New York Mets

Midseason Talent Acquisition Strategy

FINAL REPORT: ORAL PRESENTATION

Prepared for Don Wedding, GM
August 28th, 2018

New York Mets Analytics
Alexander Booth, Justin Benson,
Noah Lieberman, Michael Pallante,
Thomas Popeck Spiller

Project Overview

Current State

*As of 8/27/2018

- **Record:**
  58-73 *(4th place NL East)*
- **Payroll:**
  $149.6M (12th MLB)
- **Division:**
  Braves & Phillies
- **MiLB Ranking:**
  28th

Recommendation

Midseason Talent Acquisition Strategy

Goal:
Return to contention in 2019

CONTENTS

Project Goals & Deliverables

The goal of the Midseason Talent Acquisition Strategy is to develop predictive modeling capabilities to forecast the likelihood and future production of prospects in other organizations. Using the output, Mets front office management will be enabled to execute trades using the deliverables stated below to return to contention

Objective	Deliverables	What Defines Success
Prospect Value Projections*	• Robust data infrastructure of historical minor league performance • Predictive models for likelihood to reach majors (e.g., 'Make it') and projected career value • Final report and trade recommendations	• Successful acquisition and organization of data model • Accurate testing of model within agreed upon error bounds
Trade Scenario Dashboard*	• Dashboard that incorporates current rosters, minor league projections, and facilitates what-if trade scenarios	• Usability and successful sign-off from the GM
Mobile Application	• Mobile app that enables the analytics team and the GM to visualize results as well as the dashboard on the go	• Usability and successful sign-off from the GM

**Note: to be tailored to facilitate trades with the Mariners' organization, with the ability to go broader for other clubs*

Key Project Deliverable

Summary of Data Sources

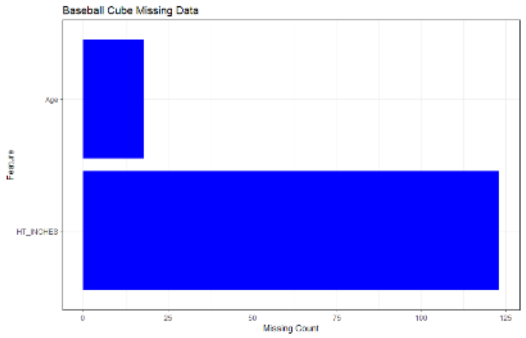
In order to deliver on project goals and deliverables, we plan to develop models based upon the following data sources

Source Name	Description	Acquisition
The Baseball Cube	• Major League Data: Player batting, fielding and pitching data from 1865 – 2017 • Minor League Data: Player batting and pitching data from 1977- 2017	Download
Baseball Reference	• Minor League Data: Player batting and pitching data from 1977 - 2017	Web scraping
Fangraphs	• Minor League Data: Player batting and pitching data from 2006 - Current	Download
Lahmans' MLB Database	• Major League Data: Player batting, fielding and pitching data from 1865 - 2017	Download
The Baseball Prospectus	• Scouting reports for recent prospects	Web scraping

Primary Data Source

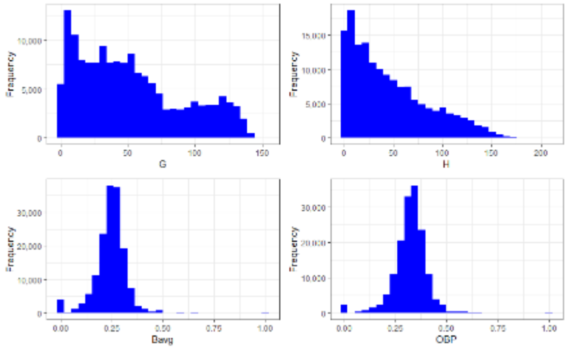
Data Overview

Robust Data Infrastructure



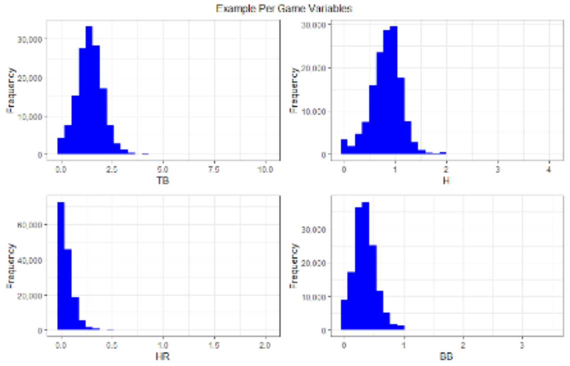
Takeaways:
- *Baseball Cube data has 156,589 non-pitcher observations representing 32,566 MiLB players from 1977-2017*
- *Only age and height are missing for a small sample of our population*

