Major League Baseball Analytics: What Contributes Most to Winning

Vignesh Arasu M05542052

BANA 8083

Capstone Project

Readers:

Dr. Yan Yu, PhD, University of Cincinnati

Matthew Risley, MA, University of Cincinnati

Abstract:

Big data and analytics has been a growing force in Major League Baseball. The principle of moneyball vitalizes the importance of two of these statistics, on-base percentage and slugging (Total Bases/Number of Bats) as the core principles for building winning franchises. This analysis of this report of data from all teams from 1962-2012 incorporating methods of multiple linear regression, logistic regression, regression and classification trees, generalized additive models, linear discriminant analysis, and k-means clustering creating the best models for number of wins by a team(linear regression response variable) and whether or not a team makes the postseason(logistic regression response variable) shows that runs scored, runs given up, on-base percentage, and slugging do have strong effects on team success of wins and making the playoffs. The in-sample best models of supervised logistic regression techniques all show great results with AUC values all over 0.90 while the unsupervised k-means clustering technique showed that the data can be effectively grouped in 3 clusters. A mix of supervised and unsupervised study techniques show that a variety of statistical techniques can be used to analyze baseball data.

Executive Summary

After performing exploratory data analysis and deciding to use 2 methodologies of multiple linear regression of having wins as a response variable and logistic regression of having playoffs as the response variable, techniques such as linear regression, generalized logistic regression, regression and classification trees, generalized additive modeling, linear discriminant analysis, and unsupervised clustering using k-means were performed on the moneyball dataset. The below tables compare the in and out of sample MSE(mean squared errors) for the linear regression methodology of wins as the response variable and the misclassification rate and AUC values for the ROC curves will be compared for the logistic regression methodology with whether a team made the playoffs or not as the response variable. The method name for the best models along with in and out of sample MSE's are the columns for the linear regression methodology **(Table 1)** and the method name, in sample MR, out of sample MR, and in and out of sample AUC will be the columns for the logistic regression methodology **(Table 2).** Lastly, the elbow curve **(Figure 1)** for k-means clustering showing the optimal number of clusters to be 3 will be shown for the clustering methodology.

Method	**In Sample MSE**	**Out of Sample MSE**
Multiple Linear Regression	14.41	15.35
Regression Tree	29.32	21.82
Generalized Additive Modeling	13.71	14.89

Table 1: Wins Linear Regression Best Model Comparison

Method	**In Sample MR**	**Out of Sample MR**	**In Sample AUC**	**Out of Sample AUC**
Generalized Logistic Regression (GLM Method)	0.110	0.080	0.966	0.978
Classification Tree	0.050	0.053	0.957	0.979

Generalized Additive Modeling	0.111	0.377	0.966	0.607
Linear Discriminant Analysis(LDA)	0.129	0.113	0.966	0.970

Table 2: Playoffs Logistic Regression Best Model Comparison

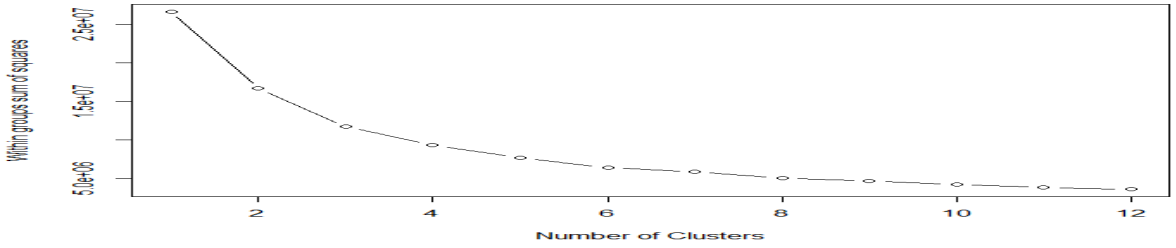

Figure 1: K-means clustering Elbow Curve

Contents

Introduction

Major League Baseball has been regarded as America's pastime since the 1920's and 1930's to the present day. The game has given society so many fond memories ranging from Babe Ruth and the championship Yankee teams to Ted Williams being the only hitter to hit over .400 in a season to Willie Mays's over the shoulder catch in the 1954 World Series [(4)]. Additional memories include Hank Aaron breaking the career home run record in 1974, the Big Red Machine in the 1970's, the Yankees dominance in the 1990's and the present day era memories[(4)]. There are presently 30 teams in Major League Baseball: 15 in the National League and 15 in the American League. [(4)] It is every team's manager, general manager, players, and fans of a team goal for a team to be successful by making the playoffs, World Series, and ultimately winning championships. To reach this goal, teams must win games on the field and organizations must do whatever possible to put the best product on the field. Plenty of factors including runs scored, runs given up, previous team records, on base percentage, number of home runs, walks, strikeouts, and many other factors help those involved in decision making optimize a team's chances of winning games. It has been known that baseball is the sport that is thriving in data analytics and is continuing to grow while causing the growth of analytics in other professional sports all over the world.

Inspiration and Reason for Topic Dataset

It is a well-conceived notion that money and power always lead to success and glory in life. Plenty of individuals have the vision of being rich and powerful and leading a happy successful life. This may seem to be true as the New York Yankees are the richest organization in baseball with an annual payroll of well over $100,000,000 [(1)] and have the most World Series championships with 27. [(2)] The second ranked team with the most World Series championships is the St. Louis Cardinals with 11 championships [(2)] not even half of that of the New York Yankees. The concept of moneyball proves that money doesn't always lead to wins and success in baseball. The Oakland Athletics and General Manager Billy Beane were one of the most successful baseball franchises from 2001 to 2005 [(4)] and were successful despite being in the bottom 5 or bottom 10 in league payrolls in most of these seasons.[(3)] For these years, the Athletics had payrolls between 30 and 40 million from 2001 to 2002 and just above 50 million from 2003 to 2005. The Athletics were consistently picked to finish with a non-competitive record during this time period being in a tough AL West division with the Seattle Mariners, Anaheim Angels, and Texas Rangers. The Oakland Athletics employed this concept of moneyball to compete against these higher market teams by looking at factors that truly optimize the value of a player. In contrast to conventional thinking of home runs, batting average, strikeouts, and power leading to the value of a player, moneyball uses analytics with on-base percentage and slugging to better determine the value of a player [(5)] The Oakland Athletics studied these concepts in detail valuing players who got on base, drew a lot of base on balls(walks), had high slugging percentages, and pitchers who prevented opponents from getting on base .[(5)] This organization studied player salaries along with these statistics and worked to obtain players who were very undervalued and overlooked by other teams but extremely useful to sign at a cheaper price to fulfill their salary constraints. In 2001, the Oakland Athletics had one of the most successful seasons in their franchise history going 102-60. They unfortunately blew a 2-0 series lead against the New York Yankees in the ALDS after winning the first two games in New York to lose the series 3-2. The A's developed superstars in 1st baseman Jason Giambi, center fielder Johnny Damon, and relief pitcher Jason Isringhausen. [(4)] Due to the low payroll of the team, the Athletics were unable to resign any of these three stars who were vital roles to the success of the season. All three stars signed lucrative contracts with Jason

