Springboard DSC Capstone Project 2 Milestone 2 Report:
Prediction of MLB player WAR statistic

Albert Chiu

March 2019

Problem

MLB teams are constantly seeking to improve their rosters through free agency signings, trades and the development of minor league players in farm systems. For these teams, player salaries represent a significant financial cost and it is important to be able to effectively evaluate a player's value.

WAR (Wins Above Replacement) is commonly used as a metric to determine player value; it is the sum of a player's total offensive and defensive contributions and attempts to measure the total number of additional wins a team can expect with the player over an average replacement level player. A simplified version of the equation is:

$$WAR = {}^{(BR + BRR + FR + PA + LA + RR)}/_{RPW}$$

BR = Batting Runs
BRR = Base Running Runs
FR = Fielding Runs
PA = Positional Adjustment
LA = League Adjustment
RR = Replacement Runs
RPW = Runs Per Win

A common challenge for MLB teams is forecasting the future production of a player. Teams will acquire a star player at the height of their careers for a large contract, only to discover that they've anchored themselves and a significant portion of payroll to a player whose production is quickly diminishing. Other times, teams will overlook a young player that will eventually become a superstar or an aging veteran that continues to provide solid on-field play. If MLB teams are able to better forecast when a player's production is on the verge of decline or improvement, it will allow them to make more informed player acquisition decisions.

Data Sets

Lahman's Baseball Database: http://www.seanlahman.com/baseball-archive/statistics/
Baseball Prospectus: https://www.baseballprospectus.com/

EDA and Data Wrangling

For this project, I obtained two csv data sets from Baseball Prospectus (containing primarily offensive stats) and Lahman's Baseball Database (containing primarily defensive stats). The csv files were read into dataframes and then merged together after converting the Baseball Prospectus player name convention (full first and last name) into the Lahman convention (first 5 letters of last name followed by first two letters of first name). There were a lot of special cases that needed to be addressed including:

- Last names that contain multiple whitespaces, ex. Tomas de la Rosa
- Names that end with Jr. or Sr., ex. Jose Cruz Jr.
- Names that have initialed first name with whitespace separation, ex. J. T. Bruett
- Names that have initialed first name without whitespace separation, ex. J.T. Snow
- Names with middle initial, ex. Bobby J. Jones
- Names that contain middle names, ex. Chan Ho Park
- Names that contain whitespace in first name, ex. La Marr Hoyt

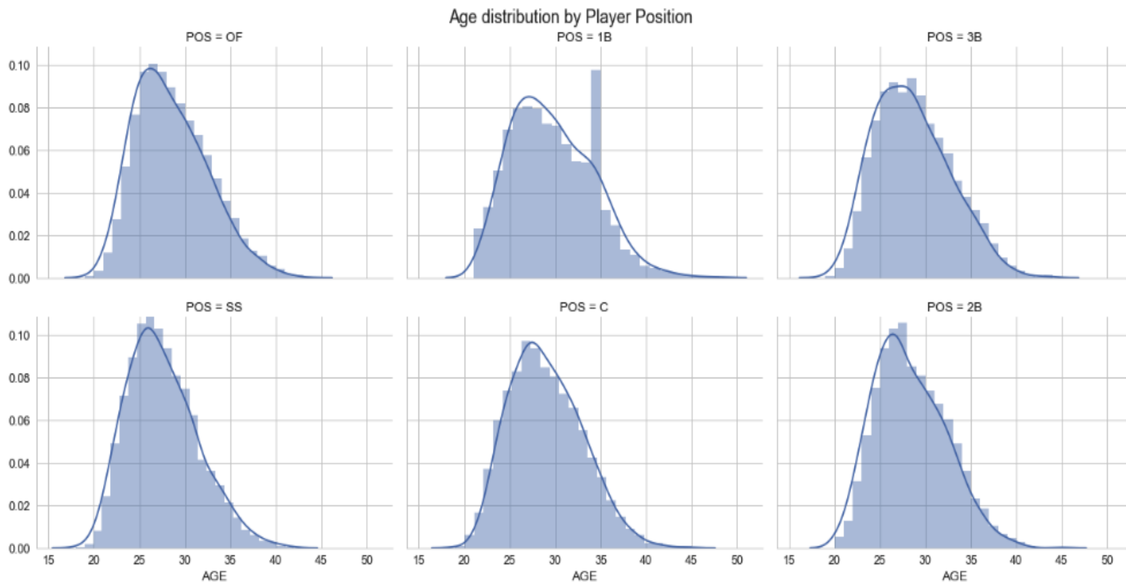
There were a number of players that played multiple positions over the course of a single season so there would be a separate row for each position played. To simplify things, I calculated the position that the player played the most inning outs for the season and aggregated all their stats to that position.

