SPORTS ANALYTICS – BASEBALL PREDICTIONS

Contents

[**EXECUTIVE SUMMARY:** 3](#_Toc434936695)

[**BUSINESS PROBLEM:** 3](#_Toc434936696)

[**STEPS FOLLOWED:** 3](#_Toc434936697)

[**KEY RESULTS:** 5](#_Toc434936698)

[**NEXT STEPS:** 5](#_Toc434936699)

[**BUSINESS CASE:** 5](#_Toc434936700)

[**IMPLEMENTATION OVERVIEW:** 5](#_Toc434936701)

[**IMPLEMENTATION IN DETAIL:** 6](#_Toc434936702)

[**DATA PREPARATION:** 6](#_Toc434936703)

[**SEMMA: Sample, Explore, Modify, Model, Assess** 10](#_Toc434936704)

[**SAMPLE:** 10](#_Toc434936705)

[**EXPLORE:** 10](#_Toc434936706)

[**Change Variable Type:** 10](#_Toc434936707)

[**Eliminate irrelevant variables:** 10](#_Toc434936708)

[**Data exploration to find patterns:** 11](#_Toc434936709)

[**MODIFY:** 13](#_Toc434936710)

[**Standardizing the variables** 13](#_Toc434936711)

[**Missing Values:** 13](#_Toc434936712)

[**Reducing dimensionality:** 16](#_Toc434936713)

[**Adding Derived variables:** 20](#_Toc434936714)

[**Analyzing potential outliers:** 20](#_Toc434936715)

[**Variable Transformation:** 22](#_Toc434936716)

[**NULL’S in SALARY:** 23](#_Toc434936717)

[**Model: Predicting Salaries** 23](#_Toc434936718)

[**Regression Model:** 23](#_Toc434936719)

[**Regression model of Cluster 1:** 24](#_Toc434936720)

[**Regression model for Cluster 2:** 24](#_Toc434936721)

[**Ensemble model:** **Linear Regression Model for Cluster 1** 24](#_Toc434936722)

[**Ensemble Model: Linear Regression model for Cluster 2** 25](#_Toc434936723)

[**MODEL:** 26](#_Toc434936724)

[**1)** **Neural Nets:** 26](#_Toc434936725)

[**2)** **Regression Model:** 27](#_Toc434936726)

[**3)** **Ensemble model 1: Logistic Regression** 27](#_Toc434936727)

[**4)** **Ensemble Model 2: Neural Nets** 28](#_Toc434936728)

[**ASSESS:** 29](#_Toc434936729)

[**TEXT MINING:** 34](#_Toc434936730)

[**Step 1: Extract twitter data** 34](#_Toc434936731)

[**Step 2: Build word cloud** 34](#_Toc434936732)

[**Step 3: Perform Sentiment Analysis** 35](#_Toc434936733)

[**TIME SERIES FORECASTING:** 36](#_Toc434936734)

[**CONCLUSION:** 37](#_Toc434936735)

[**APPENDIX:** 38](#_Toc434936736)

# **EXECUTIVE SUMMARY:**

## **BUSINESS PROBLEM:**

Our business problem deals with predicting whether a player will be inducted into Hall Of Fame or not based on the player’s career accomplishments. For this case, we gathered data and data dictionary from Lahman baseball database in CSV format.

<http://www.seanlahman.com/baseball-archive/statistics/>

We have used the below link to understand Baseball jargons.

<https://en.wikipedia.org/wiki/Baseball_statistics>

Data Dictionary



## **STEPS FOLLOWED:**

A quick rundown on the steps we followed as part of our project to develop the model to predict whether a player will be inducted into Hall of Fame or not.

* **Data Preparation:**

Lahman dataset  
26 data files  
Data Dictionary

Identify relevant tables   
(8 files)

Merged Final data set

Our integrated final data set consists of 18590 players’ records from the year 1871 to 2014.

**Challenges/Issues:**

* Identifying relevant tables/factors that affect a player’s eligibility to enter hall of fame.
* Some variables were ambiguous owing to the following:
  + Lack of clear definition in the data dictionary
  + Lack of domain knowledge from our end
* Decomposing multi-valued variables into multiple columns to fetch meaningful information that could be used in our prediction.
* Arriving at the right type of statistics such as Sum or count for each player depending on the variable.
* Converting date fields such as Debut date with inconsistent formats to uniform format for which we faced challenges and finally found a working solution using R.
* **Sampling:** We had performed random sampling to select 2880 players from those who were not inducted and had included all the 245 players who got inducted into hall of fame, out of the 18590 records in the population.

Players in the sample: 3125

Number of variables: 140

**Challenges/Issues:** We had to carefully arrive at a sample set that would not in any way lose these 245 records which would play a major role in training our model.

* **Data Exploration/Visualization:** We changed the variable types, eliminated irrelevant variables and explored the data to patterns. We found interesting patterns while exploring the data. Some of initial hypothesis that we framed based on our visualizations include:
* Right handed batter is more likely to get inducted into hall of fame
* Players who had participated in higher number of leagues are more likely to get inducted into hall of fame.
* Players who had received awards such as are more likely to get inducted into hall of fame

**Challenges:** Lot of0’s present in the data owing to players falling in exclusive buckets of pitchers, hitters and fielders. Owing to this, we knew that a lot of our variables are right skewed. The other discrepancy was that players had salary only from 1885

* **Modify:**

We found that many variables belonging to Batting, Pitching and Fielding tables had missing values for many players.

Key modifications include:

* Recoded missing values with 0 for those players who are only batters and no pitching/fielding variables are applicable to them.
* Removed 26 records which did not have any pitching/fielding/batting statistics.
* Added new variables, such as # of Awards and Years of Experience for each player and derived values from existing variables.
* Created bins for variables such as Years of Experience
* Removed some variables in each table, based on the correlation coefficients.
* Refrained from adding principal components as it reduced the explanatory power of variables.

**Challenges/Issues:**

* Distribution of most of the variables were right skewed and more than 50% of them are zeroes. We tried dealing with the issue by performing Log, Log(x+0.1) transformation but did not reduce the skewness. Then, we tried adding three additional baseball statistics which summarized each of the batting, pitching and fielding statistics. The additional statistics have minimum skew and are nearly normal distributed.
* Salary details were available for the players only after 1985. Tried imputing formula

obtained from linear regression after forming clusters of batters and pitchers.

* **Model:**

As our problem statement deals with predicting a nominal variable, i.e. a player will be inducted into hall of fame or not, we tried creating models using Logistic Regression, Neural Networks, Decision Trees, Ensemble models using Logistic Regression and Neural Networks.

* **Assess:**

We have identified Logistic Regression as the best model among those we developed based on its accuracy level and probability of occurrence of False Positive and False Negative errors. However, to corroborate our findings obtained using Logistic Regression, we also extended our modeling scope by performing Text Mining and Sentiment Analysis.

**KEY RESULTS:**

From our model results and further analysis using text mining, we could infer that our model does provide fairly accurate results. We considered the case of a single player Mike Piazza and performed sentiment analysis and text mining which confirmed our logistic model results of Mike being selected for Hall of Fame. However, in the interest of time, we could not perform sentiment analysis for all the players. We also attempted to perform Time series analysis to forecast the performance of Mike Piazza based on his pitching statistics.

## **NEXT STEPS:**

We plan to extend our case by working on the below.

1. Perform Sentiment Analysis/Text Mining for all the players.
2. Time Series Forecasting techniques other than Moving Averages to predict the player’s performance.
3. Develop models to predict the below.
4. Which School/College is most likely to produce good batters/pitchers?
5. Which player will have the best batting/pitching statistics for the next season?
6. Will a manager/player win an award or not?

# **BUSINESS CASE:**

Baseball analytics such as Sabermetrics plays a key role and is widely used to predict a player’s performance, salaries, etc. Such baseball analytics not only aids team managers to evaluate players for selection but also to help a player identify his strengths and areas of improvement. Moreover, such analytics helps the team managers to decide whether a player deserves the salary he is being paid and evaluate whether he will be able to deliver the expected performance based on the results provided by analytical models. Considering the widespread adoption of Analytics in baseball, we decided to pursue our case on baseball analytics to predict whether a player will get inducted into hall of fame or not.

# **IMPLEMENTATION OVERVIEW:**

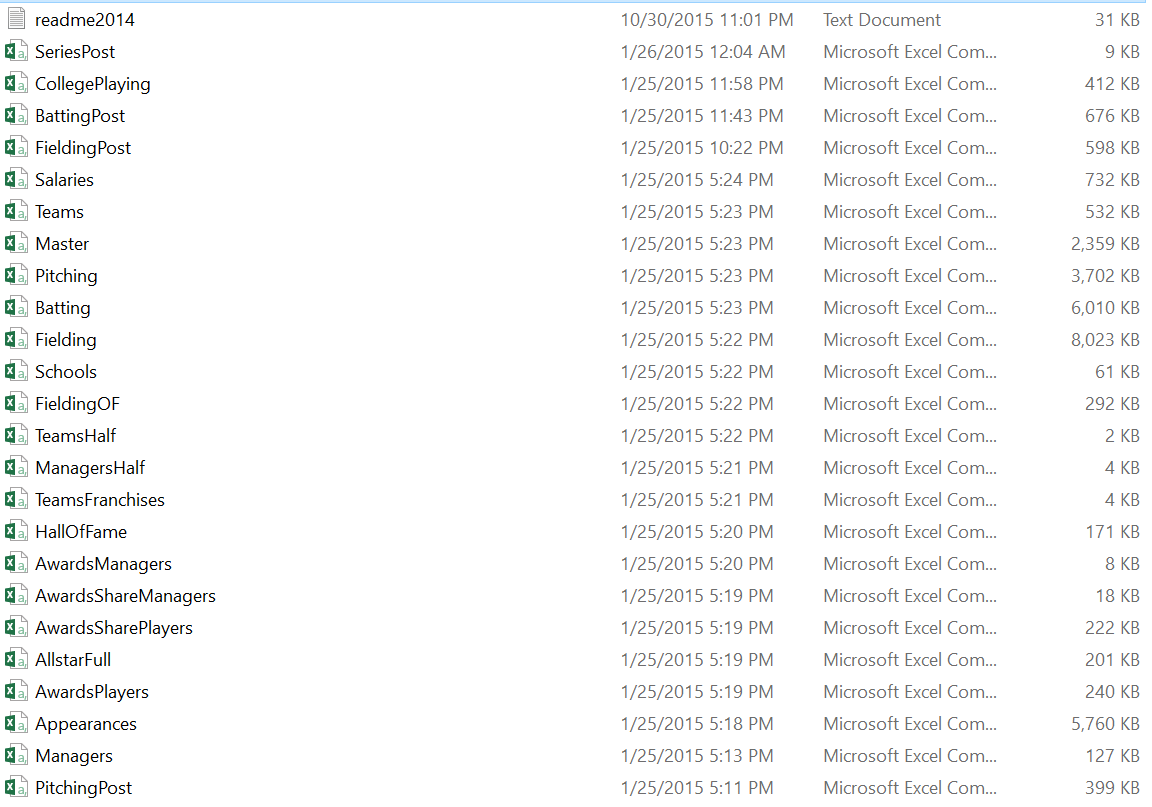
We approached the problem by identifying relevant datasets from the original Lahman dataset, joined the identified tables and followed SEMMA approach in the newly merged data set. Besides developing the models using Neural Networks, Decision Trees, Logistic Regression, etc., we also attempted Text Mining/Sentiment Analysis and Time Series Forecasting.

# **IMPLEMENTATION IN DETAIL:**

## **DATA PREPARATION:**

We had 26 .csv files in the original data set from which we identified 8 files which are relevant to finding solution for our problem statement. We faced challenges in identifying the files relevant to our problem statement, as it required understanding of the game and the criteria for a player to be eligible to enter the hall of fame. After identifying the 8 files, we joined those tables to obtain a unified data set. Our integrated data set consists of 18590 players’ records from 1871 to 2014.

**Snapshot of Lahman data files**



**Challenges/Issues:**

1. Identifying relevant tables/factors that affect a player’s eligibility to enter hall of fame.
2. We had players’ batting, pitching and fielding records available by league, team and year. We had to arrive at the right type of statistics such as Sum or count depending on the variable, as we were looking at a player’s career accomplishments till date to reflect on his eligibility to get inducted into hall of fame.

