**PREDICTIVE ANALYSIS OF MAJOR LEAGUE BASEBALL DATASET USING BINARY CLASSIFICATION TECHNIQUES**

**Anurag Ladage** aal1284@rit.edu

Center for Quality and Applied Statistics

Rochester Institute of Technology

1 Lomb Memorial Drive

Rochester, NY – 14623, USA.

**ABSTRACT**

This report ventures into the world of predictive analytics using various binary classification techniques of data mining. These techniques are applied to a custom made dataset from Major League Baseball (MLB) game statistics in order to predict the results of the playoffs a year in advance. The various techniques successfully applied include Quadratic Discriminant Analysis, Tree Classification, Random Forest, ADA Boosting and Support Vector Machines. A thorough performance comparison is made between the above-mentioned techniques and their statistical aspects are discussed in depth throughout the paper.

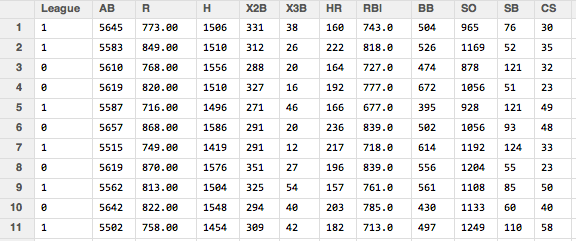
**Keywords:** Predictive Analytics, Binary Classification, Major League Baseball, Game Statistics, Support Vector Machines, Random Forest, Boosting, Tree Classification, Linear Logistics Regression, Statistical Data Mining.

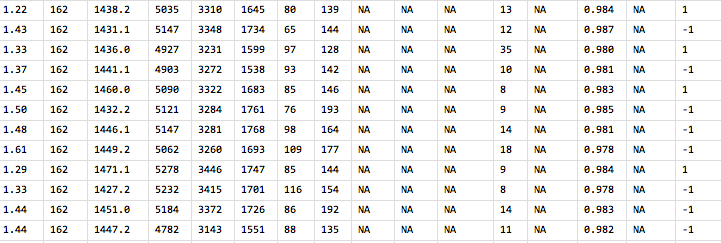
**INTRODUCTION**

A custom made dataset is used in this report. The dataset originates from Major League Baseball statistics. The problem of solving binary classification in a lower dimensionality and high complexity is tackled through this report. The results of the following years playoffs are predicted using the data from the previous year. Various challenges like data collection, filtration of unwanted data, missing values, excessive noise and high complexity are dealt with during the analysis phase. In this report we initially start with Linear Logistic Regression, as it is known to perform well on binary classification due to its inherent mathematical design. After proving conclusively that no concrete analysis can be reached by using Linear Logistic Regression we move on the much prominent techniques like Boosting, Random Forest, Tree Classification and Support Vector Machines. The techniques are performed of three major parameters of performance namely Accuracy, F- Measure and Area Under the ROC Curve.

**EXPLORATION OF THE MLB DATA SET**

A glimpse of the entire dataset is shown below which gives some initial insight into the data. The dimensionality of the dataset is n = 476 with 46 variables available for prediction. The dataset is split into a training data (3/4th) and test data (1/4th). The dataset is extracted from the Major League Baseball. The predictive vector of the dataset consists of various hitting, pitching and fielding parameters collected by the official website of Major League Baseball. These parameters can be well understood from the website [4]. The response variable is a binary variable, which represents the Result. In our case the result is defined as ‘1’ if the team makes it to the divisional series in the succeeding year and ‘-1’ if the team fails to make it to the divisional series the succeeding year. It is important to note that the while performing linear logistics regression analysis and quadratic discriminant analysis the response variable is altered to take values 0 and 1 accordingly. To maintain the legitimacy of the data the entire data is extracted only from the official website [1]. The dataset is of great importance when looked at from a practical point of view, which are discussed later.

