**Predicting MLB Hall of Fame Selection**

**(Removed the Student’s Name)**

**University of Maryland University College**

**Graduate School of Management and Technology**

**Research Paper**

**For**

**DBST 667 Fall 2014**

**Abstract**

This paper analyzed analyze a MLB Hall of Fame (HOF) database to determine which of the offensive and defensive statistics and associated sabermetrics for position players (eligible as of 2000) to determine which algorithms and which data combinations produces the best identification of players likely to be selected for the baseball HOF. The paper investigated the use of Naïve Bayes, J48 Decision Tree, and Logistics Regression algorithms in the Weka data mining software package. The paper also investigated the impact of removing the total player ranking sabermetrics from the database on the classification accuracy as well as accuracy improvements gained by reducing the attributes set analyzed. In general, logistics regression algorithm was found to provide the best performance and reducing the number of attributes considered provided a significant improvement in accuracy, and to a lesser extent the inclusion of the total player ranking.

Contents

[Introduction 5](#_Toc404525681)

[Background 5](#_Toc404525682)

[MLB Hall of Fame and Selection Process 5](#_Toc404525683)

[Sabermetrics and data mining 6](#_Toc404525684)

[Data analysis and preparation 6](#_Toc404525685)

[Data Description 6](#_Toc404525686)

[Descriptive analysis 7](#_Toc404525687)

[Other data sources 11](#_Toc404525688)

[Data cleaning and preparation. 12](#_Toc404525689)

[Methods 13](#_Toc404525690)

[Choice of a model 13](#_Toc404525691)

[Naïve Bayes classification 14](#_Toc404525692)

[J48 Decision tree 14](#_Toc404525693)

[Logistic regression 15](#_Toc404525694)

[Experimental design 15](#_Toc404525695)

[Analysis and interpretation 16](#_Toc404525696)

[Naïve Bayes classification 16](#_Toc404525697)

[J48 Decision tree 20](#_Toc404525698)

[Logistic regression 24](#_Toc404525699)

[Comparison of Alternative Solutions 28](#_Toc404525700)

[Comparison with Similar Studies 33](#_Toc404525701)

[Observations and conclusions 35](#_Toc404525702)

[References 37](#_Toc404525703)

# Introduction

This paper analyzes the Major League Baseball (MLB) Hall of Fame (HOF) database available for the class project to determine which data mining algorithms and data of offensive and defensive statistics, and associated sabermetrics, for position players (eligible as of 2000), produces the best identification of selected for the baseball HOF. In addition the paper also considers the impact of the presence or absence of one particular sabermetric variable, total player rating (TPR) on the accuracy of the different data mining algorithms evaluated.

# Background

## MLB Hall of Fame and Selection Process

The MLB Hall of Fame is intended to recognize both players who had exceptional careers and achievements, and others associated with baseball, such as executives, managers, and umpires who have made significant achievements and contributions to baseball. Players are eligible for selection to the baseball Hall of Fame if they have played at least 10 seasons and have been retired from the game for at least five years. During their initial period of eligibility, until recently, a 15 year period, they are eligible for selection by the Baseball Writers' Association of America. After that period has ended, they may be selected by the veterans committee for an indefinite period of time, which also considers and selects all non-players for the Hall of Fame. While a player’s record and playing ability, which can be measured by the statistics captured their career, are a major factor in the guidelines for selection to the Hall of Fame, the guidelines also specify that in addition to a player’s performance and contributions to the teams he played for his character, integrity, and sportsmanship, should also be considered (Election Rules, n.d.). This of course allows for an element of subjectivity in the evaluation of candidates on the part of those voting if a player is worthy of inclusion in the Hall of Fame.

## Sabermetrics and data mining

Sabermetrics is a term used to describe the statistical analysis of the performance data collected on baseball players and games. Among its goals is to determine which are the most productive and efficient players, both overall and in specific aspects of the game. The term sabermetrics was coined by Bill James in 1980 and described it as quote the search for objective knowledge about baseball” (Birnbaum, n. d.).

# Data analysis and preparation

## Data Description

The data set used was the data set used by Cochran (2000) for his paper, and includes basic statistics and several sabermetrics statistics on all position players eligible for selection to the baseball Hall of Fame in the year 2000. The data set contains 1340 instances with 24 attributes representing the players offense even defensive statistics for their career, a categorical attribute indicating which position they played (catcher (C), first base (1), second base (2), third base (3), shortstop (S) , outfield (O), or designated hitter (D)) , and a class attribute indicating if they were a member of the MLB Hall of Fame, and if elected by the baseball writers for the veterans committee, or not in the MLB Hall of Fame. A key element of the data set is the sabermetrics measure Total Player Rating (TPR) based on a proprietary formula developed by Thorn and Palmer which attempts to provide the best overall measure of a player’s value over his career.

## Descriptive analysis

Table 1 below shows the numeric attributes in the database and their range of values, while Figure 1 shows all of the attributes distributions. The first 13 attributes, from seasons played to times caught stealing, or the totals for a player’s career. The next four values represent commonly reported averages/percentages calculated from the career totals. The remaining values represent more advanced sabermetrics. This represent weighted combinations of the player’s career totals. A detailed description of how these values are calculated is in Albert’s (2010) paper with the exception of total player rating. Total player rating is a formula developed by Thorn and Palmer, which is a weighted sum of adjusted batting runs, fielding runs and base stealing runs. The formula was adjusted annually and the results were available in their annual baseball encyclopedia, titled Total Baseball, which was last published in 2004. The data is now known as batter – fielder wins, and is only available with an ESPN insiders paid subscription (Yawdoszyn, 2006).

Table 1. Numeric Attributes

| **Attribute** | **Minimum** | **Maximum** | **Mean** | **Standard Deviation** |
| --- | --- | --- | --- | --- |
| Seasons Played | 10 | 26 | 13.48 | 3.136 |
| Games Played | 140 | 3562 | 1331.26 | 519.165 |
| Official At-Bats (AB) | 252 | 14053 | 4534.61 | 2094.191 |
| Runs Scored (R) | 20 | 2246 | 635.31 | 376.41 |
| Hits (H) | 48 | 4256 | 1248.57 | 647.665 |
| Doubles (2B) | 6 | 792 | 203.22 | 116.576 |
| Triples (3B) | 0 | 309 | 50.81 | 41.038 |
| Home Runs (HR) | 0 | 755 | 85.11 | 97.93 |
| Runs Batted In (RBI) | 21 | 2297 | 565.74 | 357.164 |
| Walks (BB) | 17 | 2056 | 445.58 | 295.214 |
| Strikeouts (SO) | 0 | 2597 | 445.69 | 325.319 |
| Stolen Bases (SB) | 0 | 938 | 104.45 | 125.754 |
| Times Caught Stealing (CS) | 0 | 307 | 37.82 | 34.337 |
| Batting Average (BA) | 0.161 | 0.336 | 0.269 | 0.026 |
| On Base Percentage (OBP) | 0.194 | 0.483 | 0.336 | 0.034 |
| Slugging Percentage (SLG) | 0.201 | 0.690 | 0.385 | 0.061 |
| Fielding Average (FA) | 0.820 | 1.000 | 0.966 | 0.025 |
| Adjusted Production (AP) | 20 | 209 | 99.90 | 22.445 |
| Batting Runs (BR) | -310 | 1322 | 37.56 | 169.282 |
| Adjusted Batting Runs (ABR) | -341 | 1355 | 35.26 | 167.630 |
| Runs Created (RC) | 16 | 2838 | 657.08 | 416.119 |
| Stolen Base Runs (SBR) | -31 | 110 | -3.09 | 13.315 |
| Fielding Runs (FR) | -235 | 369 | 5.959 | 63.145 |
| Total Player Rating (TPR) | -28.9 | 105.2 | 3.53 | 15.118 |