Giambi going to the Yankees, Johnny Damon going to the Red Sox, and Jason Isringhausen going to the Cardinals, all these three teams with high payrolls. [4] The Athletics had a very uncertain future ahead that seemed to spell doom. Employing smart baseball data analytics to value the effect of on base percentage and slugging, the Athletics were able to acquire three very undervalued players: Scott Hatteberg, David Justice, and Chad Bradford to replace these stars. [4] All three of these players had blemishes such as nerve damage to Hatteberg making him switch positions from catcher to first base, old age for David Justice, and a side-winding awkward throwing motion for relief pitcher Chad Bradford that didn't attract any other teams. The Athletics were able to sign these players at a cheap price and did so because Hatteberg and Justice had good on base percentage and slugging numbers and Bradford prevented opposing batters from getting on base. [4] Having a lineup that balanced power with hitters such as Eric Chavez and Miguel Tejada with players such as Hatteberg and Justice who got on base along with building their young pitching staff with Mark Mulder, Barry Zito, and Tim Hudson along with Bradford, the Athletics surprised everyone by winning a record 20 straight games in the 2002 games in August and finishing the season at 103-59 winning 1 more game than 2001 and matching the New York Yankees in wins for the 2002 season .[4] The only downside to the moneyball story of the Athletics was lack of major playoff success by losing heartbreaking series in the playoffs. The Athletics didn't win a playoff series until 2006 and haven't won a world championship since the moneyball analytic era. [4] Despite this, the inspirational success this organization has had has shown the world that money and power alone and power statistics don't always lead to success. Looking at important analytic statistics and really seeing what is important such as on base percentage and slugging percentage lead this organization to plenty of regular season success and trips to the playoffs. The Athletics turned into a baseball powerhouse with numerous playoff appearances under Billy Beane employing the moneyball data analytic baseball concept. [4] This story and the practice of really studying data carefully lead to my interest in the dataset chosen to study analytic factors leading to success for baseball organizations.

Data Description and Methods to Use

The dataset being used for this analysis was found on Kaggle and contains Major League Baseball statistics for all the teams from 1962 to 2012. [6] They are 1232 observations and 15 variables: Team, League, Year, Runs Scored, Runs Given Up, Wins, On-Base Percentage, Slugging, Batting Average, Playoffs or Not, Rank Season, Rank Playoffs, Games Played, opponents On-Base Percentage, and Opponents Slugging Percentage. [6] The Rank Season column ranks the playoff teams by record and the Rank Playoffs column ranks how each team fared in the playoffs. The main questions to answer are what variables are significant for whether or not a team made the postseason and if on base percentage and slugging do contribute to wins and are significant factors in determining whether or not a team makes the playoffs and ultimately wins championships. The tidyverse R package and pipe operator along with the mutate functions in R will be utilized to create a few more variables such as run differential by subtracting runs scored by runs given up and differentials for on base percentage and slugging by subtracting on base percentage by opponents on base percentage and slugging percentage by opponents slugging percentage.

Supervised and unsupervised learning methods will be used in this analysis. For supervised learning, the response variable will be Playoffs which are split into two categories: 0 for a team not making the playoffs and 1 for a team making the playoffs. The first method to be incorporated will be logistic regression on the explanatory variables to see which model is the best in predicting

what variables are significant for whether a team makes the playoffs or not by using stepwise AIC and BIC selection. Generalized additive modeling and linear discriminant analysis will also be incorporated for the response variable of a team making the playoffs or not. To see which factors affect the number of wins, multiple linear regression will be performed with the number of wins as the response variable. Unsupervised learning is also very useful for data that doesn't have a specified response variable. Unsupervised learning will be employed by not having a defined and labeled response variable to perform clustering analysis. Association rules looking at groups of teams with lift and support above certain thresholds to see how successful those certain teams are could be performed for future analysis. K-means clustering can group similar teams in a certain number of clusters to see how teams are grouped together regarding the variables present in the dataset and will be the method utilized for this analysis. There are plenty of possible directions for analysis of this dataset. Baseball analytics has been a force in the field of data analytics and is continuing to grow in scope.

Exploratory Data Analysis

Major League Baseball datasets such as the one used for this analysis are extremely useful for statistical analysis and comparison of models using multiple linear regression, generalized linear regression, regression and classification trees, generalized additive models using smoothing parameters, linear discriminant analysis for a binary response variable and unsupervised learning methods such as k-means and hierarchical clustering. The model results and comparisons will be shown in the analysis of this report showing in sample and out of sample errors for the multiple linear regression models with number of wins as the response variable and the in and out of sample confusion matrices, misclassification rates and ROC curves with AUC values will be compared for models for logistic regression with the binary response variable of whether a team made the playoffs or not as the response variable. As mentioned earlier, the original dataset had 1232 observations and 15 variables. The variables were mostly numerical and provided great information, but to be more thorough, it made sense to create three new variables: RunDiff by taking the difference between a team's runs scored and runs given up in the regular season, OBPDiff by taking the difference between a team's on base percentage and opponents on base percentage, and SLGDiff by taking the difference between a team's slugging percentage and opponents slugging percentage. To view teams that had the highest and lowest values for these new variables, these variables were arranged in descending order. An additional variable, OPS combining the on base percentage and slugging percentage may be created in the next steps of this project for comparison and possible analysis using association rules. The three tables below will show the top 10 teams with the highest RunDiff, OBPDiff, and SLGDiff respectively.