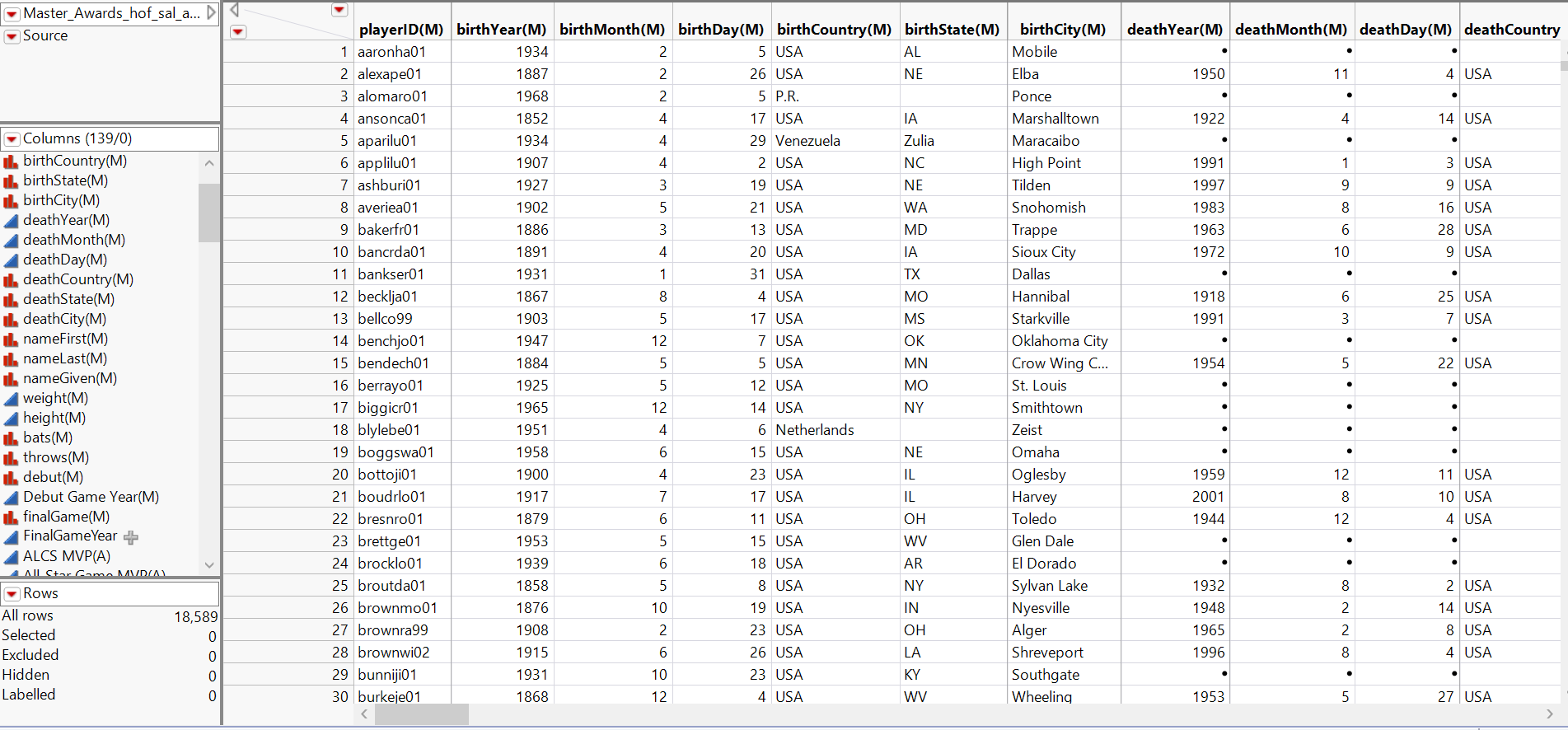
The above files are each aggregated to player level and are merged based on player ID to capture player level career statistics.

|  |  |  |  |
| --- | --- | --- | --- |
| FILENAME | LEVEL | SUMMARIZED LEVEL | JOIN KEY |
| MASTER | Player | Not summarized | Player ID |
| AWARDS | Player, Year, League | Player | Player ID |
| HALL OF FAME | Player, Year | Player | Player ID |
| SALARY | Player, Year, League | Player | Player ID |
| APPREARANCES | Player, Game, Year | Player | Player ID |
| BATTING | Player, Year | Player | Player ID |
| PITCHING | Player, Year, League | Player | Player ID |
| FIELDING | Player, Year | Player | Player ID |

**Merged File:** Master\_Awards\_hof\_sal\_app\_bat\_pit\_field

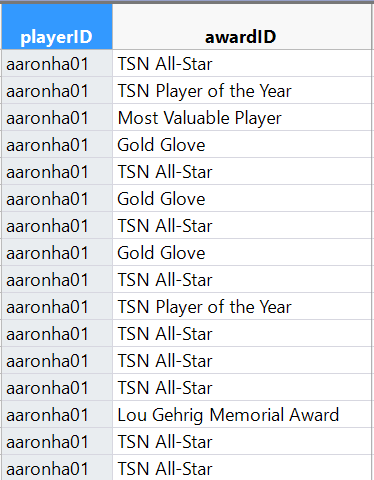
The merged file will have summarized player level baseball data with his batting, pitching, fielding, awards, appearances and salaries.

Snapshot of the merged file:

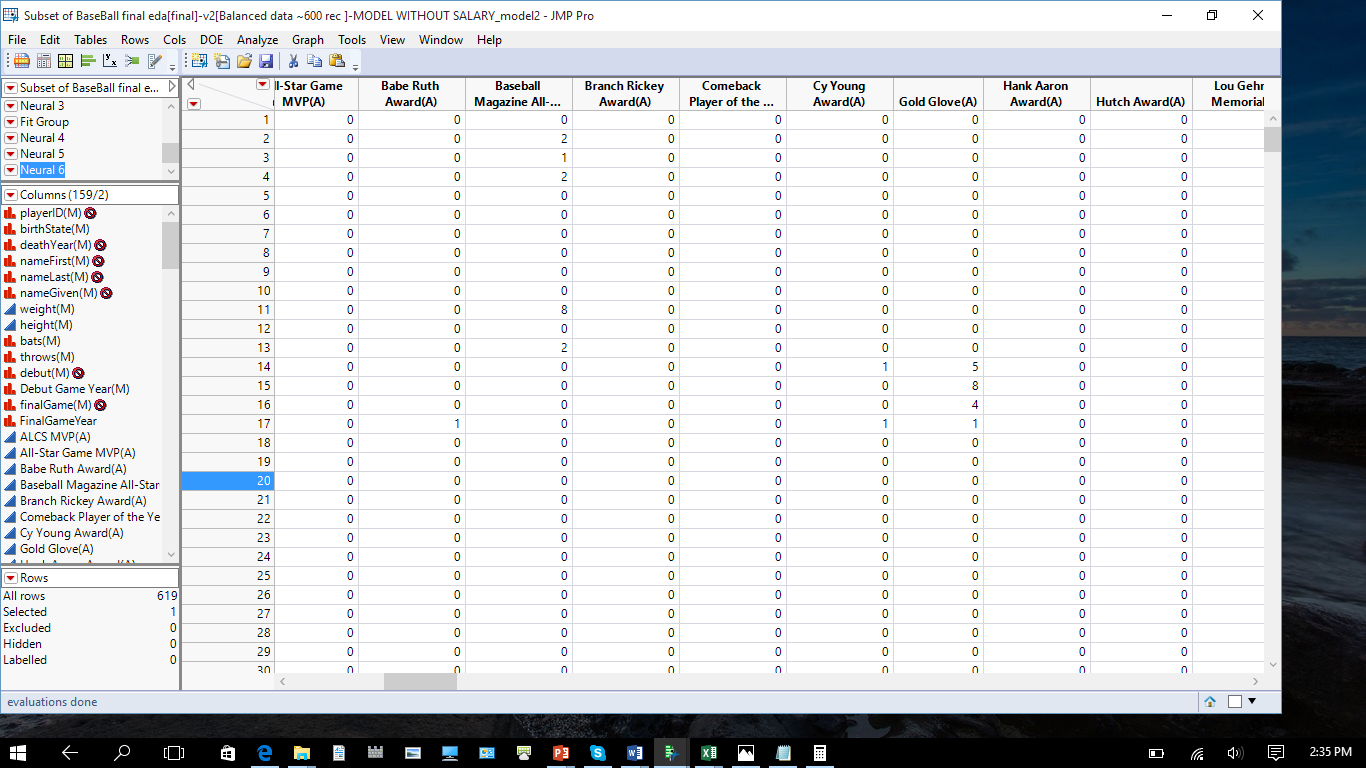


1. Some tables had multi-valued attributes such as Awards for each player, where we had to decompose those values into multiple columns to get meaningful information that could be used in our prediction.

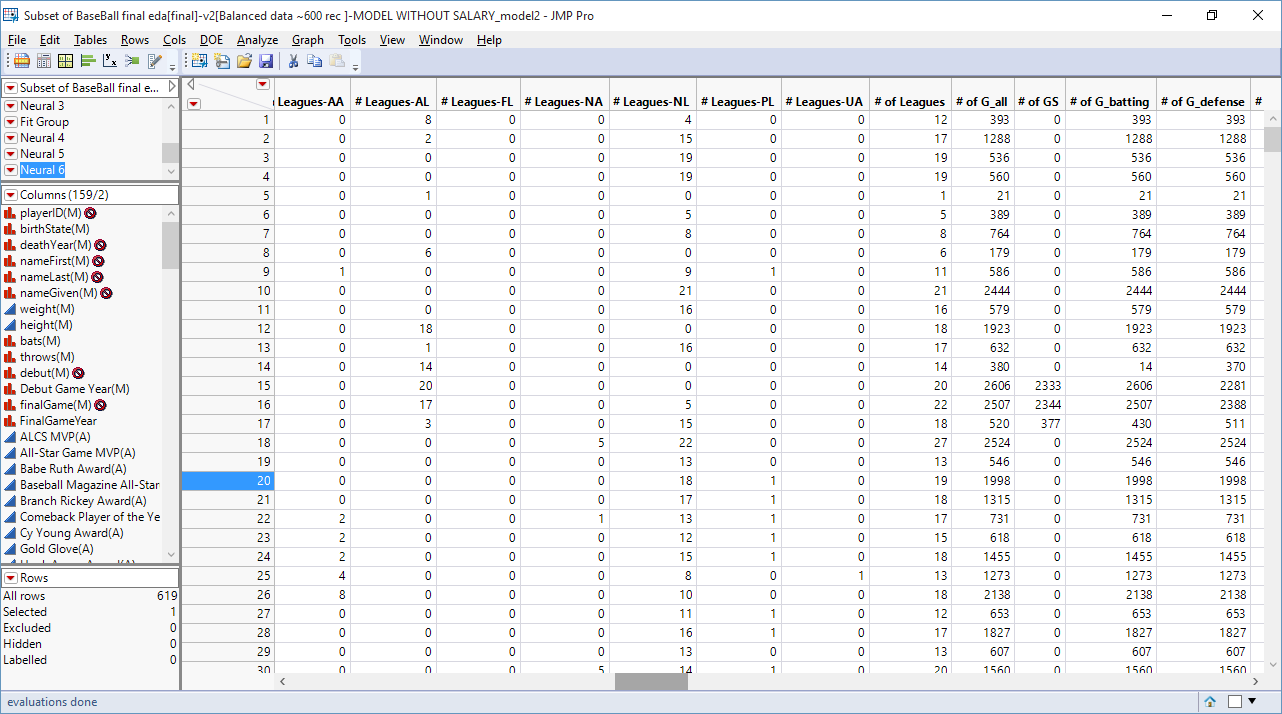
**Data in the original file:**



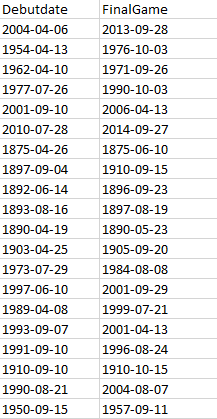
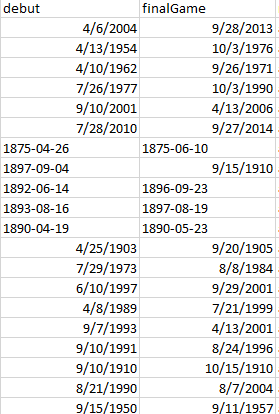
**# of different types of Awards won by each player**



**# of different leagues in which each player had participated**



1. We had fields such as Debut Date and Final Game date, whose values for players were in inconsistent format. We had challenges in converting the date to uniform format, despite using Microsoft Excel. Finally, we found a working solution to convert the dates to uniform format using R

****

**SUMMARY OF MERGED DATA SET:**

Number of rows: 18,589

Number of columns: 140

Response variables: HOF (inducted, Y)

It is a binary variable where

Not Inducted into Hall of Fame is identified as 0

Inducted into Hall of Fame is identified as 1

Number of players inducted: 245

Number of players not inducted: 18,344

## **SEMMA: Sample, Explore, Modify, Model, Assess**

## **SAMPLE:**

We have taken a sample of 3125 records and we tried to retain all the inductees into our sample.

Sampling Method: Stratified sample

Players in the sample: 3125

Number of inductees: 245

Number of non-inductees: 2880

**Challenges/Issues:**

We had to carefully arrive at a sample set that would not in any way lose these 245 records which would play a major role in training our model.

## **EXPLORE:**

### **Change Variable Type:**

Only the below variables required change in variable type.

**Debut Game Year:** Continuous to Nominal

**Final Game Year:** Continuous to Nominal

**HOF (inducted, N):** Continuous to Nominal

**HOF (inducted, Y):** Continuous to Nominal

### **Eliminate irrelevant variables:**

The below columns were excluded as they are player’s personal details and not the baseball statistics.

birthYear

birthMonth

birthday

birthCountry

birthCity

deathMonth

deathDay

deathCountry

deathState

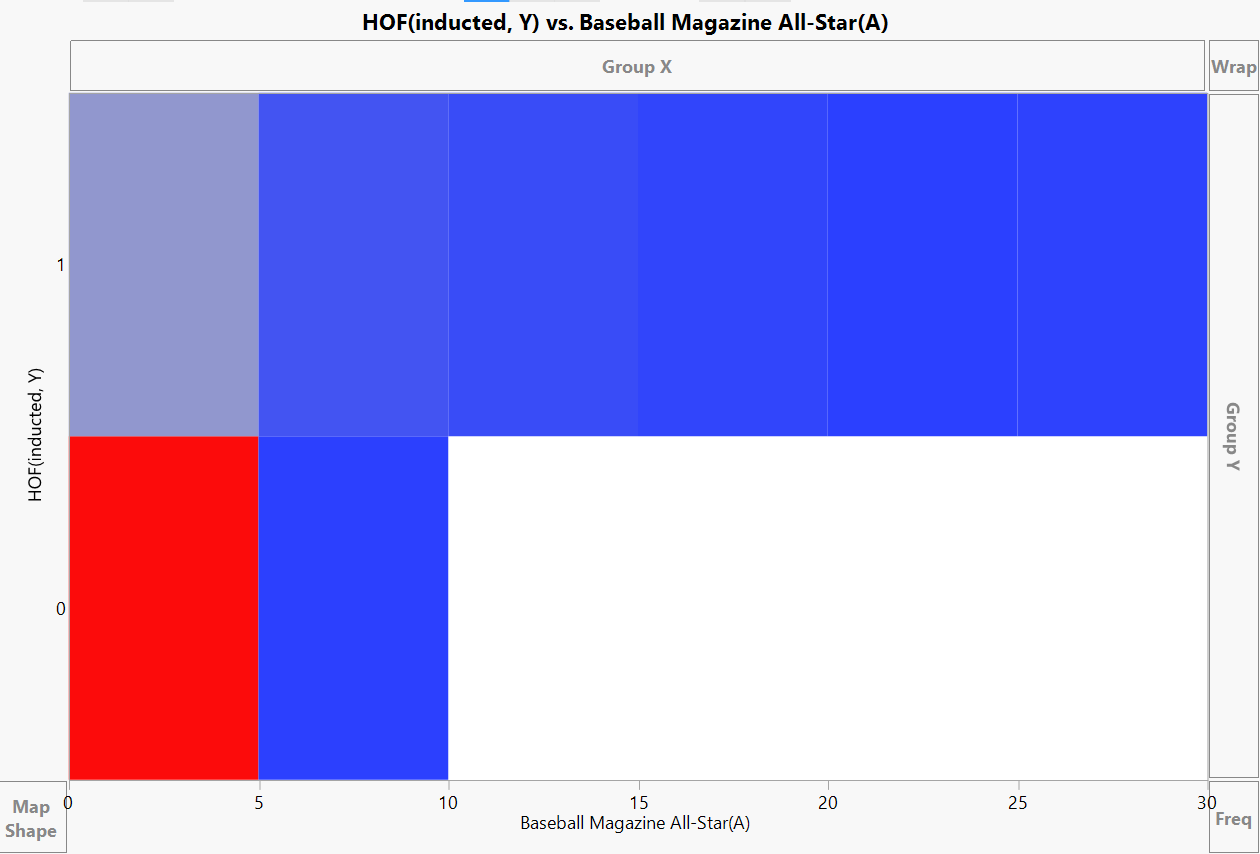
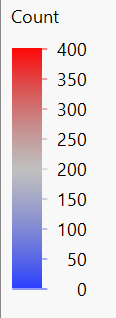
deathCity

### **Data exploration to find patterns:**

We explored the sample dataset to identify the drivers/factors that influence a player’s entry into hall of fame. We found interesting patterns while exploring the data.

