New York Mets

Midseason Talent Acquisition Strategy

Final Report

Prepared for Don Wedding, GM
August 21, 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 recommendations for the Midseason Talent Acquisition Strategy. As described in our Project Goals and Initial Findings reports, 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

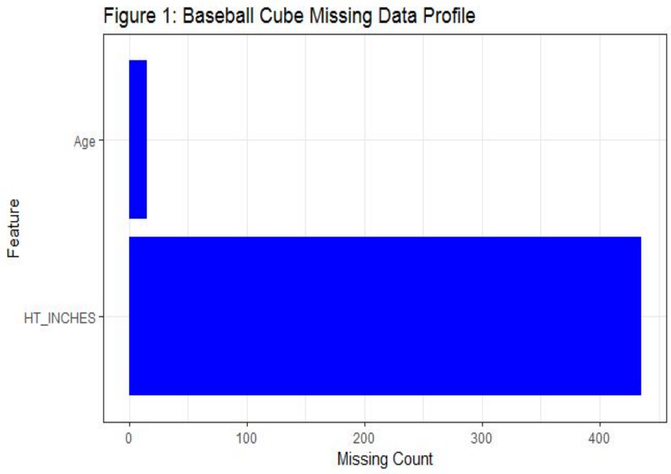
Please see the "Appendix: Statistics & Terminology" section of this document to review the key advanced batting statistics we will be focusing on for our Midseason Talent Acquisition Strategy, including WAR, wOBA, wRC+, OPS+, and wRAA.

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.

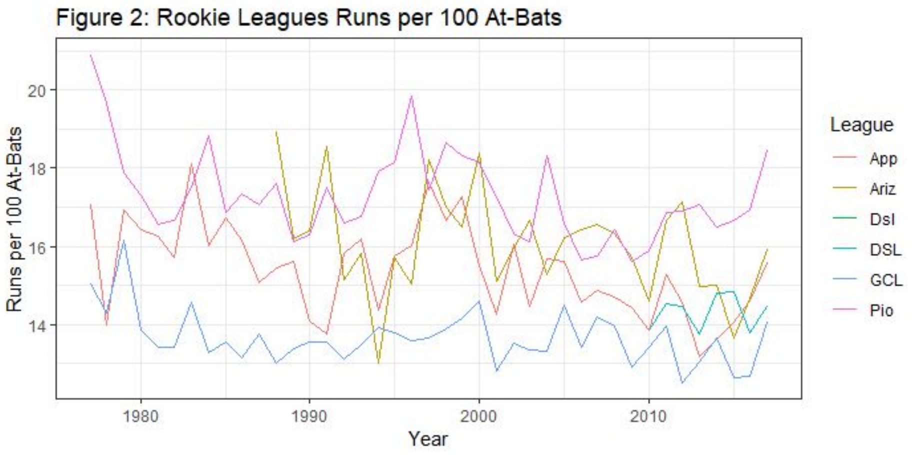
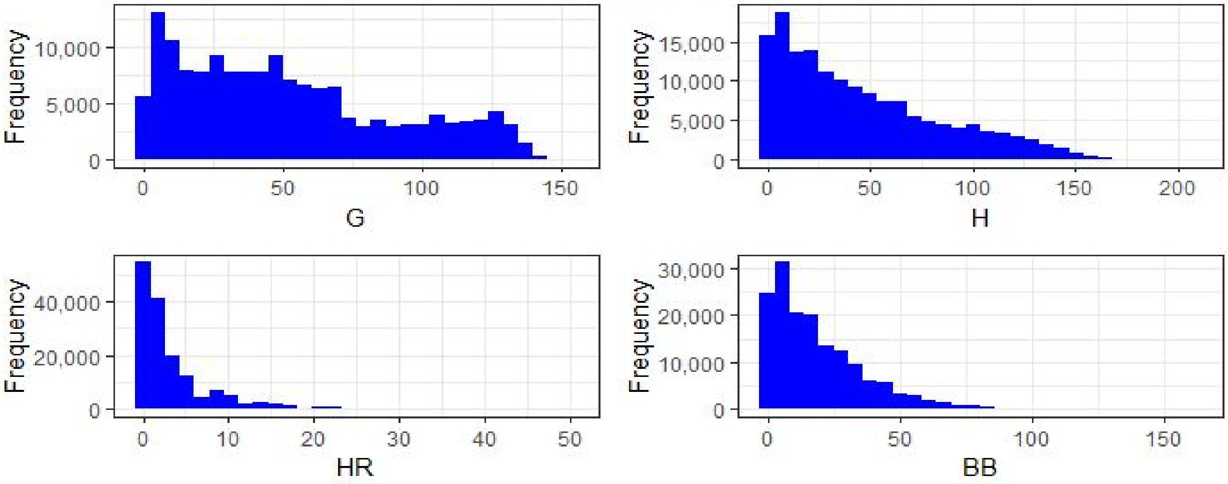
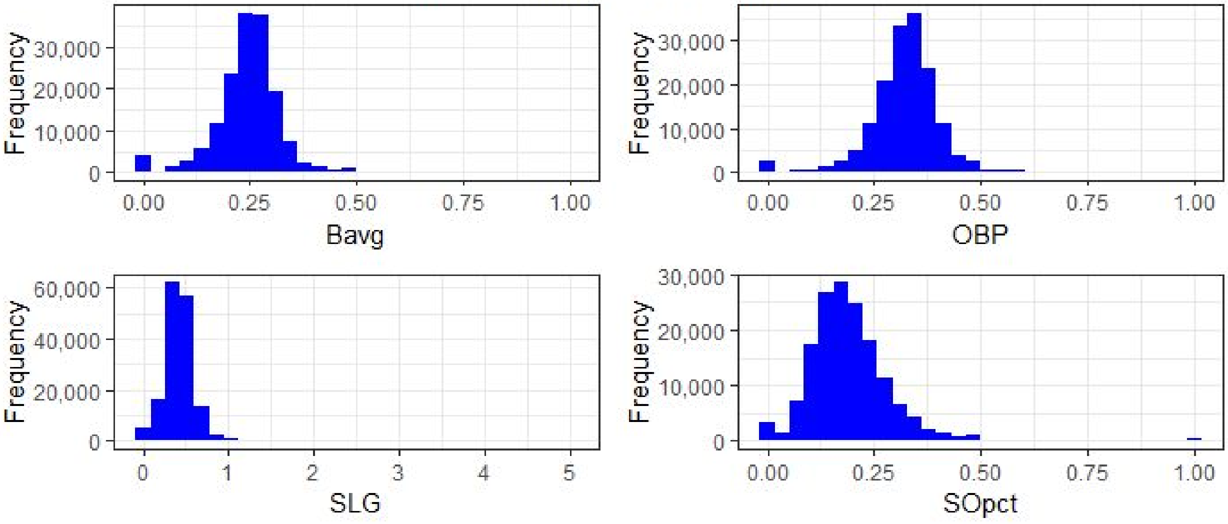