There were also numerous instances where a player would play for multiple teams over the course of one season. For these players, I merged the statistics for all teams played in a season to a single row.

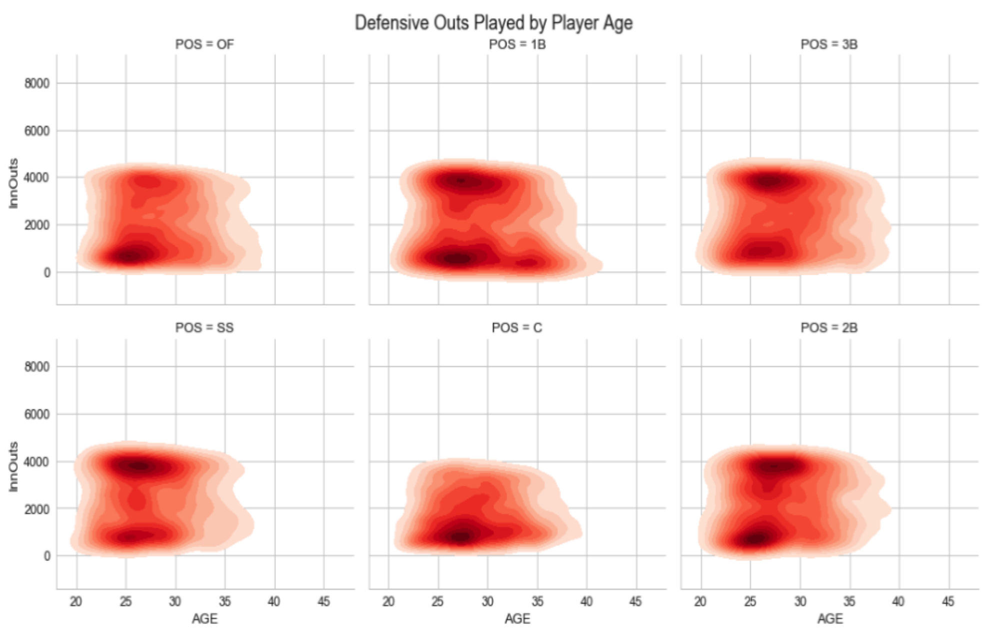
The final data set after the data wrangling is a dataframe where each row represents all the statistics of a player for a single season. It contains 18,621 unique player seasons and covers the span of 40 MLB seasons (1977 through 2017).

Github link to my data wrangling notebook: https://github.com/albertdchiu1/Capstone-2-Project-MLB-WAR-prediction/blob/master/Code/1.%20Capstone%202%20-%20Data%20Wrangling.ipynb