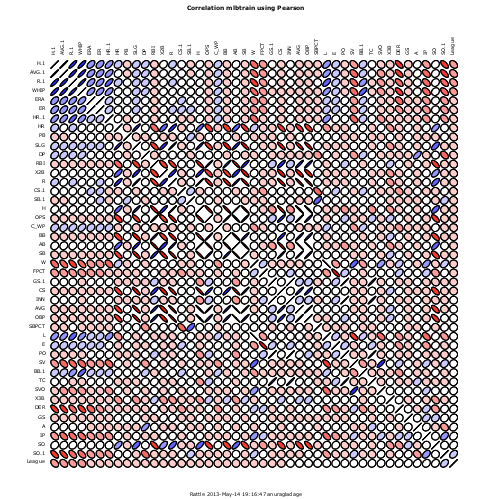


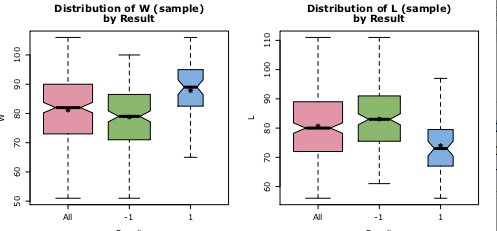
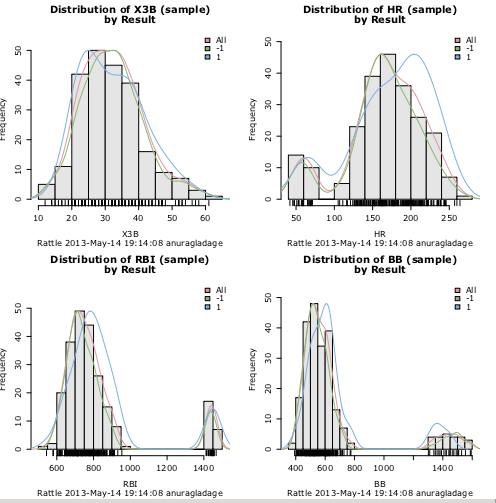


Before entering the analysis phase of the dataset some preliminary analysis is done on the data to verify any pre-assumptions and to show the complexity of the dataset.

**PRELIMINARY ANALYSIS**

Since we will be looking at a spectrum of binary classification techniques it is important to verify the model adequacies required for various techniques to be used. Further on first taking a look at the data we notice that there are some missing values present in the dataset. This is in fact a prominent problem with many of the sports related datasets. As we move towards older data many of the records are missing and this poses a problem when analyzing the data. Various methods including data imputation and neglecting the missing values are used to deal with such problems. We use the central imputation technique provided by the R-package ‘DMwR’ for this dataset. Moreover, as known by any prominent statistician or data scientist preliminary analysis using scatterplots, histogram normality plots gives a good peak into the data and further establishes the legitimacy and importance of in-depth analysis. The correlation matrix and the scatter plots of the data are shown on the following page:





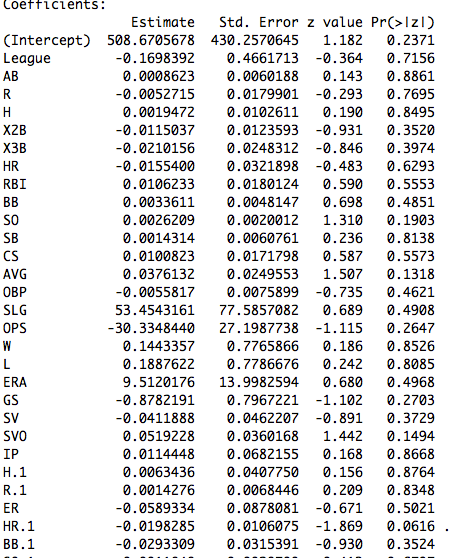
The above scatter plot shows some interesting results. There are some variables, which are related to shown as very high correlation amongst them. This may be a sign of redundancy of the predictive variables. Moreover, the dataset also set also has ample of predictive variables that are very slightly correlated to each other or completely uncorrelated. This makes the dataset a good contender for in-depth analysis. The box-plots were not that useful which suggest that the prediction of the data would be difficult due to overlapping which makes the data quite complex from predictive analysis point of view. The histograms of the predictive vector help us conclude that the data is normally distributed except for a few outliers. This assumption becomes quite crucial when classification is performed using quadratic discriminant analysis or support vector machines.

**PERFORMANCE ANALYSIS OF BINARY CLASSIFICATION TECHNIQUES**

The primary challenge with this dataset is getting a starting point. As the dataset is custom made very little is known about the complexity of the data beforehand. Thus, identifying and narrowing down the techniques to be used on this data is very important. Linear Logistics Regression and Neural Networks are two good techniques to start with as these techniques make very less assumptions regarding the distribution and complexity of the dataset.

