New York Mets

Midseason Talent Acquisition Strategy

Initial Findings Report

Prepared for Don Wedding, GM
August 5, 2018

New York Mets Analytics

Alexander Booth, Justin Benson, Noah Lieberman, Michael Pallante, Thomas Popeck Spiller

Table of Contents

Problem Statement

Dear Prof. Wedding,

As the GM of the New York Mets, we are pleased to provide you with our Initial Findings report for the Midseason Talent Acquisition Strategy. As described in our Project Goals report, the purpose of this project is to provide ideas and recommendations for developing the future on-field product of the Mets through a strengthening of youthful talent and potential in the minor leagues. As the trade deadline is approaching, we believe it is time to sell off underperforming veteran assets at the Major League level.

The Seattle Mariners have expressed interest in several major league players including infielders Asdrubal Cabrera and Todd Frazier, outfielders Jose Bautista and Jay Bruce, and pitchers Noah Syndergaard and Jacob Degrom. Using a robust data infrastructure from a combination of internal databases along with reputable supplementary sources, we plan to develop predictive models to forecast the likelihood of Mariner minor league talent reaching the Major Leagues and sustaining success. Specifically, we are looking to add offensive talent due to the current poor run production. To do so, we will focus solely on the likelihood of position players making the MLB and their projected offensive WAR (wins above replacement).

Using our model output, we will be able to identify players to target and cross reference against how the Mariners rank them according to external prospect rankings. We will provide initial recommendations along with dashboard and mobile applications to provide your front office team with trade scenario modeling techniques.

We look forward to receiving your feedback and helping return the Mets to contention.

Sincerely,

New York Mets Analytics

Alexander Booth **Justin Benson** **Noah Lieberman**
Michael Pallante **Thomas Popeck Spiller**

Description of Data

To develop models that follow our Midseason Talent Acquisition Strategy, we will utilize the data sources referenced on the chart in the "Appendix: Data Sources Chart" section of this document. Using The Baseball Cube's dataset as our primary source (as it contains the primary playerID key), we joined the additional databases to create a single dataset for modeling.

To further elaborate on the joining of our data sources, it starts with using the WAR tables available on Baseball Reference. From here, we are able to scrape the data and then link it to Lahmans' MLB Database, thus allowing us to calculate WAR in the Lahmans' MLB Database. This was done through the matching of indexes pulled from the Lahmans' MLB Database with the Baseball Reference indexes. The primary tool for used for this exercise is R, which is used to parse this data and then load the data into the database.

We then must map this data to The Baseball Cube's minor league data. This will be implemented through text matching by player names. Once completed, we are able to build models using this data set.

With the acknowledgement of our goal to add offensive talent to our organization due to our currently poor run production at the major league level, we will be specifically analyzing the batting data from our now updated and revamped Baseball Cube dataset. We will not be analyzing the pitching or fielding data that we have available to us, as it does not align with our objective. It is also important to note that we will also be removing the batting statistic of pitchers, which will not be of much help to solving our offensive struggles. We will provide a further overview of the data that is being analyzed in the next section.

The type of basic, traditional batting statistics that will be explored as we move forward are outlined by The Baseball Cube. To review detailed descriptions of these batting statistics that will be referenced, please visit the following link below to The Baseball Cube statistics glossary:

http://www.thebaseballcube.com/about/stats_glossary.asp

Additionally, we will be employing the use of some advanced batting statistics. To review expanded descriptions of these batting statistics, as well as how they are calculated, please access the following link below to the MLB advanced statistics glossary:

http://m.mlb.com/glossary/advanced-stats

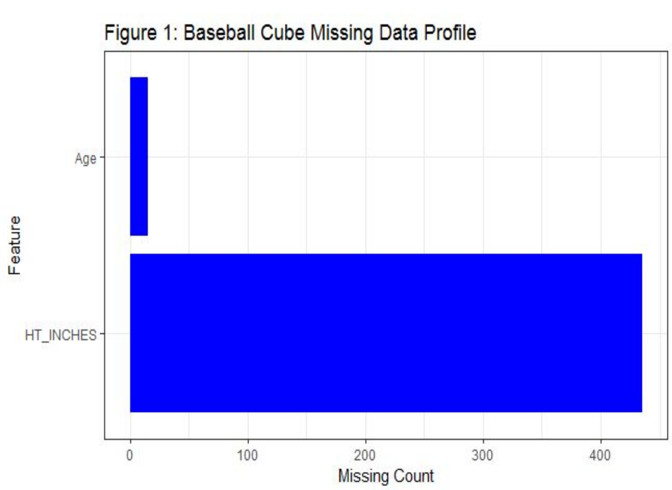
Please see the "Appendix: Statistics & Terminology" section of this document to review the key advanced batting statistics we will be focusing on for our Midseason Talent Acquisition Strategy, including WAR, wOBA, wRC+, OPS+, and wRAA.

We will then define a player as having "made it" as any player that stayed in the majors for any significant time (currently defined at 3 or more seasons, although we may change this based on further data exploration or executive directive). This will create a binary variable which will be defined as 1 for "made it" and 0 for "Didn't Make it". Additionally, we will cluster on MLB Players stats to generate a profile for the type of player they are forecasted to be. We will using the following standard WAR grouping to assess player's forecasted role within the MLB.

Overview of the Data

The New York Mets Analytics team has performed a comprehensive exploratory data analysis (EDA) of the minor league batting information in preparation of developing the models to predict likelihood of a prospect making the MLB and their cumulative wins-above-replacement (WAR). The data will be analyzed in two main groups. First, whole MiLB prospect data that is available, representing available prospect history and useful for establishing the likelihood of a prospect having an MLB career or not. The second group is the subset of prospects that have made the MLB and spent enough time in the Major Leagues to have WAR generated for them, which will be necessary to model the predict WAR for potential trade targets.

The primary tools used for the EDA are R and Alteryx. Both softwares provide the analyst a broad array of EDA tools, in R the DataExplorer package was used, along with base level functionality and in Alteryx, we were able to use its simple user interface to summarize, pivot and view the data throughout the transformation process, as well as generate correlation matrices.



To focus on the offensive side of the ball, we look only at the batting data available from The Baseball Cube. This data includes over 185,000 observations, representing 42,722 different minor league players from 1977 through 2017. We remove the batting statistics for pitchers, as they will not be the solution to the Mets current offensive struggles, leaving 156,589 observations for 32,566 position players in the minor leagues.

This data is very full, and as shown in figure 1, very few of the potential model inputs are missing. Of the standard variables provided, only age and height have any missing information. Missing age information will be backed out using the players' birthday if available. Height is missing for fewer than 500 of the 32,500 potential players and those players will be dropped from the sample of inputs for simplicity. Several players have errors in their height information (e.g., 5'91") for which manual corrections were applied.

In addition to the missing data, there are two variables that were not collected until later within the sample period. Intentional walk counts (IBB) were not available until 1981, ground-into-double-play (GDP) was not collected until 1991. To address this data problem, we evaluated model performance with and without the variables to give statistical justification to limiting the sample to the period with full data or a smaller time-frame with more robust information.

It is not uncommon for players to move around the minor leagues, either via trade, changing the organization they are being developed by, or via promotion/demotion to a higher or lower level of play. Of the 156,000 observations, there are roughly 120,000 observations based on distinct player-seasons, that is to limit the data to only one observation for each player in each year of baseball, and 23% of all player-seasons involved a prospect playing in multiple levels of minor league ball. It is uncommon for a prospect to develop much faster than that, and only 2.5% of all player-seasons involved a prospect playing in three or more separate levels.

There are several categorical variables that we will be accounting for. There are 6 different levels of minor league baseball, Rookie (Rk), A-, A, A+, AA, AAA. Among these levels, there are 20 different leagues (see Appendix for list). The competition is generally similar within each level, but specific contexts can vary greatly, for example, as shown in figure 2, the number of runs per 100 at-bats is different across the various Rookie leagues, and the Pioneer League (Pio) has more than 2 additional runs per 100 at-bats than the next closest league. Next we will flag outfield and infield players defined by their primary position. Depending on the organization need, we can isolate the type of fielder best suited to help the team.