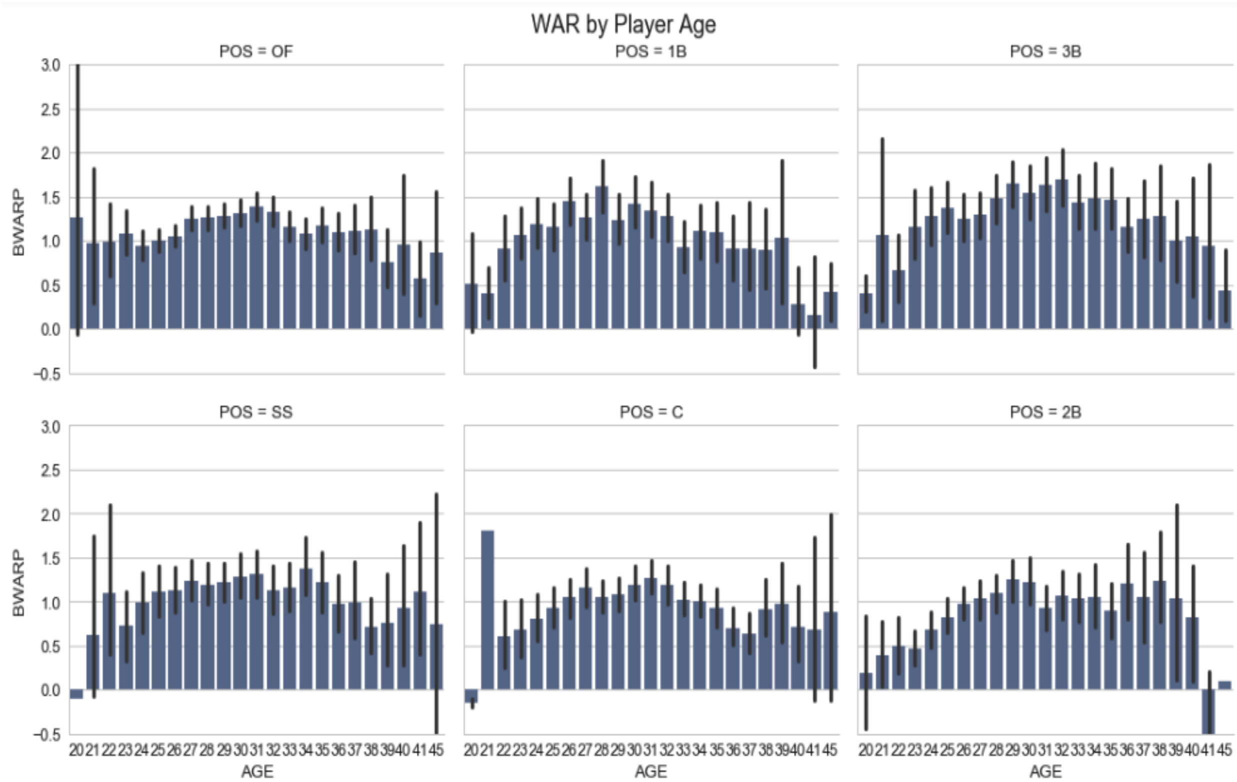
Storytelling



When we examine the age distribution of the players by position, it appears that most position are filled by players in their late 20’s and early 30s. First basemen seem to have slight longer longevity than other positions, while shortstop tend to skew younger.



When we create contour plots of the defensive outs played by players by position, we can see that different positions have different tendencies. Catchers tend to player fewer defensive outs than other positions; for 1B, 2B, 3B and SS, we see two peaks for players in their mid to late 20’s that play approx. 4000 outs and those that play several hundred outs.



When we plot the WAR by age for each position, we observe that for most positions, WAR peaks in late 20's and early 30's. Additionally, we see that there is more variance in WAR statistic for younger players (less than 25 yo) and older players (more than 35 yo) than players in their late 20's and early 30's.

Github link to data visualization notebook: https://github.com/albertdchiu1/Capstone-2-Project-MLB-WAR-prediction/blob/master/Code/2.%20Capstone%202%20-%20Data%20Visualization.ipynb

Inferential Statistics

To determine if there are statistically significant differences between players above and below specified WAR percentiles, hypothesis testing was applied to different features in the data set. Applying statistics on the data set, we gain some insights:

1. When plotting a correlation table for the features in the data set, we observe that WAR is most highly correlated with TB, DRC+, RBI, R, HR, BB and DRAA features.
2. For most positions, players in the top 5% of BWARP have a statistically different OBP than players in the bottom 95% of BWARP.
3. For shortstops, it appears that there is a large difference in DRAA between players below the 50th BWARP percentile and those above the 50th BWARP percentile.
4. For most positions, players in the top 5% of BWARP have a statistically different age than players in the bottom 95% of BWARP.

