Jack Page, Will Kennedy, Max Wassarman

Abstract 0.1

Fans, Managers, and fantasy baseball players are all interested in one thing. Which players are going to perform well above their previous contributions and ‘breakout’ in the MLB? In this paper we propose that a correctly configured LASSO regression model can be used to predict players that are projected to ‘breakout’ in the next season. Specifically, which players show significant increases in projected WAR over their previous years performance. We accomplished this task through a combination of factors including tuning our hyperparameters, formatting the aggregate data, and validating off of real statistics from years prior. This process produces breakout candidates that are projected to perform, in the default configuration, a minimum of 1.25 better in WAR than their previous available season. Through this process we hope to better understand which statistics are the most predictive of increases in WAR and more generally, which players are going to breakout this season in 2024.

Introduction 1

Fans have debated over the most predictive offensive statistics in baseball since the inception of recorded statistics. While models do exist, we set out to create a model that specifically predicts players that are intended to ‘breakout’ in a given year. Breakout here means to perform significantly better than previously recorded years in a non-linear fashion. Players that see huge increases in their wins about replacement (WAR), are the base candidates for breaking out. WAR is an aggregate statistic that is calculated to display how many wins that player will add over a ‘replacement level’ player or a player that is likely to be demoted. A player with a WAR of 6 is calculated to add 6 wins for his team over the course of a season compared to if they had played an average player working at the league minimum salary [1]. With this in mind, we set out to predict changes in WAR from year to year.

The biggest hurdle we face at this stage is deciding how to best predict changes in WAR. We settled on an approach that takes one year of total offensive statistics and fits a LASSO linear regression to predict their WAR for next season. With this, we can calculate the increase from the previous years across all instances and then filter by selection criteria to omit results from players that were linearly increasing. The breakout selection criteria that are mutable from the user are minimum age, maximum current WAR, and the WAR increase from the previous year. This format reliably produces a single digit list of players and their projected WAR increase from the previous year. We then validated this data over past seasons for which we have recorded WARs.

This process produced 6 players that are eligible for breaking out in the 2024 season that are above age 27, with a current WAR below 3, that are projected to increase by at least 1.25 in WAR based on our model. This gives us a list of players that we can expect to outperform their previous contributions to the tune of more than 1 win per season over baseline. In addition to this intended output, because we are calculating projected WAR, we can also see in our model which players are expected to regress compared to their last season numbers at breakout level swings in the negative direction. This gives us another avenue by which to validate our predictions and gives us reliable insight into the kinds of statistics that influence

changes in WAR by examining the weights of our model. Overall, several of our predicted breakout players are currently outperforming their last season statistics in convincing fashion.

Background and Related Work 1.1

In previous research, we've seen some similar studies and attempts to solve this problem. This includes Jeremy Siegel, who wrote [2], who discussed how to quantify a breakout in the first place and wrote a model to predict the probability of a breakout using a tree-based model and different weights of importance on various stats. A notable part of this model is that it exclusively looks at data from the previous year, not taking into account the possibility that the player may have simply just had a down year the year before. The most important statistic used in [2] is xwOBAcon, a stat that stands for Expected Weighted On-Base Average on Contact, which tracks how a batter performs purely when they put the ball into play. This stat combines many different metrics into one value. These different metrics "likely don't offer much of an effect isolated from each other, but the combination of all of them is very important".

Another relevant piece of literature that we found tried to predict undervalued players by looking at pitching metrics [3]. It used K-means clustering to group pitchers together then looked for undervalued pitchers who were clustered with more established arms. To identify these pitchers they computed the mean ERA of the cluster then subtracted it from the individual pitchers ERA's and looked for pitchers with the greatest difference. Along with the K-means approach this paper also used hierarchical clustering and compared the two to find which was better at accomplishing the task. They ended up finding that if you wanted to find the most valuable or most improved players the hierarchical algorithm was better but if you wanted to find the most valuable on a per inning basis the K-means was better.

The third research paper we reviewed was the most applicable to our project [4]. It tried to predict OPS+, a metric that encapsulates a hitter's pure hitting performance. In order to do so, multiple approaches were taken including unregularized and regularized regression, and support vector regression. The paper found that SVR performed well on players whose OPS+ did not vary significantly year to year but struggled with players who peaked early or late in their careers. According to the paper "the most important aspect of SVR that is lacking in linear regression is the ability to produce a non-linear prediction curve, which is the result of mapping the data into a higher dimension. Since a player's career path is not linear, the ability to produce such non-linear prediction curves is imperative" (4).

Problem 2

Every season, the problem consistently facing coaches, general managers, and fans is: "which new players will come from relative obscurity to help the team win?" Every year, players come out of the blue to take the league by storm, and seemingly no one sees it coming. But what if that could be changed? What if we could predict with decent accuracy players who could erupt to stardom before they do? This could allow general managers to give their coaches the best teams possible, the coaches

to put the team into the best position to succeed, and fans to place the best bets and assemble the best fantasy lineups known to humankind.

