction is hall of fame membership as a function of the rest of variables in the dataset (Hall.OF.Fame.Membership ~ .). The resulting tree print and plot are displayed in figure 5 and 6 respectively. The model was then tested on the test set for evaluation and the confusion matrix was generated to visualize the performance of the model on both training and test sets and evaluate the model efficiency. The confusion matrix helps determining the classification accuracy recall and precision. The accuracy is the ratio of the correctly classified records over the total number of records. The recall is given the ratio of the total number of correctly classified positive instances by the total number of positive instances. The higher the Recall the more likely the class will be correctly recognized. Last, the precision of the model is the ratio of the total number of correctly classified positive instances by the total number of predicted positive examples. High Precision confirms an instance classified as positive is indeed positive. The confusion matrix for both training and test set can be visualized in figure 7 and 8 respectively.

Results and Interpretations

Figure 5 shows that the model was built on 952 observations and the resulted tree contains 11 terminal nodes. The key variable determining the induction of a player into Hall of Fame is the total player rating, which is located at the root of the tree. This root node has two leaves: first, if the total player rating is at most 23.8 then the second splitting decision attribute is runs score. If runs score is at most 1184, runs batted in is the next determining factor. If runs batted in is less than or equal to 1187, then the model will look at the value of the total player rating again as the fourth splitting factor and so on until it reaches the terminal node 8. The path to the leaf node 8 indicates that there are 748 players on that path (following that set of rules), and of these 748 players, about 90% of them falls into the negative status for the HOF (class 0), while the other 10% falls into class 2. The decision tree follows the if ..then process until a terminal node (decision outcome) is reached. The resulting tree has 11 terminal nodes, revealing 11 set of rules impacting the election of a given player into the Hall of Fame. Each set of rules could be determined by following the path from the root node (total player rating) to a leaf node (8, 9, 10, 11, 12, 13, 14, 18, 19, 20, or 21). Each leaf node provides the number of players in its path, the x axis represents each HOF Membership status, while the y-axis represents the proportion of the total players in that path that correspond to each HOF status. The shortest paths generating a decision quickly are total player rating < = 23.8 -> runs scored > 1184 -> outcome of 0 for HOF in node 14, and total player rating > 23.8 -> home runs > 414 -> outcome of 1 for HOF in terminal node 21. Thus, from this resulting model, Total player rating, run scored and home runs can be determined as the strongest determining factors taken into consideration by the baseball voting committee for electing a player into Hall of Fame.

From the generate confusion matrix on the training set, the model has a classification accuracy of about 94%. The model has a precision of about 95% and a recall of about 99% for HOF class 0, which means that out of the times class 0 was predicted, the model was in fact correct 95% of the time. Additionally, the model correctly predicted 99% of class 0. Additionally, the system was correct in predicting class 1 100% of the times, and class 2 only 60% of the time; while out of the times class 1 or 2 should have been predicted, only 40% and 44% of class 1 and 2 were correctly predicted respectively. These results shows that the model is somehow confused between class 1 and 2 for Hall of Fame Membership, mislabeling the two classes. Overall, the model precision on the training set was 85% while the recall was 61%

The classification accuracy on the test dataset was about 92%, 2% lower than the one using the training set. The overall precision of the model here was 71% while the recall was 45%. The accuracy and the precision of the model are significantly high, while the recall is significantly low. The model is indeed correct at predicting the status for hall of fame 71% of the time, while for all instances of players that should have a given status, the model only captured 45% of the given class correctly. So, the accuracy of the model could be considered as misleading. It looks like the data in the analysis is imbalanced. This could be due to the fact that a large amount of missing values was replaced by the mean for the stolen base and stolen base runs variables in the dataset. Since these two variables have a lot of missing information, dropping these attributes from the analysis could be have been a better alternative.