Some of the hypothesis that we framed around this are:

* Right handed batter is more likely to get inducted into hall of fame
* Players who had participated in higher number of leagues are more likely to get inducted into hall of fame
* Players who had received Baseball Magazine All-Star awards such as are more likely to get inducted into hall of fame



Players who received more than 10 Baseball Magazine All-Star awards, are all inducted into Hall of Fame

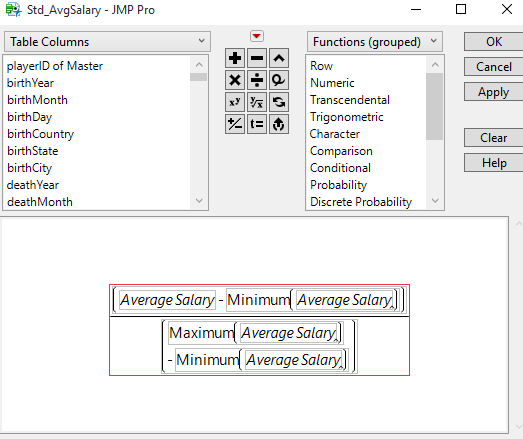
## **MODIFY:**

### **Standardizing the variables**

Better to put all continuous variables on the same scale, so that when distance measures are calculated, columns with large scale values don’t bias the output)

Standardizing columns that are continuous using the formula:-

[(X) – min(X)]/ [max(X)-min(X)]



### **Missing Values:**

We observed missing values for below variables and we imputed them as below.

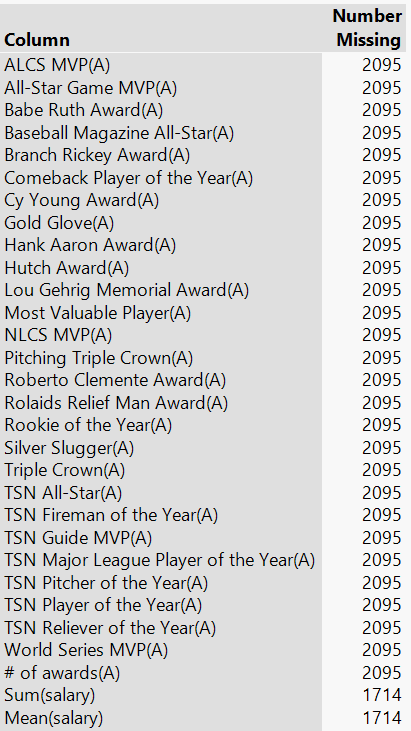
**NULL’s in Weight:** 5 NULL values. Imputed using Multivariate SVD Imputation. Imputed values are highlighted in blue.

**NULL’S Height:** 4 NULL values. Imputed using Multivariate SVD Imputation. Imputed values are highlighted in blue.

**NULL’S Bats:** 1 NULL. Set to L

**NULL’S Throws:** 1 NULL. Set to L

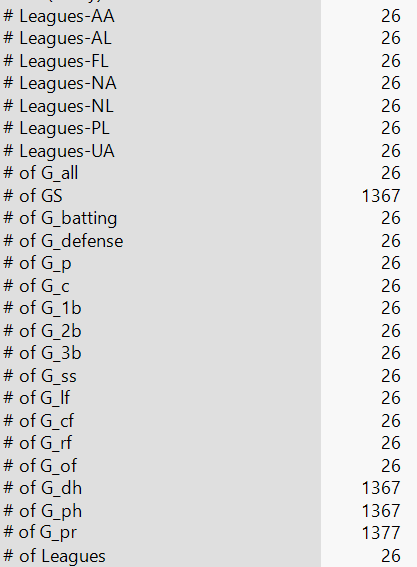
**NULL’S in Awards:** Below screenshot shows number of rows with NULLs in all the Awards variables.



2095 players did not receive any award. Therefore, these NULL’s are recoded to 0.

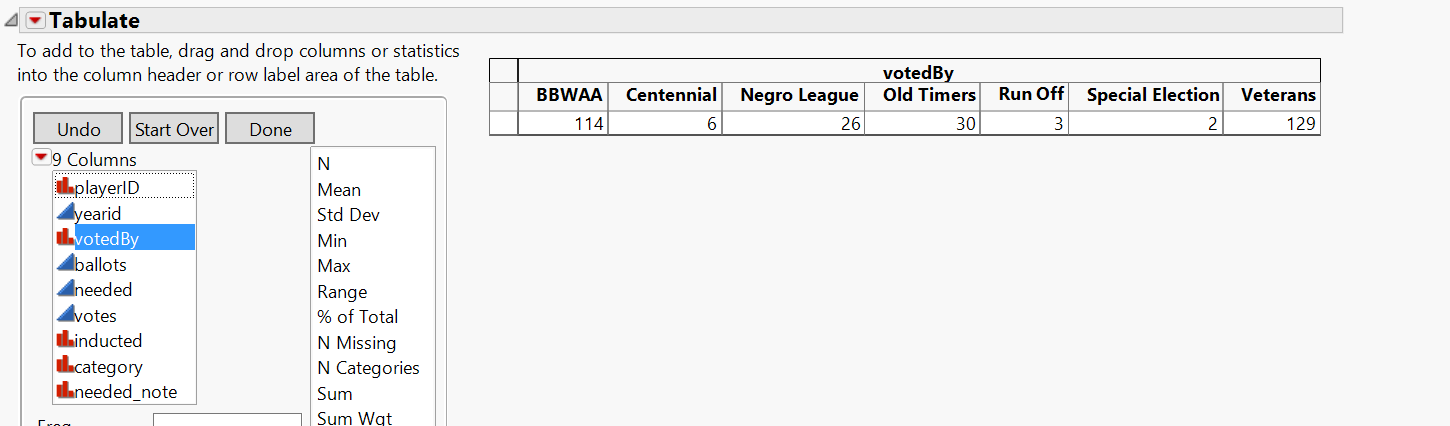
Similarly recoding the 2027 NULLS for HOF (inducted, N), HOF (inducted, Y) to 0. These players were not inducted into Hall of Fame.

**NULL’S in Appearances:**



We found that many variables belonging to Batting, Pitching and Fielding tables had missing values for many players. We recoded missing values with 0 for those players who are only batters and no pitching/fielding variables are applicable to them. Similarly, we recoded missing values with 0 for only pitchers and only fielders.

Out of 245 inductee records, Batting, Pitching and Fielding statistics were not available for 26 players but are inducted into Hall Of Fame. These 26 players play for Negro League. Since, predicting Negro League players without their batting, pitching and fielding statistics would be difficult, we deleted these rows from the dataset. Below is the screenshot from the Hall Of Fame table.



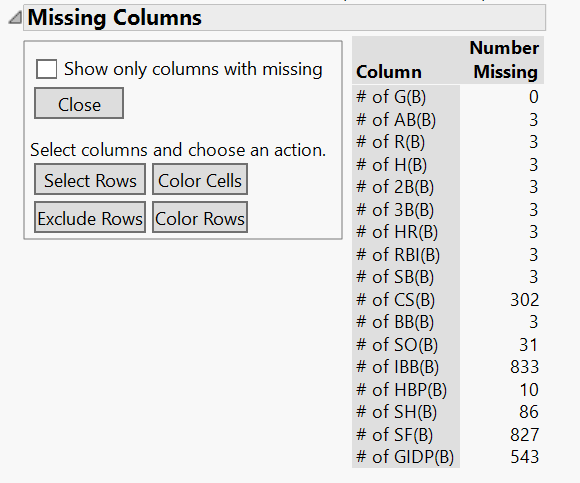
**Updated Data set Summary:**

Number of records in dataset: 3099

Number of inductees: 219

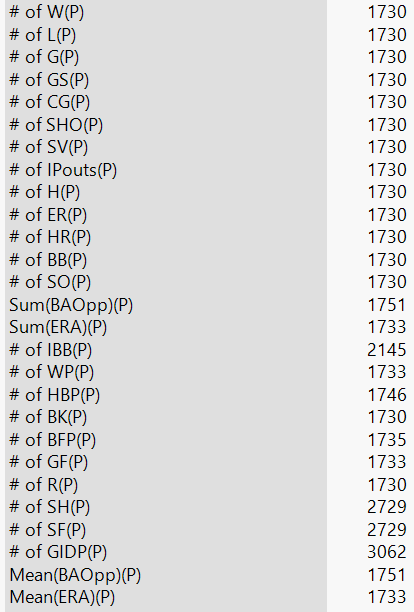
Number of non-inductees: 2880

**NULL’S in Batting Statistics:**



There are 3 players who never batted and NULL’s were formed after joining the MASTER table with the BATTING table. These values are re-coded to 0.

**NULL’S in Pitching statistics:**



There are 1730 players who are not pitchers and NULL’s were formed after joining the MASTER table with the PITCHING table. There values are recoded to 0.

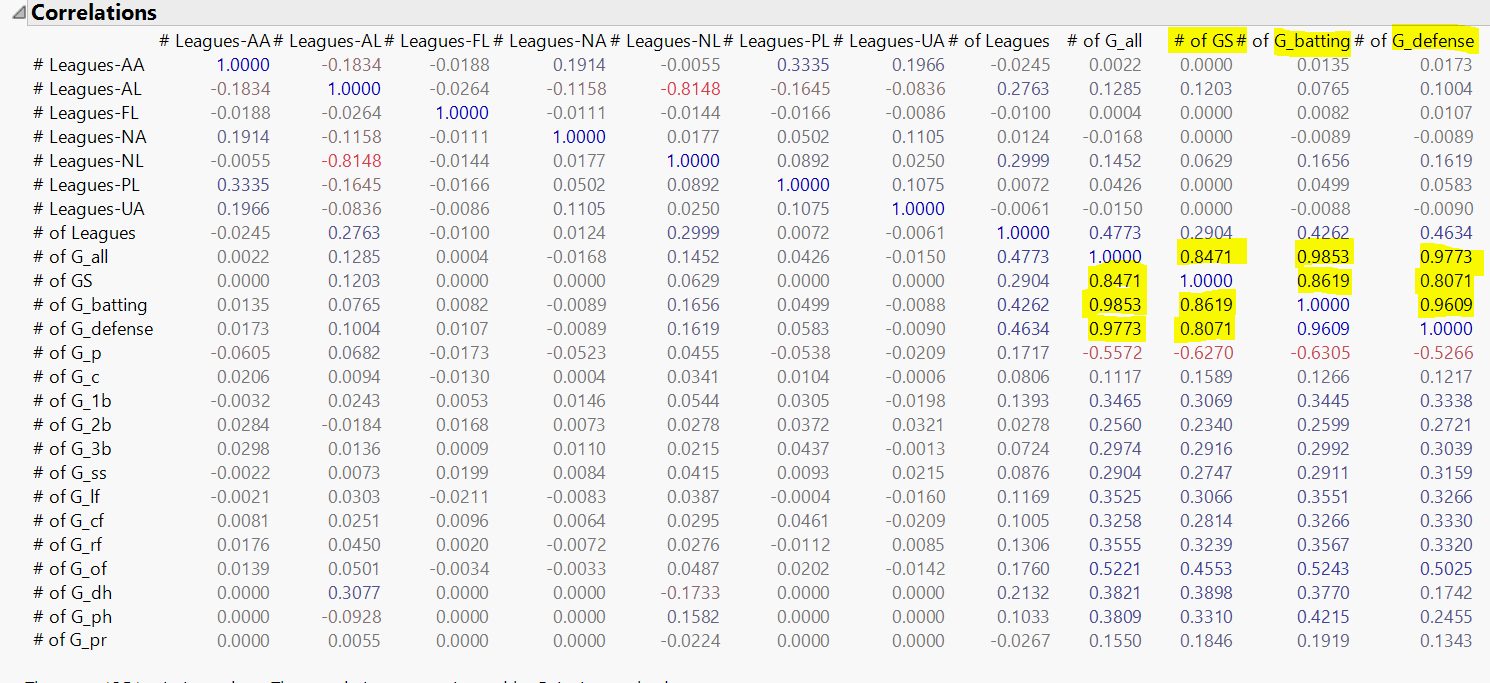
We have also removed the records of managers who were also inducted into Hall of Fame as our business problem is to predict only players.