* **Linear Logistic Regression**

We start our data analysis with our very own and one of the most widely used techniques called Linear Logistics Regression. This technique has very little assumptions about the data and is known to do well on binary classification problems. After fitting the dataset with this model the following output is achieved:



From the above table it is pretty clear that this technique fails miserably while trying to make sense of the data. This gives us a hint towards the high complexity of the data. Moreover it also helps us arrive to the conclusion that the decision boundaries between the classes are not linear due to which this technique fails to perform well. Moreover, logistic regression is also known to get lost in case there is too much noise in the data. Thus, there is no point in continuing further with this analysis. Additionally, it is important to understand that one of the reasons the Linear Logistic Classifier failed so miserably is due to the fact the separation boundary may not be linearly separable. Moreover, it is also important noting that when neural network is used for binary classification it behaves very much like logistics regression when a logistic sigmoid function is used as an activation function. Thus, as expected when neural networks are used for classification it fails miserably. This failure can be accounted for by stating the fact that the number of observations in the test data is scarce. Thus the neural network fails to create a predictive model.

Moving forward, as the normality assumptions are satisfied we look at techniques like quadratic discriminant analysis and support vector machines, which are known to do pretty well in case of binary classification problems. Moreover, it is also important to venture into non- parametric classification methods as these methods have almost no prior assumption about the data and seem to do pretty well when not much is known about the data set. Thus, we also look at some non – parametric classification techniques like tree classification, Boosting and Random Forest [3].

* **Quadratic Discriminant Analysis (QDA) and Support Vector Machines (SVM)**

The rattle package in R is used to perform Support Vector Machine classification. The R – code for QDA is as shown below:

#Performing QDA on the data

mlbtest.n <- 150 #Number of repeatitions

mlbtrain.qda.fit <- qda(Result~., data = mlbtrain)

mlbtest.qda.yhat <- predict(mlbtrain.qda.fit, mlbtest.x)$class

mlbtest.qda.conf <- table(mlbtest.y,mlbtest.qda.yhat)

mlbtest.qda.accuracy <- sum(diag(hw1data2.qda.conf))/mlbtest.n

Various parameters obtained by performing both these binary classification techniques on the dataset are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Accuracy (%) | Precision (%) | Recall (%) | F - measure |
| Quadratic Discriminant Analysis | 69.33 | 31.25 | 12.5 | 0.178 |
| SVM – ANOVA RBF | 68.66 | 39.39 | 32.5 | 0.356 |
| SVM- POLYDOT (DEGREE 2) | 66.67 | 33.33 | 25 | 0.285 |
| SVM – POLYDOT (DEGREE 3) | 74 | 52.38 | 27.5 | 0.3606 |
| SVM – POLYDOT (DEGREE 4) | 71.3 | 41.17 | 17.5 | 0.2456 |
| SVM – POLYDOT (DEGREE 5) | 70 | 33.33 | 12.5 | 0.1818 |

It is important to point out that the above parameters are tested on the test data rather than the training data as should be done in predictive analysis. It is also worth noting that two important measures are being used to compare various techniques. The first is accuracy and the second is F- measure. The formula for both these measures is as stated below:

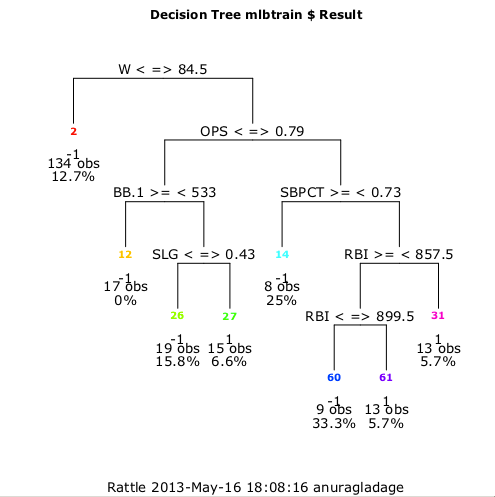
F- Measure =

Where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

Accuracy alone on its own is not sufficient to very confidently point out the correctness or effectiveness of the classifier. Thus, in order to truly compare different techniques these two parameters in combination are used in this report. From the above table it is pretty clear that both the accuracy and recall are higher for SVM classification when ‘Polydot’ kernel with degree 3. The ‘Polydot’ kernel is a polynomial kernel. It helps create a flexible separation boundary elevating the 2-D data space into higher dimensions (degrees). Thus, it will perform well in presence of a non-linear separation boundary. At this point it is important to understand that even though increasing the degrees leads to a decrease in the training error it over fits the data and eventually the model will fail to when applied on the test data. To illustrate this phenomenon we have fitted SVMs with polynomial kernels up till degree 5. It can be observed that after degree 3 both accuracy and F-measure start falling. Thus, we finally conclude that SVM using polynomial kernel of degree 3 is the most suitable till now.