Figure 2: Rookie Leagues Runs per 100 At-Bats

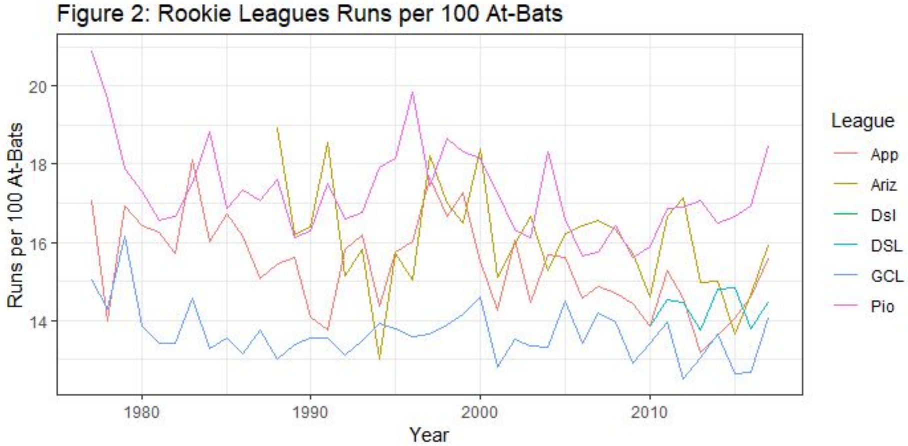
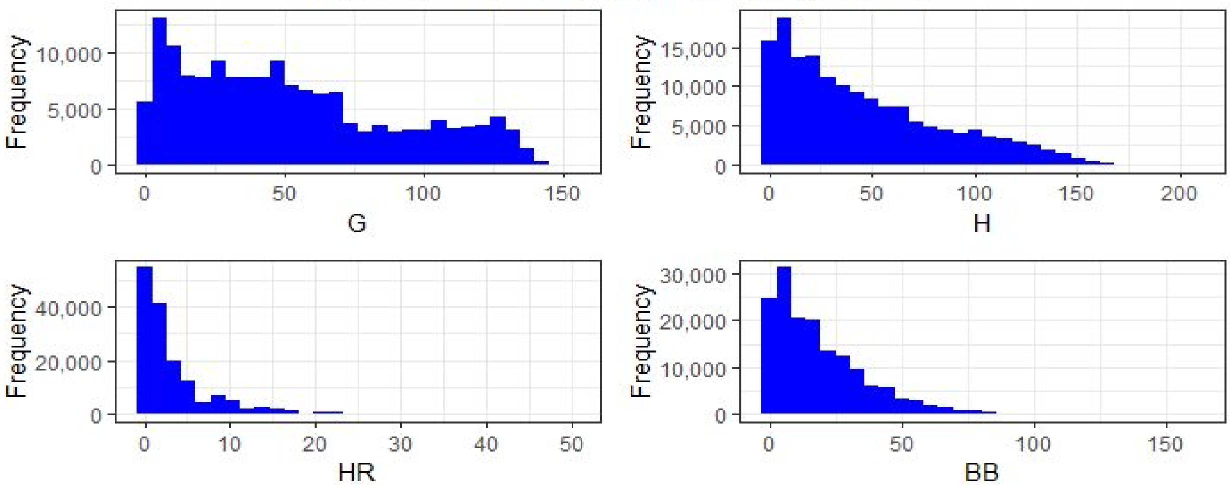
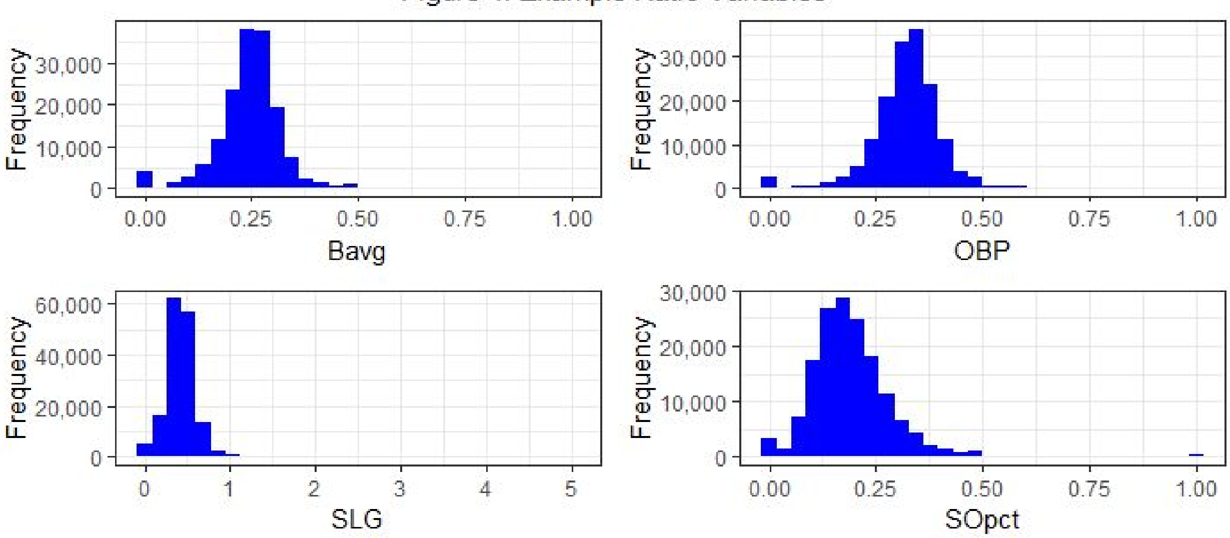


Figure 3: Example Right Skewed Counting Variables



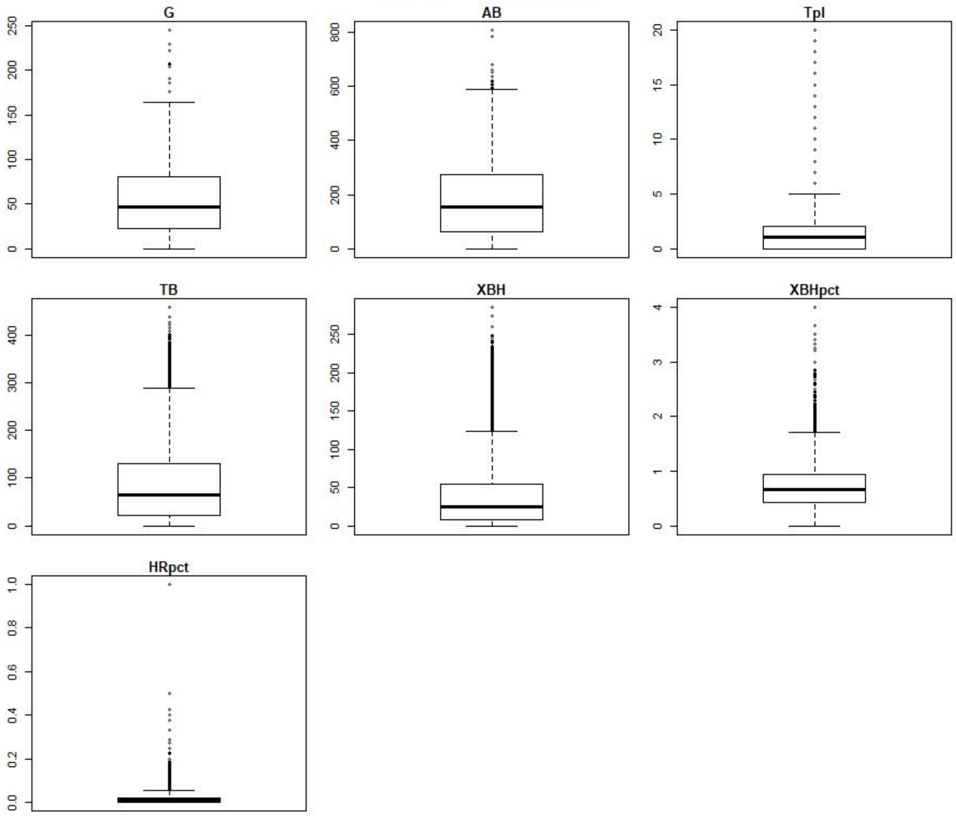
As shown in figure 3, many of the counting statistics, such as at-bats (AB), hits (H), home-runs (HR), and even walks (BB) are skewed right. This is not surprising as many of the counting statistics are heavily related to the number of games played. Figure 4, on the next page, shows that many of the ratio style variables are normally distributed, including batting average (Bavg), on base percentage (OBP), slugging percentage (SLG), and strike-out percentage (SOpct). We perform various transformations on this data, first creating advanced statistics such as weighted on-base average (wOBA) including the percentage variables listed above and secondly by taking the log of several variables to smooth the distribution. This will be further explored in the Description of Data Transformation portion.

Figure 4: Example Ratio Variables



After dealing with the skewness of the data, we are left with outliers. Figure 5 shows a select set of variables that the team has identified as meaningful to limit the range of the data. These variables showed large ranges relative to the meaningful differences in the statistics, and the outliers will be dealt with either through removal or winsorization.

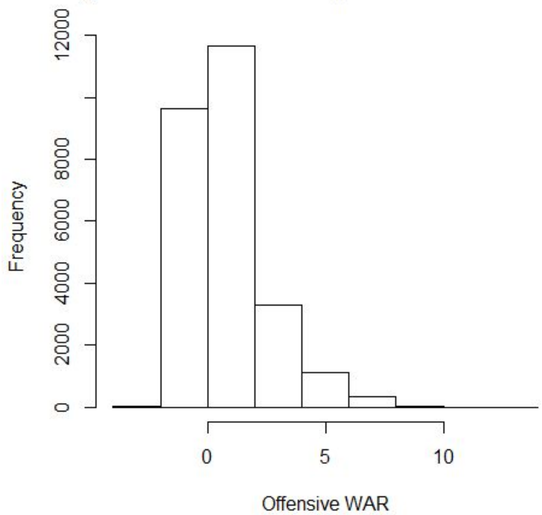
Figure 5: Select Outlier Boxplots



Of the 32,566 position players available in the data, almost 14,000 players have played at least one game in the major leagues, but many of those are injury call-ups or other short stints where a player should not be considered an impactful player for the organization. To address the non-impact of these players, we will be modeling based on whether or not a player has spent more than three years in the MLB (to have “made it”), this tends to give the player enough at-bats to be no longer considered an MLB rookie. The number of position players that have made it is only 2,700, far fewer than the players who have been on an MLB roster.