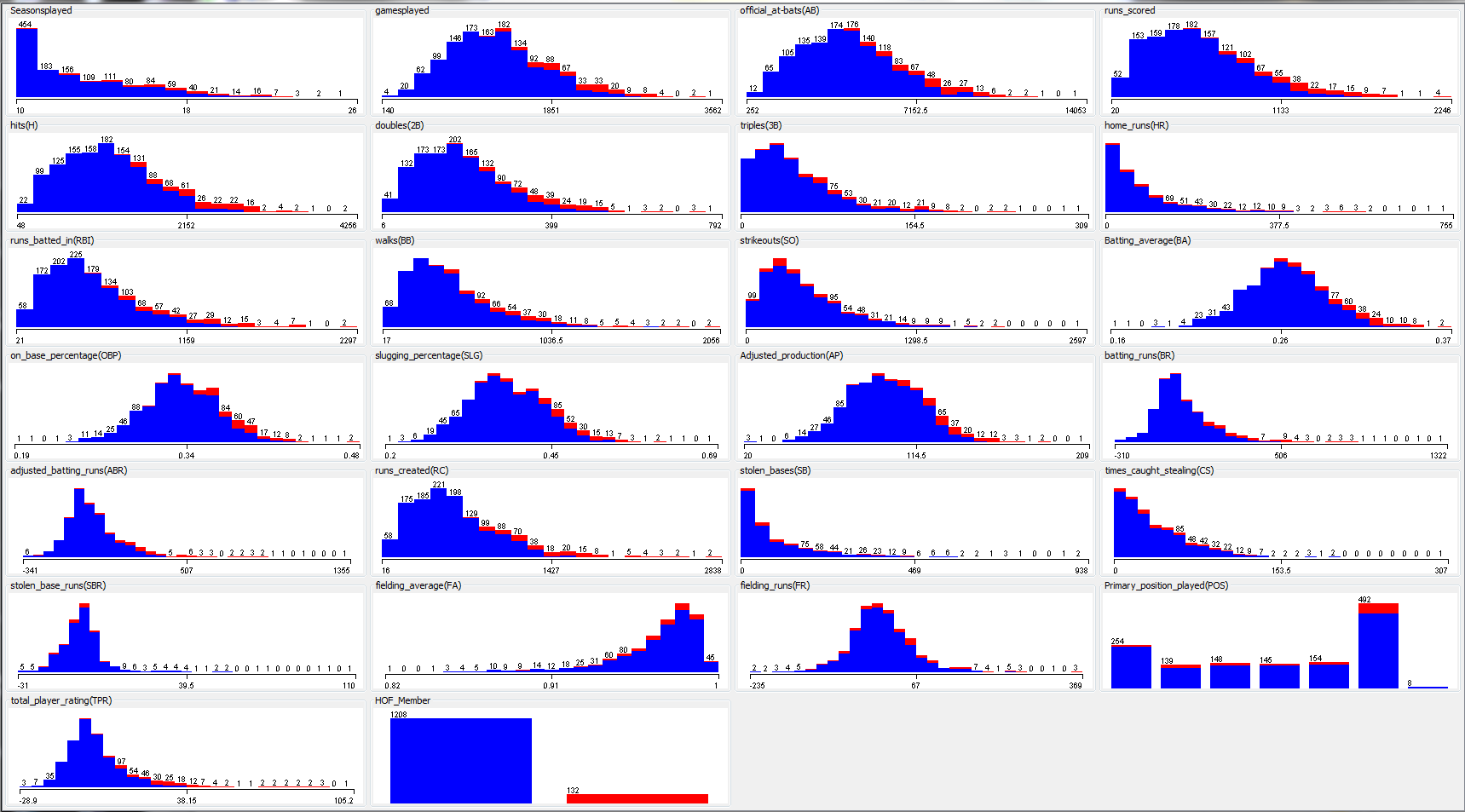


Figure 1. MLBHOF data attributes distribution

The data in table 2 shows the number of players by primary position played and the number of players at that position in the Hall of Fame. This data includes the updates made to reflect player in the database that were selected for the Hall of Fame after 2000. Overall, the distribution of players by position matches the frequency one would expect with two exceptions. First, the designated hitter position has only a small number of eligible players since it is a relatively new non-fielding position that exists only in the American League and is frequently rotated among players on a team to give them a break from their normal feeling position. Second, catchers are relatively over represented on the list of eligibles, at almost 19% of the database when the expected range would be between 11 and 13%.Overall a little over 9% of all players eligible have been selected for the Hall of Fame, however, catchers and third baseman are selected at about half the average frequency, while first baseman have the highest percentage of position players selected for the Hall of Fame. Since the American League went to the designated hitter rule, only a small number of players, who spent the majority of their career as a designated hitter, have become eligible for the Hall of Fame and none of them have been selected to the Hall of Fame when the dataset was created. Since the cutoff date for this database, one player, Paul Molitor, who spent a significant part of his career as a designated hitter has been selected for the Hall of Fame but he spend enough time as a position player that he is assigned a fielding position in this database. Currently, there exists a perception that the baseball writers Association voters downgrade career designated hitters when evaluating them as candidates for the Hall of Fame (Mills and Salaga, 2011).