### **Reducing dimensionality:**

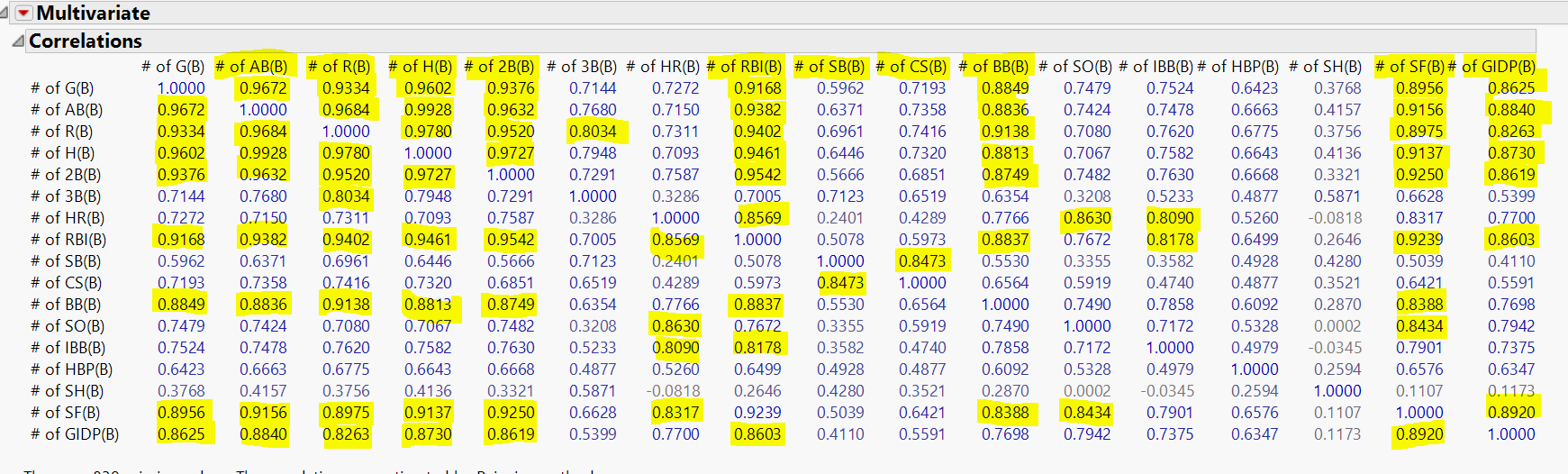
Dimensionality was reduced by following the below steps:

* A Logistic Regression was made using all the columns to identify the effect of each variable.
* Between a highly correlated pair, the least important variable (from the regression model) was identified and is excluded from the data set.

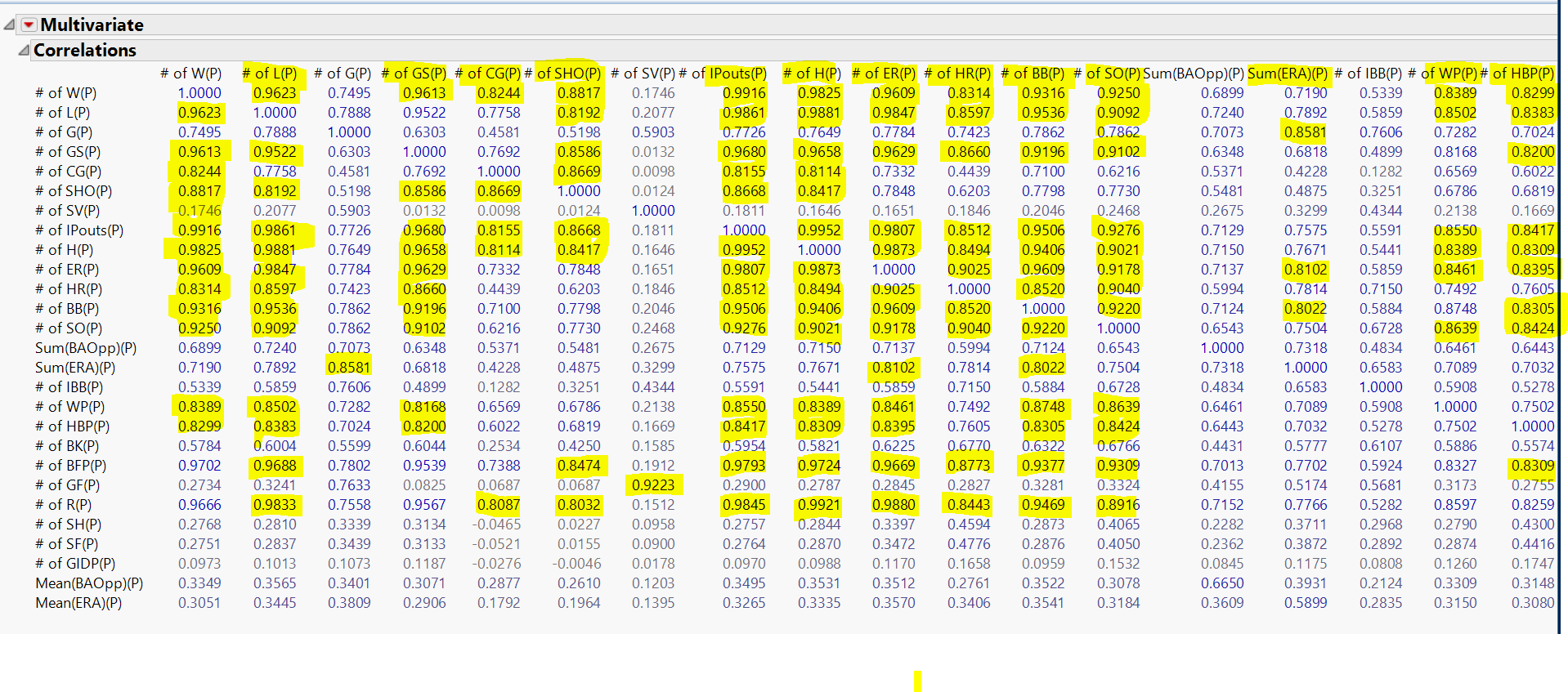
Appearances:

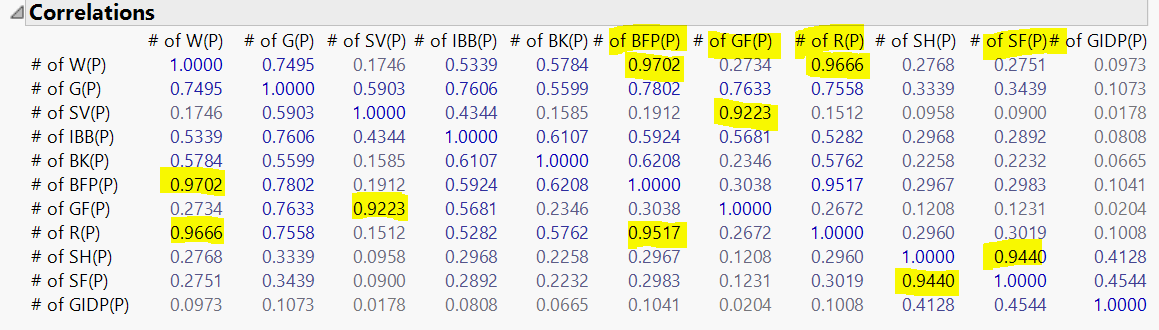


Batting data:

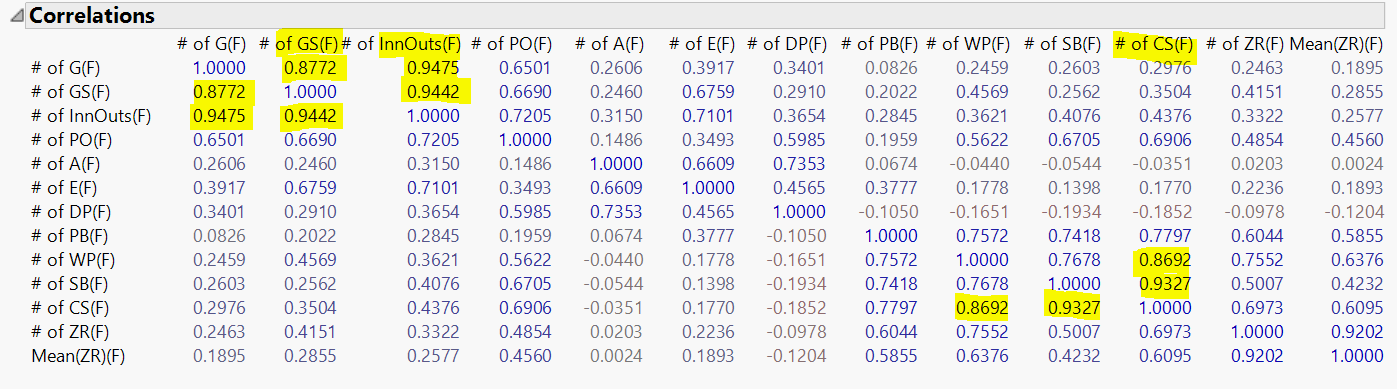


Pitching:

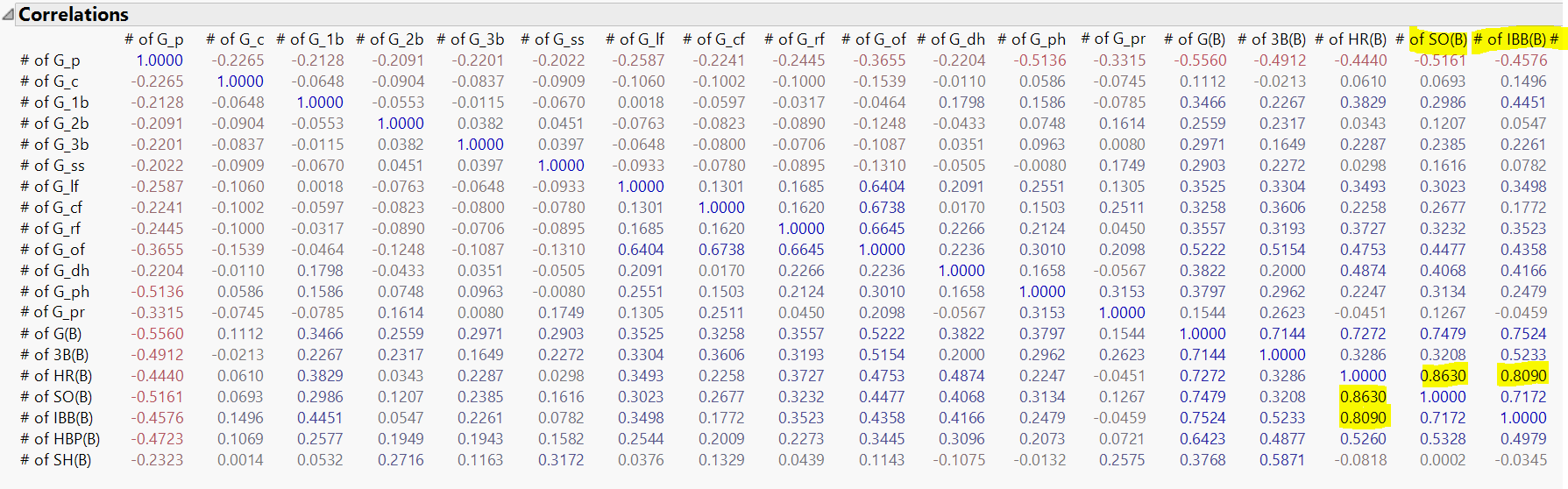


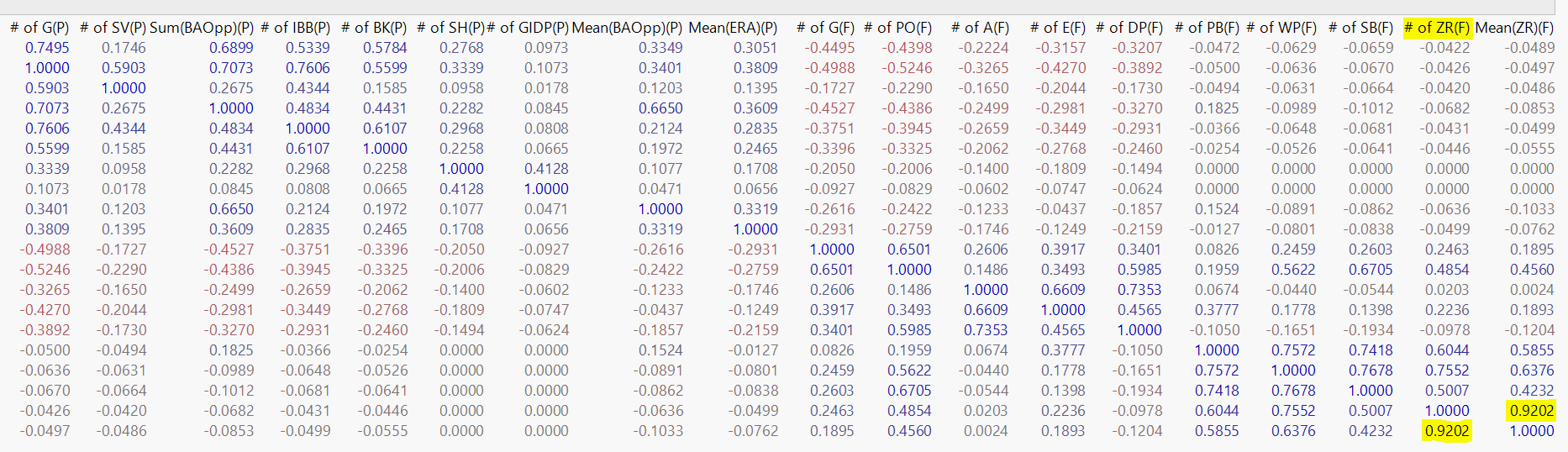


Fielding:

