From the Ballpark to the Hall of Fame: A Classification Analysis

-Shrinidhi Rajesh

Abstract:

Imagine walking through the hallowed doors of the Baseball Hall of Fame, and marveling at the plaques of legendary players like Babe Ruth, Jackie Robinson, and Hank Aaron. The honor of being inducted into the Hall of Fame is the pinnacle of achievement for any baseball player, but attaining this status is no small feat. The rigorous standards for induction are especially demanding for non-pitchers, requiring not only exceptional performance on the field but also recognition from a panel of experts. Previous research has shown that the selection process for the Baseball Hall of Fame is subjective and often influenced by personal biases and narratives. In recent years, there has been a growing interest in utilizing statistical analysis to identify potential Hall of Fame candidates objectively. This approach has the potential to make the selection process more transparent and provide a data-driven approach to identifying deserving candidates. This study focuses on one such approach known as classification analysis by utilizing logistic regression to make predictions on whether the players will make the Hall of Fame or not. The model uses every non-pitcher who has ever received a vote for Major League Baseball's Hall of Fame with 627 players and 19 different factors associated to make the induction providing a robust foundation for analysis. This classification model is further applied to the dataset of potential candidates for the 2024 Hall of Fame, to predict who may be joining the ranks of baseball's immortals. The classification suggests that there are three players among the list of 12 potential players who have higher chances of making the Hall of Fame.

Table of Contents:

Purpose of the Study:

The aim of this study is to perform classification analysis using logistic regression to accurately predict potential Hall of Fame candidates of the year 2024 in baseball. To achieve this, the study uses a training dataset of 627 non-pitchers who have previously received votes for the Hall of Fame, along with 19 different factors that are best associated with the reason for their induction. This achieved classification model is applied to the potential candidates to classify the players into Hall of fame. By doing so, this study aims to provide an objective, data-driven approach to Hall of Fame selection and contribute to the ongoing conversation around the subjective nature of the selection process. The purpose of this study further expands to:

Firstly, it can serve as a benchmark for players to strive towards, as they can see what qualities and achievements are necessary to be considered for induction into the Hall of Fame. This can motivate them to work harder and focus on improving specific aspects of their game.

Secondly, it can help teams in their player scouting and recruitment process. By analyzing the factors that contribute to Hall of Fame candidacy, teams can identify potential star players early on and make informed decisions on who to draft or sign as free agents.

Finally, it can help teams and fans appreciate the qualities that make a player truly great, beyond just their statistics. By understanding the factors that contribute to Hall of Fame candidacy, fans can gain a deeper appreciation for the game and its history, and teams can build a stronger culture around the values and traits that lead to success on and off the field.

Background of the Study:

The Baseball Hall of Fame has been honoring the greatest players, managers, umpires, and executives who have made significant contributions to the sport of baseball since 1936. The rules for induction into the Baseball Hall of Fame have evolved over time to reflect changes in the sport and its culture. The current election process for the Baseball Hall of Fame is conducted by the Baseball Writers' Association of America (BBWAA), and players must receive at least 75% of the votes to be inducted.

The selection process for the Baseball Hall of Fame has been a topic of debate and scrutiny over the years, with many questioning the subjective nature of the criteria used to determine who gets inducted. While players must meet certain statistical benchmarks there are other factors at play, including personal biases, narratives, and intangibles like leadership and sportsmanship. This has led to controversies over the years, with many deserving players overlooked or undervalued, while others with less impressive resumes are inducted.

To address this issue, there has been a growing interest in utilizing statistical analysis to identify potential Hall of Fame candidates objectively. This data-driven approach aims to remove personal biases and narratives from the selection process and provide a more transparent and objective evaluation of a player's contributions to the sport. This approach uses various statistical measures, such as Wins Above Replacement (WAR) and other advanced metrics, to evaluate a player's performance and value relative to their peers.

To conduct this study, a sample of every non-pitcher who has ever received a vote for Major League Baseball's Hall of Fame was compiled. This dataset includes 627 players and their respective factors critical for the Writers to decide who gets into the Hall of Fame. This includes players who were eventually inducted into the Hall of Fame as well as those who were not. These factors include traditional measures like batting average, home runs, and RBIs, as well as advanced metrics like WAR, adjusted OPS+, and other contextual statistics. This dataset provides a robust foundation for analysis and allows for a comprehensive evaluation of a player's career.

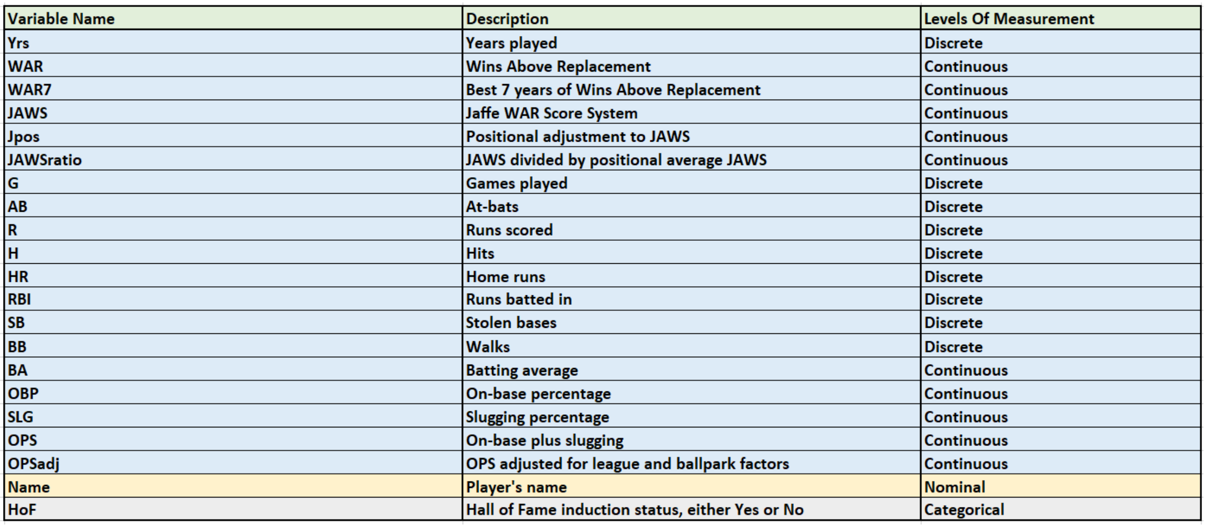
One possibility of the population of interest in this study could be all non-pitchers who have played in Major League Baseball Hall of Fame. This population would include not only the individuals in the sample dataset who have received votes for the Hall of Fame, but also all other non-pitchers who have been eligible for consideration for induction into the Hall of Fame but have not yet received a vote or have been unsuccessful in receiving enough votes for induction. It is important to note that this is only one possible prediction for the population based on the given sample. Other possible population may be all the non-pitchers who have every played in Major League Baseball. If this is the population of interest then the sample size is very small compared to all the non pitchers ever played in the history of baseball.