Table 2. Players and HOF membership by position

| **Position played** | **number in database** | **Hall of Fame members** | **Percent of players eligible** | **percent in Hall of Fame** |
| --- | --- | --- | --- | --- |
| C | 254 | 11 | 18.96% | 4.3% |
| 1 | 139 | 19 | 10.37% | 13.7% |
| 2 | 148 | 14 | 11.04% | 9.5% |
| 3 | 145 | 8 | 10.82% | 5.5% |
| S | 154 | 17 | 11.49% | 11.0% |
| O | 492 | 55 | 36.72% | 11.2% |
| D | 8 | 0 | 0.60% | 0.0% |
| **Total** | **1340** | **124** | **100.00%** | **9.3%** |

## Other data sources

Since the database was created an additional 13 players who were not included in the database have become eligible for and been selected for membership in the Hall of Fame. I created an additional test data set to use in testing the classification models producing the best results with these 13 players plus 39 additional players, who are not in the Hall of Fame and are no longer being considered by the Baseball Writers Association. I was able to access statistics at the baseballreference.com website (<http://www.baseball-reference.com/players/>) for all necessary career total information in several of the sabermetrics data sets. For those data attributes not found at the baseball reference.com website. I was able to use the formulas in Albert’s (2010) paper to calculate the missing data, with the exception of the total player rating. I was able to get the TPR for all of the new Hall of Fame players from a blog post (Yawdoszyn, 2006) discussing Hall of Fame candidates and a few other players in terms of their TPR scores. For those sabermetrics values calculated based on the Albert’s (2010) paper I was able to verify the calculations by calculating values for players in the database and confirming that the results were close matches, allowing for rounding errors. However, I was unable to consistently replicate the TPR calculations, especially for first baseman, shortstops and catchers. Therefore, I had to restrict players included in the second test data set to those players for which I could obtain a TPR score.

## Data cleaning and preparation.

For the primary data set, instances with missing values were not deleted. The data set was updated to indicate selection for the Hall of Fame for the 10 players in the database selected for the Hall of Fame since the year 2000, for their records as players. Three other people in the database were selected for the Hall of Fame, for their record as managers and were left coded as nonmembers. Player’s names are string values and were removed from the data set prior to doing the analysis. There were only two sets of categorical attributes in the data set, one for position played, and when indicating if in the Hall of Fame, and if so how elected. For this analysis I shifted the Hall of Fame attribute to two values, indicating a member or not a member, and removed the distinction between elected by the baseball writers or the veterans committee.

Given that I decided to exclude any players for which I could not obtain a TPR value as discussed in the prior section, there were no problems with missing data for the players included in the second data set of those players eligible after 2000. I also decided to exclude those players have been associated with steroid use in association with the release of the Mitchell Report. There’s been considerable speculation that the low number of votes received by those players in recent Hall of Fame voting is due to voters concerns about the players fitness for the Hall of Fame when considering integrity and sportsmanship, since the use of these performance-enhancing drugs was cheating on the part of the player (Mills and Salaga, 2011), (Yawdoszyn, 2006), (Young, Holland, & Weckman, 2008). For this reason I did not include any of those players within the post-2000 test data set.

Young et al. (2008) in their paper combined first and third base into a category labeled corner infield and second base and shortstop positions into a category called middle infield. I decided it was not appropriate to group positions played into a hierarchy, while players at second base and shortstop are selected to the Hall of Fame at approximately the same rate, and players from third-base are selected to the Hall of Fame in about half the rate is players from first base.

# Methods

## Choice of a model

In deciding which models to use, I did a literature review to see what other papers or available that addressed this topic and which models were used. The other papers are discussed in the comparison with similar studies section. I then ran those models used from the other studies, which I found interesting, along with additional models/algorithms that I was interested in exploring against the base case data set and compared the results to the results of running the ZeroR model against the base case. ZeroR is a simple classifier which assumes that everything is the most common class value; in this case that everyone is not in the Hall of Fame. This helped me to delete options which did not appear to produce reasonably accurate results. I then narrowed the models to be used for this paper down to three classification models, Naïve Bayes, J48 decision tree, and logistic regression, in order to keep the scope to a reasonable level.

### Naïve Bayes classification

Naïve Bayes is a statistical-based classification method that predicts class for an instance based upon the combined probability that the values of each individual attribute to indicate membership in the class. Naïve Bayes assumes that the value of each individual attribute is independent of the values of the other attributes. This is not always the case, and therefore faith-based classification normally works better with a reduced attribute set that eliminates redundant attributes or minimally contributing attributes. Naïve Bayes also has an advantage in that it handles missing values by simply omitting them from the calculations that specific instance and it handles cases of zero on occurrences by using the Laplace estimator which allows for a small but nonzero likelihood that it could occur in the future (Han & Kamber, 2006) (Witten, Frank, & Hall, 2011).

### J48 Decision tree

The J48 decision tree algorithm is the implementation of the C4.5 decision tree induction algorithm. These are described as greedy or non-backtracking algorithms which build their structure based on a divide and conquer strategy and those branch splitting, i.e., which attribute among those remaining will be used for the next level in the three, is based on greatest information gain. For decision trees, pruning is used to reduce the risk of overfitting, reduce complexity and increase computational efficiency. The end result is a set of rules in a tree structure that can be applied to determine which class to assign a new instance of data (Han & Kamber, 2006) (Witten, Frank, & Hall, 2011).

### Logistic regression

Regression techniques are statistical methods that are easily adaptable for classification where there is numeric attributes. Regression techniques use the training instances to calculate the probability of class membership as a function of the value for each attribute. When evaluating a new instance the probabilities are calculated and the largest is selected as the class value. Logistic regression is a generalized linear model that avoids some of the problems of linear regression. It does not assume the data for an attribute is normally distributed, which figure 1 shows is not the case for several of the attributes in this data set. Second, it does not approximate the class values (0 or 1) directly, which can produce out of range probability values, instead it approximates the values using the logit transformation function (Han & Kamber, 2006) (Witten, Frank, & Hall, 2011).

## Experimental design

In conducting the test there were two objectives. First, to determine the relative effectiveness of the three different classification systems examined and second to evaluate the effect of removing the total player ranking (TPR) score from the data set and the benefit of reducing the number of attributes in the data set. These experiments each model was evaluated against 4 data sets as shown in Table 3. One immediate impact of the removal of the TPR from the database was that the number of reduced attributes went from 7 to 10, when the TPR was removed from the data set. The Weka Explored function, using 10 fold cross validation, was used to train and evaluate each classification method against each of the data sets. The evaluation model was saved and then used on the second test data set containing player eligible since 2000. Weka experimenter function was also used to evaluate the relative utility of the four data sets free to the classification algorithms. In addition the Weka Experimenter capability was used to create an experiment to evaluate the 3 models against the four data sets.