As suggested in the initial project document, we defined MLB players based on their WAR value. We group the players by Offensive WAR on a season by season basis for position players between 1977 and 2017, using the Baseball-Reference WAR data. The gives us WAR information for 26,000 batters in the MLB. As shown in figure 6, the season by season WAR is right skewed and largely between -2 and 2 WAR on the season. The data is also effectively bounded on the lower end, as no team is likely to carry a player who is showing significant harm to the team’s success. Between 1977 and 2017, only 6 players had a season WAR below -2.

Figure 6: Wins Above Replacement Distribution



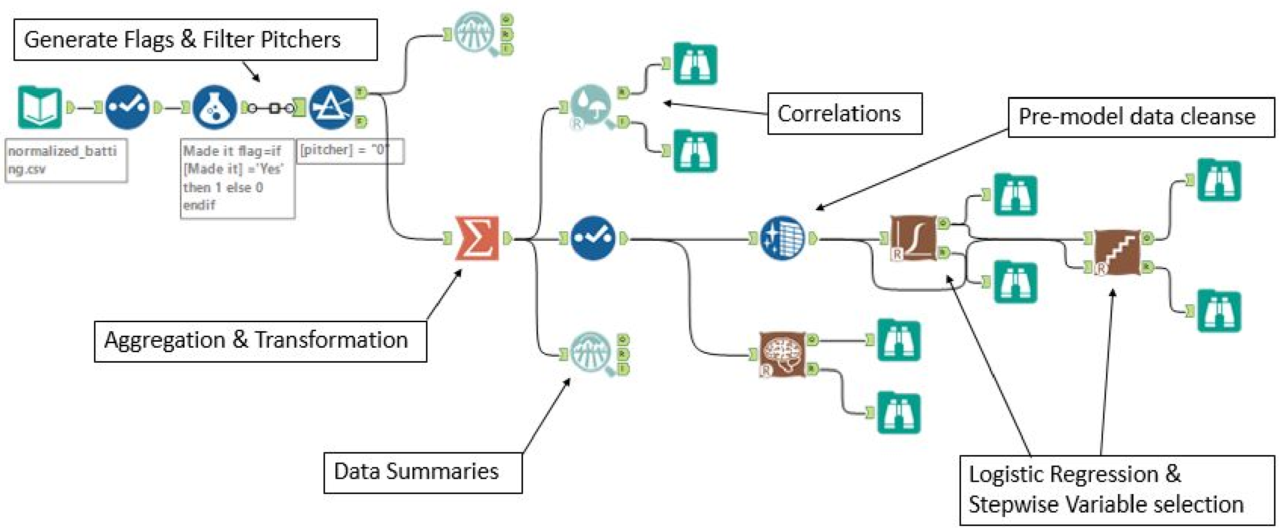
To better understand these players, we consider the MLB players using the categories in the following table to describe their peers, based on the offensive WAR. Players are grouped into 8 categories on a season by season basis. Groupings are based on the average number of players (Implied Players) per year within that band of WAR, aligning the number of All-Star or better players is in-line with the average number of position players on the All-Star rosters.

Player Value	WAR	Implied Players
Scrub	<-.25	106
Replacement Player	-0.25 to 0.25	228
Role Player	0.25 to 1	117
Solid Starter	1 to 2.5	112
Good Player	2.5 to 4	54
All-Star	4 to 6.5	32
Superstar	6.5 to 7.5	4
MVP	7.5+	2

Description of Transformation of Data

Initially, very few variables needed imputation. The next step in the data process was to determine the level of granularity for which data should be examined. Initial data came in at the player, league, level, year granularity, so a player could have played in multiple leagues or multiple levels in a given year. Additionally, they then could and would often play in multiple years in the minors before making the majors. Ideally, we want our model to be based off a granularity that makes the most sense for predicting MLB success, so we want to be able to judge a player based on their level statistics, as well as their season-by-season stats. To get to these measures, we do two things. First, we aggregate, average or sum, depending on the metric, performance for each metric we were tracking. When done by level this creates a career stat, and additionally we calculated statistics if the player moves leagues on a season-by-season basis. Next, we pivot out the career-level metrics so as to create a column for each metric at each level - this created a huge number of variables to look at, approximately 600, including variables which would track the amount of years in each level, as well as a flag for appearing at a level, as some players would skip levels.

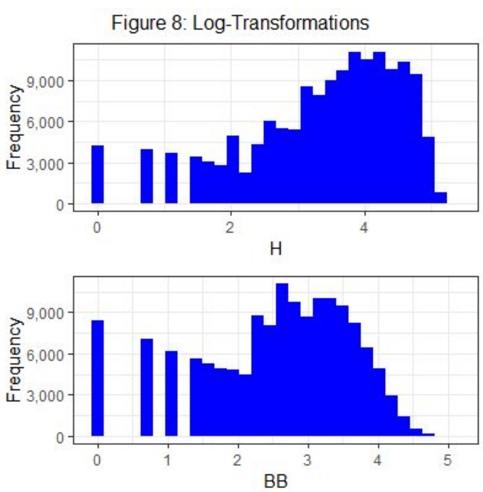
We used a tool called Alteryx to perform these initial transformations. It allowed us to quickly create workflows which could be scaled within the organization to other MiLB analysis, or transitioned to non coding resources very quickly. An example of the workflow can be seen below:



This workflow example generates summary statistics for our population, as well as allowing us to run correlation analysis, cleanse the data, and do initial logistic and neural net models with stepwise selection. It is super easy to alter these tools and add a variety of R based models for preliminary modelling before taking it into R.

All in all, this work in Alteryx created the grain where each row of our data was a player ID, and their stats were represented in a large list of columns that would be used as variables for predicting their likelihood of making the MLB. However, this also created a very large number of nulls, specifically if the player didn't play at a certain level, they would not have any data for those columns. To handle this, we used a quick R script which would transform any NULLs to 0's and create a flag to identify that that variable had been imputed with a new value. Additionally, if there were any variables which were all NULL or 0's, we would eliminate them based on insignificant impact on analysis.

As mentioned in the description of data, there are skewed variables which should be addressed. We will explore models with and without the transformations, assessing the value the transformations. Figure 8 shows how two of the transformed variables appear after transformation, and the skew is less significant, though we may still need do something else.



Another transformation being tested is the per-game statistic. When used in conjunction with the time spent in each minor league this can better compare players.

Analysis of Data

In this section of our report, we will analyze exploratory data findings that aim to provide critical insights to inform the modeling process.

Total Players Count

Figure 9 below displays the total number of unique players at the Major (blue) and Minor (red) League levels. From this plot, we are able to see that there has been vast increases in the number of MiLB players, while the MLB player pool has displayed only minor growth. Based on these trends, we would expect the average likelihood to reach the Major Leagues to decrease over time.

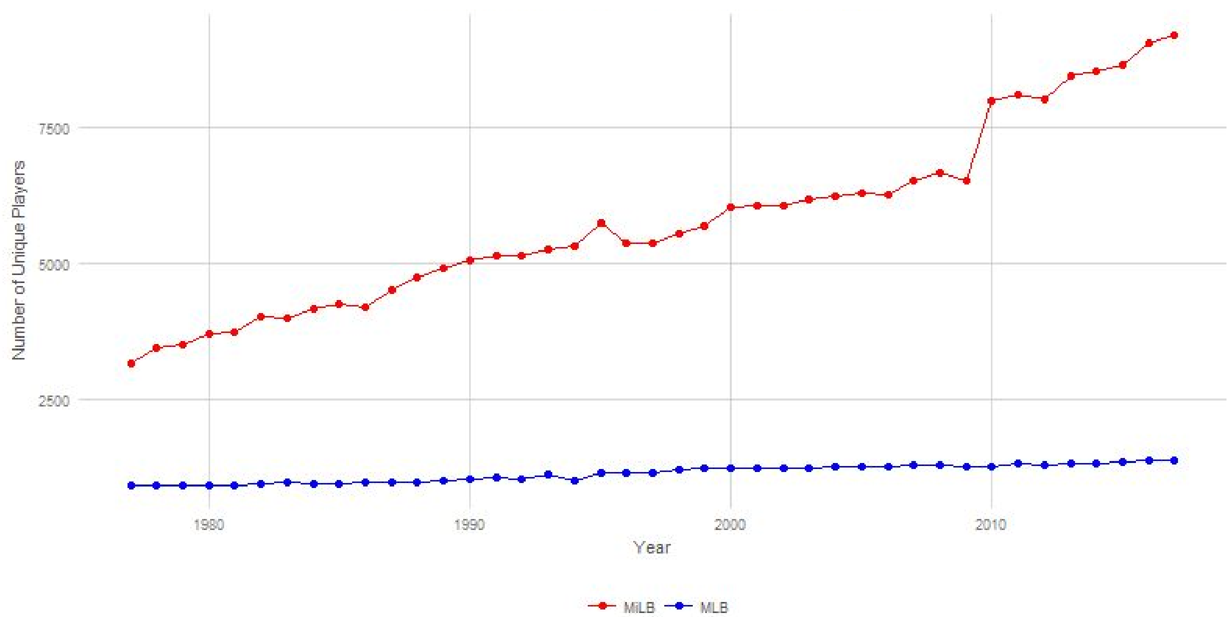


Figure 9: Unique MiLB & MLB players per year

Correlation Analysis to Response Variable

Next, correlation analysis between predictor statistics and the response variable, “Made It” (likelihood to reach the Major League level) were assessed. the following are key takeaways from the correlation plot in Figure 10:

- *Age* is the strongest predictor of a prospect being called up to the Major Leagues, however it is potentially misleading. In rookie ball, it is possible that older players are more likely to reach the majors because they are recent high college draft picks becoming accustomed to professional culture. In AAA, older players could be fringe talents that rotate up and down from the Majors. At other levels, the age factor is weaker.