Variables Used In the Study:

The dataset for the study comprises of 627 players comprising of every non-pitcher who has ever received a vote for Major League Baseball's Hall of Fame and 19 different factors associated for the induction of players into HALL of FAME

- Name: Name of the baseball players.
- HoF: This is a “Yes"/”No" indicator on whether or not the player is in the Hall of Fame.
- Yrs: How many seasons did the player play in Major League Baseball.
- WAR: Baseball References measure of Wins Above Replacement. This is a single number that describes the number of wins the player added to their teams over the course of their career.
- WAR7: The sum of the seven best seasons of WAR in the player's career. It may not be seven seasons in a row.
- JAWS: Developed by Baseball Prospectus. It contains a combination of career and 7-year peak WAR totals allowing for comparison to average Hall of Fame players by position.
- Jpos: The average JAWS score for all Hall of Fame players at this position plus overall Hall of Fame averages for positions with fewer inducted players.
- JAWSRatio: JAWS/JPOS.
- G: games played during a player's career.
- AB: at bats during a player's career.
- R: runs scored during a player's career.
- H: hits during a player's career.
- HR: home runs during a player's career.
- RBI: runs batted during during a player's career.
- SB: stolen bases during a player's career.
- BB: walks during a player's career.
- BA: batting average. This is the number of hits divided by the number
- of at bats.
- OBP: on base percentage. This is the sum of the number of hits, walks and times hit by a pitch divided by the sum of the number of at bats, walks, times hit by a pitch, and sacrifice flies.
- SLG: slugging percentage. This is the number of bases divided by the number of at bats. Every single is one base, double is two bases, triple is three bases, and home run is four bases in the numerator of this
  calculation.
- OPS: on base percentage plus slugging percentage.
- OPS+: OPS adjusted to the player's ball park. 100 is an average hitter.

  Below is a detailed description of the column variables and their levels of Measurement [Table 1]



Variable Name	Description	Levels Of Measurement
Yrs	Years played	Discrete
WAR	Wins Above Replacement	Continuous
WAR7	Best 7 years of Wins Above Replacement	Continuous
JAWS	Jaffe WAR Score System	Continuous
Jpos	Positional adjustment to JAWS	Continuous
JAWSratio	JAWS divided by positional average JAWS	Continuous
G	Games played	Discrete
AB	At-bats	Discrete
R	Runs scored	Discrete
H	Hits	Discrete
HR	Home runs	Discrete
RBI	Runs batted in	Discrete
SB	Stolen bases	Discrete
BB	Walks	Discrete
BA	Batting average	Continuous
OBP	On-base percentage	Continuous
SLG	Slugging percentage	Continuous
OPS	On-base plus slugging	Continuous
OPSadj	OPS adjusted for league and ballpark factors	Continuous
Name	Player's name	Nominal
HoF	Hall of Fame induction status, either Yes or No	Categorical

Table1: Variables and their measurement

Methodology:

The data obtained with 627 players was loaded into R Software and missing values were checked, the result suggested that there were no missing values in this dataset.

The main objective of this study was to construct a classification rule that make predictions on the chances of the 12 potential players of 2024 making the Hall Of Fame. To do this, I first tried to check for the assumptions of multivariate normality [Table 2] and homogeneity of covariance matrices[Table 5] to either proceed with discriminant analysis or Logistic Regression. The assumption of multivariate normality was tested using the Mardia test. Multivariate normality was tested for the whole data as well as by categorizing the data based on the HOF Status [Table 3, Table 4]. The test was performed for a total of three times and all three times the assumption of Multivariate Normality failed except for few Univariate normality passing the test. I proceeded to test for the assumption of homogeneity of covariance matrices using the Box's M test, which showed that the assumption of homogeneity failed for the data, indicating unequal variances. With both the assumptions failing, the only possibility to perform classification analysis was by performing Logistic Regression. As per the rule for performing Logistic Regression the response variable in this study is Binomial(HoF).

Before Logistic Regression was performed on the sample dataset, two sample ttest was performed on each of the 19 factors excluding Name, as it is a nominal data and doesn't provide any meaningful results when used. Amongst the 19 predictors, performing two sample ttest based on HoF, all the variables showed significant difference between the means of the two groups, except for the JPOS variable[Table 6]. Thus, the model of Logistic Regression was constructed using the rest of the 18 independent variables except for JPOS[Table 7].

Additionally, the variable selection methods were used on Logistic regression model to identify the best model. This was done with the help of comparing the AICs of the 4 models- Stepwise, Forward, Backward and the complete Logistic Regression Models[Table 8]. Upon comparing the AIC, the Backward and Stepwise selection method resulted in the same AIC values. Since the backward selection model is a more conservative method that involves starting with a full model and then iteratively removing variables, I preferred to go on with the backward selection model[Table 9]. Likelihood Ratio Test was performed on the backward selection logistic model to check for model fit[Table 10]. With 8 most important variables discriminating between the Hall Of Fame status, evaluation of classification rule was performed using Substitution Method [Table 11]. Once the evaluation was performed the results were graphically depicted for better understanding

[Table 12].

Furthermore, this model was used to predict the chances of new players entering the Hall Of Fame. The threshold for classification was 50%, indicating any probability greater than 0.5 suggests that the players will be inducted into Hall of Fame,2024 and any probability lesser than 0.5 says otherwise.

Lastly goodness of fit test was conducted on the Logistic Regression model using Hosmer Lemeshow Test[Table 13].

Discussion and Results:

The Logistic Regression performed using the 19 factors influencing the induction of players into the Hall Of Fame resulted in an AIC value of 381.12. Variable selection methods were performed on the Logistic Model, and the backward selection model was used for further analysis because it had the lowest AIC. Out of the 18 predictors, 8 significant predictors were obtained using the backward selection process. The Likelihood Ratio Test was conducted to check for model fit, and the p-value obtained was below the significant level of 0.05, indicating that the model is a good fit. [Table 10]

MODEL:

log(HoF)= -19.1134 + 0.1811 (Yrs)+0.0697 (JAWSratio) -0.00390 (G)+ 0.000988 (AB) -0.00575 (HR)+ 0.00280 (RBI)-0.00304 (BB)+ 31.3080 (OBP)

The backward selection process resulted in 8 significant predictors: **Yrs, JAWSRatio, G, AB, HR, RBI, BB and OBP.**

For increase in years played by each year the log odds of players being inducted into hall of fame increases by **1.198535** holding all other variables constant, the log odds of players being inducted into hall of fame is increasing at the rate of 119.9% for each year played additionally holding all other variables constant.

For increase in the JAWSratio by one unit the log odds of players being inducted into hall of fame increases by **1.072186** holding all other variables constant, the log odds of players being inducted into hall of fame is increasing at the rate of 107% for each unit increase in the JAWSRatio holding all other variables constant.

For increase in G by one unit the log odds of players being inducted into hall of fame decreases by **1.003908** holding all other variables constant, the log odds of players being inducted into hall of fame is decreasing at the rate of 103.9% for each unit increase in G holding all other variables constant.

For increase in the AB by one unit the log odds of players being inducted into hall of fame increases by **1.000988** holding all other variables constant, the log odds of players being inducted into hall of fame is increasing at the rate of 100% for each unit increase in the AB holding all other variables constant.

For increase in the HR by one unit the log odds of players being inducted into hall of fame decreases by **1.005767** holding all other variables constant, the log odds of players being inducted into hall of fame is decreasing at the rate of 100.58% for each unit increase in the HR holding all other variables constant.

For increase in the RBI by one unit the log odds of players being inducted into hall of fame increases by **1.002804** holding all other variables constant, the log odds of players being inducted into hall of fame is increasing at the rate of 100.28% for each unit increase in the RBI holding all other variables constant.

For increase in the BB by one unit the log odds of players being inducted into hall of fame decreases by **1.003045** holding all other variables constant, the log odds of players being inducted into hall of fame is decreasing at the rate of 100.3% for each unit increase in the BB holding all other variables constant.
For increase in the OBP by one unit the log odds of players being inducted into hall of fame increases by **3.95268e+13** holding all other variables constant, the log odds of players being inducted into hall of fame is increasing at the rate of 3.95268e+15% for each unit increase in the OBP holding all other variables constant.