Figure 3: Example Right Skewed Counting Variables



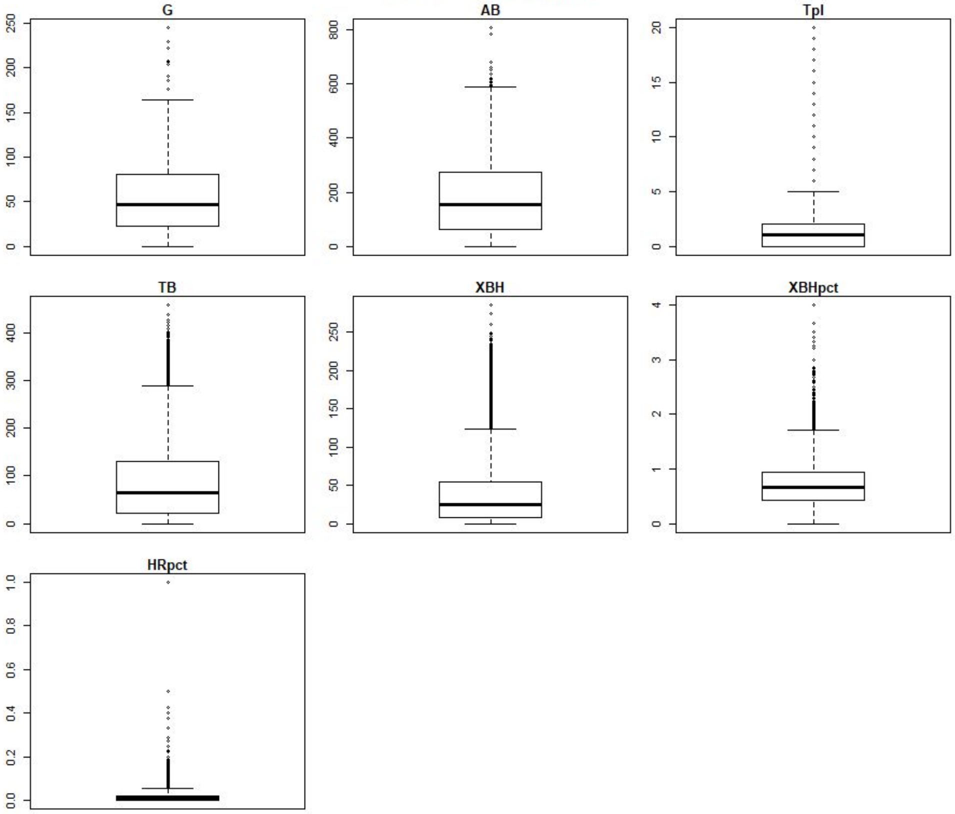
As shown in figure 3, many of the counting statistics, such as at-bats (AB), hits (H), home-runs (HR), and even walks (BB) are skewed right. This is not surprising as many of the counting statistics are heavily related to the number of games played. Figure 4, on the next page, shows that many of the 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

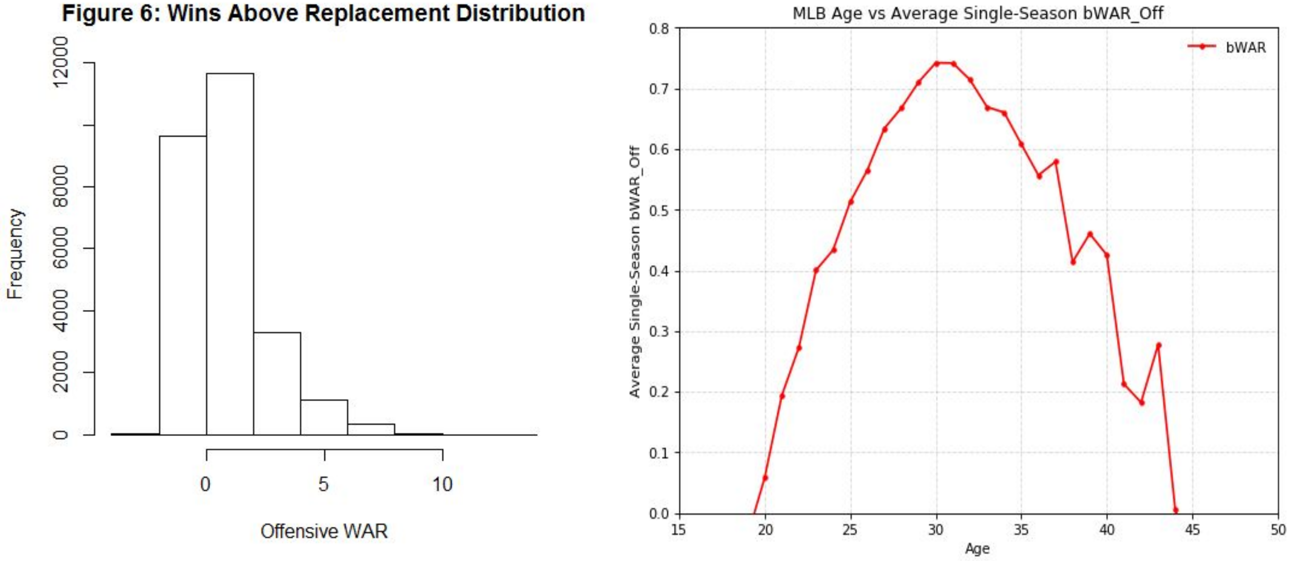


Of the 32,566 position players available in the data, almost 14,000 players have played at least one game in the major leagues, but many of those are injury call-ups or other short stints where a player should not be considered an impactful player for the organization. To address the non-impact of these 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.

Furthermore, we broke down the average single season offensive WAR by age in the dataset, also shown in Figure 6. We do see an aging curve. As players develop in the major leagues, their production increases, until around age 30, when it starts to decrease. There is survivor bias in the graphic, however, as players who stay in the major leagues until their thirties are more likely to be productive contributors.

Figure 6: Wins Above Replacement Distribution



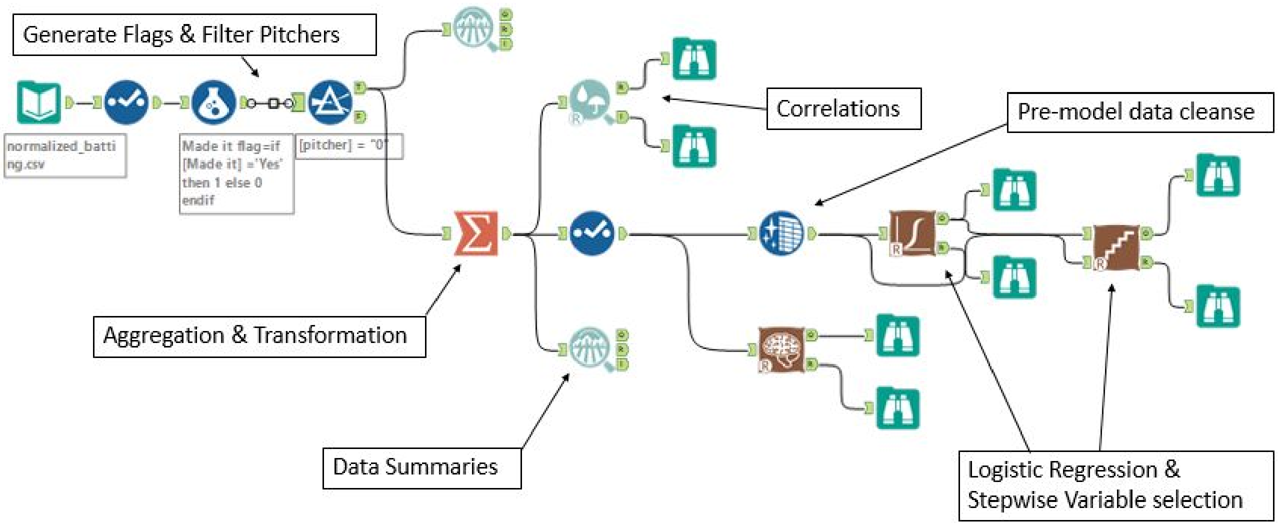
To better understand these players, we consider the MLB players using the categories in the following table to describe their peers, based on the offensive WAR. Players are grouped into 8 categories on a season by season basis. Groupings are based on the average number of players (Implied Players) per 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:

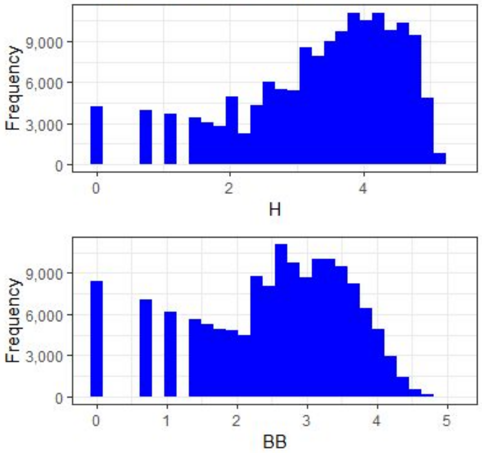

Generate Flags & Filter Pitchers
normalized_batting.csv
Made it flag=if [Made it] ='Yes' then 1 else 0 endif
[pitcher] = "0"
Correlations
Pre-model data cleanse
Aggregation & Transformation
Data Summaries
Logistic Regression & Stepwise Variable selection

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.

Figure 8: Log-Transformations



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.