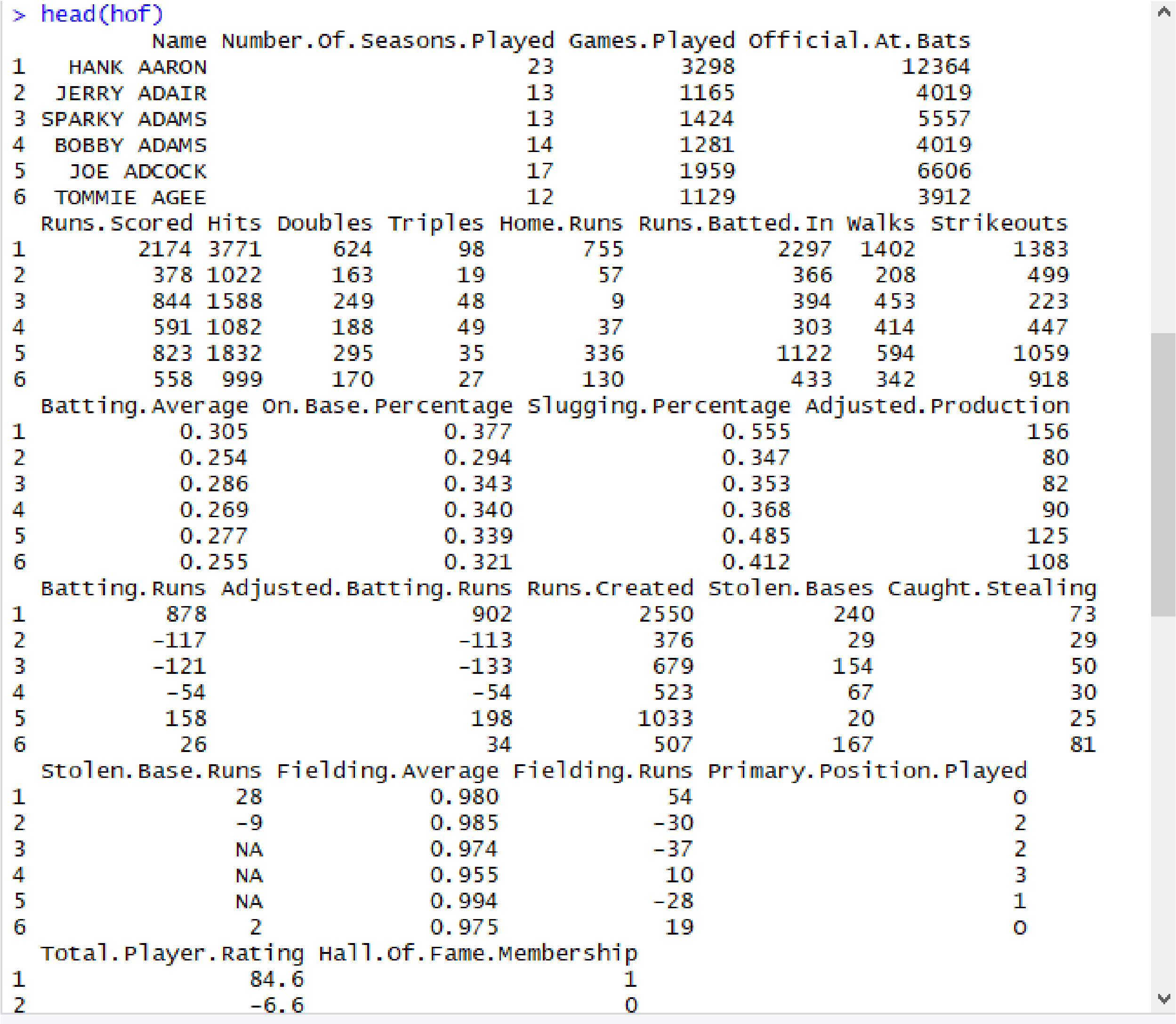
Conclusion

The derived model pointed total player rating, run scored and home runs as the strongest factors in considering an eligible former baseball player for induction into the National Baseball Hall of Fame. Unfortunately, the model did not perform very well in defining the set of rules leading to the election of a player into the Hall of Fame. Although the model has fairly high accuracy of 92% and precision of 71% in prediction unknown status, the recall statistic of 45% is significantly low, leading to suggest additional or alternative data preprocessing method(s) and fitting of the model. Dropping the variables with excessive missing information and implementing additional modeling techniques and algorithms could help optimize the model. For example, the Apriori method could be used to first evaluate the correlation between the attributes in order to derive more accurate set of rules.

References

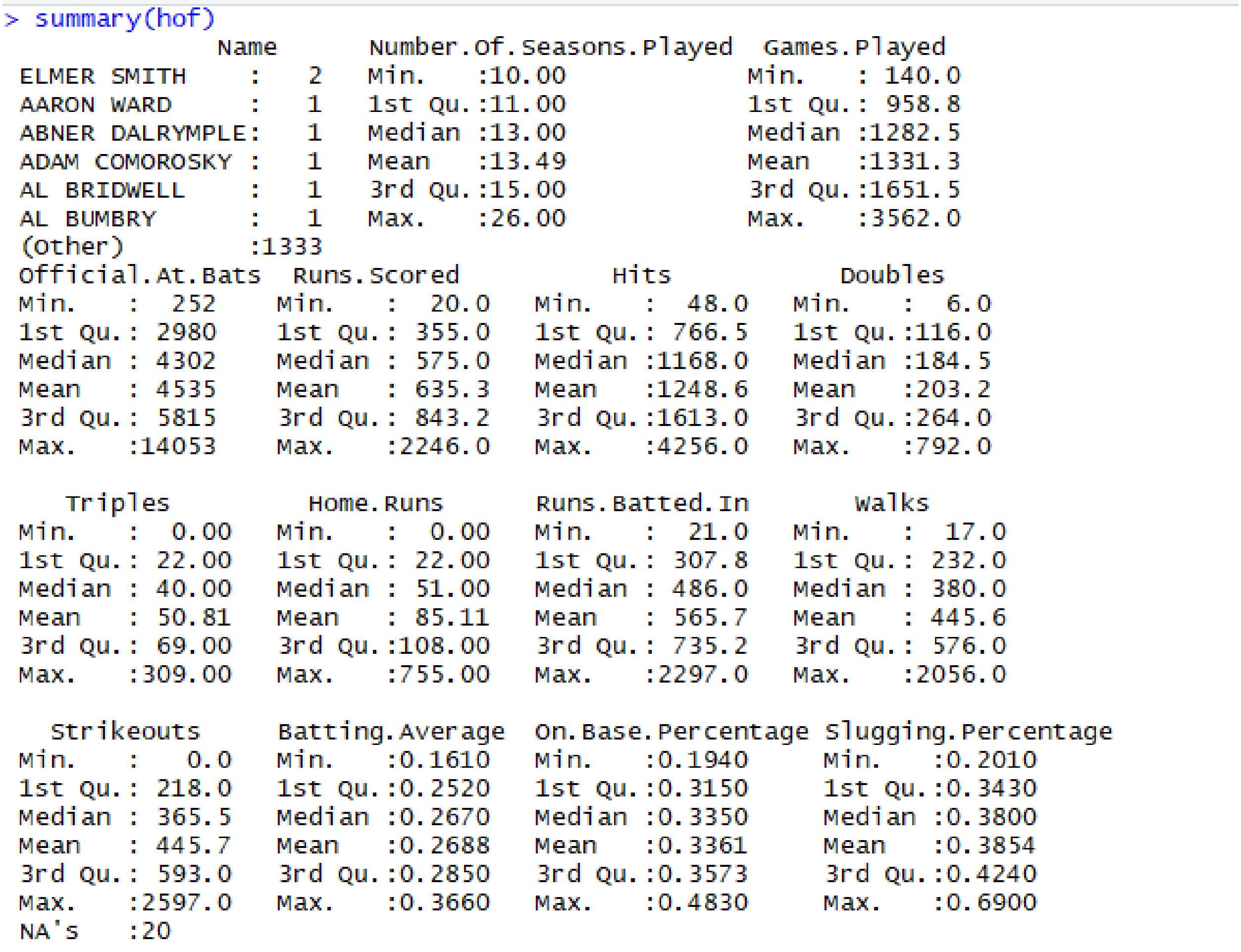
Baseball Reference. (2009). The Baseball-Reference.com Bullpen Hall of Fame. (Retrieved July, 2009 from Baseball-Reference: http://www.baseballreference .com/bullpen/Hall_of_Fame).

