Data Driven Approach to Sports Management:
A Case Study Using Major League Baseball

Jón Ragnar Guðmundsson

Thesis of 30 ECTS credits
Master of Science in Engineering Management

June 2015

Data Driven Approach to Sports Management:
A Case Study Using Major League Baseball

Jón Ragnar Guðmundsson

Thesis of 30 ECTS credits submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science in Engineering Management

June 2015

Supervisors:

Margrét Vilborg Bjarnadóttir, Supervisor,
Assistant Professor of Management Science and Statistics,
University of Maryland, USA.

Páll Jensson, Supervisor,
Professor, Reykjavík University, Iceland.

Examiner:

Sigrún B. Gunnhildardóttir, Examiner,
MSc. Engineer, AGR Ltd.

Data Driven Approach to Sports Management:
A Case Study Using Major League Baseball

Jón Ragnar Guðmundsson

30 ECTS thesis submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science in Engineering Management.

June 2015

Student:

_____

Jón Ragnar Guðmundsson

Supervisors:

_____

Margrét Vilborg Bjarnadóttir

_____

Páll Jensson

Examiner:

_____

Sigrún B. Gunnhildardóttir

The only real game, I think, in the world is baseball.

Babe Ruth

Abstract

Baseball is considered to be the national sport of the USA but its popularity has declined in the last few years, mostly due to people's interests in other sports. Not many sports come close to baseball regarding statistical analysis where everything concerning the sport is carefully registered. There is one statistic variable who has gained more attention in the later years. This variable combines all the statistics of a player into one number which depicts how many wins that player adds to the definite minimum of a replacement. This variable is called wins above replacement or WAR.

This project endeavors to see if the possibility to use WAR, exists to predict if a team reaches the playoffs or not. It subsequently attempts to see if this variable is equipped to create optimization model, which should simplify coaches' decisions in signing players.

Data from 1969 up until 2014 were used to create a database where players had been connected to the team they had started each season with. Statistical data for upcoming season were minimal in processing the data.

The study's results indicate the possibility to use WAR up to a certain point to predict whether a team qualifies to the playoffs or not. A logistic regression provided a model with a TPR close to 60% and an accuracy of approximately 70%. These are rather high percentages when there were several factors which limited the value of the calculations. Eliminating those factors would result in decreased errors and more accurate calculations.

Key words: WAR, baseball, playoffs, position players, pitchers.

Útdráttur

Hafnabolti hefur verið álitin þjóðaríþrótt Bandaríkjanna en hefur átt undir högg að sækja gagnvart öðrum íþróttum á síðustu árum. Það eru þó fáar íþróttir sem komast í námunda við hafnabolta þegar kemur að tölfræðilegum skráningum þar sem allt er viðkemur íþróttinni er nákvæmlega skráð. Þó er ein tölfræðibreyta sem hefur verið hvað mest á milli tannanna á fólki á síðustu misserum. Þessi breyta tekur saman alla tölfræði leikmanns og sameinar í eina tölu sem á að gefa til kynna hversu marga sigra ákveðinn leikmaður bætir við umfram ákveðinn lágmarks varamann. Þessi breyta kallast á ensku *wins above replacement* (WAR) eða sigrar umfram varamann.

Þetta verkefni leitast að því að athuga hvort möguleiki er á því að nýta WAR til þess að spá fyrir um hvort lið komist í úrslitakeppni eða ekki. Einnig er leitast eftir því að athuga hvort hægt er að nota þessa tölfræðibreytu til þess að útbúa bestunarlíkan sem einfaldar ákvarðanir þjálfara liða við ráðningar á leikmönnum.

Gögn frá árinu 1969 fram til ársins 2014 voru skráð í gagnabanka þar sem leikmenn voru tengdir við þau lið sem þeir byrjuðu hvert tímabil hjá. Tölfræðilegum upplýsingum fyrir komandi tímabil var haldið í lágmarki við úrvinnslu gagnanna.

Niðurstöðurnar benda til þess að hægt er að notast við WAR að einhverju leiti til þess að spá fyrir um hvort lið komist í úrslitakeppni eða ekki. Tvíundagreining sýndi fram á líkan sem var með tæplega 60% TPR og rúmlega 70% nákvæmni. Er það nokkuð hátt hlutfall í ljósi þess að nokkrir þættir takmörkuðu gildi útreikninganna. Því ætti að vera hægt að vinna enn frekar að því að lágmarka skekkjur og nákvæmni útreikninganna með því að útiloka þessa þætti.

Lykilorð: WAR, hafnabolti, úrslitakeppni, útileikmenn, kastarar.

Acknowledgment

First of all, I would like to thank my supervisor Margrét Vilborg Bjarnadóttir, Assistant Professor of Management Science and Statistics at University of Maryland, for the opportunity she presented me, to work on this project. Despite being in different country the whole time, she managed to support my through this work.

I would also like to thank my supervisor, Páll Jensson, Professor at Reykjavík University, for his support and contribution in making this study a reality. He was always ready to assist even though he did not always have fully understanding of how the game of baseball works.

I would like to thank Sean Barnes, Assistant Professor of Operations Management at University of Maryland, for providing me the data needed for this study. I am also very thankful for his time and effort to help me understanding the data and the art of baseball.

I would like to thank my friends Bjarni Salvarsson, Consultant at Miracle, and Þröstur Pétursson, Engineer at Landspítali, for all the time and assist they contributed to make this study become reality.

Furthermore I would like to thank my friend Elvar Örn Stefánsson for his time and effort, reading this study to ensure the English grammar is acceptable and appropriate.

List of Tables

List of Figures

Contents

1. Introduction

1.1. Background

Baseball is one of the most popular sports in the USA and is considered to be its national sport. In eyes of many, baseball is just a simple sport where the main purpose is to hit a ball with a bat. However, that can't be more far from the reality and the ideology behind the game. Hitting a ball is considered by many experts and professionals to be "the hardest thing to do in sports" (See figure 11 in A2).

> *"When a big league pitcher throws a 90 mph fastball, a batter has less than a quarter second to see the pitch, judge its speed and location, decide what to do, then start to swing. The bat must meet the ball within an eighth of an inch of dead center and precisely the right millisecond as the 3-inch spinning sphere whizzes by."* [1]

Apart from batting there are numerous factors and situations which affect the outcome of the game. Everything is documented and has been since the first box score appeared in the 1860s. That makes baseball one of the best and most detailed documented sport to be found. The statistical analyzing behind players' and team's performances is constantly evolving with various usability. One aspect of this evolvement is the variable *wins above replacement* (WAR). This variable attempts to combine all player's performances into a single number which is meant to demonstrate how well a player is really playing. This project's research will be based on this variable in order to check whether it is possible to use it as a predictor for teams' performances. Most researches that have used WAR have almost without exception been focusing on the relationship between payrolls and WAR. Therefore, this project's research focus is on an aspect that have hardly ever been observed before.

There is a lack of good and useful information about how WAR is calculated. Most of the information can be found in various websites and books where it is briefly described in words without mathematical presentation. A chapter in this thesis attempts to define and explain in the most accurate and detailed way on how the ideology and calculations behind WAR are.

In appendix A1 there are tables which explain baseball's statistical abbreviations. The author recommends the reader to use this appendix in order to fully understand the text and the equations in the study.

1.2. Objective of the Thesis

Which (if any) athletes to hire or recruit is a decision General Managers as well as coaches face each year. Future athletics' performance is highly uncertain, both the quality of the performance as well as the duration. Teams may also have different objective functions, while some define success as being in the top 10, for others anything but the league win may be considered a subpar performance. Therefore is the objective of this study to answer the following question:

"Can a team combination (distribution) of WAR be used to predict whether a team will qualify for the playoffs or not?"

The goal is to create a forecast model which can estimate team's probabilities of advancing to the playoffs based on the WAR of its players.

1.3. Thesis Structure

Chapter 1: Introduction

This chapter briefly provides the reader a basic information about baseball and the WAR variable which the thesis is based on. The thesis object and goal are also presented.

Chapter 2: Baseball the Game

This chapter contains information about the baseball game itself. The game is explained briefly but in detailed words so the reader can get the basic understanding of the game.

Chapter 3: What is WAR?

This chapter attempts to define and explain in the most accurate and detailed way how the ideology and calculations are behind the WAR variable.

Chapter 4: Literature Review

This chapter provides the reader an insight to several other research studies which are somewhat related to this thesis study. The history of baseball statistics is also briefly mentioned.

Chapter 5: Materials and methods

This chapter discusses how the data was processed and all assumptions made in regards to it. The tools used in this study are also mentioned and explained.

Chapter 6: Results

This chapter provides the reader with the results from the study and a short interpretation to them.

Chapter 7: Discussions

This chapter contains information about the author's thoughts of the study's results and the work done prior to it. The flaws of the study are discussed along with the possible future direction.

Chapter 8: Conclusion

This chapter summarizes the study findings.

2. Baseball the Game

2.1. How Baseball Works

Baseball game consists of two teams competing through nine innings. If teams are tied at the end of the ninth inning, more innings are played until one team outscores the other. Inning is a period which is subdivided into halves where both teams switch from being on defense and offense. The visiting team always begins each inning by batting, i.e. offense. Baseball is unique in the matter it does not have a clock but an inning ends when both teams have got three outs. This makes baseball ideal for rather long games but at same time it evaluates strategy more. When a team scores in baseball it does not score points or goals like in most other team sports, it scores runs. Team is considered to have scored a run when a player has run through all bases and back to the home base without getting outed [2]. Player gets an out in five ways [3]:

1. Strike Out - The batter fails to hit the ball in three attempts.
2. Force Out - The ball reaches the base before the player.
3. Fly Out - The ball is caught (after a hit) before it touches any surface.
4. Ground Out - The ball hits the ground and reaches first base before the player.
5. Tag Out - A defensive player is able to touch a player, who is running between bases, with the ball or the glove with the ball in it.

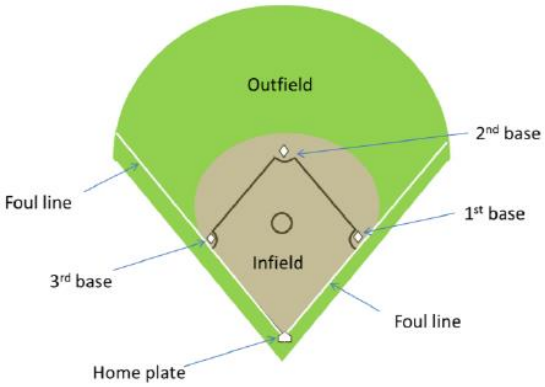

Figure 1: Key parts of a baseball field. Adapted from [4].

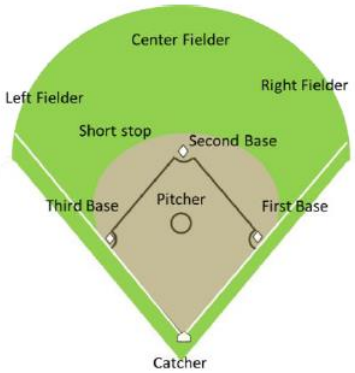

Figure 2: Locations for defensive positions on a baseball field. Adapted from [4].

While on defense, teams can have nine players on the field where each player serves a specific role (see figure 2). All named positions except for catcher, pitcher, and shortstop, field an assigned part of the playing field which holds the same name, e.g. left fielder covers

the left side of the outfield. The pitcher and the catcher are a part of the infield defensive team along with the basemen and the shortstop player. The pitcher pitches (i.e. throws the ball) towards the catcher, who is located behind the batter, in attempt to make the batter strike out. If the batter manages to hit the ball, the defensive team collaborates to get him and other players on bases (called base runners) out with one of the outs mentioned before. On the other hand if the batter hits the ball over the outfield fence he can run through all the bases without interference, which is called home run[1]. When the defensive team has collected three outs then it is its turn to bat [5].

On offense the sole purpose is to advance through all bases and score runs. When a batter is at the plate, i.e. at bat, he tries to hit the ball fairly out of the strike zone, often called to get "*a ball"*. The strike zone is a fictional rectangular box which is defined by the width of the batting plate and the length between the batter's chest and knees. A strike is when a pitch gets through the strike zone unbatted or when *a ball* hits the ground outside the foul lines. If the batter has already collected two strikes and hits the ball outside the foul line he is allowed to bat again until the ball ends within the foul lines or is out. The defenders are allowed to catch *a ball* in the foul territory which results in the batter being out. However, a batter does have other possibilities of reaching first base. If a pitcher throws the ball and hits the batter, has pitched four fair balls without three strikes or a hit, or a defensive player commits a foul, the batter is awarded first base. The second process mentioned is often called base on balls, usually referred to as a *walk*. When a batter gets *a ball* and stops at first base it is called a *single*, a *double* if he gets past first base and over to second base, and a *triple* if he gets to third base through both first and second base [6].