Table 4. Data Sets used

| **Data Set Name** | **Modifications** |
| --- | --- |
| Base case | data was updated to reflect players added to the Hall of Fame after the cutoff date |
| Reduced attributes | CFS Subset Evaluator (Best First) to select attributes retained: runs\_scored, hits(H), Batting\_average(BA), batting\_runs(BR), runs\_created(RC), total\_player\_rating(TPR), and HOF membership |
| No TPR | Base case with TPR attribute removed |
| Reduced attributes- no TPR | CFS Subset Evaluator (Best First) used on No TPR data set. Attributes retained: runs\_scored, hits(H), triples(3B), home\_runs(HR), runs\_batted\_in(RBI), Batting\_average(BA), batting\_runs(BR), adjusted\_batting\_runs(ABR), runs\_created(RC), fielding\_runs(FR), HOF\_Member |

## Analysis and interpretation

### Naïve Bayes classification

The summary output, accuracy results and confusion matrix results for the Naïve Bayes results are shown in table 5. In general the Naïve Bayes classification appears to favor accurately classifying a high percentage of HOF members at the expense of including a high number of false positives (classifying non-HOF members as members). This trend was evident in both the training set and the test set.

Table 5. Naive Bayes classification results

| **Data Set** | **Test mode:10-fold**  **cross-validation** | **Test data** |
| --- | --- | --- |
| **Base case**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1187 88.5821 %  Incorrectly Classified Instances 153 11.4179 %  Kappa statistic 0.5368  Mean absolute error 0.1157  Root mean squared error 0.3326  Relative absolute error 64.9711 %  Root relative squared error 111.5997 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.889 0.856 0.886  0.144 0.111 0.141  0.983 0.457 0.931  0.889 0.856 0.886  0.934 0.596 0.9  0.954 0.953 0.954  **Confusion Matrix**  a b <-- classified as  1074 134 | a = 0  19 113 | b = 1 | **Summary**  Correctly Classified Instances 48 92.3077 %  Incorrectly Classified Instances 4 7.6923 %  Kappa statistic 0.814  Mean absolute error 0.0723  Root mean squared error 0.2618  Relative absolute error 19.0567 %  Root relative squared error 60.4415 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.897 1 0.923  0 0.103 0.026  1 0.765 0.941  0.897 1 0.923  0.946 0.867 0.926  0.994 0.994 0.994  **Confusion Matrix**  a b <-- classified as  35 4 | a = 0  0 13 | b = 1 |
| **Reduced attributes**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1226 91.4925 %  Incorrectly Classified Instances 114 8.5075 %  Kappa statistic 0.6146  Mean absolute error 0.0875  Root mean squared error 0.2729  Relative absolute error 49.1213 %  Root relative squared error 91.5842 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.923 0.841 0.915  0.159 0.077 0.151  0.982 0.544 0.938  0.923 0.841 0.915  0.951 0.661 0.923  0.963 0.963 0.963  **Confusion Matrix**  a b <-- classified as  1115 93 | a = 0  21 111 | b = 1 | **Summary**  Correctly Classified Instances 44 84.6154 %  Incorrectly Classified Instances 8 15.3846 %  Kappa statistic 0.6596  Mean absolute error 0.1467  Root mean squared error 0.3588  Relative absolute error 48.961 %  Root relative squared error 78.2491 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted  0.795 1 0.846  0 0.205 0.051  1 0.619 0.905  0.795 1 0.846  0.886 0.765 0.855  0.99 0.99 0.99  **Confusion Matrix**  a b <-- classified as  31 8 | a = 0  0 13 | b = 1 |
| **TRP attribute removed**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1172 87.4627 %  Incorrectly Classified Instances 168 12.5373 %  Kappa statistic 0.4975  Mean absolute error 0.1223  Root mean squared error 0.3402  Relative absolute error 68.6403 %  Root relative squared error 114.1456 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.881 0.818 0.875  0.182 0.119 0.176  0.978 0.429 0.924  0.881 0.818 0.875  0.927 0.563 0.891  0.95 0.948 0.95  **Confusion Matrix**  a b <-- classified as  1064 144 | a = 0  24 108 | b = 1 | **Summary**  Correctly Classified Instances 40 76.9231 %  Incorrectly Classified Instances 12 23.0769 %  Kappa statistic 0.5294  Mean absolute error 0.2235  Root mean squared error 0.4625  Relative absolute error 74.6158 %  Root relative squared error 100.8613 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.692 1 0.769  0 0.308 0.077  1 0.52 0.88  0.692 1 0.769  0.818 0.684 0.785  0.986 0.977 0.984  **Confusion Matrix**  a b <-- classified as  27 12 | a = 0  0 13 | b = 1 |
| **Reduced attributes no TRP**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1204 89.8507 %  Incorrectly Classified Instances 136 10.1493 %  Kappa statistic 0.5594  Mean absolute error 0.1031  Root mean squared error 0.3043  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.907 0.818 0.899  0.182 0.093 0.173  0.979 0.491 0.931  0.907 0.818 0.899  0.942 0.614 0.909  0.955 0.955 0.955  **Confusion Matrix**  a b <-- classified as  1096 112 | a = 0  24 108 | b = 1 | **Summary**  Correctly Classified Instances 44 84.6154 %  Incorrectly Classified Instances 8 15.3846 %  Kappa statistic 0.6596  Mean absolute error 0.1494  Root mean squared error 0.3735  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.795 1 0.846  0 0.205 0.051  1 0.619 0.905  0.795 1 0.846  0.886 0.765 0.855  0.966 0.966 0.966  **Confusion Matrix**  a b <-- classified as  31 8 | a = 0  0 13 | b = 1 |

Table 6 shows the results the T-test results (.05 level of confidence) for selected measures of effectiveness (MOE) using the Weka experiment function to determine if the different data sets had an impact on the quality of the prediction made, The table below shows that the results achieved with the two reduced attribute data sets were significantly better than performance for the base case, and that removing the TPR from the data set produced significantly worst results.

Table 6. Naive Bayes T-Test results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Base Case |** | **Reduced Attributes** | **Base case-TPR removed** | **No TPR reduced attributes** |
| **Percent\_correct** | 88.53(2.80) | | 91.49(2.39) v | 87.59(2.97) \* | 89.79(2.66) v |
| (v/ /\*) | | (1/0/0) | (0/0/1) | (1/0/0) |
| **Area\_under\_ROC** | 0.95(0.02) | | 0.96(0.02) v | 0.95(0.02) \* | 0.96(0.02) |
| (v/ /\*) | | (1/0/0) | (0/0/1) | (0/1/0) |
| **F\_measure** | 0.93(0.02) | | 0.95(0.01) v | 0.93(0.02) \* | 0.94(0.02) v |
| (v/ /\*) | | (1/0/0) | (0/0/1) | (1/0/0) |

Weka experiment rank results are shown in table 7. The ranking function showed that when using the Naïve Bayes classification there was a significant difference in the percent correct and F-measure results for each of the four data sets. The results show that a reduced attribute set is most important and that the TPR ranking also provides significant information.