Appendix

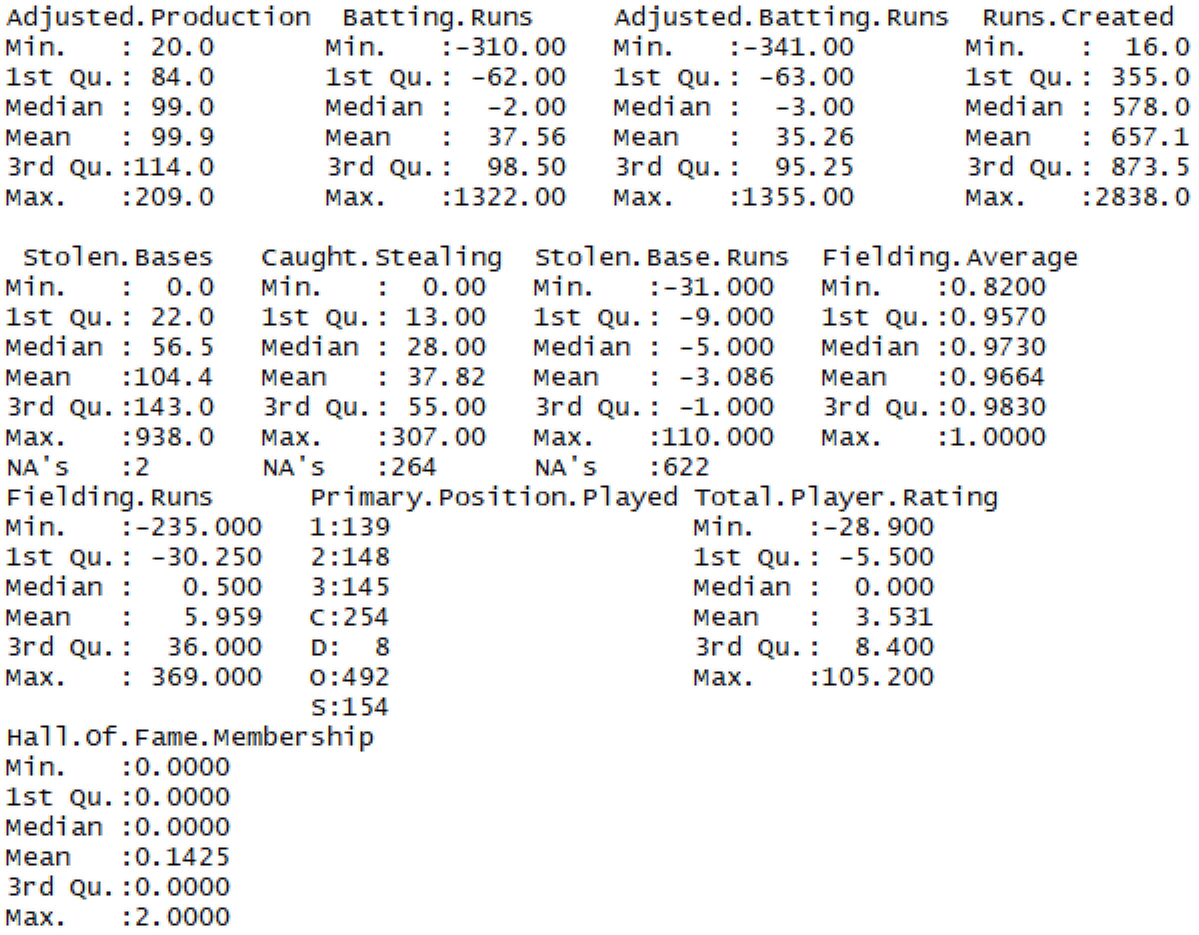


```
> head(hof)
          Name Number.Of.Seasons.Played Games.Played Official.At.Bats
1   HANK AARON                       23         3298            12364
2  JERRY ADAIR                       13         1165             4019
3 SPARKY ADAMS                       13         1424             5557
4  BOBBY ADAMS                       14         1281             4019
5   JOE ADCOCK                       17         1959             6606
6  TOMMIE AGEE                       12         1129             3912
  Runs.Scored Hits Doubles Triples Home.Runs Runs.Batted.In Walks Strikeouts
1        2174 3771     624      98       755           2297  1402       1383
2         378 1022     163      19        57            366   208        499
3         844 1588     249      48         9            394   453        223
4         591 1082     188      49        37            303   414        447
5         823 1832     295      35       336           1122   594       1059
6         558  999     170      27       130            433   342        918
  Batting.Average On.Base.Percentage Slugging.Percentage Adjusted.Production
1           0.305              0.377               0.555                 156
2           0.254              0.294               0.347                  80
3           0.286              0.343               0.353                  82
4           0.269              0.340               0.368                  90
5           0.277              0.339               0.485                 125
6           0.255              0.321               0.412                 108
  Batting.Runs Adjusted.Batting.Runs Runs.Created Stolen.Bases Caught.Stealing
1          878                   902         2550          240              73
2         -117                  -113          376           29              29
3         -121                  -133          679          154              50
4          -54                   -54          523           67              30
5          158                   198         1033           20              25
6           26                    34          507          167              81
  Stolen.Base.Runs Fielding.Average Fielding.Runs Primary.Position.Played
1               28            0.980            54                       0
2               -9            0.985           -30                       2
3               NA            0.974           -37                       2
4               NA            0.955            10                       3
5               NA            0.994           -28                       1
6                2            0.975            19                       0
  Total.Player.Rating Hall.Of.Fame.Membership
1                84.6                       1
2                -6.6                       0
```

