MLB xBA Prediction

Austin Kingsley
MSDS 696 - Practicum II

Research Question

Is the expected batting average of MLB players able to be predicted by using recent performance of both batters and pitchers, as well as the matchup potential between the relevant parties using available Statcast pitch metrics?

Data

- All game files are retrieved from the BaseballSavant API, which are supplied by the MLB.
- Statcast pitch data, which measures various pitch metrics, is a recently available metric which will be used in the pitcher-batter matchup component of the project.
- 47 columns are removed due to deprecation, irrelevance, or redundancy.
- 23 columns for Statcast pitch data are included in the assignment, and will be the basis for determining batter-pitcher matchup.
- 3 columns (expected batting average, strikeouts, and walks) will be formatted in various ways to represent batter and pitcher performance over their appearances.
- This assignment is narrowed to regular season games to handle the majority of missing data and for clearer scope.

Methodology

For this project, there are three components that will be done individually and brought together at the end:

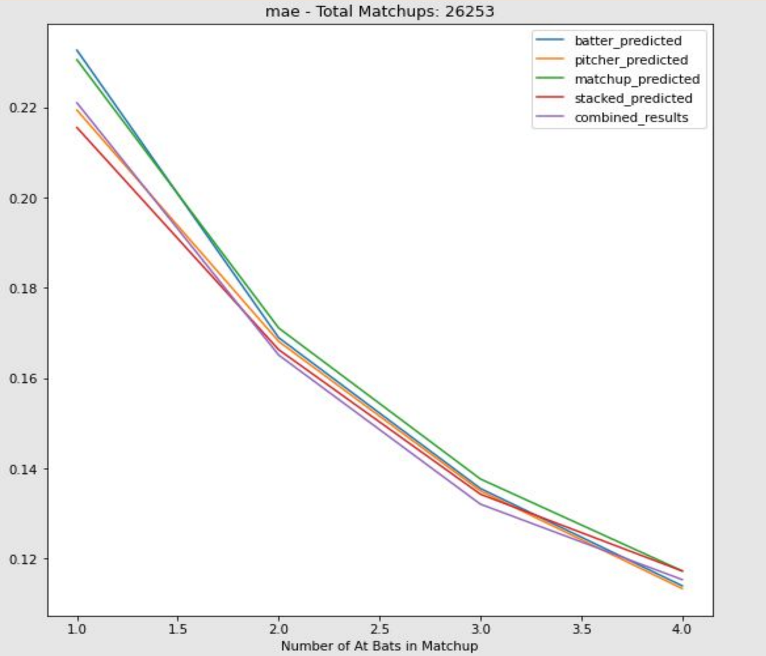
- Pitcher performance based on previous appearances, involving differentiation of relief appearances and starts. This will involve separation and combination of days going back so many games.
  - Changes in pitch metrics have proven to be a worthwhile option to explore as well. In future iterations.
- Batter performance based on previous at-bats, which will take into account days off and study the impact of rest days for athletes in relation to performance.
- Pitcher-Batter Matchup will involve recently available Statcast data to compare pitcher repertoire and batter performance against these thrown pitches historically.
  - Soft clustering and hard clustering have been incorporated into the final version.

Results

The error rate is the mean absolute error weighted by plate appearance. This captures the overall impact of model.

- Matchup Specific Modeling was compared using five different models
    - When using Matchup Specific Information alone, an overall error rate of 0.183 was achieved
    - When using Pitcher Specific Information alone, an overall error rate of 0.176 was achieved
    - When using Batter Specific Information alone, an overall error rate of 0.182 was achieved
    - When Batter and Pitcher Predictions were added to the Matchup Specific Model, an overall error rate of 0.175 was achieved.
    - When predictions for the Batter, Pitcher, and Matchup Specific Models were put together, an overall error rate of 0.174 was achieved.
- As the number of at bats per matchup increases, the error rate reduces by approximately 0.100 for all models.

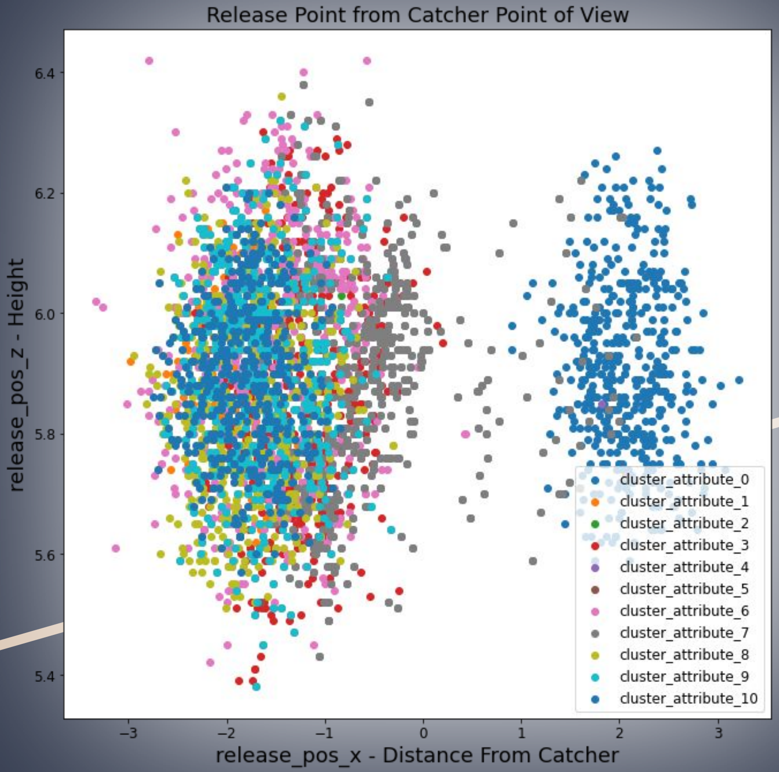
Model Performance by Matchup PA



- The test set contained 26253 matchups.
    - 64.3% of matchups contained one plate appearance.
    - 16.1% of matchups contained two plate appearances.
    - 18.4% of matchups contained three plate appearances.
    - 0.79% of matchups contained four plate appearances.
    - There are no qualifying matchups above containing five or more plate appearances.

Appearance per Matchup	Batter Model	Pitcher Model	Matchup Model	Stacked Model	Combined Model
1	0.2404647	0.2247453	0.2362080	0.2112090	0.2226849
2	0.1777279	0.1765643	0.1780783	0.1713727	0.1715528
3	0.1401068	0.1392430	0.1408975	0.1366073	0.1346946
4	0.1167547	0.1171960	0.1180898	0.1171464	0.1133636

Conclusions



- Conventional wisdom that “pitching wins games” appears to be true for xBA as well.
- Matchup details introduced improvements to player performance, even if minimal.
- There are a number of ways to improve this model, which include:
  - The clusters used to represent pitches is inadequate, because it does not fully represent left handed pitchers.
  - Batter metrics against certain pitches, is traditionally calculated based on the last pitch of an at bat. Instead, representing all pitches through a weighted method could represent all pitches seen while maintaining focus on the most impactful pitch.
  - Random Forest was used for all predictive modeling due to its reliability and robustness. Different models can be tested for better suited models.

Contact

Thank you for reviewing my MLB xBA Prediction Project.

If you have any comments or feedback, I can be reached at:

- akingsley319@yahoo.com
- (773) 341-8026
- www.linkedin.com/in/austin-kingsley

To follow future updates, this project can be found at:

- https://github.com/akingsley319/MLB_xBA_Prediction