As a group of three baseball players, we know that baseball seems, at times, to be a very unpredictable sport. Still, as has been well documented over the years, there is often more data of statistical interest than any other sport. Starting with the sabermetrics revolution begun by Bill James in the late 1970s and advanced throughout the years, baseball statistics have gone from simply batting average or "does he hit the ball?" to advanced equations calculating just how much better this player is than the average replacement player. This additional information allows many more people to identify not just how good a player is but which players are better than others. Though this advance in statistical prowess in the game allows easier access to more people, it has also made it much more difficult to get ahead in any facet of the game. With all of these stats available, whatever someone sees in a player, everyone else can see as well, at least on the surface. This leads to the problem: how do you get a leg up on other interested parties when it comes to predicting a player that others might be undervaluing or overvaluing? This is the issue we have tackled using machine learning. Since players break out from unassuming stats to one of the better players in the league, while others go from some of the best players in the league to simply average, or slightly better. We looked to see if we could find a pattern in their impressive stats. A way to tell that a breakout or regression is near.

Solution 3

In order to accomplish our task, we have been dealing with data from fangraphs, which contains data on past seasons and some projections. We will be primarily concerned with WAR as it is the most encompassing statistic. It combines many others into one number, which can be used to evaluate how a player performs.

WAR stands for Wins Above Replacement, and the calculation predicts how many wins a player generates for his team compared to a replacement-level player (Someone in the minor leagues or free agency) at his position. This is obviously a very significant statistic since it allows a team to see how much a player helps their team to win.

We use this data to train our model, in which we sort and adjust the data to add the value of the next year's WAR to the previous season's stats, giving us a label to predict in a similar way to the regression models from homework 5. After this, we had to work with the data for a bit, since the value of this next_WAR variable was NaN for every player that either retired or wasn't in the league the next year, or, even more simply, every player's stats from last year. We dealt with this by removing each of these past players from the dataset, and then we split past years versus current for our training and test datasets. This obviously being a real-world dataset, we don't know the correct amount of WAR for this upcoming season, so we can't calculate MAE or any other method of finding the quality of these predictions. However, we can compare our top couple of breakout and regression candidates with how they are performing so far this season, and how they are projected to do for the rest of the season.

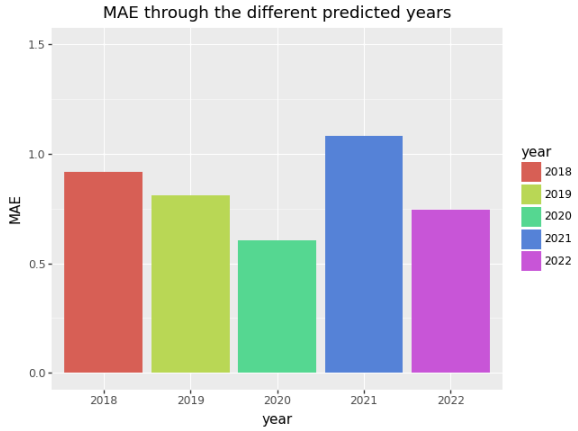
For our model, we used a lasso model, but as we have learned for lasso

models, we need to find an optimal alpha for the model. To find the optimal value for our alpha we used cross validation. Cross validation trained multiple models then evaluated them on a subset of our data. It looked for the model with the least error across folds. This allowed us to determine the optimal alpha value to use in our lasso regression model.

Experimental Setup 3.1

For our experiment, we trained the lasso model that we had previously found the optimal alpha value for on all of the batting data of players with at least three hundred plate appearances in a season from 2017 to 2022, with the model's goal being to predict their WAR in the next season. This model then outputs our predicted WAR values for each player who qualified by our metrics. Finally, to find our breakout and regression candidates, we calculate the difference between our predicted WAR for the following year and the WAR that each player had the year before.

We confirmed that this method gave us a decent prediction method by testing it on the years before 2023 to see that the mean absolute error is not particularly large for each year, with almost all years having a MAE below 1.0. This allowed us to be confident that our model will give us at least a decent prediction of each player's production for the upcoming year.



Results 4

For example, our model predicts a significant regression for Freddie Freeman of the Los Angeles Dodgers, predicting that his WAR will drop from almost 8 in 2023 to only 4 in 2024. We are additionally encouraged by this result from our model because if we check popular projections for Freeman, his projected WAR varies between the values 3.7 and 4.2, which backs our model's results.