	Team	League	Year	RS	RA	RunDiff	W	OBP	OOBP	OBPDiff	SLG	OSLG	SLGDiff
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	NYY	AL	1998	965	656	309	114	0.364	0.331	0.0330	0.460	0.419	0.0410
2	SEA	AL	2001	927	627	300	116	0.360	0.301	0.0590	0.445	0.378	0.0670
3	BAL	AL	1969	779	517	262	109	0.343	0.331	0.0120	0.414	0.419	-0.0050
4	HOU	NL	1998	874	620	254	102	0.356	0.331	0.0250	0.436	0.419	0.0170
5	CIN	NL	1975	840	586	254	108	0.353	0.331	0.0220	0.401	0.419	-0.0180
6	ATL	NL	1998	826	581	245	106	0.342	0.331	0.0110	0.453	0.419	0.0340
7	OAK	AL	2001	884	645	239	102	0.345	0.308	0.0370	0.439	0.380	0.0590
8	LAD	NL	1974	798	561	237	102	0.342	0.331	0.0110	0.401	0.419	-0.0180
9	ARI	NL	1999	908	676	232	100	0.347	0.320	0.0270	0.459	0.402	0.0570
10	CIN	NL	1976	857	633	224	102	0.357	0.331	0.0260	0.424	0.419	0.0050

Table 1: Highest Run Differential

	Team	League	Year	RS	RA	RunDiff	W	OBP	OOBP	OBPDiff	SLG	OSLG	SLGDiff
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	SEA	AL	2001	927	627	300	116.0	0.360	0.301	0.0590	0.445	0.378	0.0670
2	BOS	AL	2007	867	657	210	96.0	0.362	0.314	0.0480	0.444	0.392	0.0520
3	NYY	AL	2002	897	697	200	103.0	0.354	0.309	0.0450	0.455	0.395	0.0600
4	BOS	AL	2004	949	768	181	98.0	0.360	0.318	0.0420	0.472	0.408	0.0640
5	NYY	AL	2003	877	716	161	101.0	0.356	0.314	0.0420	0.453	0.407	0.0460
6	ARI	NL	2002	819	674	145	98.0	0.346	0.305	0.0410	0.423	0.397	0.0260
7	CHC	NL	2008	855	671	184	97.0	0.354	0.316	0.0380	0.443	0.395	0.0480
8	CLE	AL	1996	952	769	183	99.0	0.369	0.331	0.0380	0.475	0.419	0.0560
9	NYY	AL	2006	930	767	163	97.0	0.363	0.326	0.0370	0.461	0.413	0.0480
10	BOS	AL	2002	859	665	194	93.0	0.345	0.308	0.0370	0.444	0.385	0.0590

Table 2: Highest On-Base-Percentage Differential

	Team	League	Year	RS	RA	RunDiff	W	OBP	OOBP	OBPDiff	SLG	OSLG	SLGDiff
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	BOS	AL	2003	961	809	152	95.0	0.360	0.327	0.0330	0.491	0.415	0.0760
2	ATL	NL	2003	907	740	167	101.0	0.349	0.327	0.0220	0.475	0.401	0.0740
3	NYY	AL	2009	915	753	162	103.0	0.362	0.327	0.0350	0.478	0.408	0.0700
4	SFG	NL	2002	783	616	167	95.0	0.344	0.319	0.0250	0.442	0.372	0.0700
5	BOS	AL	2011	875	737	138	90.0	0.349	0.322	0.0270	0.461	0.392	0.0690
6	TEX	AL	2011	855	677	178	96.0	0.340	0.307	0.0330	0.460	0.392	0.0680
7	SEA	AL	2001	927	627	300	116.0	0.360	0.301	0.0590	0.445	0.378	0.0670
8	CLE	AL	2005	790	642	148	93.0	0.334	0.302	0.0320	0.453	0.387	0.0660
9	SEA	AL	1997	925	833	92	90.0	0.355	0.331	0.0240	0.485	0.419	0.0660
10	SEA	AL	1996	993	895	98	85.0	0.366	0.331	0.0350	0.484	0.419	0.0650

Table 3: Highest Slugging Differential

Table 1 displays the finding that all the top 10 teams with the highest run differential won more than 100 regular season games. Winning 100 regular season games in a regular season is a fairly rare and amazing accomplishment for an organization. The goal of a game is to score more runs than an opponent and run differential directly shows this. The concept of moneyball places utmost importance on OBP(on base percentage) and SLG(slugging). **Table 2** shows that three of the top 10 teams in on base percentage differential: the 2001 Seattle Mariners, the 2002 New York Yankees and the 2003 New York Yankees all had over 100 wins. All three teams made the playoffs with the 2001 Seattle Mariners advancing to the ALCS before losing to the New York Yankees and the 2003 New York Yankees advancing all the way to the World Series before falling to the Florida Marlins in 6 games[4]. All of the top 10 teams in **Table 2** had at least 93 wins and two of those teams, the 2007 Boston Red Sox ranked at number 2 in the table and the 2004 Boston Red Sox ranked at number 4 in the table won the World Series[4]. **Table 3** also shows some excellent findings as three of the teams in the table with the highest slugging percentage differential won more than 100 games and three of the teams made the World Series: the 3rd ranked 2009 New York Yankees, the 4th ranked 2002 San Francisco Giants and the 6th ranked 2011 Texas Rangers[4]. The 2009 New York Yankees won the World Series over the Philadelphia Phillies in 6 games and the 2002 San Francisco Giants and the 2011 Texas Rangers were so close to winning the world series as both teams lost in 7 games to the Anaheim Angels and St. Louis Cardinals respectively[4]. These tables show that moneyball concepts of looking at on base percentage and slugging percentages are extremely important factors for an organization's success in winning games, making the postseason, and ultimately winning championships.

Some final modifications made to the dataset dealt with missing values and how to categorize the teams RankSeason and RankPlayoffs for teams that didn't make the playoffs. The values for OOBP(opponents on base percentage) and OSLG(opponents slugging percentage) had values of NA for years before 1982 in this dataset as it is likely statistics weren't closely tabulated for an opposing's team on base percentage and slugging and were only calculated for each individual team observation. When values have NA's in dataset, using imputation is the method practiced most commonly to deal with this missingness. Mean or median imputation is most commonly used. For this analysis, median imputation was employed converting the NA values to the median opponents on base and slugging percentage values for the values that had OOBP and OSLG. This is a reasonable method and likely will give values close to the actual values, but some error will be present as each decade has a different style of play and the values of OOBP and OSLG may differ in the 1960's and 1970's as it does in the 1990's and 2000's more power home run ball era. The RankSeason and RankPlayoffs variable columns had the ranks for the standings of playoff teams and where each playoff team fared in the playoffs respectively. Since the majority of teams in Major League Baseball do not make the playoffs each season, most of the values in these columns didn't have variables. These columns were numerically ranked 1, 2, 3, and so on. For all the teams that didn't make the playoffs, the columns would be numerically ranked to 0 which makes it easy to identify the teams that didn't make the postseason for a season. The final major decisions for this dataset were deciding to split this dataset into 80% in sample training and 20% out of sample testing for model comparison. The training dataset had 985 observations and 18 variables and the testing dataset had 247 observations and 18 variables. By analyzing this dataset, it was decided that number of Wins and whether a team made the playoffs were the two most important variables in this dataset. The number of wins is a numerical continuous variable and whether a team made the playoffs or not is a binary categorical variable. For the number of wins, a multiple linear regression model approach, regression tree, and generalized additive modeling will be performed and for whether or not a team made the playoffs, a generalized logistic regression, classification tree, generalized additive modeling, and linear discriminant analysis will be studied. For the dataset, clustering methods to decide and group the teams into an optimal number of clusters will be performed to study the differences between each cluster group.