The evaluation of the Classification Rule was performed using Substitution Method[Table11]. Based on the substitution method results, the logistic regression model correctly classified 72.29% of the players who were inducted into the Hall of Fame and 95.44% of the players who were not inducted into the Hall of Fame, using 0.5 as the threshold for prediction.

With the prediction classification exceeding 70 % accuracy, the test data of 12 potential candiates for the 2024 Hall Of Fame were categorized based on the Logistic Regression Model Obtained. The results indicate that three of the 12 players have higher chances of entering the Hall of Fame. 0.959,0.844,0.587[Table 14]

Adrian Beltre(Predicted Probability= 0.959): Beltre was a third baseman who played in the Major Leagues for 21 seasons. He was known for his excellent defense and his ability to hit for both power and average. He finished his career with 3,166 hits, 477 home runs, and a .286 batting average. He won five Gold Glove awards and four Silver Slugger awards, and was named to four All-Star teams. Beltre is considered one of the greatest third basemen of all time, and his impressive statistics and longevity in the game make him a likely candidate for induction into the Hall of Fame.

Joe Mauer(Predicted Probability=0.844): Mauer was a catcher who played for the Minnesota Twins for 15 seasons. He was known for his excellent defensive skills and his ability to hit for a high average. He won three batting titles and was named to six All-Star teams. He finished his career with a .306 batting average, 2,123 hits, and 143 home runs. Mauer is considered one of the greatest catchers of his generation, and his impressive career statistics and accolades make him a likely candidate for induction into the Hall of Fame.

Chase Utley(Predicted Probability=0.587): Utley was a second baseman who played in the Major Leagues for 16 seasons. He was known for his excellent defense, his ability to hit for power and average, and his aggressive style of play. He finished his career with a .275 batting average, 1,885 hits, and 259 home runs. He won four Silver Slugger awards and was named to six All-Star teams. Utley is considered one of the greatest second basemen of his generation, and his impressive career statistics and reputation as a hard-nosed player make him a possible candidate for induction into the Hall of Fame.

The goodness of test performed using Hosmer-Lemeshow test. The resulted p value was greater than 0.05 indicating that the observed counts are what to be expected.

Summary Statistics:

The summary statistics were calculated for the sample dataset[Table 15, Table 16, Table 17]. In_HOF dataset contains players who were inducted into hall of fame with HoF = Yes. The NotIn_HOF dataset consist of players who were not inducted into the hall of fame with HoF=No. In the In_HOF data, the players have a minimum of 10 years of experience, while in the NotIn_HOF data, some players have as little as 4 years of experience. The mean and median WAR for players in the In_HOF is higher than for those in the NotIn_HOF, with a minimum WAR of 16.2 and -5.3, respectively. In the same way, the mean and median for WAR7 is also higher compared to that of NotIn_HOF group. The JAWS score, which combines a player's career WAR with their peak WAR, is also higher for the In_HOF group, with a mean of 54.14 compared to a mean of 27.45 for the NotIn_HOF group. Additionally, players in the In_HOF tend to have higher career totals for statistics such as runs, hits, home runs, RBIs, and stolen bases, as well as higher batting average, on-base percentage, slugging percentage, and OPS. Finally, the minimum JAWS ratio for players in the In_HOF is 30.77, while in the NotIn_HOF group, the minimum is -4.525. This suggests that the players in the In_HOF tend to have a higher ratio of JAWS score to the average JAWS score of players at their position than those who are not in the Hall of Fame.Overall, it appears that players who are in the Hall of Fame (In_HOF) have certain statistical advantages over those who are not (NotIn_HOF). These advantages include higher mean and median WAR, WAR7, JAWS score, career totals for various statistics, and higher batting averages, on-base percentages, slugging percentages, and OPS. Additionally, players in the In_HOF tend to have a higher ratio of JAWS score to the average JAWS score of players at their position, suggesting that they are more dominant within their respective positions than those who are not in the Hall of Fame.

Size: In_HOF contains data for 166 players who have been inducted into the Baseball Hall of Fame, while NotIn_HOF contains data for 461 players who have not been inducted.
Mean values: In_HOF generally has higher mean values for statistics such as WAR, WAR7, JAWS, JAWSratio, R, H, HR, RBI, SB, BB, BA, OBP, SLG, OPS, and OPSadj.
Range: The range of values for statistics is generally wider in the NotIn_HOF dataset.
Minimum values: In_HOF has minimum values for some statistics such as WAR, WAR7, and JAWS that are higher than the minimum values for these statistics in the NotIn_HOF dataset.

Summary statistics of the test data consisting of 12 potential players was performed by segregating the data into two groups, Predicted players with higher chances to be inducted into Hall of Fame and players who have lesser chances to be inducted [Table 18, Table 19]. The comparison of statistics between predicted Hall of Fame players and non-Hall of Fame players in the test data indicates that players who perform well across multiple metrics are more likely to be inducted into the Hall of Fame. The summary stats of players with high chances to be inducted consistently coincides with the training dataset. The three players who were predicted to enter the hall of fame have generally played longer, had higher WAR, JAWS, JAWS ratio, played more games, had more at-bats, scored more runs, had more hits, hit more home runs, had more RBIs, stolen more bases, and had more walks than non-Hall of Fame players. Additionally, Hall of Fame players had slightly higher batting averages, OPS, and adjusted OPS compared to non-Hall of Fame players. These statistics suggest that players who have performed well across a wide range of metrics are more likely to be inducted into the Hall of Fame.

Model Assumptions:

The assumptions of multivariate normality and homogeneity of covariance matrices were performed on the trained dataset of 627 players. The results showed both the assumptions of multivariate normality and homogeneity of covariance matrices failing, thus logistic regression model was used to construct the classification rule.

The assumption of Independence/Autocorrelation was also checked[Table 20]. The assumption was tested using Durbin Watson test and the resulting p-value is greater than 0.05, failing to reject the null hypothesis of no autocorrelation in the residuals. Therefore, we can assume that the independence assumption of the linear regression model is not violated by autocorrelation in the residuals.

Conclusion:

This study identified the key factors that influence a player's induction into the Baseball Hall of Fame using logistic regression. The results showed that the number of years played, JAWSRatio, runs batted during during a player's career, walks during a player's career, on base percentage, games played during a player's career, at bats during a player's career and home runs during a player's career were the most significant variables in determining a player's likelihood of being inducted. The logistic regression model was able to correctly classify 72.29% of the players who were inducted into the Hall of Fame and 95.44% of the players who were not inducted. Using this model, 12 potential candidates for the 2024 Hall of Fame were evaluated, and three players, Adrian Beltre, Joe Mauer, and Chase Utley, were identified as having a higher chance of being inducted.

This project tends to have practical implications for the selection process of the Baseball Hall of Fame. By identifying the key factors that influence a player's induction, this study can assist in the fair recognition of deserving baseball players who may have otherwise been overlooked. The three players identified by this study to be inducted into Hall of Fame, 2024 align with the predictions made by other official websites, suggesting the validity of this study's findings. In particular, it is worth mentioning that Adrian Beltre's induction to the Hall of Fame in 2024 is more prominent and unanimous with other studies. While this study represents one statistical analysis to predict a player's chances of being inducted into the Hall of Fame, there are many more analyses that can be performed.

The use of data and statistical models can help ensure a more objective and transparent selection process, which is crucial for maintaining the integrity and credibility of the Hall of Fame. This study can serve as a model for similar data analysis in other sports and industries where objective decision-making is paramount. Overall, this study provides insights into how data analytics can contribute to decision-making processes and promote fairness and transparency.