Table 7. Weka rank Naive Bayes

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | **Result set** |
| 3 | 3 | 0 | Reduced Attributes |
| 1 | 2 | 1 | Reduced attributes No TPR |
| -1 | 1 | 2 | Base Case |
| -3 | 0 | 3 | Base case-TPR removed |

### J48 Decision tree

Table 8 below shows the summary output, accuracy results and confusion matrix results for the for the J48 decision tree algorithm. The J 48 algorithm produced trees of size 17 to 13 with 9 to 7 leaves. Unlike the Naïve Bayes algorithm, the J 48 decision tree algorithm tended to produce confusion matrices with a smaller percentage of Hall of Fame members correctly identified, but also with a significant lower number of false positives (nonmembers identified as members). When the J48 models for each of the data sets were tested against the test data set. It tended to do better than the Naïve Bayes by correctly predicting a high percentage of Hall of Fame members with a low number of false positives. The one exception was for the base case data set with the TPR attribute removed. In that particular instance, the model only identified half of the Hall of Fame members as members, with no false positives.

Table 8. J48 decision tree classification results

| **Data Set** | **Test mode:10-fold**  **cross-validation** | **Test data** |
| --- | --- | --- |
| **Base case**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1255 93.6567 %  Incorrectly Classified Instances 85 6.3433 %  Kappa statistic 0.6099  Mean absolute error 0.0776  Root mean squared error 0.2342  Relative absolute error 43.5833 %  Root relative squared error 78.602 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.975 0.583 0.937  0.417 0.025 0.378  0.955 0.72 0.932  0.975 0.583 0.937  0.965 0.644 0.934  0.834 0.834 0.834  **Confusion Matrix**  a b <-- classified as  1178 30 | a = 0  55 77 | b = 1 | **Summary**  Correctly Classified Instances 52 100 %  Incorrectly Classified Instances 0 0 %  Kappa statistic 1  Mean absolute error 0  Root mean squared error 0  Relative absolute error 0 %  Root relative squared error 0 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg  1 1 1  0 0 0  1 1 1  1 1 1  1 1 1  1 1 1  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  0 13 | b = 1 |
| **Reduced attributes**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1254 93.5821 %  Incorrectly Classified Instances 86 6.4179 %  Kappa statistic 0.6068  Mean absolute error 0.0796  Root mean squared error 0.2272  Relative absolute error 44.6969 %  Root relative squared error 76.2463 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.974 0.583 0.936  0.417 0.026 0.378  0.955 0.713 0.931  0.974 0.583 0.936  0.965 0.642 0.933  0.871 0.871 0.871  **Confusion Matrix**  a b <-- classified as  1177 31 | a = 0  55 77 | b = 1 | **Summary**  Correctly Classified Instances 49 94.2308 %  Incorrectly Classified Instances 3 5.7692 %  Kappa statistic 0.85  Mean absolute error 0.1065  Root mean squared error 0.2117  Relative absolute error 35.5374 %  Root relative squared error 46.1752 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted  0.949 0.923 0.942  0.077 0.051 0.071  0.974 0.857 0.945  0.949 0.923 0.942  0.961 0.889 0.943  0.978 0.978 0.978  **Confusion Matrix**  a b <-- classified as  37 2 | a = 0  1 12 | b = 1 |
| **TRP attribute removed**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1254 93.5821 %  Incorrectly Classified Instances 86 6.4179 %  Kappa statistic 0.598  Mean absolute error 0.0796  Root mean squared error 0.2401  Relative absolute error 44.6663 %  Root relative squared error 80.5629 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.977 0.561 0.936  0.439 0.023 0.398  0.953 0.725 0.931  0.977 0.561 0.936  0.965 0.632 0.932  0.832 0.832 0.832  **Confusion Matrix**  a b <-- classified as  1180 28 | a = 0  58 74 | b = 1 | **Summary**  Correctly Classified Instances 45 86.5385 %  Incorrectly Classified Instances 7 13.4615 %  Kappa statistic 0.5625  Mean absolute error 0.1232  Root mean squared error 0.282  Relative absolute error 41.1274 %  Root relative squared error 61.5077 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  1 0.462 0.865  0.538 0 0.404  0.848 1 0.886  1 0.462 0.865  0.918 0.632 0.846  0.915 0.915 0.915  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  7 6 | b = 1 |
| **Reduced attributes no TRP**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1250 93.2836 %  Incorrectly Classified Instances 90 6.7164 %  Kappa statistic 0.5915  Mean absolute error 0.0832  Root mean squared error 0.2448  Relative absolute error 46.7122 %  Root relative squared error 82.1546 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.972 0.576 0.933  0.424 0.028 0.385  0.954 0.691 0.929  0.972 0.576 0.933  0.963 0.628 0.93  0.87 0.87 0.87  **Confusion Matrix**  a b <-- classified as  1174 34 | a = 0  56 76 | b = 1 | **Summary**  Correctly Classified Instances 49 94.2308 %  Incorrectly Classified Instances 3 5.7692 %  Kappa statistic 0.8421  Mean absolute error 0.1093  Root mean squared error 0.2287  Relative absolute error 36.4754 %  Root relative squared error 49.8746 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.974 0.846 0.942  0.154 0.026 0.122  0.95 0.917 0.942  0.974 0.846 0.942  0.962 0.88 0.942  0.957 0.957 0.957  **Confusion Matrix**  a b <-- classified as  38 1 | a = 0  2 11 | b = 1 |

The T-test results, using the Weka experiment function for the J48 decision tree, show there were no significant differences in the MOEs from the different data sets used in this experiment. In addition, there was no rank differentiation between the poor data sets for this algorithm.

Table 9. J48 decision tree T-Test results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Base Case |** | **Reduced Attributes** | **Base case-TPR removed** | **No TPR reduced attributes** |
| **Percent\_correct** | 94.03(1.70)| | 94.10(1.67) | 93.40(1.81) | 93.70(1.76) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |
| **Area\_under\_ROC** | 0.85(0.09)| | 0.88(0.09) | 0.83(0.09) | 0.87(0.09) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |
| **F\_measure** | 0.97(0.01) | | 0.97(0.01) | 0.96(0.01) | 0.97(0.01) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |

### Logistic regression

Logistic regression results are shown in table 10 below. As was the case with the J48 decision tree, for the training data logistic regression classification developed a classification scheme that tended toward lower false positives at the expense of more false negatives relative to the Naïve Bayes classification scheme. One run against the test data set, as the confusion matrices show, logistic regression models had a very high accuracy rate, correctly identifying all our all but one Hall of Fame members, with no false positives with one exception. The classification model developed from the reduced attributes set when total player rating was not available had no false positives, but only correctly identified eight out of 13 Hall of Fame members.