This moneyball dataset has Year, RS, RA, RunDiff, W, OBP, OOBP, OBPDiff, SLG, OSLG, SLGDiff, BA, RankSeason, RankPlayoffs, G and the Team, League, and Playoffs variables are all categorical. The summary statistics for the in sample training split of the numerical variables are shown below in **Table 4.**

Variable	**Min**	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**
Year	1962	1976.0	1990.0	1989.0	2002.0	2012.0
RS	463.00	652.00	711.30	715.30	774.00	993.00
RA	472.00	649.00	709.00	714.40	773.00	1103.0
RunDiff	-331.00	-71.000	3.0000	0.8934	74.000	300.00
W	40.000	73.000	81.000	81.000	89.000	116.00
OBP	0.2770	0.3170	0.3260	0.3264	0.3370	0.3690
OOBP	0.2940	0.3310	0.3310	0.3314	0.3310	0.3840
OBPDiff	-0.0540	-0.0170	-0.0060	-0.0050	0.0070	0.0590
SLG	0.3010	0.3750	0.3960	0.3974	0.4200	0.4910
OSLG	0.3460	0.4190	0.4190	0.4191	0.4190	0.4990

SLGDiff	-0.118	-0.0460	-0.0240	-0.0220	0.0030	0.0760
OPS	0.5840	0.6940	0.7220	0.7237	0.7550	0.8510
OOPS	0.6460	0.7500	0.7507	0.7500	0.7500	0.8830
OPSDiff	-0.166	-0.0610	-0.0310	-0.0270	0.0070	0.1260
BA	0.2140	0.2510	0.2600	0.2593	0.2680	0.2940
RankSeason	0.0000	0.0000	0.0000	0.6173	0.0000	8.0000
RankPlayoffs	0.0000	0.0000	0.0000	0.5340	0.0000	5.0000
G	158.00	162.00	162.00	161.90	162.00	165.00

Table 4: Numerical Variable Summary Statistics Training Data

Of the teams in the in sample training dataset, 194 teams made the playoffs and 791 failed to make the playoffs.

It is really useful to see how numerical values are related to each other in a dataset by seeing if one variable has a strong effect on another variable. Plotting a correlation matrix is the most effective way to apply these findings. **Figure 1** shows strong correlations between some variables. Some strong positive correlations present in **Figure 1** are RS and OBP with a correlation of 0.90, RS and SLG with a correlation of 0.91, RS with SLGDiff and BA with correlations of 0.82 and 0.83 respectively, RunDiff with W having a 0.94 correlation, OBP with SLG with a correlation of 0.79, OBP with BA with a correlation of 0.84 and OOBP with OSLG with a correlation of 0.84. The variables OPS taken by adding OBP and slugging along with OOPS(Opponents OPS) had a few correlations with variables other than OBP, SLG, OOBP, OSLG, OBPDiff and SLGDiff. OPS was highly correlated with RS and BA with correlations of 0.96 and 0.85 while OPSDiff(OPS – OOPS) was also highly correlated with RS with a correlation of 0.84.All these values listed are correlation coefficients which take on a value between -1 and 1 to see how strongly two variables are correlation. Usually, correlations with values higher than 0.8 and -0.8 have high strong positive and strong negative correlations respectively.

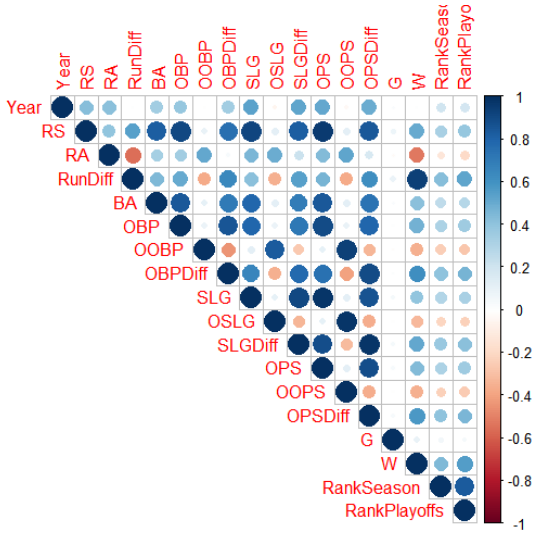

Figure 1: Correlation Matrix of Numerical Variables

By studying the summary statistics in **Table 4** and the histograms of the numerical variables in **Figure 2,** one can see the distribution and skewness for each of the variables in the histogram.