Limitations:

1)The threshold value used to categorize players is a crucial factor to consider as it can significantly impact the classification results. In this data a threshold value of 0.5 and above is used to categorize the players into Hall of Fame. However, lowering the threshold increases the number of players predicted as inducted, but it also increases the chances of false positives. Conversely, increasing the threshold reduces the number of predicted inductees, but it also increases the chances of false negatives. Therefore, the choice of threshold should be considered cautiously.

2)Second limitation of the model is the absence of prior probabilities. Prior probabilities are often based on prior knowledge, assumptions, and biases, which can be difficult to set for this dataset given the lack of prior information. The results are solely based on the data and model assumptions and do not account for any external information or prior knowledge that may affect the predictions.

3) The logistic regression model is limited by the variables and sample data used, and it may not capture all the factors that influence the induction process, such as sportsmanship, impact on the game, and character. Therefore, the model's results should be used in conjunction with other considerations, and not as the sole criterion for induction.

4) It is ultimately up to the Baseball Writers' Association of America to decide whether or not a player meets the criteria for induction. While the logistic regression model can provide insights into the factors that influence induction, it is not a definitive predictor and should be used in conjunction with other considerations.

5) It is worth noting that this logistic regression model is built using data only from non-pitcher players. Therefore, it may not be appropriate to use this model to predict Hall of Fame induction for pitcher players, as the factors influencing their induction may differ from non-pitcher players. This limitation should be considered when interpreting the results and applying the model to new data. Future research could explore the factors influencing Hall of Fame induction specifically for pitcher players and develop a separate model if necessary.

References:

[1] Britannica:
Title: Baseball Hall of Fame
Website: https://www.britannica.com/topic/Baseball-Hall-of-Fame

[2] National Baseball Hall of Fame and Museum:
Title: BBWAA Rules for Election
Website: https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election

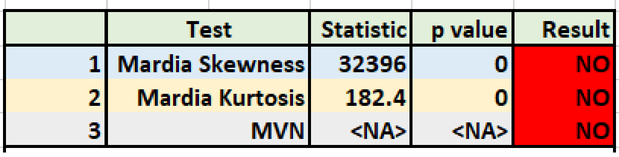
[3] Baseball Reference:
Title: Hall of Fame
Website: https://www.baseball-reference.com/awards/hof.shtml

[4] MLB.com:
Title: 2024 Hall of Fame ballot breakdown
Website: https://www.mlb.com/news/2024-hall-of-fame-ballot-breakdown

[5] theScore:
Title: Who's on the 2024 Hall of Fame ballot? Breaking down the new candidates
Website: https://www.thescore.com/mlb/news/2526592

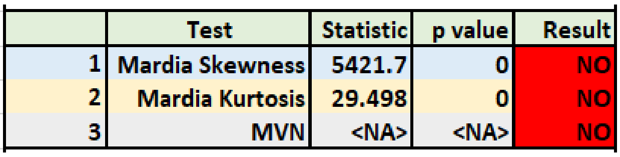
Tables and Plots:

Table 2: Mardia Test for checking the assumption of Multivariate Normality on the complete data set



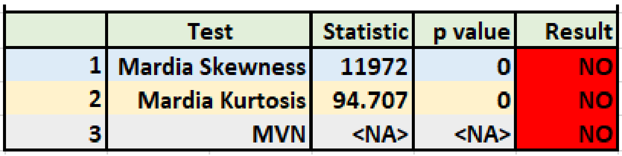
	Test	Statistic	p value	Result
1	Mardia Skewness	32396	0	NO
2	Mardia Kurtosis	182.4	0	NO
3	MVN	<NA>	<NA>	NO

Table 3: Mardia Test for checking the assumption of Multivariate Normality on the dataset of players inducted into Hall OF Fame



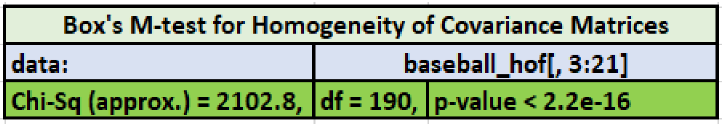
	Test	Statistic	p value	Result
1	Mardia Skewness	5421.7	0	NO
2	Mardia Kurtosis	29.498	0	NO
3	MVN	<NA>	<NA>	NO

Table 4: Mardia Test for checking the assumption of Multivariate Normality on the dataset of players not inducted into Hall OF Fame



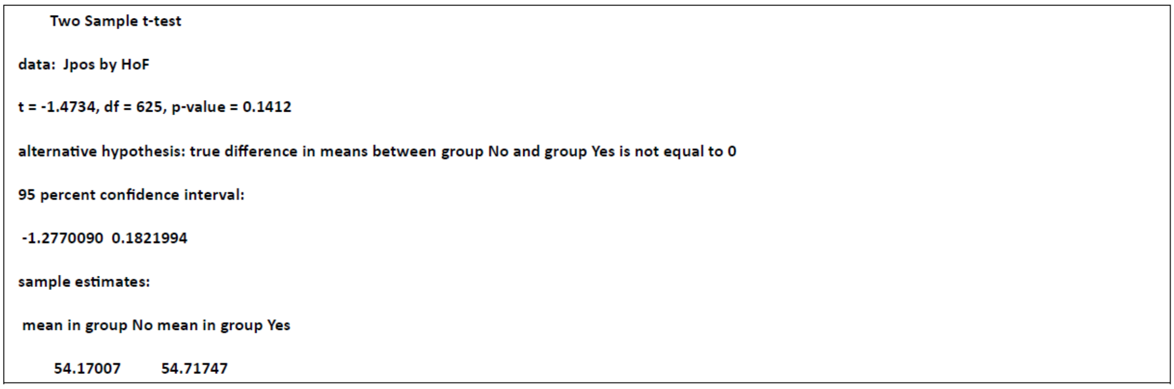
	Test	Statistic	p value	Result
1	Mardia Skewness	11972	0	NO
2	Mardia Kurtosis	94.707	0	NO
3	MVN	<NA>	<NA>	NO

Table 5: Box's M test for checking the assumption of homogeneity of covariance matrices



Box's M-test for Homogeneity of Covariance Matrices		
data:	baseball_hof[, 3:21]	
Chi-Sq (approx.) = 2102.8,	df = 190,	p-value < 2.2e-16

Table 6: T test performed to compare the means of Jpos between players inducted and not inducted in the Hall of Fame

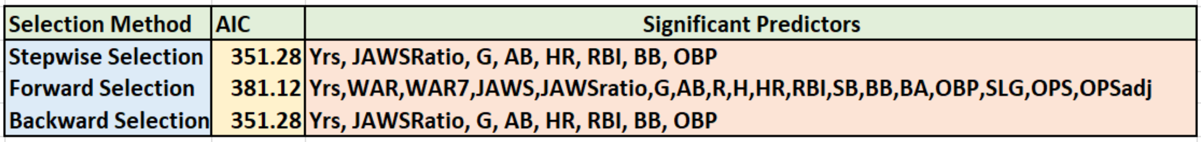


```
      Two Sample t-test

data:  Jpos by HoF

t = -1.4734, df = 625, p-value = 0.1412

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

 -1.2770090  0.1821994

sample estimates:

 mean in group No mean in group Yes

       54.17007        54.71747
```

