Max Opper

5/7/2023

CS480

OpperStats Final Report

OpperStats (Open-source Player Projection Estimation Report and Statistics System) was created to provide the user with accurate player projections for the 2023 season. There are many projection systems that already exist, but most of them are not open-source, meaning that the inner mechanisms of the systems are secret for the user. This can make choosing a projection system difficult for fans and analysts alike, as it is up to them to determine the strengths and weaknesses of each model individually. By building a model from the ground up with easy to obtain Statcast data, the user can easily see how and why a player is projected for specific stats. These projections are useful for comparing a player's future offensive or pitching value, which is nice for fantasy baseball purposes, sports betting, or simply seeing if a player's hot start is legitimate.

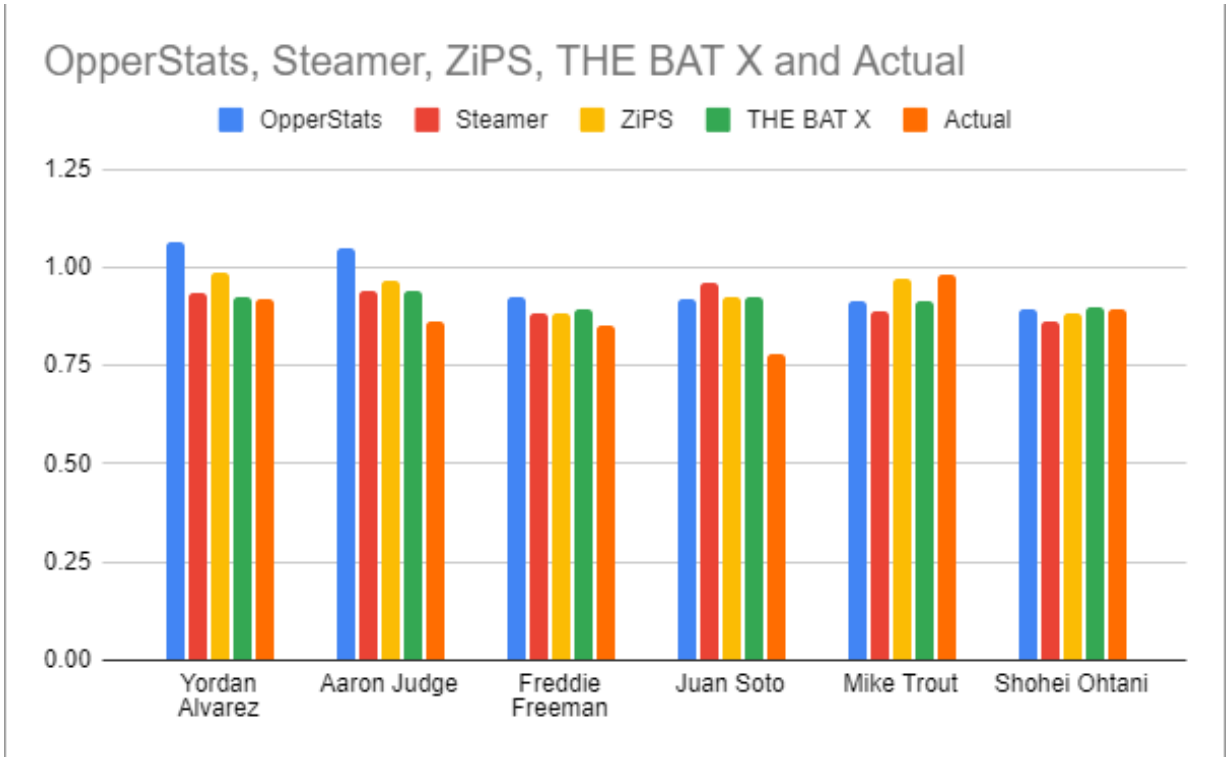
OpperStats makes its predictions by utilizing the random forest regression algorithm from the sklearn library in python. Originally a decision tree regression algorithm was going to be utilized for the model, but that method proved to be inaccurate and prone to overfitting. The random forest regression model was much more accurate, especially once the ideal parameters were specified, that being a max depth of 10 nodes and the number of estimators being 100. The model for batters and pitchers uses a plethora of expected stats, which are generated using automated tools at each ballpark (for a full list of stats used, see glossary). This data is oftentimes more accurate and practical for predicting future value, as they are observation based values compared to results based values. An example of this would be xBA, batting average is

traditionally calculated by taking the number of hits and dividing by the number of at bats, while xBA looks at all balls put in play by the batter and compares them to similar batted ball events. This is more accurate as batting average doesn't account for things such as BABIP luck or stellar defensive plays. Cross validation was also utilized to check for overfitting in the model, which is when a model gives accurate predictions for training data but not for new information. For cross validation, the training data used was Statcast data from the 2018-2021 seasons, and for the validation data the 2022 season was used. Mean square error was used for the cross validation score, the MSE for the training and validation data is shown below.