Figure 1: Data Preview

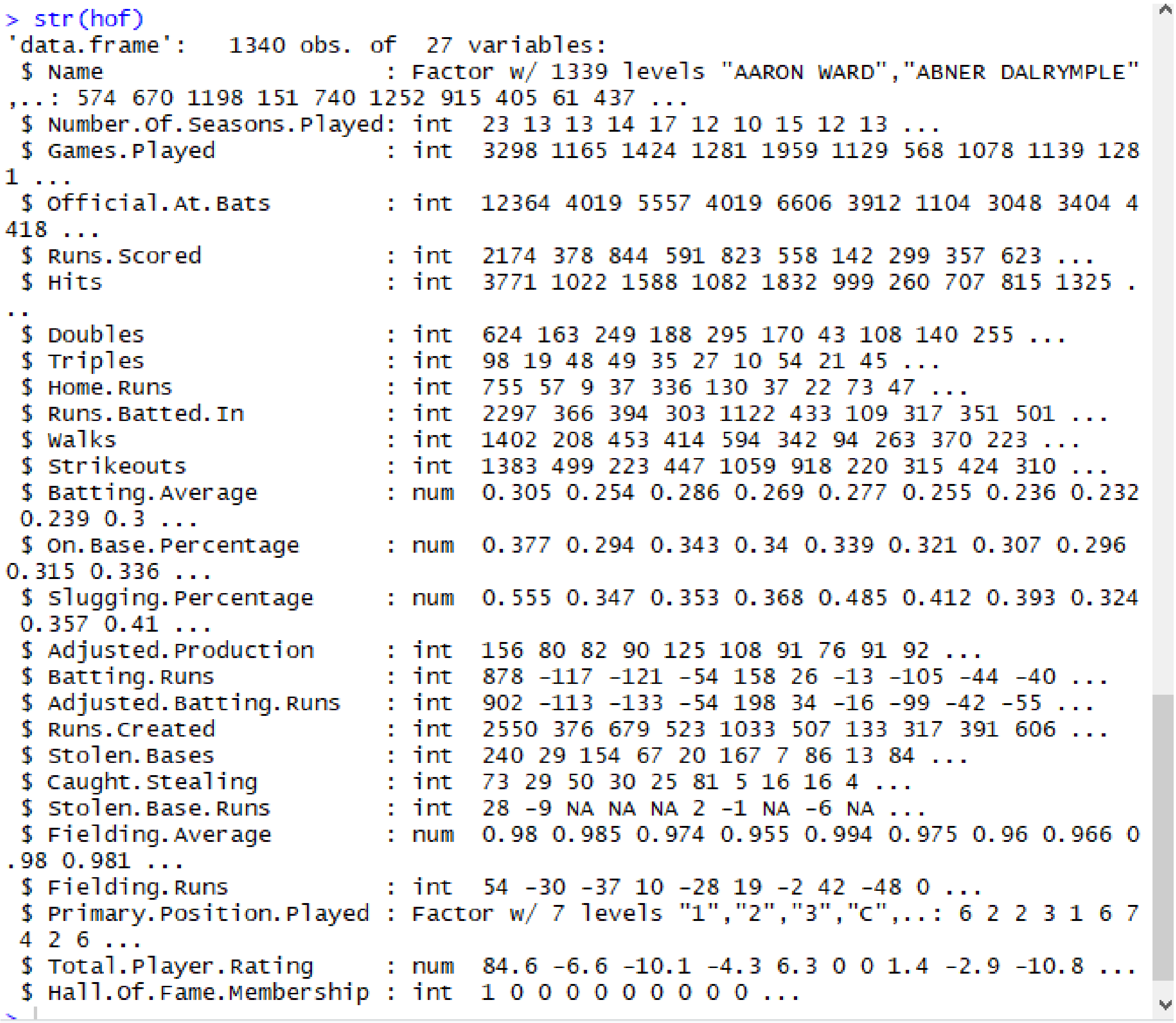


```
> summary(hof)
              Name       Number.Of.Seasons.Played  Games.Played   
 ELMER SMITH    :   2   Min.   :10.00             Min.   : 140.0  
 AARON WARD     :   1   1st Qu.:11.00             1st Qu.: 958.8  
 ABNER DALRYMPLE:   1   Median :13.00             Median :1282.5  
 ADAM COMOROSKY :   1   Mean   :13.49             Mean   :1331.3  
 AL BRIDWELL    :   1   3rd Qu.:15.00             3rd Qu.:1651.5  
 AL BUMBRY      :   1   Max.   :26.00             Max.   :3562.0  
 (Other)        :1333                                             
 Official.At.Bats  Runs.Scored          Hits           Doubles     
 Min.   :  252    Min.   :  20.0   Min.   :  48.0   Min.   :  6.0  
 1st Qu.: 2980    1st Qu.: 355.0   1st Qu.: 766.5   1st Qu.:116.0  
 Median : 4302    Median : 575.0   Median :1168.0   Median :184.5  
 Mean   : 4535    Mean   : 635.3   Mean   :1248.6   Mean   :203.2  
 3rd Qu.: 5815    3rd Qu.: 843.2   3rd Qu.:1613.0   3rd Qu.:264.0  
 Max.   :14053    Max.   :2246.0   Max.   :4256.0   Max.   :792.0  
                                                                   
    Triples          Home.Runs       Runs.Batted.In       Walks       
 Min.   :  0.00   Min.   :  0.00   Min.   :  21.0   Min.   :  17.0  
 1st Qu.: 22.00   1st Qu.: 22.00   1st Qu.: 307.8   1st Qu.: 232.0  
 Median : 40.00   Median : 51.00   Median : 486.0   Median : 380.0  
 Mean   : 50.81   Mean   : 85.11   Mean   : 565.7   Mean   : 445.6  
 3rd Qu.: 69.00   3rd Qu.:108.00   3rd Qu.: 735.2   3rd Qu.: 576.0  
 Max.   :309.00   Max.   :755.00   Max.   :2297.0   Max.   :2056.0  
                                                                    
   Strikeouts     Batting.Average  On.Base.Percentage Slugging.Percentage
 Min.   :   0.0   Min.   :0.1610   Min.   :0.1940     Min.   :0.2010     
 1st Qu.: 218.0   1st Qu.:0.2520   1st Qu.:0.3150     1st Qu.:0.3430     
 Median : 365.5   Median :0.2670   Median :0.3350     Median :0.3800     
 Mean   : 445.7   Mean   :0.2688   Mean   :0.3361     Mean   :0.3854     
 3rd Qu.: 593.0   3rd Qu.:0.2850   3rd Qu.:0.3573     3rd Qu.:0.4240     
 Max.   :2597.0   Max.   :0.3660   Max.   :0.4830     Max.   :0.6900     
 NA's   :20                                                              
```