Table 7: Logistic Regression on the trained dataset of 627 players

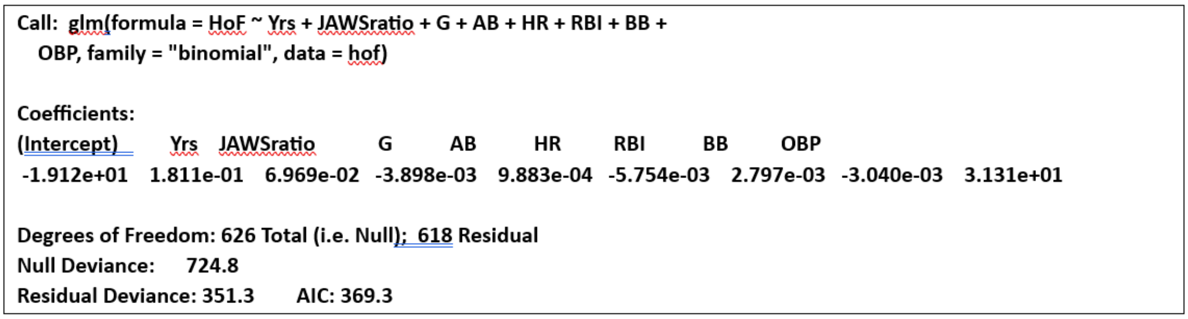
```
Call:
glm(formula = HoF ~ Yrs + WAR + WAR7 + JAWS + JAWSratio + G +
   AB + R + H + HR + RBI + SB + BB + BA + OBP + SLG + OPS +
   OPSadj, family = "binomial", data = hof)
Deviance Residuals:
   Min      1Q   Median      3Q      Max
-3.3953  -0.4281  -0.1443   0.0522   2.6236
Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.910e+01  1.001e+01  -2.908  0.00363 **
Yrs          2.238e-01  1.025e-01   2.183  0.02904 *
WAR          2.328e+00  1.953e+00   1.192  0.23341
WAR7         2.456e+00  1.959e+00   1.254  0.21001
JAWS        -4.774e+00  3.915e+00  -1.219  0.22269
JAWSratio    8.694e-02  3.104e-02   2.801  0.00509 **
G           -3.475e-03  2.649e-03  -1.312  0.18963
AB           1.962e-03  1.486e-03   1.321  0.18664
R            1.450e-03  1.954e-03   0.742  0.45811
H           -4.084e-03  4.342e-03  -0.941  0.34691
HR          -5.403e-03  3.884e-03  -1.391  0.16421
RBI          3.673e-03  1.276e-03   2.878  0.00400 **
SB          -7.281e-04  1.308e-03  -0.557  0.57772
BB          -4.428e-03  2.064e-03  -2.145  0.03193 *
BA           1.625e+01  3.507e+01   0.463  0.64317
OBP          1.343e+02  1.087e+02   1.235  0.21696
SLG          7.796e+01  1.064e+02   0.733  0.46359
OPS         -8.273e+01  1.058e+02  -0.782  0.43423
OPSadj      -1.782e-02  2.107e-02  -0.846  0.39781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 724.77  on 626  degrees of freedom
Residual deviance: 343.12  on 608  degrees of freedom
AIC: 381.12
Number of Fisher Scoring iterations: 7
```

Table 8: AICs of the three variable selection process along with the significant predictors identified by each model.



Selection Method	AIC	Significant Predictors
Stepwise Selection	351.28	Yrs, JAWSRatio, G, AB, HR, RBI, BB, OBP
Forward Selection	381.12	Yrs,WAR,WAR7,JAWS,JAWSratio,G,AB,R,H,HR,RBI,SB,BB,BA,OBP,SLG,OPS,OPSadj
Backward Selection	351.28	Yrs, JAWSRatio, G, AB, HR, RBI, BB, OBP

Table 9: Backward Selection Model of Logistic Regression

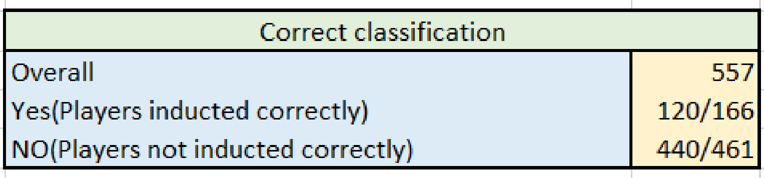


```
Call:  glm(formula = HoF ~ Yrs + JAWSratio + G + AB + HR + RBI + BB +
   OBP, family = "binomial", data = hof)

Coefficients:
(Intercept)     Yrs   JAWSratio       G          AB         HR        RBI         BB        OBP
 -1.912e+01  1.811e-01  6.969e-02  -3.898e-03  9.883e-04  -5.754e-03  2.797e-03  -3.040e-03  3.131e+01

Degrees of Freedom: 626 Total (i.e. Null);  618 Residual
Null Deviance:      724.8
Residual Deviance: 351.3       AIC: 369.3
```