There is one vital rule regarding base runners. Each base may only include one base runner. That forces the base runners to leave their base when another offensive player runs towards it, but they do not have to if there is an empty base or bases behind them. Base runners have several ways of advancing through bases:

- when a batter hits
- with walks
- when a pitcher throws the ball into the batter
- by stealing a base

Base runner who starts running from his base before the batter gets a hit or walk, and is able to reach next base without being tagged out is said to have stolen a base [6].

[1] A batter can also get a home run while the ball is in play on the field but that is exceptionally rare.

2.2. Player Positions

Most of the player positions in baseball have already been mentioned in last subsection. Full list of the positions are shown in table 1 where the positions have been categorized into pitchers and batters along with where on the field each position is played. The positions which have not been discussed and need further explanation will be outlined in this subsection.

Table 1: Detailed list of player positions.

	Batting	**Pitching**
Infielder	1B – First Baseman 2B – Second Baseman 3B – Third Baseman C – Catcher CI – Corner Infielder IF – Inner Fielder MI – Middle Infielder SS – Shortstop	CL - Closer RP – Relief Pitcher SP – Starting Pitcher
Infielder / Outfielder	UT – Utility Player	
Outfielder	CF – Center Fielder LF – Left Fielder OF - Outfielder RF – Right Fielder	
Batter	DH – Designated Hitter	

Table 2: A description for several player positions.

Position	**Description**
Corner Infielder	A player who can play either first or third base [7].
Closer	A relief pitcher who pitches to get the final outs of a game which is close to an end and his team is in the lead.
Designated hitter	A player who is only allowed to play offense and, if used, may replace the starting pitcher and all subsequent pitchers in the batting order. He must be selected before the game and if he is replaced, then the replacer will become a designated hitter. The designated hitter is only used in the American League [8, 9].
Middle Infielder	A player who can play either shortstop or second base [10].
Outfielder	A player who can play in all the outfield positions.
Relief Pitcher	A player who comes into games started by another pitcher [11].
Utility Player	A player who can play multiple positions, both infield and outfield [12].

2.3. The Season

MLB consists of two leagues, the American League (AL) and National League (NL), in which 30 teams participate. Each league is subdivided into three divisions, Central, East, and West, with five teams each. The regular season is from last days in March or early April until last week in September, around 180 days. At this time the teams play 162 games each where the teams in same division play each other at least 17 times. Each club can play up to 20 games against teams in the other league. When interleague teams are competing, the designated hitter can only be used if the game is held at the home of the AL team [13].

Figure 3 shows how the season lengths in USA sports vary, e.g. does a MLS club play at least 34[2] games in 312 days which is far less and over longer season than a baseball team.

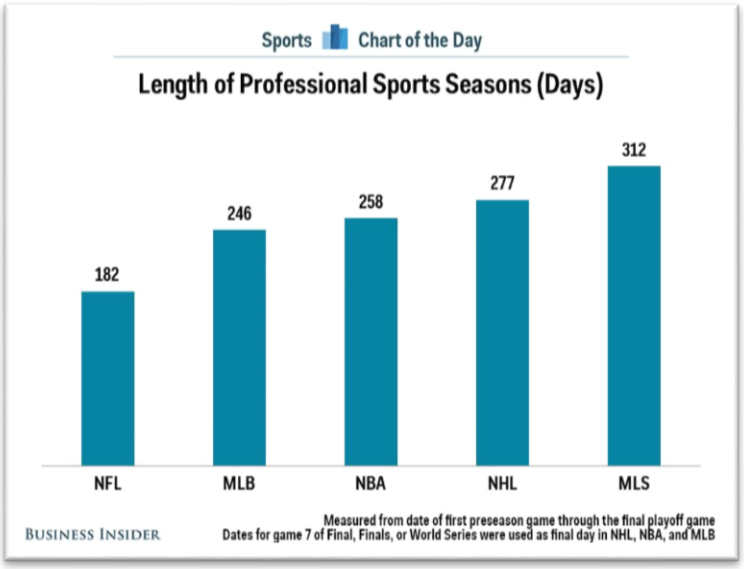

Figure 3: Comparison of full season's lengths in USA sports in 2014. Adapted from [14].

2.4. The Playoffs

In 1969, for the first time, the leagues were both subdivided into two divisions, East and West. That led to an increased number of playoffs teams – four teams instead of two. The divisions' winners within each league competed for the League Championship where the winner would then compete against the winner from the other league in the World Series. This structure remained unchanged until 1994 when the Central division was added to each league and eight teams made the playoffs [15]. The division winners from each league qualified to the playoffs along with the two teams with the best record of the remaining teams

[2] Regular season is 34 games. Some teams also play in the playoffs and/or other domestic and continental cups which increase the number of games.

(one from each league). Those teams were called “wild-card teams” [16]. This league setup has not changed since but the playoffs structure changed before the 2012 season. Then a second wild-card team, from each league, was added to the playoffs, thereby allowing ten teams in the postseason – five from each league. The playoffs formation was similar to the eight team postseason but added a game called “The Major League Baseball Wild Card Game” to the postseason. The two wild-card teams in each league played against each other. The winner from this game would advance to the Division Series [17].

It is necessary to mention two seasons’ alternatives in the playoffs. Both alternatives were due to a baseball players’ strike. The first one occurred in the middle of the 1981 season and caused the playoffs to have different strategy. The season standings were split into two halves and the teams with the best record in their division in each half competed internally to decide which teams should advance to the League Series [18]. The second strike took place in 1994 and resulted in an early season ending without a postseason and shortened the season after [19].

3. What is WAR?

Wins above Replacement (WAR) is an attempt to combine all of baseball player's on-field performances into a single number. Its purpose is to give knowledge about how much better a player is than what a team would typically have to replace that player. WAR simplifies comparison between different positions and unlike many other baseball stats it does not only focus on offense, defense is as crucial as offense. Runs are the foundation of WAR calculations where they are compared to the average player and then to a replacement player [20]. The replacement player in this context is a theoretical player which, according to Fangraphs [21], is a player that can be defined as a

> *"…one who costs no marginal resources to acquire. This is the type of player who would fill in for the starter in case of injuries, slumps, alien abductions, etc."*

He is usually a player who is in the non-roster or is a six year minor league free agent. The idea of using a comparison to a replacement level player instead of an average player can be rationalized by the fact that teams seldom find an average player in midseason to fill in for an unplayable starter. In addition they don't have minimum salaries and are therefore more expensive. Apart from those two reasons, the average player is a poor criterion for comparison as average performances aren't always equal.

In its simplest way, ten runs are equal to one win but that isn't always the fact. The following equation broadly sums up how the runs above replacement (RAR) are computed [20].

$$\begin{aligned} \mathrm{RAR} &= \mathrm{Runs}_{\mathrm{Player}} - \mathrm{Runs}_{\mathrm{Replacement}} \\ &= \left(\mathrm{Runs}_{\mathrm{Player}} - \mathrm{Runs}_{\mathrm{Avg.\ Player}}\right) + \left(\mathrm{Runs}_{\mathrm{Avg.\ Player}} - \mathrm{Runs}_{\mathrm{Replacement}}\right) \\ &= \mathrm{RAA} + \mathrm{RAR}_{\mathrm{Avg.\ Player}} \end{aligned} \tag{1}$$

There are more complex calculations with multiple attributes and factors behind the variables and they will be described in the best possible way in the two following subsections. WAR for position players and pitchers are calculated with different methods as the roles in the game differ substantially. Equations and explanations will be provided when possible but keep in mind there are more than one ways of calculating WAR.

3.1. WAR for Position Players

Wins above replacement for position players is a combination of six factors where the first five factors (a. – e.) are used to calculate runs above average (RAA). The sixth factor, replacement level runs, is found in subsection f.

a. Batting Runs (R_{Bat})

Batting Runs is a good indicator in regards to added or lost value in relation to situational performance. It depends on weighted runs above average (wRAA) which is then built on weighted on-base average (wOBA). wOBA is a linear weighted formula which intends to weight each aspect of hitting in proportion to their real value.

$$\text{wOBA} = \frac{\alpha_1 \cdot \text{uBB} + \alpha_2 \cdot \text{HBP} + \alpha_3 \cdot 1\text{B} + \alpha_4 \cdot 2\text{B} + \alpha_5 \cdot 3\text{B} + \alpha_6 \cdot \text{HR} + \alpha_7 \cdot \text{SB} - \alpha_8 \cdot \text{CS}}{\text{AB} + \text{BB} - \text{IBB} + \text{HBP} + \text{SF}} \quad (2)$$

$\alpha_{1,..,8}$ are weighting coefficients and vary between years, where it is calculated for both leagues and then scaled for each season and league. Then the next step is to calculate wRAA.

$$\text{wRAA} = \frac{\text{wOBA} - \text{wOBA}_{\text{League}}}{\text{wOBA\_Scale}} \cdot \text{PA} \quad (3)$$

$$\text{PA} = \text{AB} + \text{BB} + \text{HBP} + \text{SF} + \text{SH} \quad (4)$$

The $\text{wOBA}_{\text{League}}$ is the league's average for the relevant year and the wOBA_Scale value depends on the coefficients used in wOBA [22]. Generally, the scale changes little, e.g. 1.21 in 2008 and 2009 and then 1.25 in 2010 [23].

b. Baserunning Runs (R_{Baser})

This part demonstrates whether the player's number of runs are better or worse than the average player was for all baserunning events.

Consists of two parts:

- Stolen bases (SB) and caught stealing runs (CS).
- Non-base stealing base running (PB and WP).

Baserunners have three bases to occupy or advance from where the opportunities to advance are multiple and vary; 1st base with 14 opportunities, 2nd base with 11 opportunities and 3rd base with eight opportunities. The guideline for the calculations of the baserunning runs is rather difficult to understand and for those who want to read more about it, and the list of all the opportunities, can find it here [24].

c. Grounded into Double Play Runs (R_{Dp})

This statistics presents the number of times a defense executes a double play without a fielding error on either of the putouts. If a play occurs where the batter hits into double play then he and the pitcher get a ground into double play [25]. In other words, this section finds out how many runs a player contributes more or less than the average he was at avoiding grounding into double plays.

$$R_{Dp} = DP_{Difference} \cdot \left(\left(GIDP_{Player\ Opportunities} \cdot GIDP_{League\ Rate} \right) - GIDP_{Player} \right) \quad (5)$$

$$DP_{Difference} = (DP - DP_{Avoided})_{League\ Average} = 0.44 \quad (6)$$

. Infield ground ball having one runner on first base, less than two outs and with minimum one out recorded on the play is called GIDP player opportunities [24].

d. Fielding Runs (R_{Field})

The formula for fielding runs estimates how many runs a fielder saves. Average fielder contributes zero fielding runs, which means a fielder can have a negative or positive value. There are two different methods used in the calculations, defensive runs saved (DRS) and total zone rating (TZR). DRS was introduced in 2003 and all fielding runs calculations for each season since then use it but all seasons prior to that use TZR.

Total zone rating is based on the available data, sometimes it uses only basic fielding and pitching stats but sometimes there are more detailed play-by-play data available. The play-by-play data often includes information regarding the hit location, where the field has been split into zones. Fielder gets credit if he makes a play on a ball that is outside of his zone [26].

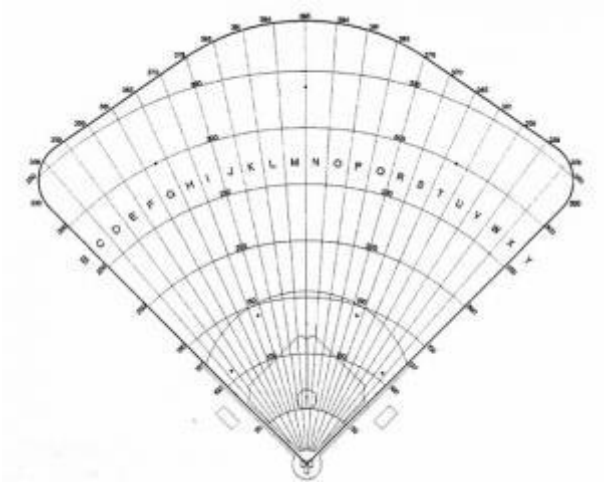

Figure 4: Stats zone chart. Adapted from [26].

TZR consists of four different categories where each applies to certain positions on the field:

- *Fielder's fielding range (R_{Tz})* – The number of runs above or below average the fielder saves based on fielding plays made. In this context the outfield stops and double plays turned are not counted for.
- *Outfield arms (R_{Of})* – The number of runs above or below average an outfielder saves based on baserunner stops and advances. This factor compares the player to the league average of attributes like assists, errors due to an advancement and rate of holding.
- *Double Play runs (R_{Dp})* – The number of runs above or below average an infielder is by turning double plays and giving opportunities. This focuses on the DP rate in situations where there is a groundball fielded by an infielder with a baserunner at first base and less than two outs [27].
- *Catcher defense (R_{Ctch})* – The number of runs above or below average a catcher saves based on baserunner stops and advances. Catchers play a position that requires multiple defensive abilities. Wild pitches, errors, pickoffs, passed balls, stealings caught and stolen bases allowed are all aspects that define the catcher's defense. Like with the other previous mentioned numbers the catcher defense aspects are compared to the league averages before converting it to runs.