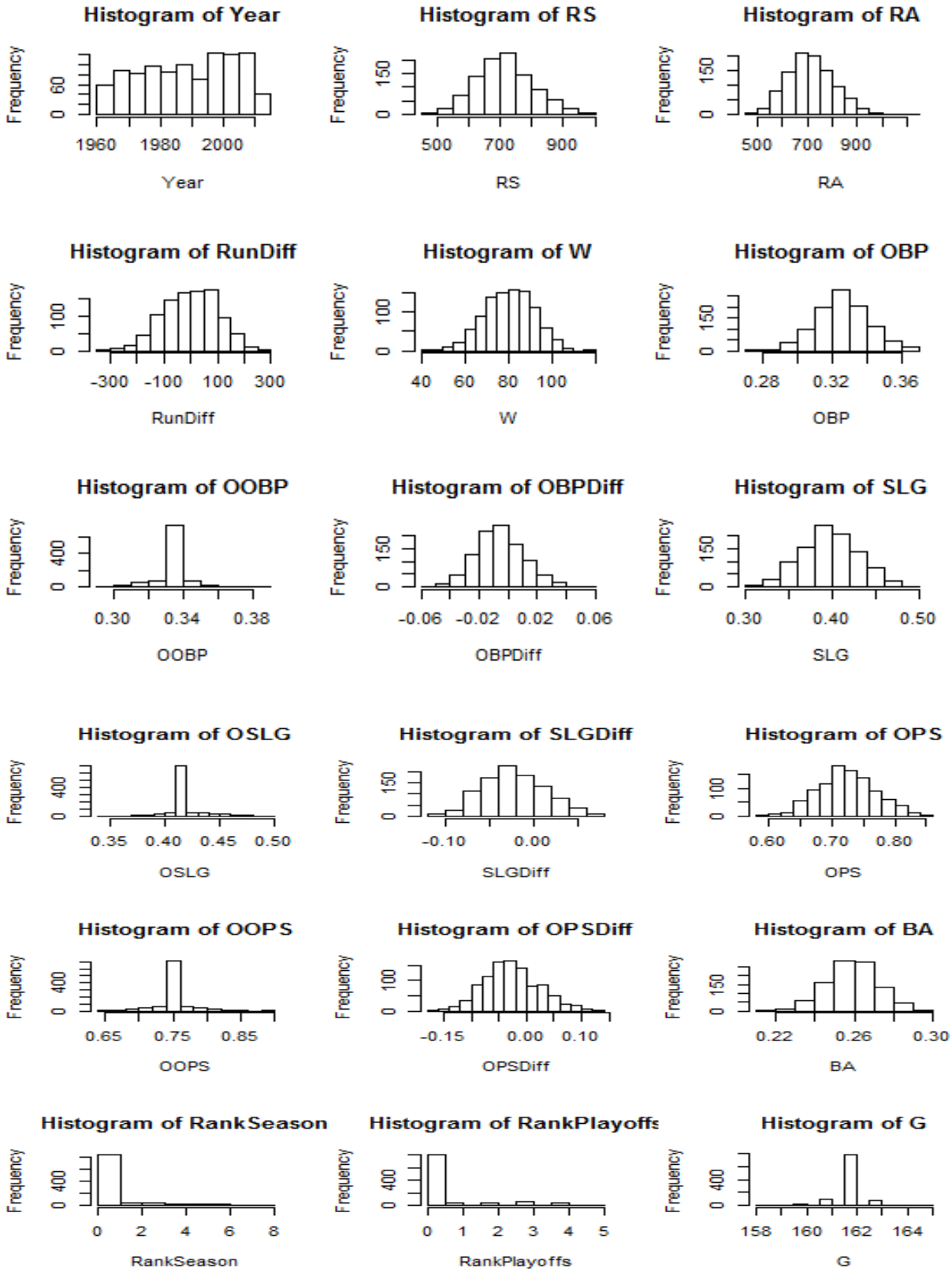

Figure 2: Histograms of In Sample Numerical Variables

Figure 2 shows a nice representation of the numerical variables of the in sample dataset. From this figure and the summary statistics, one can conclude that run differential is slightly left skewed with mean value less than the median and the RankSeason and RankPlayoff variables are right skewed with mean values greater than the median values. The rest of the variables show a fairly normally skewed distribution with similar mean and median values.

Logistic Regression and Multiple Linear Regression

As mentioned previously, the models in this analysis will be analyzed in two categories: one as logistic with Playoffs as the binary response variable and one as multiple linear regression with Wins as the response variable. Winning baseball games to lead organizations to the playoffs and further glory of possibly making it to the World Series to win a title is the most important goal of every organization. The optimal model found for logistic regression using p-values less than 0.05 being significant and the stepwise BIC method led to the model of **Playoffs~Wins+Year** for the in sample training data and **Playoffs~Wins+Year+G** for the out of sample testing data. **Table 5** shows the confusion matrix for this in sample training data. The in sample misclassification rate was found to be 0.11. **Table 6** shows the confusion matrix for the out of sample testing data which had a misclassification rate of 0.08.

Predicted		
Truth	0	1
0	696	95
1	14	180

Table 5: In Sample Confusion Matrix GLM Method

Predicted		
Truth	0	1
0	180	17
1	3	47

Table 6: Out of Sample Confusion Matrix GLM Method

The in sample ROC curve and AUC value of 0.966 is shown in **Figure 3** while the out of sample ROC curve and AUC value of 0.978 is shown in **Figure 4.**

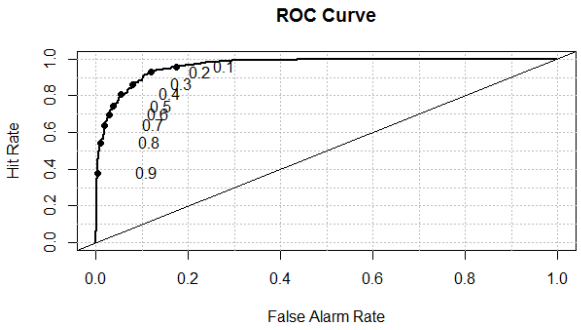

Figure 3: In Sample GLM ROC Curve

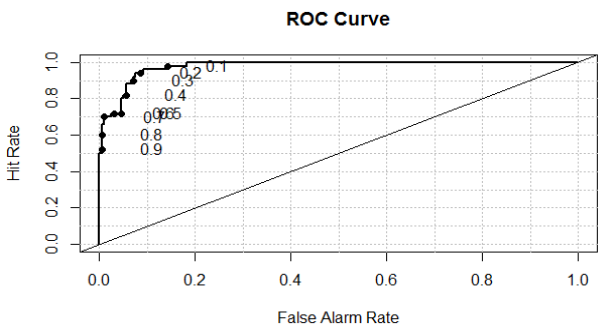

Figure 4: Out of Sample GLM ROC Curve

The multiple linear regression analysis using the lm command shows the optimal model for the in sample data to be **W~RS+RA+Playoffs** and to be **W ~ RS + RA + G + Playoffs** for the

out of sample data. The in sample mean squared error for the optimal model is 14.41 and the out of sample mean squared error is 15.35.

Regression and Classification Trees

For the regression tree with wins as the response variable, the variables RS and RA help to branch the trees to a certain number of wins. The regression tree for in sample data showing branching is shown in **Figure 5** below.

