Next In Line: Forecasting MiLB Prospects' Path to the MLB

A Thesis Submitted in Partial Fulfillment of the
Requirements of the Renée Crown University Honors Program at
Syracuse University

Brett Cerenzio

Candidate for Bachelor of Science in Sport Analytics (B.S.)
and Renée Crown University Honors
April 15th, 2025

Honors Thesis Faculty Advisor: Dr. Rodney Paul

Honors Thesis Reader: Dr. Jeremy Losak

Abstract

The one question that plagues many MLB front offices is determining the appropriate time to call up their prospects. Calling them up too early can hurt their future confidence but accelerating them through the minor league system will allow them to make a longer impact at the MLB level. Prospect scouting today, at least in the public sphere, still relies heavily on the "eye test," especially compared to how stats have revolutionized the rest of the baseball industry. Is there a way we can project a prospect's future performance at the next level of the minor leagues simply by looking at their current level stats?

In this paper, I employ several modeling techniques to answer this question and to bring a more "objective" view to minor league scouting. Utilizing prospect data from Fangraphs and TheBaseballCube, I utilized GAM, BART, & XGBoost models to create aging curves for major league promotion likelihood, predict performance at the next level of the minor leagues, and predict the probability of a given prospect to make the MLB. I then created a Shiny App that looks at prospects for the 2025 season and predicts their stat line at the next level and the likelihood they make the MLB. This will allow teams to blend the objective & subjective nature of scouting when making decisions about their potential future players.

Executive Summary

This project investigates how we can utilize prospects' minor league stat lines to project future performance at higher levels of professional baseball. Minor league scouting today has remained relatively subjective despite baseball leading one of the more aggressive movements towards advanced analytical thinking in the sports field. Analytical departments among baseball teams have been constantly increasing ever since the publication of *Moneyball,* and as of recently many other sports industries are following suit after seeing its success in baseball. Despite this, many prospect rankings today are often done exclusively through video analysis, which is inherently very subjective. This is fine in a vacuum, but perhaps there is more we can do to add an objective component to evaluate these prospects, especially as more data/stats come in for the minor leagues. Using level-by-level data from the 2014-2024 seasons from Fangraphs, I sought to evaluate whether we can predict how a player will perform at the next level based on their current stat line as well as predict the probability they make it to the MLB.

During my initial exploration of the data, I identified levels of the minor league system that can prove to have the hardest transition for batters & pitchers, which can give fans the best insight into whether a player has what it takes to make it to the MLB. For batters, the hardest transition within the minor league system was the jump from A+ to AA, as batters on average see a moderate dip in batting average and a slight increase in K% during this transition. However, unsurprisingly, this doesn't compare to the transition from AAA to the MLB, where batters see a drastic decrease in performance on average. For pitchers, the hardest transition seems to be the move from AA to AAA, where pitchers see a drastic increase in metrics like ERA & FIP. This could be because AAA contains a lot of former MLB players as well as each team's top prospects, which makes it the first level where pitchers can't pitch around guys as effectively as they are too disciplined & there are much fewer holes in a lineup at this level. This means that pitchers have to put an extra emphasis on executing every pitch at this level or else they will be hit hard. I also created aging curves for batters, starting pitchers, and relief pitchers that outline how their likelihood of making the MLB decreases as they age based on their current level and controlling for their performance at a given level. For batters, staying an extra year at the lower levels (levels A & A+) sharply decreases your chances of making the big leagues while at the higher levels, performance means much more relative to a player's age. The same can be said for a pitchers' chances to make the MLB. However, this effect is less drastic, as pitchers seem to be given a longer grace period to "find their form."

To predict stat lines at the next level, I trained BART & XGBoost models, which I chose not only for their predictive capabilities but also because they come to their predictions in different ways. BART seemed to perform best at predicting stats at the next level largely due to being more equitable in its use of the covariates. For batters, I predicted AVG, OBP, ISO, K%, BB%, and SwStr% at the next level, while for pitchers I predicted ERA, FIP, K%, BB%, and SwStr%. I then created an XGBoost model that predicts the likelihood of a player to make the MLB based on age, level, prospect ranking, and performance. This was all published on a Shiny App for the 2025 season, where I give my predictions on how players will perform if/when they get called up to the next level of the minor leagues. This paper aims to apply a more "objective" view to the scouting process, that when combined with traditional scouting, can help front offices come to more accurate conclusions about their future players.

Table of Contents

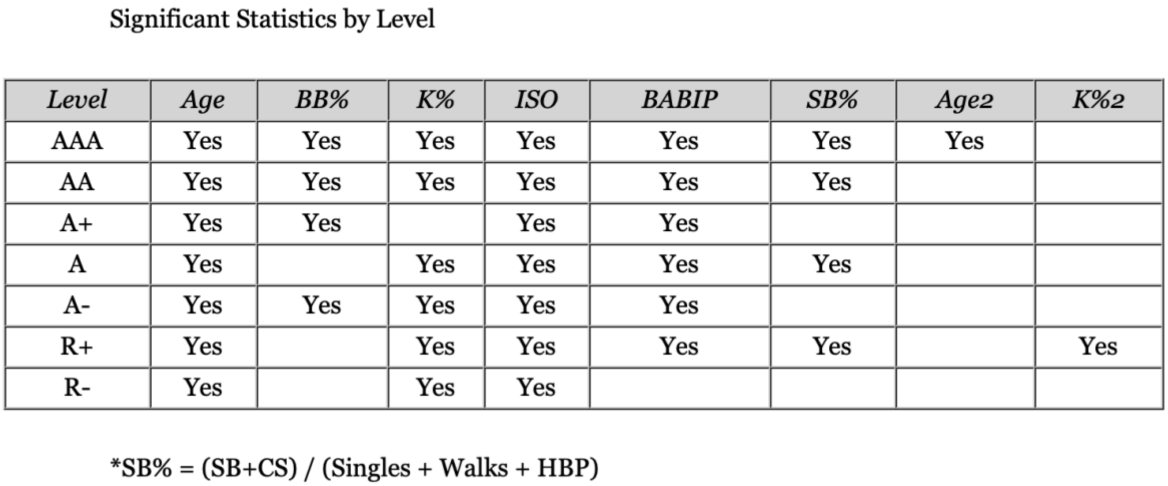
I. Introduction

Throughout the past 10-20 years, the sports world has seen a rapid acceleration into the world of analytics, and baseball has been at the forefront. Since *Moneyball* was published in 2003, baseball teams have been constantly increasing their analytics departments in hopes of getting an edge against the other 29 teams in the MLB. Whereas in the past when many hypotheses were put up to the “eye”/”feel” test, every front office decision today needs to be backed with numerous advanced metrics that are put forth to meticulously evaluate what kind of value a player can add to a team. Despite this advancement coming in droves at the MLB level, this approach has not been heavily applied when evaluating MLB prospects’ development throughout the minor leagues. Minor league scouting today is still very video-focused, and public rankings are determined on how a scout interprets a player’s hit tool rather than their stat line at the minor league level. This is alright in a vacuum, but as more minor league statistics come to light, shouldn’t we look to bridge the gap between “objective” & “subjective” just like front offices have in other facets of the game? In this paper, I will attempt to bridge this gap by using player stats at the previous level to then predict their stats at the next level as well as predict their likelihood of making it to the MLB based on their current situation. This will provide teams with some extra information based on past outcomes which they can use in conjunction with conventional wisdom provided by their scouts.

II. Literature Review

Current Projection Systems

When I looked into MiLB projection systems, I could not find much research that had been done on the subject. However, I did find one that was posted on The Hardball Times: KATOH. Chris Mitchell created the KATOH projection system in 2014 after Yankees prospect Gosuke Katoh, who ran a strikeout rate close to 40% in the 2014 season (Mitchell, 2014). However, Katoh was 2-3 years younger than most of his competition, which made Chris realize he had no clue what was truly important when evaluating a minor league player's performance and how that might impact their chances of making the MLB. Utilizing minor league data dating back to 1990, he found that these variables were statistically significant:



Significant Statistics by Level

Level	*Age*	*BB%*	*K%*	*ISO*	*BABIP*	*SB%*	*Age2*	*K%2*
AAA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
AA	Yes	Yes	Yes	Yes	Yes	Yes		
A+	Yes	Yes		Yes	Yes			
A	Yes		Yes	Yes	Yes	Yes		
A-	Yes	Yes	Yes	Yes	Yes			
R+	Yes		Yes	Yes	Yes	Yes		Yes
R-	Yes		Yes	Yes				

***SB% = (SB+CS) / (Singles + Walks + HBP)**

Table 1: Displays which variables were significant in determining who would make MLB at each level.

From this table, you can see that the most important factors in determining which players will make the MLB were Age, K%, ISO, and BABIP. He also mentions that BB% wasn't that impactful at the lower levels because that's where pitchers struggle with command the most (Mitchell, 2014).

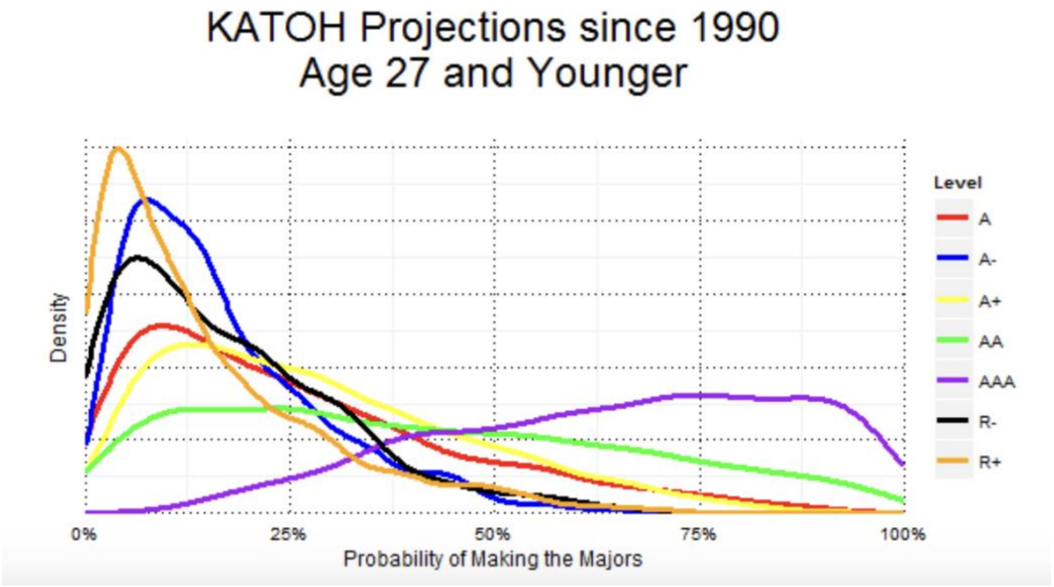


Figure 1: Density plots showing the likelihood of a player making the MLB at each level.

The above graph demonstrates the density plots of the probabilities a player makes the MLB depending on which level they are at. Unsurprisingly, the density plots are extremely skewed right at the lower levels but begin to become less skewed as the level increases.

The next part of my research led me to investigate some of the methodologies behind MLB projection systems. Piper Slowinski (Slowinski, 2010) published an article in 2010 where she explained why projection systems are important and how they're used in today's game to forecast future performance. At the end of the article, she listed a couple of projection systems that I decided to do more research on. This led me to three new articles: “Statistical Projections and You, Part I” (Druschel, 2016), “A Guide to the Projection Systems” (Druschel, 2016), and “Fangraphs Prep: Build and Test Your Own Projection System” (Cross & Mailhot, 2020). The first two articles gave summaries of how the popular projection systems today are calculated as well as how they are similar and different to one another. These projection systems included Marcel, PECOTA, Steamer, and ZIPS, which are the main projection systems used on Fangraphs and Baseball Prospectus. The final article walked me through how I could make my own projection system. While this approach is much more basic than the one I plan to implement in my model, it’s important to remember the basics before moving on to more complicated models.

Machine Learning

While I know of numerous machine learning methods such as Random Forest, XGBoost, K-Means, and PCA, there are also numerous models that I’m not familiar with that could help with building my project. One of the modeling techniques I was particularly interested in was Bayesian Additive Regression Trees (BART). Hugh Chipman, Edward George, & Robert McCulloch published a paper in 2010 to explain how BART trees are built and how they’re similar and different from other modeling techniques(Chipman et. al, 2010). BART trees are fitted utilizing an iterative Bayesian backfitting Monte Carlo Markov Chain (MCMC) algorithm to generate samples from a posterior distribution. One of the biggest things that set BART models apart from other Machine Learning methods is that BART models utilize heavy regularization techniques, making each tree a “weak learner.” Thus, it is much less prone to overfitting than other machine learning models.

I also took a deeper dive into neural networks to see how these can be used to find insights that other machine learning methods might miss. In a research paper titled “Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network,” authors William Young II, William Holland, and Gary Weckman did a deep-dive on how neural networks are developed (Young et. al 2008). This paper allowed me to learn about the structure of a neural network as well as the different activation functions and hyperparameters that need to be tuned before deploying a model. I supplemented this research paper with 3 other articles that gave a description of different types of neural networks (Coursera Staff, 2024), different activation functions (Baheti, 2021), and a real life example that allowed me to create my own neural network in Python using the Keras package (Brownlee, 2022). I also looked into how an NLP model was built since it was connected to a project where the authors were able to predict the outcomes of very uncertain events with very limited data (Heaton & Mitra, 2021). However, the scope of the project didn't make a ton of sense with my goals. Another article I read talked about utilizing an LSTM-CNN hybrid model for text classification, which I found particularly interesting because of another article I read earlier in my research which I will discuss in a later section (Zhan et al, 2018).

Recently, a new clustering format has been becoming more popular in the data science main frame besides K-Means: a method called DBSCAN. DBSCAN is especially useful for finding shapes in the data, and creates clusters based on areas that see high density that are separated by low density points in the multi-dimensional space (Deng, 2020). I then read another research paper that showed how DBSCAN can be used in R and how the different hyperparameters can affect the output of the model. I also learned about another package that rivals DBSCAN, OPTIC, that rivals DBSCAN in the sense that it allows for clusters with varying degrees of density, something that DBSCAN struggles with (Hahsler et al, 2019).

Finally, I wanted to do a deeper dive into Naive Bayes modeling, which is another modeling technique that has taken over the sports landscape in recent years, yet is a modeling technique I was not too familiar with. The first article I read on Naive Bayes first talked about the math and the assumptions that went into creating the model while showing how you can use Naive Bayes in R utilizing the e1071 package (Zhang, 2016). I then looked at papers that sought to compare Naive Bayes models to other models, including Random Forest, SVMs

(Pranckevičius & Marcinkevičius, 2017), and XGBoost (Hendrawan et. al, 2022). While they found that Naive Bayes performed similarly to the rest of the modeling techniques with balanced data, XGBoost outperformed the model with unbalanced data.

Aging Curves

One thing that I found in common with all projection models was that they had some way to account for age after making their initial projections. Obviously, as a player ages, they are going to become less athletic which will inhibit their ability to perform at the highest level the next season. Therefore, I found it necessary to at least take a brief look at the current research regarding aging curves in the MLB, as I couldn't find anything for the MiLB. In the first paper I read, Kevin Ng looked at players after their 7th season in the majors (the player's first year of free agency) and looked at age, years pro, and WAR to predict how they will perform in the future (Ng, 2017). He found that although players reach their athletic peak at around 24-26, players will continue to develop because of added experience in the MLB until their age 29 season on average (Ng, 2017). This is when their diminishing athletic ability seems to catch up with them. He also found that players improve at a much greater rate due to age than they decline after their age 29 season.

The second article I read utilized the Lahman database to evaluate player aging curves using OPS as their response variable (Nguyen & Matthews, 2024). They hypothesized that current MLB aging curves overestimated the peak age for baseball players because only really good, established players make it on MLB teams in their 30s. He utilized multiple imputation to fill in the expected peaks for players who performed poorly in their first few years in the league so that they could get a better idea of the true peak age. They found that this age was around 26 years old rather than 29. They went on to present their findings at CMSAC.

MiLB to MLB Projections/Research

While I could not find any research that attempted to predict how players would perform at the next level of the minor league system, there was plenty of past research that attempted to answer which players were more likely to make the MLB. In one paper, Gabriel Chandler and Guy Stevens built a Random Forest model to determine the probability that a player had a Major League impact, which they described as playing at least 320 games (Chandler & Stevens, 2012). They evaluated this topic in 3 different layers: what stats at the Minor League levels are more correlated with MLB Success, how draft status impacts a prospect's progression through the minors, and factors each franchise uses in evaluating its minor league players. They found that, unsurprisingly, stats at the higher levels were more predictive of MLB success than those at the lower levels. However, he also found that higher draft picks were much more likely to be promoted because of their name-value rather than because of their pure stats.

Another paper I read looked to predict the probability a player would make the MLB debut rather than simply being successful in the MLB. Chung-Hao Lee & Woei-jyh Lee utilized data containing players drafted between 2000-2010 to determine which stats were the most important in predicting whether a player would make the MLB or not. He utilized 4 different machine learning algorithms to answer this question: XGBoost, Random Forest, Decision Trees, and Support Vector Machines (SVMs) (Lee & Lee, 2023). XGBoost was the best at predicting who would make the MLB, with metrics like Draft Pick, AVG, OPS, and AB being the most impactful variables. Their research also supports the conclusion drawn upon by Chandler and

Stevens that players drafted earlier are generally promoted quicker as they are more highly touted by the organization.

