Capstone Project: Prediction of MLB WAR Statistic

Albert Chiu - March 2019

Problem

- MLB teams are constantly seeking to improve their rosters.
- A big challenge for teams is predicting future production of a player:
    - A star player is signed to a large contract only to have his on field production immediately deteriorate.
    - Teams overlook a young player that eventually becomes a superstar or an aging veteran that continues to provide solid performance.

WAR (Wins Above Replacement)

- WAR is a commonly used metric to determine player value.
- It is a summation of a player's offensive and defensive contributions and attempts to measure the total number of wins a team can expect with the player over an average replacement level player.

CALIBER OF PLAYER	WINS ABOVE REPLACEMENT
BENCH GUY	0-1 WAR
ROLE PLAYER	1-2 WAR
SOLID STARTER	2-3 WAR
ABOVE-AVERAGE	3-4 WAR
ALL-STAR	4-5 WAR
SUPERSTAR	5-6 WAR
MVP	6+ WAR

Data Sets

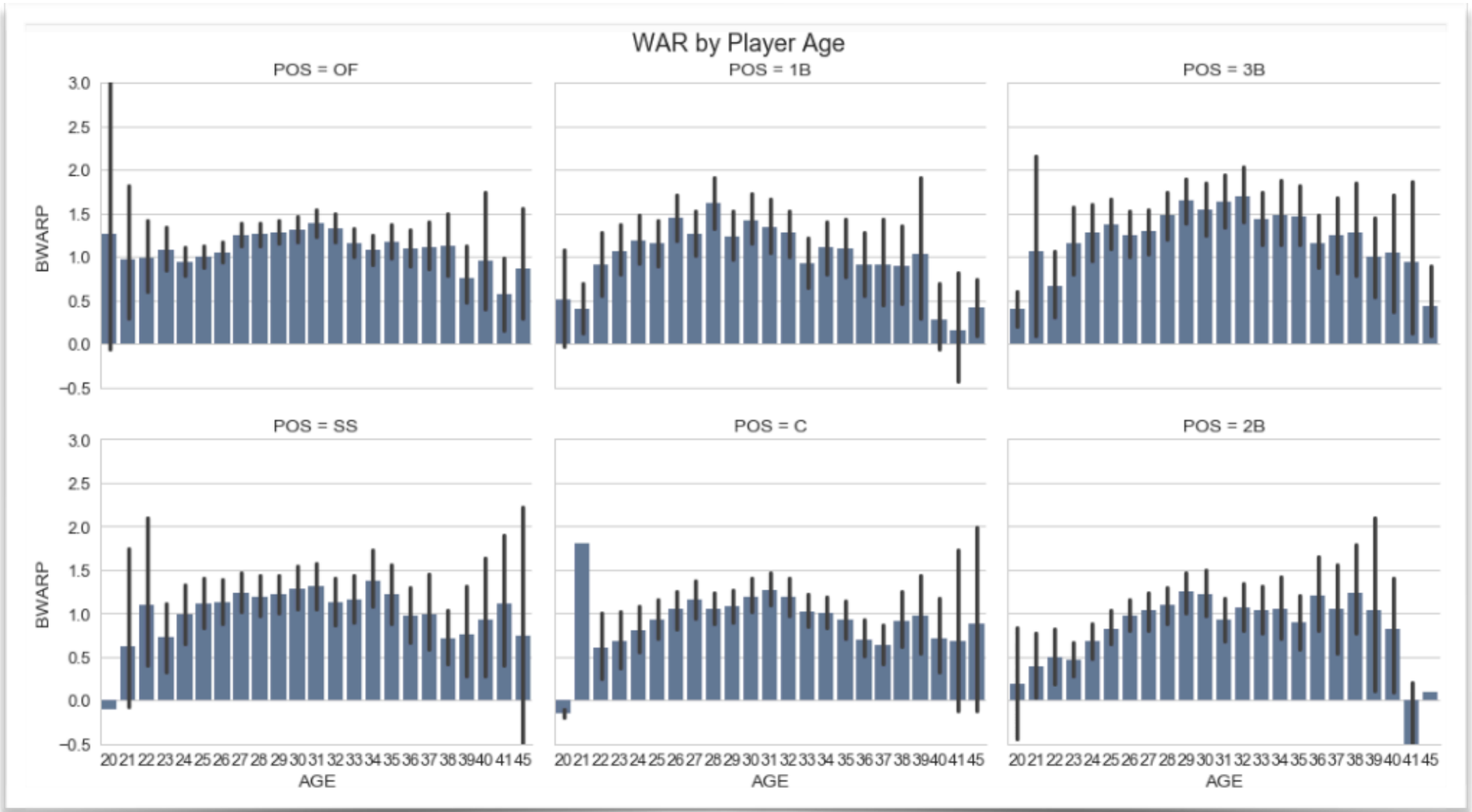
- Two data sets were used for this project:
  - Baseball Prospectus: https://legacy.baseballprospectus.com/sortable/index.php?cid=2762830
  - Lahman's Baseball Database: http://seanlahman.com/files/database/baseballdatabank-master_2018-03-28.zip
- The merged dataframe contains 18,621 unique player seasons and covers a span of 40 MLB seasons (1977 through 2017).