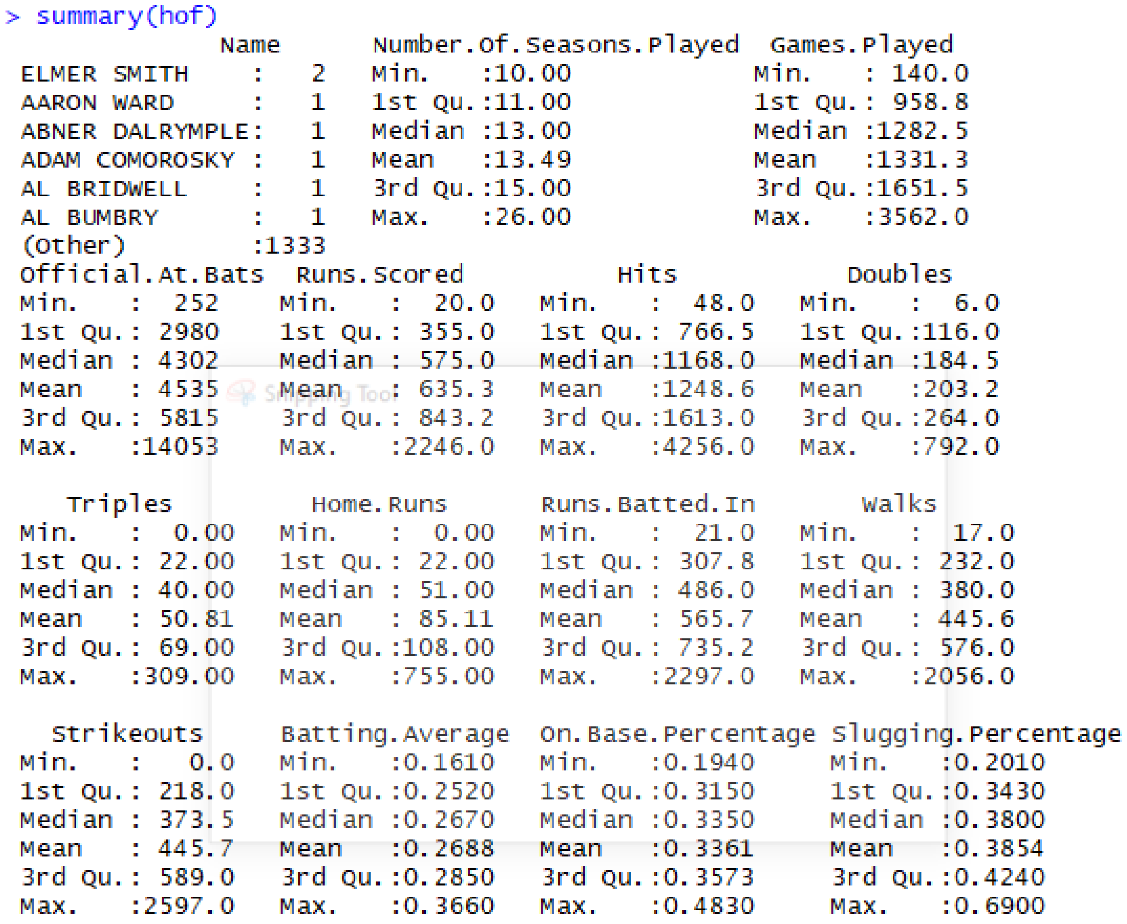
```
Adjusted.Production  Batting.Runs     Adjusted.Batting.Runs  Runs.Created
Min.   : 20.0       Min.   :-310.00   Min.   :-341.00       Min.   :  16.0
1st Qu.: 84.0       1st Qu.: -62.00   1st Qu.: -63.00       1st Qu.: 355.0
Median : 99.0       Median :  -2.00   Median :  -3.00       Median : 578.0
Mean   : 99.9       Mean   :  37.56   Mean   :  35.26       Mean   : 657.1
3rd Qu.:114.0       3rd Qu.:  98.50   3rd Qu.:  95.25       3rd Qu.: 873.5
Max.   :209.0       Max.   :1322.00   Max.   :1355.00       Max.   :2838.0

 Stolen.Bases   Caught.Stealing  Stolen.Base.Runs  Fielding.Average
Min.   :  0.0   Min.   :  0.00   Min.   :-31.000   Min.   :0.8200
1st Qu.: 22.0   1st Qu.: 13.00   1st Qu.: -9.000   1st Qu.:0.9570
Median : 56.5   Median : 28.00   Median : -5.000   Median :0.9730
Mean   :104.4   Mean   : 37.82   Mean   : -3.086   Mean   :0.9664
3rd Qu.:143.0   3rd Qu.: 55.00   3rd Qu.: -1.000   3rd Qu.:0.9830
Max.   :938.0   Max.   :307.00   Max.   :110.000   Max.   :1.0000
NA's   :2       NA's   :264      NA's   :622
Fielding.Runs      Primary.Position.Played Total.Player.Rating
Min.   :-235.000   1:139                   Min.   :-28.900
1st Qu.: -30.250   2:148                   1st Qu.: -5.500
Median :   0.500   3:145                   Median :  0.000
Mean   :   5.959   C:254                   Mean   :  3.531
3rd Qu.:  36.000   D:  8                   3rd Qu.:  8.400
Max.   : 369.000   O:492                   Max.   :105.200
                   S:154
Hall.Of.Fame.Membership
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.1425
3rd Qu.:0.0000
Max.   :2.0000
```