When it comes to calculating the total fielding runs there are different variables accounted for, based on the position being calculated.

$$R_{Field_{Infielder}} = R_{Tz} + R_{Dp}$$
$$R_{Field_{Outfielder}} = R_{Tz} + R_{Of} \quad (7)$$
$$R_{Field_{Catcher}} = R_{Ctch}$$

Defensive runs saved is combined of eight factors [24]:

- *Fielding range* – See description in the TZR paragraph.
- *Outfield arms* – See description in the TZR paragraph.
- *Double Play runs* – See description in the TZR paragraph.
- *Good play / bad play values* – This evaluates fielder's efforts to convert a batted ball to an out. This includes

  > *"…28 positive play types like HR-saving catches, backing up a play, blocking a pitch in the dirt and 54 misplays like missing the cutoff man, failing to anticipate the wall and allowing extra bases, not covering a base, pulling a foot off the bag, etc."* [24]
- *Bunt*[3] *fielding* – The number of bunt runs saved. This classifies how a fielder handles bunted balls in play.
- *Catcher SB/CS data*
- *Pitcher SB/CS data*
- *Catcher handling of the pitching staff*

e. Positional Adjustment Runs (R_{Pos})

Some positions are more difficult to play and the positional adjustment is an attempt to compensate this difficulty difference, using a position multiplier. This multiplier is based on the batting performance difference between positions and the fielding performance when players change positions in a game. The calculations surrounding the multiplier lack information but the numbers haven't changed since 2003. In order to calculate R_{Pos}, for positional players, the following equation is used:

[3] Bunting is where a batter places the bat in the strike zone and lets the ball hit it without swinging it.

Table 3: Position multipliers for positional adjustment runs calculations. Adapted from [24].

Position	Position multiplier (runs)
1B	-9.5
2B	3
3B	2
C	9
CF	2.5
DH	-15
LF	-7
RF	-7
SS	7

$$R_{Pos_{Batter}} = \frac{\text{Position multiplier} \cdot \text{Innings played}}{1350} \tag{8}$$

The denominator value represents number of innings ($9 \cdot 150 = 1350$).

Sometimes pitchers also bat so there is a specific approach to calculate their positional adjustment. The pitcher's positional adjustment varies between years and depends on the pitchers performances while they are playing on offense.

$$Pos_{\text{League Adj.}\,\text{Pitcher}} = \left(\frac{-600 \cdot (R_{Bat} + R_{Baser} + R_{Dp} + R_{Rep})}{PA_{Sum}}\right)_{League} = \left(\frac{-600 \cdot R_{Sum}}{PA_{Sum}}\right)_{League} \tag{9}$$

The R_{Sum} and PA_{Sum} in the equation are only numbers from pitchers who bat, not all players in the league. This factor is then used to calculate the positional adjustment runs for individual pitchers by multiplying it to the pitcher's individual plate appearances [24].

$$R_{Pos_{Pitcher}} = Pos_{\text{League Adj.}\,\text{Pitcher}} \cdot PA \tag{10}$$

f. Replacement Level Runs (R_{Rep})

The most recent replacement level factor is 0.294, this indicates the proportion of wins the replacement players would have won, 47.7 games in 162 games [28].

$$\text{Replacement Level} = \text{Wins}_{\text{Ratio}} - \frac{\text{WAR}_{\text{League}}}{\text{Teams} \cdot \text{Games}} = 0.5 - \frac{1000}{30 \cdot 162} = 0.294 \qquad (11)$$

$\text{Wins}_{\text{Ratio}}$ represents ratio for the total number of wins available in a season, which is half of the overall number of games because a baseball game can only end with a win or loss. $\text{WAR}_{\text{League}}$ is set at 1000 league total due to an agreement between Baseball Reference and FanGraphs to establish a common replacement level [29]. Furthermore, the number of wins given by the replacement level is split between pitchers, 41%, and position players, 59%. It is based on the two groups' difference in free agents salaries over the last four seasons [24]. With this breakdown, the replacement level corresponds to 20.5 runs per 600 plate appearances (PA), where this number for runs is called replacement level multiplier.

$$R_{\text{Rep}} = \frac{\text{PA}}{600} \cdot 20.5 \qquad (12)$$

The difference in strength between the two leagues must also be taken into account. The replacement level for MLB and the WAR assigned to each league often varies between years. This distribution can be found in appendix A4. When all calculations have been computed once, the league total WAR is compared to the distributed WAR. If it differs, then a fractional replacement runs are either added or subtracted from each player in relation to their playing time and how it differs between the leagues [24].

g. Runs above Average (RAA) and Runs above Replacement (RAR)

If all previous steps have been completed, then it is possible to calculate the runs above average (RAA) and the runs above the replacement (RAR) which were mentioned earlier in equation 1. RAA is a value combined of batting runs, baserunning runs, fielding runs, grounded into double plays runs and positional adjustment runs.

$$\text{RAA}_{\text{Pos. Player}} = R_{\text{Bat}} + R_{\text{Baser}} + R_{\text{Dp}} + R_{\text{Field}} + R_{\text{Pos}} \qquad (13)$$

RAR is a value that adds RAA to the calculated replacement level runs.

$$RAR_{Pos.\ Player} = RAA_{Pos,\ Player} + R_{Rep} \tag{14}$$

h. Converting Runs to Wins & WAR

The players' runs above average (RAA) is converted to wins above average (WAA) with a PythagenPat win-loss estimator. It is a floating exponent subject to total runs per game (RPG) in the league and the relevant player's RAA and RAR per game.

$$waaWL\% = \frac{RPG_{League}^{x}}{RPG_{League}^{x} + \left(RPG_{League} - \left(\frac{RAA}{G}\right)_{Player}\right)^{x}} \tag{15}$$

$$x = \left(2 \cdot RPG_{League} - \left(\frac{RAA}{G}\right)_{Player}\right)^{0.285} \tag{16}$$

After calculating the waaWL% then it is finally possible to calculate the wins above average with equation 17.

$$WAA = (waaWL\% - 0.5) \cdot G \tag{17}$$

Then a similar calculations are made for the replacement player's runs to get the difference between the average player and the replacement level. Summing the outcome with WAA gives the WAR for the position player [30].

$$WAR_{Position\ Player} = WAA + WAR_{Replacement} \tag{18}$$

3.2. WAR for Pitchers

Calculations for pitcher's WAR depends on its simplest form of runs allowed and innings pitched. Pitcher's performance is best measured with those two factors but other performances also need reasonable involvement in the total calculations. Complex park factors and a

leverage multiplier index are new variables that are not used in the WAR calculations for the position player and need a considerable time and effort to calculate.

a. Runs Allowed (RA9)

This is similar to ERA but this factor includes the unearned runs as well, i.e. takes into account all runs.

$$\mathrm{RA9} = \frac{9 \cdot \mathrm{Runs}}{\mathrm{IP}} \tag{19}$$

By multiplying with nine, this factor is changed to runs allowed per nine innings pitched [31].

b. Level of Opposition and Handling Interleague ($RA9_{Opp}$)

It is vital to consider the different strength of each team since one team can be more difficult to play against than others. By looking at the opponent's scoring and put that number into a neutralized context with park factors makes it possible to resolve the expected runs against that team. Since it is only allowed to use designated hitters in the AL and at the home of the AL team when playing interleague (see subsections 2.2 and 2.3) then it is important to examine the impact of those scenarios. This type of interleague games is excluded from the team's average but when those games are played the team(s) with a DH get additional 0.2 runs per nine innings added to their pre calculated runs scored per nine innings. This added number is expected to be the number of allowed runs by a league average pitcher. Information from the last 365 days are used to calculate the team's average [32].

c. Adjusting for Team Defense ($RA9_{Def}$)

It can be difficult to determine the runs allowed by the pitcher because it is often not the pitcher‘s fault. One way is to measure the performance of the pitcher and account for the defense differently. The calculations for the pitcher‘s defense is based on how well his team‘s defense is and the proportion between the balls in play allowed (BIP) by the pitcher and team is intended to pinpoint the pitcher‘s performance.

$$\mathrm{RA9_{Def}} = \frac{\mathrm{BIP_{Pitcher}}}{\mathrm{BIP_{Team}}} \cdot (\mathrm{DRS_{Team}} \text{ or } \mathrm{TZR_{Team}}) \tag{20}$$

The defensive part of the equation depends on what year is being calculated for. As mentioned earlier, in subsection 3.2.e, the DRS has only been used since 2003 [32].

d. Adjusting Averages for Starters and Relievers ($RA9_{Role}$)

This segment calculates adjustments on whether a player played an inning as a starter or reliever. Starters have much higher earned run average (ERA) than relievers, which is linked to the fact that relievers get on field after the game begins and play for an inning or less and then leave. The same adjustment factor for the difference between starters and relievers ERAs has been used since 1974, 0.1125 $^{\text{runs}}/_{\text{game}}$, but in the years 1969 – 1973 it was significantly lower, 0.0583 $^{\text{runs}}/_{\text{game}}$ [32]. No further information on how these numbers are computed were found.

e. Custom Pitching Park Factors (PPFp)

There are multiple different park factors that need to be considered when estimating the customized park factor. The information used in calculating these factors span three years and are weighted by batters faced in each park. The calculations for the pitching park factor (PPF) are somewhat complicated.

The first step is to find the correlation between the differences of runs scored and allowed at home and away.

$$\text{InitialFactor} = \frac{(\text{Runs}_{\text{Scored}} - \text{Runs}_{\text{Allowed}})_{\text{Home}}}{\text{Games}_{\text{Home}}} \cdot \frac{\text{Games}_{\text{Away}}}{(\text{Runs}_{\text{Scored}} - \text{Runs}_{\text{Allowed}})_{\text{Away}}} \quad (21)$$

Next is necessary to make two corrections regarding the innings and parks. First is to consider the difference in playing a pitch at home or away. It's often considered to be more difficult to play on the road. This correction is called Innings Pitched Corrector (IPC) and is described in equation 22. This corrector is used later to calculate the Others Park Corrector (OPC). Please notice that if IPC is greater than one, it is due to the fact the innings pitched away are higher since the other team is batting more often in the ninth inning. The constant 18.5 is the average of half-innings per game if the home team always bats in the ninth inning. Second is to consider other road parks (equation 23). The total difference from the league

average is balanced with the park rating of the rated club. Here represents NT_{League} the number of teams faced in the same league.

$$IPC = \frac{18.5 - \left(\frac{Wins}{Games}\right)_{Home}}{18.5 - \left(\frac{Losses}{Games}\right)_{Away}} \tag{22}$$

$$OPC = \frac{NT_{League}}{\frac{InitialFactor}{IPC} + NT_{League} - 1} = \frac{NT_{League}}{TeamRating + NT_{League} - 1} \tag{23}$$

The next step is to determine the scoring factors. The OPC is not necessary to find the scoring factors but can in some situations simplify the calculations. The scoring factors are two, one is the park factor (SF) and the other deals with the fact a player does not play his own team (SF1).

$$SF = TeamRating \cdot OPC = \frac{InitialFactor \cdot NT_{League}}{InitialFactor + IPC \cdot \left(NT_{League} - 1\right)} \tag{24}$$

$$SF1 = 1 - \left|\frac{SF - 1}{NT_{League} - 1}\right| \tag{25}$$

Runs scored and allowed per home or away game have also to be taken into account to get to the final solution. Runs scored per home game (RHT) and runs allowed per home game (OHT) have parallel equations for away fixtures (RAT and OAT).

$$RHT = \left(\frac{Runs_{Scored}}{Games}\right)_{Home} \tag{26}$$

$$OHT = \left(\frac{Runs_{Allowed}}{Games}\right)_{Home} \tag{27}$$

$$RAL = \frac{Runs_{Total}}{NT_{League} \cdot \frac{Games_{Total}}{2}} = \frac{2 \cdot Runs_{Total}}{NT_{League} \cdot Games_{Total}} \tag{28}$$

RAL stands for total runs (in the league) per game for half of the games in the league.