Github link to inferential statistics notebook: https://github.com/albertdchiu1/Capstone-2-Project-MLB-WAR-prediction/blob/master/Code/3.%20Capstone%202%20-%20Inferential%20Statistics.ipynb

In-Depth Analysis with Machine Learning

The goal for this project will be making a prediction of the WAR statistic for baseball players in the MLB. Because of this, regression algorithms were used for modeling.

The data set was split into training and test sets (80:20 split) and several regression algorithms were fitted to the training set, including XGBoost, Random Forest and Linear Regression. For the Random Forest and XGBoost models, hyperparameter tuning was used to optimize model performance.

To measure the performance of the model, several metrics were calculated: root mean square error, R2, mean absolute error and mean absolute percentage error. Additionally, a histogram of the residual distribution was plotted for each model to visualize the distribution of the deviation from the actual WAR value.

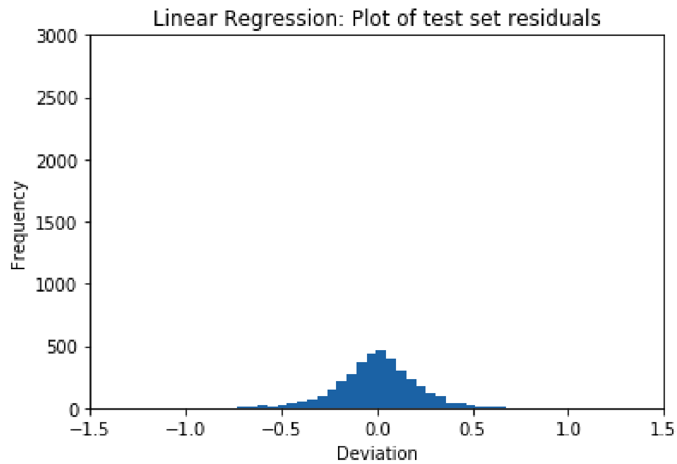
Of the three algorithms implemented, XGBoost was the top performer. We can see that it achieves the best scores in RMSE, MAE and MAPE for both training and test sets and the plot of the residuals of the XGBoost model indicates that it has the smallest deviation from the true WAR value.

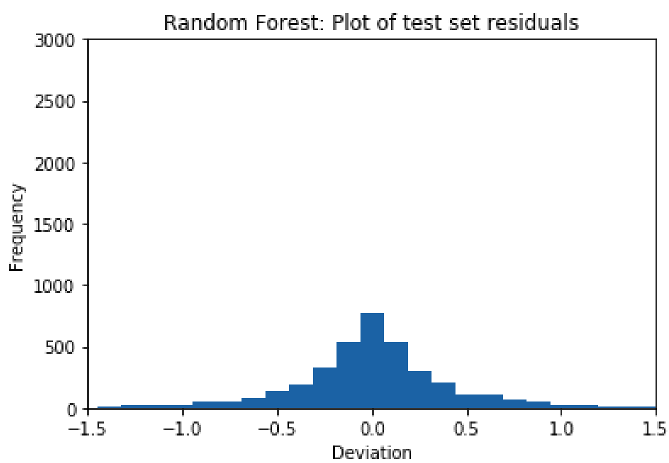
Model Performance:

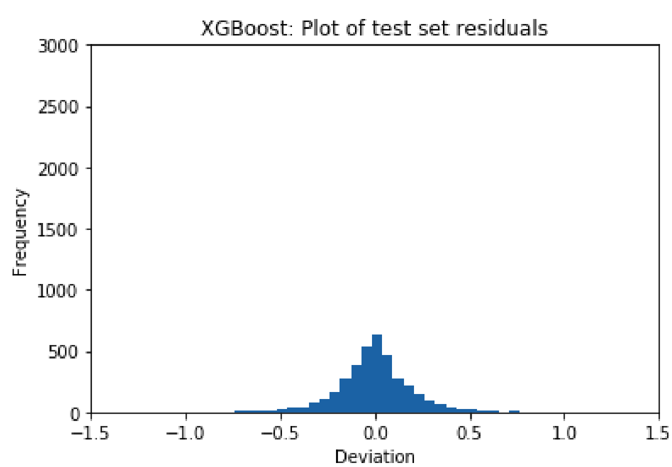
	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Train	MAE Test
Linear Reg	0.0449	0.0454	0.9839	0.9839	0.153	0.1563
Random Forest	0.0335	0.2312	0.9879	0.9182	0.1201	0.3079
XGBoost	0.0135	0.04	0.9952	0.9858	0.0859	0.1331