Figure 2: Dataset Summary

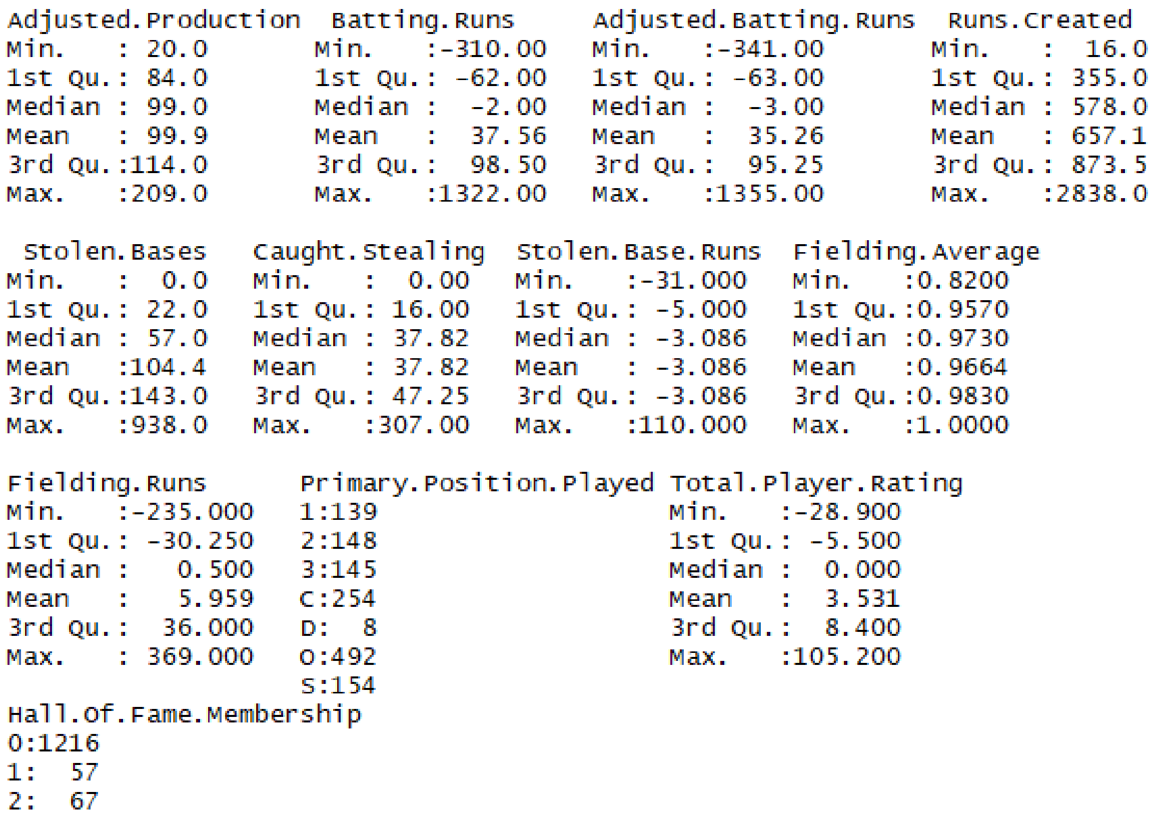


```
> str(hof)
'data.frame':   1340 obs. of  27 variables:
 $ Name                   : Factor w/ 1339 levels "AARON WARD","ABNER DALRYMPLE"
,..: 574 670 1198 151 740 1252 915 405 61 437 ...
 $ Number.Of.Seasons.Played: int  23 13 13 14 17 12 10 15 12 13 ...
 $ Games.Played           : int  3298 1165 1424 1281 1959 1129 568 1078 1139 128
1 ...
 $ Official.At.Bats       : int  12364 4019 5557 4019 6606 3912 1104 3048 3404 4
418 ...
 $ Runs.Scored            : int  2174 378 844 591 823 558 142 299 357 623 ...
 $ Hits                   : int  3771 1022 1588 1082 1832 999 260 707 815 1325 .
..
 $ Doubles                : int  624 163 249 188 295 170 43 108 140 255 ...
 $ Triples                : int  98 19 48 49 35 27 10 54 21 45 ...
 $ Home.Runs              : int  755 57 9 37 336 130 37 22 73 47 ...
 $ Runs.Batted.In         : int  2297 366 394 303 1122 433 109 317 351 501 ...
 $ Walks                  : int  1402 208 453 414 594 342 94 263 370 223 ...
 $ Strikeouts             : int  1383 499 223 447 1059 918 220 315 424 310 ...
 $ Batting.Average        : num  0.305 0.254 0.286 0.269 0.277 0.255 0.236 0.232
 0.239 0.3 ...
 $ On.Base.Percentage     : num  0.377 0.294 0.343 0.34 0.339 0.321 0.307 0.296
0.315 0.336 ...
 $ Slugging.Percentage    : num  0.555 0.347 0.353 0.368 0.485 0.412 0.393 0.324
 0.357 0.41 ...
 $ Adjusted.Production    : int  156 80 82 90 125 108 91 76 91 92 ...
 $ Batting.Runs           : int  878 -117 -121 -54 158 26 -13 -105 -44 -40 ...
 $ Adjusted.Batting.Runs  : int  902 -113 -133 -54 198 34 -16 -99 -42 -55 ...
 $ Runs.Created           : int  2550 376 679 523 1033 507 133 317 391 606 ...
 $ Stolen.Bases           : int  240 29 154 67 20 167 7 86 13 84 ...
 $ Caught.Stealing        : int  73 29 50 30 25 81 5 16 16 4 ...
 $ Stolen.Base.Runs       : int  28 -9 NA NA NA 2 -1 NA -6 NA ...
 $ Fielding.Average       : num  0.98 0.985 0.974 0.955 0.994 0.975 0.96 0.966 0
.98 0.981 ...
 $ Fielding.Runs          : int  54 -30 -37 10 -28 19 -2 42 -48 0 ...
 $ Primary.Position.Played : Factor w/ 7 levels "1","2","3","C",..: 6 2 2 3 1 6 7
 4 2 6 ...
 $ Total.Player.Rating    : num  84.6 -6.6 -10.1 -4.3 6.3 0 0 1.4 -2.9 -10.8 ...
 $ Hall.of.Fame.Membership : int  1 0 0 0 0 0 0 0 0 0 ...
```