- Among on-field production statistics, there are a few combination offensive statistics (*OPS, wOBA, SLG*) that are the strongest positive predictors of likelihood to make it to the Majors. These statistics grow stronger as prospects move up the Minor League ranks, meaning the production is more significant as they move closer to the Major Leagues.
- Striking out (*K_BB, Kpct*) is the strongest negative indicator of likelihood to make the Major Leagues. This is logical as players that make less contact at the Minor League level will be less likely to sustain success at the Major League level.

Correlation (R) to "Made It" (Making it to the Major League Level)

Statistic	Rk	A-	A	A+	AA	AAA	Overall
Age	0.20	0.03	0.01	0.15	-0.07	0.15	**0.25**
OPS	0.13	0.17	0.19	0.20	0.22	0.21	**0.16**
wOBA	0.13	0.17	0.18	0.20	0.22	0.21	**0.16**
SLG	0.13	0.16	0.17	0.19	0.21	0.21	**0.16**
SecA	0.12	0.15	0.17	0.18	0.20	0.20	**0.15**
OBP	0.11	0.15	0.17	0.17	0.18	0.17	**0.14**
Bavg	0.12	0.15	0.17	0.17	0.18	0.15	**0.14**
ISO	0.12	0.13	0.14	0.16	0.19	0.19	**0.14**
wRAA	0.07	0.14	0.18	0.14	0.21	0.21	**0.13**
HRpct	0.09	0.09	0.10	0.11	0.14	0.17	**0.10**
BABIP	0.06	0.10	0.11	0.11	0.11	0.09	**0.09**
XBH	0.01	0.08	0.14	0.05	0.15	0.17	**0.08**
Homeruns	0.02	0.07	0.12	0.05	0.14	0.16	**0.08**
TB	0.00	0.07	0.14	0.04	0.15	0.17	**0.08**
Runs	-0.01	0.07	0.13	0.05	0.14	0.17	**0.08**
XBHpct	0.06	0.06	0.06	0.08	0.12	0.14	**0.07**
IBB	0.04	0.06	0.10	0.08	0.11	0.12	**0.07**
RBI	-0.01	0.07	0.13	0.03	0.14	0.17	**0.07**
Doubles	-0.01	0.07	0.13	0.03	0.13	0.15	**0.07**
Hits	-0.02	0.06	0.13	0.03	0.13	0.15	**0.07**
Triples	0.01	0.06	0.11	0.07	0.09	0.11	**0.06**
SB	0.00	0.04	0.11	0.06	0.12	0.10	**0.06**
BBpct	0.03	0.06	0.08	0.08	0.09	0.11	**0.06**
BB	-0.03	0.04	0.09	0.02	0.11	0.15	**0.05**
SF	-0.01	0.04	0.09	0.02	0.08	0.13	**0.05**
CS	-0.02	0.03	0.10	0.02	0.10	0.10	**0.04**
PA	-0.05	0.03	0.10	-0.01	0.10	0.13	**0.04**
At-Bats	-0.05	0.02	0.10	-0.01	0.10	0.13	**0.04**
HBP	-0.04	0.02	0.05	-0.01	0.06	0.09	**0.02**
GDP	-0.05	0.02	0.07	-0.02	0.05	0.10	**0.02**
Games	-0.07	0.00	0.07	-0.04	0.05	0.09	**0.01**
K	-0.09	-0.03	0.03	-0.06	0.04	0.09	**-0.01**
AB_HR	-0.04	0.01	0.00	-0.05	0.00	0.02	**-0.01**
SH	-0.05	-0.05	0.00	-0.05	0.00	0.01	**-0.02**
K_BB	-0.09	-0.08	-0.09	-0.14	-0.10	-0.09	**-0.09**
Kpct	-0.12	-0.12	-0.13	-0.15	-0.12	-0.12	**-0.11**

Figure 10: Correlation (R) to "Made It"

Scouting Report Word Cloud

A word cloud based on scouting reports was created as shown in Figure 11 below. Words such as “Speed, strength, arm, hands, quick, frame, plus, and build” jump out immediately as characteristics that scouts are looking for. This indicates that it may make sense to also add body type metrics such as BMI into our modelling to help include some of the scouting knowledge that normal stats will miss. Additionally, the prevalence of words like “average” and “aboveaverage” indicate that more logical processing may be needed to determine what feature is being described this way.

Figure 10: Scouting Report Word Cloud

Historical WAR Classification Assessment

Based on the WAR classification of players defined in the Overview of the Data, the following table shows the top 3 players by number of seasons within the top WAR bands. The listed players are well known and this is consistent with the Mets’ understanding of the traditional offensive value of these MLB players.

MVP	Superstar	All-Star
Alex Rodriguez (8)	Derek Jeter (5)	Manny Ramirez, Paul Molitor, Rickey Henderson (13)
Barry Bonds (7)	Albert Belle, Albert Pujols, Mike Schmidt, Todd Helton (4)	Dave Winfield, Lou Whitaker, Rafael Palmeiro (12)
Mike Trout (5)	Nine Players (3)	Eight Players (11)

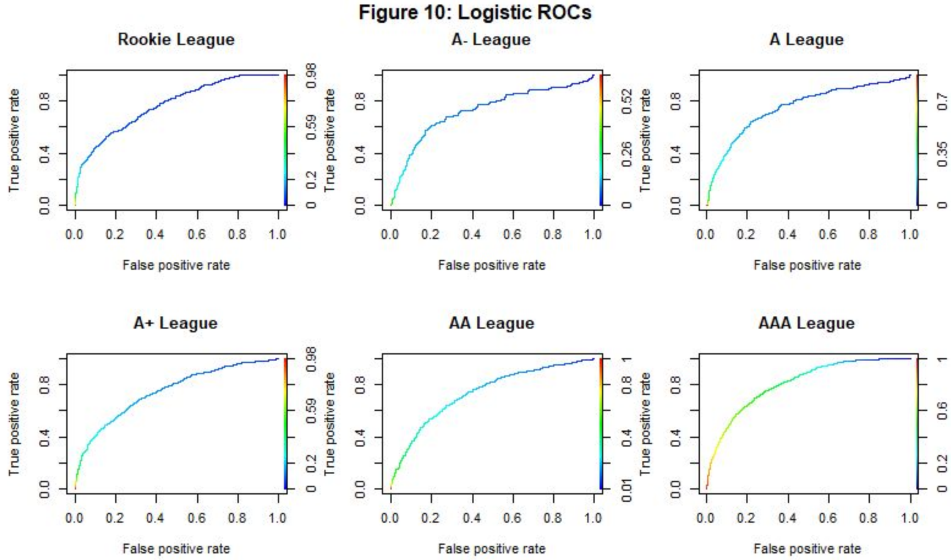
Logistic Model Findings

The first model explored is a logistic model aimed at identifying if a MiLB prospect will play in the MLB for three or more years (Made.it). Six models are generated independently for each level of the minors, Rookie (Rk), Low-A (A-), A, High A (A+), AA, and AAA. For the initial models, we start the analysis data in 1996, grabbing the post strike[1] information, and stopped in 2014, to allow time for the prospects we model on to have three years of time to have a valid Made.it value by 2017, our most current data. The sample data is broken out in to an 80/20 training and validation data, and the 2017 prospect information is used to generate potential trade targets. When combined with the WAR prediction model, we can best identify prospects that will have the most successful careers in the future.

Logistic models are generated using the per-game data, each observation is a season in the specific level of the minors. A model is generated for each level of play, using the basic batting statistics (G, R, H, Dbl, Tpl, HR, RBI, SB, CS, BB, IBB, SO, SH, SF, HBP, GDP), Age, and variables for the specific league with the minor league level, such as the American League (Amer) of the AAA level. If a player played in more than one level in a year, he would be included in any level played in.

The coefficients for the logistic models are available in the appendix, and figure 10 shows the ROC curve. There are some confusing results, such as in the AAA model, ground-into-double-play is a positive indicator of making it to the MLB, we would not expect such a negative batting result, generating two outs, would be a good thing. Interactions will have to be further explored.