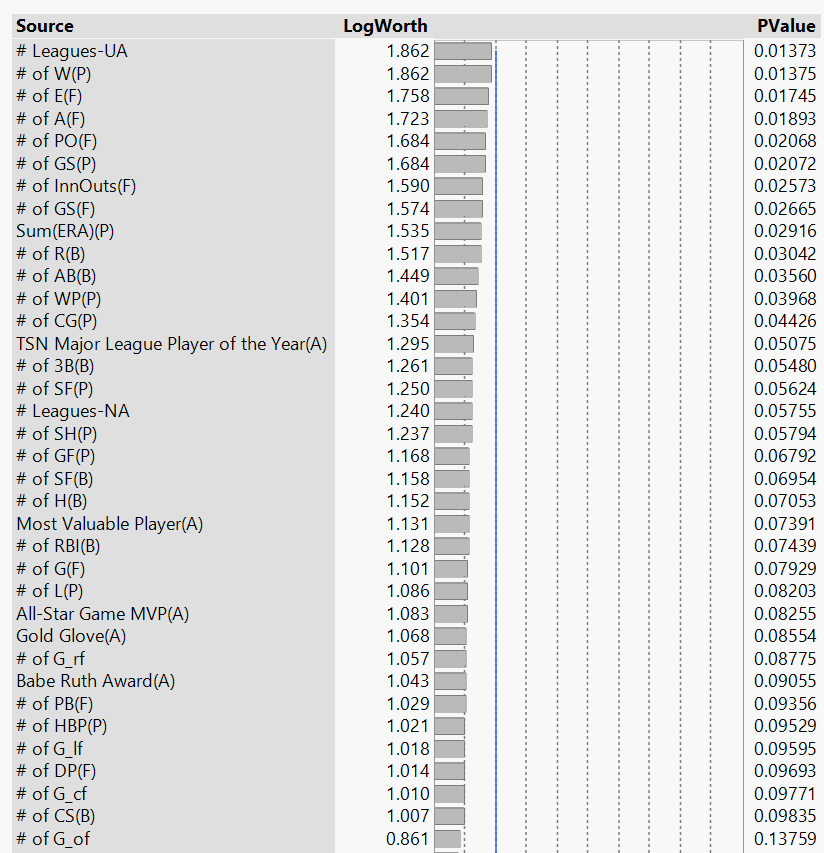


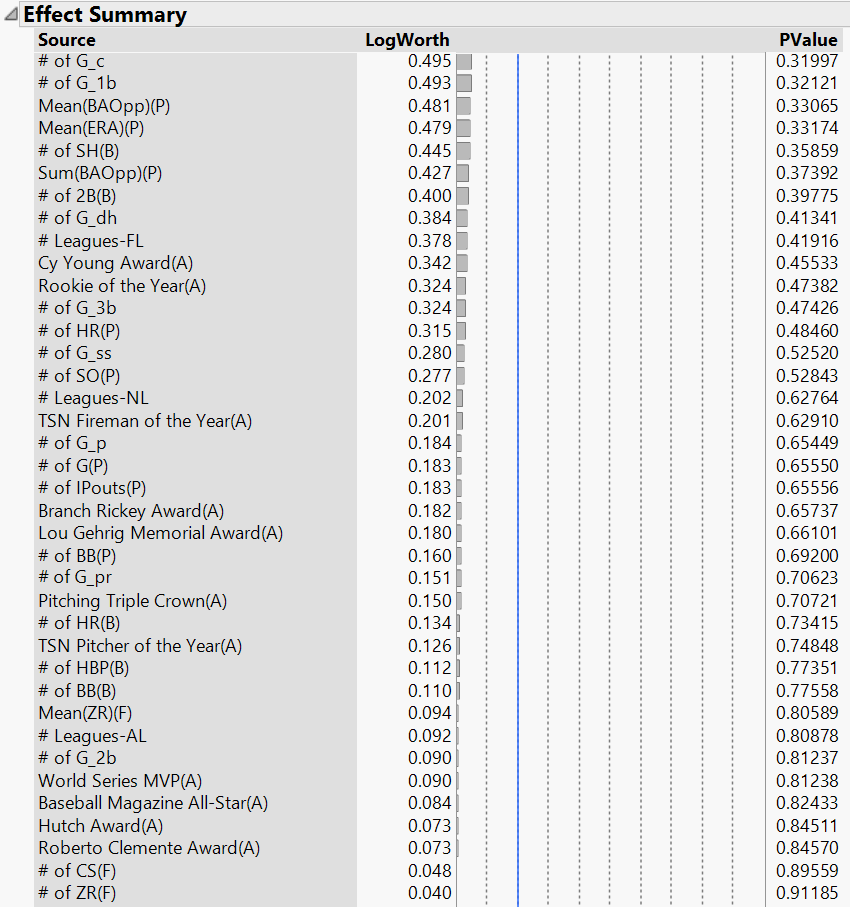
2nd cycle:





Effect of each variable in the Logistic Regression model.



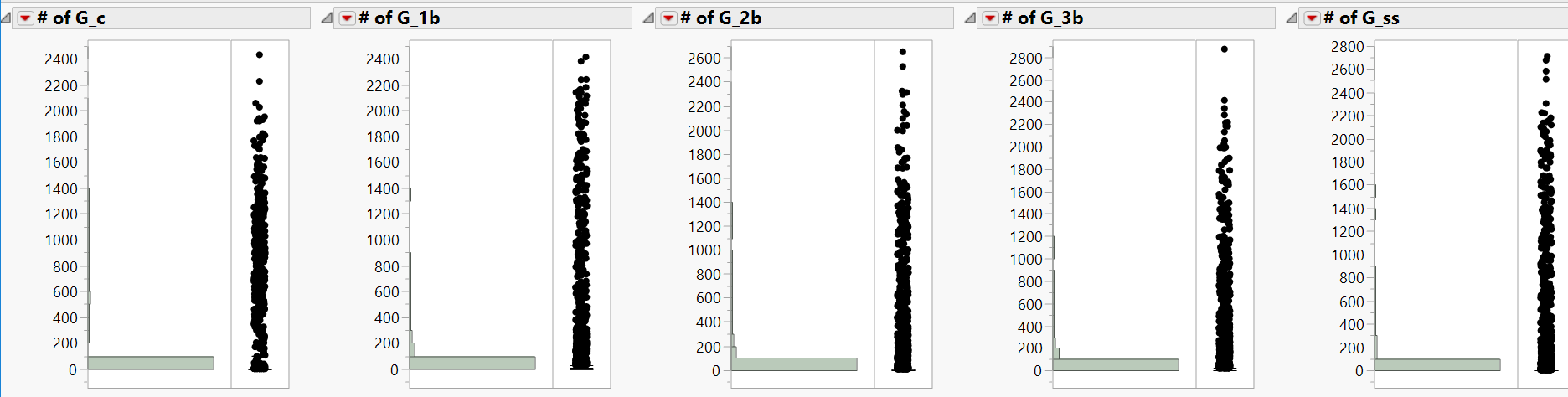


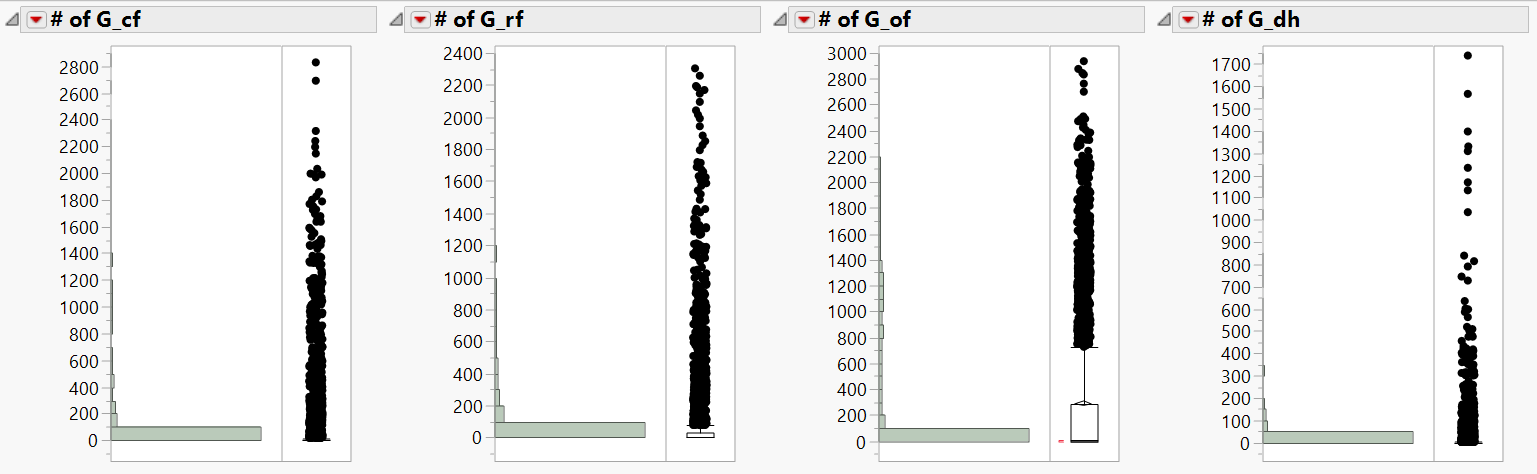
### **Adding Derived variables:**

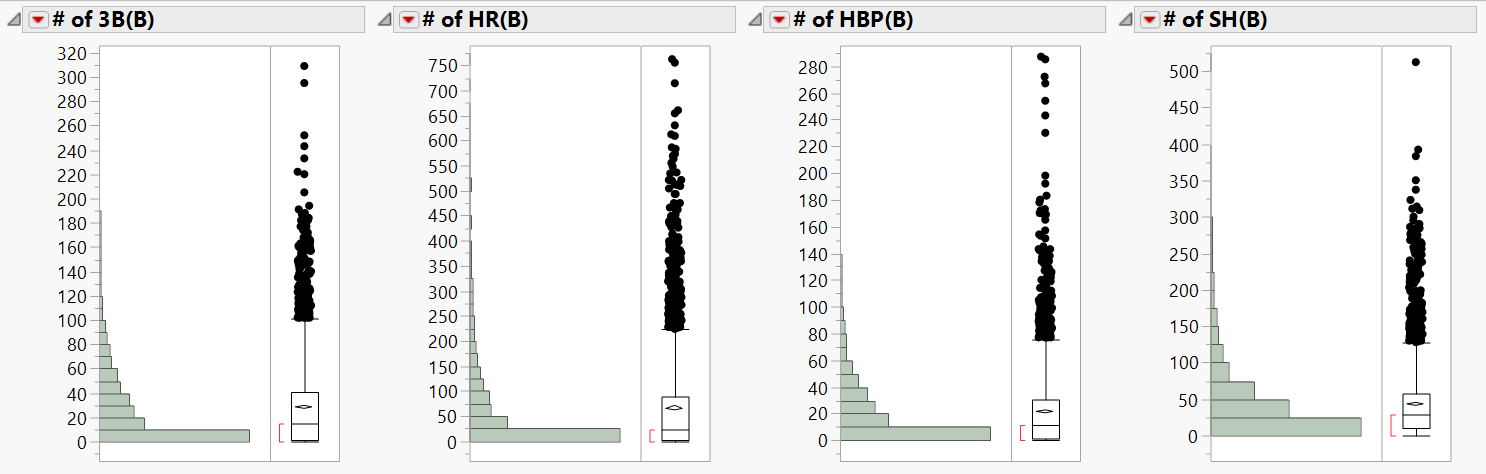
We added new variables, such as # of Awards and Years of Experience for each player and derived values from existing variables.

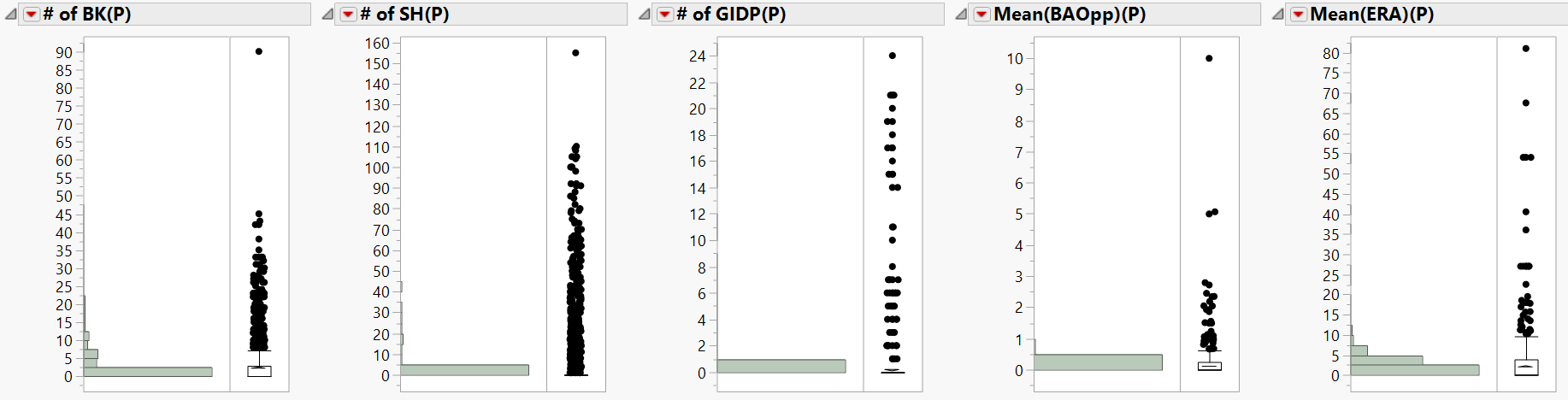
### **Analyzing potential outliers:**

Most of the variables have outliers and they are strongly right skewed as shown below.











More than 50% of the records in each variable are 0’s. Although, many variables have outliers, they are valid records.

**Variable Transformation:**

Since we had to replace missing values with 0, distribution of the variables were right skewed in most cases. We tried dealing with the issue by performing Log(x+0.1) transformation, which did not aid in getting the issue resolved.

Instead of applying a transform function like LOG, Square root etc., to reduce the skewness, new baseball statistics were created using the below formulae.

**On Base Percentage (OBP):**

Measure on how a batter reaches base

OBP =

**Slugging Percentage (SLG %):**

Measures power of a hitter

SLG% =

**On Base plus Slugging (OPS):**

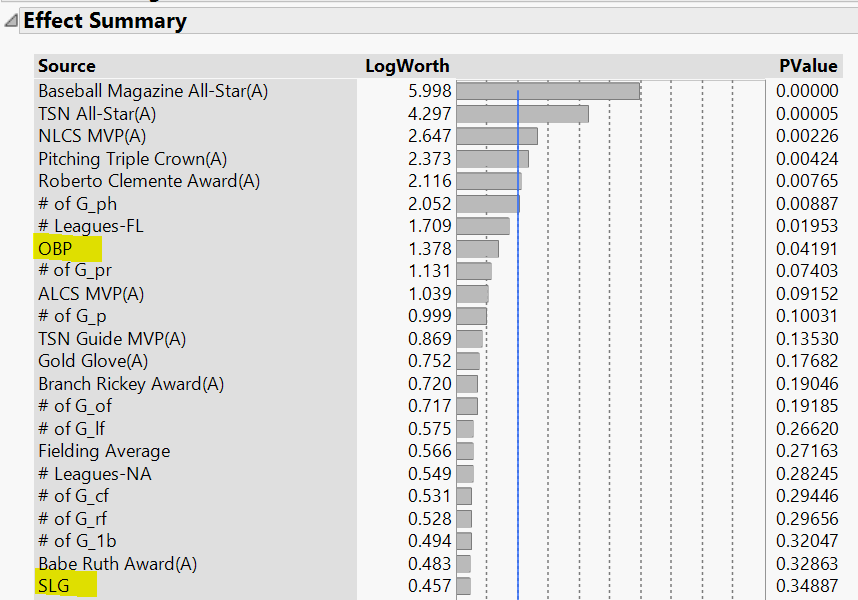
OPS = OBP + SLG %

**Fielding Independent Pitching (FIP):**

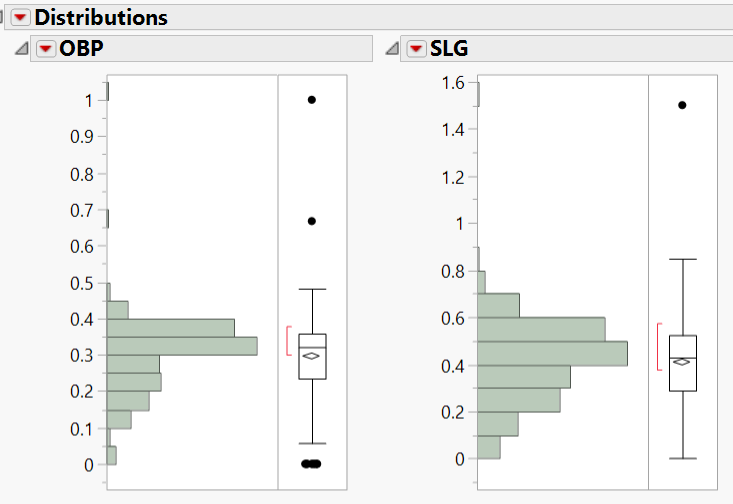
Includes Statistics the pitched has total control over.