Skewed Counting Variables and Normally Distributed Ratio



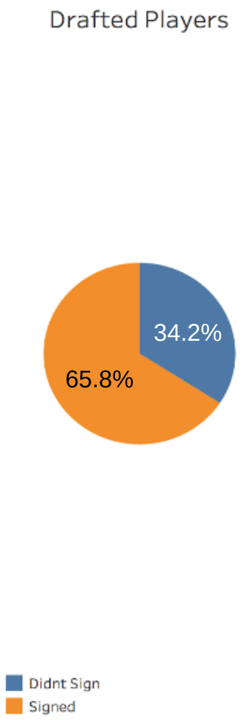
Takeaways:
- *Many counting statistics (e.g., AB, H, HR, BB) are skewed right*
- *High correlation between counting statistics and games played*
- *Ratio variables (e.g., Bavg, OBP, SLG, SOpct) display a normal distribution*
- *Correlation between ratio variables and games played is not as significant as it is for counting variables*

Per Game Variables



Takeaways:
- *Converting to a per game basis will de-skew many of the variables.*
- *Some, like HR may need further adjustment.*

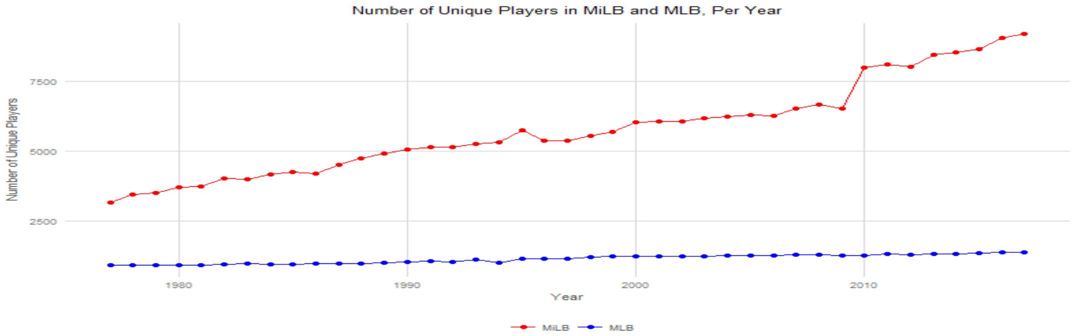
Exploratory Data Analysis
Percentage Distribution to Reach MLB Level

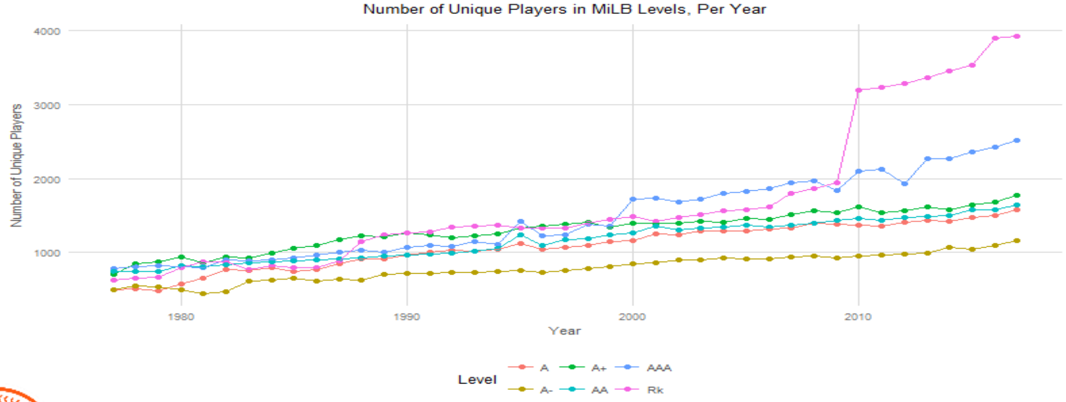


Observations

- Approximately 2/3 of players drafted in the MLB draft sign with their professional organization
- Out of the players that sign, slightly more than 60% of players will never surpass High A ball
- Only 11% of players will reach the Major Leagues