* **Non-Parametric Classification (Decision Trees, Boosting and Random Forest)**

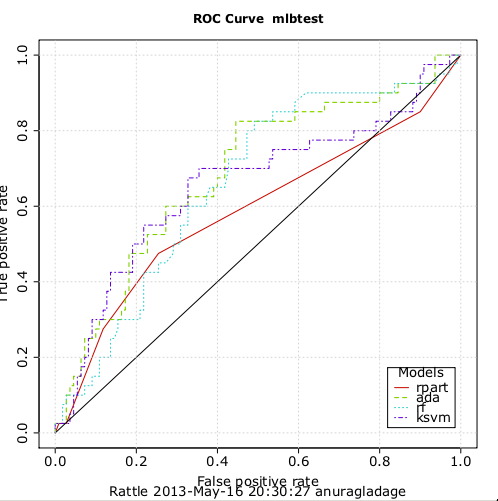
We start by applying the decision tree classification on our dataset. The decision tree thus obtained is shown below:



The accuracy, prediction, recall and F-Measure of the decision tree are shown in the table below. It is also worth noting that both Boosting and Random Forest work on the principle of averaging trees and thus have a chance of performing well were decision trees perform well. Thus, after fitting these techniques and testing their predictive performance on the test data we collect the following performance measures:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Accuracy (%) | Precision (%) | Recall (%) | F - measure |
| Decision Tree | 72 | 44.44 | 20 | 0.276 |
| Boosting | 73.33 | 50 | 25 | 0.333 |
| Random Forest | 71.33 | 36.36 | 10 | 0.157 |

To conclusively prove the best among the above technique we make use of the Receiver Operating Characteristic (ROC) curve. The true positive rate is plotted against the false positive rate in order to obtain ROC curve. ROC is generally a preferred way to compare different techniques as it is an unbiased parameter and the output can be represented in a much-preferred graphical manner. A comparative ROC curve of SVM (degree 3 polydot), Decision Tree Classifier, Ada Boosting and Random Forest is shown below:



**PRACTICAL SIGNIFICANCE, CONCLUSION AND FUTURE STUDY:**

From a practical perspective, the correct classification of this data will dramatically change the game of baseball. Just imagine for a second if the teams could predict beforehand whether or not they could enter the divisional series the following year. It would give a major boost to the confidence of the players knowing they have a chance to improve their chance in the years to come by playing better this year. It would definitely give a psychological boost. The advertisers, fan and sponsors could plan in advance and save a lot of money by avoiding wrong investments and at the same time earn more by making apt investment policies. The practical gratifications of this study are endless!

The table below represents the area under the ROC curve for various techniques used above. From the table below it can be seen that the area under the curve is maximum for Ada Boosting Classifier. Moreover, its accuracy and F-measure are also very close to SVD Classifier. Thus, we conclude that ADA boosting works well on data with less dimensionality and high complexity. This is due to the fact that ADA boosting in very general terms is a technique, which takes multiple weak classifiers and boosts their overall aggregated predictive error, which tends to be less than the individual weak classifiers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Ada  Boosting | Random  Forest | SVM  Classifier | Decision Trees |
| Area Under the ROC Curve | 0.6857 | 0.6598 | 0.6552 | 0.5855 |

Even though we managed to classify data with an accuracy of nearly 75% this is still not good enough from real–world applications. The complexity of data was much more than initially estimated. One major drawback of this dataset was the scarcity of data (n = 480), which is one of the major reasons the techniques (primarily neural networks) were able to achieve only a mediocre predictive performance. Moreover, if somehow data could be filtered to reduce the overlapping in the predictive variables and noise in the data the techniques may be able to reduce the predictive error significantly. Moreover, even anomaly detection analysis can be performed on the given dataset, as the number of instances of class 1 is significantly less when compared to class -1. Performing regression analysis on the dataset would also give interesting results if a prediction in terms of which team has higher chances of entering the divisional series the following year could be achieved.

**REFERENCES:**

**1] Major league baseball stats. (2013/5/12). Retrieved from**

**http://mlb.mlb.com/stats/sortable.jsp?c\_id=mlb&tcid=mm\_mlb\_stats**

**2] Clarke, Fokoué, and Zhang (2009), Principles and Theory for Data Mining and**

**Machine Learning, Springer Verlag, New York.**

**3] Schapire, R. (2005, 5 1). Boosting, Department of Computer Science,**

**Princeton University. Retrieved from** [**http://videolectures.net/mlss05us\_schapire\_b/**](http://videolectures.net/mlss05us_schapire_b/)

**4] *Major league baseball parameters*. (2013, 05 11). Retrieved from**

**http://mlb.mlb.com/stats/sortable.jsp?c\_id=mlb&tcid=mm\_mlb\_stats**