Equations 24 – 28 are calculated variables which are used to compute the Team Pitching Rating (TPR) and the Team Batting Rating (TBR). In next two equations things get a little complicated. Equations 29 and 30 are part of an iterative process that is repeated three times (n = 1, 2, 3). The initial TPR (TPR_0) is set as 1 when calculating the TBR_1 for the first time. Then the result from TBR_1 is used in TPR_1 which is then used in TBR_2, and so on.

$$TBR_n = \frac{\left|\frac{RAT}{SF1} + \frac{RHT}{SF}\right| \left|1 + \frac{TPR_{n-1} - 1}{NT_{League} - 1}\right|}{RAL} \tag{29}$$

$$TPR_n = \frac{\left|\frac{OAT}{SF1} + \frac{OHT}{SF}\right| \left|1 + \frac{TBR_n - 1}{NT_{League} - 1}\right|}{RAL} \tag{30}$$

After the iteration process the Pitcher's Park Factor may be calculated [33]. PPF differs from PPFp in that manner PPFp is 100 times PPF.

$$PPFp = 100 \cdot \frac{SF + SF1}{\left|2 \cdot \left|1 + \frac{TBR - 1}{NT - 1}\right|\right|} = 100 \cdot PPF \tag{31}$$

A one game PPFp can be calculated with the following formula:

$$PPFp_{One\ game} = 2 \cdot PPFp_{Team} - 1 \tag{32}$$

f. League Average Pitcher Performance ($RA9_{Avg}$)

When all adjustments and factors are known then it is possible to calculate the actual runs per nine innings for the average pitcher [31].

$$RA9_{Avg} = \frac{PPFp}{100} \cdot \left(RA9_{Opp} - RA9_{Def} + RA9_{Role}\right) \tag{33}$$

g. Runs above Average (RAA)

This segment provides the number of runs a player is better than the average player. The results from subsections a. and g. are used to calculate it and because the results have been calculated for per nine innings there is a need to convert that to get the actual runs.

$$\mathrm{RAA} = \frac{\mathrm{IP}}{9} \cdot \left(\mathrm{RA9}_{\mathrm{Avg}} - \mathrm{RA9}\right) \tag{34}$$

The outcome is then centered to keep the total league average as zero [31].

h. Replacement Level Runs (R_{Rep})

The replacement level runs for pitchers is calculated with the following equation.

$$\mathrm{R_{Rep}} = \left(\frac{\mathrm{R_{Allowed}}}{\mathrm{Out}}\right)_{\mathrm{League}} \cdot \left(\frac{20.5 - 1.8}{100}\right) \cdot \mathrm{Out_{Player}} = 0.187 \cdot \left(\frac{\mathrm{R_{Allowed}}}{\mathrm{IP}}\right)_{\mathrm{League}} \cdot \mathrm{IP_{Player}} \tag{35}$$

The value 20.5 is the replacement level multiplier and 1.8 is an empirical factor which aligns the overall number of replacement runs to a desired league total [24]. Further information regarding the replacement level runs can be found in subsection 3.1.f.

i. Wins above Average Adjustment (WAA_{Adj})

There is one problem with all the calculations prior to the WAA. They make many of the relievers and average starters more valuable than they really are. This is due to the fact that closers and relievers tend to play in more essential situations which can determine the winner or loser of the game. To level the discrepancy a leverage multiplier, called game entering leverage index (gmLI), is used. The index focuses on the pressure which a pitcher faces in particular games when performing relief appearances. Index of 1.0 is thought as average, lower or higher than average indicate less or more pressure.

$$\mathrm{WAA_{Adj}} = \mathrm{WAA} \cdot \frac{1 + \mathrm{gmLI}}{2} \tag{36}$$

The outcome is then recentered, like in the RAA calculations. The leverage index is very complex and will not be analyzed further but for those who seek out further information regarding it can find it here [34].

j. Converting Runs to Wins & WAR

Converting pitcher's runs above average and replacement level runs uses same methods and equations as the position players, which can be found in subsection 3.1.h. The equation for the pitcher's WAR is slightly different than the position player's equation. The pitcher's equation has an additional variable, wins above average adjustment (see subsection 3.2.i.), added to it [32].

$$WAR_{Pitcher} = WAA + WAR_{Rep} + WAA_{Adj} \quad (37)$$

4. Literature Review

4.1. The History of Baseball Statistics

The birth of baseball statistics for analyzing was in the 1860s when Henry Chadwick established the box score. Since then multiple aspects of the game have been analyzed with various effects. In the early twentieth century the batting average was considered meaningless and a run value to hits was first calculated. This calculations influenced Bill James many decades later when creating the metric of Runs Created and the concept of linear weights. Around the end of the World War II the on-base percentage (OBP) was used to evaluate a player's talent and it was used in the mid-1950s to calculate the Offensive Performance Statistics (OPS = OBP + SLG). Few years later the run expectancy matrix was developed and led to different results for run values to hits. In 1964, the book Percentage Baseball was published. It included the first developed Scoring Index which pointed out the importance of OBP and slugging. Eight years later a program, which evaluated how player's hits influenced their teams' probabilities of winning the game, was evolved. However, it was never tried in the MLB.

In 1979 the first sabermetrician[4] was hired by the Houston Astros. The San Francisco Giants hired a statistical analyst a year later. Then the Texas Rangers and several other teams in the coming years and decades [36].

4.2. Optimization Literature for Team Selection

Aqil Burney et al. [37] used a generic approach to find the optimal solution for team selection by using a genetic algorithm. Their focus was on cricket but their method was converted into a generic model for other types of team sports. They approached with some new extensions such as attributes regarding personal and team performances and team's combination of players. The algorithm processed supports team selection with previous outcomes, i.e. win or lose. The ideology behind genetic algorithm is that you start with a set of solutions called population and these solutions are used to create a new population. The new population is supposed to be better than the old one. The solutions selected to create the new population are

[4] Sabermetrician is a sabermetric researcher and sabermetrics are *"the search for objective knowledge about baseball"* [35].

selected according to their fitness, i.e. the fitter they are, the more increased chance to reproduce [38]. In this scenario "to reproduce" means "to win".

Britz and von Maltiz [39] aimed to find an optimal solution to a team selection via the Hungarian algorithm. Their focus was not on selecting the best player in each position to create the strongest team but to optimize the best team for a specific task, i.e. attributes needed will vary while facing different teams throughout a season. The project's goal was to create a useful solution to select a team without any match data. It focused on the authors own testing where each player position had different emphasis.

Keener [40] created a model that predicted the number of wins that each MLB team will have in an oncoming season. The model used various baseball statistics at the beginning of a season, e.g. total salary and age of pitchers. The data used in the model was from only one season and even though the results were promising, there was a lack of data to fully determine the model reliability.

Kleinbard [41] examined the relationship between payrolls and wins to see whether the wins can be predicted by the payrolls. Many baseball fans and professionals have argued for years that payroll disparity creates a competitive imbalance in professional baseball. He came to the conclusion that variation in wins explained by salary have been decreasing in later years even though the salary disparity has been steadily increasing. This was reasoned with the following argument:

> *"As the pro-ready age for young athletes continues to fall and leagues continue to enact stricter regulations against substances that had historically disproportionately helped older players, policies that ensure cheap young labor—the league's most valuable currency—will do more to reduce the effects of payroll on winning than restrictions on overall team salary."*

Perhaps the most famous analytics in baseball took place in the Oakland's Athletics camp in the 2002 season. Their performance was so phenomenal that a book and a movie were published to tell the story surrounding their performance. With a low budget to sign players the Athletics were forced to search for players with a different method to be able to make a reasonable effort to reach their goal - to qualify for the playoffs. They identified that players with a good ability to get on base frequently had been overlooked when evaluated for contracts. In same manner they discovered that players with higher batting averages and slugging percentages were overvalued. By using this assertions the team won more games than was predicted by its payroll [4].

4.3. Statistical Analysis of Baseball Performance

Most statistical analysis which are related to performance in baseball try to demonstrate and show connections between money and performances, e.g. which players are overpaid and which players are underpaid by taking various factors into account. Big part of this "money vs. performance" analysis has been worked on free agents. As an example of an analysis is a study done by Barnes and Bjarnadóttir [42] where two methods were used to estimate players' salaries based on their performances. First, Barnes and Bjarnadóttir predicted a player salary using data collected about his current and career performances, his current salary, and other key features. The second method estimated WAR for each player to convert into a salary based on performance. By using WAR this approach relied on more number of statistics for each player than other older studies.

Baumer et al. [43] proposed a new aggregate measure called openWAR. It was meant to be competitive to the former WARs and have some different methodology. The goal was to use public data and develop it with more accuracy and clarify the standard for a replacement player which they thought was nebulous. The four main differences between openWAR and the former WAR were as follows:

1. It is not a measure of player's ability to use in forecasting.
2. Uses plate appearance level to control for defensive position in both batting and fielding estimates. That offers more precise comparisons of players to their peer group.
3. Share credit and blame for hits on balls in play between the fielders and pitcher. It uses the location of the batted ball to decide what player(s) should share the credit.
4. It proposes a new definition of replacement level based on distribution of performance beyond all the active major league players that play regularly.

Their research included some limitations that need further research. The MLBAM data used tended to have misprinted statistics and many accounting of baserunner movement for non-batting events, e.g. stolen bases and errors, need further work. Its defensive models may also be improved with more solid data.

Baumer and Zimbalist [36] made an attempt to model the effectiveness of sabermetric statistics. They modeled team performance as a function of payroll and created metrics which purpose was to indicate the influence of sabermetrics on team's composition and performance. They came to the following conclusion.

"…there is no reason to believe, independent of sabermetrics, that a team with high sabermetric index would be more successful than their payroll would indicate. Yet the regression model above shows that our sabermetric index explains nearly 37 percent of the variation in team winning percentage that is not explained by payroll."

Studnitzer [44] researched how wins above replacement and more traditional statistics effect on salaries in both factors. His results indicated that WAR had a large effect on average annual value but his exact words were:

"The only significant variable with a large effect on average annual value was WAR in both arbitration and free agency, confirming that front offices are effectively evaluating players and devoting resources to player who have shown the capability to provide more production in the past. Further, there is a large positive interaction term between free agents and WAR, indicating that the presence of an arbitrator may be hindering the market tendency to provide even more emphasis on such comprehensive evaluations."

5. Materials and Methods

5.1. Introduction

The data used in this project is collected from www.Baseball-Reference.com and holds statistical information regarding each player from the 2014 season and going back to the 1969 season. It consists of two different datasets, one for batting players and one for pitchers. The data contains statistical information regarding thousands of players who have all played at least one game in the Major Baseball League. Some players can be found in both datasets since pitchers do occasionally bat, and batters pitch. When collecting the relevant data there are different statistical information to focus on for batters and pitchers. This leads to different methods in calculating WAR, the value this project focuses on. There are multiple procedures to calculate and determine WAR but there are three implementations that are most often referenced. They are FanGraphs's WAR, Baseball-Reference's WAR, and Baseball Prospectus's WARP (Wins above Replacement Player). This project will use Baseball-Reference's WAR in relation to the data used.

The computer software used in all the data processing was Microsoft SQL Server (MSSQL), Microsoft Excel and R.

5.2. The Strategy

The main object is to see whether it is possible to use WAR as a playoffs estimator when putting together a team before each season. One way of doing that is using the players' last registered data and connect it to the oncoming season in a new dataset. It is not enough to look at data from last season since there are always players who miss at least one season due to injuries or not being offered a contract from any team. This new dataset is then used to create two different datasets, a baseline dataset and the project's dataset. The latter dataset will be processed with various tools, e.g. logistic regression and prediction trees, in order to create a model which beats the baseline.

5.3. Assumptions

5.3.1. Connecting the Players to Teams

There were few assumptions necessary in order to put together a dataset which connected the last registered player data with the correct team for the oncoming seasons. First of all no trades, which took place after the season started, were taken into account when estimating the teams' WAR. This was due to the fact that general managers and coaches cannot predict before the season what trades will occur. When a trade is made, the original dataset registers the relevant players twice for that season, see figure 5 for an example. In the same manner a player is sometimes waived and picked up by another team which the datasets collect similarly. However, there is a slight difference in how the datasets collects the salary data for those players. This difference was used to assume in which team the player started the season with. If the salary was a number larger than zero then the player was registered to that team. That was done like that due to the fact that the datasets seem to register the salary as zero for the new teams. Several[5] players were handpicked and checked online to see if that assumption was reasonable. That checking proved the assumption to be correct.

	OriginalID	Player	Year	Team	Pos	Age	G	WAR	Salary	First Year
	127	Adam Dunn	2008	ARI	UT	28	44	-0,4	0	2001
	128	Adam Dunn	2008	CIN	LF	28	114	1,1	13000000	2001
NEW	128	Adam Dunn	2008	CIN	LF	28	158	0,7	13000000	2001

Figure 5: An example of a collected data for a player who played for more than one team in a season.

Occasionally players who had played for more than one team were found in the datasets but had none the less zero salary registered for all the relevant teams. In those cases the amount of games played (G) for each club was looked at. The team the players played most games for that season was the team they were assumed to have started the season with. The reason being that most trades and waives take place shortly before trade deadline [45]. In both cases mentioned before, the WAR was calculated in same way. The players' WAR for all played teams, in that season, was summed since there was no weighting necessary. Sometimes players play different positions for each team but there was no need to take that into consideration since WAR strives to be uniform. Figure 5 shows an example of how the data was combined into one ID to use in the new dataset (the yellow row).

[5] It would have been too time consuming to check each double registered player since the only way was to check by hand.

Regarding rookies and players who had played more than one season and tried a new position, e.g. a player that had only batted but tried to pitch for the first time and vice versa, a different approach was needed to use for their first season in the MLB or in the new position. A total mean WAR for each position was calculated for all the years and then assigned to each player, i.e. all players playing the same position got the same WAR in each year.