Exploratory Data Analysis
Unique Player Counts By Level

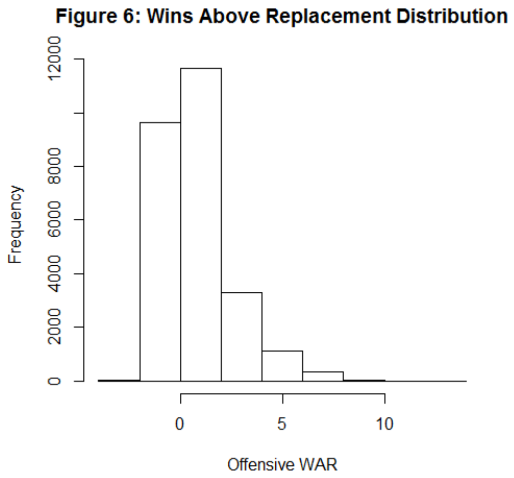




Observations

- MiLB player growth has significantly outpaced MLB growth (which has remained relatively constant), especially since 2010
- The largest increase in MiLB growth has been at the Rookie (Rk) level (driven by the addition of Dominican Summer League statistics)

Exploratory Data Analysis
Wins Above Replacement (WAR) Distribution

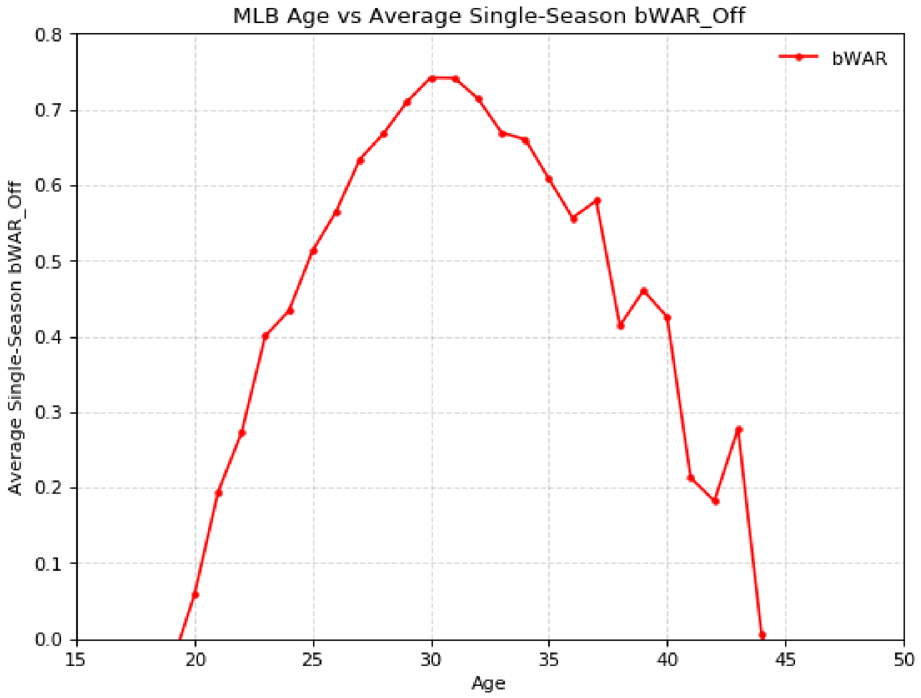


Player Value	WAR	Implied # Players
Scrub	<-.25	106
Replacement Player	-0.25 to 0.25	228
Role Player	0.25 to 1	117
Solid Starter	1 to 2.5	112
Good Player	2.5 to 4	54
All-Star	4 to 6.5	32
Superstar	6.5 to 7.5	4
MVP	7.5+	2

Observations

- Through examination of MLB player WAR values from 1977-2017, we find a skewed left distribution of “Scrubs”, “Replacement Players”, “Role Players”, and “Solid Starters”
- On a year-by-year basis, based on the number of players in the MLB, we may imply there will by 32 “All-Star”, 4 “Superstar”, and 2 “MVP” equivalent statistical seasons

Exploratory Data Analysis
Wins Above Replacement (WAR) Distribution



Observations

- Further examination of MLB player WAR values from 1977-2017, shows a well defined aging curve. On average, the best single-season in terms of WAR is achieved when a player is 30-31.
- With such a stark distribution, we hypothesize that age will play a large role in defining the relationship between whether a player will “make-it” as well as their total career WAR, due simply to access to accumulation.