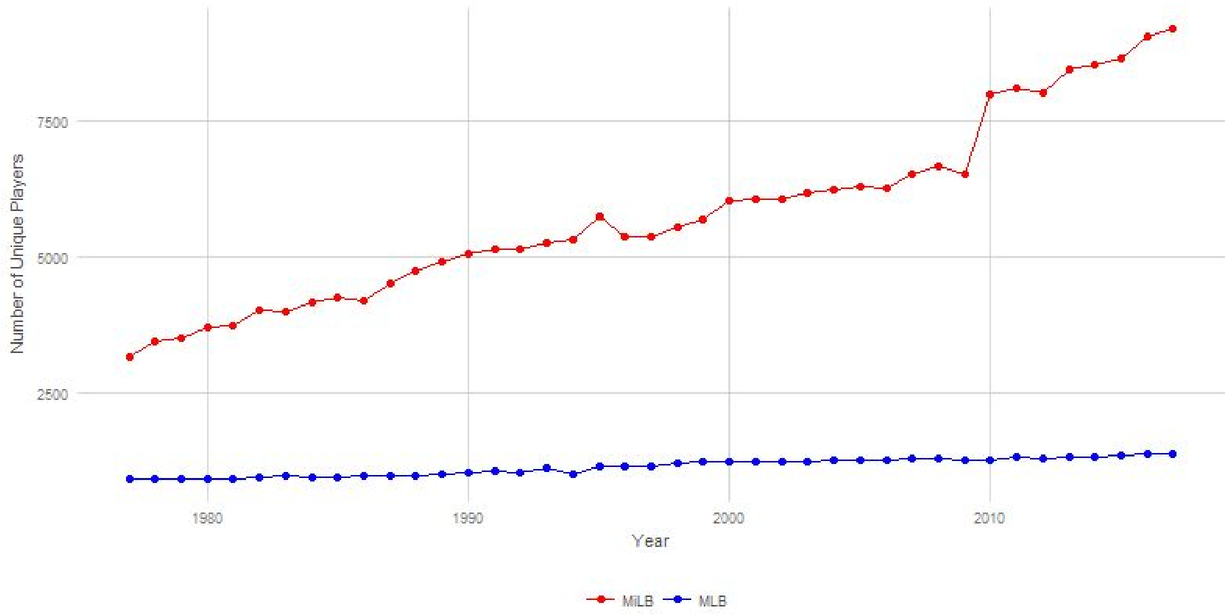


Figure 9: Unique MiLB & MLB players per year

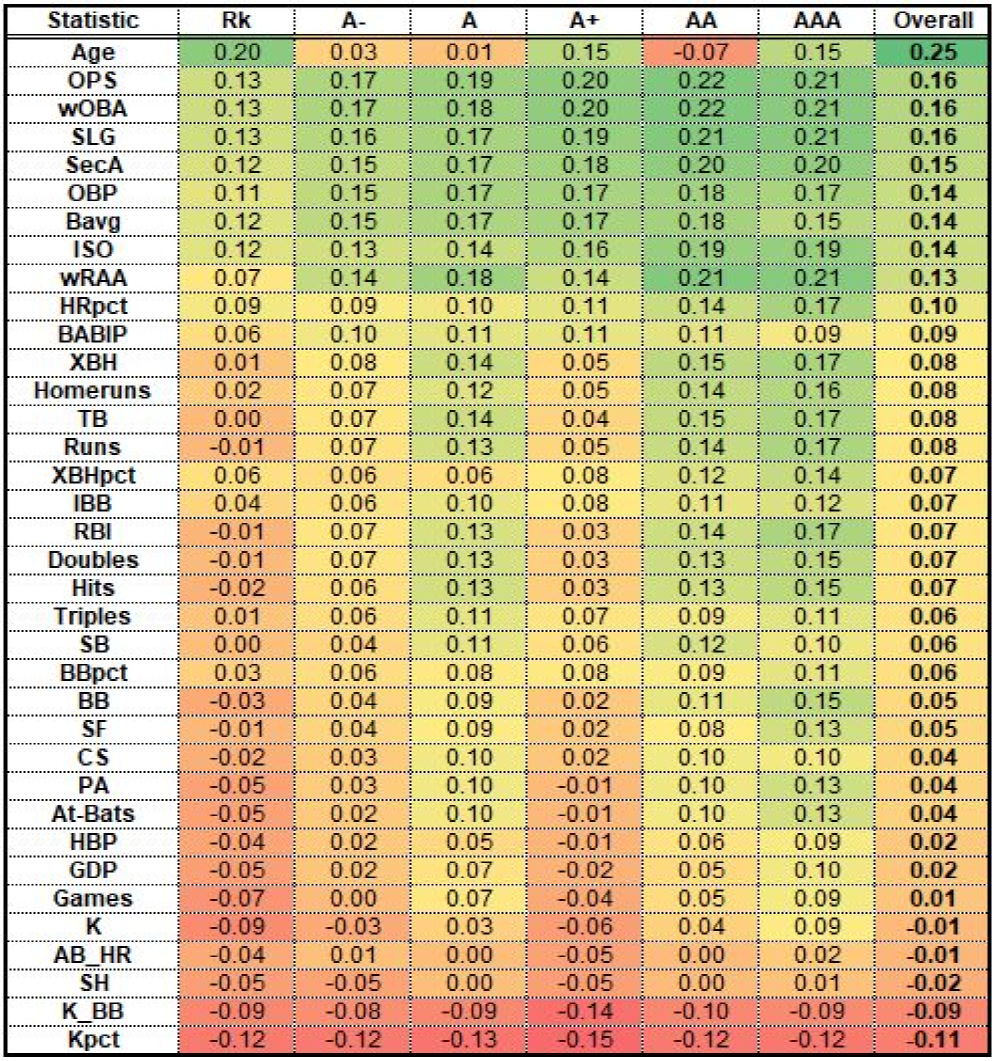
Correlation Analysis to Response Variable

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 “above average” indicate that more logical processing may be needed to determine what feature is being described this way.

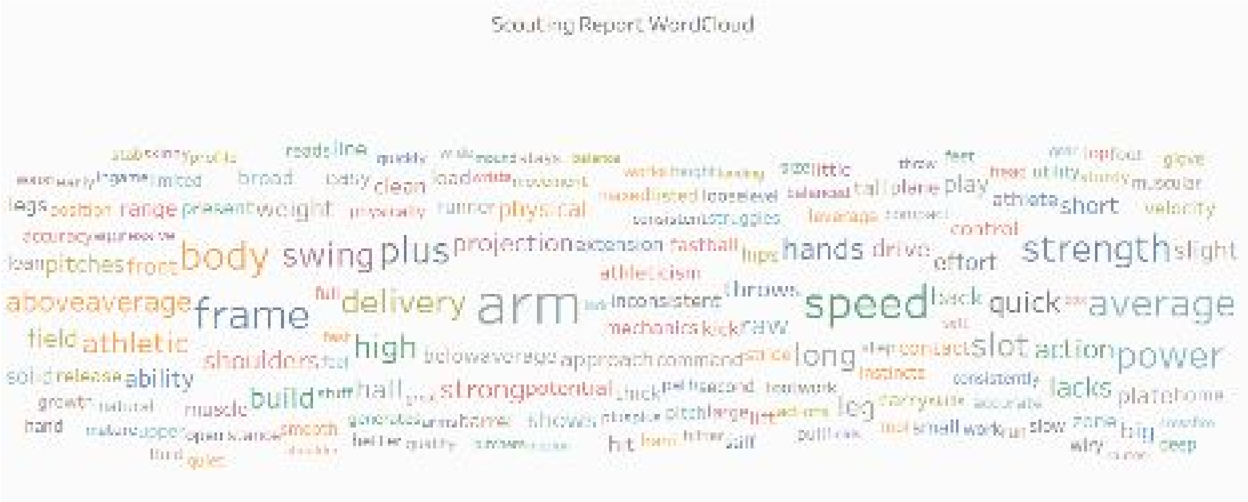


Figure 10: Scouting Report Word Cloud

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)

Modeling

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 per game career batting statistics, including previous levels of play. If a player missed a level of play, such as skipping the Rookie League, then a flag is generated and included in the model. If a player played in more than one level in a year, he would be included in the training data for any level played in, but only using the career statistics in the lower league. A player who played in Rookie League (*Rk*), A+, and AAA would be included in the *Rk* model, using only their Rookie League data, in the A+ model using Rookie League and A+ career statistics, and finally the AAA model, using all three levels cumulative data.

The coefficients for the logistic models are available in the appendix, and figure 11 shows the ROC curve for the base level logistic model. The ROC shows the cumulative correct classifications based on a rank ordered probability. That is showing, by order of predicted probability, the share of the total that have actually made it to the MLB (the y-axis) compared to players the model says have a high likelihood, but failed to stay for at least 3 years in the MLB, the faster the model moves up the left hand side, before moving right, the better the model. As we expect the AAA model is superior in correctly identifying MLB players than the Rookie League model, competition is closer to the MLB level and there is generally more information available about the player in that level.