Alexander Gow used a different approach to answer how MiLB performance impacted MLB performance. Unlike the previous examples, Gow utilized Neural Networks to try to answer this question and compared his results to linear models and XGBoost models that he created. He found that Neural Networks were significantly more accurate than the XGBoost model, but I'm skeptical of his analysis. First, he doesn't mention if he tuned the hyperparameters of the XGBoost model, which could drastically change the accuracy of an XGBoost model. But more importantly, his analysis does not include age as a factor, which I consider to be a big blindspot in his analysis. Another article that utilizes Neural Networks was "Trouble with the Curve: Predicting Future MLB Players Using Scouting Reports" by Jacob Danovitch. However, in contrast with the previous analyses, Danovitch's model looked exclusively at scouting reports rather than minor league stats (Danovitch, 2019). That is, he used a CNN neural network that read the text of each scouting report from numerous websites including Fangraphs and MLB. Overall, the model had a 73.2% accuracy rate, which is pretty good considering it's only going off inherently subjective measures of talent.

The final article I looked at analyzed which teams were the best at developing talent at the minor league levels. Patrick Brennan utilized season-level minor league data at each level for every season since 2007, using wRC+ for hitters and FIP- for pitchers (Brennan, 2021). To evaluate over different levels, he created MLB equivalences in each of these statistics based on level, such that having a 130 wRC+ at Triple-A would be weighted as being better than having a 130 wRC+ at Single-A, for example. He also created aging curves for each, where players increasingly improved in terms of wRC+ from ages 17-24 before seeing diminishing returns. He found that the 5 best teams for developing talent were the Royals, Yankees, Blue Jays, Dodgers, and Astros.

Other Projection Systems

Since there weren't many papers that projected performance after changing leagues in baseball, I decided to take a look at other sports to see if there were any papers that could have some relevance to my own research. The first sport I looked at was the NBA, where Alex Hussey analyzed the probability each NBA draft pick would a) be an all-star, b) the best player in the draft, c) win an MVP, and d) be a bust (Hussey, 2021). He looked at each draft pick's NBA stats including points, assists, rebounds, and net rating to come to his conclusions. For the number one pick, he found that there was a 75% chance they would become an all-star, a 30% chance at being the best player in the draft, a 20% chance at becoming an MVP, and a 20% chance of being a bust. One thing to note is that he does not use college stats in his analysis, which prohibits me from using this research as a way to project into the future.

I then chose to look at soccer, which I thought would be a great avenue to study league-to-league variance considering Europe has 4 top leagues that are all deemed to be the best in the world with their own play styles. However, I was surprised to see that most of the research didn't involve player performance from league to league, but rather financial fair play regulations. Regardless, I was able to find this article published on DataBuckets in 2015 which evaluated how interleague transfers between the 4 top leagues (Premier League, LaLiga, Serie A, & Bundesliga) impacted minutes per 90 and goals per 90 on their new club (DataBuckets, 2015). They found that, unsurprisingly, players involved in interleague transfers varied much more widely in their performance with their new team compared to players transferred domestically, giving credibility to the notion that each league has their own unique style of play. They also found that players typically earned more minutes after their transfer except for when they were transferred to the Premier League, supporting the notion that the Premier League is the most competitive league in Europe.

In the NFL sphere, Michael Gallagher aimed to answer the question of which mattered more when predicting NFL performance: their combine stats or their college stats/performance (Gallagher, 2019). Utilizing draft data from 2007-2012 for first-round draft picks, he found that the answer depended on position group. For CBs, the 40 yard dash and the vertical jump were highly correlated to NFL success, while for DT college stats were much more significant in predicting future NFL success. One con with this study is that it only used the pearson coefficient in its analysis, and didn’t mention statistical significance once in the entire paper.

Finally, I took it back to baseball, but instead looked into how players from the Nippon League tend to transfer to the MLB. Similar to soccer, I found it hard to find anything pertaining to player projections, as many focused on the financial aspect of the posting system. However, I finally stumbled across an article written by CJ Lu Sing, who compared OPS and ERA across the two leagues (Sing, 2022). Since there have only been a few players who have made the transition from the Nippon League to the MLB, the datasets were fairly small. However, he did find that, as expected, players saw a minor decrease in OPS after transitioning to the MLB and a minor increase in ERA after transitioning to the MLB.

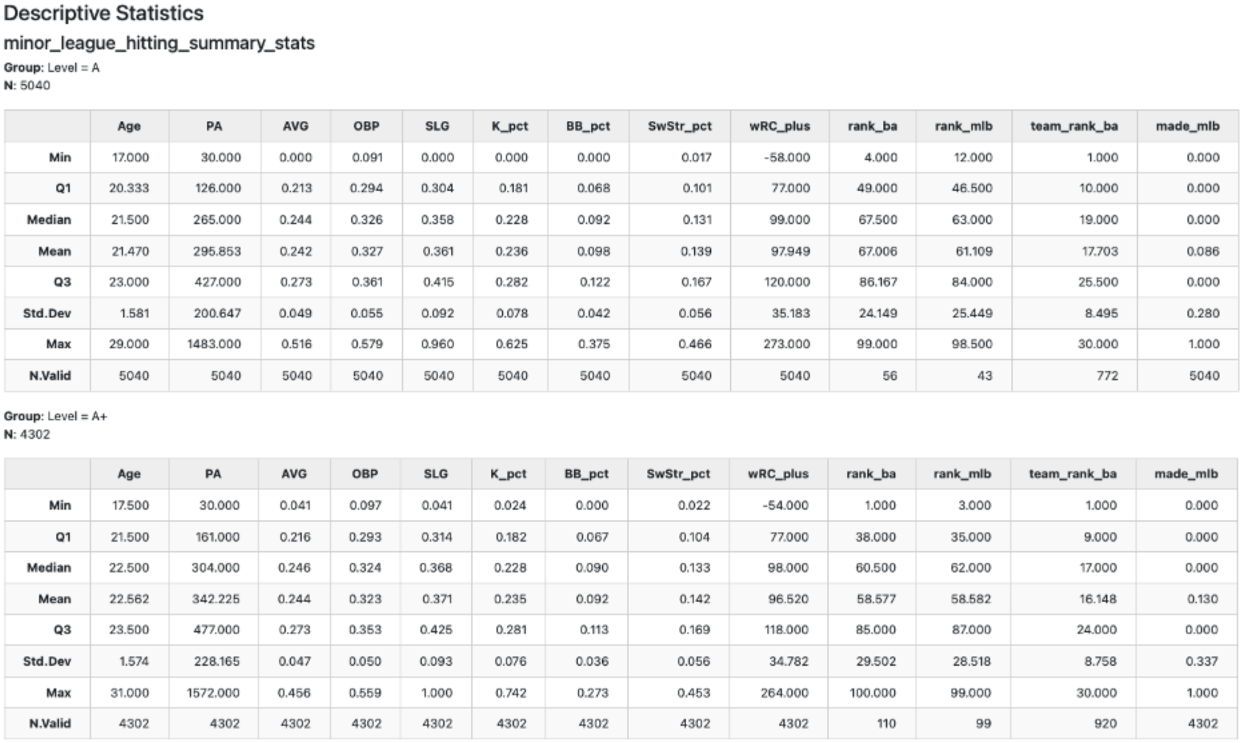
III. Data

I collected yearly data from Fangraphs for the 2014-2024 seasons for each of the top 4 levels in affiliated minor league baseball: A, A+, AA, & AAA as well as the MLB rookie data from the 2014-2024 seasons. Hitters needed to have a minimum of 30 PA in a given season at the level to be included in the dataset while pitchers needed to pitch a minimum of 10 innings in a given season at the minor league level to be included. To be included at the MLB level, a batter needs at least 100 PA, a SP needs at least 50 IP as a starter, and a RP needs at least 10 IP as a reliever. This data contained baseline statistics such as AVG & OBP for batters & ERA Pitchers as well as advanced stats like wRC+ (batters), K%, BB%, & FIP (pitchers). For each player, I took the total or weighted average (depending on the statistic) of their stats at that level. Since I was focused on a continuous path to the MLB, I removed any minor league seasons where the last year they played at that level was more than 2 years ago. For example, if a player was demoted to AA in 2017, and had not been in AA since 2014, that player's 2017 stat line would not be considered when evaluating that player's performance in AA and how that might translate to AAA. This also helped remove any observations where players were sent on rehab assignments from their major league team. For player ages, I took the average age of the player while he was at that level. If a player was called up to AA when he was 25 and played there until he was called up at the age of 26, his age in the data would be 25.5. Finally, I removed any players who played in the MLB from 2004-2013 to remove any players who had already had significant playing time in the MLB but were sent back down to the minors (following the same minimum qualifications as stated above). Finally, for pitchers, I differentiated whether a pitcher got promoted to the MLB as a RP or a SP based on whether they started a game in their rookie season.

I also obtained the top 100 prospect rankings from Baseball America and MLB Pipeline for the 2014-2024 seasons as well as Baseball America's top 30 prospect rankings for each team from the 2014-2024 seasons (note that these are the rankings prior to the start of each season).

This should give us some insight as to how scouts view these players beyond what the stats may tell us (basically, how does the eye test project a certain player). Obviously, only a few players in the dataset will be included in these rankings, but that should give us an idea of which players show the most potential to outside evaluators relative to their peers. These rankings are well-regarded in the baseball community, and it has a good track record for correctly identifying the top prospects in the sport.

For this paper, I wanted to focus on projecting both the standard stats (like AVG/OBP/SLG for hitters and ERA for pitchers) and what some would consider "process based stats" that don't directly translate to results/success but are highly correlated with it (K%, BB%, & SwStr% for hitters and FIP, K%, BB%, & SwStr% for pitchers). To project these, I utilized their performance in the previous level as well as their age, Baseball America Top 100 Ranking (rank_ba), MLB Pipeline Top 100 Ranking (rank_mlb), and their top 30 ranking among their team's farm system (team_rank_ba).



Descriptive Statistics

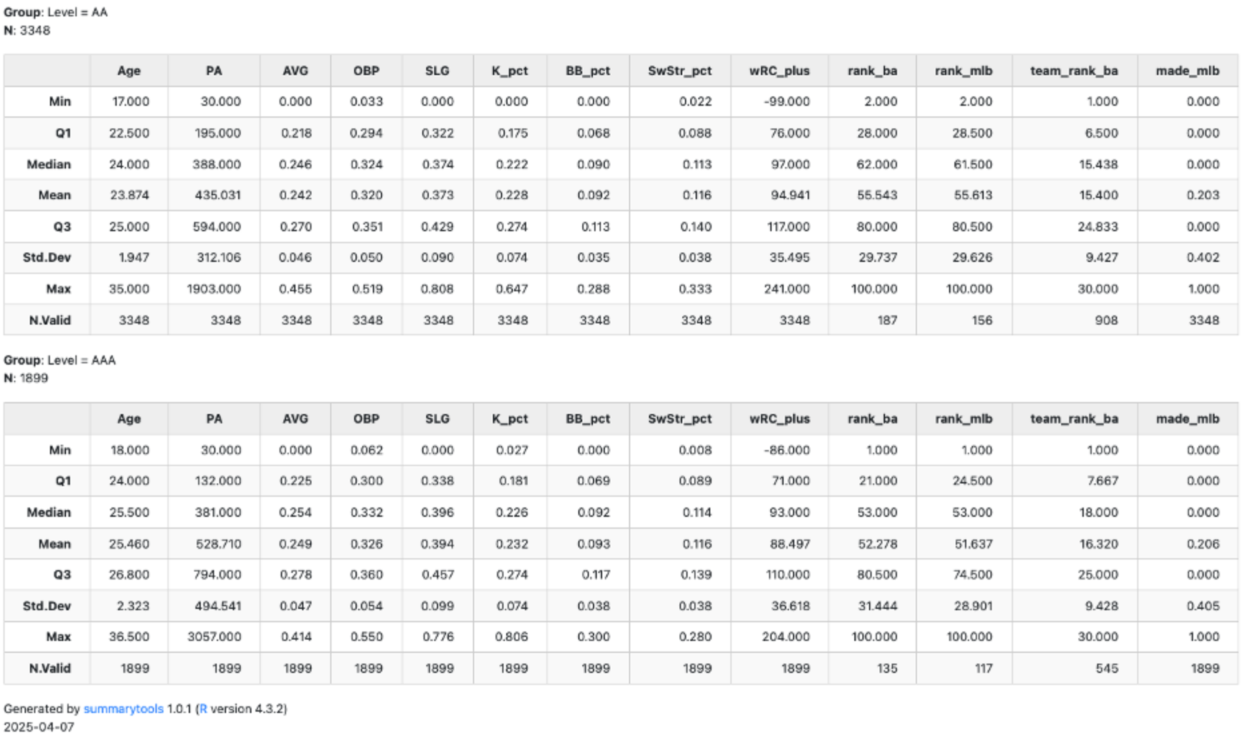
minor_league_hitting_summary_stats

Group: Level = A
N: 5040

	Age	PA	AVG	OBP	SLG	K_pct	BB_pct	SwStr_pct	wRC_plus	rank_ba	rank_mlb	team_rank_ba	made_mlb
Min	17.000	30.000	0.000	0.091	0.000	0.000	0.000	0.017	-58.000	4.000	12.000	1.000	0.000
Q1	20.333	126.000	0.213	0.294	0.304	0.181	0.068	0.101	77.000	49.000	46.500	10.000	0.000
Median	21.500	265.000	0.244	0.326	0.358	0.228	0.092	0.131	99.000	67.500	63.000	19.000	0.000
Mean	21.470	295.853	0.242	0.327	0.361	0.236	0.098	0.139	97.949	67.006	61.109	17.703	0.086
Q3	23.000	427.000	0.273	0.361	0.415	0.282	0.122	0.167	120.000	86.167	84.000	25.500	0.000
Std.Dev	1.581	200.647	0.049	0.055	0.092	0.078	0.042	0.056	35.183	24.149	25.449	8.495	0.280
Max	29.000	1483.000	0.516	0.579	0.960	0.625	0.375	0.466	273.000	99.000	98.500	30.000	1.000
N.Valid	5040	5040	5040	5040	5040	5040	5040	5040	5040	56	43	772	5040

Group: Level = A+
N: 4302

	Age	PA	AVG	OBP	SLG	K_pct	BB_pct	SwStr_pct	wRC_plus	rank_ba	rank_mlb	team_rank_ba	made_mlb
Min	17.500	30.000	0.041	0.097	0.041	0.024	0.000	0.022	-54.000	1.000	3.000	1.000	0.000
Q1	21.500	161.000	0.216	0.293	0.314	0.182	0.067	0.104	77.000	38.000	35.000	9.000	0.000
Median	22.500	304.000	0.246	0.324	0.368	0.228	0.090	0.133	98.000	60.500	62.000	17.000	0.000
Mean	22.562	342.225	0.244	0.323	0.371	0.235	0.092	0.142	96.520	58.577	58.582	16.148	0.130
Q3	23.500	477.000	0.273	0.353	0.425	0.281	0.113	0.169	118.000	85.000	87.000	24.000	0.000
Std.Dev	1.574	228.165	0.047	0.050	0.093	0.076	0.036	0.056	34.782	29.502	28.518	8.758	0.337
Max	31.000	1572.000	0.456	0.559	1.000	0.742	0.273	0.453	264.000	100.000	99.000	30.000	1.000
N.Valid	4302	4302	4302	4302	4302	4302	4302	4302	4302	110	99	920	4302



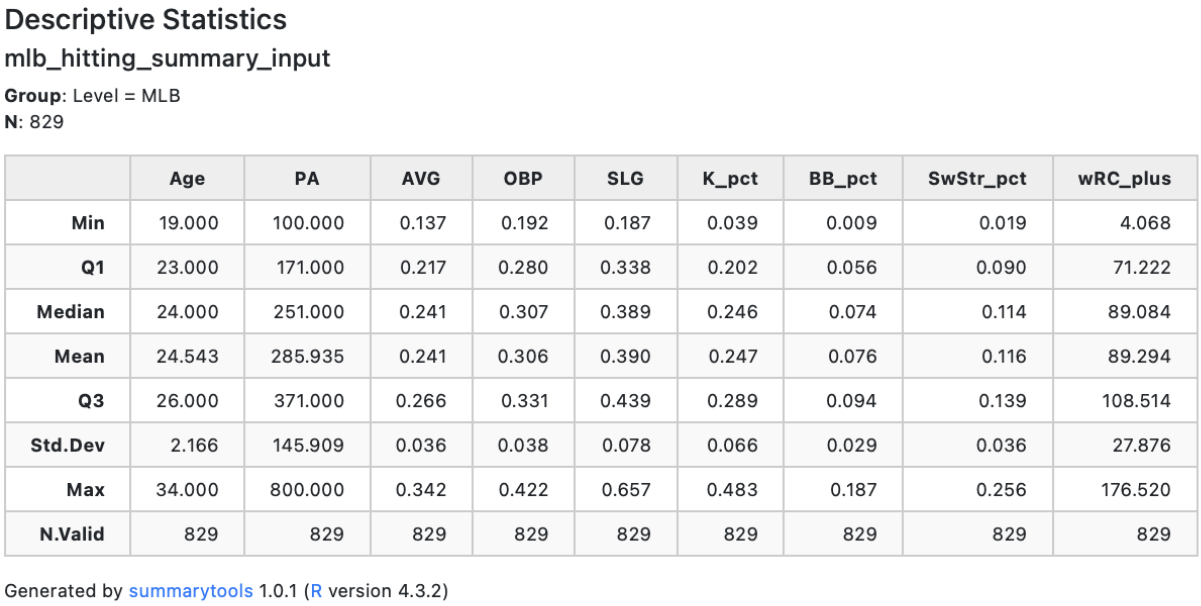
Group: Level = AA
N: 3348

	Age	PA	AVG	OBP	SLG	K_pct	BB_pct	SwStr_pct	wRC_plus	rank_ba	rank_mlb	team_rank_ba	made_mlb
Min	17.000	30.000	0.000	0.033	0.000	0.000	0.000	0.022	-99.000	2.000	2.000	1.000	0.000
Q1	22.500	195.000	0.218	0.294	0.322	0.175	0.068	0.088	76.000	28.000	28.500	6.500	0.000
Median	24.000	388.000	0.246	0.324	0.374	0.222	0.090	0.113	97.000	62.000	61.500	15.438	0.000
Mean	23.874	435.031	0.242	0.320	0.373	0.228	0.092	0.116	94.941	55.543	55.613	15.400	0.203
Q3	25.000	594.000	0.270	0.351	0.429	0.274	0.113	0.140	117.000	80.000	80.500	24.833	0.000
Std.Dev	1.947	312.106	0.046	0.050	0.090	0.074	0.035	0.038	35.495	29.737	29.626	9.427	0.402
Max	35.000	1903.000	0.455	0.519	0.808	0.647	0.288	0.333	241.000	100.000	100.000	30.000	1.000
N.Valid	3348	3348	3348	3348	3348	3348	3348	3348	3348	187	156	908	3348