FIP =

The below screenshot shows the strong effect of OBP and SLG (stronger than other batting statistics) in the regression model.



**Distributions of OBP and SLG:**

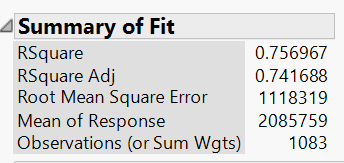
As shown below, the new statistics have reduced skew and are near to normal.

### **NULL’S in SALARY:**

Salary information is not present for players before 1985. Instead of imputing the NULL values with a statistic like Mean, we made a regression model which is trained and validated on the players who have salary information and we used the same formula to impute NULL’S for the players without salaries.

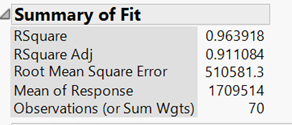
### **Model: Predicting Salaries**

### **Regression Model:**

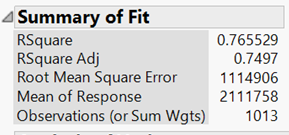


Our dataset has batting, fielding and pitching statistics for all the players. Only pitchers will not have batting statistics, only batters will not have pitching statistics etc. Therefore, data is clustered into 2 using K-means clustering and then we modeled separately for each cluster.

### **Regression model of Cluster 1:**

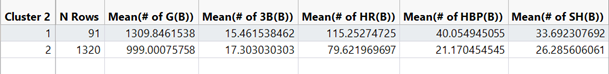


### **Regression model for Cluster 2:**



Model is more accurate for Cluster 1. A summary table is calculated to identify the type of players in each cluster.

Batting statistics for Cluster 1 and Cluster 2:

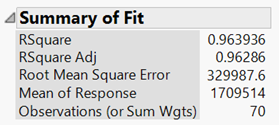


Pitching statistics for Cluster 1 and Cluster 2:

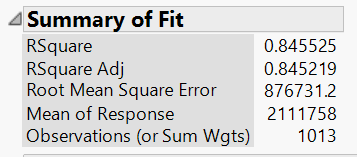


As shown in the above summary tables, players who belong to Cluster 1 do not pitch and are only batters.

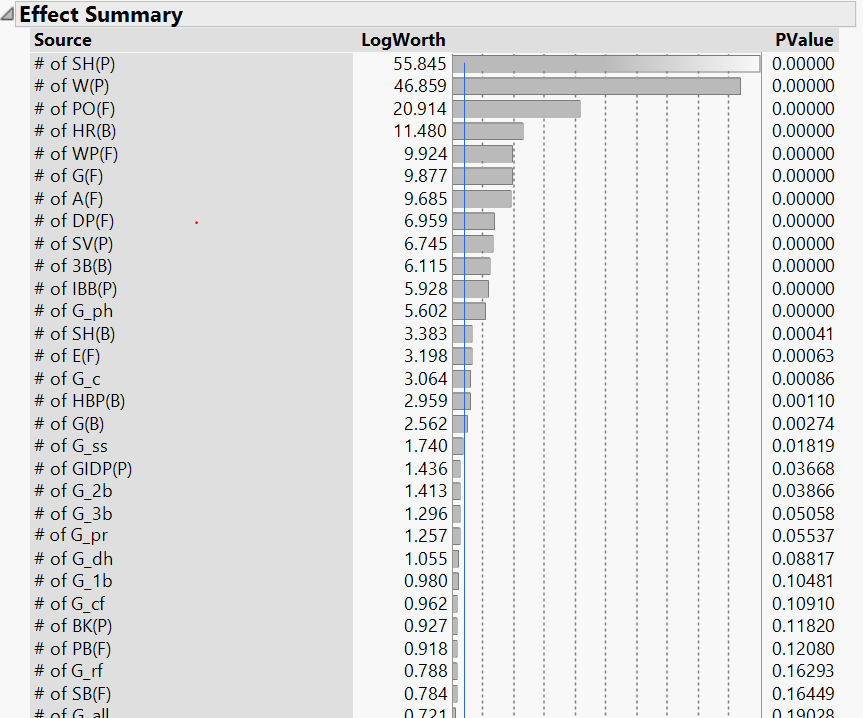
### **Ensemble model:** **Linear Regression Model for Cluster 1**



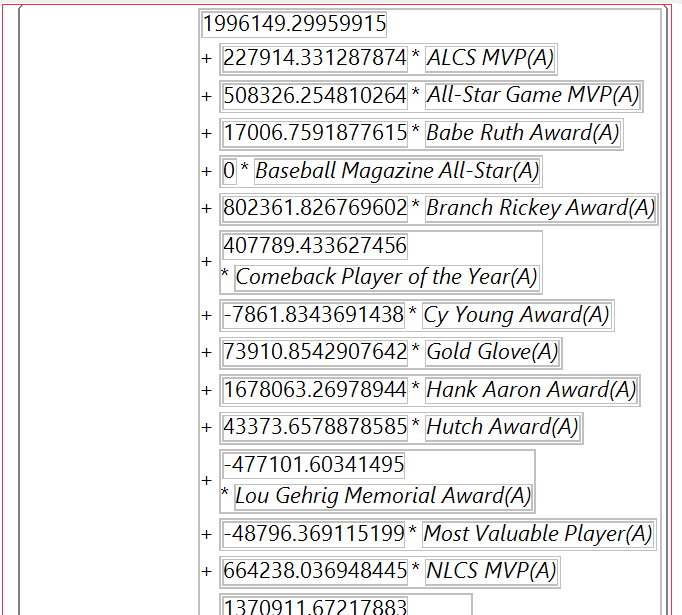
### **Ensemble Model: Linear Regression model for Cluster 2**



Factors determining the Salaries:



The below regression formula is used to impute salaries for the players who does not have salary data.



## **MODEL:**

As our problem statement deals with predicting a nominal variable, i.e. a player will be inducted into hall of fame or not, we tried creating models using Logistic Regression, Neural Networks, Decision Trees, Ensemble models using Logistic Regression and Neural Networks.

Balanced dataset:

Includes 214 inductees and 405 non-inductees.

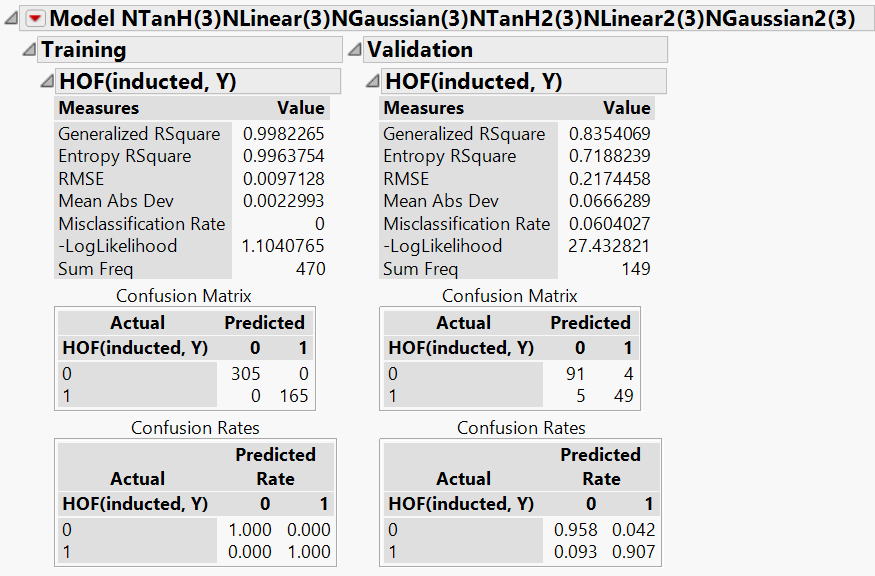
Total number of records: 619

75 %-> Training

25 %-> Validation

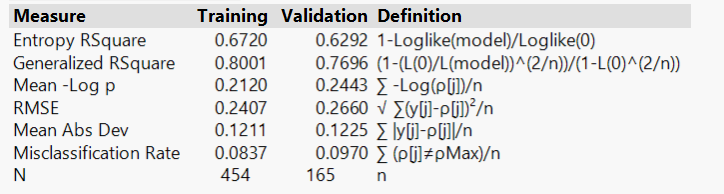
1. **Neural Nets:**

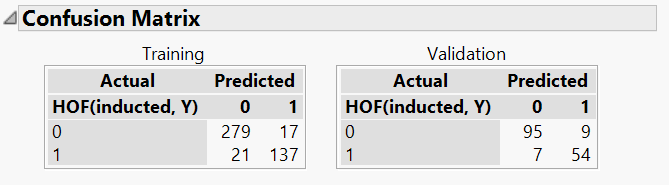
Salary information is not available before 1985. Therefore, salary is not included for modeling.



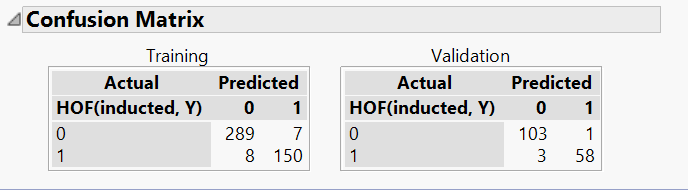
1. **Regression Model:**

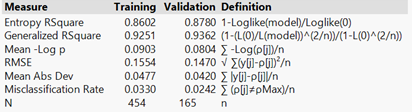
Salary is not considered for modeling.



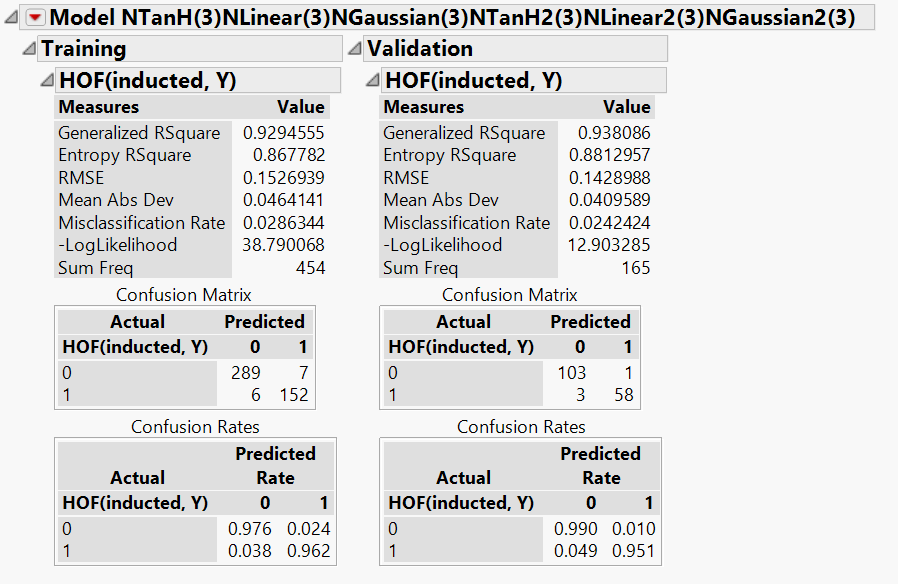


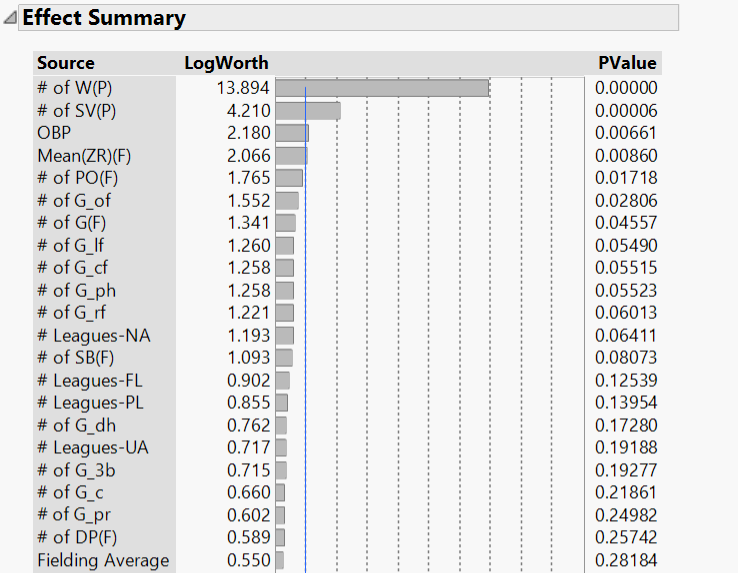
1. **Ensemble model 1: Logistic Regression**



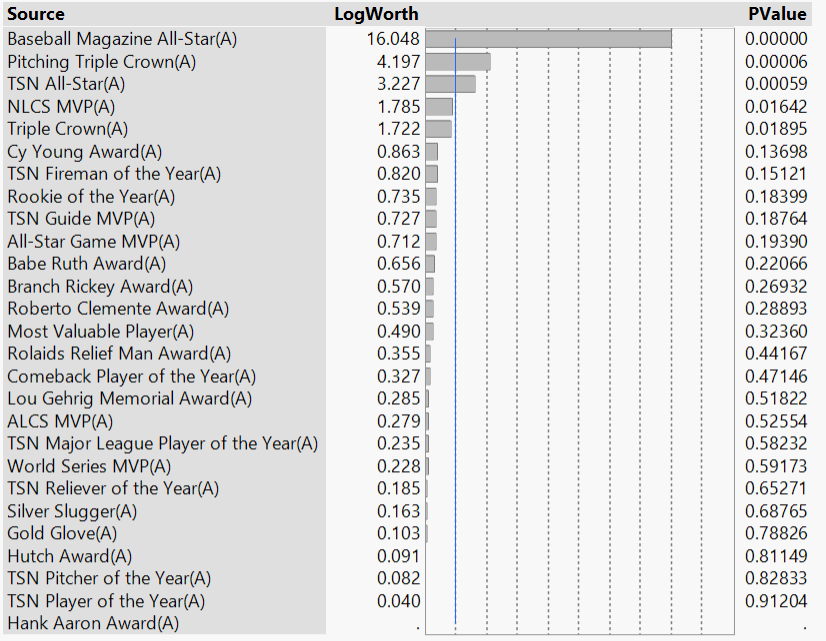


1. **Ensemble Model 2: Neural Nets**



Important Statistics for Hall Of Fame:

Important awards for Hall Of Fame:



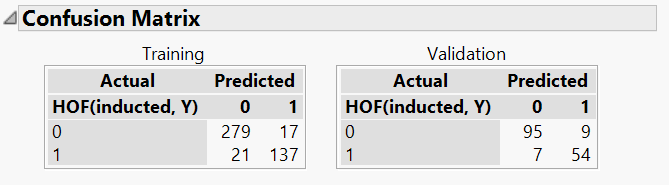
## **ASSESS:**

To select the best model to predict whether a player will be inducted to the hall of fame, we need to compare the confusion matrices for Initial Regression Analysis, Neural Net Analysis, Decision Trees and Ensemble models developed using Linear Regression and Neural Net Analysis.