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

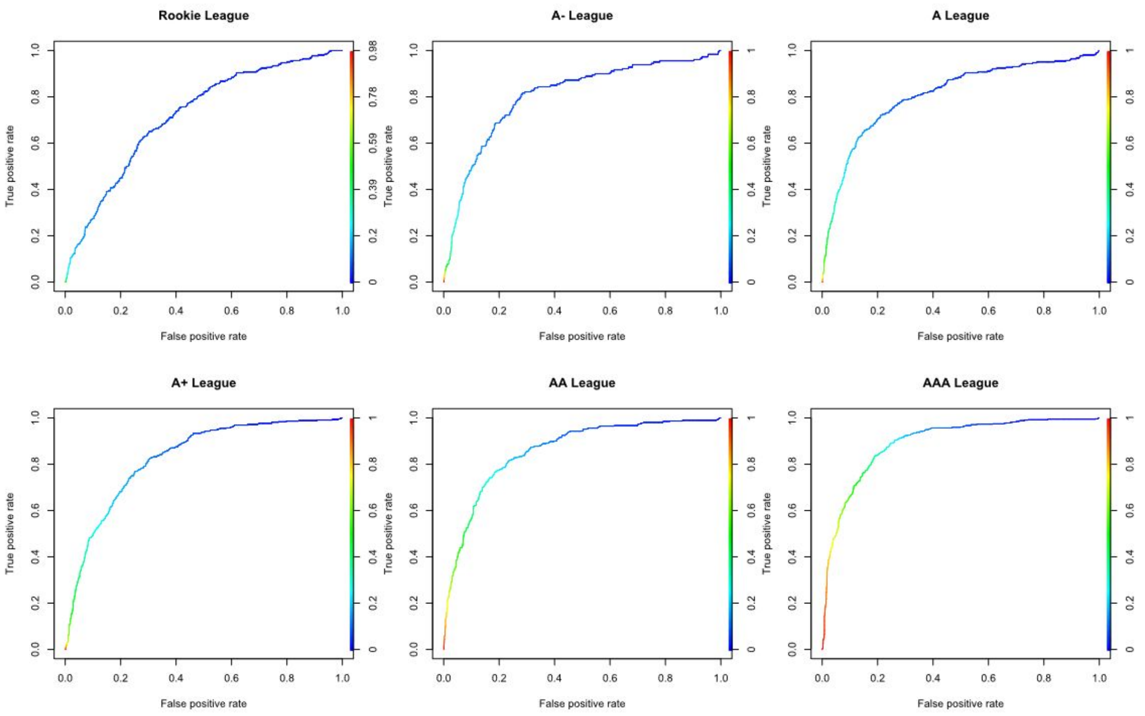


Figure 11: Logistic ROC – "Made It"

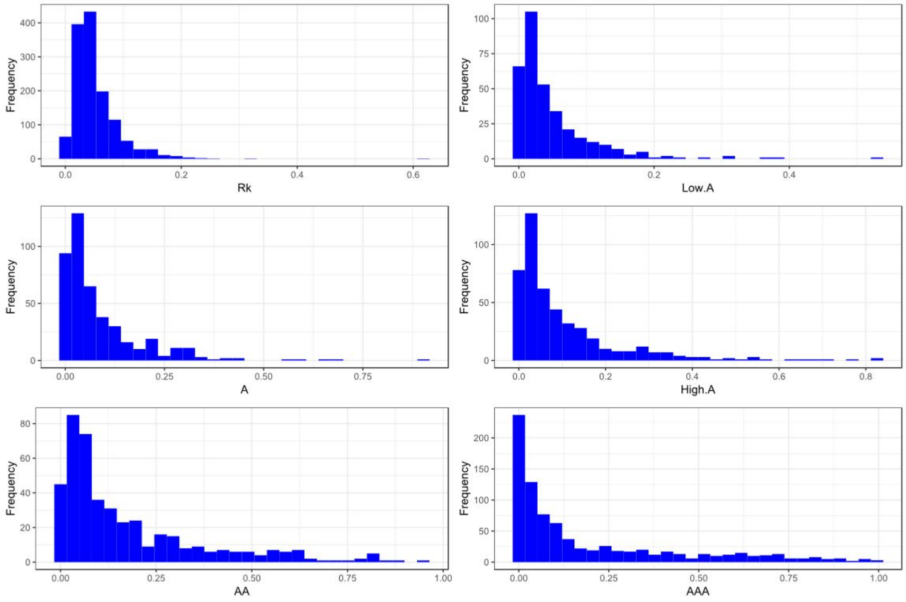


Figure 12: Logistic "Made It" Histogram

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 12, 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.

Random Forest Model Findings

The next suite of models explored are Random Forests, a machine learning technique that frequently will outperform traditional statistical models, such the logistic model above. Using the same set of inputs as the logistic, a large number of Random Forests were trained and compared to each other. Again using 80/20 training/validation split, but each forest was trained to optimize the tuning variables. Figure 13 shows the ROC for the best random forest models and figure 14 shows the distribution of the likelihood across the different levels of play.

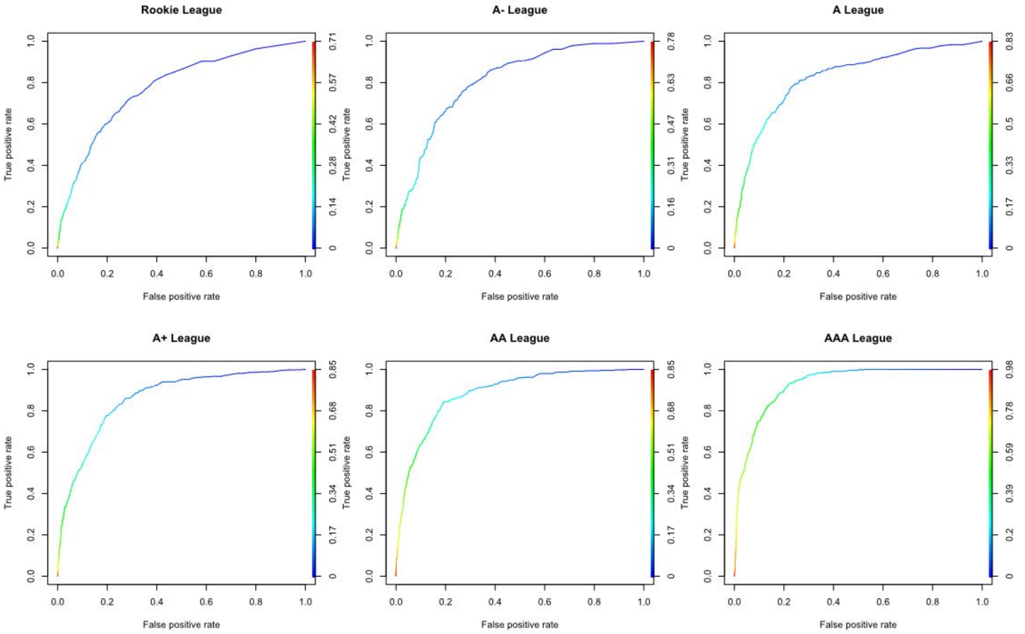


Figure 13: Random Forest ROC - "Made It"