However, a big error was found in the data regarding the pitchers positions. Multiple pitchers had been recorded with wrong positions, e.g. numerous starting pitchers appeared to be relief pitchers in the data. The data showed that the teams had on average three starting pitchers while in the reality they tend to be around four to six. This error in the data affected how to approach the data selection when creating the project's dataset. How it affected will be described in next subsection.

5.3.2. The Baseline Dataset and the Project's Dataset

The main idea was to create datasets which used as little data as possible from the oncoming seasons, and at the same time focus on using the last recorded data of each player. The only data used from oncoming seasons are the previous mentioned rookies' data and the players' positions. It can be reasoned for with the fact that general managers and coaches usually buy or get new players for certain positions they are looking to improve or fill.

Players for the baseline was selected based on their last season's number of games played regardless of their position. The players were split into two groups; positional players and pitchers. Each group were ranked in descending order based on their amount of games played and the top 12 players from each group were selected. If selected players had the same number of games played, a random selection decided who should be included and who should not. Despite that selection method there were 23 teams which had the overall number of pitchers under 12. Those teams had a combination of 22-23 players and the total average number of players included in the teams in the baseline dataset was 23.98 player.

When selecting players for the project's dataset a different selection method was used. The number of players to be selected was the same as in the baseline approach but using the player's WAR as the selection variable. The positional players had two rounds of selection. First were the players who played one of the eight positions mentioned in figure 2 (dismiss the pitcher) split by their position and ranked by their WAR and the players with the highest WAR from previous season were selected. If a team didn't have a player in one of those

positions then players who played similar position were ranked and then selected, e.g. if a team was missing a CF after the first round the OFs it was ranked and the best selected (see table 4). Considering the fact that teams in the American League almost always have DH, were the DH selected as a one of the group for those teams. The AL teams prior to 1973 did not include a DH since they were first allowed in 1973. To fill in the last available spots, all remaining positional players were ranked as a single group and the top players selected. When more multiple players in the last spot had the same WAR a random selection decided whom to pick.

Table 4: Backup selections for missing defensive positions.

Position	**1st Backup selection**	**2nd Backup selection**	**3rd Backup selection**
1B, 3B	CI	IF	UT
2B, SS	MI	IF	UT
CF, LF, RF	OF	UT	

As mentioned in last subsection, an error regarding the pitchers was found in the original data. This error caused the selection method for the pitchers to be more imprecise than the positional players' method. The pitchers were ranked as one group and the top 12 players selected. Like in the baseline were 23 teams that had fewer than 12 players. The project's dataset were split randomly into two new datasets, training and testing. The training dataset included 30 seasons and was used to create the models which were then applied on the testing data for validation.

5.3.3. Playoffs Spots

Seeing as the division winners always get a spot in the playoffs, the teams' estimated WAR must be considered for each division before each league. As mentioned in section 2.4, for the first time in 1994, non-champion teams were allowed to qualify in addition to the divisions champions when the number of playoffs spots was increased. This led to the assumption to pick a non-champion team with the best estimated WAR in each league, in this scenario the divisions had no significance.

The number of playoffs spots, the number of teams and the leagues setups have changed several times in the years the data covers (see appendix A3). A simple way to approach these changes is to use the modern day playoffs setup and apply it backwards to

each year. This gives ten playoffs spots in each of the 45 seasons or a total of 450 playoffs spots available to the 1234 teams which participated in the major league from 1969 – 2014. The seasons with strikes are included.

5.4. Tools

5.4.1. Classification Matrix

To compare actual data to predicted data is ideal to use a classification matrix. It contains the information done by the classification system and the matrix is often used to evaluate the performance of the system. A classification matrix is of size NxN, where N is the number of different label values [46]. The following matrix shown in table 5 will be used in this project to evaluate the playoffs qualification estimations.

Table 5: 2x2 Confusion Matrix. Adapted from [46].

		Predicted		
		No	Yes	
Actual	No	a	b	a + b
	Yes	c	d	c + d
		a + c	b + d	Total

The a (TN) and d (TP) entries indicate numbers when the predictions are correct but b (FP) and c (FN) when they are incorrect [46]. Several standard terms will be used to evaluate data in the matrix:

1. True positive rate (TPR), or sensitivity, is the proportion of positive cases, calculated with the equation [47]:

$$TPR = \frac{d}{c + d} \tag{38}$$

2. Accuracy (AC) is the proportion of the total number of positive cases, calculated with the equation [47]:

$$AC = \frac{a + d}{a + b + c + d} \tag{39}$$

3. Precision (P) is the proportion of the positive predicted cases that were correct or in other words it measures the accuracy. It is calculated with the equation [47]:

$$P = \frac{d}{b + d} \tag{40}$$

However, in the project's cases P is equal to TPR since there are just either no's or yes's if a team reaches playoffs, i.e. each FP leads to FN and causes the incorrect predictions to be equal.

5.4.2. Prediction Trees

A classification tree and a regression tree are nonlinear predictive models which fall under the category of prediction trees. Shalizi [48], a former assistant professor at Carnegie Mellon University, explained very well, and in a simple way, how prediction trees work:

> *"We want to predict a response or class Y from inputs X_1, X_2, . . . X_p. We do this by growing a binary tree. At each internal 1 node in the tree, we apply a test to one of the inputs, say X_i . Depending on the outcome of the test, we go to either the left or the right sub-branch of the tree. Eventually we come to a leaf node, where we make a prediction. This prediction aggregates or averages all the training data points which reach that leaf."*

Classification trees attempt to predict the class instead of numerical values, i.e. the endnotes indicate either a yes or no, while regression trees focus on the numerical values. Both these trees give a rather simple graphical solutions regardless of the number and complexity of the variables and data.

5.4.3. Logistic Regression

Logistic regression is somewhat similar to a linear regression but it uses a non-continuous dependent binary variable (1 is true and 0 is false) to estimate the probabilities of an event occurring. It allows a use of multiple independent variables and the error differences tend to follow a logistic distribution. With a logistic regression, an attempt is made to find the best fitting model in which describes the relationship between the dependent and independent variables. The model strives to forecast the likelihood (odds) of something happening, which is making the playoffs in this project.

$$\text{logit}(p) = b_0 + \sum_{i=1}^{n} b_i \cdot X_i \tag{41}$$

Here are b_i the coefficients and X_i the variables. Where the dependent variables have only a "yes" or a "no", there are only two possible outcomes with different odds of occurring – probabilities of event A taking place and the probabilities of event B taking place. This relationship is best described with the following equation.

$$\text{odds} = \frac{\text{Prob. of event A}}{\text{Prob. of event B}} = \frac{p}{1-p} \tag{42}$$

Taking the natural logarithm of the equation gives a logit transformation of the odds which relate to logit(p) in equation 41 [49], [50]. This can be presented as

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \tag{43}$$

Simple derivation gives the desired equation to calculate the probabilities (p).

$$p = \frac{e^{\text{logit}(p)}}{1 + e^{\text{logit}(p)}} = \frac{1}{1 + e^{-\text{logit}(p)}} \tag{44}$$

Logistic regression will be used in order to create a simple model which is able to beat the baseline. In R there are functions who work well with logistic regression and offer methods to determine the logistic method with automatic variable selection. The step function provides three different stepwise approaches to find the best fitted model; backward, forward or both. With backward direction the independent variables are removed one by one, determined by their AIC[6], until the best combination is left. The forward direction starts only with a constant and adds one variable in each step until the best fitted model is achieved. Both direction use previously mentioned methods to get the ideal model.

[6] The Akaike's information criterion measures the quality of a statistical model for the related data.

6. Results

6.1. Classification Matrix and Logistic Regression

There is one simple way of checking the difference between the project's model and the baseline model. The procedure is to calculate each teams' total average for the estimated WAR (EstWAR) and rank the teams in each season in descending order. The available playoffs spots in each season can then be "handed out" in accordance with the playoffs rules.

Table 6: Classification matrices for the project's and the baseline model.

Project's model - Total			**Baseline model - Total**		
	No	*Yes*		*No*	*Yes*
No	598	186	*No*	595	189
Yes	186	264	*Yes*	189	261
	TPR	58.7%		TPR	58.0%
	ACR	69.9%		ACR	69.4%

A comparison of the classification matrices in table 6 gives almost identical results. The project's model has an insignificantly higher true positive rate and accuracy than the baseline. Strictly speaking the model has beaten the baseline but the difference is too little to reach the conclusion that this model is in fact better. Figure 6 provides a simple way to predict the probabilities a team, with certain estimated WAR, has of reaching the playoffs. The figure is based on the project's model in table 6.

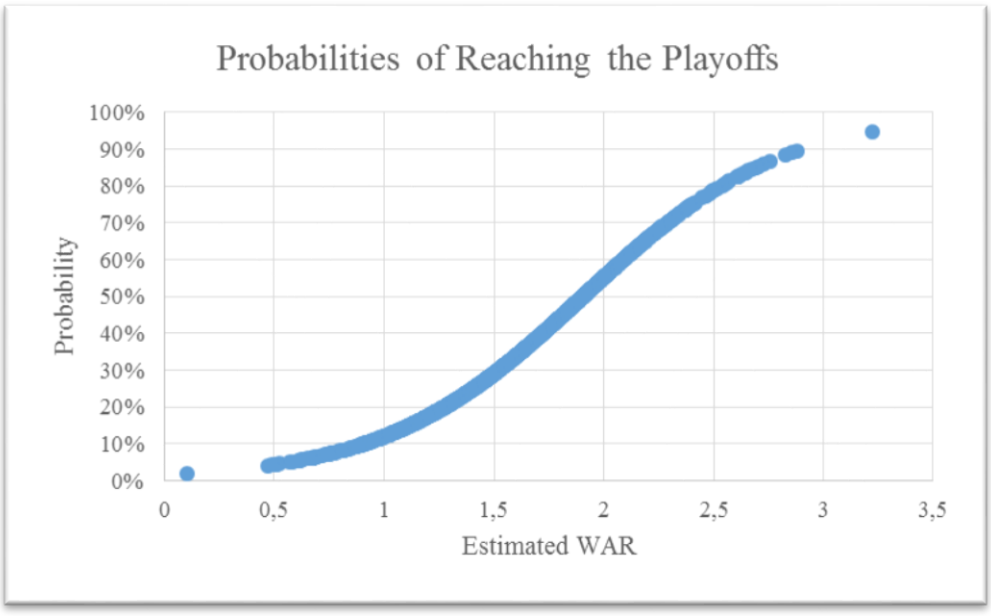

Figure 6: The probabilities of reaching the playoffs, based on the project's model in table 5.

A logistic regression applied on the training data provided an equation.

$$\begin{aligned}\text{logit}(p) = & -3.819 - 0.277 \cdot \text{MinWAR} + 0.795 \cdot \text{Pitchers} + 0.109 \cdot 1\text{B} + 0.086 \cdot 2\text{B} \\ & + 0.091 \cdot 3\text{B} + 0.197 \cdot \text{C} + 0.098 \cdot \text{CF} + 0.126 \cdot \text{LF} + 0.180 \cdot \text{RF} + 0.117 \cdot \text{SS}\end{aligned} \quad (45)$$

All the variables, except MinWAR, in the equation above demonstrate the team's overall average WAR for the appropriate position. The overall average for batters, EstWAR, RP, CL and SP were not included in the regression process. This is because all the batters positions should give more precise information since the overall average for batters is just one number combined from all the positions. This idea was not possible for the pitchers because of the error in the data. The same reasoning was with EstWAR because the variable includes all the positions. Summarized information for the equation is shown in table 7.

Table 7: Summary of the logistic regression.

	Coeff. b	**Std. Error**	**e^b**	**z-value**	**p-value**	
Intercept	-3.819	0.366	0.022	-10.431	< 2e-16	***
MinWAR	-0.277	0.160	0.758	1.726	0.084	.
AvgWARPitchers	0.795	0.174	2.215	4.558	5.2e-06	***
AvgWAR1B	0.109	0.042	1.115	2.597	0.009	**
AvgWAR2B	0.086	0.044	1.090	1.982	0.047	*
AvgWAR3B	0.091	0.038	1.095	2.365	0.018	*
AvgWARC	0.197	0.057	1.217	3.445	0.001	***
AvgWARCF	0.098	0.039	1.103	2.498	0.012	*
AvgWARLF	0.126	0.041	1.134	3.083	0.002	**
AvgWARRF	0.180	0.041	1.197	4.344	1.4e-05	***
AvgWARSS	0.117	0.403	1.124	2.893	0.004	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 1040.68 on 790 degrees of freedom.

Residual deviance: 915.88 on 780 degrees of freedom.

AIC = 937.88

The p value associated with the null deviance model is 4.6e-9, which is way below 5%, indicating that the data does not plausibly emanate from a logistic regression model with a constant term only. The same can be said about the residual deviance model which has a p value of 5.2e-4. However, the difference in the chi-square statistics for the null model and the residual model is zero. This means that it is possible to reject the null hypothesis that the

deviance of the model with only the constant term and the model with the independent variables added to it are the same. The deviances are not the same and in fact is the deviance of the model with the independent variables is statistically significantly lower.