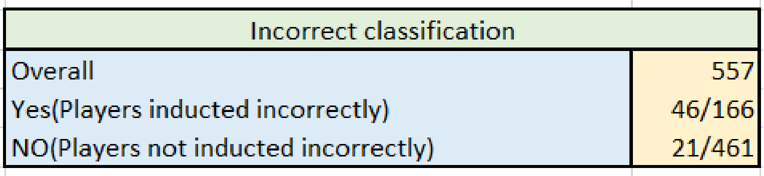
Table 10: Likelihood Ratio Test

```
baseball_hof_LRT <- baseball_hof_LR$null.deviance-baseball_hof_LR$deviance
baseball_hof_LRT
[1] 381.6519
baseball_hof_df <- baseball_hof_LR$df.null-baseball_hof_LR$df.residual
baseball_hof_df
[1] 18
1-pchisq(baseball_hof_LRT,baseball_hof_df)
[1] 0

Likelihood ratio test

Model 1: HoF ~ 1
Model 2: HoF ~ Yrs + WAR + WAR7 + JAWS + JAWSratio + G + AB + R + H +
   HR + RBI + SB + BB + BA + OBP + SLG + OPS + OPSadj
 #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -362.39
2  19 -171.56 18 381.65  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

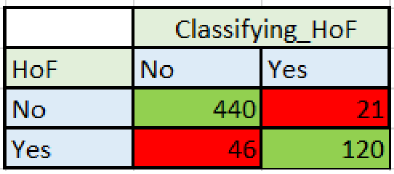
Table 11: Classification Performed on the training dataset along with the Confusion Matrix – Substitution Method



Correct classification	
Overall	557
Yes(Players inducted correctly)	120/166
NO(Players not inducted correctly)	440/461



Incorrect classification	
Overall	557
Yes(Players inducted incorrectly)	46/166
NO(Players not inducted incorrectly)	21/461



	Classifying_HoF	
HoF	No	Yes
No	440	21
Yes	46	120

Table 12: Graphical Representation of the dataset with the 672 non-pitchers on the X-axis and the probability on the   Y-axis

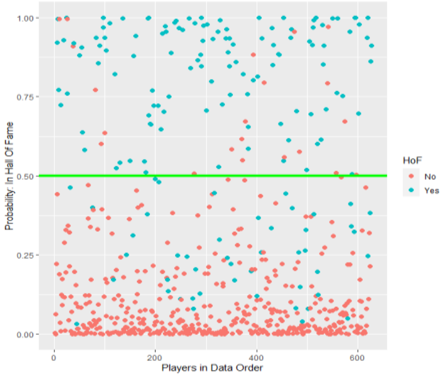
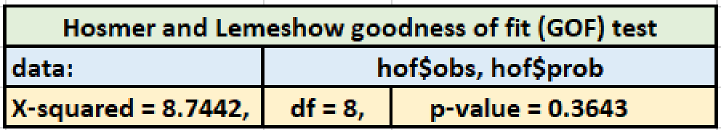
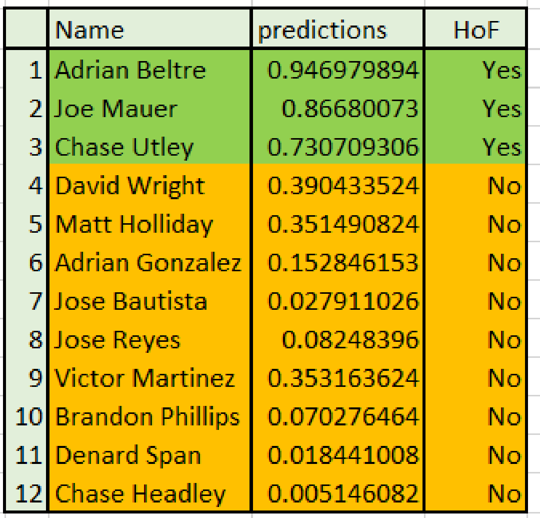


Table 13: Goodness Of Fit Test – Hosmer And Lemeshow



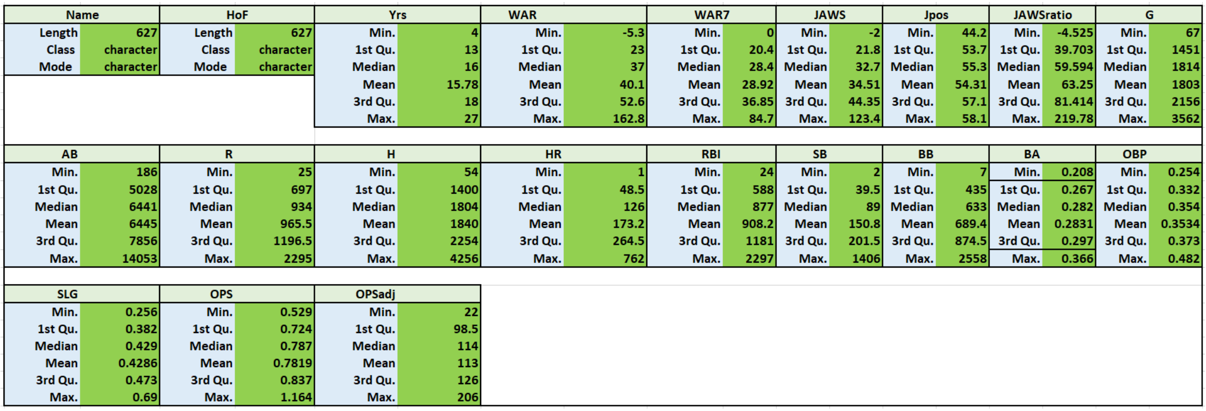
Hosmer and Lemeshow goodness of fit (GOF) test		
data:	hofobs, hofprob	
X-squared = 8.7442,	df = 8,	p-value = 0.3643

Table 14: List of 12 potential players with their predicted probability based on the classification rule using the trained dataset of 627 players.



	Name	predictions	HoF
1	Adrian Beltre	0.946979894	Yes
2	Joe Mauer	0.86680073	Yes
3	Chase Utley	0.730709306	Yes
4	David Wright	0.390433524	No
5	Matt Holliday	0.351490824	No
6	Adrian Gonzalez	0.152846153	No
7	Jose Bautista	0.027911026	No
8	Jose Reyes	0.08248396	No
9	Victor Martinez	0.353163624	No
10	Brandon Phillips	0.070276464	No
11	Denard Span	0.018441008	No
12	Chase Headley	0.005146082	No

Table 15: Summary Statistics of the 627 players

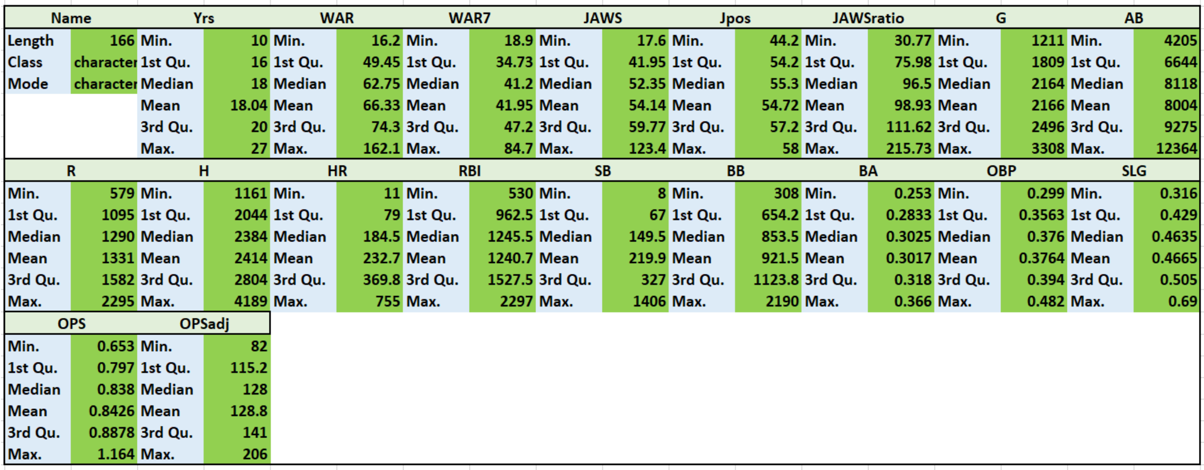


Name		HoF		Yrs		WAR		WAR7		JAWS		Jpos		JAWSratio		G	
Length	627	Length	627	Min.	4	Min.	-5.3	Min.	0	Min.	-2	Min.	44.2	Min.	-4.525	Min.	67
Class	character	Class	character	1st Qu.	13	1st Qu.	23	1st Qu.	20.4	1st Qu.	21.8	1st Qu.	53.7	1st Qu.	39.703	1st Qu.	1451
Mode	character	Mode	character	Median	16	Median	37	Median	28.4	Median	32.7	Median	55.3	Median	59.594	Median	1814
				Mean	15.78	Mean	40.1	Mean	28.92	Mean	34.51	Mean	54.31	Mean	63.25	Mean	1803
				3rd Qu.	18	3rd Qu.	52.6	3rd Qu.	36.85	3rd Qu.	44.35	3rd Qu.	57.1	3rd Qu.	81.414	3rd Qu.	2156
				Max.	27	Max.	162.8	Max.	84.7	Max.	123.4	Max.	58.1	Max.	219.78	Max.	3562

AB		R		H		HR		RBI		SB		BB		BA		OBP	
Min.	186	Min.	25	Min.	54	Min.	1	Min.	24	Min.	2	Min.	7	Min.	0.208	Min.	0.254
1st Qu.	5028	1st Qu.	697	1st Qu.	1400	1st Qu.	48.5	1st Qu.	588	1st Qu.	39.5	1st Qu.	435	1st Qu.	0.267	1st Qu.	0.332
Median	6441	Median	934	Median	1804	Median	126	Median	877	Median	89	Median	633	Median	0.282	Median	0.354
Mean	6445	Mean	965.5	Mean	1840	Mean	173.2	Mean	908.2	Mean	150.8	Mean	689.4	Mean	0.2831	Mean	0.3534
3rd Qu.	7856	3rd Qu.	1196.5	3rd Qu.	2254	3rd Qu.	264.5	3rd Qu.	1181	3rd Qu.	201.5	3rd Qu.	874.5	3rd Qu.	0.297	3rd Qu.	0.373