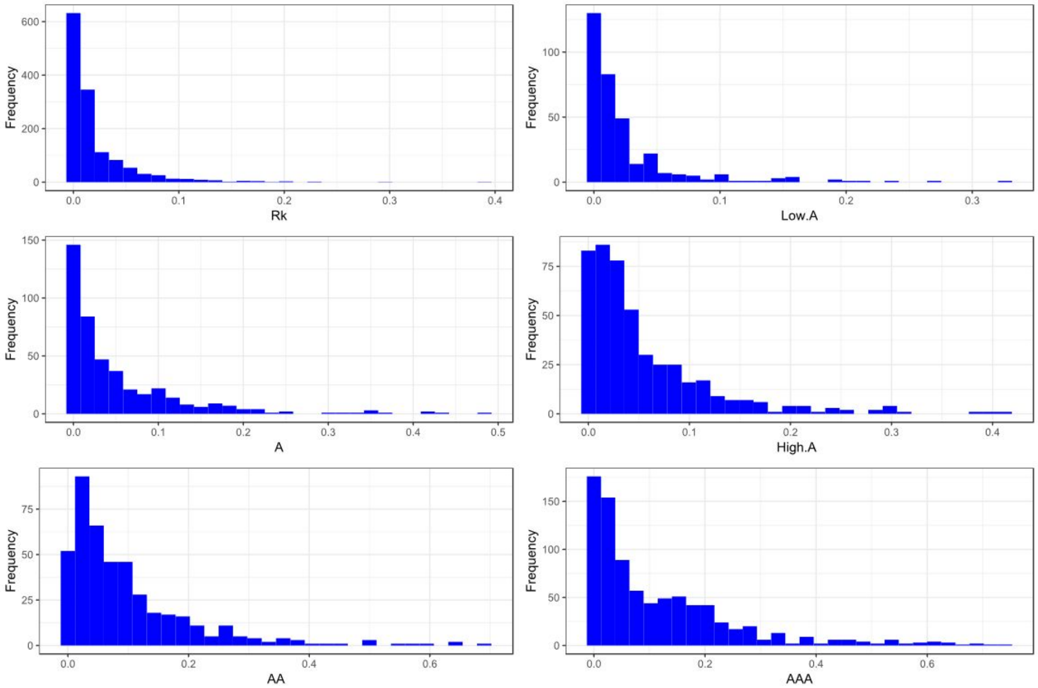


Figure 14: Random Forest "Made It" Histogram

Support Vector Machine (SVM) Findings

The SVM model was attempted but ultimately entirely unsuccessful. Despite tuning both linear and radial kernels, varying the subset and available inputs, the SVM model was unable to generate a meaningful model. The models for any level below AA never gave any player a meaningful chance of making it.

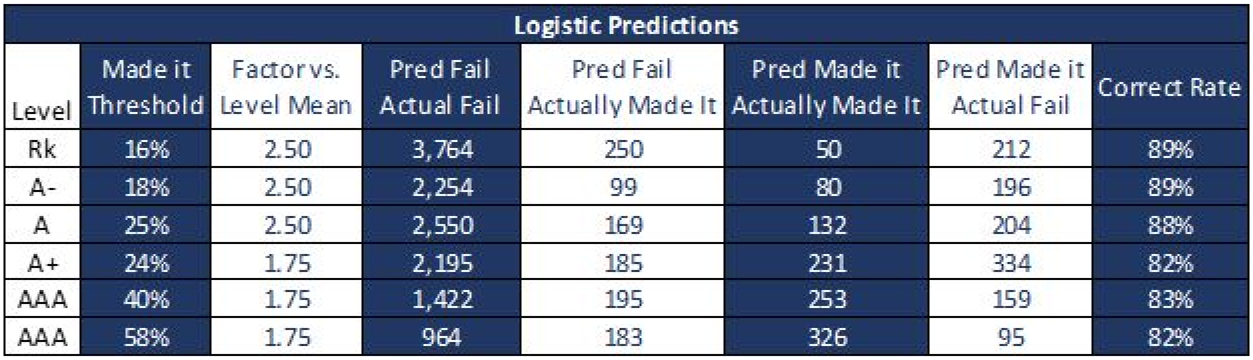
Additionally, this model took such an extensive amount of time and resources to run and tune that it would not be advisable to rely on the SVM predictions in any way.

Made-It Results

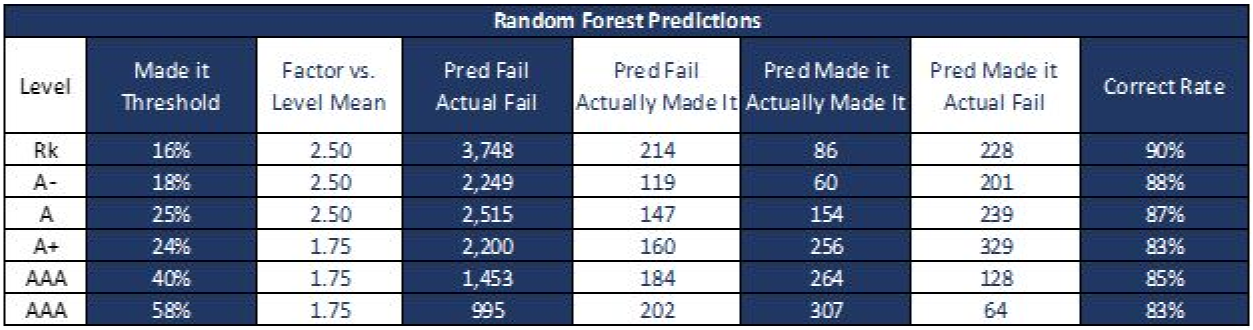
The logistic and random forest models did not have too significant of differences, and ultimately the logistic model is the best for the Mets to implement. The accuracy is generally comparable across the levels, however the Random Forest is better able to identify a more pure pool of prospects that make the major leagues, whereas the logistic has some statistical information that can be leveraged to gain extra information. For example, it is easier to understand the immediate impact of a player improving their batting average or runs per game, with internal knowledge of our developmental strengths, we could look at marginal contributions to better identify a player in an inefficient development system. In a softer approach, combining this information with scouting reports and our developmental programs can let us more simply select players who could thrive in the Mets' system. Recommendations will be based off the random forest model, but information for all models is available in the appendix.

The following tables report the accuracy the the two differing models. These results represent the 20%-sample validation data as predicted by the final models. The threshold at which a player is considered to have made it is defined differently at each level, and are based on the factor in the third column, as we discussed with the scouting department as to which prospects they would be most interested in doing a deeper analysis on.

In most levels, the Random Forest has better accuracy and a higher number of correct prediction of players that will make it. This can give the scouting department a set of prospects to focus on more directly to augment the model.



Logistic Predictions							
Level	Made it Threshold	Factor vs. Level Mean	Pred Fail Actual Fail	Pred Fail Actually Made It	Pred Made it Actually Made It	Pred Made it Actual Fail	Correct Rate
Rk	16%	2.50	3,764	250	50	212	89%
A-	18%	2.50	2,254	99	80	196	89%
A	25%	2.50	2,550	169	132	204	88%
A+	24%	1.75	2,195	185	231	334	82%
AAA	40%	1.75	1,422	195	253	159	83%
AAA	58%	1.75	964	183	326	95	82%



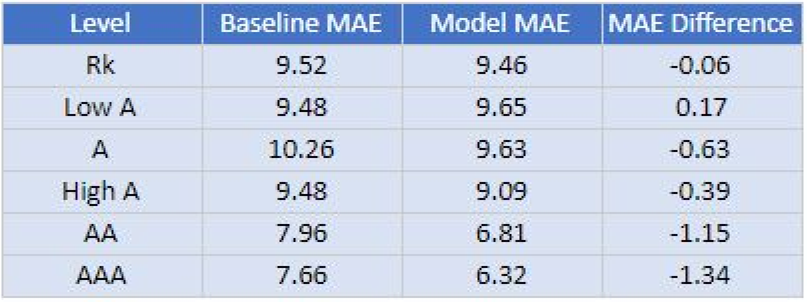
Random Forest Predictions							
Level	Made it Threshold	Factor vs. Level Mean	Pred Fail Actual Fail	Pred Fail Actually Made It	Pred Made it Actually Made It	Pred Made it Actual Fail	Correct Rate
Rk	16%	2.50	3,748	214	86	228	90%
A-	18%	2.50	2,249	119	60	201	88%
A	25%	2.50	2,515	147	154	239	87%
A+	24%	1.75	2,200	160	256	329	83%
AAA	40%	1.75	1,453	184	264	128	85%
AAA	58%	1.75	995	202	307	64	83%

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. Similar to the Made It model above, six models are generated independently for each level of the minors, Rookie (Rk), Low-A (A-), A, High A (A+), AA, and AAA. For each level, we created a gradient boosting regressor comprised of an 80-20% train-validation split including 50 decision trees and a learning rate of 1%. Gradient boosting is a machine learning technique similar to random forests, as they both use decision trees. We found that a gradient boosting model outperformed a simple random forest across all levels.

The initial training 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 models seemed to provide fairly robust predictions. This number was broken down even further, especially at the lower levels, due to the small number of players who have played that low and still made the major leagues.