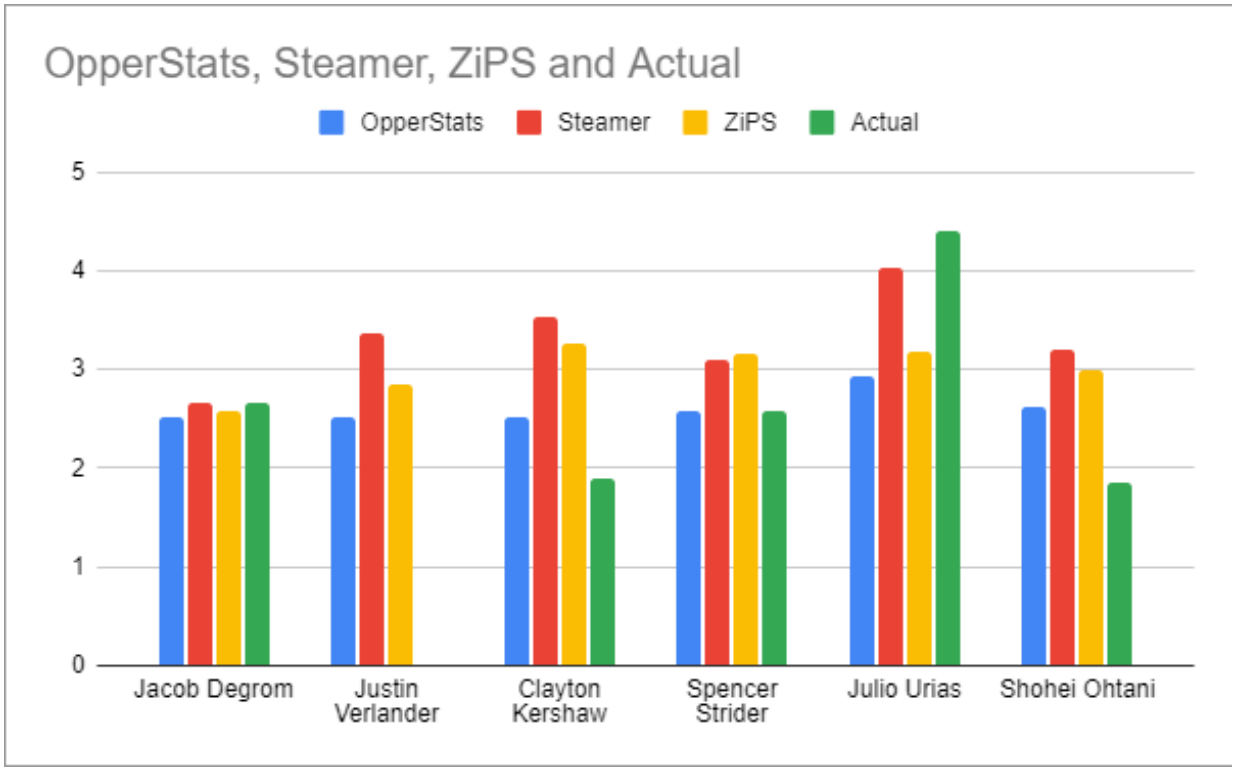
	BA	OBP	SLG	ERA	BAA	wOBA
Training	0.0005	0.0004	0.002	1.1625	0.0006	0.0007
Validation	0.00048	0.00039	0.0018	0.9424	0.00052	0.00058
Difference	4.1%	2.5%	10.5%	20.9%	14.2%	18.75%

Predictions made for batters were much more accurate when compared to the pitching counterparts, which could be due to ERA being a somewhat unreliable predictor of a pitcher's ability as it doesn't take into account park factors or the defensive abilities of his teammates. This is further illustrated when comparing projections from different systems. Looking at batting statistics, OpperStats is fairly comparable to other popular projection systems, at least among the top predicted performers by OPS (OBP + SLG). This differs heavily from the pitching projections. Various outlets have vastly different projections for pitcher ERA, which could be attributed to the aforementioned issues with using ERA for pitcher value.

OPS projections for top 6 qualified batters by OpperStats:



ERA projections for 6 qualified starters by OpperStats:



In conclusion, OpperStats has successfully achieved its goal of providing an open-source player projection system for the 2023 baseball season. By utilizing the random forest regression algorithm, the model is able to provide accurate predictions for both batters and pitchers based on a wide range of expected stats. The model is also designed to avoid overfitting and was tested using cross validation, which showed that the model's predictions were accurate not only for the training data but also for new information. Although pitcher projections were less accurate than batter projections, this can be attributed to the limitations of using ERA as a predictor for pitcher value. Overall, OpperStats provides a valuable resource for fans, analysts, and anyone interested in projecting player performance for various purposes, such as fantasy baseball or sports betting.

Glossary:

Batter Predictors:

xBA:

xSLG

xwOBA

xOBP

xISO

xwOBACON

xBACON

Average Exit Velocity

Average Launch Angle

Barrel Rate

Hard Hit %

Pitcher Predictors:

xBA

xSLG

xwOBA

xOBP

xISO

Average Exit Velocity

Average Launch Angle

Sweet Spot %

Barrel Rate

Hard Hit %

GB%, LD%, FB%