Group: Level = AAA
N: 1899

	Age	PA	AVG	OBP	SLG	K_pct	BB_pct	SwStr_pct	wRC_plus	rank_ba	rank_mlb	team_rank_ba	made_mlb
Min	18.000	30.000	0.000	0.062	0.000	0.027	0.000	0.008	-86.000	1.000	1.000	1.000	0.000
Q1	24.000	132.000	0.225	0.300	0.338	0.181	0.069	0.089	71.000	21.000	24.500	7.667	0.000
Median	25.500	381.000	0.254	0.332	0.396	0.226	0.092	0.114	93.000	53.000	53.000	18.000	0.000
Mean	25.460	528.710	0.249	0.326	0.394	0.232	0.093	0.116	88.497	52.278	51.637	16.320	0.206
Q3	26.800	794.000	0.278	0.360	0.457	0.274	0.117	0.139	110.000	80.500	74.500	25.000	0.000
Std.Dev	2.323	494.541	0.047	0.054	0.099	0.074	0.038	0.038	36.618	31.444	28.901	9.428	0.405
Max	36.500	3057.000	0.414	0.550	0.776	0.806	0.300	0.280	204.000	100.000	100.000	30.000	1.000
N.Valid	1899	1899	1899	1899	1899	1899	1899	1899	1899	135	117	545	1899

Generated by summarytools 1.0.1 (R version 4.3.2)
2025-04-07



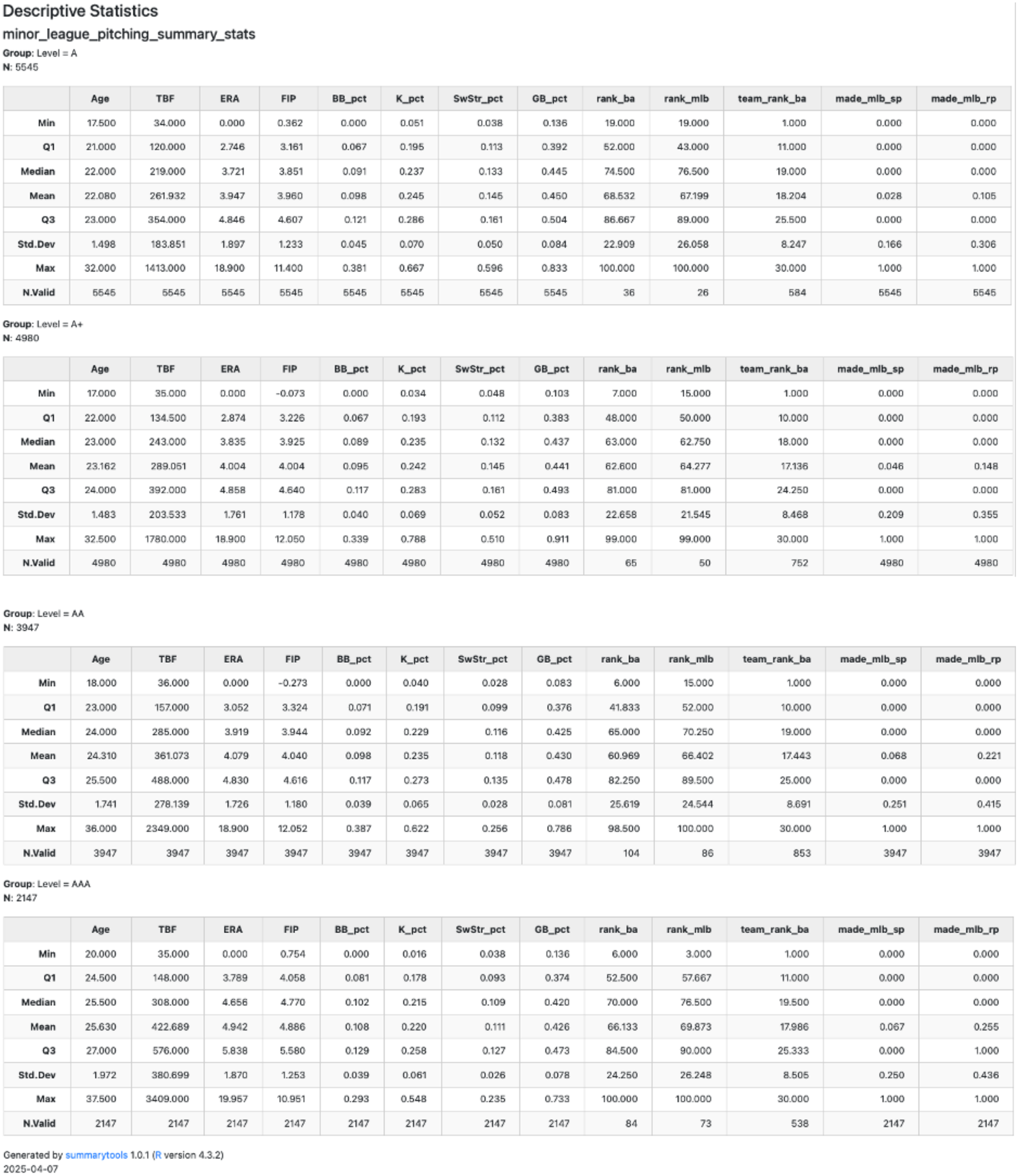
Descriptive Statistics

mlb_hitting_summary_input

Group: Level = MLB
N: 829

	Age	PA	AVG	OBP	SLG	K_pct	BB_pct	SwStr_pct	wRC_plus
Min	19.000	100.000	0.137	0.192	0.187	0.039	0.009	0.019	4.068
Q1	23.000	171.000	0.217	0.280	0.338	0.202	0.056	0.090	71.222
Median	24.000	251.000	0.241	0.307	0.389	0.246	0.074	0.114	89.084
Mean	24.543	285.935	0.241	0.306	0.390	0.247	0.076	0.116	89.294
Q3	26.000	371.000	0.266	0.331	0.439	0.289	0.094	0.139	108.514
Std.Dev	2.166	145.909	0.036	0.038	0.078	0.066	0.029	0.036	27.876
Max	34.000	800.000	0.342	0.422	0.657	0.483	0.187	0.256	176.520
N.Valid	829	829	829	829	829	829	829	829	829

Generated by summarytools 1.0.1 (R version 4.3.2)



Descriptive Statistics

minor_league_pitching_summary_stats

Group: Level = A
N: 5545

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct	rank_ba	rank_mlb	team_rank_ba	made_mlb_sp	made_mlb_rp
Min	17.500	34.000	0.000	0.362	0.000	0.051	0.038	0.136	19.000	19.000	1.000	0.000	0.000
Q1	21.000	120.000	2.746	3.161	0.067	0.195	0.113	0.392	52.000	43.000	11.000	0.000	0.000
Median	22.000	219.000	3.721	3.851	0.091	0.237	0.133	0.445	74.500	76.500	19.000	0.000	0.000
Mean	22.080	261.932	3.947	3.960	0.098	0.245	0.145	0.450	68.532	67.199	18.204	0.028	0.105
Q3	23.000	354.000	4.846	4.607	0.121	0.286	0.161	0.504	86.667	89.000	25.500	0.000	0.000
Std.Dev	1.498	183.851	1.897	1.233	0.045	0.070	0.050	0.084	22.909	26.058	8.247	0.166	0.306
Max	32.000	1413.000	18.900	11.400	0.381	0.667	0.596	0.833	100.000	100.000	30.000	1.000	1.000
N.Valid	5545	5545	5545	5545	5545	5545	5545	5545	36	26	584	5545	5545

Group: Level = A+
N: 4980

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct	rank_ba	rank_mlb	team_rank_ba	made_mlb_sp	made_mlb_rp
Min	17.000	35.000	0.000	-0.073	0.000	0.034	0.048	0.103	7.000	15.000	1.000	0.000	0.000
Q1	22.000	134.500	2.874	3.226	0.067	0.193	0.112	0.383	48.000	50.000	10.000	0.000	0.000
Median	23.000	243.000	3.835	3.925	0.089	0.235	0.132	0.437	63.000	62.750	18.000	0.000	0.000
Mean	23.162	289.051	4.004	4.004	0.095	0.242	0.145	0.441	62.600	64.277	17.136	0.046	0.148
Q3	24.000	392.000	4.858	4.640	0.117	0.283	0.161	0.493	81.000	81.000	24.250	0.000	0.000
Std.Dev	1.483	203.533	1.761	1.178	0.040	0.069	0.052	0.083	22.658	21.545	8.468	0.209	0.355
Max	32.500	1780.000	18.900	12.050	0.339	0.788	0.510	0.911	99.000	99.000	30.000	1.000	1.000
N.Valid	4980	4980	4980	4980	4980	4980	4980	4980	65	50	752	4980	4980

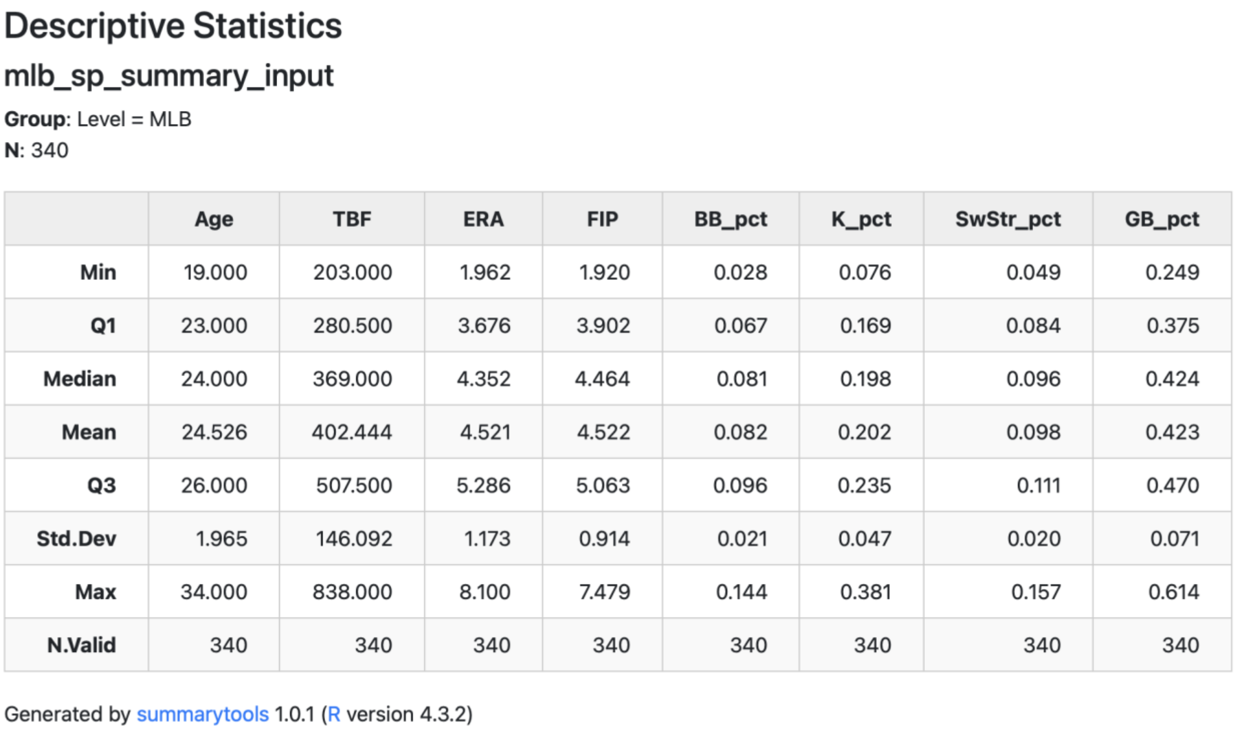
Group: Level = AA
N: 3947

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct	rank_ba	rank_mlb	team_rank_ba	made_mlb_sp	made_mlb_rp
Min	18.000	36.000	0.000	-0.273	0.000	0.040	0.028	0.083	6.000	15.000	1.000	0.000	0.000
Q1	23.000	157.000	3.052	3.324	0.071	0.191	0.099	0.376	41.833	52.000	10.000	0.000	0.000
Median	24.000	285.000	3.919	3.944	0.092	0.229	0.116	0.425	65.000	70.250	19.000	0.000	0.000
Mean	24.310	361.073	4.079	4.040	0.098	0.235	0.118	0.430	60.969	66.402	17.443	0.068	0.221
Q3	25.500	488.000	4.830	4.616	0.117	0.273	0.135	0.478	82.250	89.500	25.000	0.000	0.000
Std.Dev	1.741	278.139	1.726	1.180	0.039	0.065	0.028	0.081	25.619	24.544	8.691	0.251	0.415
Max	36.000	2349.000	18.900	12.052	0.387	0.622	0.256	0.786	98.500	100.000	30.000	1.000	1.000
N.Valid	3947	3947	3947	3947	3947	3947	3947	3947	104	86	853	3947	3947

Group: Level = AAA
N: 2147

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct	rank_ba	rank_mlb	team_rank_ba	made_mlb_sp	made_mlb_rp
Min	20.000	35.000	0.000	0.754	0.000	0.016	0.038	0.136	6.000	3.000	1.000	0.000	0.000
Q1	24.500	148.000	3.789	4.058	0.081	0.178	0.093	0.374	52.500	57.667	11.000	0.000	0.000
Median	25.500	308.000	4.656	4.770	0.102	0.215	0.109	0.420	70.000	76.500	19.500	0.000	0.000
Mean	25.630	422.689	4.942	4.886	0.108	0.220	0.111	0.426	66.133	69.873	17.986	0.067	0.255
Q3	27.000	576.000	5.838	5.580	0.129	0.258	0.127	0.473	84.500	90.000	25.333	0.000	1.000
Std.Dev	1.972	380.699	1.870	1.253	0.039	0.061	0.026	0.078	24.250	26.248	8.505	0.250	0.436
Max	37.500	3409.000	19.957	10.951	0.293	0.548	0.235	0.733	100.000	100.000	30.000	1.000	1.000
N.Valid	2147	2147	2147	2147	2147	2147	2147	2147	84	73	538	2147	2147

Generated by summarytools 1.0.1 (R version 4.3.2)
2025-04-07



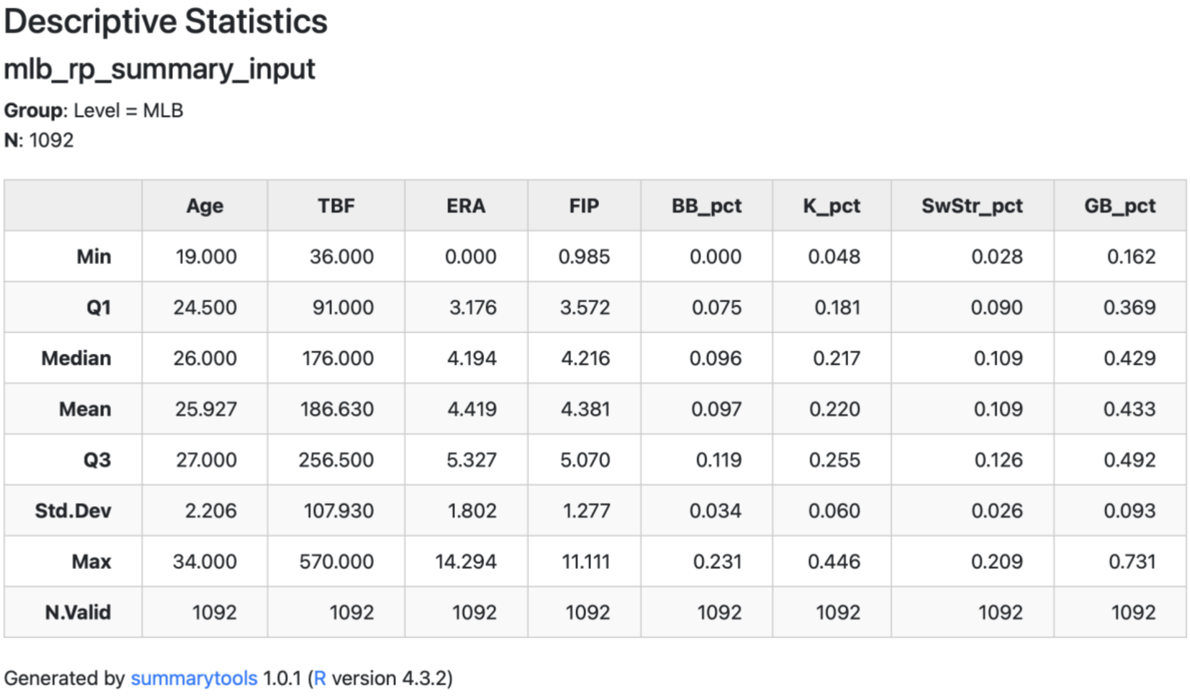
Descriptive Statistics

mlb_sp_summary_input

Group: Level = MLB
N: 340

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct
Min	19.000	203.000	1.962	1.920	0.028	0.076	0.049	0.249
Q1	23.000	280.500	3.676	3.902	0.067	0.169	0.084	0.375
Median	24.000	369.000	4.352	4.464	0.081	0.198	0.096	0.424
Mean	24.526	402.444	4.521	4.522	0.082	0.202	0.098	0.423
Q3	26.000	507.500	5.286	5.063	0.096	0.235	0.111	0.470
Std.Dev	1.965	146.092	1.173	0.914	0.021	0.047	0.020	0.071
Max	34.000	838.000	8.100	7.479	0.144	0.381	0.157	0.614
N.Valid	340	340	340	340	340	340	340	340

Generated by summarytools 1.0.1 (R version 4.3.2)



Descriptive Statistics

mlb_rp_summary_input

Group: Level = MLB
N: 1092

	Age	TBF	ERA	FIP	BB_pct	K_pct	SwStr_pct	GB_pct
Min	19.000	36.000	0.000	0.985	0.000	0.048	0.028	0.162
Q1	24.500	91.000	3.176	3.572	0.075	0.181	0.090	0.369
Median	26.000	176.000	4.194	4.216	0.096	0.217	0.109	0.429
Mean	25.927	186.630	4.419	4.381	0.097	0.220	0.109	0.433
Q3	27.000	256.500	5.327	5.070	0.119	0.255	0.126	0.492
Std.Dev	2.206	107.930	1.802	1.277	0.034	0.060	0.026	0.093
Max	34.000	570.000	14.294	11.111	0.231	0.446	0.209	0.731
N.Valid	1092	1092	1092	1092	1092	1092	1092	1092

Generated by summarytools 1.0.1 (R version 4.3.2)

Figure 2: Summary tables for each level of professional baseball.