To begin our analysis, we created a rudimentary baseline of predicted WAR for each level to ensure our model was providing reasonable predictions. We calculated this baseline by simply averaging the actual career WAR in the training set for that particular level. We then calculated the average absolute error, MAE, by subtracting this prediction from the actual WAR values. The MAE then, is the average difference in career WAR from our predictions and the actual values. If our model predictions could not beat this absolute error from simply predicting the average for every data point, then we would need to rethink our methodology.



Level	Baseline MAE	Model MAE	MAE Difference
Rk	9.52	9.46	-0.06
Low A	9.48	9.65	0.17
A	10.26	9.63	-0.63
High A	9.48	9.09	-0.39
AA	7.96	6.81	-1.15
AAA	7.66	6.32	-1.34

Our gradient boosting model provided better predictions than the baseline, in terms of a lower mean absolute error, for every level higher than or equal to A. Unsurprisingly, our model performed around the same as the baseline for the lowest levels, Rookie and Low A. This can be attributed to the small amount of data available to train on at these levels, as well as the fewer features available to our model, since Rookie and Low A are usually the starting point on a player's development journey. As in the Made It model, the AAA model performed the best, decreasing the MAE by 1.34 WAR. The full set of feature importances per model can be found in the attached appendix.

Next, we created a scatter plot of predicted career WAR versus the actual career WAR in our validation set for each of the levels. 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 were no patterns in the residuals.

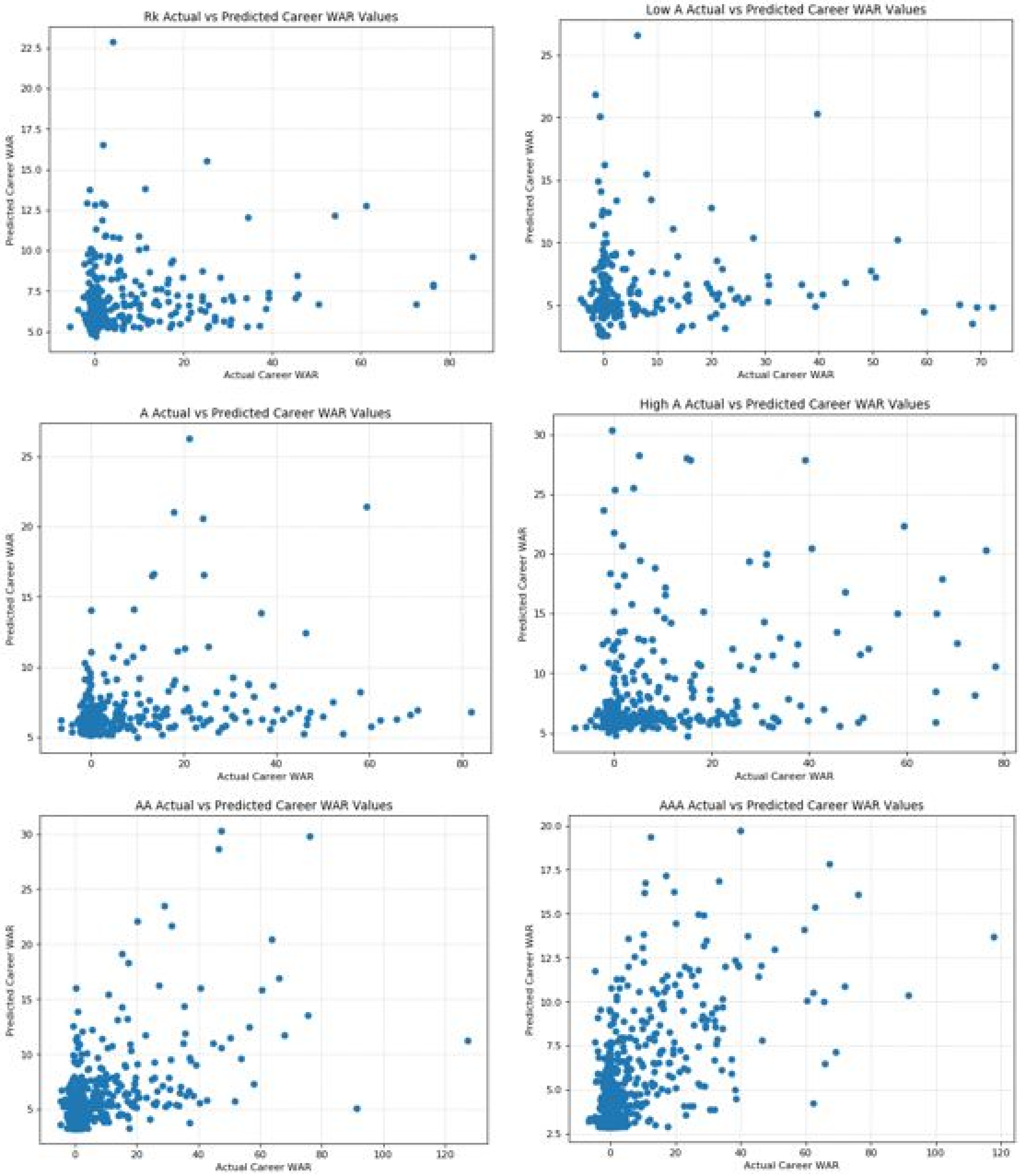


Figure 15: Gradient Boosting Model Predictions vs Actual Career WAR, per Level

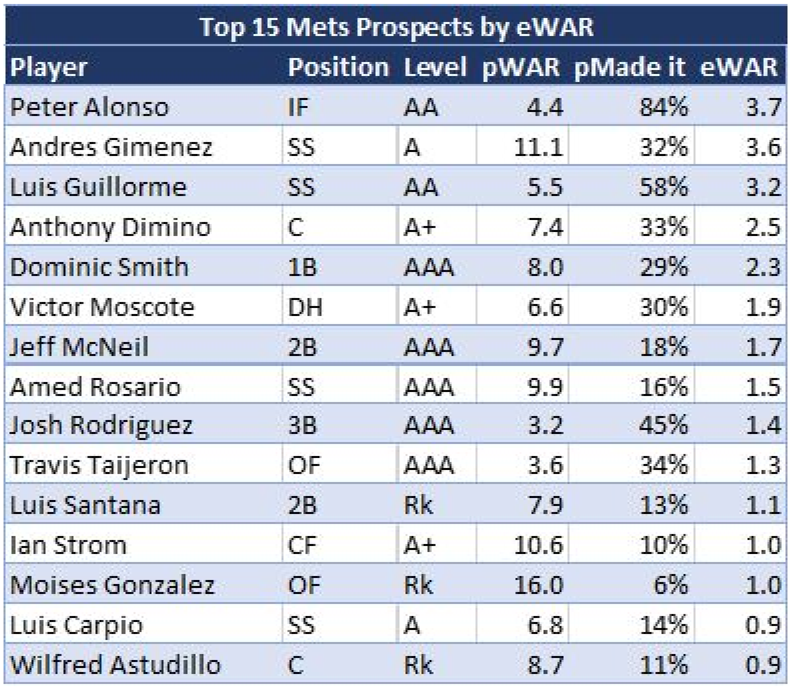
Our model does a great job of predicting career WAR of less than 20. Unfortunately, it seems like outliers are not predicted as robustly as they could be. Most of the outliers are conservatively predicted, however, they are still predicted as “all-stars” or “above-average”. Since these WAR predictions could 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 expect that more trees and more folding on the training set would reduce the absolute mean error in the predictions. This would require further investment in our computing power and technology. While we wish to have more data in our training and validation sets, especially for the Rookie ball and Low A levels, we know that the small conversion rate of MiLB players making the MLB makes this wish impossible. Finally, if we did obtain more data, 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.

We finally retrained each model per level on the full training/validation set and then used those models to predict the career WAR on each of the targeted players.

Model Interactions

The final act is to combined the models generated for a player's likelihood of making it and their predicted WAR. By multiplying the results for each player we can generate an expected value of each player, accounting for their likelihood based on the highest level they have played in as of 2017 and the resulting WAR, we call the product expected-WAR, or eWAR. By judging players by eWAR instead of likelihood or WAR alone, we can factor in the boom or bust nature of some prospects. This also reduces the variance in some of the predictions, especially at the lower levels.