Assigning the equation to the training data, testing data and the overall data gives the following classification matrices. Same ranking method as in table 6 was used, except the teams were ranked by the probabilities provided by the equation.

Table 8: Classification matrices based on the equation from the logistic regression.

Training			**Testing**			**Overall**		
	No	*Yes*		*No*	*Yes*		*No*	*Yes*
No	393	123	*No*	207	61	*No*	600	184
Yes	123	177	*Yes*	61	89	*Yes*	184	266
	TPR	59.0%		TPR	59.3%		TPR	59.1%
	ACR	69.9%		ACR	70.8%		ACR	70.2%

The equation seems to apply properly on the testing data since the difference is close to none between the ratios. The overall true positive rate and the accuracy point towards the assumption that the baseline can be beaten. Both the matrices in table 6 got beaten by the results from the equation. The difference is still not large but the baseline has approximately one percentage lower true positive rate and less than one percentage difference in accuracy.

Figure 7 shows how the correlation between the variables in the data is. There is very little correlation between the variables which can explain why the difference between the regression model and the model in table 6 is insignificant.

Figure 7: Correlation between all variables in the data available or used in the forecasting models.

6.2. Classification Tree

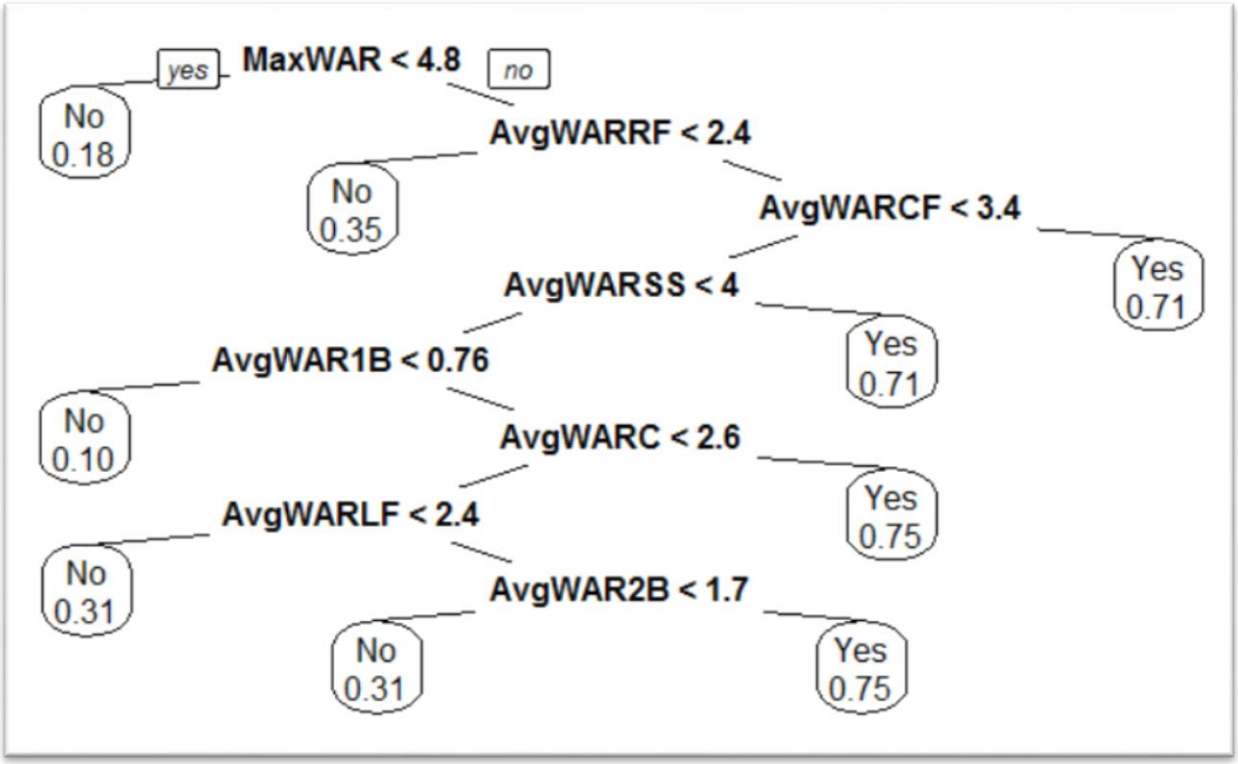

Figure 8: A classification tree of the model.

The figure above provides a visualized result of the model which is rather easy to understand. If a variable satisfies a condition then it continues to the left, otherwise to the right.

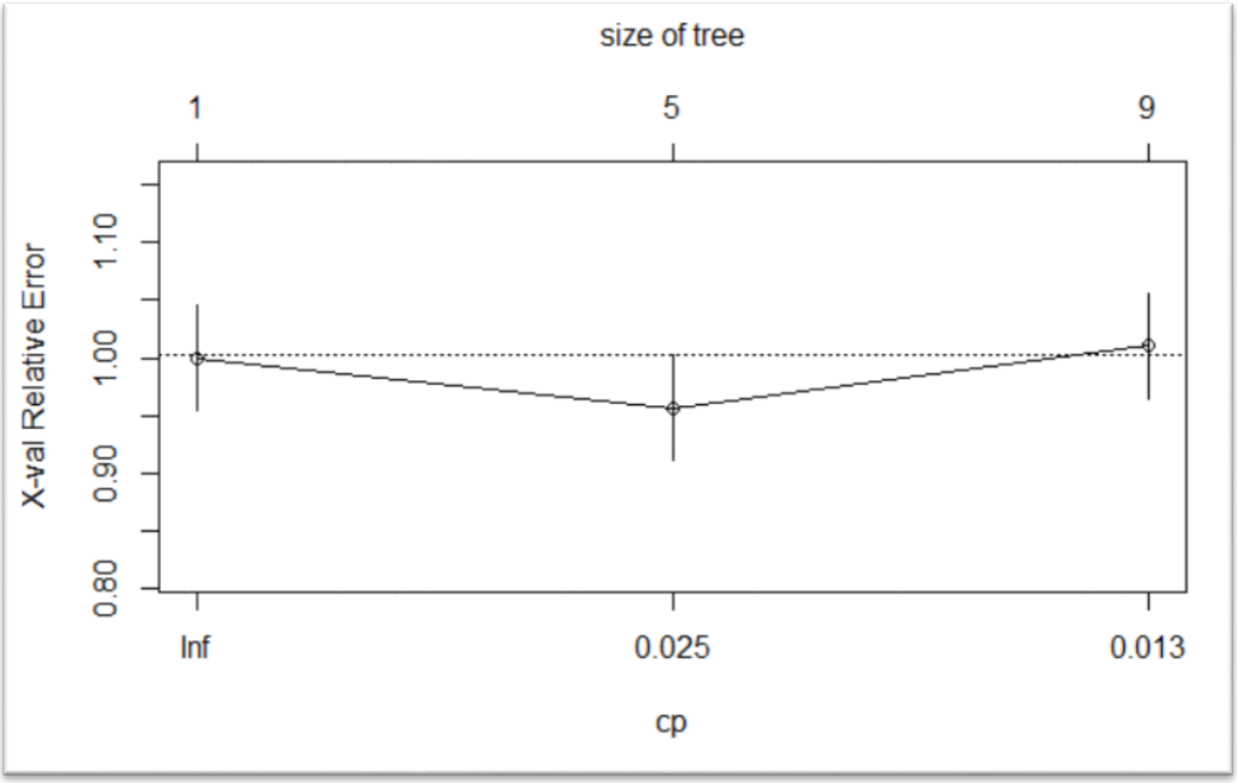

Figure 9: Cross-validation results of the classification tree.

The plot above provides a graphical representation to the cross validated error summary. The cp values are plotted against the geometric mean to depict the deviation until the minimum value is reached. The best choice of cp value for pruning is a value below the mean line (the dot line) and leftmost. In this case there is no pruning available since the leftmost cp value is insignificant.

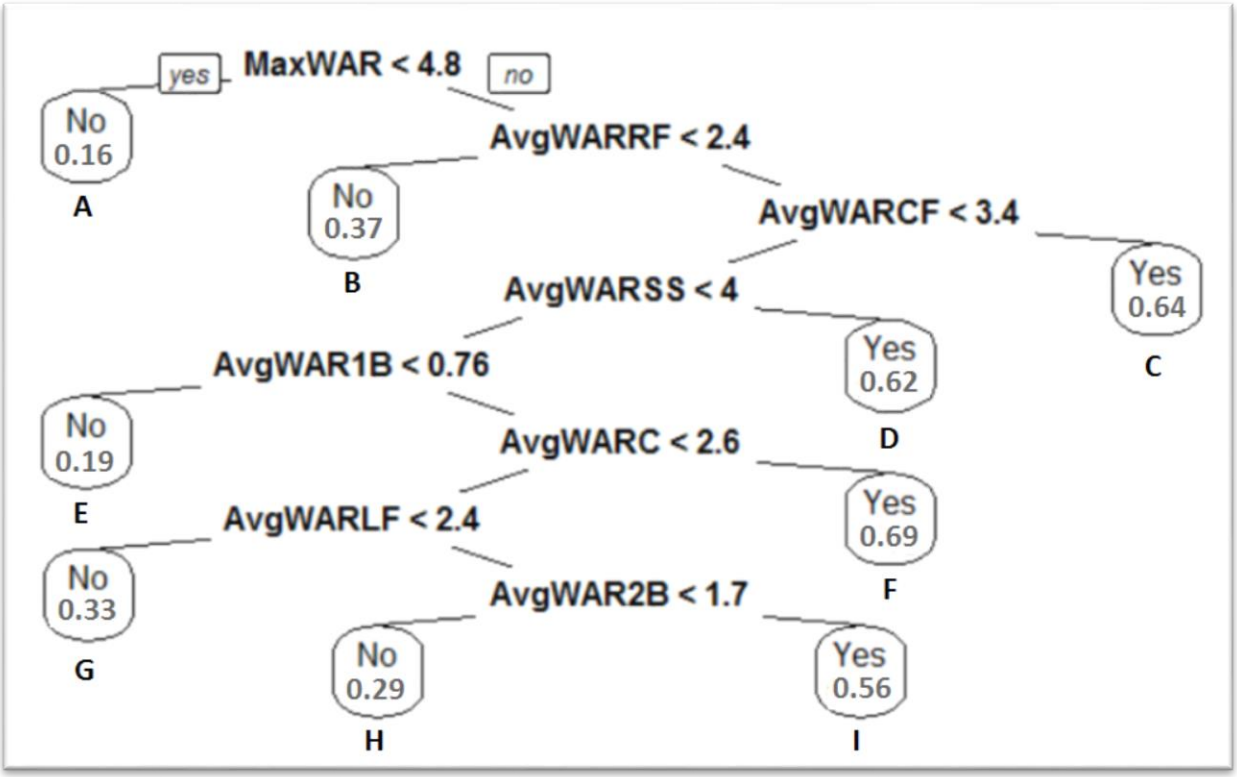

Figure 10: A classification tree of the model with ratios according to the dataset.

In figure 10, the classification tree's structure from figure 8 has been applied to the dataset and the probabilities in each endnote have changed. However, these number are more like ratios than probabilities. This tree's probabilities should be more reliable than in the original tree from the R program. That can be rationale with the fact, that the R program could not provide a tree with exactly 450 playoffs spots which is needed to get a more precise results. The distribution in the tree can be found in table 9.

Table 9: The distribution in each endnote from figure 10.

Endnote	Nr. of Teams	Nr. of Playoffs Teams	Ratio
A	263	43	16.3%
B	550	202	36.7%
C	130	83	63.8%
D	50	31	62.0%
E	52	10	19.2%
F	35	24	68.6%
G	89	29	32.6%
H	31	9	29.0%
I	34	19	56.0%

7. Discussions

Even the simplest model using only the average WAR of the top 24 players to predict whether a team qualifies for playoffs or not gives a rather good prediction – a TPR close to 60% and an accuracy of 70%. The other attempts made with the logistic regression and the classification tree had minor success. However, when the original tree structure was applied on the data a better or perhaps more useful tree was created. The endnotes' probabilities are more like ratios between how many teams actually qualified for the playoffs of all the number of teams that ended in the note. The tree is very simple to use and gives results quickly.

The lack of improvement can in some way be explained with the fact that the correlations between the variables are pretty low. Figure 7 shows that very well. This lack of correlation affects the models and leads to higher misclassification errors, which again decreases the reliability of the models.

It must not be forgotten to think of other unpredictable variables and circumstances which could not be included in the model but would most certainly affect the outcome of it. Some examples of those variables are injuries, trades and changes in players' day forms.

An attempt was made to apply a logistic regression on the AL to see the effects of the designated hitter. That did not give a good results and further checking was not made. The results can be found in appendix A6.