Table 10. Logistic regression classification results

| **Data Set** | **Test mode:10-fold**  **cross-validation** | **Test data** |
| --- | --- | --- |
| **Base case**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1279 95.4478 %  Incorrectly Classified Instances 61 4.5522 %  Kappa statistic 0.7201  Mean absolute error 0.066  Root mean squared error 0.1939  Relative absolute error 37.0318 %  Root relative squared error 65.0573 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.985 0.674 0.954  0.326 0.015 0.295  0.965 0.832 0.952  0.985 0.674 0.954  0.975 0.745 0.952  0.965 0.965 0.965  **Confusion Matrix**  a b <-- classified as  1190 18 | a = 0  43 89 | b = 1 | **Summary**  Correctly Classified Instances 1280 95.5224 %  Incorrectly Classified Instances 60 4.4776 %  Kappa statistic 0.7277  Mean absolute error 0.0708  Root mean squared error 0.1924  Relative absolute error 39.7378 %  Root relative squared error 64.5775 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  1 1 1  0 0 0  1 1 1  1 1 1  1 1 1  1 1 1  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  0 13 | b = 1 |
| **Reduced attributes**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1280 95.5224 %  Incorrectly Classified Instances 60 4.4776 %  Kappa statistic 0.7277  Mean absolute error 0.0708  Root mean squared error 0.1924  Relative absolute error 39.7378 %  Root relative squared error 64.5775 %  Total Number of Instances 1340    **Detailed Accuracy By Class**  0 1 Weighted Avg  0.984 0.689 0.955  0.311 0.016 0.282  0.967 0.827 0.953  0.984 0.689 0.955  0.975 0.752 0.953  0.969 0.969 0.969  **Confusion Matrix**  a b <-- classified as  1189 19 | a = 0  41 91 | b = 1 | **Summary**  Correctly Classified Instances 51 98.0769 %  Incorrectly Classified Instances 1 1.9231 %  Kappa statistic 0.9474  Mean absolute error 0.1146  Root mean squared error 0.2007  Relative absolute error 38.2517 %  Root relative squared error 43.7717 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted  1 0.923 0.981  0.077 0 0.058  0.975 1 0.981  1 0.923 0.981  0.987 0.96 0.981  0.994 0.994 0.994  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  1 12 | b = 1 |
| **TRP attribute removed**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1277 95.2985 %  Incorrectly Classified Instances 63 4.7015 %  Kappa statistic 0.713  Mean absolute error 0.069  Root mean squared error 0.1984  Relative absolute error 38.7335 %  Root relative squared error 66.5628 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg  0.983 0.674 0.953  0.326 0.017 0.295  0.965 0.817 0.95  0.983 0.674 0.953  0.974 0.739 0.951  0.963 0.963 0.963  **Confusion Matrix**  a b <-- classified as  1188 20 | a = 0  43 89 | b = 1 | **Summary**  Correctly Classified Instances 51 98.0769 %  Incorrectly Classified Instances 1 1.9231 %  Kappa statistic 0.9474  Mean absolute error 0.0559  Root mean squared error 0.1359  Relative absolute error 18.6586 %  Root relative squared error 29.6285 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  1 0.923 0.981  0.077 0 0.058  0.975 1 0.981  1 0.923 0.981  0.987 0.96 0.981  0.998 0.998 0.998  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  1 12 | b = 1 |
| **Reduced attributes no TRP**  Class  TP Rate FP Rate Precision Recall  F-Measure ROC Area | **Summary**  Correctly Classified Instances 1265 94.403 %  Incorrectly Classified Instances 75 5.597 %  Kappa statistic 0.6427  Mean absolute error 0.0804  Root mean squared error 0.2053  Relative absolute error 45.1549 %  Root relative squared error 68.8919 %  Total Number of Instances 1340  **Detailed Accuracy By Class**  0 1 Weighted Avg.  0.983 0.583 0.944  0.417 0.017 0.377  0.956 0.794 0.94  0.983 0.583 0.944  0.969 0.672 0.94  0.956 0.956 0.956  **Confusion Matrix**  a b <-- classified as  1188 20 | a = 0  55 77 | b = 1 | **Summary**  Correctly Classified Instances 47 90.3846 %  Incorrectly Classified Instances 5 9.6154 %  Kappa statistic 0.7059  Mean absolute error 0.1445  Root mean squared error 0.2586  Relative absolute error 48.2254 %  Root relative squared error 56.3942 %  Total Number of Instances 52  **Detailed Accuracy By Class**  0 1 Weighted Avg.  1 0.615 0.904  0.385 0 0.288  0.886 1 0.915  1 0.615 0.904  0.94 0.762 0.895  0.972 0.972 0.972  **Confusion Matrix**  a b <-- classified as  39 0 | a = 0  5 8 | b = 1 |

The results from the Weka experiment function, in table 11 below, show that there was no significant difference in the MOE’s from the different data sets when the base case data set was used as the point of reference. That is the results from the other data sets were not significantly better or worse than the results achieved with the base case.

Table 11. Logistic regression T-Test results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Base Case |** | **Reduced Attributes** | **Base case-TPR removed** | **No TPR reduced attributes** |
| **Percent\_correct** | 0.97(0.02) | | 0.97(0.02) | 0.97(0.02) | 0.96(0.02) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |
| **Area\_under\_ROC** | 0.97(0.02) | | 0.97(0.02) | 0.97(0.02) | 0.96(0.02) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |
| **F\_measure** | 0.97(0.01) | | 0.97(0.01) | 0.97(0.01) | 0.97(0.01) |
| (v/ /\*) | | (0/1/0) | (0/1/0) | (0/1/0) |

However, as table 12 shows, when the t-test was used to rank the four different data sets. Overall, the reduced attribute data set was ranked first in the base case without the TPR attribute was ranked last. The reduced attribute data set was significantly better than the base case with the TPR attribute removed, while there was no significant difference between it and the other two data sets or between them and the base case with TPR removed. This applies when the data sets were evaluated on percent correct, F – measure, and area under ROC MOE’s.

Table 12. Weka rank logistic regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | Result set |
| 1 | 1 | 0 | Reduced Attributes |
| 0 | 0 | 0 | No TPR reduced attributes |
| 0 | 0 | 0 | Base Case |
| -1 | 0 | 1 | Base case-TPR removed |

# Comparison of Alternative Solutions

As an initial comparison of the three models I did one comparison using the Weka experiment environment where I compared to three models with the ZeroR algorithm as a base case for comparison. The ZeroR algorithm uses a rule that assumes all entities are assigned to the most common class value, and serves as a minimum standard which any algorithm under consideration should exceed. Table 13 below shows the ranked results for the four algorithms with the base case data set, assessed for the F – measure and area under the ROC MOE’s. The percent cranked ranking was similar except that the positions of J48 decision tree and Naïve Bayes were reversed. This demonstrates that all three of the algorithms produced significantly better results than the ZeroR algorithm.