Max.	14053	Max.	2295	Max.	4256	Max.	762	Max.	2297	Max.	1406	Max.	2558	Max.	0.366	Max.	0.482

SLG		OPS		OPSadj	
Min.	0.256	Min.	0.529	Min.	22
1st Qu.	0.382	1st Qu.	0.724	1st Qu.	98.5
Median	0.429	Median	0.787	Median	114
Mean	0.4286	Mean	0.7819	Mean	113
3rd Qu.	0.473	3rd Qu.	0.837	3rd Qu.	126
Max.	0.69	Max.	1.164	Max.	206

Table 16: Summary statistics of the non-pitchers inducted into the HALL of FAME

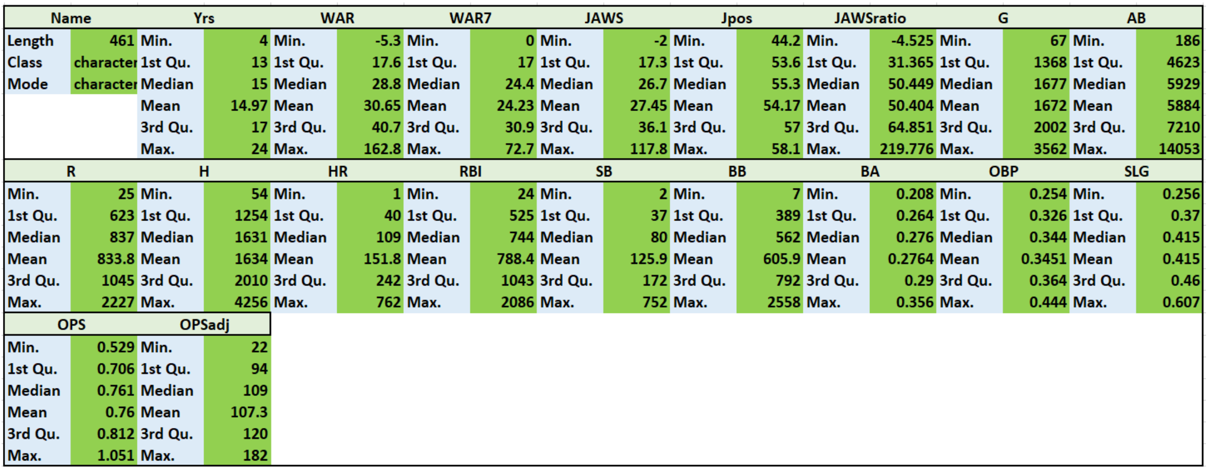


Name		Yrs		WAR		WAR7		JAWS		Jpos		JAWSratio		G		AB	
Length	166	Min.	10	Min.	16.2	Min.	18.9	Min.	17.6	Min.	44.2	Min.	30.77	Min.	1211	Min.	4205
Class	character	1st Qu.	16	1st Qu.	49.45	1st Qu.	34.73	1st Qu.	41.95	1st Qu.	54.2	1st Qu.	75.98	1st Qu.	1809	1st Qu.	6644
Mode	character	Median	18	Median	62.75	Median	41.2	Median	52.35	Median	55.3	Median	96.5	Median	2164	Median	8118
		Mean	18.04	Mean	66.33	Mean	41.95	Mean	54.14	Mean	54.72	Mean	98.93	Mean	2166	Mean	8004
		3rd Qu.	20	3rd Qu.	74.3	3rd Qu.	47.2	3rd Qu.	59.77	3rd Qu.	57.2	3rd Qu.	111.62	3rd Qu.	2496	3rd Qu.	9275
		Max.	27	Max.	162.1	Max.	84.7	Max.	123.4	Max.	58	Max.	215.73	Max.	3308	Max.	12364

R		H		HR		RBI		SB		BB		BA		OBP		SLG	
Min.	579	Min.	1161	Min.	11	Min.	530	Min.	8	Min.	308	Min.	0.253	Min.	0.299	Min.	0.316
1st Qu.	1095	1st Qu.	2044	1st Qu.	79	1st Qu.	962.5	1st Qu.	67	1st Qu.	654.2	1st Qu.	0.2833	1st Qu.	0.3563	1st Qu.	0.429
Median	1290	Median	2384	Median	184.5	Median	1245.5	Median	149.5	Median	853.5	Median	0.3025	Median	0.376	Median	0.4635
Mean	1331	Mean	2414	Mean	232.7	Mean	1240.7	Mean	219.9	Mean	921.5	Mean	0.3017	Mean	0.3764	Mean	0.4665
3rd Qu.	1582	3rd Qu.	2804	3rd Qu.	369.8	3rd Qu.	1527.5	3rd Qu.	327	3rd Qu.	1123.8	3rd Qu.	0.318	3rd Qu.	0.394	3rd Qu.	0.505
Max.	2295	Max.	4189	Max.	755	Max.	2297	Max.	1406	Max.	2190	Max.	0.366	Max.	0.482	Max.	0.69

OPS		OPSadj	
Min.	0.653	Min.	82
1st Qu.	0.797	1st Qu.	115.2
Median	0.838	Median	128
Mean	0.8426	Mean	128.8
3rd Qu.	0.8878	3rd Qu.	141
Max.	1.164	Max.	206

Table 17: Summary statistics of the players not inducted into the HALL of FAME

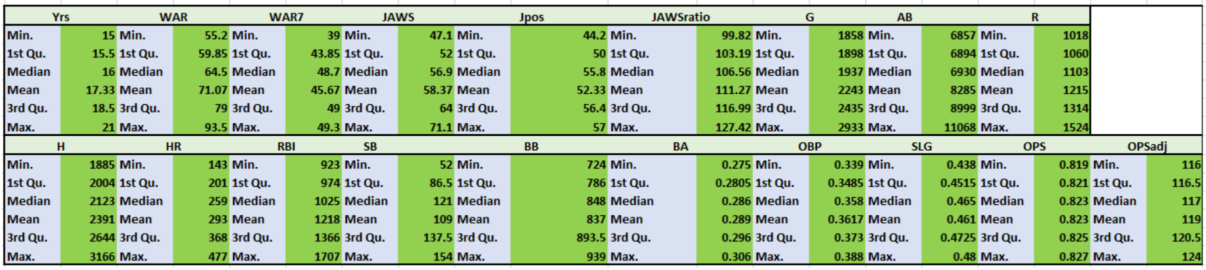


Name		Yrs		WAR		WAR7		JAWS		Jpos		JAWSratio		G		AB	
Length	461	Min.	4	Min.	-5.3	Min.	0	Min.	-2	Min.	44.2	Min.	-4.525	Min.	67	Min.	186
Class	character	1st Qu.	13	1st Qu.	17.6	1st Qu.	17	1st Qu.	17.3	1st Qu.	53.6	1st Qu.	31.365	1st Qu.	1368	1st Qu.	4623
Mode	character	Median	15	Median	28.8	Median	24.4	Median	26.7	Median	55.3	Median	50.449	Median	1677	Median	5929
		Mean	14.97	Mean	30.65	Mean	24.23	Mean	27.45	Mean	54.17	Mean	50.404	Mean	1672	Mean	5884
		3rd Qu.	17	3rd Qu.	40.7	3rd Qu.	30.9	3rd Qu.	36.1	3rd Qu.	57	3rd Qu.	64.851	3rd Qu.	2002	3rd Qu.	7210
		Max.	24	Max.	162.8	Max.	72.7	Max.	117.8	Max.	58.1	Max.	219.776	Max.	3562	Max.	14053

R		H		HR		RBI		SB		BB		BA		OBP		SLG	
Min.	25	Min.	54	Min.	1	Min.	24	Min.	2	Min.	7	Min.	0.208	Min.	0.254	Min.	0.256
1st Qu.	623	1st Qu.	1254	1st Qu.	40	1st Qu.	525	1st Qu.	37	1st Qu.	389	1st Qu.	0.264	1st Qu.	0.326	1st Qu.	0.37
Median	837	Median	1631	Median	109	Median	744	Median	80	Median	562	Median	0.276	Median	0.344	Median	0.415
Mean	833.8	Mean	1634	Mean	151.8	Mean	788.4	Mean	125.9	Mean	605.9	Mean	0.2764	Mean	0.3451	Mean	0.415
3rd Qu.	1045	3rd Qu.	2010	3rd Qu.	242	3rd Qu.	1043	3rd Qu.	172	3rd Qu.	792	3rd Qu.	0.29	3rd Qu.	0.364	3rd Qu.	0.46
Max.	2227	Max.	4256	Max.	762	Max.	2086	Max.	752	Max.	2558	Max.	0.356	Max.	0.444	Max.	0.607

OPS		OPSadj	
Min.	0.529	Min.	22