Exploratory Data Analysis
Correlation Plots

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

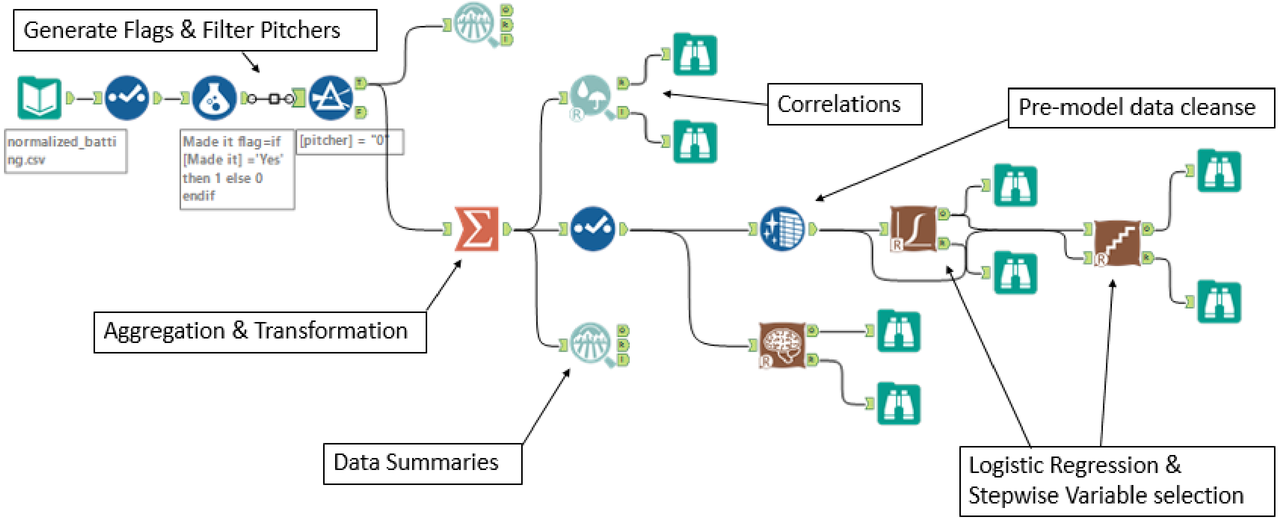
Statistic	Rk	A-	A	A+	AA	AAA	Overall
Age	0.20	0.03	0.01	0.15	-0.07	0.15	**0.25**
OPS	0.13	0.17	0.19	0.20	0.22	0.21	**0.16**
wOBA	0.13	0.17	0.18	0.20	0.22	0.21	**0.16**
SLG	0.13	0.16	0.17	0.19	0.21	0.21	**0.16**
SecA	0.12	0.15	0.17	0.18	0.20	0.20	**0.15**
OBP	0.11	0.15	0.17	0.17	0.18	0.17	**0.14**
Bavg	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**

Observations

- *Age* is the strongest predictor of a prospect being called up to the Major Leagues, however it is potentially misleading due to several factors
- 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
- Striking out (*K_BB, Kpct*) is the strongest negative indicator on likelihood to make the Major Leagues

Description of Transformation of Data

Alteryx was used to perform data transformations and quickly create workflows to be scaled within the organization to other MiLB analysis



Data Modeling
“Made It” Model

The first analytical output created was a “Made It” model to assess the likelihood of players to reach the Major Leagues for at least 3 seasons. Several techniques were tested and ultimately a random forest model proved most accurate

	Area Under ROC Curve	
Level	Random Forest	Logistic
Rk	0.77	0.72
A-	0.81	0.80
A	0.83	0.81
A+	0.86	0.83
AA	0.88	0.86
AAA	0.93	0.89