Looking at the summary tables, we can see that both Baseball America and MLB Pipeline seem to favor batters when building their rankings despite more pitchers being present in the dataset. This could suggest two things: one being that pitching prospects are more top-heavy compared to hitting prospects with the other being that positional variability gives batters much more opportunities to be promoted to the MLB roster which comes across in the prospect rankings. We can also see that the mean probability to make the MLB is the same at both AA and AAA for both pitchers and batters, suggesting that AA is the first “big test” for prospects.

IV. Level-To-Level Observed Changes

To get a brief introduction to the data, I wanted to see how player stats change from level-to-level to see which transition has the largest impact on performance. This can give some insight into which level jump proves to be the most challenging for batters and pitchers. Logically, their performance in a metric at the previous level is going to be the best predictor of their performance in that metric at the next level. Using this, I can compare how well a metric translates to the next level.

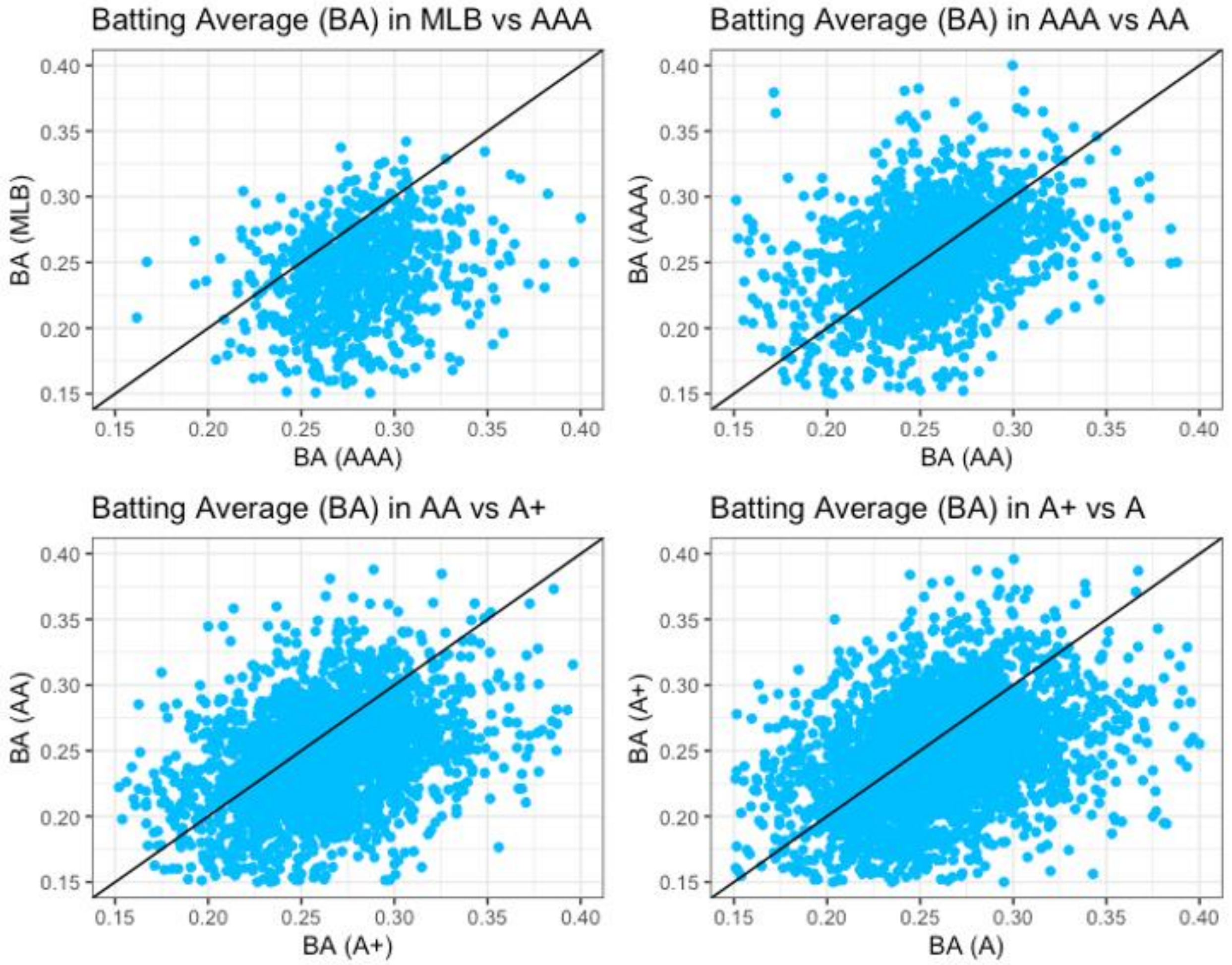


Figure 3: Comparing current level batting average (BA) to next level BA to compare the transition difficulty from level-to-level throughout professional baseball. The black diagonal line represents all points where current level BA equals next level BA. The transition from AAA to MLB seems to have the harshest effect on BA while the transition from A+ to AA also has a moderate effect.

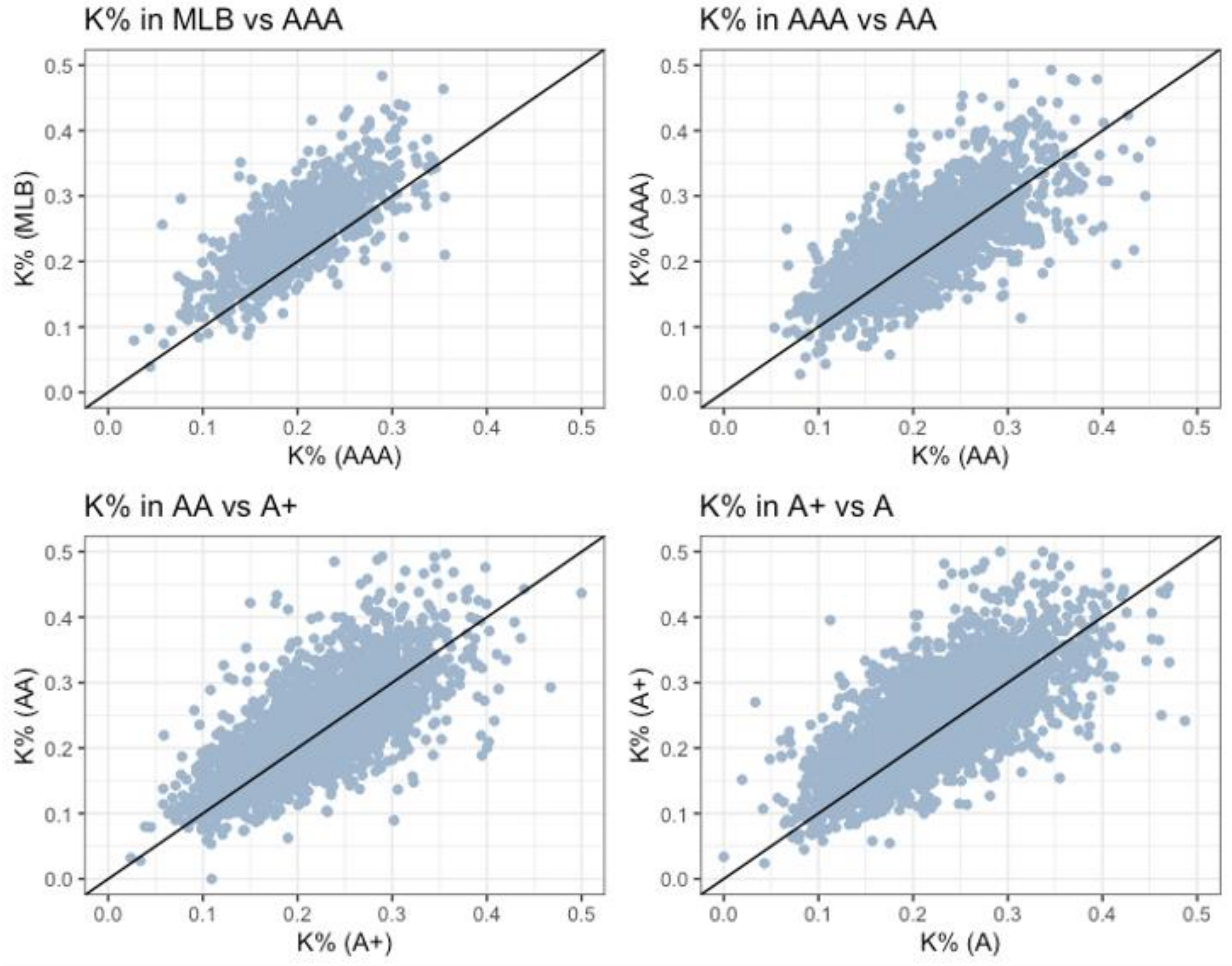


Figure 4: Similar to Figure 3, but for K%. We see similar trends between transitioning to different levels of professional baseball where AAA to MLB is the toughest for batters while A+ to AA also shows a moderate effect (although to less of a degree than BA).

When comparing K% and batting average (BA) across levels, we can see that the hardest transition for a batter is from AAA to MLB (as one would expect). BA and K% sharply decrease once a batter transitions to the MLB as they start facing more elite, high-end pitching. We can also see that the transition from A+ to AA can be challenging for batters, although this transition is much more subtle and is particularly prevalent in BA. This suggests that the transition from A+ to AA is the big transition among the minor leagues that can show whether or not a batter is good enough for the MLB.

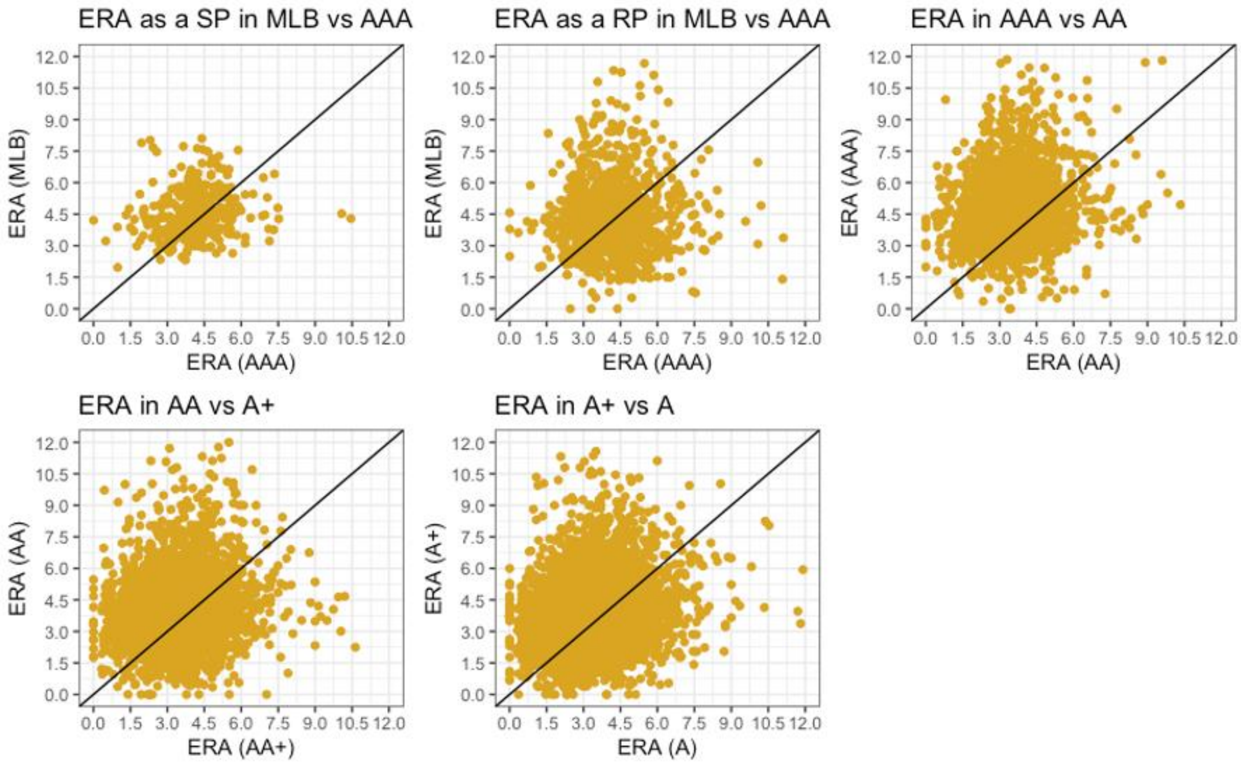


Figure 5: Illustrates the overall impact on ERA after getting promoted to the next level on professional baseball, where the black diagonal line represents all points where ERA remains the same from level to level. Here, we can see that the transition from AA to AAA seems to have the largest effect on a pitcher's ERA.

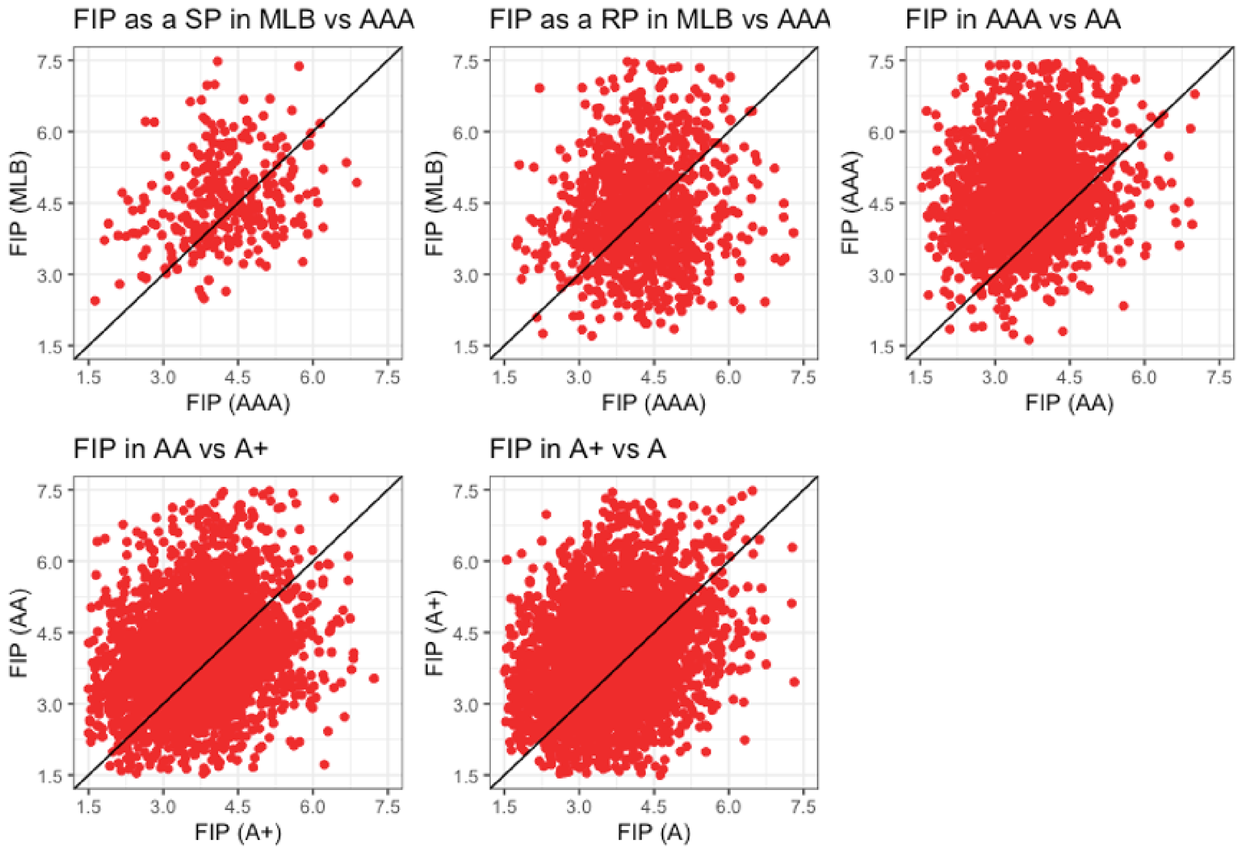


Figure 6: Similar to Figure 5, but for FIP instead of ERA. We see similar results where the transition from AA to AAA has the largest impact on a pitcher's FIP.

For pitchers, we can see that the first real test occurs when getting called up to AAA. This level is often filled with former MLB players, “AAAA” players, and top prospects, which means there are much fewer holes in a lineup. This forces pitchers to execute their pitches at a higher rate and those who are unable to will be exposed.

V. Aging Curves in MLB Promotion

To expand on my data exploration, I wanted to look at how age impacts a player’s likelihood of getting called up to the MLB. As a player gets older, his value to an MLB organization wanes. It’s a sign that he doesn’t have what it takes to make a big league roster or be a part of the team’s long-term plans.