1st Qu.	0.706	1st Qu.	94
Median	0.761	Median	109
Mean	0.76	Mean	107.3
3rd Qu.	0.812	3rd Qu.	120
Max.	1.051	Max.	182

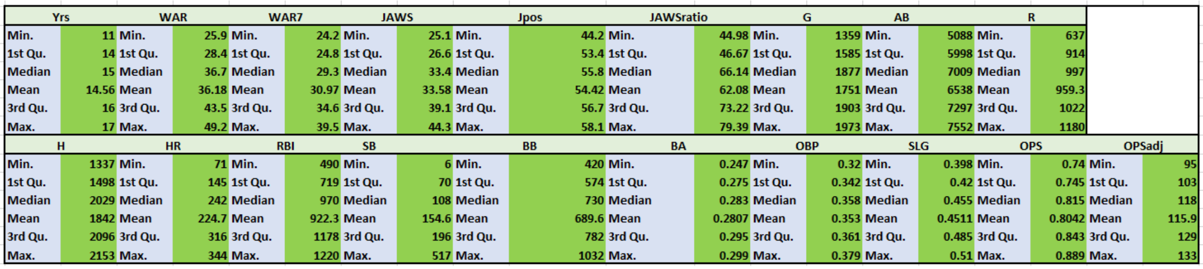
Table 18: Summary statistics of the test data - Players with higher chances of getting inducted into the HALL of FAME



Yrs		WAR		WAR7		JAWS		Jpos		JAWSratio		G		AB		R	
Min.	15	Min.	55.2	Min.	39	Min.	47.1	Min.	44.2	Min.	99.82	Min.	1858	Min.	6857	Min.	1018
1st Qu.	15.5	1st Qu.	59.85	1st Qu.	43.85	1st Qu.	52	1st Qu.	50	1st Qu.	103.19	1st Qu.	1898	1st Qu.	6894	1st Qu.	1060
Median	16	Median	64.5	Median	48.7	Median	56.9	Median	55.8	Median	106.56	Median	1937	Median	6930	Median	1103
Mean	17.33	Mean	71.07	Mean	45.67	Mean	58.37	Mean	52.33	Mean	111.27	Mean	2243	Mean	8285	Mean	1215
3rd Qu.	18.5	3rd Qu.	79	3rd Qu.	49	3rd Qu.	64	3rd Qu.	56.4	3rd Qu.	116.99	3rd Qu.	2435	3rd Qu.	8999	3rd Qu.	1314
Max.	21	Max.	93.5	Max.	49.3	Max.	71.1	Max.	57	Max.	127.42	Max.	2933	Max.	11068	Max.	1524

H		HR		RBI		SB		BB		BA		OBP		SLG		OPS		OPSadj	
Min.	1885	Min.	143	Min.	923	Min.	52	Min.	724	Min.	0.275	Min.	0.339	Min.	0.438	Min.	0.819	Min.	116
1st Qu.	2004	1st Qu.	201	1st Qu.	974	1st Qu.	86.5	1st Qu.	786	1st Qu.	0.2805	1st Qu.	0.3485	1st Qu.	0.4515	1st Qu.	0.821	1st Qu.	116.5
Median	2123	Median	259	Median	1025	Median	121	Median	848	Median	0.286	Median	0.358	Median	0.465	Median	0.823	Median	117
Mean	2391	Mean	293	Mean	1218	Mean	109	Mean	837	Mean	0.289	Mean	0.3617	Mean	0.461	Mean	0.823	Mean	119
3rd Qu.	2644	3rd Qu.	368	3rd Qu.	1366	3rd Qu.	137.5	3rd Qu.	893.5	3rd Qu.	0.296	3rd Qu.	0.373	3rd Qu.	0.4725	3rd Qu.	0.825	3rd Qu.	120.5
Max.	3166	Max.	477	Max.	1707	Max.	154	Max.	939	Max.	0.306	Max.	0.388	Max.	0.48	Max.	0.827	Max.	124

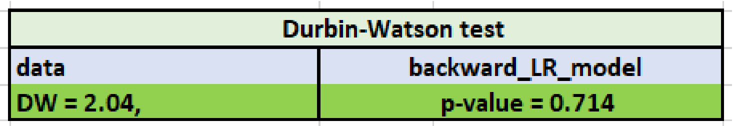
Table 19: Summary statistics of the test data - Players with lesser chances of getting inducted into the HALL of FAME



Yrs		WAR		WAR7		JAWS		Jpos		JAWSratio		G		AB		R	
Min.	11	Min.	25.9	Min.	24.2	Min.	25.1	Min.	44.2	Min.	44.98	Min.	1359	Min.	5088	Min.	637
1st Qu.	14	1st Qu.	28.4	1st Qu.	24.8	1st Qu.	26.6	1st Qu.	53.4	1st Qu.	46.67	1st Qu.	1585	1st Qu.	5998	1st Qu.	914
Median	15	Median	36.7	Median	29.3	Median	33.4	Median	55.8	Median	66.14	Median	1877	Median	7009	Median	997
Mean	14.56	Mean	36.18	Mean	30.97	Mean	33.58	Mean	54.42	Mean	62.08	Mean	1751	Mean	6538	Mean	959.3
3rd Qu.	16	3rd Qu.	43.5	3rd Qu.	34.6	3rd Qu.	39.1	3rd Qu.	56.7	3rd Qu.	73.22	3rd Qu.	1903	3rd Qu.	7297	3rd Qu.	1022
Max.	17	Max.	49.2	Max.	39.5	Max.	44.3	Max.	58.1	Max.	79.39	Max.	1973	Max.	7552	Max.	1180

H		HR		RBI		SB		BB		BA		OBP		SLG		OPS		OPSadj	
Min.	1337	Min.	71	Min.	490	Min.	6	Min.	420	Min.	0.247	Min.	0.32	Min.	0.398	Min.	0.74	Min.	95
1st Qu.	1498	1st Qu.	145	1st Qu.	719	1st Qu.	70	1st Qu.	574	1st Qu.	0.275	1st Qu.	0.342	1st Qu.	0.42	1st Qu.	0.745	1st Qu.	103
Median	2029	Median	242	Median	970	Median	108	Median	730	Median	0.283	Median	0.358	Median	0.455	Median	0.815	Median	118
Mean	1842	Mean	224.7	Mean	922.3	Mean	154.6	Mean	689.6	Mean	0.2807	Mean	0.353	Mean	0.4511	Mean	0.8042	Mean	115.9
3rd Qu.	2096	3rd Qu.	316	3rd Qu.	1178	3rd Qu.	196	3rd Qu.	782	3rd Qu.	0.295	3rd Qu.	0.361	3rd Qu.	0.485	3rd Qu.	0.843	3rd Qu.	129
Max.	2153	Max.	344	Max.	1220	Max.	517	Max.	1032	Max.	0.299	Max.	0.379	Max.	0.51	Max.	0.889	Max.	133

Table 20: Durbin Watson Test for the assumption of independence/autocollinearity



Durbin-Watson test	
data	backward_LR_model
DW = 2.04,	p-value = 0.714