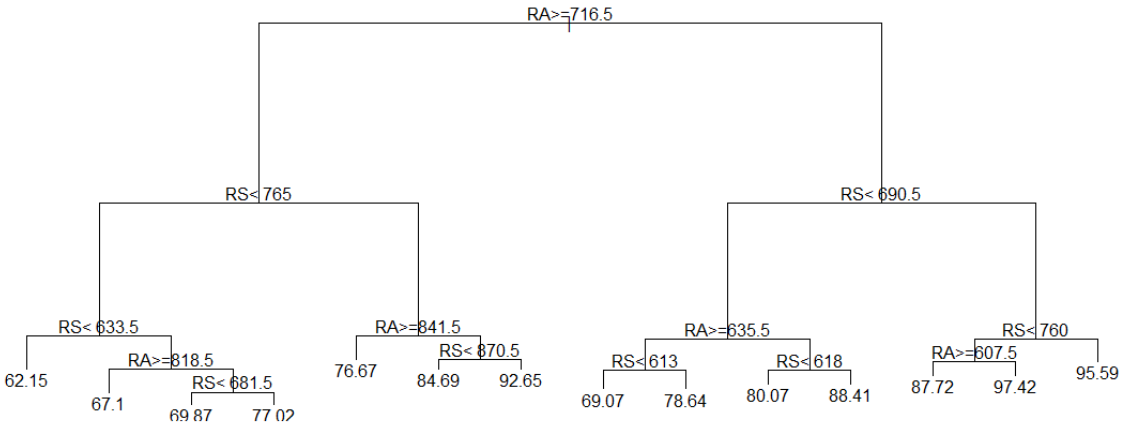

Figure 5: Regression Tree for Wins In Sample Data

Figure 6 shows the complexity parameter for the in sample data to be about 0.016 corresponding to 10 splits and a tree size of 12.

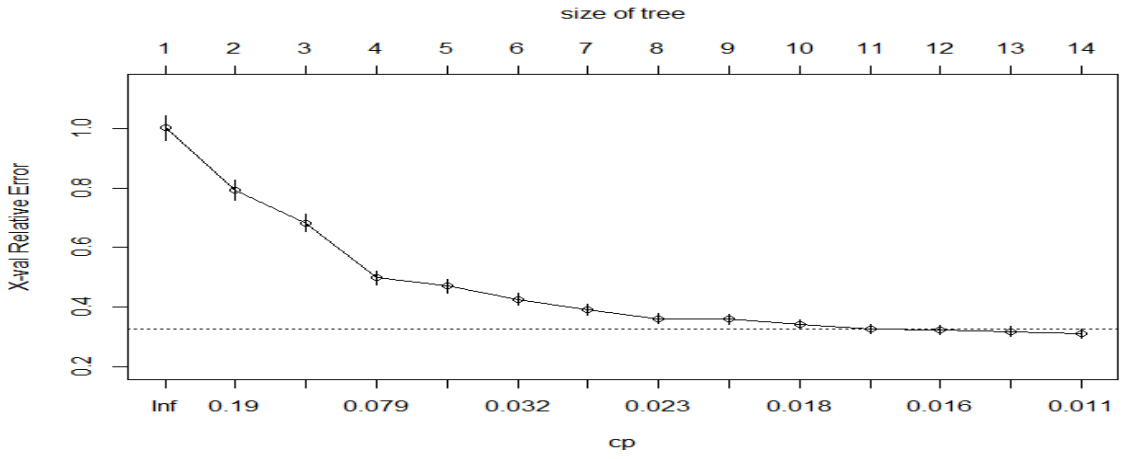

Figure 6: Complexity Parameter and Tree Size of Wins Response Variable

The mean squared error for the in sample regression tree is about 29.32 and is about 21.82 for the out of sample testing data.

For the logistic regression model with playoffs as the response variable, **Figure 7** shows the complexity parameter to be 0.17 and the size of tree to be 2.

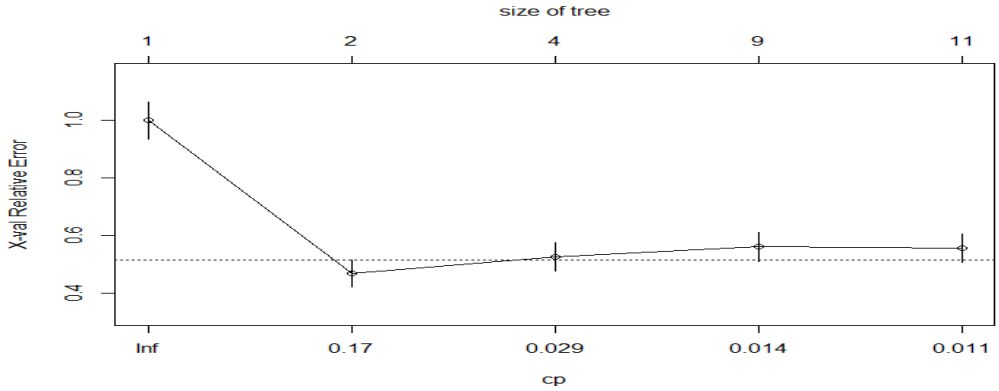

Figure 7: Complexity Parameter and Tree Size of Playoffs Response Variable

The in sample misclassification rate is 0.050 and the out of sample misclassification rate is 0.053. **Figure 8** shows the in sample AUC to be 0.957 and **Figure 9** shows the out of sample AUC to be 0.979.

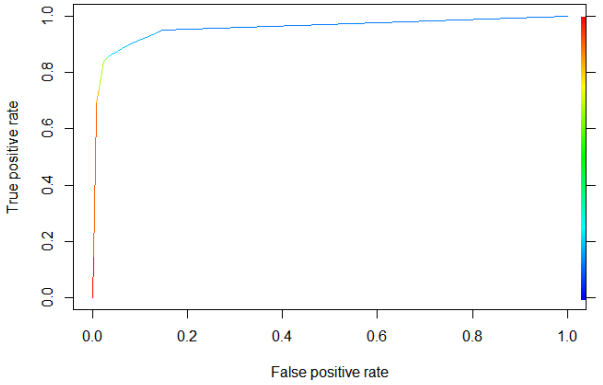

Figure 8: In Sample ROC Classification Tree

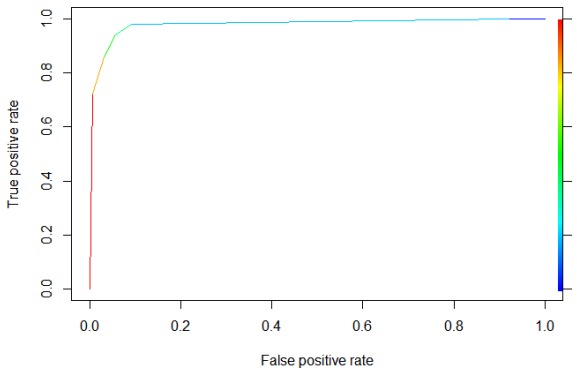

Figure 9: Out of Sample ROC Classification Tree

Generalized Additive Models

Generalized additive models model terms in a non-linear fashion and employ the concept of smoothing numerical continuous variables to reach an optimal model. For wins as the response variable, the optimal model was found to be **W~s(RS)+s(RA)+s(OBP)+s(SLG)+Playoffs+G**. The fact that OBP and SLG are in the optimal model shows strong evidence in this case that the moneyball concept of focusing on OBP and SLG does in fact have a factor in leading to more wins. SLG was on the edge of being significant with a p-value of 0.0541 in the full model. OBP had a p-value of 0.0472 and was also on the edge of being significant. It is usually a judgement call to see if these values should be included in the optimal model. For the optimal model, OBP had a p-value of 0.0412 and SLG had a p-value of 0.1120 and it could be argued that OBP should be included and SLG doesn't needed to be included in the optimal model. The optimal model