	MAPE Train	MAPE Test
Linear Reg	38.126%	41.97%
Random Forest	23.341%	55.391%
XGBoost	22.358%	28.245%

Plots of residuals:







In the next step, the top performing model, XGBoost, was used to determine feature importance. The following statistics were identified as contributing most to predicting WAR value: R, TB, DRC+, BRR, InnOuts, Def_A, Def_CS, Def_DRAA, Def_FRAA, POS_1B.

Github link to machine learning notebook: https://github.com/albertdchiu1/Capstone-2-Project-MLB-WAR-prediction/blob/master/Code/4.%20Capstone%202%20-%20Machine%20Learning.ipynb

Conclusions and Future Work

From this project, I learned that we can predict WAR with a fair degree of accuracy using XGBoost (averaging MAPE of 22.4% for training data and 28.2% for test data). We observe that the test data performs slightly worse than the training data, suggest there is some overfitting. While Linear Regression comes in a close second, we see that the difference in performance between training and test sets is much smaller than in XGBoost.

In a future update of this project, I will be attempting to apply time series forecasting techniques to predict WAR statistics for specific players in an upcoming season.

Additionally, I would like to take into consideration player salaries to determine how teams can maximize their projected wins given their player salary budget. This will help teams identify players that can provide the most bang for buck while avoiding players that are overcompensated for their performance.

Recommendations for Client

From the results of this project, we have obtained several insights in addition to building a model that predicts a players WAR statistic:

1. While first basemen average more home runs than other positions, they see a drop off when they hit their early 30's. While second basemen average fewer home runs, they maintain the number of home runs hit thru their 30's better than other positions.
2. First basemen have greater longevity in the league than players in other positions. Shortstops tend to skew younger (average age of MLB SS is 27.5 yo while other positions average 28 to 29.5 yo).
3. Compared to other positions, third basemen average the highest WAR value (1.41); second basemen (0.95) and catchers (1.04) average the lowest WAR values.

Using this information, along with the specific needs and budget of each organization, MLB teams can make more informed personnel decisions.

General Statistics:
POS - Position
G - Games Played
GS - Games Started
League - Identifies American or National League
BWARP – Wins Above Replacement

Offensive Statistics:
PA - Plate Appearances
AB - At Bats
R - Runs
H - Hits
1B - Singles
2B - Doubles
3B - Triples
HR - Home Runs
TB - Total Bags
BB - Walks
IBB - Intentional Walks
SO - Strike Outs
SF - Sacrifice Flys
SH - Sacrifice Bunt
RBI - Runs Batted In
DP - Double Plays
SB - Stolen Bases
CS - Caught Stealing
AVG - Batting Average
OBP - On Base Percentage
SLG - Slugging
OPS - On Base Plus Slugging
ISO - Isolated Power
oppOPS - Aggregate OPS of pitchers faced
DRC+ - Deserved Runs Created Plus
BRR - Base Running Runs

Defensive Statistics:
InnOuts - Defensive Outs Played
Def_PO - Put Outs
Def_A - Assists
Def_E - Errors
Def_DP - Double Play
Def_PB - Passed Ball (Catchers only)
Def_WP - Wild Pitches (Catchers only)
Def_SB - Stolen Bases (Catchers only)
Def_CS - Caught Stealing (Catchers only)
Def_ZR - Zone Rating (Catchers only)
Def_DRAA - Defensive Runs Above Avg
Def_FRAA - Fielding Runs Above Avg