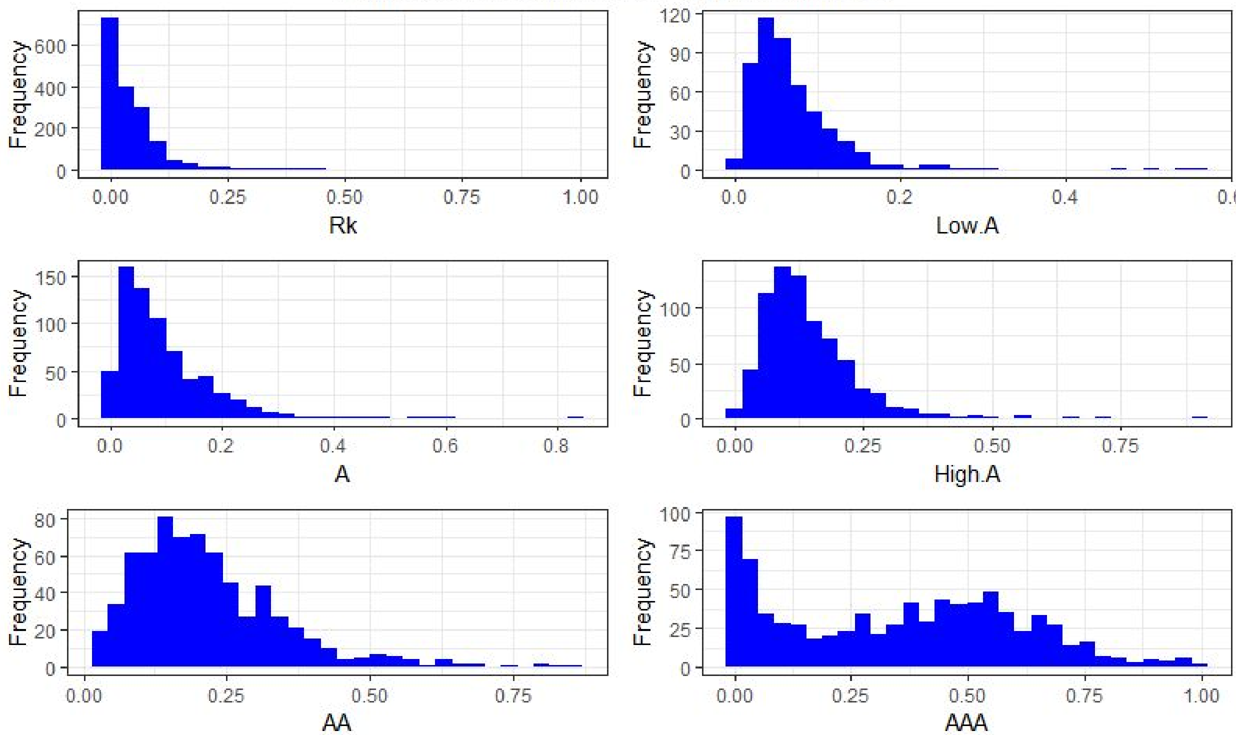
Figure 10: Logistic ROCs



[1] https://en.wikipedia.org/wiki/1994%E2%80%9395_Major_League_Baseball_strike

With the logistic results for each league, we predict the likelihood of all the 2017 MiLB prospects to have at least a three year MLB career. The results are shown in figure 11, as generated from the validation data. As expected, the Rookie League players have the lowest overall likelihood, however there are a few players that stand out compared to their competition and have high probabilities of making it. As we move up in the levels, the distribution becomes less skewed and has a higher central tendency. One caveat is that AAA has more players who have long careers while never making it to the majors than the lower leagues, and so the spike at 0 has more to do with AAA-lifers than traditional prospects with low-likelihoods.

Figure 11: Histogram of 2017 Prospect Likelihoods



The model could be improved by including advanced statistics or other information, building the models based on career as opposed to season statistics, and using other statistical learning techniques.

WAR Prediction Findings

Our next step comprised of attempting to predict the career contributions, or total value, of a player based on their minor league data. We created a simple random forest comprised of an 80-20% train-test split including 500 trees, folded and clustered 25 times.

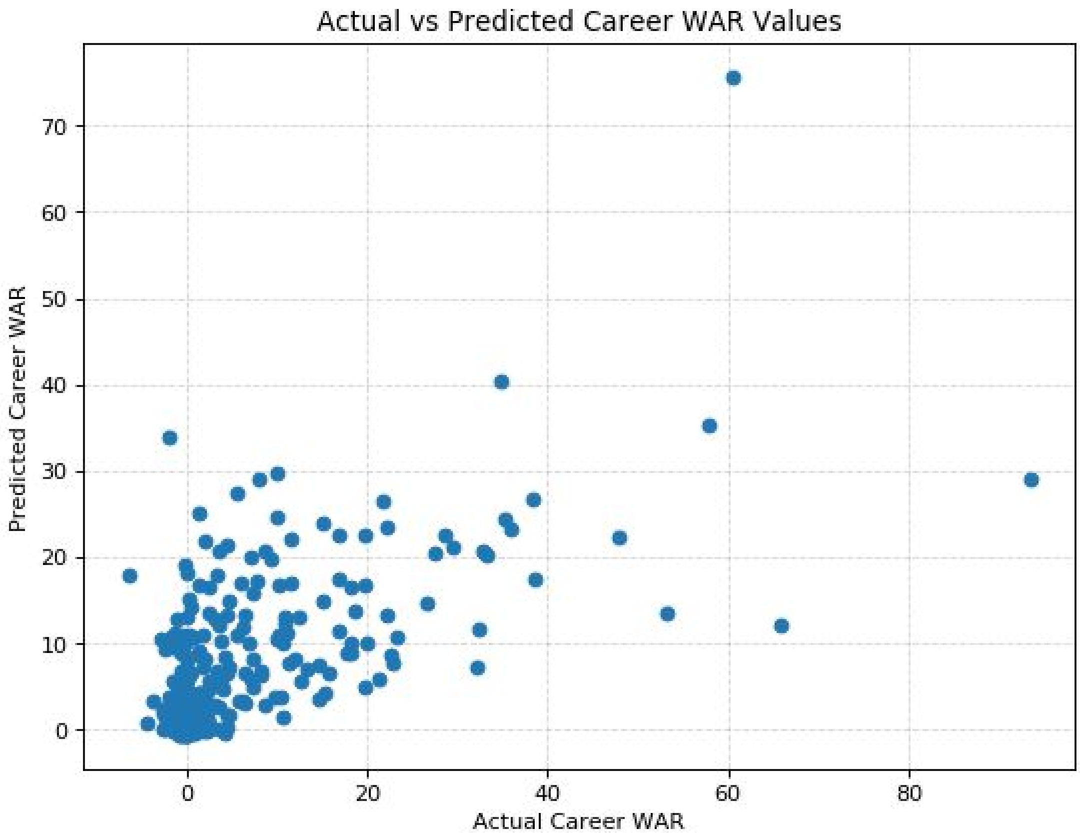
The initial data set of minor league players only consisted of around 5,000 players, around five percent of our entire sample. This small sample definitely provided some drawbacks in our model initially, mainly bias and a higher risk of over-fitting. However, our initial model seemed to provide fairly robust predictions.

To begin with, we created a rudimentary baseline of predicted WAR to ensure our model was providing reasonable predictions. We calculated this baseline by simply averaging the actual career WAR in the training set. This value ended up being around 5.1 WAR. This baseline created an absolute error in our validation set of 6.9. If our random forest could not beat this absolute error from simply predicting the average for every data point, then we would need to rethink our methodology.

Luckily, our random forest provided better predictions than the baseline, in terms of a lower absolute error. Our model had an average absolute error of 4.34, almost 65% lower than the baseline. Unsurprisingly, the AAA player statistics had the most impact in the model. Perhaps surprising to some, however, was the importance of the average age in high A of the player. Below lists the top 15 features in order of importance:

```
Variable: AAA_Avg_GDP_norm     Importance: 0.08
Variable: AAA_Avg_Age          Importance: 0.05
Variable: AAA_Avg_AB_HR_norm   Importance: 0.04
Variable: AAA_Avg_IBB_norm     Importance: 0.04
Variable: AAA_Sum_G            Importance: 0.02
Variable: AAA_Sum_K_BB         Importance: 0.02
Variable: AAA_Sum_SOpct        Importance: 0.02
Variable: High_A_Avg_Age       Importance: 0.02
Variable: AAA_Avg_AB           Importance: 0.01
Variable: AAA_Avg_Bavg         Importance: 0.01
Variable: AAA_Avg_Bavg_norm    Importance: 0.01
Variable: AAA_Avg_IBB          Importance: 0.01
Variable: AAA_Avg_K_BB_norm    Importance: 0.01
Variable: AAA_Avg_OPS_norm     Importance: 0.01
Variable: AAA_Avg_orgs_norm    Importance: 0.01
```

Next, we created a scatter plot of predicted career WAR versus the actual career WAR in our validation set. Perfect predictions would be located on the diagonal. If there was a pattern in the error, then we would also have to rescale some of our variables and retrain the model in order to eliminate bias. In this case, there was no pattern in the residuals.



Our model does a great job of predicting career WAR of less than 20. Unfortunately, it seems like outliers are not predicted as robustly as they could be. Most of the outliers are conservatively predicted, however, they are still predicted as “all-stars” or “above-average”. Since these WAR predictions should be used to bucket, or estimate, the potential worth of a player, a conservative prediction also allows the Mets, a typically risk-averse organization, to quantify and minimize risk in a prospect. Our model identifies these players as having a high career WAR, however, just not as high as their actual career WAR.