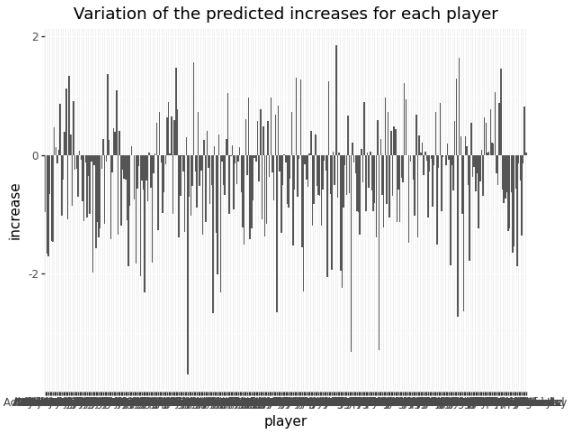
As for the breakout candidates our model predicts, two great examples of our model immediately jump off the page: Shea Langeliers and Bryan De La Cruz. For Langeliers, he was a slightly above-average player in 2023 with a WAR of 0.7. Our model has predicted a significant breakout for Langeliers, with a predicted WAR of 2 this year. When we compare this to how he is playing this season, Langeliers already has a WAR of 1.1 and is projected for a season WAR somewhere along the lines of 1.5-1.8 WAR for the whole season.

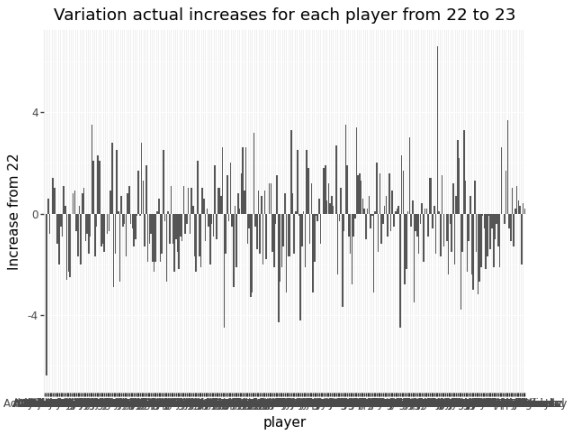
	Season	Name	Age	G	AB	PA	H	1B	2B	3B	...	Barrels	Barrel%	maxEV	HardHit	HardHit%	Events	CStr%	CSW%	Prediction	Increase
2403	2023	Keibert Ruiz	24	136	523	562	136	94	24	0	...	27	0.058	110.1	148	0.317	467	0.163	0.229	0.674955	1.274955
1585	2023	Shea Langeliers	25	135	448	490	92	47	19	4	...	41	0.133	113.4	136	0.442	308	0.146	0.303	1.988846	1.288846
2197	2023	Bryan De La Cruz	26	153	579	626	149	98	32	0	...	39	0.088	111.0	188	0.424	443	0.158	0.292	1.165109	1.365109
1169	2023	Vladimir Guerrero Jr.	24	156	602	682	159	103	30	0	...	56	0.111	116.7	249	0.492	506	0.128	0.236	2.755076	1.455076
1178	2023	Spencer Torkelson	23	159	606	684	141	75	34	1	...	62	0.141	112.7	222	0.505	440	0.166	0.269	2.939004	1.639004
1870	2023	MJ Melendez	24	148	533	602	125	75	29	5	...	42	0.114	113.2	182	0.496	367	0.157	0.306	2.145168	1.845168

As we can see below, the overall consistency of our predictions is good, especially when compared to the previous year's actual increases in WAR. The spikes regarding

player growth or regression are pretty similar with a couple of outliers. If we were to remove Aaron Judge (the large negative spike) and Ronald Acuña (the large positive spike) from the previous season, our predictions would look even better.

Regarding Judge and Acuña, both of their changes in WAR are incredibly large. For example, Judge's decrease of 6.4 WAR and Acuña's increase of 6.6, are about the same as the difference between Ivan Rodriguez's MVP season in 1999 and a replacement-level season, something a minor leaguer would produce. These changes obviously don't take place often at all, so it is not surprising to see the model not allow for any of these.





Conclusions and Future Work 5

In this study, we addressed the challenge of predicting MLB players who are likely to have breakout seasons. Specifically, we identified players who our model expects to have significant increases in their Wins Above Replacement (WAR) when compared to their previous season. To do so we developed a LASSO regression model, which we validated against historical data, to identify players projected to see a substantial increase in WAR. Our model effectively filtered players based on criteria such as age, current WAR and projected WAR increase. This left us with a short list of breakout candidates. Notably, our predictions for players like Shea Langeliers and Bryan De La Cruz, who are already producing well above their previous seasons numbers, validated our models findings. Along with the projection of breakout players, our model also appears to have the ability to accurately predict regression from the previous season as well.

For future work, there are several different directions that can be taken. First could be the implementation of statcast data into the model. We initially had tried to accomplish this but found that it was too complex for the scope of the project due to the time constraints we had. Second would be experimenting with different machine learning models such as random forests or neural networks which might be able to find relationships that our LASSO regression could not. A third possibility could be modeling over multiple seasons in order to predict a players total future value not just one season.

Sources

[1]Sabermetrics 101: Understanding the Calculation of WAR
https://www.samford.edu/sports-analytics/fans/2023/Sabermetrics-101-Understanding-the-Calculation-of-WAR

[2] https://pitcherlist.com/how-to-quantify-breakout-hitters/

[3]https://cs229.stanford.edu/proj2016/report/Ishii-UsingMachineLearningAlgorithmsToIdentifyUndervaluedBaseballPlayers-report.pdf

[4] https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26585755.pdf