showed the edf(expected degree of freedom) of OBP and SLG to be 1.00 meaning that these variables have the potential to be linearized. For the logistic regression model with playoffs as the binary response variable, the gam function had the family = "binomial" argument in the code function and the best model was **Playoffs~s(W)+Year**. The in sample mean squared error for the wins response variable model was 13.71 and the out of sample mean squared error the optimal out of sample model of **W~s(RS)+s(RA)+G+Playoffs** was found to be 14.89.

For the logistic regression optimal models, the optimal cutoff probability was found to be 0.20 using the search grid and the in sample misclassification was found to be 0.111. The out of sample best model of **Playoffs~Team** misclassification rate was found to be 0.377. **Figure 10** shows the in sample AUC to be 0.966 and **Figure 11** shows the out of sample AUC to be 0.607.

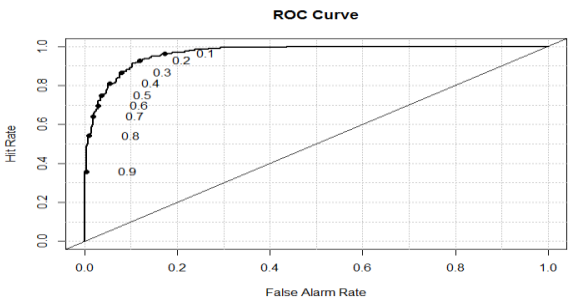

Figure 10: GAM In Sample ROC

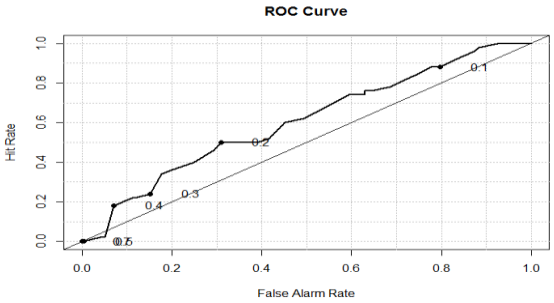

Figure 11: GAM Out of Sample ROC

Linear Discriminant Analysis

The goal of linear discriminant analysis is to group data together to have the in group data to be as similar as possible to data within that group and data outside that group to be as different as possible as in group data. This is a useful method to compare these metrics and is used when the response variable has a binary category as in logistic regression. The in sample misclassification rate for this analysis was found to be 0.185 and the out of sample misclassification rate was found to be 0.162. **Figure 12** shows the in sample AUC to be 0.966 and **Figure 13** shows the out of sample AUC to be 0.970.

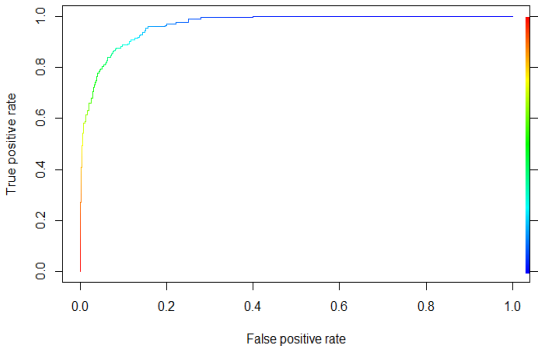

Figure 12: In Sample ROC LDA

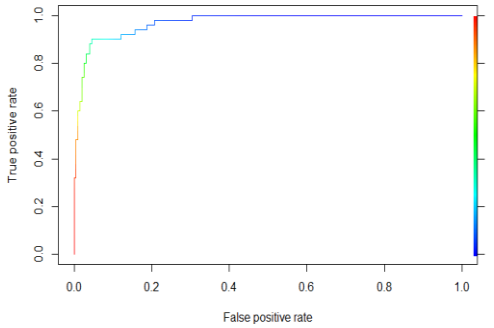

Figure 13: Out of Sample ROC LDA

Clustering: Unsupervised Learning

When there is no specific label for a y or response dependent variable, the type of learning is classified as unsupervised learning. This is in contrast to the earlier supervised method learning methods of multiple linear regression, generalized logistic regression, regression and CART trees, generalized additive modeling, and linear discriminant analysis in which there were response variables labeled as wins for the linear regression cases and playoffs for the logistic regression cases. K-means and hierarchical clustering are effective methods to find the optimal number of clusters to group the data points into. In this analysis, K-means clustering will be used. K-means clustering takes the whole set of data points and uses an algorithm to find the optimal number of clusters to group the data points into. Plotting the within group sum of squares vs Number of clusters and using the elbow curve approach shown in **Figure 14** shows the optimal number of clusters to be 2 or 3 with 3 being a slightly better option as 2 clusters is likely too few for valid comparisons to be analyzed. To confirm the optimal number of clusters, plotting the silhouette width vs the optimal number of clusters in **Figure 15** shows the optimal number of clusters to be 2 or 3, with 3 being chosen to have enough variability in comparison. The number of clusters with optimal silhouette width should be the optimal number of clusters chosen for analysis.