To improve this model, we may investigate a similar strategy to the logistic model. Since most of the important features are statistics from AAA, players in rookie ball and low A have poorly predicted career MLB WAR. By creating a unique model per level, similar to the logistic model, we would be able to mitigate this issue. Additionally, we expect that more trees and more folding on the training set would reduce the absolute mean error in the predictions. Finally, we hope to test other models, including a neural net for example, to see if we can improve the absolute mean error in the validation set further. While we wish to have more data in our training and validation sets, we know that the small conversion rate of MiLB players making the MLB makes this wish impossible.

Preliminary Conclusions & Deliverables

Logistic Models

From the logistic models, we have generated a list of 55 prospects who have a greater than 75% likelihood to stay in the majors for three or more years. Focusing our scouting departments' energy on these players would yield more benefit if we can make a beneficial trade with their organization. Additionally, we rank the 111 available Mariner prospects and the Mets farm system in the appendix, as they have expressed interest in a trade.

The table below shows some of the top prospects in the Mets' minor league system, as ranked by SBNation (Other Ranking). Each numerical column represents the likelihood of a 3+ year career, as measured by the prospects' 2017 MiLB batting statistics. As you can see, we are lacking in top flight prospects, particularly at first base (1B). In the Mariner's appendix table, we see that Seth Mejias-Brean is a 1B/3B with a 62% chance of success. Mejias-Brean should be a primary trade target to bolster our farm as he has a high likelihood of making the majors, and would be our most likely prospect to make the majors in either position.

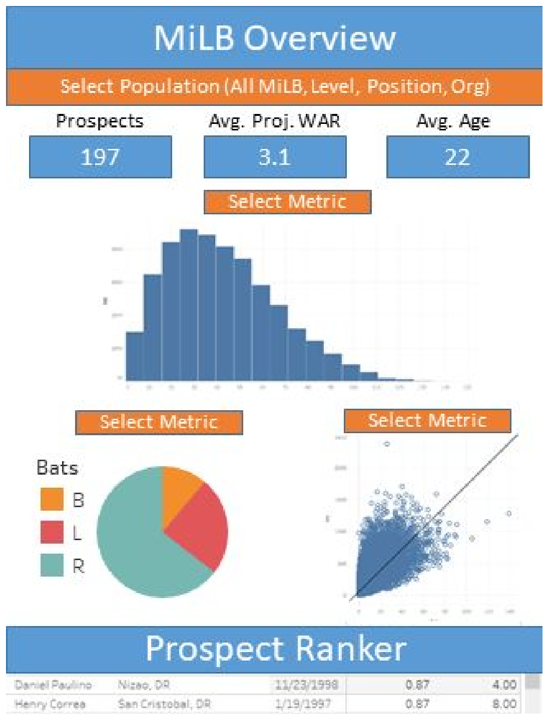
Player Name	Position	Rookie	A-	A	A+	AA	AAA	Other Ranking
Luis Guillorme	SS					27%		B-
Peter Alonso	IF				23%	52%		B-/B
Andres Gimenez	SS			21%				B/B+
Jhoan Urena	3B				22%		43%	C+
Tomas Nido	C					19%		C+
David Thompson	OF					21%		C+
Desmond Lindsay	OF			8%				C+/B-

By interacting the likelihood of making the MLB for 3 or more years and the anticipated career WAR, we can refine trade targets based on a Likely-WAR measure. As the WAR model is further developed, we can expand the expected the likely-WAR to better improve the Mets' farm system.

Dashboard Creation

To aid in the management's use of these models, we are developing a set of dashboards to report prospect statistics and value. As shown in the MiLB Overview wireframe below, we will show the available prospect pool as shown by a variety of metrics. The orange boxes represent places where management will be able to drill down on information about the specific prospects. At the bottom will be a filtered list of prospects with relevant information.

Management will also be able to drill down on an individual prospect and see more granular information about that player. As shown in the player details tab, it contain the basic information, projected likelihood of making the MLB, projected WAR, and a tag for identifying what type of player he is expected to be.



Project Status

The New York Mets' Analytics team remains on track for key deliverables and milestones according to our timeline as displayed below. Upon completion of initial findings in this report, we will shift focus towards finalizing the Midseason Talent Acquisition Strategy project over the next 4 weeks. In addition to the Final Report and Executive Summary, our team will provide an interactive dashboard and mobile application for the front office team to examine trade possibilities.

NYM Analytics Midseason Talent Acquisition Timeline:

APPENDIX: Statistics & Terminology

The Baseball Cube Statistics Glossary (Traditional Batting Statistics)

http://www.thebaseballcube.com/about/stats_glossary.asp

MLB Advanced Statistics Glossary (Advanced Batting Statistics)

http://m.mlb.com/glossary/advanced-stats

Key Advanced Batting Statistics:

- Wins Above Replacement (WAR): “WAR measures a player's value in all facets of the game by deciphering how many more wins he's worth than a replacement-level player at his same position (e.g., a Minor League replacement or a readily available fill-in free agent).

  For example, if a shortstop and a first baseman offer the same overall production (on offense, defense and the basepaths), the shortstop will have a better WAR because his position sees a lower level of production from replacement-level players” (MLB Advanced Media).

- Weighted On-base Average (WOBA): “wOBA is a version of on-base percentage that accounts for how a player reached base -- instead of simply considering whether a player reached base. The value for each method of reaching base is determined by how much that event is worth in relation to projected runs scored (example: a double is worth more than a single).

  For instance: In 2014, a home run was worth 2.101 times on base, while a walk was worth 0.69 times on base. So a player who went 1-for-4 with a home run and a walk would have a wOBA of .558 -- (2.101 + 0.69 / 5 PAs)” (MLB Advanced Media).

- Weighted Runs Created Plus (wRC+): “wRC+ takes the statistic Runs Created and adjusts that number to account for important external factors -- like ballpark or era. It's adjusted, so a wRC+ of 100 is league average and 150 would be 50 percent above league average.

  For example, a player who plays his home games at hitter-friendly Coors Field will have a lower wRC+ than a player who posts identical stats at pitcher-friendly O.co Coliseum. The production of the player at Coors Field is deemed less impressive because of his ballpark's hitter-friendly nature” (MLB Advanced Media).

- On-Base Plus Slugging Plus (OPS+): "OPS+ takes a player's on-base plus slugging percentage and normalizes the number across the entire league. It accounts for external factors like ballparks. It then adjusts so a score of 100 is league average, and 150 is 50 percent better than the league average.

  For example, Miguel Cabrera's .895 OPS in 2014 was 50 percent better than the MLB average after being adjusted for league and park factors. As a result, his OPS+ was 150" (MLB Advanced Media).

- Weighted Runs Above Average (wRAA): "wRAA measures how many runs a hitter contributes, compared with an average player -- so a player with a 0 wRAA would be considered league average, offensively. It's calculated by finding the difference in the number of runs contributed between a player and the league average (which is determined by the league average wOBA).

  Because wRAA uses wOBA to determine how many runs a player is worth, a player with an above-average wOBA will have an above-average wRAA. But -- unlike wOBA -- wRAA is a counting stat. As a result, players with a higher number of plate appearances can accrue a higher wRAA than an equal player with fewer plate appearances" (MLB Advanced Media).

APPENDIX: Data Sources Chart

NYM Analytics Data Sources:

Source Name	Description	Location	Acquisition
Fangraphs	Minor League Player batting and pitching data from 2006-Current	https://www.fangraphs.com/minorleaders.aspx	Download
Baseball Reference	Minor League Player batting and pitching data from 1977 - 2017	https://www.baseball-reference.com/register/	Web scraping
The Baseball Cube	Major League Data Player batting, fielding and pitching data from 1865 - 2017 & Minor League batting and pitching data from 1977- 2017	http://www.thebaseballcube.com	Download
Lahmans' MLB Database	Major League Data Player batting, fielding and pitching data from 1865 - 2017	http://www.seanlahman.com/baseball-archive/statistics/	Download
The Baseball Prospectus	Scouting reports for recent prospects	https://legacy.baseballprospectus.com/prospects/eyewitness.php	Web scraping
Sentiment Analysis - word list	List of words implying positive/negative sentiment analysis, will be augmented to include "baseball terms"	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html	Download