Table 13. Weka classification algorithm ranking with ZeroR base case

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | Result set |
| 3 | 3 | 0 | Logistic regression |
| 1 | 2 | 1 | Naïve Bayes |
| -1 | 1 | 2 | J48 decision tree |
| -3 | 0 | 3 | ZeroR |

The ROC visual threshold curves for these four algorithms are shown in Figure 2 below. The ZeroR results were in the upper left and show a straight 45° line reflecting the low quality of its simplistic classification rule. The curves for the Naïve Bayes and the J48 decision tree functions are almost identical as are there area under ROC values, while that for the logistics regression is not quite as good, reflecting its tendency for more false negatives in order to minimize false positives.

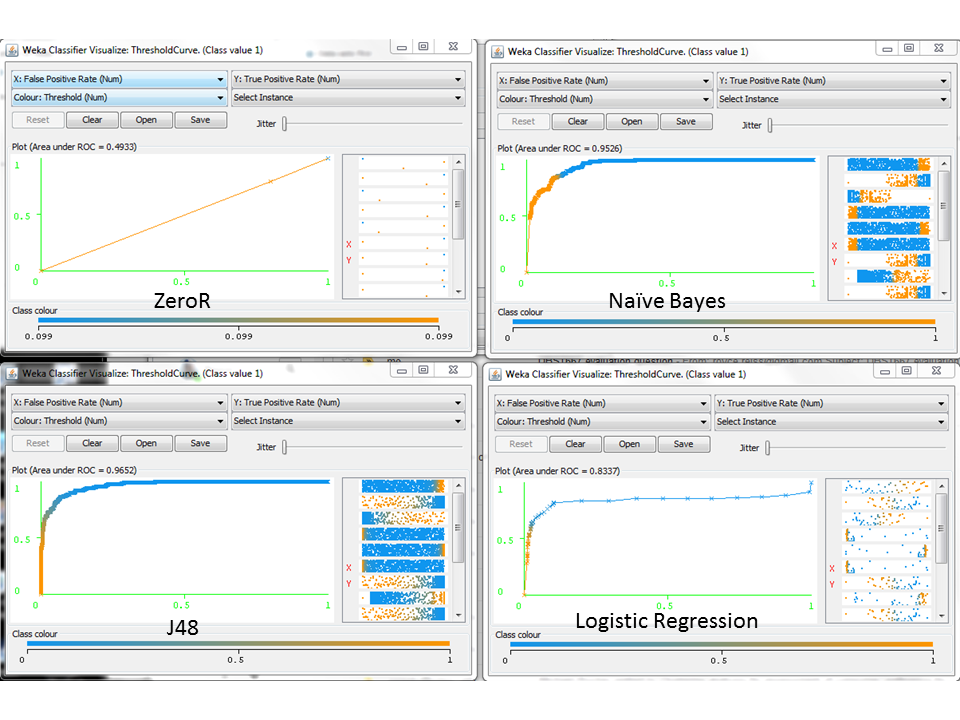


Figure 2. ROC curves

The Weka experimenter capability was used to set up a 3 x 4 experiment that compared the three classification algorithms against each of the four data sets and did a combined evaluation of the results. All evaluations used the corrected T-test algorithm in Weka at the .05 level of confidence. Focusing on the models first, the results below were developed using the J48 decision tree algorithm is the base for comparison. This was selected in order show maximum discrimination between the three algorithms. Tables 14, 15, and 16 show the results for the three primary MOE’s percent correct, F – measure, and area under ROC. The results of the ranking algorithm are shown in tables 17 (percent correct and F – measure) and 18 (area under ROC). The overall results showed that the logistics regression algorithm performed best, followed by the J48 decision tree, and then the Naïve Bayes. However when area under ROC is considered then the Naïve Bayes algorithm performs better than the J48 decision tree algorithm.

Table 14. T-Test Classification models Percent correct

| **Dataset** | **J48 Tree |** | **Naïve Bayes** | **Logistic Regression** |
| --- | --- | --- | --- |
| **Base case** | 94.03(1.70) | | 88.53(2.80) \* | 95.34(1.77) v |
| **Reduced attributes** | 94.10(1.67) | | 91.49(2.39) \* | 95.34(1.75) v |
| **No TPR** | 93.40(1.81) | | 87.59(2.97) \* | 95.07(1.75) v |
| **Reduced attributes- no TPR** | 93.70(1.76) | | 89.79(2.66) \* | 94.54(1.70) |
|  | (v/ /\*) | | (0/0/4) | (3/1/0) |

Table 15. T-Test Classification models F-Measure

| **Dataset** | **J48 Tree |** | **Naïve Bayes** | **Logistic Regression** |
| --- | --- | --- | --- |
| **Base case** | 0.97(0.01) | 0.93(0.02) \* | 0.97(0.01) v |
| **Reduced attributes** | 0.97(0.01) | 0.95(0.01) \* | 0.97(0.01) v |
| **No TPR** | 0.96(0.01) | 0.93(0.02) \* | 0.97(0.01) v |
| **Reduced attributes- no TPR** | 0.97(0.01) | 0.94(0.02) \* | 0.97(0.01) |
|  | (v/ /\*) | | (0/0/4) | (3/1/0) |

Table 16. T-Test Classification models Area\_under\_ROC

| **Dataset** | **J48 Tree |** | **Naïve Bayes** | **Logistic Regression** |
| --- | --- | --- | --- |
| **Base case** | 0.85(0.09) | | 0.95(0.02) v | 0.97(0.02) v |
| **Reduced attributes** | 0.88(0.09) | | 0.96(0.02) v | 0.97(0.02) v |
| **No TPR** | 0.83(0.09) | | 0.95(0.02) v | 0.97(0.02) v |
| **Reduced attributes- no TPR** | 0.87(0.09) | | 0.96(0.02) v | 0.96(0.02) v |
|  | (v/ /\*) | | (0/0/4) | (4/0/0) |

Table 17. Weka classification algorithm ranking F – measure and percent correct

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | Result set |
| 7 | 7 | 0 | Logistic regression |
| 1 | 4 | 4 | J48 decision tree |
| -8 | 0 | 8 | Naïve Bayes |

Table 17. Weka classification algorithm ranking area under ROC

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | Result set |
| 6 | 6 | 0 | Logistic regression |
| 2 | 4 | 2 | Naïve Bayes |
| -8 | 0 | 8 | J48 decision tree |

The relative utility of the data sets used has been discussed as part of the discussion for each algorithm. As one would expect the combined results, shown in table 18, are in line with the individual results. Using the Weka ranking function, with a T-test confidence interval of 0.5, there was a clear and distinct ranking of the data sets, for each data set was significantly different from the others. Overall, using a reduced attribute data set was more important than the presence or absence of the total player ranking, however, having it in a data set was significantly better than not having it.