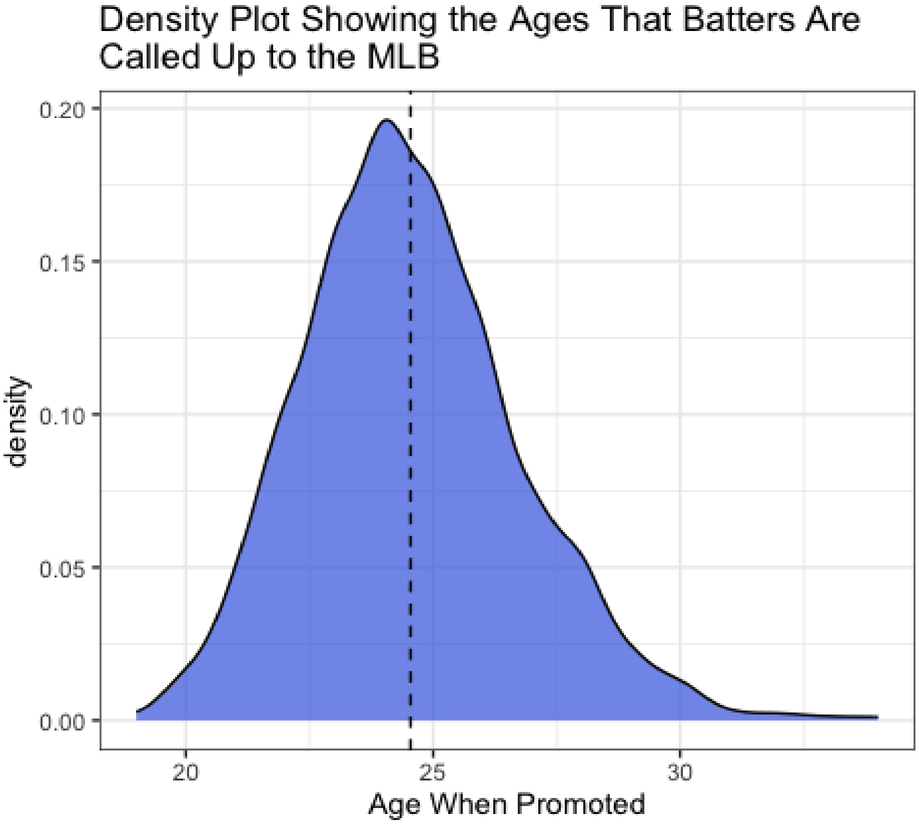


Figure 7: Density Plot of the ages batters are called up to the MLB

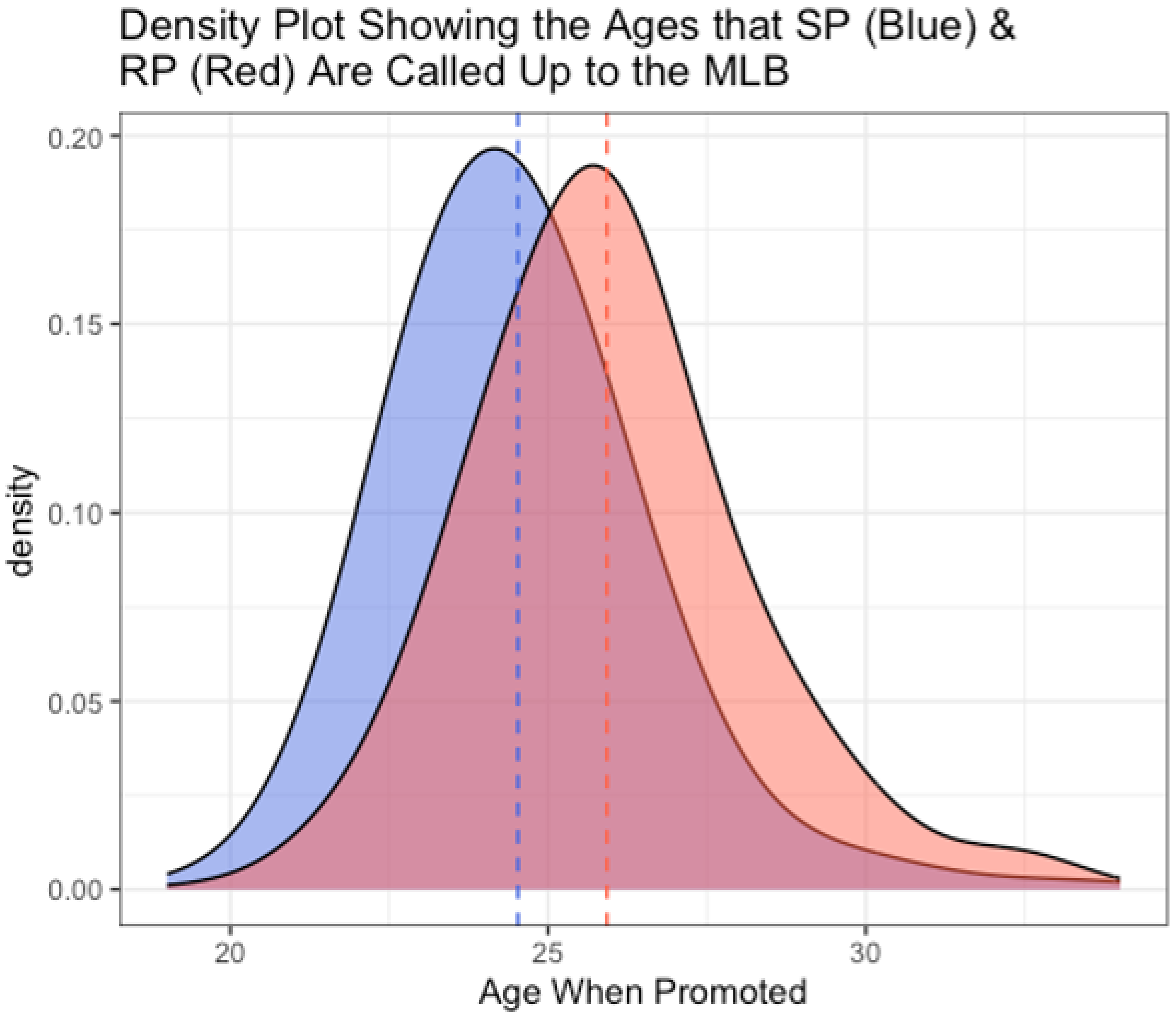


Figure 8: Density plot of the average ages SP (blue) and RP (red) are called up to the MLB

Looking at the first density plot above, we can see that a minor league batter is called up by the time he's 24 on average, and the likelihood he gets called up starkly decreases as he ages, with only a few batters getting called up once they hit 30. For pitchers, the distribution looks a bit different depending on whether a player gets called up as a SP or RP. To be called up as a SP, you need to be called up sooner in your professional career while age isn't nearly as important for RP. This makes sense, as SPs are supposed to be your top 5-6 arms on your roster so they should make their way through the minor league system much quicker.

While these density plots give good insight into how age impacts prospects' likelihood of getting called up, I wanted to dig a bit deeper and see if I could isolate the impact of age on MLB promotion. To do this, I utilized generalized additive models (GAM) to create aging curves at each level regressing against whether a player made the MLB. I also controlled for top 100 ranking, top 30 team ranking, and performance metrics at each level to isolate the effects of age on major league likelihood (see Appendix). I also removed any players who haven't had sufficient time to make the MLB based on the median amount of time it took players who made the MLB to make it. For instance, in A-ball the median was 4 so I only included players who played in A-Ball from 2014-2019 (Since 2020 didn't have minor league seasons) while in AAA the median was 1 so I included players who played from the 2014-2023 seasons.

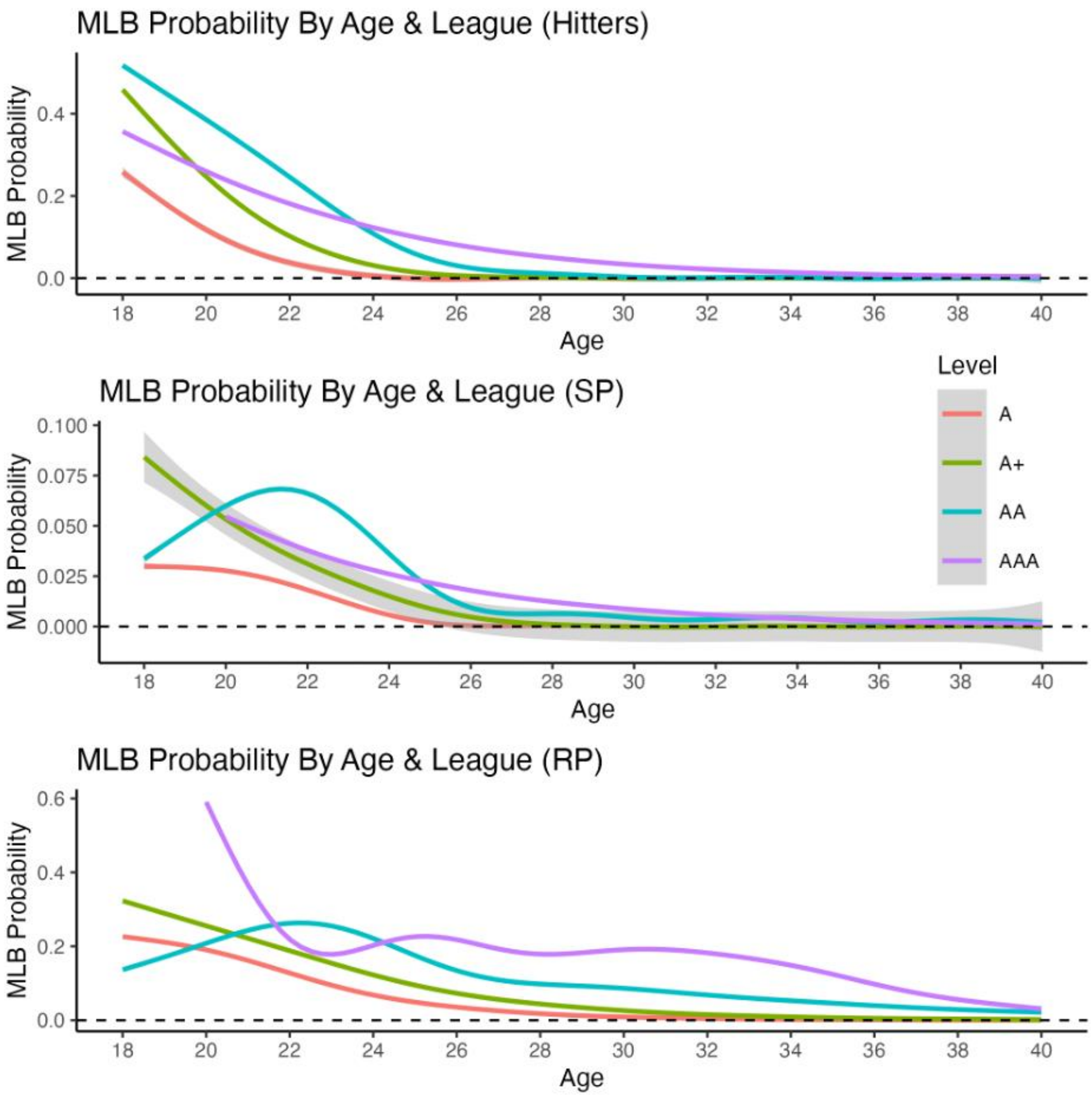


Figure 9: Displays how age impacts a player's chances of making the MLB as a batter, SP, or RP. For batters, age plays a big role in the likelihood at the the lower levels (A & A+) but becomes less important at the higher levels. For SP, getting to AA by the time they're 20-22 will maximize their chances to make a team's rotation in the future. For RP, age plays a relatively negligible role compared to batters and SP and is based more on performance.

These plots tell some pretty interesting stories. In the hitters' plot, we can see that age is really important at the lower levels and even spending 1 extra year at these lower levels significantly decreases the likelihood of getting called up in the future. However, at AAA we can see that age plays much less of a role in a player's likelihood of making the MLB and that overall performance plays a much bigger role. For SP, we see a bit more leniency at A-ball in regards to age and don't see the likelihood decrease until their age 22 season at that level. We can also see a big hump in AA for SP where MLB probability peaks at around 21-22. This can probably be explained by college pitchers who are drafted in the first rounds, as these guys often skip A & A+ ball entirely and go straight to AA ball, which is probably what the model is picking up on. Finally, for relief pitchers, age has very little to do with a pitcher's probability of making the MLB relative to hitters and batters, especially at the AAA. Most prospects start out as SP, so age being less of a factor for RP isn't too surprising. Especially in recent years, we've seen numerous RP completely revamp their career after an offseason or mid-season change. On the Yankees alone, we have seen RP like Ian Hamilton, Lucas Luetge, and Luke Weaver come out of nowhere to become vital parts of the Yankees bullpen, showing how many players can become impactful RP even when they're older.

VI. Predicting Performance at the Next Level & the Probability of Making MLB Using ML

Now to get into what I'm really interested in: can we predict a batter's performance at the next level just based on their metrics at their current level? Scouting (at the public sphere) is very heavily based on film analysis which is inherently subjective, and this is fine, but is there a way we can add some objectivity by analyzing a player's past performance? Or at the very least can we find some stats that are much more reliable when projecting performance at the next level? To do this, I utilized Bayesian Additive Regression Trees (BART) and XGBoost models and compared their performances to see which would lead to the best outcomes given their different skill sets. BART models utilize a sum-of-trees approach where each tree contributes a small amount to the final prediction. They also heavily regularize themselves based on their priors, which helps prevent them from overfitting on the training data, which makes them suitable for small datasets or noisy data like minor league stat lines. XGBoost models utilize a boosting algorithm that aims to improve predictions by minimizing the residuals from the previous iteration. These models are often highly efficient and predictive in large datasets but have a high tendency to overfit on small datasets due to their extremely greedy nature in variable selection when building the model. For batters, I aimed to predict AVG, OBP, ISO, wRC+, K%, BB%, and SwStr% at the next level while for pitchers, I predicted ERA, FIP, K%, BB%, and SwStr%. The variables used in each model varied slightly based on external factors that impacted each stat individually, but the 4 key variables that each model had included: average top 100 rank from MLB & Baseball America (BA), top 30 team ranking from BA, Age, & a metric-specific stat that considers both rate & volume/sample size (see appendix). After running these models, BART outperformed XGBoost on all outcomes.

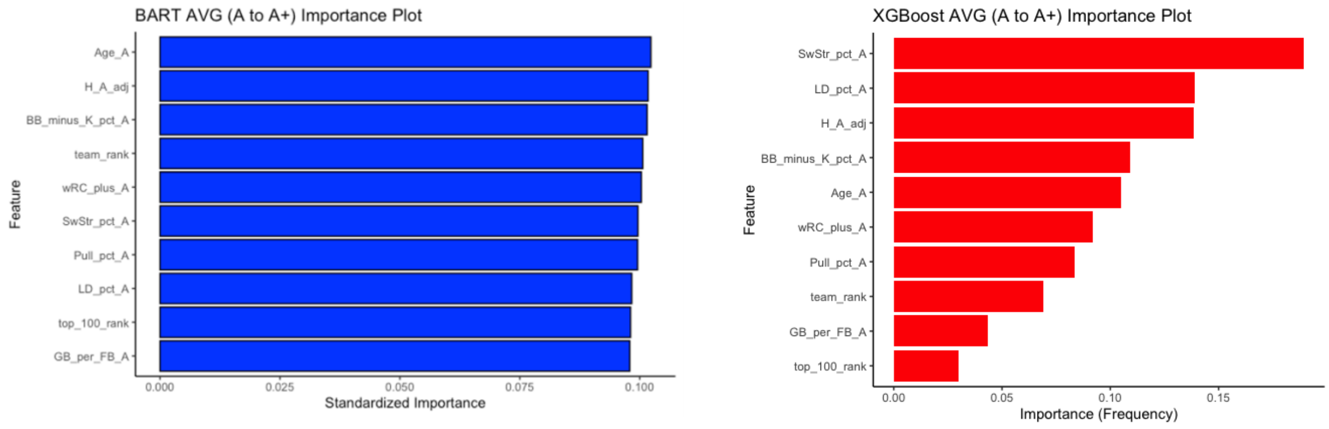
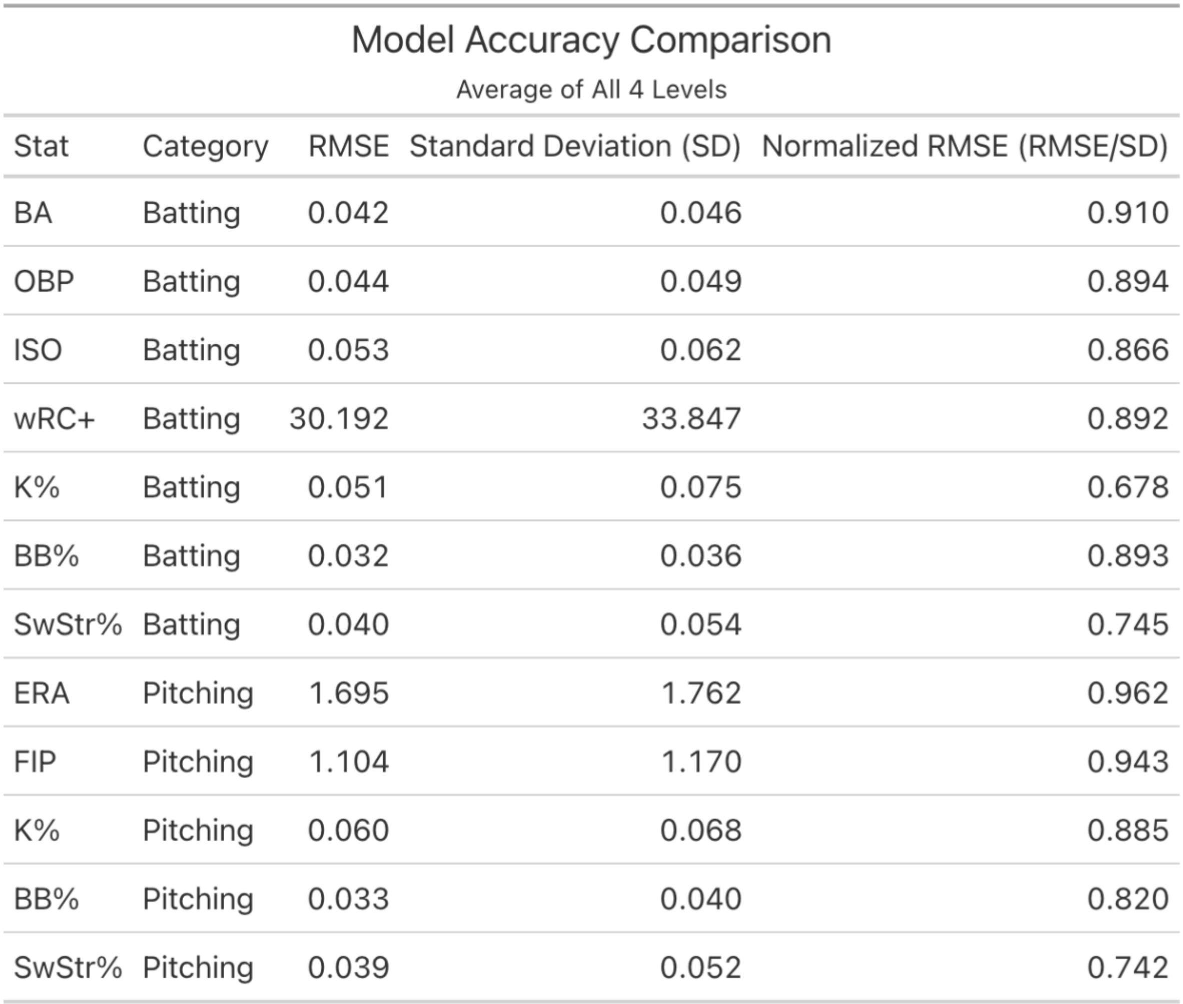


Figure 10: Importance plots for BART and XGBoost models based on how frequently each variable was used to build the model. Overall, BART was much more equitable in how it utilized the inputs given which prevented it from overfitting like XGBoost.