Project Goal

- To develop a model that predicts a position player's (non-pitcher) future WAR statistic.
- Use feature importance to determine which features contribute most to a high value of WAR.

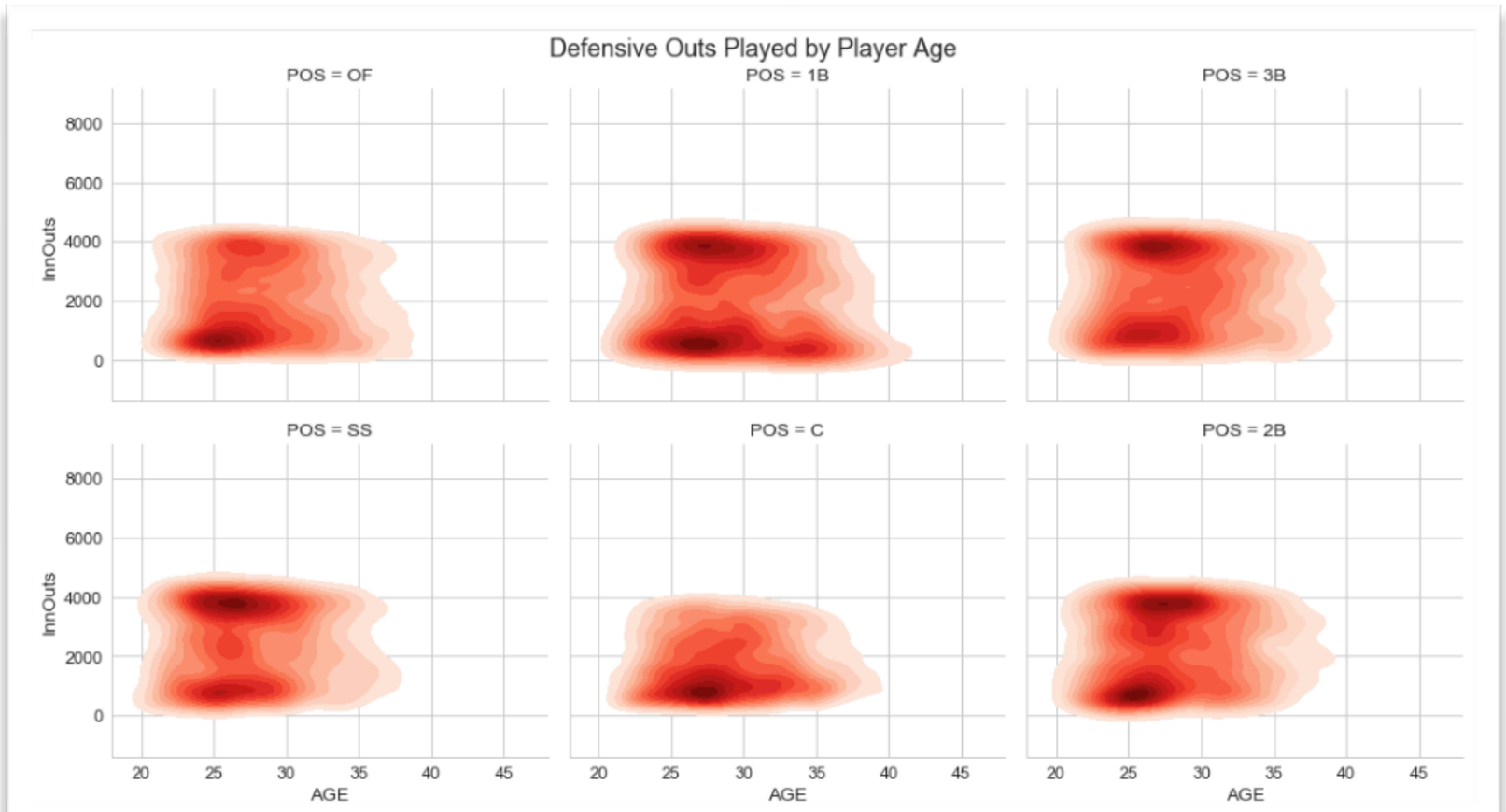
WAR by Player Age

- We observe that for most positions, WAR peaks in late 20's and early 30's.



Defensive Outs by Player Age

- Using a contour plot, we see that catchers tend to play the fewest outs.
- We observe for 1B, 2B, 3B and SS that there is a clustering of mid-20 yo players that play 4000 outs and a clustering that play several hundred outs.



Inferential Statistics

- Performing inferential statistics on the data set, we gain several 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.