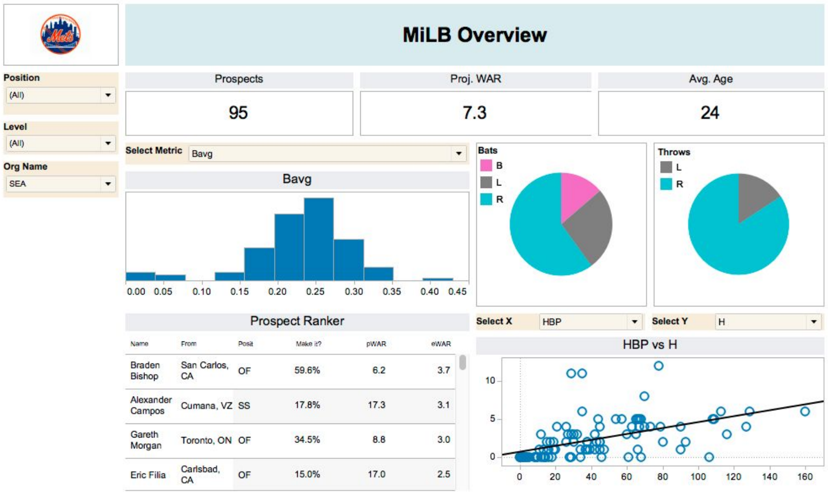
Top 15 Mets Prospects by eWAR

Player	Position	Level	pWAR	pMade it	eWAR
Peter Alonso	IF	AA	4.4	84%	3.7
Andres Gimenez	SS	A	11.1	32%	3.6
Luis Guillorme	SS	AA	5.5	58%	3.2
Anthony Dimino	C	A+	7.4	33%	2.5
Dominic Smith	1B	AAA	8.0	29%	2.3
Victor Moscote	DH	A+	6.6	30%	1.9
Jeff McNeil	2B	AAA	9.7	18%	1.7
Amed Rosario	SS	AAA	9.9	16%	1.5
Josh Rodriguez	3B	AAA	3.2	45%	1.4
Travis Taijeron	OF	AAA	3.6	34%	1.3
Luis Santana	2B	Rk	7.9	13%	1.1
Ian Strom	CF	A+	10.6	10%	1.0
Moises Gonzalez	OF	Rk	16.0	6%	1.0
Luis Carpio	SS	A	6.8	14%	0.9
Wilfred Astudillo	C	Rk	8.7	11%	0.9

The current value of the top 15 prospects in the Mets' farm system are shown to the right. There are only a few players with significant likelihood of making it, and those have a somewhat lackluster upside. To short, our cupboards are rather bare, with only a few players showing significant likelihood of major contributions. As suggested above, though, we can see a player like Moises Gonzalez in the Rookie League has the highest predicted WAR (pWAR), but given the low likelihood of a Rookie League player contributing for a long time, he only has 1 eWAR.

Dashboard - Desktop & Mobile

To aid in the management's use of these models and ultimately aid in acquiring new and undervalued players for the Mets, we have developed a dashboard application to report prospect statistics and value. A snapshot of the Desktop/Tablet version can be seen below:



MiLB Overview

Position: (All)
Level: (All)
Org Name: SEA

Prospects	Proj. WAR	Avg. Age
95	7.3	24

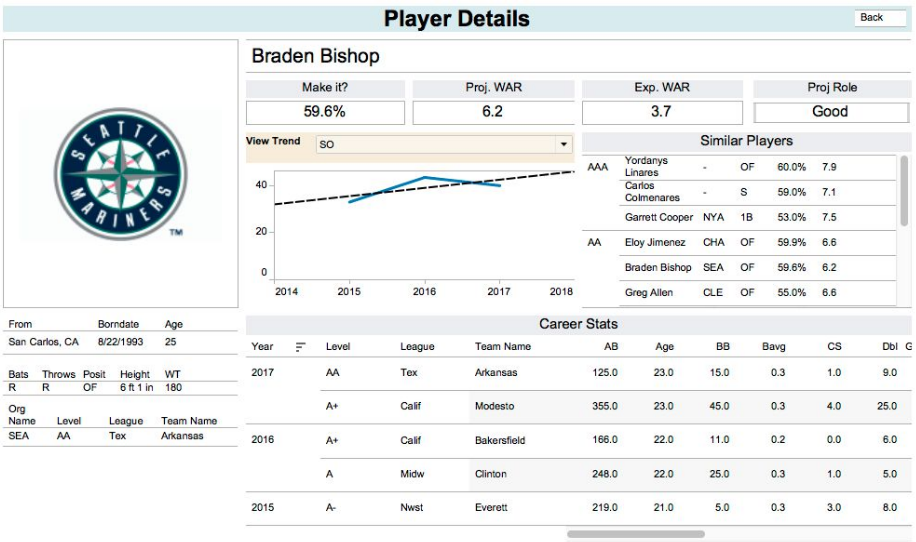
Select Metric: Bavg

Select X: HBP
Select Y: H

Prospect Ranker

Name	From	Posit	Make it?	pWAR	eWAR
Braden Bishop	San Carlos, CA	OF	59.6%	6.2	3.7
Alexander Campos	Cumana, VZ	SS	17.8%	17.3	3.1
Gareth Morgan	Toronto, ON	OF	34.5%	8.8	3.0
Eric Filia	Carlsbad, CA	OF	15.0%	17.0	2.5

As shown above, 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. Additionally, every graph acts as a filter for the user so if you are interested in slicing the data in other ways, you simply click on the population you want to look at and the dashboard will respond. Upon selecting a prospect to look more deeply into, again simply select their name and you will be directed to their player page (seen below).



Player Details

Braden Bishop

Make it?	Proj. WAR	Exp. WAR	Proj Role
59.6%	6.2	3.7	Good

View Trend: SO

Similar Players

AAA	Yordanys Linares	-	OF	60.0%	7.9
	Carlos Colmenares	-	S	59.0%	7.1
	Garrett Cooper	NYA	1B	53.0%	7.5
AA	Eloy Jimenez	CHA	OF	59.9%	6.6
	Braden Bishop	SEA	OF	59.6%	6.2
	Greg Allen	CLE	OF	55.0%	6.6

From	Borndate	Age
San Carlos, CA	8/22/1993	25

Bats	Throws	Posit	Height	WT
R	R	OF	6 ft 1 in	180

Org Name	Level	League	Team Name
SEA	AA	Tex	Arkansas

Career Stats

Year	Level	League	Team Name	AB	Age	BB	Bavg	CS	Dbl	G
2017	AA	Tex	Arkansas	125.0	23.0	15.0	0.3	1.0	9.0	
	A+	Calif	Modesto	355.0	23.0	45.0	0.3	4.0	25.0	
2016	A+	Calif	Bakersfield	166.0	22.0	11.0	0.2	0.0	6.0	
	A	Midw	Clinton	248.0	22.0	25.0	0.3	1.0	5.0	
2015	A-	Nwst	Everett	219.0	21.0	5.0	0.3	3.0	8.0	

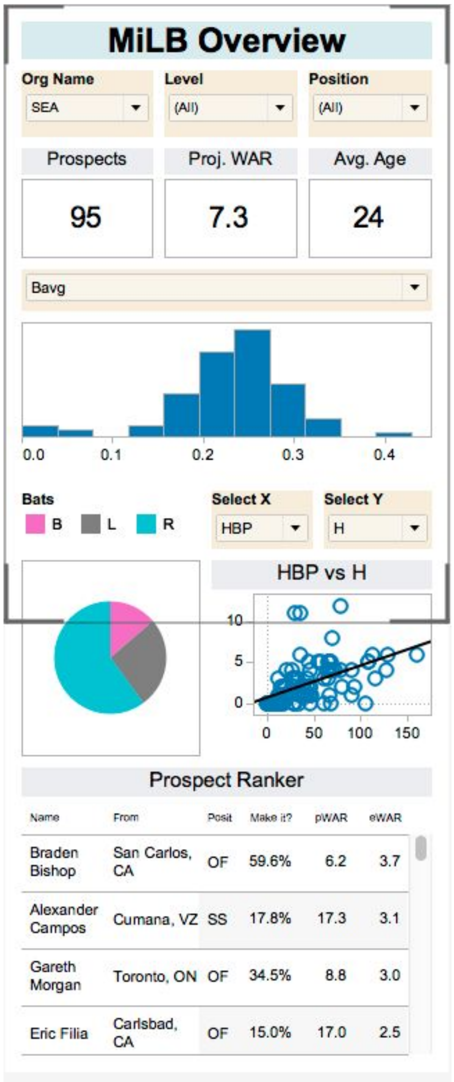
The player details page contains the basic information for that prospect, projected likelihood of making the MLB, projected WAR, and a tag for identifying what type of player he is expected to be. Additionally, a trending view of whatever metric you'd like to view is available along with a career stats table for more detailed views. To go back to the main page, click the back button in the upper right corner.