Looking at the BART and XGBoost importance plots when predicting AVG from A to A+, it's clear why the BART models outperformed XGBoost. BART was a lot more equitable in its variable selection, while XGBoost put a lot of emphasis on SwStr%, LD%, & adjusted hit rate which caused it to overfit. For these reasons, BART did a much better job capturing the high variability of stats at the next level.



Model Accuracy Comparison

Average of All 4 Levels

Stat	Category	RMSE	Standard Deviation (SD)	Normalized RMSE (RMSE/SD)
BA	Batting	0.042	0.046	0.910
OBP	Batting	0.044	0.049	0.894
ISO	Batting	0.053	0.062	0.866
wRC+	Batting	30.192	33.847	0.892
K%	Batting	0.051	0.075	0.678
BB%	Batting	0.032	0.036	0.893
SwStr%	Batting	0.040	0.054	0.745
ERA	Pitching	1.695	1.762	0.962
FIP	Pitching	1.104	1.170	0.943
K%	Pitching	0.060	0.068	0.885
BB%	Pitching	0.033	0.040	0.820
SwStr%	Pitching	0.039	0.052	0.742

Table 2: Average Model Accuracy across all 4 levels for each stat. We can see that for batters the model performed the best at projecting K% & SwStr% while struggling to predict BA at future levels (makes sense as BA is the most influenced by luck from the stats above). For pitchers, it struggled to predict ERA and FIP but did much better with K%, BB%, and SwStr%.

Table 2 gives us a better sense of how the model performed against each statistic when projecting future performance and can give us better insight into not only how reliable these predictions are but also what stats are more reliable from level-to-level. For batters, the model had the easiest time projecting K% at the future levels. Even at the MLB level, K% remains consistent from year-to-year so it makes sense that the same would hold true at the minor league levels. Conversely, the model seemed to have the toughest time predicting BA, most likely due to its volatile nature. Of the batting stats listed above, BA is the one that is influenced by luck the most, so it makes sense that the model would have the toughest time predicting BA at the future levels. For pitchers, the model had the toughest time predicting the two main stats used in pitcher evaluation (ERA & FIP) while performing much better with K%, BB%, and SwStr%. This highlights the importance of using supplemental stats besides ERA & FIP when evaluating pitchers at the next level, as these stats can change dramatically from level to level.

Next, I created a model that predicted the probability of someone making the MLB. For this model, I decided to stick with just an XGBoost model due to its improved predictive accuracy with unbalanced response variables compared to other models such as Naive Bayes (Hendrawan et. al, 2022). For batters, this was a simple logit regression. For pitchers, this was a multinomial logistic regression with 3 levels: didn't make MLB, made MLB as a RP, & made MLB as a SP (see Appendix for more details).

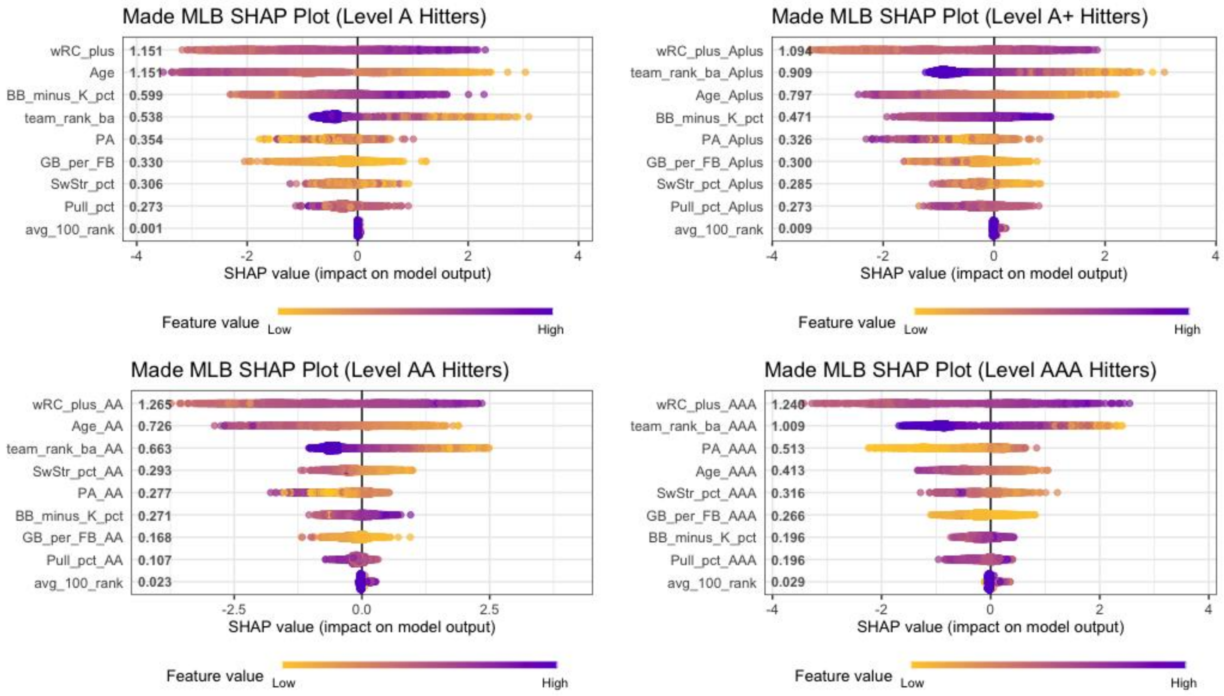
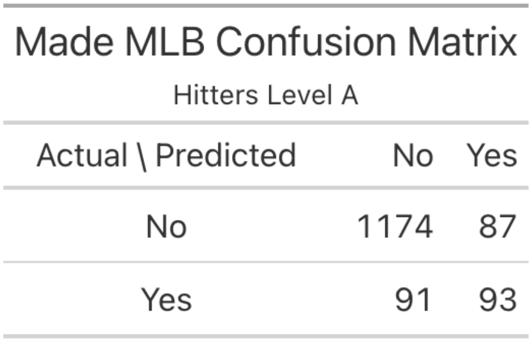


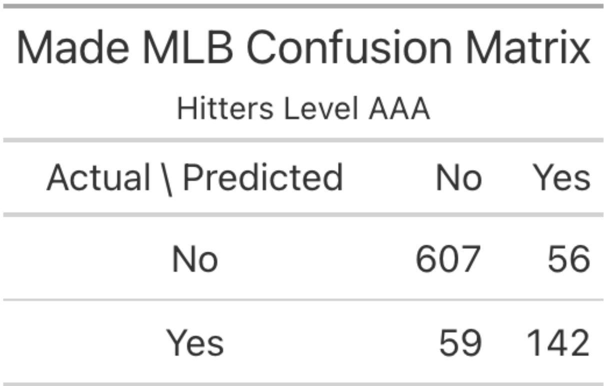
Figure 11: SHAP plot of probability to make the MLB as a hitter.



Made MLB Confusion Matrix

Hitters Level A

Actual \ Predicted	No	Yes
No	1174	87
Yes	91	93



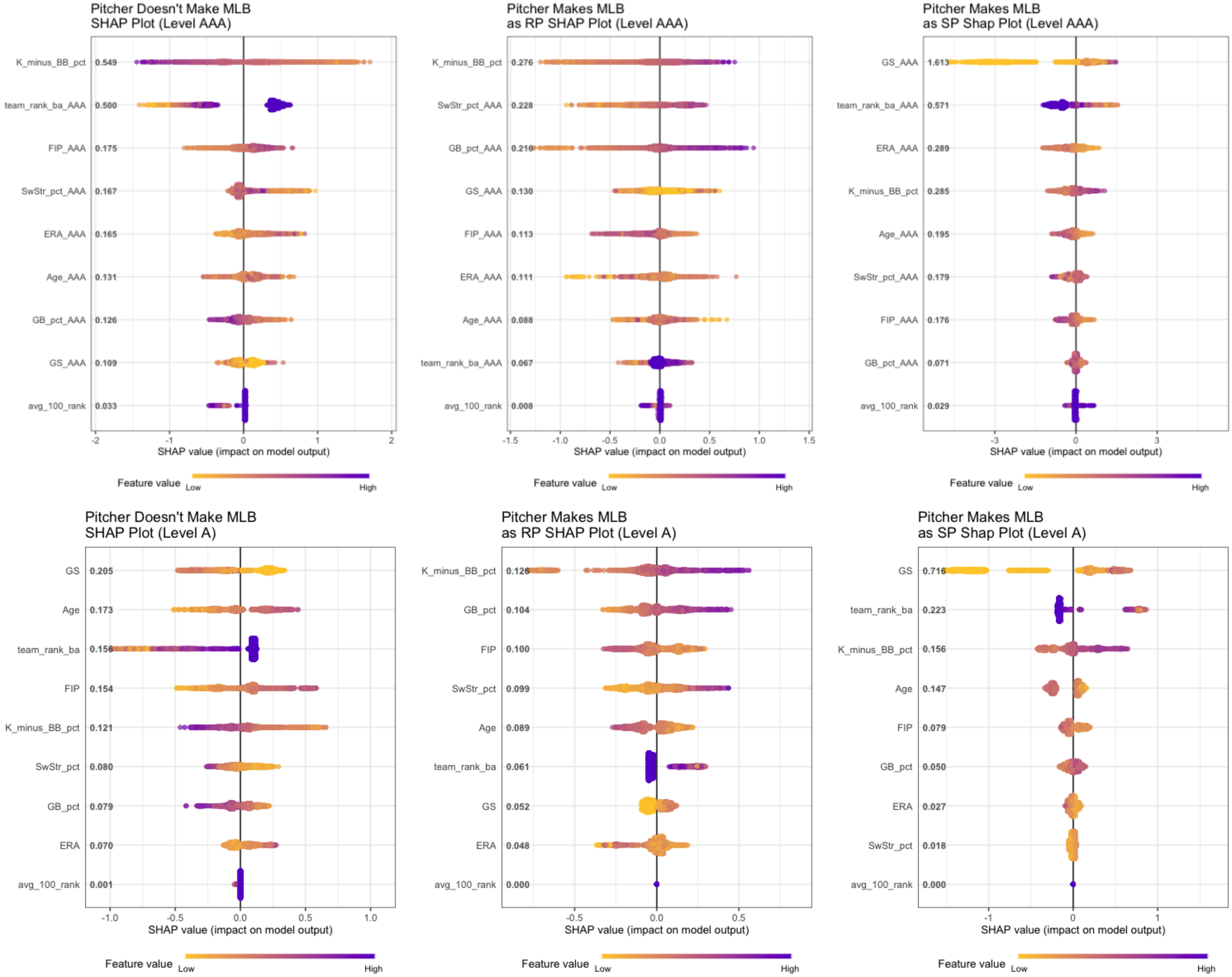
Made MLB Confusion Matrix

Hitters Level AAA

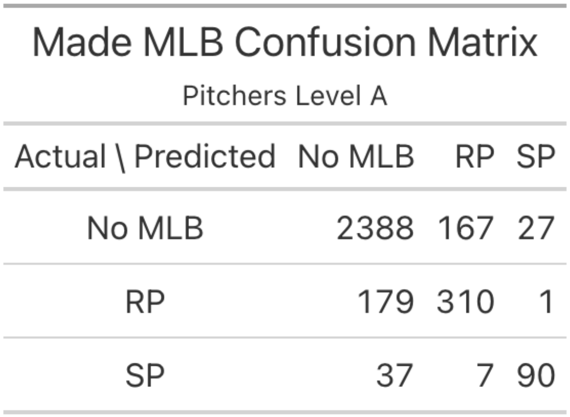
Actual \ Predicted	No	Yes
No	607	56
Yes	59	142

Tables 3 & 4: Confusion Matrix showing accuracy of predictions for hitters in single-A & AAA.

Looking at Figure 11, we see many of the same trends as the aging curves for batters in Figure 9. Age plays a big role in predicting MLB likelihood at the A, A+, & AA levels, but doesn't play a huge role in determining MLB success at the AAA level. wRC+ remains the most important factor in determining whether a batter makes the MLB, which makes sense since this is widely considered the best stat for hitters when considering overall production, followed by their top 30 ranking on Baseball America. Looking at the confusion matrices in Tables 3 & 4, we can see that the overall accuracy of these models needs to be improved in the future. Due to the volume of models needed for this project, it was hard to put too much time on any one model, which meant a cost in performance for some models. For these models in particular, it heavily overfitted on the training set and in the future, I would put more time into regularizing these models to limit overfitting. I tried to use the Synthetic Minority Oversampling Technique (SMOTE) to bring more balance to the response variable, but the model still overfitted and performance still fell. Regardless, logical principles still hold when comparing these models, as it is much easier to project who will make the MLB at the AAA level compared to A. Players in Level A still have a lot of time to improve which means that poor performance at that level isn't necessarily a death sentence while at AAA players are much closer to their full potential which gives evaluators a better idea of whether or not they are capable of playing at the MLB level.



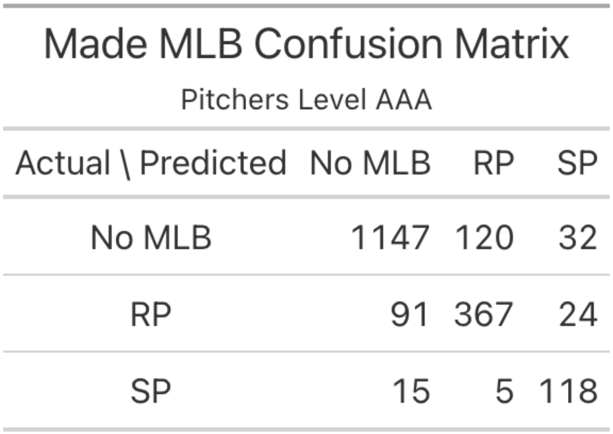
Figures 12 & 13: SHAP plots of a pitchers probability of making the MLB at the Single-A (A) level and the AAA level.



Made MLB Confusion Matrix

Pitchers Level A

Actual \ Predicted	No MLB	RP	SP
No MLB	2388	167	27
RP	179	310	1
SP	37	7	90



Made MLB Confusion Matrix

Pitchers Level AAA

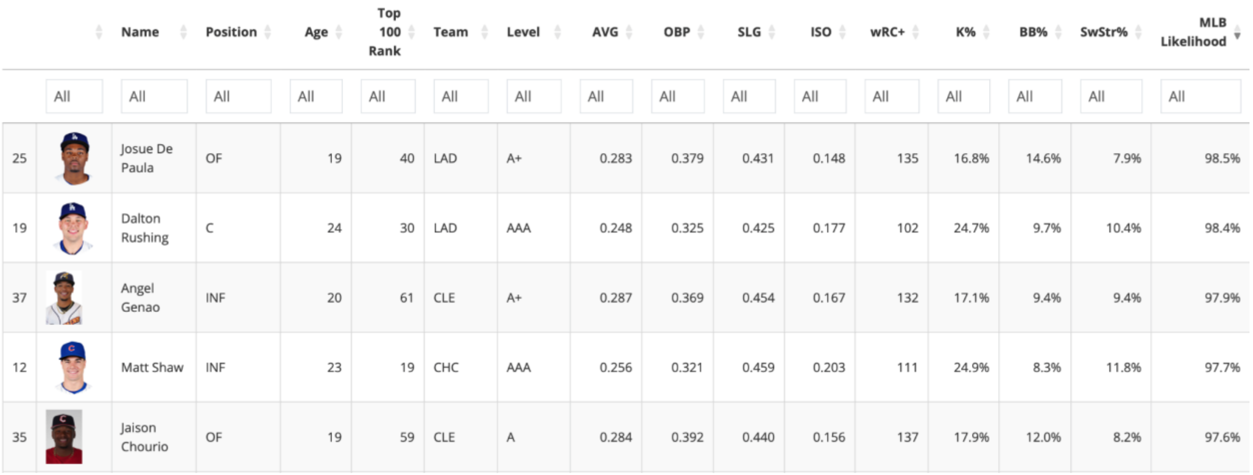
Actual \ Predicted	No MLB	RP	SP
No MLB	1147	120	32
RP	91	367	24
SP	15	5	118

Tables 5 & 6: Confusion Matrices showing accuracy of predictions for pitchers at Single-A & AAA.