Figure 3: Data Structure

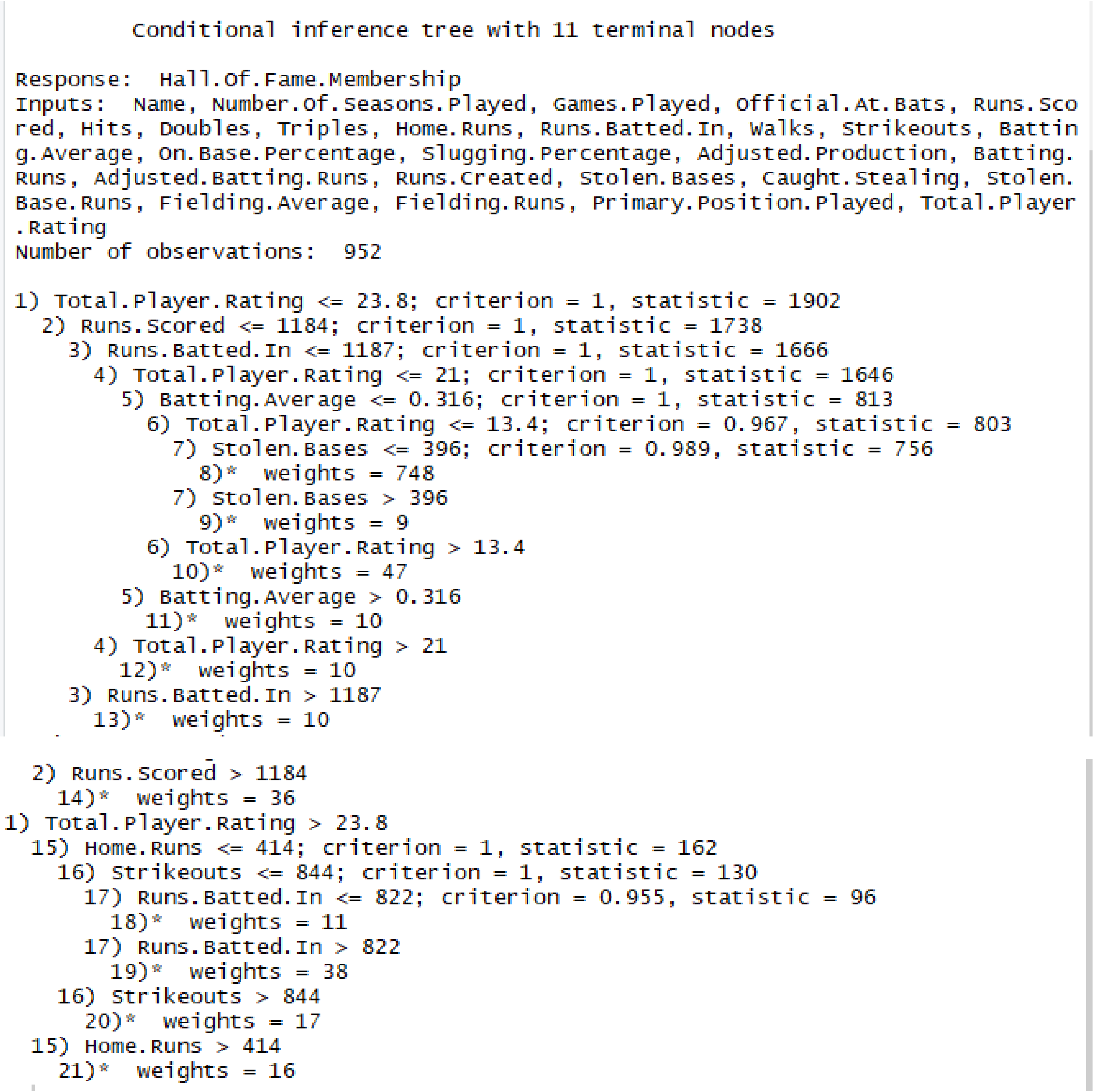


```
> summary(hof)
              Name      Number.Of.Seasons.Played  Games.Played
 ELMER SMITH    :   2   Min.   :10.00            Min.   : 140.0
 AARON WARD     :   1   1st Qu.:11.00            1st Qu.: 958.8
 ABNER DALRYMPLE:   1   Median :13.00            Median :1282.5
 ADAM COMOROSKY :   1   Mean   :13.49            Mean   :1331.3
 AL BRIDWELL    :   1   3rd Qu.:15.00            3rd Qu.:1651.5
 AL BUMBRY      :   1   Max.   :26.00            Max.   :3562.0
 (Other)        :1333
 Official.At.Bats  Runs.Scored          Hits           Doubles
 Min.   :  252    Min.   :  20.0   Min.   :  48.0   Min.   :  6.0
 1st Qu.: 2980    1st Qu.: 355.0   1st Qu.: 766.5   1st Qu.:116.0
 Median : 4302    Median : 575.0   Median :1168.0   Median :184.5
 Mean   : 4535    Mean   : 635.3   Mean   :1248.6   Mean   :203.2
 3rd Qu.: 5815    3rd Qu.: 843.2   3rd Qu.:1613.0   3rd Qu.:264.0
 Max.   :14053    Max.   :2246.0   Max.   :4256.0   Max.   :792.0

    Triples          Home.Runs      Runs.Batted.In        Walks
 Min.   :  0.00   Min.   :  0.00   Min.   :  21.0   Min.   :  17.0
 1st Qu.: 22.00   1st Qu.: 22.00   1st Qu.: 307.8   1st Qu.: 232.0
 Median : 40.00   Median : 51.00   Median : 486.0   Median : 380.0
 Mean   : 50.81   Mean   : 85.11   Mean   : 565.7   Mean   : 445.6
 3rd Qu.: 69.00   3rd Qu.:108.00   3rd Qu.: 735.2   3rd Qu.: 576.0
 Max.   :309.00   Max.   :755.00   Max.   :2297.0   Max.   :2056.0

   Strikeouts     Batting.Average  On.Base.Percentage Slugging.Percentage
 Min.   :   0.0   Min.   :0.1610   Min.   :0.1940     Min.   :0.2010
 1st Qu.: 218.0   1st Qu.:0.2520   1st Qu.:0.3150     1st Qu.:0.3430
 Median : 373.5   Median :0.2670   Median :0.3350     Median :0.3800
 Mean   : 445.7   Mean   :0.2688   Mean   :0.3361     Mean   :0.3854
 3rd Qu.: 589.0   3rd Qu.:0.2850   3rd Qu.:0.3573     3rd Qu.:0.4240
 Max.   :2597.0   Max.   :0.3660   Max.   :0.4830     Max.   :0.6900
```



```
Adjusted.Production  Batting.Runs      Adjusted.Batting.Runs  Runs.Created
Min.    : 20.0       Min.    :-310.00  Min.    :-341.00       Min.    :  16.0
1st Qu.: 84.0        1st Qu.: -62.00   1st Qu.: -63.00        1st Qu.: 355.0
Median : 99.0        Median :  -2.00   Median :  -3.00        Median : 578.0
Mean   : 99.9        Mean   :  37.56   Mean   :  35.26        Mean   : 657.1
3rd Qu.:114.0        3rd Qu.:  98.50   3rd Qu.:  95.25        3rd Qu.: 873.5
Max.   :209.0        Max.    :1322.00  Max.    :1355.00       Max.    :2838.0

 Stolen.Bases    Caught.Stealing   Stolen.Base.Runs   Fielding.Average
Min.   :  0.0   Min.   :  0.00    Min.    :-31.000   Min.   :0.8200
1st Qu.: 22.0   1st Qu.: 16.00    1st Qu.: -5.000    1st Qu.:0.9570
Median : 57.0   Median : 37.82    Median : -3.086    Median :0.9730
Mean   :104.4   Mean   : 37.82    Mean   : -3.086    Mean   :0.9664
3rd Qu.:143.0   3rd Qu.: 47.25    3rd Qu.: -3.086    3rd Qu.:0.9830
Max.   :938.0   Max.   :307.00    Max.   :110.000    Max.   :1.0000

Fielding.Runs       Primary.Position.Played Total.Player.Rating
Min.   :-235.000    1:139                   Min.   :-28.900
1st Qu.: -30.250    2:148                   1st Qu.: -5.500
Median :   0.500    3:145                   Median :  0.000
Mean   :   5.959    C:254                   Mean   :  3.531
3rd Qu.:  36.000    D:  8                   3rd Qu.:  8.400
Max.   : 369.000    O:492                   Max.   :105.200
                    S:154
Hall.Of.Fame.Membership
0:1216
1:  57
2:  67
```

