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

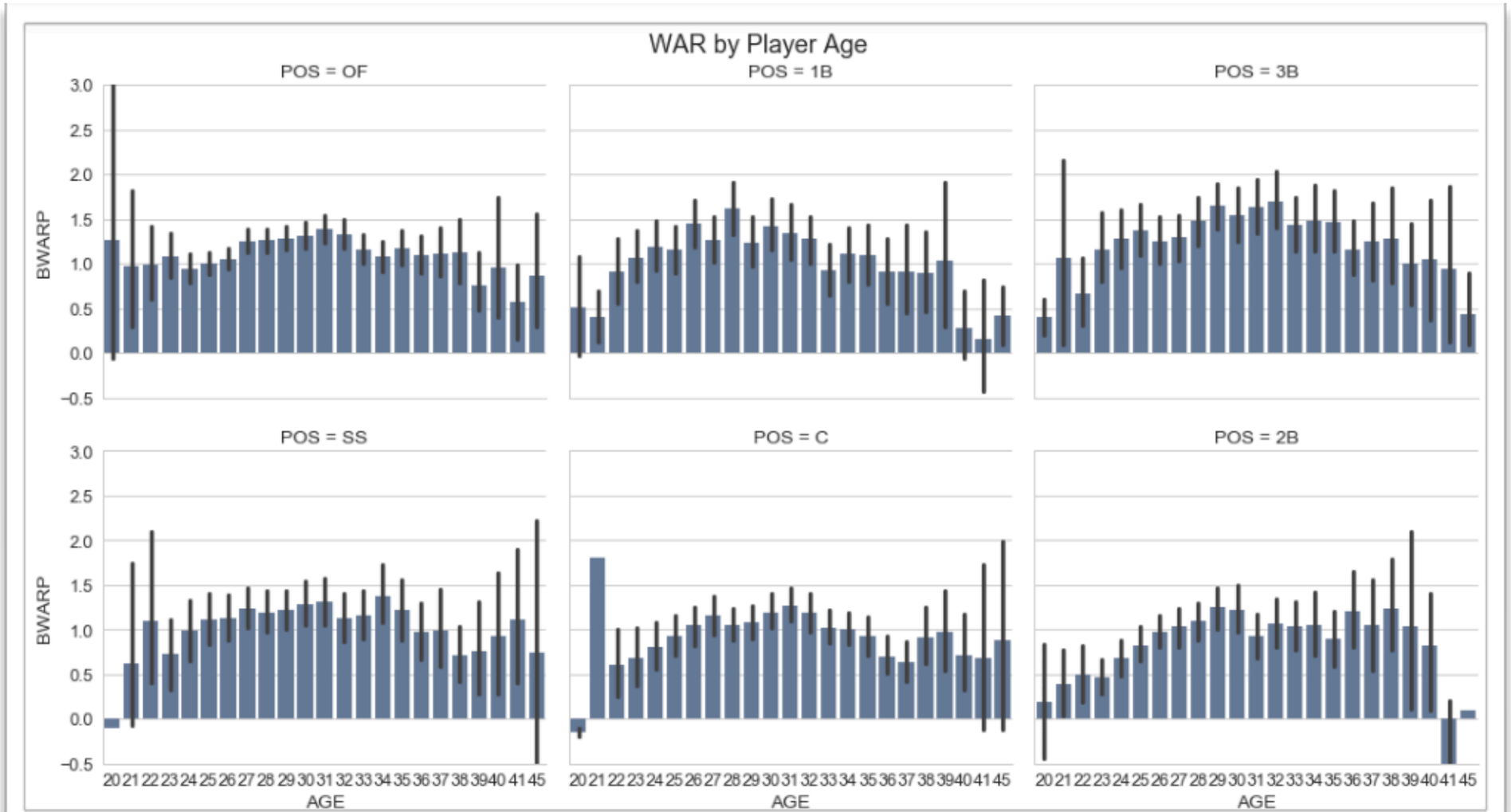
- Two data sets were used for this project:
  - Baseball Prospectus: https://legacy.baseballprospectus.com/sortable/index.php?cid=2762830
  - Lahman's Baseball Database: http://seanlahman.com/files/database/baseballdatabank-master_2018-03-28.zip
- The merged dataframe contains 18,621 unique player seasons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