Figures 12 & 13 show the SHAP values for predicting MLB likelihood at the single-A & AAA levels. However, they differ from Figure 11 in the sense that the far left SHAP plot shows which variables were more important in determining who would not make the MLB, the middle shows which variables shows which were most important in making the MLB as a RP, and the far right SHAP plot shows the variables that were most important in making the MLB as a SP. In Level A, we can see that the number of games started (GS) plays a big role in determining whether a player will make the MLB, which makes sense considering most pitchers start out as SP in their professional careers before potentially moving to a RP. It’s also probably why the model has a tough time identifying RP at the A level, since most pitchers are being trained as SP at that level before potentially being transitioned into a relief role later in their professional career. In AAA, K%-BB% seems to be the best predictor of future MLB career, with it proving to be the most important variable in determining whether a pitcher wouldn’t make the MLB or make the MLB as a RP. Team ranking also proved to have a significant effect on determining whether a pitcher wouldn’t make the MLB or make the MLB as a SP, but didn’t have a huge impact on the probability of making the MLB as a RP. Unlike with the hitter models, utilizing SMOTE really helped the pitcher models identify trends to correctly forecast a pitcher’s likelihood of making the MLB while also identifying whether they were more likely to make the MLB as a RP or a SP.

VII. Shiny App

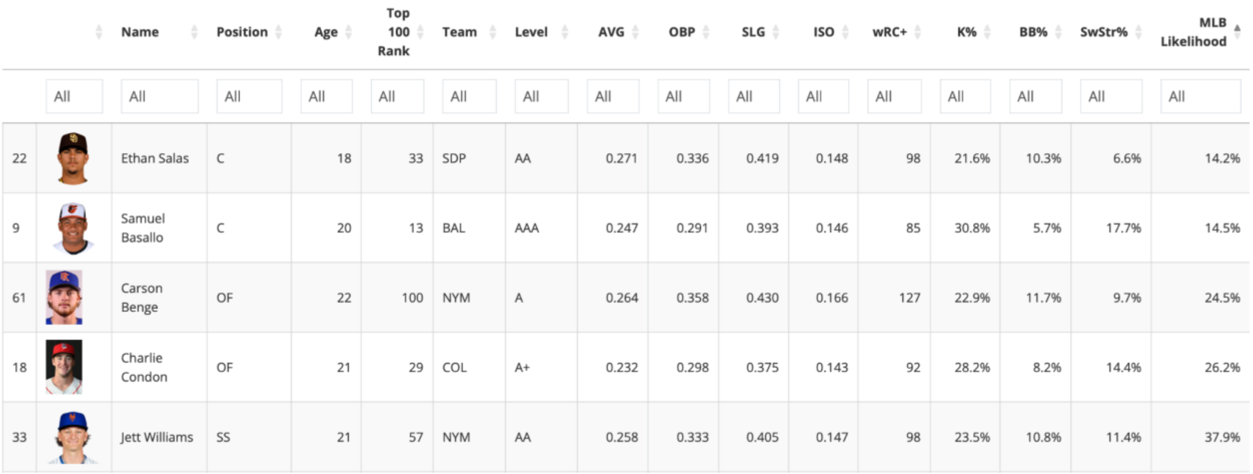
Using these models, I created a Shiny App that looks at my predictions for 2025 prospects based on my models. Teams can use this information to get a better sense of how their prospects will fare at the next level of the minors, and then compare these predictions based on what their intuition/scouting report suggests to either confirm belief or identify what they might be omitting in their initial thought process and how they can change the player's development plan to give them a better chance at the next level. This can be a great tool for evaluating whether the time is right to promote a certain player or if it's best to wait and give them more time to develop at their current level. It can also be used to evaluate whether a player is underperforming based on his stat line at the previous level of the minors. This would inform front offices to keep the prospect at that level since we can expect him to improve as we would expect him to improve as he plays more games and there would be little benefit to sending him back down. If the prospect continues to struggle, this could be a sign of another issue that was not present or noticeable at the lower levels, which player development staff can catch onto early (could be a mechanical issue, mental issue, etc.).



		Name	Position	Age	Top 100 Rank	Team	Level	AVG	OBP	SLG	ISO	wRC+	K%	BB%	SwStr%	MLB Likelihood
	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All
25		Josue De Paula	OF	19	40	LAD	A+	0.283	0.379	0.431	0.148	135	16.8%	14.6%	7.9%	98.5%
19		Dalton Rushing	C	24	30	LAD	AAA	0.248	0.325	0.425	0.177	102	24.7%	9.7%	10.4%	98.4%
37		Angel Genao	INF	20	61	CLE	A+	0.287	0.369	0.454	0.167	132	17.1%	9.4%	9.4%	97.9%
12		Matt Shaw	INF	23	19	CHC	AAA	0.256	0.321	0.459	0.203	111	24.9%	8.3%	11.8%	97.7%
35		Jaison Chourio	OF	19	59	CLE	A	0.284	0.392	0.440	0.156	137	17.9%	12.0%	8.2%	97.6%

Table 7: Table showing batters my model likes among the top 100 prospects (in terms of MLB Likelihood).

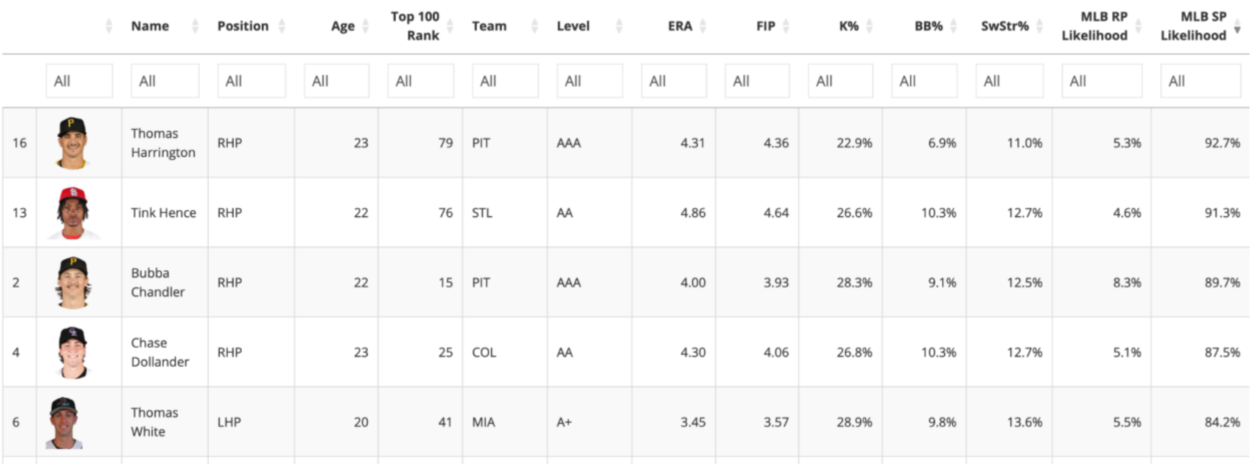
Table 7 displays the top 5 batters with the best chance of making the MLB according to my model, with Josue De Paula and Dalton Rushing at the top of the leaderboard. The stat line predictions above represent their predicted stat line at the next level if they were called up. For instance, Josue De Paula would be expected to slash .283/.379/.431 with a 16.8% K% and a 14.6% BB% at AA if he were called up. Josue De Paula had a wRC+ of 131 at the A+ level in 2024 at the age of 19, an extremely impressive achievement for his age. He also walked more often than he struck out (21.7% BB% vs 16.5% K%), which is another thing the model likes. In 2025, he has started right where he left off hot with a 159 wRC+ while also increasing his BB% to 23.2% and lowering his K% to 15.9%. Dalton Rushing is another player that the model likes a ton. In 2024, he had a wRC+ of 133 while maintaining a high walk rate (15.4%), a moderate strikeout rate (20.1%), and a low SwStr% (9.1%).



		Name	Position	Age	Top 100 Rank	Team	Level	AVG	OBP	SLG	ISO	wRC+	K%	BB%	SwStr%	MLB Likelihood
	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All
22		Ethan Salas	C	18	33	SDP	AA	0.271	0.336	0.419	0.148	98	21.6%	10.3%	6.6%	14.2%
9		Samuel Basallo	C	20	13	BAL	AAA	0.247	0.291	0.393	0.146	85	30.8%	5.7%	17.7%	14.5%
61		Carson Benge	OF	22	100	NYM	A	0.264	0.358	0.430	0.166	127	22.9%	11.7%	9.7%	24.5%
18		Charlie Condon	OF	21	29	COL	A+	0.232	0.298	0.375	0.143	92	28.2%	8.2%	14.4%	26.2%
33		Jett Williams	SS	21	57	NYM	AA	0.258	0.333	0.405	0.147	98	23.5%	10.8%	11.4%	37.9%

Table 8: Table showing batters my model doesn't like among the top 100 prospects (in terms of MLB Likelihood).

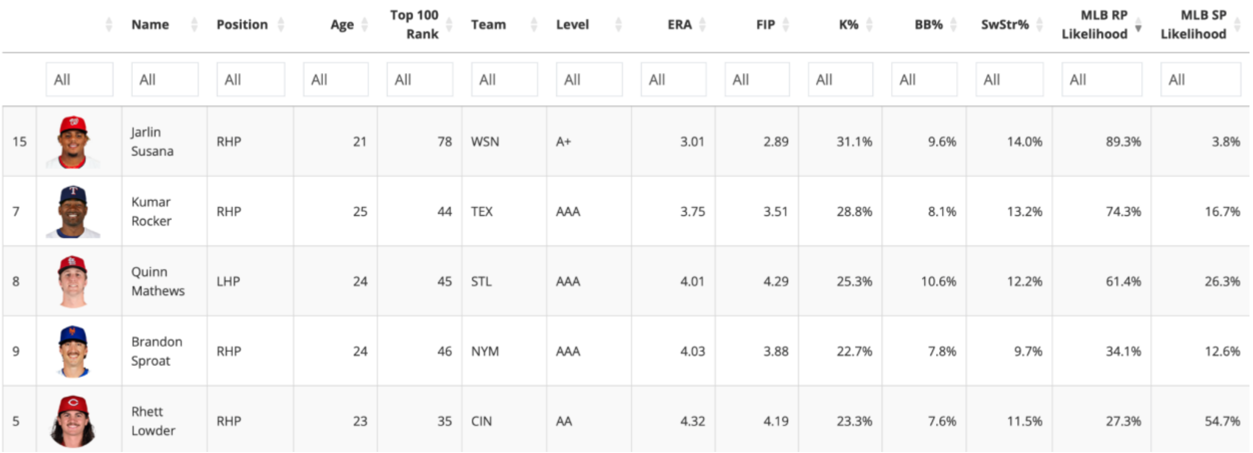
Table 8 displays the batters in the top 100 rankings with the lowest probability of making the MLB according to my model. Ethan Salas is an extremely interesting case given that he's in AA as an 18-year-old, you would expect the model to project him super highly given he's so young. However, he performed poorly in AA, albeit he only had 33 PA at the level. He had a 59 wRC+ at the level with a .179 AVG & a .214 SLG in 2023. He was sent down to A+ in 2024 and spent the entire season there. Since being called back up, he has slightly improved but not by much, having only a 75 wRC+ with a .188 AVG & a .219 SLG. The projection models give him a little grace given the low sample size at the AA level & his high prospect ranking, which is why his projected wRC+ at the AAA is 98 and his projected AVG is .271, although these are probably overestimating his actual hitting ability. However, Salas is mainly known for his fielding ability, not his bat, which is something the MLB probability model doesn't take into account. MLB Pipeline gives him 70-grade fielding on the 20-80 scouting scale along with a 60-grade arm, which probably explains why he's rated so highly. Especially at the catcher position where defense is so important, it makes sense why some scouts would rank him as highly as they do despite his less than impressive bat. Samuel Basallo is another one who ranked poorly in my rankings, and unlike Salas, he isn't a prospect who is known for his glove. Rather, this low rating is due to his poor performance at the AAA level last year, where he had a wRC+ of 66 with a K% of 31.4%, a near 10% increase from his AA K%. He also had a SwStr% of 17.7% in AAA last year, well above average and not likely to translate to the MLB level. However, this projection can be taken with a grain of salt, as Basallo has mashed at every other level, with a 125 wRC+ in AA last year and a 197 wRC+ in A+ (2023). And although he only has 22 PA so far this year, he currently has a 131 wRC+, a 18.2% K%, and a 9.8% SwStr% in AAA. Also, based on his AA stats, my model predicts that he will have a 121 wRC+ at the AAA level, further suggesting that his poor performance in AAA was a fluke. This suggests a flaw in my original model design by not including performance at previous levels of the minor leagues. While this may not be feasible at the lower levels (A & A+) due to lack of data, this could be possible at the AA & AAA levels so that the models don't put too much value on one bad season at a given level when a prospect has had a proven track record at previous levels of the minors.



		Name	Position	Age	Top 100 Rank	Team	Level	ERA	FIP	K%	BB%	SwStr%	MLB RP Likelihood	MLB SP Likelihood
	All	All	All	All	All	All	All	All	All	All	All	All	All	All
16		Thomas Harrington	RHP	23	79	PIT	AAA	4.31	4.36	22.9%	6.9%	11.0%	5.3%	92.7%
13		Tink Hence	RHP	22	76	STL	AA	4.86	4.64	26.6%	10.3%	12.7%	4.6%	91.3%
2		Bubba Chandler	RHP	22	15	PIT	AAA	4.00	3.93	28.3%	9.1%	12.5%	8.3%	89.7%
4		Chase Dollander	RHP	23	25	COL	AA	4.30	4.06	26.8%	10.3%	12.7%	5.1%	87.5%
6		Thomas White	LHP	20	41	MIA	A+	3.45	3.57	28.9%	9.8%	13.6%	5.5%	84.2%

Table 9: Table showing pitchers in the top 100 rankings my model thinks will be an MLB SP.

Table 9 displays pitchers my model believes have the best chance to become SP at the MLB level. Thomas Harrington started 8 games at the AAA level in 2024 with a 3.33 ERA, 21.2% K%, and a 4.5% BB%. His ability to limit the amount of walks he gives up while also striking out batters at a decent rate makes him a good option. He's also a very young pitcher at the AAA level, which when complemented with his plus stat line, makes him someone the model likes moving forward. Tink Hence is another person the model likes for many of the same reasons as Harrington. In 2024, Tink started 20 games with a 2.71 ERA, a 34.1% K%, & a 8.1% BB%. His nasty stuff makes him a really viable option for a future rotation spot, although the model does expect to struggle upon the transition from AA to AAA, mostly due to his high WHIP in 2023 (also because many pitchers struggle transitioning from AA to AAA).



		Name	Position	Age	Top 100 Rank	Team	Level	ERA	FIP	K%	BB%	SwStr%	MLB RP Likelihood	MLB SP Likelihood
	All	All	All	All	All	All	All	All	All	All	All	All	All	All
15		Jarlin Susana	RHP	21	78	WSN	A+	3.01	2.89	31.1%	9.6%	14.0%	89.3%	3.8%
7		Kumar Rocker	RHP	25	44	TEX	AAA	3.75	3.51	28.8%	8.1%	13.2%	74.3%	16.7%
8		Quinn Mathews	LHP	24	45	STL	AAA	4.01	4.29	25.3%	10.6%	12.2%	61.4%	26.3%
9		Brandon Sproat	RHP	24	46	NYM	AAA	4.03	3.88	22.7%	7.8%	9.7%	34.1%	12.6%
5		Rhett Lowder	RHP	23	35	CIN	AA	4.32	4.19	23.3%	7.6%	11.5%	27.3%	54.7%

Table 10: Table showing pitchers in the top 100 rankings my model thinks will be an MLB RP.

Table 10 displays the pitchers who are most likely to become RP among the top 100 prospects. The most likely being Jarlin Susana, particularly because of his extremely low FIP (2.25) in 2024 as well as his extremely high GB% of 53.8%, something that shows a lot of importance for projecting RP at the A+ level but is not nearly as important for SP (see SHAP Plot in Appendix). He also had an extremely high K% in A+ last year (36.8%), which is something the model likes as well. Kumar Rocker is second most likely to become a RP according to the model, but this is likely due to small sample size. He only started 2 games at the AAA level (pitching 10 innings exactly) and only has 36.2 IP throughout his entire minor league career spanning from the complex league to AAA, so I would not take this prediction at face value.

VIII. Limitations

As I alluded to before, the biggest flaw in my model design is the lack of utilizing stat lines at multiple levels for the higher levels of the minor league system. This led to misleading predictions for Samuel Basallo as mentioned earlier but also could be affecting other batters who may have performed poorly on their initial transition to a new level but can presumably perform better once they settle in based on their performance at the previous level of the minor leagues. I also mention defense as being a big factor that is omitted in this analysis, particularly because there aren't any good defensive metrics that are usable at the minor league level. While offensive production has been valued much more highly by MLB teams in recent years, defensive production is still a factor in whether a player gets promoted, especially for catchers. Model performance can also be improved as better data from the minor league levels come out, specifically metrics like exit velocity, launch angle, barrel rate, pitch movement, etc. These metrics have proven to be more stable at the MLB level, so it would be safe to assume that these metrics would also provide better predictive power when applied to minor league development.