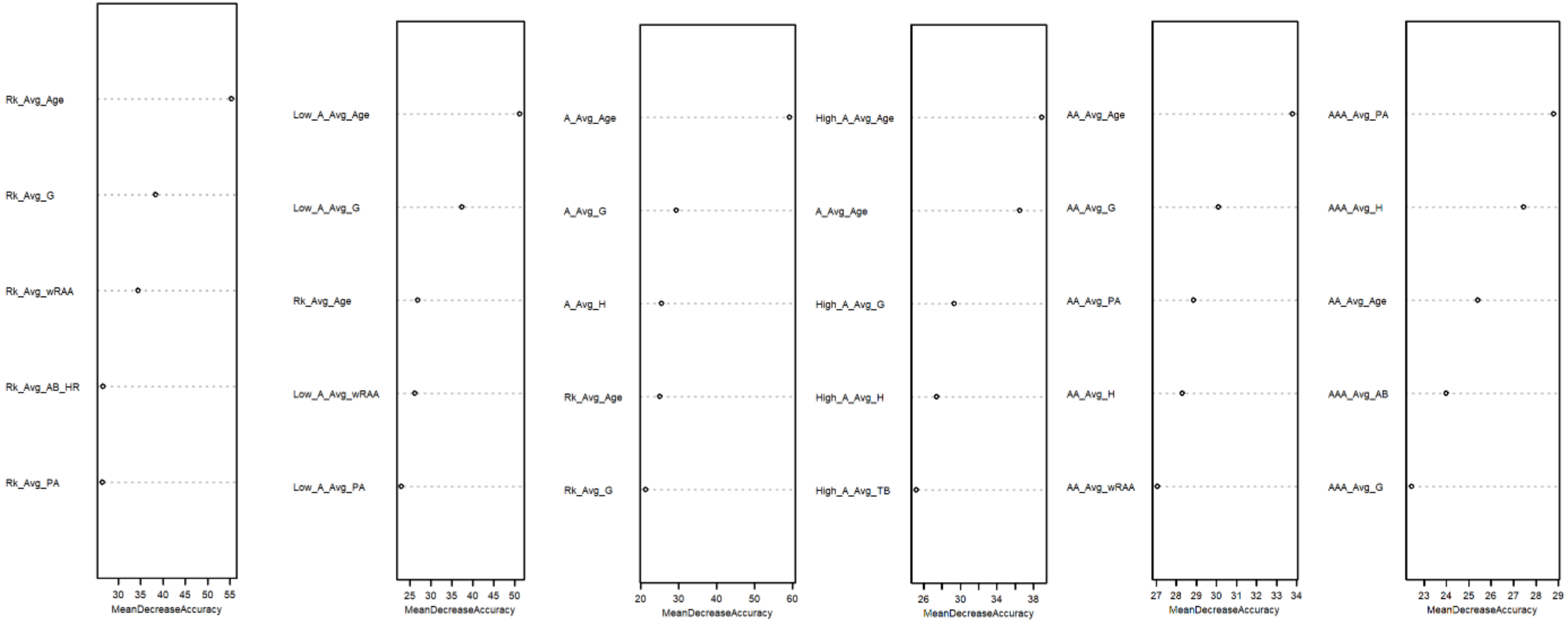
Logistic Predictions

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

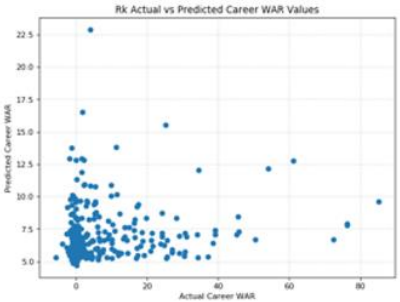
Level	Made it Threshold	Factor vs. Level Mean	Pred Fail Actual Fail	Pred Fail Actually Made It	Pred Made it Actually Made It	Pred Made it Actual Fail	Correct Rate
Rk	16%	2.50	3,748	214	86	228	90%
A-	18%	2.50	2,249	119	60	201	88%
A	25%	2.50	2,515	147	154	239	87%
A+	24%	1.75	2,200	160	256	329	83%
AAA	40%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