The project's methods which attempted to improve and beat the baseline had some limitations. Both the logistic regression and the classification trees can forecast datasets which have dependable binary variables. Those binary variables are usually not restricted of anything other than they can only take two values. The dependent variable in this project had multiple external rules which could not be included in the calculations and limited the accuracy of the models created. The number of playoffs spots is a constant number which means that in each year a specific amount of 1 (positive outcome) have to be included. In this project there were 450 playoffs spots available over 45 seasons, i.e. 10 spots each year. That number is then split between the leagues and again between the leagues' subdivision where the subdivision's winners advance to the playoffs even though there is a team with a better record in any of the other league's subdivisions. This could not be implemented in the creation of the classification tree or the logistic regression model, or at least the author did not have enough technical knowledge to do that. If this could be established in the model creation then the results would most likely improve. Therefore it would be a good foundation and reasoning for further researches in relation to the project's material.

7.1. Future Work

The first step is to come up with a correctly recorded data for the pitchers and use similar methods to see if or how the results differ from the results in this project. It is necessary to check what affects the pitcher's positions truly have on the outcome to predict a qualification to the playoffs. The pitcher position is considered to be one of the most important, or the most important, position in the game of baseball since every action made in the game can be linked, in some manner, back to how well the ball was thrown to the batter.

This project is a part of a bigger project which is being worked on and aims to create a data driven optimization model to support decisions which general managers face when hiring or recruiting individuals. This project was originally meant to implement a simple optimization model to identify what kind of players a team should have hired and compare the model with actual hiring decisions. Unfortunately there was not enough time to complete all the necessary steps towards the optimization model but the work towards it can continue.

8. Conclusion

WAR as a single forecasting variable has every reason to advance further in the future but it needs more researching. Despite the fact that the methods in this project attempted to beat the baseline without any great evidence of success, there are various things which indicate that it is possible, e.g. better data for pitchers would increase the likelihood of more precise and dependable results to base future works and researches on.

References

[1] “Hitting A Baseball – ‘The Hardest Thing To Do In Sports’ | Axon Sports.” [Online]. Available: http://www.axonpotential.com/hitting-a-baseball-the-hardest-thing-to-do-in-sports/. [Accessed: 23.05.2015].

[2] “How Baseball Works,” *HowStuffWorks*. [Online]. Available: http://entertainment.howstuffworks.com/baseball.htm. [Accessed: 26.04.2015].

[3] H. K. (Organization) and T. W. Hanlon, *Sports Rules Book-3rd Edition, The*. Human Kinetics, 2009.

[4] S. Barnes and M. Bjarnadottir, “Baseball - Discovering the Moneyball Effect.” .

[5] “Baseball Defense,” *HowStuffWorks*. [Online]. Available: http://entertainment.howstuffworks.com/baseball.htm. [Accessed: 26.04.2015].

[6] “Baseball Offense,” *HowStuffWorks*. [Online]. Available: http://entertainment.howstuffworks.com/baseball.htm. [Accessed: 26.04.2015].

[7] “Infield Positioning for Baseball Situations,” *Pro Baseball Insider*. [Online]. Available: http://probaseballinsider.com/baseball-instruction/infield-situational-positioning/. [Accessed: 26.04.2015].

[8] “Rules, Regulations and Statistics,” *Major League Baseball*. [Online]. Available: http://mlb.mlb.com/mlb/official_info/about_mlb/rules_regulations.jsp. [Accessed: 26.04.2015].

[9] “Designated hitter,” *baseball-reference.com*. [Online]. Available: http://www.baseball-reference.com/bullpen/Designated_hitter. [Accessed: 26.04.2015].

[10] “In Baseball, What Is a Middle Infielder?,” *wiseGEEK*. [Online]. Available: http://www.wisegeek.com/in-baseball-what-is-a-middle-infielder.htm. [Accessed: 26.04.2015].

[11] “Relief pitcher - BR Bullpen,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/bullpen/Relief_pitcher. [Accessed: 26.04.2015].

[12] “Utility Player Definition - Sporting Charts.” [Online]. Available: /dictionary/mlb/utility-player/. [Accessed: 26.04.2015].

[13] Major League Clubs and Major League Baseball Players Association, “Basic Agreement.” 12.12.2011.

[14] “Major League Baseball’s Season Is Not As Long As You May Think.” [Online]. Available: http://www.businessinsider.com/major-league-baseball-season-2014-9. [Accessed: 26.04.2015].

[15] “World Series Winners, Stats, and Results and Postseason Series,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/postseason/. [Accessed: 25.04.2015].

[16] “Wild Card Definition - Sporting Charts.” [Online]. Available: /dictionary/mlb/wild-card/. [Accessed: 25.04.2015].

[17] B. M. Bloom, “Addition of Wild Card berths finalized for 2012,” *Major League Baseball*. [Online]. Available: http://m.mlb.com/news/article/26927024/. [Accessed: 25.04.2015].

[18] “1981 MLB Season: A Strike-Split Year & The Comeback Dodgers,” *TheSportsNotebook*. [Online]. Available: http://www.thesportsnotebook.com/2014/05/1981-mlb-season-sports-history-articles/. [Accessed: 26.04.2015].

[19] “Longest strike in Major League Baseball history ends - Mar 31, 1995,” *HISTORY.com*. [Online]. Available: http://www.history.com/this-day-in-history/longest-strike-in-major-league-baseball-history-ends. [Accessed: 26.04.2015].
[20] “Baseball-Reference.com WAR Explained,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained.shtml. [Accessed: 17.04.2015].
[21] “Replacement Level | FanGraphs Sabermetrics Library.” [Online]. Available: http://www.fangraphs.com/library/misc/war/replacement-level/. [Accessed: 10.05.2015].
[22] “wRAA For Position Player WAR Explained,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained_wraa.shtml. [Accessed: 18.04.2015].
[23] “Custom wOBA and Linear Weights Through 2010: Baseball Databank Data Dump 2.1,” *Beyond the Box Score*. [Online]. Available: http://www.beyondtheboxscore.com/2011/1/4/1912914/custom-woba-and-linear-weights-through-2010-baseball-databank-data. [Accessed: 18.04.2015].
[24] “Position Player WAR Calculations and Details,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained_position.shtml. [Accessed: 20.04.2015].
[25] “Ground Into Double Play - GIDP Definition - Sporting Charts.” [Online]. Available: /dictionary/mlb/ground-into-double-play-gidp/. [Accessed: 03.05.2015].
[26] D. Basco and J. Zimmerman, “Measuring Defense: Entering the Zones of Fielding Statistics,” *http://sabr.org*. [Online]. Available: http://sabr.org/research/measuring-defense-entering-zones-fielding-statistics.
[27] “Total Zone Data,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/total_zone.shtml. [Accessed: 18.05.2015].
[28] “WAR Comparison Chart,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained_comparison.shtml. [Accessed: 19.04.2015].
[29] “Unifying Replacement Level | FanGraphs Baseball.” [Online]. Available: http://www.fangraphs.com/blogs/unifying-replacement-level/. [Accessed: 20.04.2015].
[30] “Baseball-Reference.com WAR Explained, Converting Runs to Wins,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained_runs_to_wins.shtml. [Accessed: 22.04.2015].
[31] “Baseball Reference,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/. [Accessed: 17.05.2015].
[32] “Pitcher WAR Calculations and Details,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/war_explained_pitch.shtml. [Accessed: 11.05.2015].
[33] “Park Adjustments,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/parkadjust.shtml. [Accessed: 10.05.2015].
[34] “Win Expectancy (WE) and Run Expectancy (RE) Stats,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/about/wpa.shtml. [Accessed: 11.05.2015].
[35] “A Guide to Sabermetric Research,” *sabr.org*. [Online]. Available: http://sabr.org/sabermetrics/single-page. [Accessed: 20.05.2015].
[36] B. Baumer and A. Zimbalist, *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*, 1st edition. Philadelphia, Pennsylvania 19104-4112: University of Pennsylvania Press, 2014.

[37] S. M. Aqil Burney, N. Mahmood, K. Rizwan, and A. Usman, “A Generic Approach for Team Selection in Multiplayer Games using Genetic Algorithm,” *Int. J. Comput. Appl.*, vol. 40, no. 17, pp. 11–17, Feb. 2012.
[38] M. Obitko, “Genetic Algorithm Description - Introduction to Genetic Algorithm,” *http://www.obitko.com*, 1998. [Online]. Available: http://www.obitko.com/tutorials/genetic-algorithms/ga-basic-description.php. [Accessed: 26.04.2015].
[39] S. S. Britz and M. J. von Maltiz, “Application of the Hungarian Algorithm in Baseball Team Selection and Assignment.” University of the Free State.
[40] M. H. Keener, “The econometrics of baseball: A statistical investigation,” The University of Tampa, Research.
[41] M. Kleinbard, “Can’t Buy Much Love: Why money is not baseball’s most valuable currency,” Columbia Business School, Research, Feb. 2014.
[42] S. Barnes and M. Bjarnadottir, “Great Expectations: An Analysis of Free Agent Performance,” presented at the MIT Sloan Sports Analytics Conference, Boston, 2014.
[43] B. S. Baumer, S. T. Jensen, and G. J. Matthews, “openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball,” Mar. 2015.
[44] J. Studnitzer, “Simplicity Versus WAR: Examining Salary Determinations in Major League Baseball’s Arbitration and Free Agent Markets.” Haverford College, 01.05.2014.
[45] S. Barnes, “Regarding the ‘new data,’” [Email]. 29.03.2015.
[46] R. Kohavi and F. Provost, “Glossary of Terms,” *robotics.stanford.edu*, 1998. [Online]. Available: http://robotics.stanford.edu/~ronnyk/glossary.html.
[47] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, Feb. 2011.
[48] C. Shalizi, “Classification and Regression Trees.” 11-Jun-2009.
[49] Author unknown, “What is Logistic Regression,” *strath.ac.uk*. [Online]. Available: https://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/whatislogisticregression/. [Accessed: 19.05.2015].
[50] Author unknown, “Logistic regression,” *medcalc.org*. [Online]. Available: https://www.medcalc.org/manual/logistic_regression.php. [Accessed: 19.05.2015].
[51] “Baseball statistics - BR Bullpen,” *Baseball-Reference.com*. [Online]. Available: http://www.baseball-reference.com/bullpen/Baseball_statistics. [Accessed: 28.04.2015].
[52] “RMSD | Root mean squared deviations | statistics of fit | reference manual,” *Spider Financial | spiderfinancial.com*. [Online]. Available: http://www.spiderfinancial.com/support/documentation/numxl/reference-manual/descriptive-stats/rmsd. [Accessed: 27.04.2015].

8. Appendices

A1 - Baseball Abbreviations

This appendix includes five categorized tables that each hold descriptions and the real name behind each abbreviation used in www.baseball-reference.com statistics. This is taken directly from their website but includes minor additions, e.g. some equations to give a deeper understanding [51].

Table 10: Baserunning statistics abbreviations. Adapted from [51].

CS	*Caught Stealing*	Times tagged out when attempting to steal.
R	*Runs Scored*	Times reached home base legally and safely.
SB	*Stolen Base*	Number of bases advanced other than on batted balls, walks, or hits by pitch.

Table 11: Batting statistics abbreviations. Adapted from [51].

1B	*Single*	Hits on which the batter reached first base safely without the contribution of a fielding error.
2B	*Double*	Hits on which the batter reached second base safely without the contribution of a fielding error.
3B	*Triple*	Hits on which the batter reached third base safely without the contribution of a fielding error.
AB	*At Bat*	Batting appearances, not including bases on balls, hit by pitch, sacrifices, interference, or obstruction.
AB/HR	*At Bats per Home Run*	At bats divided by home runs.
AO/GO	*Ground Ball Fly Ball ratio*	Number of ground balls outs divided by number of fly balls out.
BA	*Batting Average*	Hits divided by at bats. ($BA = {}^{H}/_{AB}$)
BB	*Base on Balls ("Walk")*	Time receiving four balls and advancing to first base.

BB/K	*Walk-to-strikeout ratio*	Number of base on balls divided by number of strikeouts.
FC	*Fielder's Choice*	Times reaching base when a fielder chose to try for an out on another runner.
GDP or GIDP	*Grounded into Double Play*	Number of ground balls hit that became double plays.
GS	*Grand Slam*	A home run with the bases loaded, resulting in four runs scoring, and four RBI credited to the batter.
H	*Hits*	Time reached base because of a batted, fair ball without error by the defense.
HBP	*Hit by Pitch*	Times touched by a pitch and awarded first base as a result.
HR	*Home Runs*	Hits on which the batter successfully touched all four bases, without the contribution of a fielding error.
IBB	*Intentional Base on Balls*	A base on balls (see BB above) deliberately thrown by the pitcher. Also known as IW (intentional walk).
K	*Strike Out*	Number of times that strike three is taken or swung at and missed, or bunted foul.
LOB	*Left on Base*	Number of runners not out nor scored at the end of an inning.
OBP	*On Base Percentage*	Times reached base (H + BB + HBP) divided by at bats plus walks plus hit by pitch plus sacrifice flies (AB + BB + HBP + SF).
OPS	*On-Base Plus Slugging*	On-base percentage plus slugging average.
PA	*Plate Appearance*	Number of completed batting appearances.
RBI	*Runs Batted in*	Number of runners who scored due to a batters' action, except when batter grounded into double play or reached on an error.
RC	*Runs Created*	Statistic that attempts to measure how many runs a player has contributed to his team.