Table 18. Weka data set ranking for F – measure (all classification algorithms)

|  |  |  |  |
| --- | --- | --- | --- |
| **Difference**  **>-<** | **# Significantly**  **>** | **# Significantly <** | Result set |
| 4 | 4 | 0 | Reduced Attributes |
| 0 | 2 | 2 | No TPR reduced attributes |
| -1 | 1 | 2 | Base Case |
| -3 | 0 | 3 | Base case-TPR removed |

# Comparison with Similar Studies

Braun, K., Hartz, B., Leyhane, J., & McGee, D (2006) in their paper used data that included number of times a player was selected as an All-Star and number of awards for leading in statistical categories that a player received. But they did not consider data on base running and fielding performance. They also ultimately decided to only consider players in the post-World War II era since, as they note, some argue that many of the players selected from the earlier are essentially mistakes that do not deserve to be in the Hall of Fame based on their career statistics (p. 3). In their initial test, they use Naïve Bayes, JRip, and random forest classifiers. However, they dropped Naïve Bayes and used the other two both individually and with meta-classifiers, AdaBoost with JRip and GainRatio with random forest. Their results used the F measure as their primary measure of effectiveness and produced values of 0.72 for JRip and 0.75 for random forest, which were comparable to my results using J48 and the full attribute data set.

Young et al. (2008) investigated the use of neural networks to forecast Hall of Fame selection for position players. The attributes considered included player position, basic career batting totals, base running totals and fielding totals plus total performance and character awards received by the player. They also excluded designated hitters due to their lack of fielding statistics. They used under supervised K means algorithm, which initially produced 10 clusters approximately half a which contained Hall of Fame players. They reported after considerable testing that they were able to achieve approximately a 98% accuracy rate for classifying players as to whether they were in the Hall of Fame or not. This was something that I only managed with the test data set for the J 48 decision tree and logistic regression models, but did not meet with the training set.

Mills and Salaga (2011) use the random forest classification algorithm to forecast probability of Hall of Fame induction for current and recently retired players who have played for at least 10 years. They used data available from baseball reference.com’s subscription database to extract what they consider to be the traditional batting and baserunning career totals and averages along with the total number of times a player was selected as an All-Star. The players were divided into two data sets, first, a training data set of all players who retired after 1950 and were eligible for or in the Hall of Fame, and second, a test data set consisting of all players who retired after 1989 and were not in the Hall of Fame, as of 2009. They observed that home runs in total All-Star selections were two of the most important attributes in forecasting Hall of Fame selection. They ran multiple models, all of which had a very low misclassification rate for identifying non-Hall of Fame players. However, the models misclassified between 10% and 23% of the Hall of Fame players as Non-Hall of Fame, with an overall OOP rate of 1% to 2.6%. These were similar to the results achieved with the J48 decision tree in the logistic regression models.

# Observations and conclusions

The analysis conducted as part of this project as well as the other related studies show that selecting a Hall of Fame quality player is not a clear-cut statistical based system, however I accuracy’s can be achieved. The experiment and analysis showed that the logistics regression algorithm performed better than the other two algorithms over the full range of cases. In general, the J48 decision tree performed better than the Naïve Bayes algorithm, however, under some measures naïve Bayes performed better. One issue for consideration in deciding whether Naïve Bayes or J48 decision tree would be preferred could depend upon which standard one might wish to apply. Specifically, is it more important to capture a high percentage of Hall of Fame members correctly, while accepting a high number of false positives which could be equated to identifying people not in the Hall of Fame who should be considered, which reflects the Naïve Bayes classification process. Alternatively, as reflected by the J48 decision tree, would it be better to have a more restrictive process which perhaps identifies Hall of Fame members who could be considered mistakes as suggested by Braun et al. (2006) in their paper.

In addition to addressing the advantage of reducing the number of attributes, the paper also looked at the value of TPR, as an example of a more complex sabermetrics score. The neck conclusion is that a reduced attributes set is more important for improving accuracy; however complex sabermetrics such as TPR provide significant information.

The database that was used for this study did not address metrics that some of the other studies did such as number of achievement awards which play a significant role in the selection process (Mills and Salaga, 2011), (Young et al., 2008). If that data were available. It would be worthwhile to include it in the analysis.

The conduct of this project proved to be a good experience in learning the capabilities of the Weka data mining system and several the algorithms that uses. One area where I would like to devote more time is working with the clustering algorithm. I did some initial work with it but could not extract the statistical data I wanted to use to more fully understand the results. I need to work more with Weka to determine if was simply my lack of knowledge of Weka’s capabilities, or if it’s simply something that it does not support.

# References

Albert, J. (2010). Sabermetrics: The Past, the Present, and the Future. Mathematics and Sports, (43), 15. Retrieved from: <http://www-math.bgsu.edu/~albert/papers/saber.html> (chapter 1)

Baseball Encyclopedia of Players. (n.d.). Retrieved from: http://www.baseball-reference.com/players/

Birnbaum, P. (n.d.). A Guide to Sabermetrics. Society for American Baseball Research. Retrieved from: http://sabr.org/sabermetrics

Braun, K., Hartz, B., Leyhane, J., & McGee, D. (2006). *Determining a Baseball Hall of Fame Candidate*. Retrieved from: <http://www.toofishes.net/uploads/baseball-hof.pdf>

Cochran, J. "Career records for all modern position players eligible for the major league baseball hall of fame." *Journal of Statistics Education 8.2 (2000)*. Retrieved from: <http://www.amstat.org/PUBLICATIONS/JSE/secure/v8n2/datasets.cochran.new.cfm>

Election Rules. (n.d.). National Baseball Hall of Fame. Retrieved from: http://baseballhall.org/hall-of-famers/bbwaa-rules-for-election

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Mills, B. M., & Salaga, S. (2011). Using tree ensembles to analyze National Baseball Hall of Fame voting patterns: an application to discrimination in BBWAA voting. *Journal of Quantitative Analysis in Sports*, *7*(4). DOI: 10.2202/1559-0410.1367 . Retrieved from: http://www.brianmmills.com/uploads/2/3/9/3/23936510/1mills\_\_salaga\_-\_random\_forest\_baseball\_hof\_jqas\_2011.pdf

Witten, I., Frank, E., & Hall, M. (2011). Data Mining : Practical Machine Learning Tools and Techniques, third edition. Morgan Kaufmann.

Yawdoszyn, M. (2006). Hall of Famers' Sabermetrics Rankings. [Blog post]. Retrieved from: <http://mysite.verizon.net/vze2x57w/sabermetrics/id1.html>

Young, W. A., Holland, W. S., & Weckman, G. R. (2008). Determining hall of fame status for major league baseball using an artificial neural network. *Journal of Quantitative Analysis in Sports*, *4*(4). DOI: 10.2202/1559-0410.1131