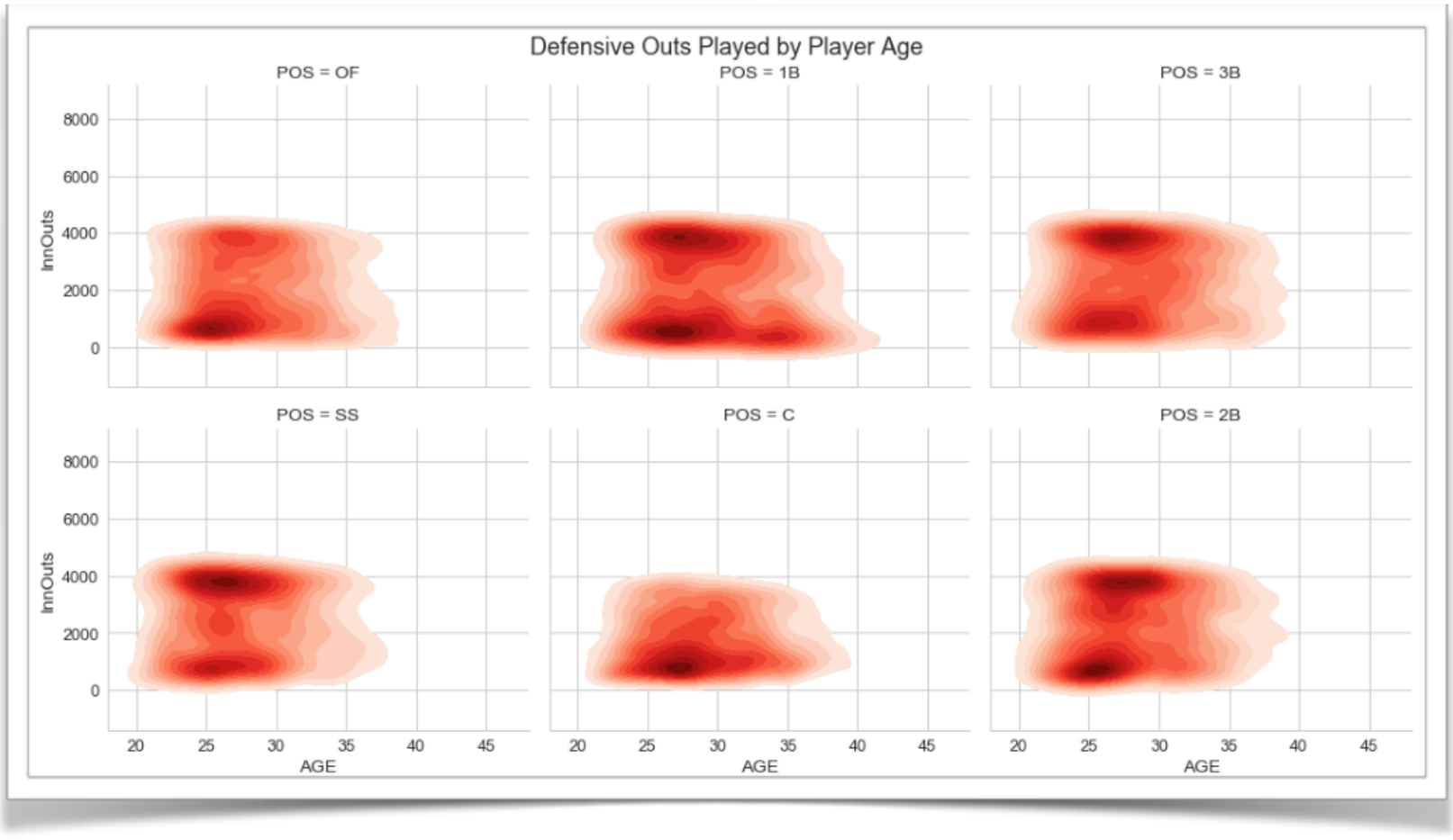
WAR by Player Age

- 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.