SF	*Sacrifice Fly*	Number of fly ball outs which allow another runner to advance on the base paths or score.
SH	*Sacrifice Hit*	Number of sacrifice bunts which allows another runner to advance on the base paths or score.
SLG	*Slugging Average*	Total bases divided by at-bats.
TA	*Total Average*	Total bases plus walks, plus steals, divided by plate appearances plus caught stealing. $\left(TA = \frac{TB + BB + Steals}{PA + CS} \right)$
TB	*Total Bases*	One for each single, two for each double, three for each triple, and four for each home run.
TOB	*Times on Base*	Times reaching base as a result of hits, walks, and hit-by-pitches.
XBH	*Extra Base Hits*	XBH = *Doubles* + *Triples* + HR

Table 12: Fielding statistics abbreviations. Adapted from [51].

A	*Assists*	Number of outs recorded on a play where a fielder touched the ball, except if such touching is the putout.
DP	*Double Plays*	One for each double play during which the fielder recorded a putout or an assist.
E	*Errors*	Number of times a fielder fails to make a play he should have made with common effort, and the offense benefits as a result.
FP	*Fielding Percentage*	Total plays (chances minus errors) divided by the number of total chances. $\left(FP = \frac{A + PO}{A + PO + E} = \frac{A + PO}{TC} \right)$
INN	*Innings*	Number of innings that a player is at one certain position.
PB	*Passed Ball*	Charged to the catcher when the ball is dropped and one or more runners advance.

PO	*Putout*	Number of times the fielder tags, forces, or appeals a runner and he is called out as a result.
RF	*Range Factor*	Used to determine the amount of field that the player can cover. $\left(\text{RF} = \frac{(\text{PO} + \text{A}) \cdot 9}{\text{Number of Innings}} \right)$
SB	*Stolen Bases*	Number of times a runner advanced on the pitch without being thrown out by the catcher.
TC	*Total Chances*	Assists plus putouts plus errors. (TC = A + PO + E)
TP	*Triple Play*	One for each triple play during which the fielder recorded a putout or an assist.

Table 13: General statistics abbreviations. Adapted from [51].

G	*Games Played*	Number of games where the player played, in whole or in part.

Table 14: Pitching statistics abbreviations. Adapted from [51].

BABIP	*Batting Average on Balls in Play*	Batting average against a pitcher on batted balls ending a plate appearance, excluding home runs.
BB	*Base on Balls ("walk")*	Times pitching four balls, allowing the batter-runner to advance to first base.
BB/9		Base on balls times nine divided by innings pitched (Bases on balls per 9 innings pitched).
BF	*Total Batters Faced*	Opponent's total plate appearances.
BK	*Balk*	Number of times pitcher commits an illegal pitching action or other illegal action while in contact with the pitching rubber, thus allowing baserunners to advance.
BS	*Blown Save*	Number of times entering the game in a save situation, and being charged the run which ties the game.

CERA	*Component ERA*	An estimate of a pitcher's ERA based upon the individual components of his statistical line (K, H, 2B, 3B, HR, BB, and HBP).
CG	*Complete Game*	Number of games where player was the only pitcher for his team.
DICE	*Defense-Independent Component ERA*	An estimate of a pitcher's ERA based upon the defense-independent components of his statistical line (K, HR, BB, and HBP).
ER	*Earned Run*	Number of runs that did not occur as a result of errors or passed balls.
ERA	*Earned Run Average*	Earned runs times innings in a game (usually nine) divided by innings pitched.
G	*Games Pitched*	Number of times a pitcher pitches in a season.
G/F	*Ground Ball Bly Ball ratio*	Ground balls allowed divided by fly balls allowed.
GF	*Games Finished*	Number of games pitched where player was the final pitcher for his team.
GS	*Starts*	Number of games pitched where player was the first pitcher for his team.
H	*Hits Allowed*	Total hits allowed.
H/9	*Hits per Nine Innings*	Hits allowed times nine divided by innings pitched (also known as H/9IP - Hits allowed per 9 innings pitched).
HB	*Hit Batsman*	Times hit a batter with pitch, allowing runner to advance to first base.
HLD	*Hold*	Number of games entered in a save situation, left in save situation, recorded at least one out, and not having surrendered the lead.
HR	*Home Runs Allowed*	Total home runs allowed.
IBB	*Intentional Walks Allowed*	

IP	*Innings Pitched*	Number of outs recorded while pitching divided by three. $\left(IP = {}^{Outs}/_{3} \right)$
IP/GS		Average number of innings pitched per game started.
IR	*Inherited Runners*	Number of runners on base when the pitcher enters the game.
IRA	*Inherited Runs Allowed*	Number of inherited runners allowed to score.
K	*Strikeout*	Number of batters who received strike three.
K/9	*Strikeouts per Nine Innings*	Strikeouts times nine divided by innings pitched (Strikeouts per 9 innings pitched).
K/BB	*Strikeout-to-Walk ratio*	Number of strikeouts divided by number of base on balls.
L	*Loss*	Number of games where pitcher was pitching while the opposing team took the lead, never lost the lead, and went on to win.
OBA	*Opponents' Batting Average*	Hits allowed divided by at-bats faced.
PIT	*Pitches Thrown*	(Pitch count)
RA	*Run Average*	Number of runs allowed times nine divided by innings pitched.
RAA	*Runs Against Average*	A sabermetric statistic to predict win-percentage.
SO	*Shutout*	Number of complete games pitched with no runs allowed.
SV	*Save*	Number of games where the pitcher enters a game led by the pitcher's team, finishes the game without surrendering the lead, is not the winning pitcher, and either (a) the lead was three runs or less when the pitcher entered the game; (b) the potential tying run was on base, at bat, or on deck; or (c) the pitcher pitched three or more innings.

W	*Win*	Number of games where pitcher was pitching while his team took the lead and went on to win (also related: winning percentage).
WP	*Wild Pitches*	Charged when a pitch is too high, low, or wide of home plate for the catcher to field, thereby allowing one or more runners to advance or score.

A2 – The Science of the Swing

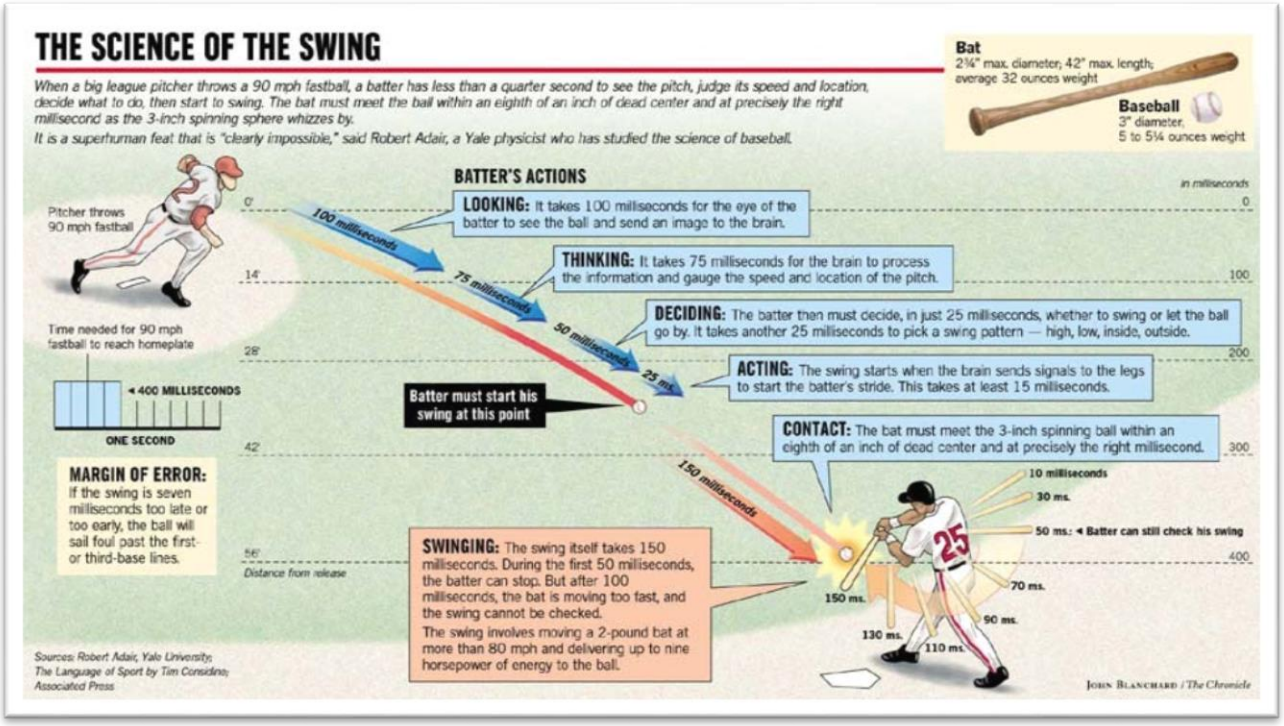

Figure 11: The science of the swing. Adapted from [1].

A3 – Changes in the MLB Setup

Table 15: How number of teams and playoffs spots have changed through the years.

Year	Nr. of Teams			Nr. of Divisions	Nr. of Playoffs Teams (Wild Card Teams)		
	AL	*NL*	*Total*		*AL*	*NL*	*Total*
1969	12	12	24	4	2	2	4
1977	14	12	26	4	2	2	4
1993	14	14	28	4	2	2	4
1994	14	14	28	6	4 (1)	4 (1)	8
1998	14	16	30	6	4 (1)	4 (1)	8
2012	14	16	30	6	5 (2)	5 (2)	10
2013	15	15	30	6	5 (2)	5 (2)	10

A4 – Distribution of WAR_{League}

Table 16: Distribution of WAR_{League} between the leagues through the years. Adapted from [24].

Year	WAR NL	WAR AL	Affective changes
2013	475	525	HOU => AL
1998 – 2012	487	513	Expansion in 1998
1996 – 1997	420	513	
1995	373	456	Strike
1994	298	364	Strike
1993	420	513	Expansion
1992	353	513	
1989 – 1991	377	490	
1988	398	465	
1982 – 1987	400	467	
1981	264	308	Strike
1980	399	466	
1979	376	489	
1977 – 1978	377	490	Expansion in 1977
1975 – 1976	379	419	
1973 – 1974	380	420	
1972	364	402	Strike
1971	399	399	
1969 – 1970	400	400	

A5 – Root Mean Square Deviation

Root mean square deviation (RMSD), also known as root mean square error (RMSE) or population standard deviation, is used to compare variation between two datasets with the following equation [52].

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_{Predicted} - x_{Actual})^2}{N}} \quad (46)$$

a. $x_{Predicted}$ is the predicted team position.
b. x_{Actual} is the real life team position.
c. N is the number of teams in the comparison.

This method is ideal to compare the predicted final positions of each team to their actual final position for all the years.

Table 17: Comparison of two models' RMSD for league positions.

	AL	NL	Overall
The Baseline Model	3.72	4.11	3.92
The Project's Simple Model	3.76	4.16	3.96

Applying the RMSD on the baseline model and the project's simple model gives almost identical results. Both have a standard deviation close to four for the league as a whole and less than four for AL and approximately four for the NL.

A6 – Logistic Regression for AL with DH included

Table 18: Classification matrix for AL where DH variable is included.

DH model for AL - Total		
	No	*Yes*
No	283	89
Yes	89	121
	TPR	57.6%
	ACR	69.4%

This classification matrix shows results for a logistic model for the American League where it includes designated hitters in the stepwise regression process. However, the DH variable is not included in the final equation. The results are worse than the baseline model and the simple project's model, due to the facts that no further modelling with DH was needed.

Table 19: Summary of the logistic regression for the AL's designated hitter model.

	Coeff. b	**Std. Error**	**e^b**	**z-value**	**p-value**	
Intercept	-9.698	2.681	6.1e-05	-3.618	< 3e-04	***
MinWAR	-0.401	0.249	0.670	-1.610	0.108	
AvgAge	0.216	0.097	1.241	2.232	0.026	*
AvgWARPitchers	0.910	0.271	2.485	3.363	7.7e-04	***
AvgWAR1B	0.096	0.064	1.101	1.499	0.134	
AvgWARC	0.196	0.082	1.217	2.386	0.017	*
AvgWARCF	0.170	0.058	1.185	2.922	0.003	**
AvgWARRF	0.224	0.065	1.252	3.463	5.4e-04	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 491.84 on 376 degrees of freedom.

Residual deviance: 415.37 on 369 degrees of freedom.

AIC = 431.37