This analysis also doesn't directly account for the possibility of injuries throughout a player's career and how that can impact future performance as well as the player's future likelihood of making the MLB. This mainly because injury data is not readily available at the minor league level but is also because the nature of this analysis is looking at performance from a level-by-level basis rather than performance over time. It would be interesting to see how injuries can impact a player's success at a given level by looking at the performance before and after the injury and would likely lead to better results when projecting how we can expect a player to bounce back post injury.

It’s also worth noting that each organization has their own philosophy regarding player development and how they want to utilize their players at the MLB level. The Rays, for example, want their RP to pitch multiple innings out of the bullpen, which is fairly unconventional compared to the rest of the league. Thus, we can expect their prospects to see more multi-inning outings to prepare them for this that we wouldn’t expect from prospects in other systems. These philosophies are kept behind closed doors for the most part which make them hard to quantify but could be useful if you’re building a similar projection model for a team. Implementing these philosophies into the model can give a more holistic view of each prospect within the farm system and be informative to a front office.

Finally, when looking at the models, all of them put a high emphasis on team ranking when determining their projected stat line or the likelihood they would make the MLB. This makes sense, as better prospects are going to be ranked higher and thus are more likely to perform better. However, this naively assumes that each team’s farm system is on the same level, which is simply not true (i.e. it assumes that the #2 prospects on each team are on the same talent level, which is not necessarily the case). For instance, the Cubs have 7 prospects on MLB’s top 100 rankings while A’s have 1. However, the top 30 rankings suggest that Cade Horton (the Cubs #2 prospect) is on the same talent level as Colby Thomas (A’s #2 prospect) even though Horton is ranked #48 on the top 100 rankings while Colby Thomas is unranked. Furthermore, the top 30 rankings suggest that Jefferson Rojas (Cubs #7 prospect) is worse than Colby Thomas even though Jefferson Rojas is ranked #93 on the top 100 rankings. To get a better representation of each player’s talent level, all prospects need to be ranked on the same scale. Scouting reports can be used to do this; however, scouting reports are only available for each team’s top 30 prospects, making it unsuitable for this type of analysis where all minor league prospects are considered.

IX. Conclusions, Discussions, & Future Work

This paper offers a new perspective into scouting prospective talent and how we can expect them to transition throughout the minor league system. When paired with conventional wisdom from scouts, it can help give better context into where a given player is in his developmental process and should give front offices more insight into who has a better long-term trajectory into the MLB. In future developments, we can expand this system to project how a player might perform multiple levels above when compared to his current baseline stats (for instance, we could try to predict how a player might perform in AA or AAA based on their level A stats).

There could also be some expansion in this project to investigate the temporal trends of a given player. A minor leaguer's performance in his last 100 PA provides a more informative story into how a batter might perform at the next level compared to 100 PA before that point, something this current model doesn’t consider. Some time series analysis on these trends could offer some more informative insights into how a player can be expected to perform at the next level rather than just taking their career statistics at the level before promotion. These predictions will also get better as more and more data gets published from the minor league levels. Ball tracking data for AAA has started to be published on Baseball Savant starting in the 2024 season. This will give us more access to underlying stats such as exit velocity, barrels, pitch movement/stuff, etc. As access to these stats becomes more accessible & available, projecting future performance will become much more reliable & accurate.

However, the overall goal of this project was to help bridge the gap between objective & subjective scouting across baseball's minor league system, and I think this offers a good stepping stone into improving this process in the future.

X. References

8 common types of neural networks. Coursera. (2024, April 23). https://www.coursera.org/in/articles/types-of-neural-networks

Baheti, P. (2021, May 27). *Activation functions in neural networks [12 types & use cases]*. V7. https://www.v7labs.com/blog/neural-networks-activation-functions

Brennan, P. (2021, December 31). *Quantifying player development in the minor leagues*. Wordpress. https://patrickbrennan33.wordpress.com/2021/12/31/quantifying-player-development-in-the-minor-leagues/

Brownlee, J. (2022, August 16). *Your first deep learning project in python with keras step-by-step*. MachineLearningMastery.com. https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

Chandler, Gabriel & Stevens, Guy. (2012). An Exploratory Study of Minor League Baseball Statistics. Journal of Quantitative Analysis in Sports. 8. 10.1515/1559-0410.1445.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-aoas285

Cross, J., & Mailhot, J. (2020, May 20). *Fangraphs prep: Build and test your own projection system*. FanGraphs. https://blogs.fangraphs.com/fangraphs-prep-build-and-test-your-own-projection-system/

Danovitch, J. (2019). Trouble with the curve: Predicting future MLB players using scouting reports. *arXiv preprint arXiv:1910.12622*.

Deng, D. (2020, September). DBSCAN clustering algorithm based on density. In *2020 7th international forum on electrical engineering and automation (IFEEA)* (pp. 949-953). IEEE.

Druschel, H. (2016, February 22). *A guide to the projection systems*. Beyond the Box Score. https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun

Foley, J. (2020, February 28). *Statistical Projections and you, part 1*. Twinkie Town. https://www.twinkietown.com/2020/2/28/21148335/mlb-baseball-statistical-projections-primer-zips-steamer-marcel-pecota-minnesota-twins-2020

Gallagher, M. (2019). *A Better Predictor of NFL Success: Collegiate Performance or the NFL Draft Combine?*(Order No. 28268674). Available from ProQuest Dissertations & Theses Global. (2456872688). https://libezproxy.syr.edu/login?url=https://www.proquest.com/dissertations-theses/better-predictor-nfl-success-collegiate/docview/2456872688/se-2

Gow, Alexander, "Using Machine Learning to predict MLB success Based on MILB performance" (2019). *IPHS 300: Artificial Intelligence for the Humanities: Text, Image, and Sound.* Paper 18. https://digital.kenyon.edu/dh_iphs_ai/18

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software*, *91*(1), 1–30. https://doi.org/10.18637/jss.v091.i01

Heaton, C., & Mitra, P. (2023). *Learning contextual event embeddings to predict player performance in the MLB*. MIT Sloan Sports Analytics Conference. https://www.sloansportsconference.com/research-papers/learning-contextual-event-embeddings-to-predict-player-performance-in-the-mlb

Hendrawan, I. R., Utami, E., & Hartanto, A. D. (2022). Comparison of naïve bayes algorithm and XGBoost on local product review text classification. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 143-149.

How do soccer players perform after transferring to different teams and leagues?. DataBuckets. (2015, May 7). https://databuckets.org/databucket/how-do-soccer-players-perform-after

Hussey, A. (2021). *NBA Draft Analysis* (Doctoral dissertation, Dublin, National College of Ireland).

Lee, CH., Lee, Wj. (2023). Baseball Informatics—From MiLB to MLB Debut. In: Sharma, V., Maheshkar, C., Poulose, J. (eds) Analytics Enabled Decision Making. Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-19-9658-0_5

Mitchell, C. (2014, December 30). *Katoh: Forecasting Major League hitting with minor league stats*. The Hardball Times. https://tht.fangraphs.com/katoh-forecasting-a-hitters-major-league-performance-with-minor-league-stats/

Ng, K. (2017). Analyzing Major League Baseball player's performance based on age and experience. *Journal of Sports Economics & Management*, 7(2), 78-100.

Nguyen, Q., & Matthews, G. J. (2024). Filling the gaps: A multiple imputation approach to estimating aging curves in baseball. *Journal of Sports Analytics*, *10*(1), 77-85.

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

Sing, C. L. (2022, November 29). *From Japan to MLB: The players before and after signing: Sports analytics group at Berkeley*. From Japan to MLB: The Players Before and After Signing | Sports Analytics Group at Berkeley. https://sportsanalytics.studentorg.berkeley.edu/articles/japan-to-mlb.html

Slowinski, P. (2010, February 25). *Projection Systems*. Fangraphs. https://library.fangraphs.com/principles/projections/

Young, W., Holland, W. & Weckman, G. (2008). Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network. *Journal of Quantitative Analysis in Sports*, *4*(4). https://doi.org/10.2202/1559-0410.1131

Zhang, J., Li, Y., Tian, J., & Li, T. (2018, October). LSTM-CNN hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (pp. 1675-1680). IEEE.

Zhang Z. (2016). Naïve Bayes classification in R. *Annals of translational medicine*, *4*(12), 241. https://doi.org/10.21037/atm.2016.03.38

XI. Appendix

- Github Page with Code & Any Additional Work to the Project
- Data Engineering
  - For players outside the top 100 ranking I imputed a value of 110
  - For players outside the top 30 ranking, I imputed a value of 35
  - To create the metric-specific stat for each of the projection models, I used a fairly simple that follows the following logic: $\frac{Evidence}{(Sample\ Size)^x}$. For example, for AVG, "Evidence" would represent "Hits" & "Sample Size" would represent "AB". Using this formula, you can put a weight on counting stats & rate stats by changing the value of "x" (the exponent). For example, if you set x to 0, you would get a pure counting stat (ie Hits) & if you set x to 1, you would get a pure rate stat (ie AVG). For this project, I put more weight on rate stats, but the amount of weight I put on rate stats varied by the type of stat that was used. If a higher value was better (ie AVG), I set x to 0.8. However, if a lower value is (ie ERA), I set x to 0.6 and included the raw rate stat as a supplement. The only stat that didn't follow this logic was SwStr% since I didn't have swing-by-swing counting data, just the pure rate stat.
- Pitchers (GAM Model):
  - $Made\ MLB = \beta_1(Top\ 100\ Rank) + \beta_2(Top\ 30\ Rank) + \beta_3(Batters\ Faced) + \beta_4(Games\ Started) + \beta_5(K\%) + \beta_6(GB\%) + \beta_7(BB\%) + \beta_8(SwStr\%) + \beta_9(FIP) + s(Age) + \varepsilon$
- Batters (GAM Model):

- $Made\ MLB = \beta_1(Top\ 100\ Rank) + \beta_2(Top\ 30\ Rank) + \beta_3(PA) + \beta_4(wRC+) + \beta_5(K\%) + \beta_6(GB\ per\ FB) + \beta_7(BB\%) + \beta_8(SwStr\%) + s(Age) + \varepsilon$

- Pitcher Made MLB XGBoost
  - Covariates: Average Top 100 Ranking (Baseball America & MLB), Top 30 Team Ranking, Age, Games Started, ERA, FIP, GB%, K%-BB%, SwStr%
- Batter Made MLB XGBoost
  - Covariates: Top 100 Ranking (Baseball America & MLB), Top 30 Team Ranking, Age, wRC+, PA, BB%-K%, Pull%, GB_per_FB, SwStr%

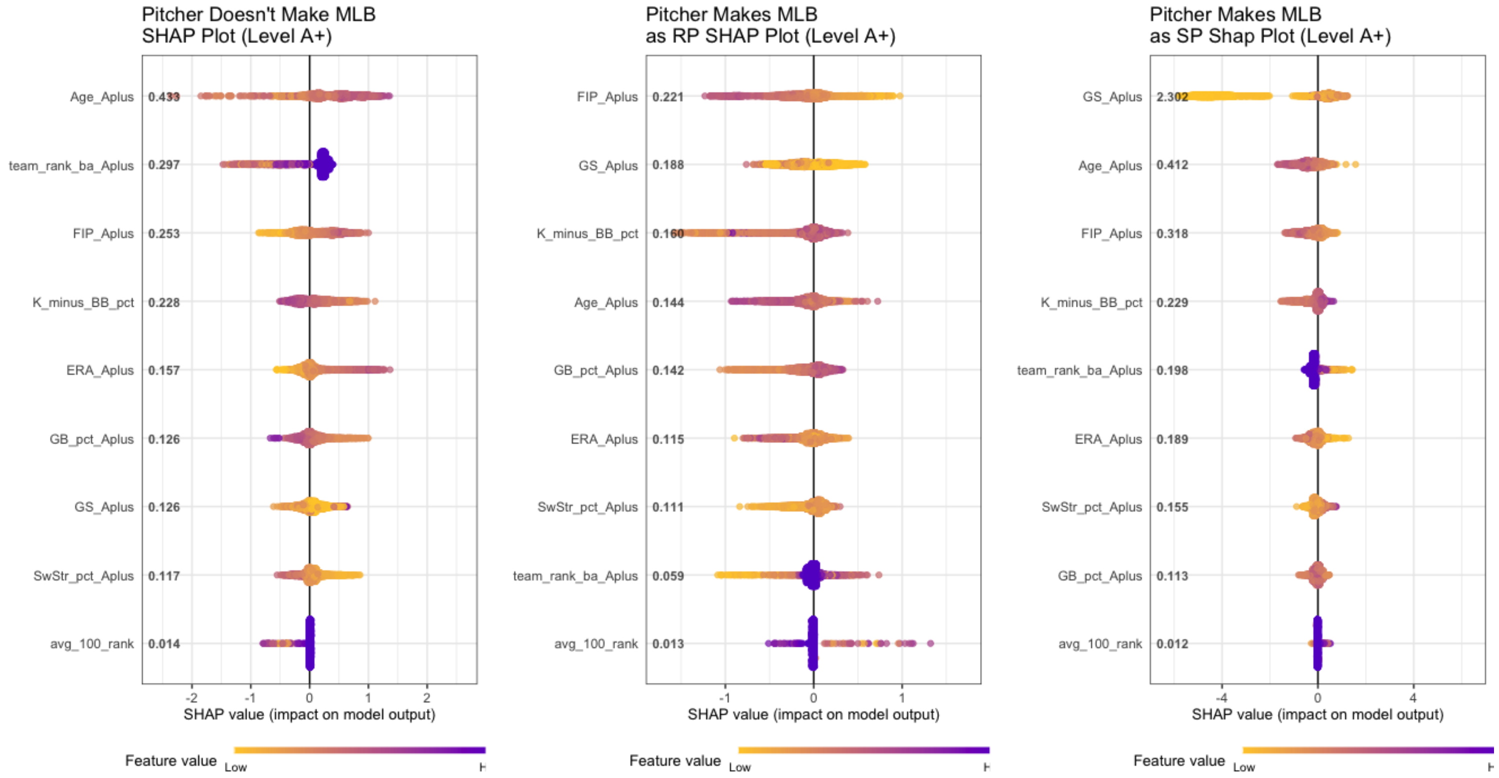


Figure 14: SHAP plot of pitcher’s probability to make the MLB (A+)

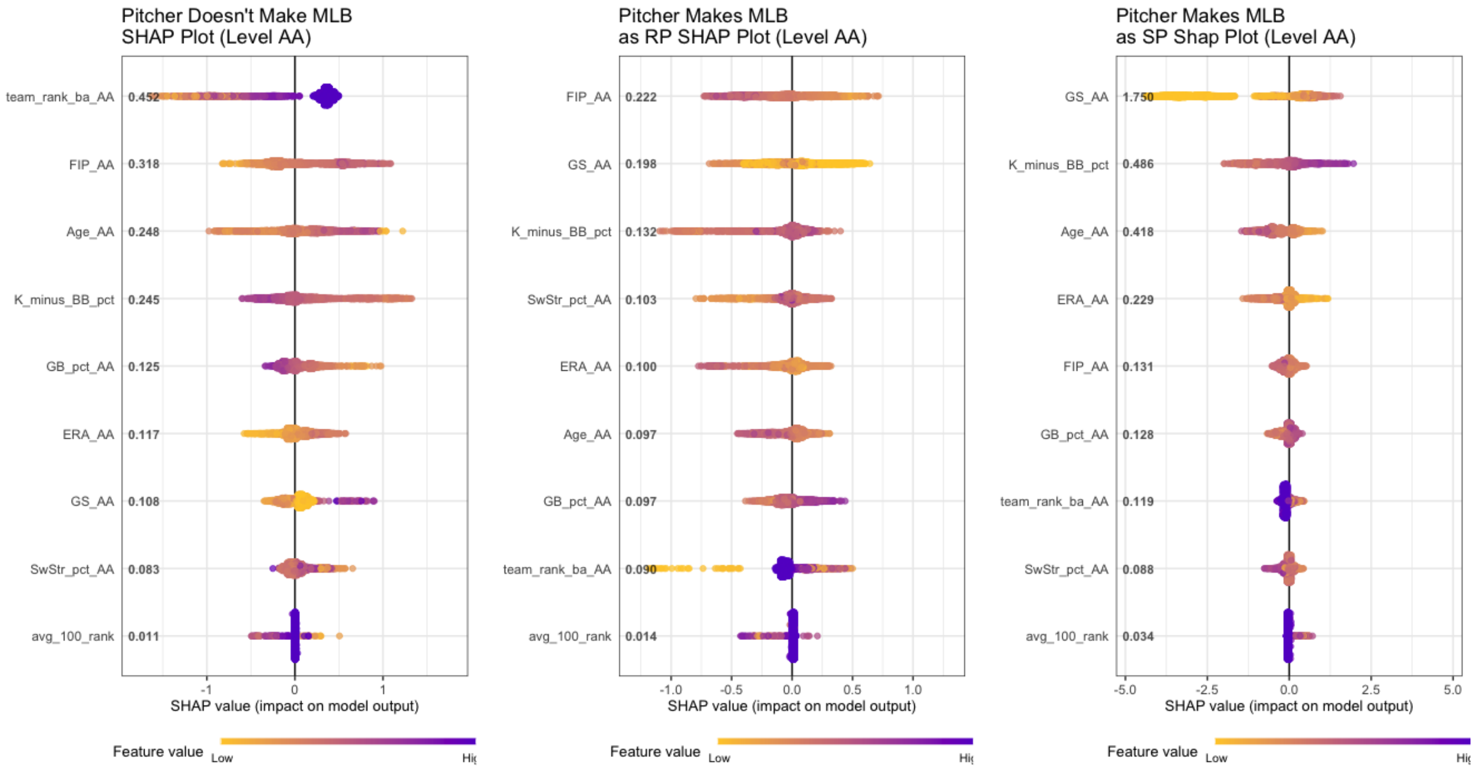


Figure 15: SHAP plot of pitcher’s probability to make the MLB (AA)