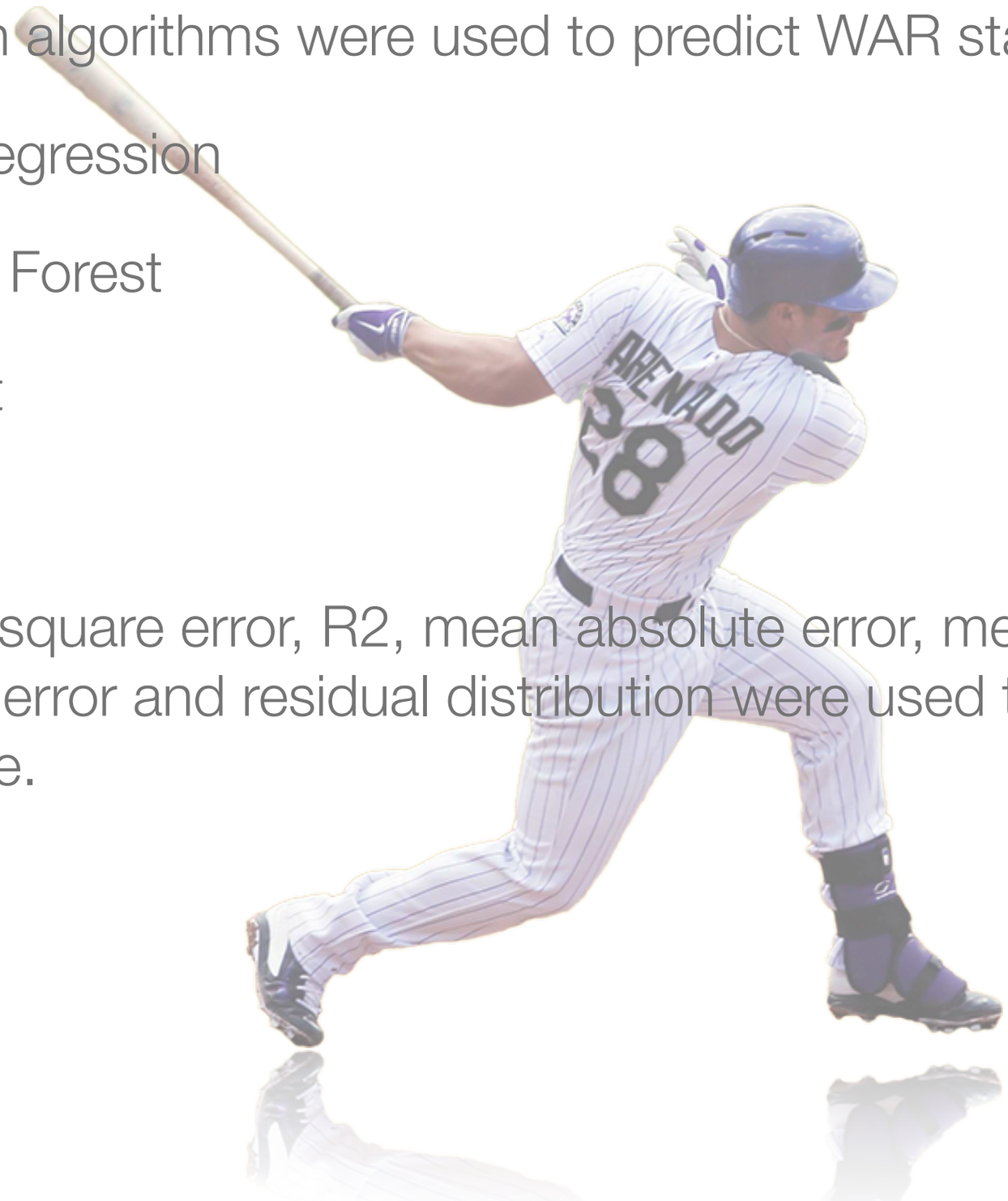
2. For most positions, players in the top 5% of BWARP have a statistically different OBP than players in the bottom 95% of BWARP.

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.

Machine Learning

- 3 regression algorithms were used to predict WAR statistic:
  - Linear Regression
  - Random Forest
  - XGBoost



- Root mean square error, R2, mean absolute error, mean absolute percentage error and residual distribution were used to determine model performance.

Machine Learning

- XGBoost is the top performing model:

Model Performance:

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Train	MAE Test
Linear Reg	0.0449	0.0454	0.9839	0.9838	0.1528	0.1569
Random Forest	0.0346	0.2312	0.9876	0.9177	0.1208	0.3157
XGBoost	0.022	0.0437	0.9921	0.9844	0.1084	0.1451

	MAPE Train	MAPE Test
Linear Reg	38.719%	39.258%
Random Forest	23.648%	53.725%
XGBoost	26.018%	29.778%

- Linear Regression comes in a close second and has a less of a performance drop off between training and test data sets.

Recommendations

- We've gained several insights that will help MLB teams make more informed personnel decisions:
    - First basement average more home runs than other positions but see a decline in their early 30's.
    - Third basemen average the highest WAR value (1.41); second basemen (0.95) and catchers (1.04) average the lowest WAR values
    - First basemen have greater longevity in the league than other position players. Shortstops tend to skew younger (average age of MLB shortstop is 27.5 yo while other positions average between 28 to 29.5 yo)