Primary Data Source

APPENDIX: Logistic Results

	Dependent variable: MLB Career >= 3 Years					
	Rookie	A-	A	A+	AA	AAA
G	0.001	-0.004*	0.003***	-0.003***	-0.01***	-0.01***
	-0.002	-0.002	-0.001	-0.001	-0.001	-0.001
R	0.04	0.7**	-0.1	0.6***	0.7***	0.5***
	-0.2	-0.3	-0.3	-0.2	-0.2	-0.2
H	0.9***	1.6***	1.8***	1.0***	1.7***	1.7***
	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1
Dbl	0.5**	0.5	1.5***	0.6**	1.1***	0.5**
	-0.3	-0.4	-0.4	-0.3	-0.3	-0.2
Tpl	-0.1	0.9	2.5**	2.1***	0.5	1.2*
	-0.5	-1	-1	-0.7	-0.7	-0.7
HR	1.5***	-0.3	4.1***	1.8***	2.5***	2.3***
	-0.4	-0.8	-0.7	-0.5	-0.5	-0.4
RBI	0.2	0.4	-0.04	0.1	-0.2	0.2
	-0.2	-0.3	-0.3	-0.2	-0.2	-0.2
SB	1.0***	0.01	1.8***	1.6***	2.1***	2.0***
	-0.2	-0.4	-0.3	-0.3	-0.3	-0.3
CS	-1.1*	1.3	-0.8	-0.5	-1.2**	0.1
	-0.6	-0.9	-0.8	-0.5	-0.6	-0.5
BB	0.7***	1.1***	1.7***	0.9***	1.4***	1.4***
	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1
IBB	2.3**	2.7	3.3**	4.6***	2.0**	1.9**
	-1	-2.1	-1.5	-1	-0.9	-0.9
SO	-0.8***	-0.7***	-1.3***	-0.8***	-0.5***	0.2**
	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1
SH	-2.9***	-2.3*	-1.3	-3.1***	-1.8***	-1.8***
	-1	-1.3	-1	-0.7	-0.6	-0.5
SF	2.6***	0.4	1.8*	1.6**	2.0***	1.9***
	-0.5	-1.2	-1	-0.7	-0.7	-0.5
HBP	-0.3	0.7	0.7	0.2	0.1	1.2***
	-0.4	-0.6	-0.6	-0.3	-0.5	-0.4
GDP	-0.1	-0.5	0.5	0.7**	0.6	0.9***
	-0.5	-0.7	-0.7	-0.3	-0.4	-0.3
Age	0.2***	-0.1*	-0.2***	0.1***	-0.02*	0.2***
	-0.01	-0.03	-0.02	-0.01	-0.01	-0.01
Observations	14,038	6,260	8,935	10,084	9,307	12,856
Log Likelihood	-2,835.40	-1,474.20	-2,635.10	-4,091.90	-4,536.70	-6,734.70
Akaike Inf. Crit.	5,714.80	2,986.30	5,308.30	8,223.90	9,115.40	13,511.30
Note:	*p<0.1; **p<0.05; ***p<0.01					

	Dependent variables League Dummies					
	MLB Career >= 3 Years					
	Rookie	A-	A	A+	AA	AAA
DSL	-2.6***					
	-0.4					
Pio	-0.2*					
	-0.1					
App	-0.3**					
	-0.1					
GCL	0.5***					
	-0.1					
NYPL		0.2				
		-0.1				
Midw			0.01			
			-0.1			
Calif				-0.4***		
				-0.1		
Caro				-0.3***		
				-0.1		
East	-0.01				-0.01	
	-0.1				-0.1	
SAL	-0.3				-0.3	
	-0.2				-0.2	
Tex	-0.1				-0.1	
	-0.1				-0.1	
Amer						0.5***
						-0.2
IL						0.3***
						-0.04
Mex						-2.5***
						-0.1
Constant	-6.7***	-3.2***	-0.4	-5.0***	-2.8***	-6.8***
	-0.4	-0.7	-0.5	-0.3	-0.3	-0.2
Observations	14,038	6,260	8,935	10,084	9,307	12,856
Log Likelihood	-2,835.40	-1,474.20	-2,635.10	-4,091.90	-4,536.70	-6,734.70
Akaike Inf. Crit.	5,714.80	2,986.30	5,308.30	8,223.90	9,115.40	13,511.30

Note: *p<0.1; **p<0.05; ***p<0.01

APPENDIX: Mets' Prospect Made.it Likelihood Tables

The following tables show the likelihood of the Mets' prospects, based on their 2017 statistics in each level of the minor leagues in which a player appeared and the model for that level and have appeared in less than two MLB seasons. Prospects may be called up to replace injured players while not being ready for a full time position, limiting the number of years they have appeared in the MLB helps clarify a prospect versus a player who has successfully graduated to the major leagues.
In the tables each player is placed at their highest minor league level, for example, Jeff McNeil (2B), a player who maxed out in AAA in the minors while playing in A+ during the same year, will only appear in the table below. He had a 30% and 59% likelihood of having a three year career in the MLB, based on the respective models. He was able to improve his relative chance of success against higher competition, and that bodes well for his future.

AAA 2017 Mets' Prospects							
Player Name	Position	Rookie	A-	A	A+	AA	AAA
Travis Taijeron	OF						70%
Jeff McNeil	2B				30%		59%
Amed Rosario	SS						57%
Josh Rodriguez	3B						57%
Cody Decker	1B					28%	54%
Xorge Carrillo	C						54%
Dominic Smith	1B						51%
L.J. Mazzilli	2B					24%	45%
Jhoan Urena	3B				22%		43%
Jayce Boyd	1B						39%
Phillip Evans	SS						35%
Victor Cruzado	OF						34%
Jeffrey Glenn	C						30%
John Mora	OF				18%		29%
Jio Mier	SS					18%	26%
Dale Burdick	SS				10%	14%	16%
Arnaldo Berrios	CF			3%	10%		12%

AA 2017 Mets' Prospects						
Player Name	**Position**	**Rookie**	**A-**	**A**	**A+**	**AA**
Peter Alonso	IF				23%	52%
Patrick Mazeika	C-1B				21%	44%
Luis Guillorme	SS					27%
Kevin Taylor	2B					24%
Kevin Kaczmarski	OF					23%
David Thompson	OF					21%
Colton Plaia	C					21%
Matt Oberste	1B					20%
Jean Rodriguez	3B-SS				11%	19%
Tomas Nido	C					19%
Champ Stuart	OF					17%
Patrick Biondi	OF					15%
Gustavo Nunez	2B-SS					14%
Tyler Moore	C-IF					9%

A+ 2017 Mets' Prospects					
Player Name	**Position**	**Rookie**	**A-**	**A**	**A+**
Anthony Dimino	C	15%		83%	23%
Tim Tebow	OF			1%	21%
Nick Sergakis	3B				20%
Jacob Zanon	CF			17%	17%
Eudor Garcia	IF				15%
Michael Paez	SS			19%	15%
Ian Strom	CF		6%	17%	13%
Jeff Diehl	1B-RF				13%
J.J. Franco	SS			6%	12%
Wuilmer Becerra	RF-OF				11%
Colby Woodmansee	IF	10%		1%	11%
Daniel Rizzie	C			4%	10%
Leon Byrd	2B		7%		9%
Jose Garcia	C				8%
Enmanuel Zabala	OF				7%
Vinny Siena	IF			1%	7%
Victor Moscote	DH	4%			4%

A 2017 Mets' Prospects				
Player Name	**Position**	**Rookie**	**A-**	**A**
Andres Gimenez	SS			21%
Luis Carpio	SS			17%
Dash Winningham	1B			10%
Desmond Lindsay	OF			8%
Gene Cone	OF			8%
Oliver Pascual	SS	3%		8%
Ali Sanchez	C			7%
Ricardo Cespedes	OF	4%	4%	7%
Milton Ramos	SS			6%
Blake Tiberi	IF			6%
Brandon Brosher	OF			5%
Reed Gamache	IF		9%	5%
Jay Jabs	3B			5%
Natanael Ramos	C	6%		2%

A- 2017 Mets' Prospects			
Player Name	**Position**	**Rookie**	**A-**
Walter Rasquin	1B		13%
Matt Winaker	OF		11%
Wagner Lagrange	OF	12%	10%
Scott Manea	C		7%
Jeremy Vasquez	1B-OF	10%	5%
Jeremy Wolf	OF		5%
Guillermo Granadillo	OF	14%	4%
Edgardo Fermin	SS	6%	4%
Jose Maria	C		3%
Dylan Snypes	SS		3%
Franklin Correa	2B		3%
Carlos Sanchez	C		3%
Cecilio Aybar	SS		2%
Carl Stajduhar	C-IF		2%

Rookie League 2017 Mets' Prospects		
Player Name	**Position**	**Rookie**
Rigoberto Terrazas	3B	11%
Matthew Foley	C	10%
Domingo Martinez	C	10%
Luis Montero	3B	9%
Raphael Gladu	OF	8%
Dionis Paulino	OF	7%
Hansel Moreno	SS	6%
Kenneth Bautista	OF	6%
Juan Uriarte	C	6%
Kenny Hernandez	SS	5%
Ranfy Adon	OF	5%
Kevin Hall	C	5%
Raul Beracierta	OF	5%
Mark Vientos	3B	4%
Grabiel Jimenez	OF	4%
Luis Lebron	C	4%
Danny Hoy	IF	4%
Angel Manzanarez	SS	4%
Gregory Guerrero	SS	4%
Edinson Valdez	OF	3%
Jack Schneider	CF	3%
Robby Kidwell	C	3%
Anthony Dirocie	OF	3%
Gavin Garay	SS	3%
Yeffry De Aza	IF	2%
Cristopher Pujols	3B	1%
Yoel Romero	SS	1%

Rookie League 2017 Mets' Prospects		
Player Name	**Position**	**Rookie**
Wilmer Reyes	SS	1%
Luis Santana	2B	1%
Jhoander Saez	OF	1%
Shervyen Newton	SS	1%
David Lozano	2B	1%
Jorge Martinez	C	1%
Anderson Bohorquez	SS	1%
Pedro Ventura	SS	1%
Juan De La Rosa	OF	0%
Jose Peroza	3B	0%
Yordin Araujo	OF	0%
Wilfred Astudillo	C	0%
Sebastian Espino	SS	0%
Kevin Torres	C	0%
Alexis Marquez	3B	0%
Andres Regnault	C	0%
Jean Carlos Soto	OF	0%
Moises Gonzalez	OF	0%
Jeison Rodriguez	OF	0%
Tulio Garcia	OF	0%
Eulises Sanchez	OF	0%
Gilberto Espinoza	OF	0%
Julio Rene	OF	0%
Wilmy Valdez	1B	0%
Alejandro Medina	DH	0%
Ezequiel Pena	OF	0%
Rafael Valdez	SS	0%
Jose Mena	C	0%

APPENDIX: Mariners' Prospect Made.it Likelihood Tables

The following tables show the likelihood of the Mariners' prospects, based on their 2017 statistics in each level of the minor leagues in which a player appeared and the model for that level and have appeared in less than two MLB seasons. Prospects may be called up to replace injured players while not being ready for a full time position, limiting the number of years they have appeared in the MLB helps clarify a prospect versus a player who has successfully graduated to the major leagues.