Figure 4: Dataset Summary After Transformation



```
         Conditional inference tree with 11 terminal nodes

Response:  Hall.Of.Fame.Membership
Inputs:  Name, Number.of.Seasons.Played, Games.Played, Official.At.Bats, Runs.Sco
red, Hits, Doubles, Triples, Home.Runs, Runs.Batted.In, Walks, Strikeouts, Battin
g.Average, On.Base.Percentage, Slugging.Percentage, Adjusted.Production, Batting.
Runs, Adjusted.Batting.Runs, Runs.Created, Stolen.Bases, Caught.Stealing, Stolen.
Base.Runs, Fielding.Average, Fielding.Runs, Primary.Position.Played, Total.Player
.Rating
Number of observations:  952

1) Total.Player.Rating <= 23.8; criterion = 1, statistic = 1902
  2) Runs.Scored <= 1184; criterion = 1, statistic = 1738
    3) Runs.Batted.In <= 1187; criterion = 1, statistic = 1666
      4) Total.Player.Rating <= 21; criterion = 1, statistic = 1646
        5) Batting.Average <= 0.316; criterion = 1, statistic = 813
          6) Total.Player.Rating <= 13.4; criterion = 0.967, statistic = 803
            7) Stolen.Bases <= 396; criterion = 0.989, statistic = 756
              8)*  weights = 748
            7) Stolen.Bases > 396
              9)*  weights = 9
          6) Total.Player.Rating > 13.4
            10)*  weights = 47
        5) Batting.Average > 0.316
          11)*  weights = 10
      4) Total.Player.Rating > 21
        12)*  weights = 10
    3) Runs.Batted.In > 1187
      13)*  weights = 10
  2) Runs.Scored > 1184
    14)*  weights = 36
1) Total.Player.Rating > 23.8
  15) Home.Runs <= 414; criterion = 1, statistic = 162
    16) Strikeouts <= 844; criterion = 1, statistic = 130
      17) Runs.Batted.In <= 822; criterion = 0.955, statistic = 96
        18)*  weights = 11
      17) Runs.Batted.In > 822
        19)*  weights = 38
    16) Strikeouts > 844
      20)*  weights = 17
  15) Home.Runs > 414
    21)*  weights = 16
```

Figure 5: ctree Method Results

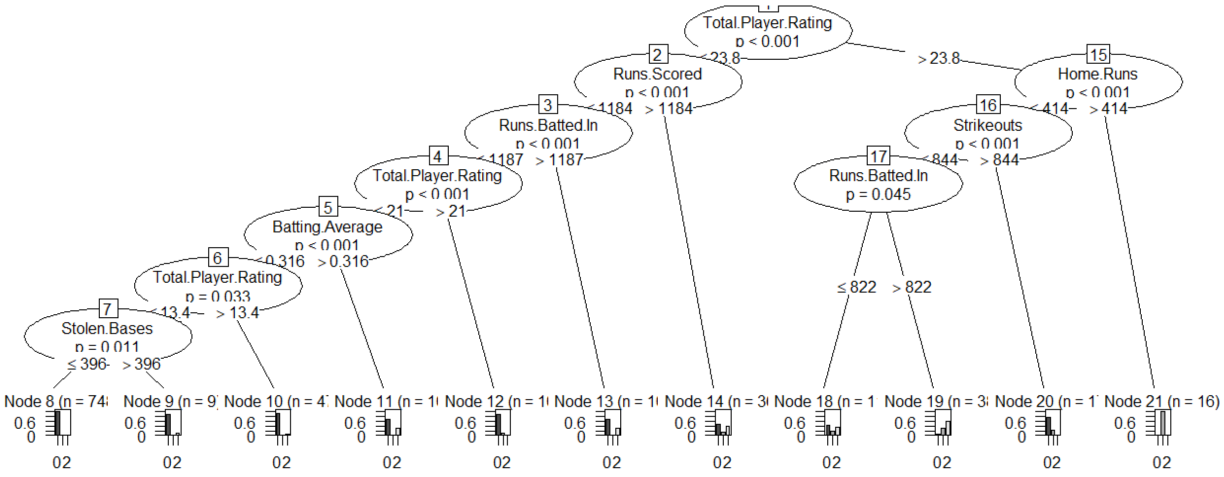
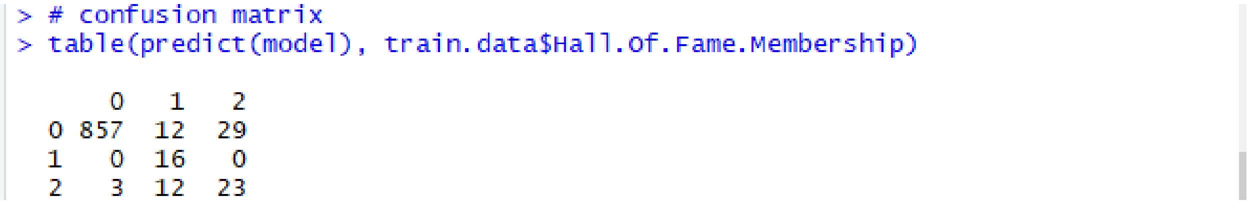
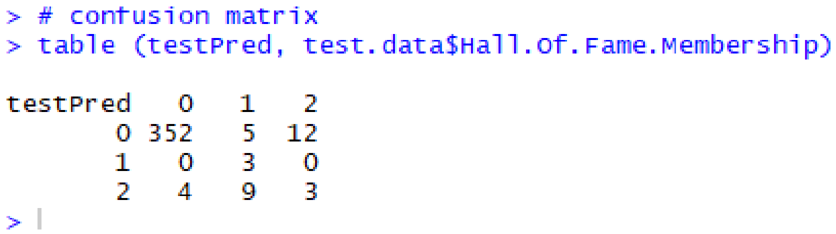


Figure 6: Decision Tree Plot



```
> # confusion matrix
> table(predict(model), train.data$Hall.of.Fame.Membership)

      0   1   2
  0 857  12  29
  1   0  16   0
  2   3  12  23
```

Figure 7: ctree Model Confusion Matrix of Training Set



```
> # confusion matrix
> table (testPred, test.data$Hall.of.Fame.Membership)

testPred   0   1   2
       0 352   5  12
       1   0   3   0
       2   4   9   3
> 
```

Figure 8: ctree Model Confusion Matrix of Test Set