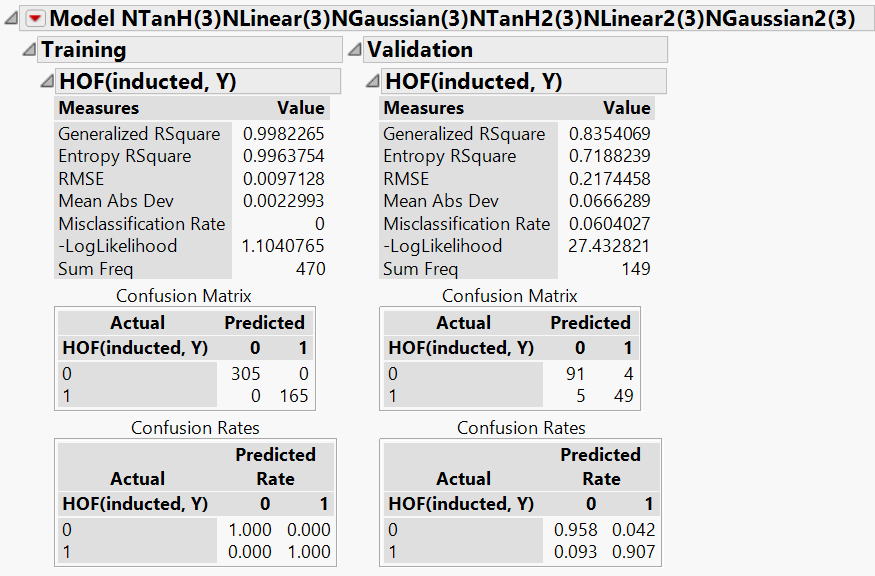
Below are the factors that need to be considered while selecting the best model.

1. Model should be less complex.
2. Model should be easily explained.
3. Should incur relatively lesser total cost.
4. Model should perform the same for both Training and Validation.
5. Should incur less cost to build and maintain the model.

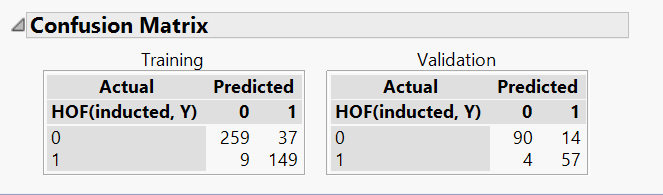
**Logistic Regression:**



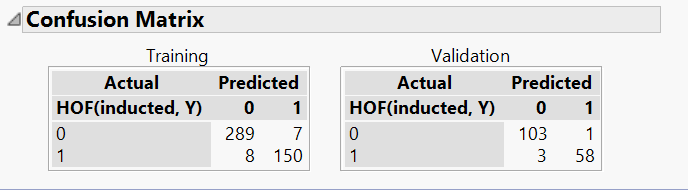
**Neural Net Analysis:**



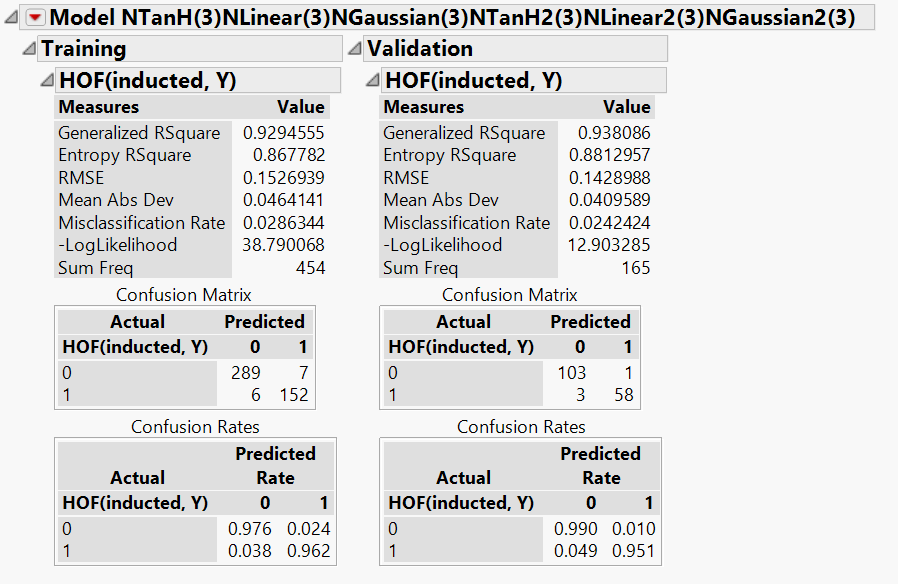
**Decision Tree Analysis:**

****

**Ensemble model using Logistic Regression:**



**Ensemble model using Neural Net Analysis:**



We can evaluate the models using 3 parameters.

**Overall Accuracy of the model:**

**Accuracy= ((Inducted, Inducted) + (Not Inducted, Not Inducted))/Total Number of Predictions**

**FALSE POSITIVE:** Predicting a player will get inducted into hall of fame when he is not actually inducted**.**

*False Positive probability = (Not Inducted, Inducted)/Total number of Inducted predictions*

Unit cost of wrong prediction\* Probability of Error \* Number of incorrect predictions

**FALSE NEGATIVE:** Predicting a player will not get inducted when he gets inducted into hall of fame.

*False Negative probability = (Inducted, Not Inducted)/Total number of Not Inducted predictions*

Unit cost of wrong prediction \* Probability of Error \* Number of incorrect predictions

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Training** | | | **Validation** | | | |
| **Accuracy** | **False Positive** | **False Negative** | | **Accuracy** | **False Positive** | **False Negative** |
| Neural Networks | 100% | 0% | 0% | | 94% | 8% | 9% |
| Logistic Regression | 90% | 11% | 18% | | 90% | 14% | 11% |
| Decision Trees | 90% | 20% | 6% | | 89% | 20% | 7% |
| Ensemble Model using Logistic Regression | 97% | 4% | 5% | | 98% | 2% | 5% |
| Ensemble Model using Neural Networks | 97% | 4% | 4% | | 98% | 2% | 5% |

Probabilities computed based on the above formulas and confusion matrices.

**Logistic Regression vs Decision Tree Analysis vs Neural Net Analysis**

Based on the above probability calculations, it could be seen that among the initial standalone models, Neural Networks offers better accuracy than Logistic Regression and Decision Trees, for both Training and Validation data. In case of False Positive error for validation data, Decision Tree has the least probability of occurrence than Logistic Regression and Neural Networks. To the Contrary, In case of False Negative error for validation data, Decision Tree has the highest probability of occurrence than Logistic Regression and Neural Networks. Overall, Neural Networks seems to offer better accuracy. However, since Neural Networks cannot be easily explained because it involves layers of functions, it is better to select Logistic Regression, which provides relatively better accuracy than Decision Trees.

**Ensemble Model using Logistic Regression vs Neural Networks**

Among the Ensemble models, both Logistic Regression and Neural Networks offer the same level of accuracy. Besides, both the models have the same probability of False Positive and False Negative error occurring. Again, since neural networks cannot be easily explained, we can select Ensemble Model developed using Logistic Regression.

**Initial Standalone models vs Ensemble models**

When evaluating the performance of standalone vs Ensemble models, we can consider the case of Logistic Regression alone for standalone models, as it stands better among the initial models. We know that standalone Logistic Regression offers an accuracy of 90% and Ensemble Logistic Regression model offers an accuracy of 98%. Despite the fact that Ensemble model offers better accuracy than Logistic Regression, since the Ensemble model is relatively complex and involves more cost to build and maintain and involves Neural networks, standalone Logistic Regression model can be selected.  
  
***Note***: Cost benefit analysis is not possible in this case as arriving at a cost of wrong prediction may vary depending on the usage of prediction. Hence, we have only considered the probability of each type of error occurring and the number of incorrect predictions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Complexity** | **Cost to build & maintain** | **Explicable** | **Overfitting** | **Accuracy** |
| Neural | **-** | **+** | **-** | **-** | **+** |
| Logistic | **+** | **+** | **+** | **+** | **+** |
| Decision | **+** | **+** | **+** | **-** | **-** |
| Ensemble - Logistic | **-** | **-** | **-** | **+** | **+** |
| Ensemble Neural | **-** | **-** | **-** | **+** | **+** |

**FINAL RESULT:** Based on the above analysis, it can be concluded that Initial Logistic Regression model performs better for predicting whether a player will be inducted into baseball hall of fame. However, we tried to corroborate our results by performing text mining and sentiment analysis by extending our modeling scope.

**BUSINESS VALUE:** Arriving at cost benefit analysis is not feasible in this case, as awards or honors are generally a motivation factor for existing/new players, we could not identify a cost associated with it. Our model to predict players inducted to Hall of Fame offers the below advantages which cannot be quantified.

For sponsors looking for players to endorse their products, this model might serve as a platform for identification of such players.

For people interested in forming new leagues for All-Start/Fantasy leagues, this model gives a statistical edge in selecting players.

## **TEXT MINING:**

We have used Text Mining to assess the validity of our model using a sample case.

The model predicts that Mike Piazza (a Major league base catcher who mostly played for the New York Mets and the Los Angeles dodgers) will be entering the hall of fame in the years to come.

We thought of using Text Mining as a means of adding credibility to the prediction of our model.

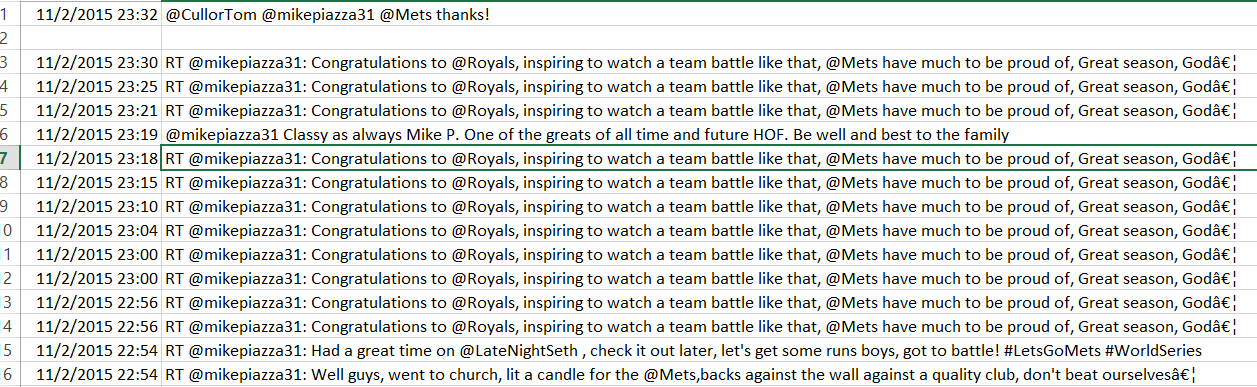
Hence, we thought of collecting social media data from Twitter and analyzing the trends, sentiments and popular words in all of those tweets, in order to understand the real world estimations of Mike Piazza’s skills and his chances to enter the hall of fame.

### **Step 1: Extract twitter data**

Extracting data from Mike Piazza’s twitter profile using the twitter API.

We have used the SEARCH API offered by Twitter to get data from Mike Piazza’s twitter handle [**@mikepiazza31**](https://twitter.com/mikepiazza31)to fetch tweets . The twitter API can only go back two weeks in time period and has a limit of 50,000 tweets.

The programming language Python was used to implement the API and fetch the data. The final form of the data is in the form of a CSV file



### **Step 2: Build word cloud**

Building a word cloud to understand what the document majorly consists of. The word cloud was built using [www.wordle.net](file:///C:\Users\raviteja\Desktop\www.wordle.net)



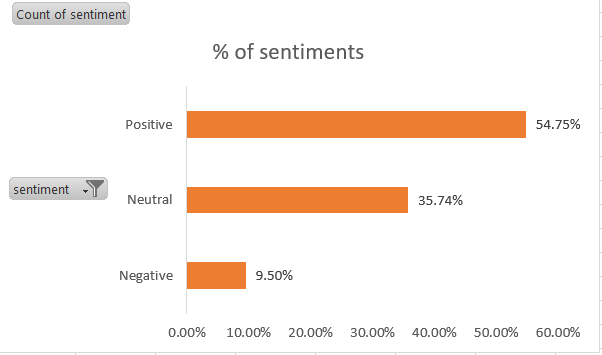
From the word cloud we can see the word “Cooperstown” is the most biggest word in terms of size. Hence we can say that it is the most commonly occurring word in the tweets.