In the tables each player is placed at their highest minor league level, for example, Kevin Santa (SS), a player who maxed out in AAA in the minors while playing in Rookie League, and A+ during the same year, will only appear in the table below. He had a 20%, 10%, and 55% likelihood of having a three year career in the MLB, based on the respective models. He was able to improve his relative chance of success against higher competition, and that bodes well for his future, the drop in A+ ball might be explained by limited time in the league.

AAA 2017 Mariners' Prospects							
Player Name	Position	Rookie	A-	A	A+	AA	AAA
Boog Powell	OF						71%
Danny Muno	2B						66%
Seth Mejias-Brean	1B-3B					20%	62%
Mike Marjama	C						58%
Kevin Santa	SS	20%			10%		55%
Ian Miller	OF					51%	54%
Joey Wong	SS-3B					14%	53%
D.J. Peterson	3B						46%
Andrew Aplin	OF	13%					46%
Tyler O'Neill	OF						44%
Dario Pizzano	OF					36%	40%
Steven Baron	C					11%	37%
Tyler Smith	SS						36%
Logan Taylor	IF	8%			10%		31%
Zach Shank	2B-SS						29%
Alexander Capriata	C						26%
Austin Grebeck	OF		4%		8%		26%
Eugene Helder	SS		12%				15%
Ryan Scott	C			4%	3%	38%	8%
Brayan Hernandez	OF	6%	7%				7%
Sebastian Valle	C	6%					6%
Joseph Rosa	SS		16%	44%		19%	4%
Gianfranco Wawoe	2B				11%	34%	4%
Rayder Ascanio	SS			8%	8%		3%

AA 2017 Mariners' Prospects						
Player Name	Position	Rookie	A-	A	A+	AA
Braden Bishop	OF				21%	62%
Chuck Taylor	OF					31%
Ryan Casteel	C					23%
Chris Mariscal	SS				19%	18%
Tyler Marlette	C					18%
Keury De la cruz	OF					18%
Chantz Mack	OF					17%
Marcus Littlewood	C					17%
Kyle Petty	1B					16%
Nelson Ward	IF					14%
Jay Baum	IF-OF				20%	14%
Willie Argo	OF				23%	11%
Justin Seager	1B-3B					11%
Brock Hebert	SS-2B					9%
Adam Law	OF-3B					7%

A+ 2017 Mariners' Prospects					
Player Name	Position	Rookie	A-	A	A+
Eric Filia	OF				27%
Joey Curletta	RF-OF				15%
Donnie Walton	SS	38%			14%
Kyle Lewis	OF	7%			13%
Jordan Cowan	2B-SS				12%
Ricky Eusebio	OF				11%
Luis Liberato	OF			9%	10%
Joe DeCarlo	3B				10%
Arturo Nieto	C				9%
Daniel Torres	C				7%
Joe Rizzo	3B			17%	4%

A 2017 Mariners' Prospects				
Player Name	**Position**	**Rookie**	**A-**	**A**
Luis Rengifo	2B			23%
Anthony Jimenez	OF	5%		16%
Johan Quevedo	C			11%
Bryson Brigman	SS			11%
Nick Zammarelli	3B			10%
Gareth Morgan	OF			6%
Dimas Ojeda	LF			6%
Jhombeyker Morales	SS			5%
Johnny Slater	CF		23%	5%
Louis Boyd	IF	14%	11%	4%
Conner Hale	IF		3%	4%
Nick Thurman	C			4%
Billy Cooke	OF-IF		2%	3%
Kristian Brito	1B			3%

A 2017 Mariners' Prospects				
Player Name	**Position**	**Rookie**	**A-**	**A**
Luis Rengifo	2B			23%
Anthony Jimenez	OF	5%		16%
Johan Quevedo	C			11%
Bryson Brigman	SS			11%
Nick Zammarelli	3B			10%
Gareth Morgan	OF			6%
Dimas Ojeda	LF			6%
Jhombeyker Morales	SS			5%
Johnny Slater	CF		23%	5%
Louis Boyd	IF	14%	11%	4%
Conner Hale	IF		3%	4%
Nick Thurman	C			4%
Billy Cooke	OF-IF		2%	3%
Kristian Brito	1B			3%

Rookie League 2017 Mariners' Prospects		
Player Name	**Position**	**Rookie**
Ryan Costello	IF	13%
Jack Larsen	OF	10%
Caleb Eldridge	1B-OF	8%
Ryan Garcia	1B	7%
Connor Hoover	SS	6%
DeAires Moses	OF	6%
Ismerling Mota	C	5%
Jose Sandoval	OF	3%
Sebastian Ochoa	OF	1%
Nolan Perez	3B	0%
Alexander Campos	SS	0%
Luis Joseph	2B	0%
Cesar Izturis	2B	0%
Oberto Munoz	C	0%
Danny Contreras	OF	0%
Daniel Santos	C	0%
Robert Perez	OF	0%
Freuddy Batista	C	0%
Steve Branche	SS	0%
Luis Veloz	OF	0%
Jepherson Garcia	DH	0%
Jose Cano	3B	0%
Miguel Perez	OF	0%

APPENDIX: References

Baseball Reference. (n.d.). MLB Stats, Scores, History, & Records. Retrieved July 9, 2018, from https://www.baseball-reference.com/

Fangraphs Baseball. (n.d.). Baseball Statistics and Analysis | FanGraphs Baseball. Retrieved July 9, 2018, from https://www.fangraphs.com/

Gaines, C. (2013, June 7). Most Baseball Draft Picks Will Still Be In The Minors Four Years From Now. Retrieved July 7, 2018, from http://www.businessinsider.com/chart-how-long-it-takes-a-draft-pick-to-reach-major-league-baseball-2013-6

High School Baseball Web. (n.d.). Probability Of Playing College and Professional Baseball. Retrieved July 9, 2018, from http://www.hsbaseballweb.com/probability.htm

Major League Baseball. (n.d.). The Official Site of Major League Baseball. Retrieved July 9, 2018, from https://www.mlb.com/

Mitchell, C. (2014, December 30). KATOH: Forecasting Major League Hitting with Minor League Stats. Retrieved July 9, 2018, from https://www.fangraphs.com/tht/katoh-forecasting-a-hitters-major-league-performance-with-minor-league-stats/

MLB Advanced Media, LP. (n.d.) Advanced Stats | Glossary. Retrieved July 31, 2018, from http://m.mlb.com/glossary/advanced-stats

NCAA Research. (2018, April 20). Estimated probability of competing in professional athletics. Retrieved July 9, 2018, from http://www.ncaa.org/about/resources/research/estimated-probability-competing-professional-athletics

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review, 14*(3), 693-709. Retrieved July 9, 2018, from https://www.jstor.org/stable/2525981?seq=1#page_scan_tab_contents.

OKennedy, P. (2013, March 5). What are the odds of making it to the major leagues? Retrieved July 9, 2018, from https://www.blessyouboys.com/2013/3/5/3977782/what-are-the-odds-of-making-it-to-the-major-leagues

Owens, F. (2017, June 26). What percentage of minor league baseball players will ever reach the majors? Retrieved July 9, 2018, from https://www.quora.com/What-percentage-of-minor-league-baseball-players-will-ever-reach-the-majors

Rosenbaum, M. (2012, June 12). Examining the Percentage of MLB Draft Picks Who Reach the Major Leagues. Retrieved July 9, 2018, from https://bleacherreport.com/articles/1219356-examining-the-percentage-of-mlb-draft-picks-that-reach-the-major-leagues#slide0

Sickels, J. (2018, February 15). New York Mets Top 20 prospects for 2018. Retrieved from https://www.minorleagueball.com/2018/2/14/17015250/new-york-mets-top-20-prospects-for-2018

The Baseball Cube. (n.d.) Stats Glossary. Retrieved July 30, 2018, from http://www.thebaseballcube.com/about/stats_glossary.asp

The Baseball Cube. (n.d.). The Baseball Cube - MLB, Minor League, College Statistics, Data and the Draft. Retrieved July 9, 2018, from http://www.thebaseballcube.com/

APPENDIX: Code

Code and internal analysis information available in included Zip file.