This application has been optimized for both Desktop & mobile viewing and can be viewed here:

https://public.tableau.com/profile/justin.benson#!/vizhome/MLBEDA/MainPage

Additionally, we've created a quick user guide for you to spread within the organization which goes into more detail about functionality of the dashboard.

Screenshots of the mobile version can be seen below.



MiLB Overview

Org Name	Level	Position
SEA	(All)	(All)

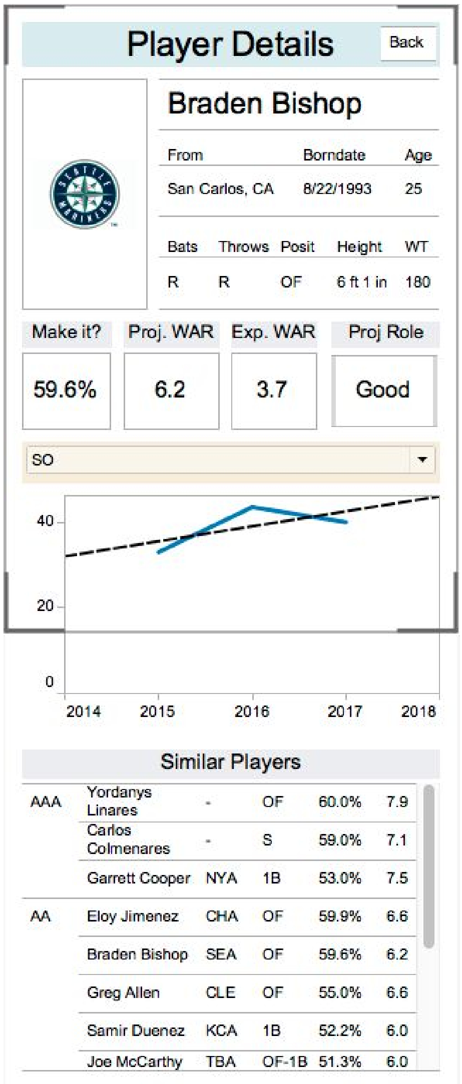
Prospects	Proj. WAR	Avg. Age
95	7.3	24

Bavg

Select X	Select Y
HBP	H

Prospect Ranker

Name	From	Posit	Make it?	pWAR	eWAR
Braden Bishop	San Carlos, CA	OF	59.6%	6.2	3.7
Alexander Campos	Cumana, VZ	SS	17.8%	17.3	3.1
Gareth Morgan	Toronto, ON	OF	34.5%	8.8	3.0
Eric Filia	Carlsbad, CA	OF	15.0%	17.0	2.5



Player Details

Braden Bishop

From	Borndate	Age
San Carlos, CA	8/22/1993	25

Bats	Throws	Posit	Height	WT
R	R	OF	6 ft 1 in	180

Make it?	Proj. WAR	Exp. WAR	Proj Role
59.6%	6.2	3.7	Good

SO

Similar Players

AAA	Yordanys Linares	-	OF	60.0%	7.9
	Carlos Colmenares	-	S	59.0%	7.1
	Garrett Cooper	NYA	1B	53.0%	7.5
AA	Eloy Jimenez	CHA	OF	59.9%	6.6
	Braden Bishop	SEA	OF	59.6%	6.2
	Greg Allen	CLE	OF	55.0%	6.6
	Samir Duenez	KCA	1B	52.2%	6.0
	Joe McCarthy	TBA	OF-1B	51.3%	6.0

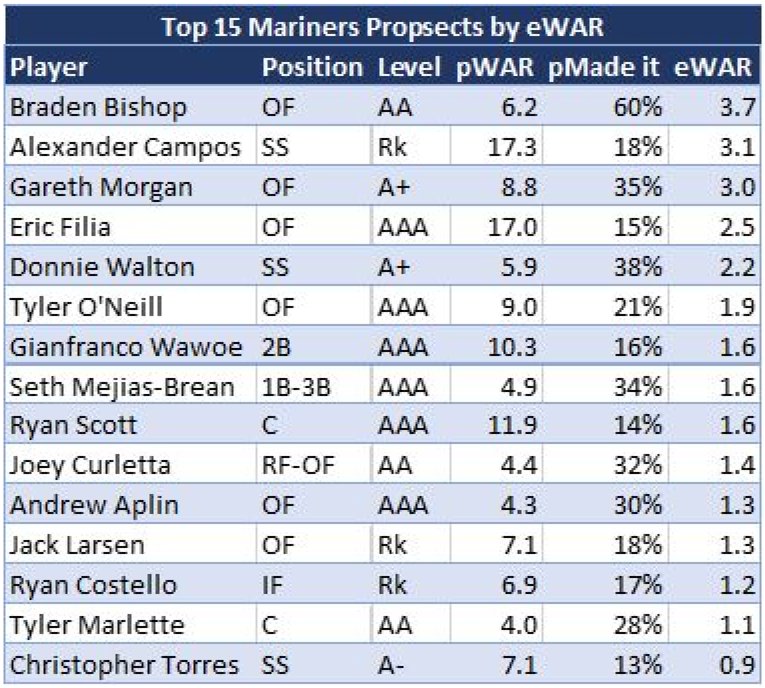
Technology Resources

The following table shows the technology resources used. All of the software was already in use within the analytics department and so no extra budget was spent on tools. R and Python are open source in addition to already being imbedded. Alteryx and Tableau have been used for prior projects and have shown their value. We use Tableau's capacity for both mobile and desktop dashboards to generate the final tool that management and the scouting department can use to better identify valuable trade targets.

Activity	Tool	Description	Libraries/Applications
Data Wrangling and EDA	Python	Common open source software with packages for data manipulation and extensive community support.	Pandas, numpy
	R/R Studio	Common open source software with packages for data scraping, manipulation, and extensive community support.	rvest, tidyverse, ggplot2, ROCR, stargazer, DataExplorer, data.table
	Alteryx	Proprietary Data blending, transforming and light analytics software	R Code macros & Correlation tools
Model Development	Python (Spyder, Anaconda, Jupyter)	Common open source software with packages for data modeling and extensive community support.	SciKit-Learn, Matplotlib, Seaborn
	R/R Studio	Common open source software with packages for data modeling and extensive community support.	tidyverse, caret, e1071, MASS
Reporting	Tableau	Proprietary software already embedded in the organization.	Desktop application & Public server

Recommendations

After the analysis performed we come up the following information for the Mariners prospects. Given these results it would be wise to consider Braden Bishop (OF) and Gareth Morgan (OF), and Seth Mejias-Brean (1B-3B) as safer prospects to trade for. If the first two are available for players that Peter Alonso (Mets - IF) can replace, then we would trade our infield for outfield prospects, and for Seth, if we could replace the current player with Anthony Domino (C) or Travis Taijeron (OF) that could help the team develop.



Top 15 Mariners Propsects by eWAR

Player	Position	Level	pWAR	pMade it	eWAR
Braden Bishop	OF	AA	6.2	60%	3.7
Alexander Campos	SS	Rk	17.3	18%	3.1
Gareth Morgan	OF	A+	8.8	35%	3.0
Eric Filia	OF	AAA	17.0	15%	2.5
Donnie Walton	SS	A+	5.9	38%	2.2
Tyler O'Neill	OF	AAA	9.0	21%	1.9
Gianfranco Wawoe	2B	AAA	10.3	16%	1.6
Seth Mejias-Brean	1B-3B	AAA	4.9	34%	1.6
Ryan Scott	C	AAA	11.9	14%	1.6
Joey Curletta	RF-OF	AA	4.4	32%	1.4
Andrew Aplin	OF	AAA	4.3	30%	1.3
Jack Larsen	OF	Rk	7.1	18%	1.3
Ryan Costello	IF	Rk	6.9	17%	1.2
Tyler Marlette	C	AA	4.0	28%	1.1
Christopher Torres	SS	A-	7.1	13%	0.9

Conclusion

Although the New York Mets are in the midst of a disappointing season, there is much hope for the future. The team is only a few years removed from a trip to the world series, and the pitching rotation is mostly still under control. With an infusion of offensive talent, we expect a swift return to contention. The scouting department should focus their efforts on prospects identified from our models within the Mariners Minor League system. These players provide the greatest opportunity to bring value to our Major League roster in the future.