Upon cross-checking we find that cooperstown , New York is where the baseball hall of fame is located and people have used the hashtag #cooperstown to indicate that Mike is favorite to enter the hall of fame.

### **Step 3: Perform Sentiment Analysis**

Sentiment analysis of Mike’s Tweets.

Upon doing a sentiment analysis of Mike’s tweets. We get positive, negative and neutral category of tweets.



We see that 55% of Mike’s tweets are positive indicating that the general opinion of Mike amongst the public is positive and that he is a good baseball player. This corresponds to his increased chances of entering the hall of fame.

Based on the positive sentiments, we try to extract the most commonly occurring positive words and we create a word cloud out of it.



Based on the two word clouds and the amount of positive sentiment associated with Mike Piazza, we can say that he is a strong candidate to enter the hall of fame and that is what our model predicts too.

## **TIME SERIES FORECASTING:**

We attempted a time series forecasting to predict the Home Runs and Batting Average to answer the question: “What If Mike continued to play after 2007? Would he increase or decrease his chances of getting into hall of fame?”

Findings: Our forecast says, had he played after 2007, he would have diminished his chances of hitting the Hall of fame

CHALLENGES – FORECASTING

* Moving Average forecast – Biased to the most recent data points. Not very robust
* Difficult to identify the cycle for predictions
* Adjusted R-squared was negative – as they there were too few data points

## **CONCLUSION:**

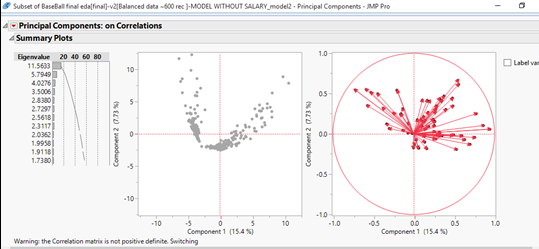
* We have followed the SEMMA approach quite meticulously and have adopted text mining and forecasting methods to further increase our understanding of the model results
* As next steps, we will perform the text mining analysis for all players, who have been predicted to be 1 by our model
* We will also forecast the batting, pitching and fielding statistics using a different and more robust forecasting method

## **APPENDIX:**

**Methods that did not work:**

* **Dimension reduction using PCA:**

We tried Principal Component Analysis to identify variables that would contain most of the information to reduce dimensionality.



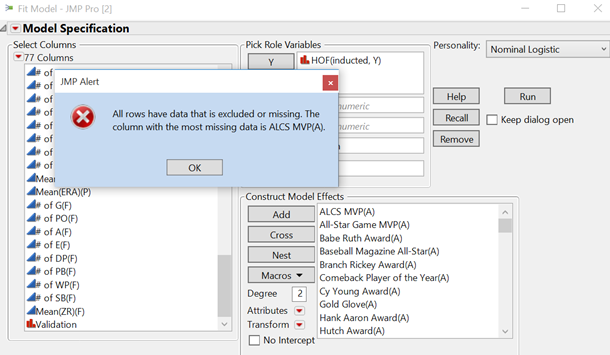
From the Eigen values, it could be seen that out of 73 variables, 57 variables are needed to explain 99% variability. Also, adding required number of principal components, would in turn reduce the explanatory power of variables as each principal component would be a linear expression of all these variables. Hence, we did not follow the principal component approach as it makes it hard to identify the factors that influence the prediction, with the addition of principal components.

**Eigenvalues**

| **Number** | **Eigenvalue** | **Percent** |  | **Cum Percent** |
| --- | --- | --- | --- | --- |
| 1 | 11.5633 | 15.418 |  | 15.418 |
| 2 | 5.7949 | 7.727 |  | 23.144 |
| 3 | 4.0276 | 5.370 |  | 28.514 |
| 4 | 3.5006 | 4.667 |  | 33.182 |
| 5 | 2.8380 | 3.784 |  | 36.966 |
| 6 | 2.7297 | 3.640 |  | 40.605 |
| 7 | 2.5618 | 3.416 |  | 44.021 |
| 8 | 2.3117 | 3.082 |  | 47.104 |
| 9 | 2.0362 | 2.715 |  | 49.818 |
| 10 | 1.9958 | 2.661 |  | 52.480 |
| 11 | 1.9118 | 2.549 |  | 55.029 |
| 12 | 1.7380 | 2.317 |  | 57.346 |
| 13 | 1.5430 | 2.057 |  | 59.403 |
| 14 | 1.4707 | 1.961 |  | 61.364 |
| 15 | 1.4146 | 1.886 |  | 63.250 |
| 16 | 1.3142 | 1.752 |  | 65.003 |
| 17 | 1.2745 | 1.699 |  | 66.702 |
| 18 | 1.2268 | 1.636 |  | 68.338 |
| 19 | 1.1458 | 1.528 |  | 69.865 |
| 20 | 1.1412 | 1.522 |  | 71.387 |
| 21 | 1.0923 | 1.456 |  | 72.843 |
| 22 | 1.0264 | 1.369 |  | 74.212 |
| 23 | 1.0166 | 1.355 |  | 75.567 |
| 24 | 0.9878 | 1.317 |  | 76.884 |
| 25 | 0.9667 | 1.289 |  | 78.173 |
| 26 | 0.9446 | 1.259 |  | 79.433 |
| 27 | 0.8854 | 1.181 |  | 80.613 |
| 28 | 0.8282 | 1.104 |  | 81.718 |
| 29 | 0.8136 | 1.085 |  | 82.802 |
| 30 | 0.8031 | 1.071 |  | 83.873 |
| 31 | 0.7543 | 1.006 |  | 84.879 |
| 32 | 0.7202 | 0.960 |  | 85.839 |
| 33 | 0.7061 | 0.941 |  | 86.781 |
| 34 | 0.6854 | 0.914 |  | 87.694 |
| 35 | 0.6546 | 0.873 |  | 88.567 |
| 36 | 0.5915 | 0.789 |  | 89.356 |
| 37 | 0.5843 | 0.779 |  | 90.135 |
| 38 | 0.5725 | 0.763 |  | 90.898 |
| 39 | 0.5247 | 0.700 |  | 91.598 |
| 40 | 0.4786 | 0.638 |  | 92.236 |
| 41 | 0.4748 | 0.633 |  | 92.869 |
| 42 | 0.4639 | 0.619 |  | 93.488 |
| 43 | 0.4323 | 0.576 |  | 94.064 |
| 44 | 0.4154 | 0.554 |  | 94.618 |
| 45 | 0.3701 | 0.494 |  | 95.111 |
| 46 | 0.3666 | 0.489 |  | 95.600 |
| 47 | 0.3592 | 0.479 |  | 96.079 |
| 48 | 0.3238 | 0.432 |  | 96.511 |
| 49 | 0.2876 | 0.383 |  | 96.894 |
| 50 | 0.2778 | 0.370 |  | 97.264 |
| 51 | 0.2372 | 0.316 |  | 97.581 |
| 52 | 0.2121 | 0.283 |  | 97.863 |
| 53 | 0.2041 | 0.272 |  | 98.136 |
| 54 | 0.1912 | 0.255 |  | 98.391 |
| 55 | 0.1811 | 0.241 |  | 98.632 |
| 56 | 0.1711 | 0.228 |  | 98.860 |
| 57 | 0.1663 | 0.222 |  | 99.082 |
| 58 | 0.1579 | 0.210 |  | 99.292 |
| 59 | 0.1200 | 0.160 |  | 99.452 |
| 60 | 0.1124 | 0.150 |  | 99.602 |
| 61 | 0.0775 | 0.103 |  | 99.706 |
| 62 | 0.0676 | 0.090 |  | 99.796 |
| 63 | 0.0658 | 0.088 |  | 99.884 |
| 64 | 0.0506 | 0.067 |  | 99.951 |
| 65 | 0.0409 | 0.055 |  | 100.006 |
| 66 | 0.0235 | 0.031 |  | 100.037 |
| 67 | 0.0122 | 0.016 |  | 100.053 |
| 68 | 0.0076 | 0.010 |  | 100.063 |
| 69 | 0.0054 | 0.007 |  | 100.071 |
| 70 | 0.0012 | 0.002 |  | 100.072 |
| 71 | 0.0001 | 0.000 |  | 100.072 |
| 72 | 0.0001 | 0.000 |  | 100.073 |
| 73 | 0.0000 | 0.000 |  | 100.073 |

* **Regression model on Log transformed data:**

Standardized all the columns to the scale [01] and did a log transformation to normalize the data. This method did not work out as the LOG transformation created many NULL values and Logistic regression ignored all the records.



* **LOG(x+0.1) transformation:**

Since the above LOG(x) transformation created many NULL values, LOG(x+0.1) transformation was applied to the data after standardizing it to the scale [0 1]. Though this method reduced the skewness to some extent, regression model is less accurate than the model built with original variables.

* **Model after clustering the data using K-means clustering:**

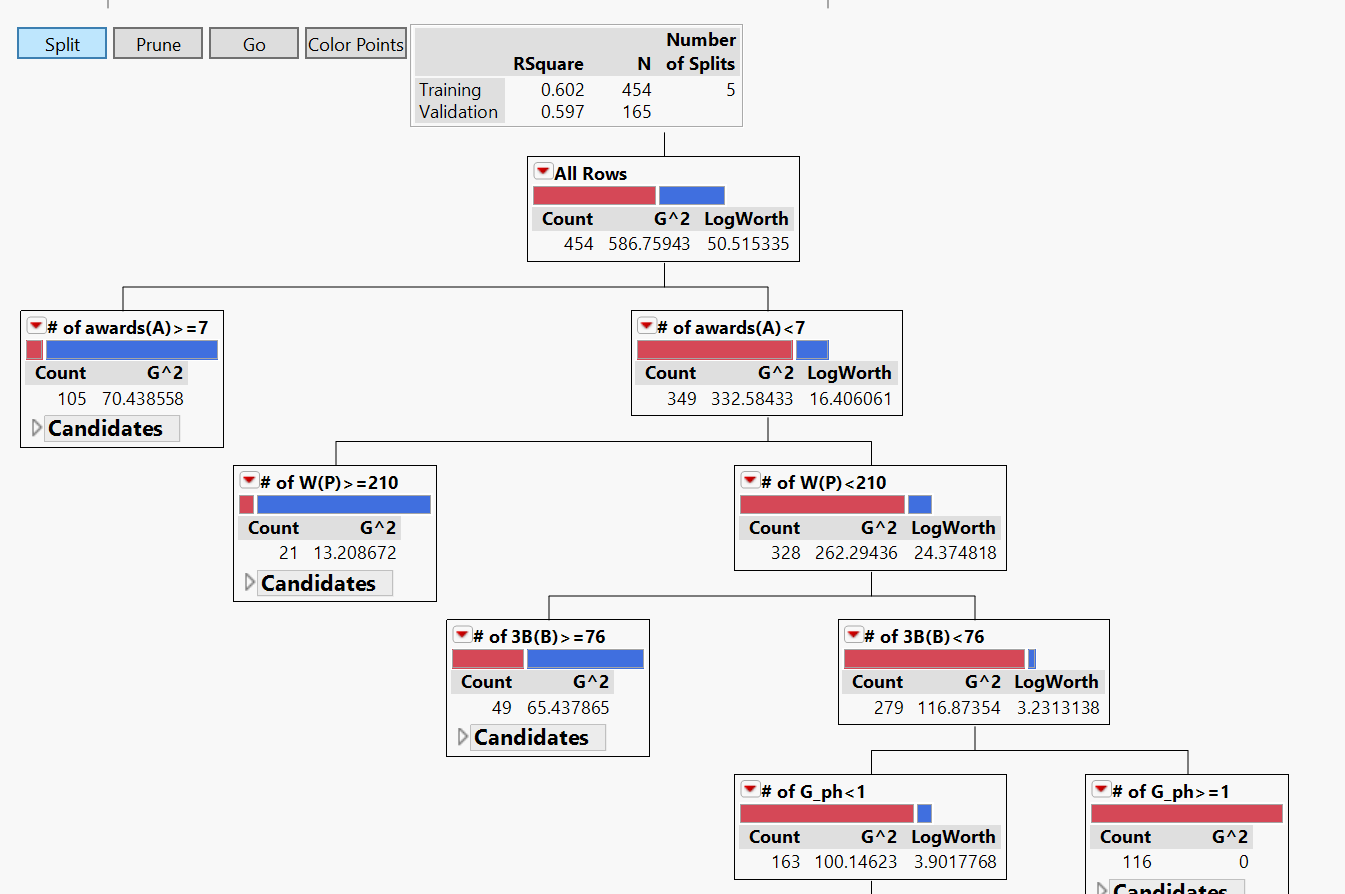
The data is clustered into 2 and modeled separately. But, this model less accurate than the model built on original variables.

* **Separate variables into 0’s and non 0’s and model separately:**

We made an attempt to model separately for 0’s and non 0’s. Log transform for non 0’s. But, since the dataset has around 110 variables we were not able to classify data into 0’s and non 0’s.

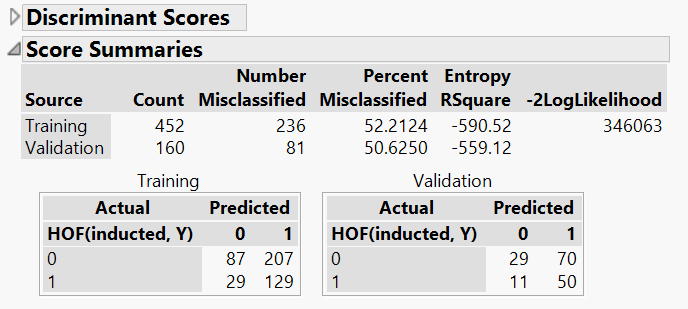
**Decision Trees to predict Hall Of Fame:**

Both RSquare and accuracy of the decision tree is very low.



**Discriminatory Analysis to predict Hall Of Fame:**

Both the R2 and accuracy was very low as shown in the below screenshot.



**Model only using Batting, Pitching, and Fielding and Appearances variables:**

In order to check if AWARDS variables have huge impact on the models, we tried to build a model by not considering awards. As shown below in the confusion matrix, model accuracy has decreased slightly but not much. Therefore, we concluded that batting, pitching and fielding statistics play a major role.