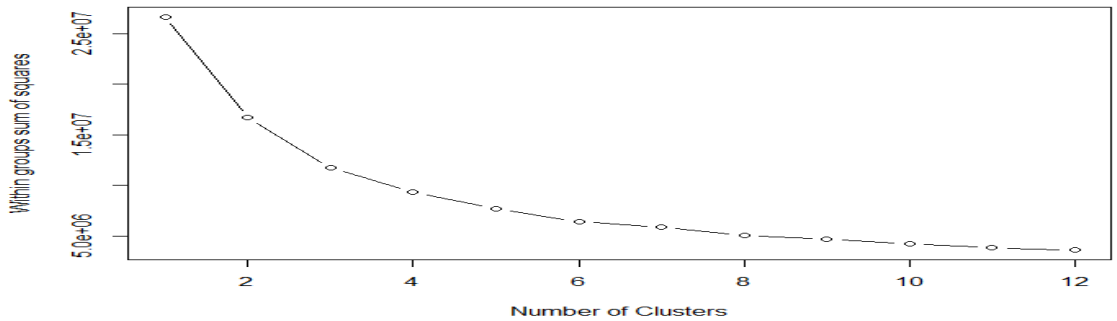

Figure 14: K-means clustering elbow curve

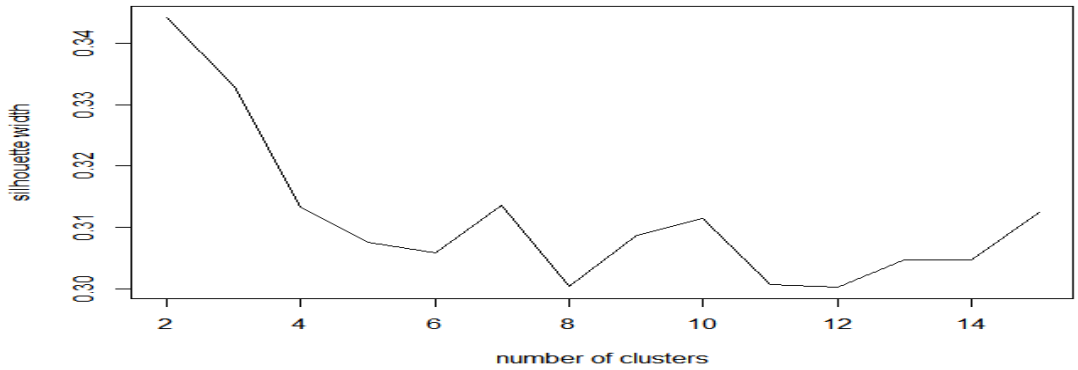

Figure 15: Silhouette Width vs Number of Clusters

For the in sample 80% training data, the 3 group clusters showed there to be 311 observations in Cluster 1, 337 observations in Cluster 2 and 337 observations in Cluster 3. The mean for the numerical variables for each of these clusters are shown below in **Table 7.**

Group	Year	RS	RA	RunDiff	SLG	OSLG	
1	1992.344	694.3023	799.9035	-105.60129	0.3942862	0.4268039	
2	1993.050	803.2077	713.1513	90.05638	0.4254036	0.4161869	
3	1981.828	646.8309	636.8220	10.00890	0.3721662	0.4147715	
	SLGDiff	OBP	OOBP	OBPDiff	OPS	OOPS	OPSDiff
1	-0.032517685	0.3230675	0.3362701	-0.01320257	0.7173537	0.7630740	-0.04572026
2	0.009216617	0.3396914	0.3292938	0.01039763	0.7650950	0.7454807	0.01961424
3	-0.042605341	0.3162641	0.3290089	-0.01274481	0.6884303	0.7437804	-0.05535015
	G	BA	W	RankSeason	RankPlayoffs		
1	161.8650	0.2569293	70.06109	0.04180064	0.02572347		
2	162.0059	0.2693887	90.03264	1.39169139	1.24925816		
3	161.9258	0.2514985	82.05045	0.37388724	0.28783383		

Table 7: Mean values for each of the 3 clusters

By looking at **Table 7,** one can see that group 2 has the highest RS, RunDiff, SLG, OBP, OPS, BA, lowest RA and teams in that cluster having an average regular season record of 90-72. Cluster 3 performed the second strongest and Cluster 1 was clearly the weakest of the three clusters having the lowest RunDiff and having the highest RA, OSLG, OOBP and OOPS indicating that teams in Cluster 1 had anemic pitching staffs. Cluster 1 did seem to outperform Cluster 3 in hitting, but the anemic pitching in Cluster 1 is the reason that the teams averaged a record of 70-92 as compared to a record of 82-80 for teams in Cluster 3. The moneyball methodology of clustering teams with the strongest OBPDiff and SLGDiff into the strongest groups shows that the moneyball concept can apply really well to clustering baseball analytics analysis.

Conclusion and Possible Future Analysis

The number of runs scored and runs given up are the factors that ultimately decide whether teams win games or not and the number of wins a team accumulate in a season decide whether or not a team makes the postseason or not. RS (runs scored) and RA(runs given up) show up as explanatory variables in nearly every optimal model for the linear regression methods with wins as the response variable. The number of wins is the strongest factor in determining in whether or not a team makes the postseason and is present as an explanatory variable in almost all of the optimal models for the logistic regression test. The k-means clustering test grouped the massive

50 years of baseball team data into 3 well organized clusters: Cluster 1 grouping overall poor teams, Cluster 2 grouping overall good teams, and Cluster 3 grouping overall average teams. Multiple linear regression and generalized additive modeling both performed well and close to each other in metrics in terms of mean squared error for in and out of sample testing with regression trees performing a bit worse. For logistic regression, all of the in and out of sample misclassification rates and AUC curves performed really well with the exception being the out of sample generalized additive modeling with a misclassification rate of 0.377 and AUC of 0.607. The team variable being the only significant variable for the binary playoffs response variable is the likely cause of these out of sample statistics. Further supervised model techniques such as neural networks and support vector machines can be run to see the metrics of mean squared error, misclassification rates, and AUC values of optimal models while hierarchical clustering, and association rules unsupervised learning models can be further studied to group teams and find differences of teams between and within groups. It is great to see the growth of analytics in sport and the possibilities are endless to dive into to find important factors leading to the success of organizations.

References

1. "New York Yankees Payroll." http://www.spotrac.com/mlb/new-york-yankees/payroll/ Spotrac 2018. Web. 20 May 2018.

2. "Number of World Series championships won by team from 1903 to 2017." https://www.statista.com/statistics/235618/mlb--number-of-world-series-championships-by-team/. Number of World Series championships won by team from 1903 to 2017. Statista 2017. 22 May 2018.

3. "MLB Team Payrolls." http://www.stevetheump.com/Payrolls.htm. Steve's Menu 2018. 22 May 2018.

4. "Baseball Stats and History The complete source for current and historical baseball players, teams, scores and leaders." https://www.baseball-reference.com/ Baseball Info Solutions, 2010-2018. 23 May 2018.

5. Martyn, Paul. "Supply Chain Analytics: 'Field of Dreams, ' Moneyball and Beyond." https://www.forbes.com/sites/paulmartyn/2018/04/03/supply-chain-insights-from-baseball-field-of-dreams-moneyball-and-beyond/#37070d8d2787 Logistics and Transportation/Big Data. 3 April 2018. 24 May 2018.

6. Duckett, Wes. "Moneyball MLB Statistics 1962-2012." https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012. Kaggle 2017. 10 May 2018.