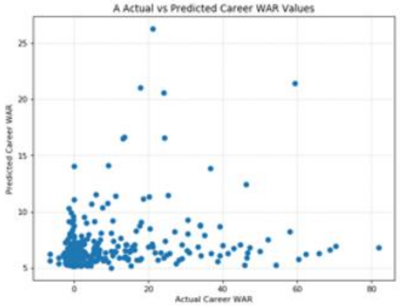
Data Modeling
“Made It” Model

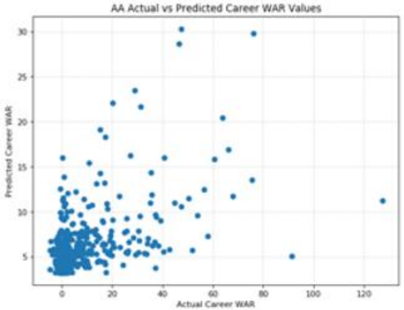


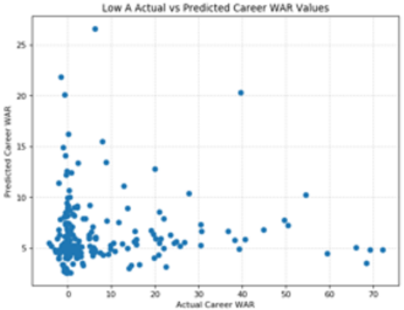
Data Modeling
WAR Model

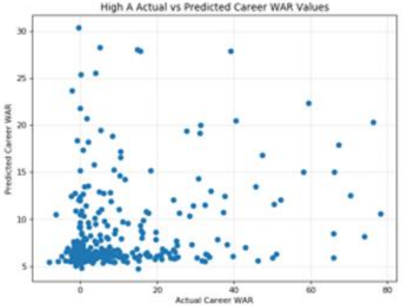
The second analytical output created was a WAR model to assess the future MLB value of Minor League prospects. A gradient boosting model (GBM) provided the best predictions relative to a baseline

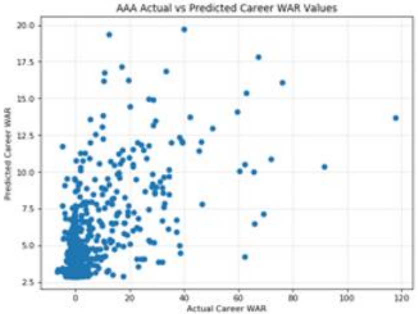












Level	Baseline MAE	Model MAE	MAE Difference
Rk	9.52	9.46	-0.06
Low A	9.48	9.65	0.17
A	10.26	9.63	-0.63
High A	9.48	9.09	-0.39
AA	7.96	6.81	-1.15
AAA	7.66	6.32	-1.34

Data Modeling
WAR Model

Rookie League:

Variable: Rk_Avg_Age	Importance: 0.07
Variable: Rk_Avg_CS_norm	Importance: 0.06
Variable: Rk_Avg_AB_HR_norm	Importance: 0.05
Variable: Rk_Avg_IBB_norm	Importance: 0.03
Variable: Rk_Avg_wOBA_norm	Importance: 0.03

Low A:

Variable: Low_A_Avg_AB_HR	Importance: 0.05
Variable: Low_A_Avg_HBP_norm	Importance: 0.04
Variable: Rk_Avg_GDP_norm	Importance: 0.03
Variable: Low_A_Avg_OPS_norm	Importance: 0.03
Variable: Low_A_Sum_SecA	Importance: 0.03

A:

Variable: A_Avg_Age	Importance: 0.09
Variable: A_Avg_wOBA_norm	Importance: 0.06
Variable: A_Avg_PA	Importance: 0.05
Variable: A_Avg_Tpl_norm	Importance: 0.03
Variable: A_Sum_BBpct	Importance: 0.03

High A:

Variable: High_A_Avg_Age	Importance: 0.11
Variable: High_A_Sum_SecA	Importance: 0.05
Variable: High_A_Avg_BBpct_norm	Importance: 0.04
Variable: High_A_Avg_SecA	Importance: 0.04
Variable: High_A_Avg_HRpct_norm	Importance: 0.03

AA:

Variable: High_A_Avg_Age	Importance: 0.1
Variable: AA_Avg_wRAA	Importance: 0.06
Variable: AA_Avg_Age	Importance: 0.04
Variable: AA_Avg_BABIP_norm	Importance: 0.02
Variable: AA_Avg_OBP	Importance: 0.02

AAA:

Variable: AAA_Avg_AB_HR_norm	Importance: 0.07
Variable: AAA_Avg_GDP_norm	Importance: 0.06
Variable: AAA_Avg_IBB_norm	Importance: 0.05
Variable: AAA_Sum_G	Importance: 0.04
Variable: AAA_Avg_Age	Importance: 0.03

Recommendations
Top Mets Prospects within the Organization

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

Observations

- A total future value metric, eWAR, was created based on the “Made It” (pMade it) and WAR (pWAR) models
- Dominic Smith and Jeff McNeil, are the prospects closest having full time career with the Mets organization and necessary parts of the future infield
- We should be willing to part with Veteran players in those positions and target prospects at other positions

[1]Source: MLB Prospect Rankings

Recommendations
Top Mariners Prospects to Target

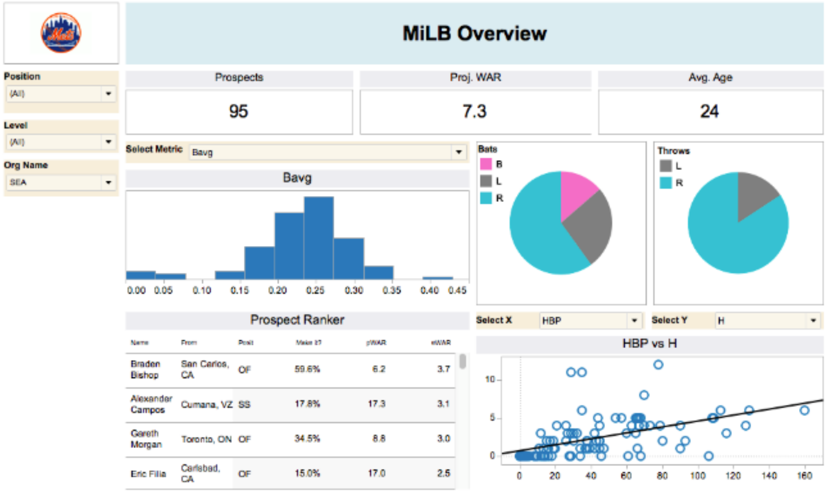
Top 15 Mariners Propsects by eWAR					
Player	**Position**	**Level**	**pWAR**	**pMade it**	**eWAR**
Braden Bishop	OF	AA	6.2	60%	3.7
Alexander Campos	SS	Rk	17.3	18%	3.1
Gareth Morgan	OF	A+	8.8	35%	3.0
Eric Filia	OF	AAA	17.0	15%	2.5
Donnie Walton	SS	A+	5.9	38%	2.2
Tyler O'Neill	OF	AAA	9.0	21%	1.9
Gianfranco 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

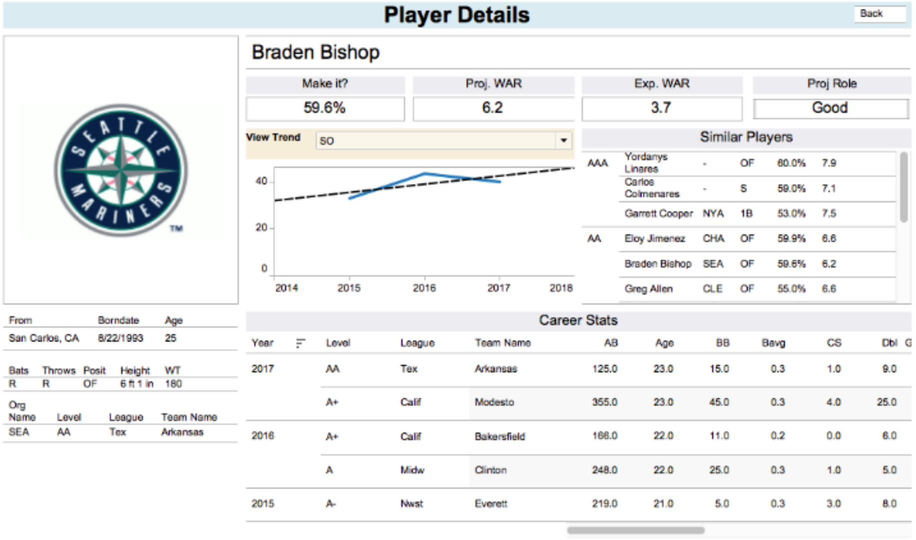
Observations

- A total future value metric, eWAR, was created based on the “Made It” (pMade it) and WAR (pWAR) models
- Braden Bishop, Alexander Campos, and Gareth Morgan have the highest eWAR’s within the Mariners organization
- However, given the need for close to MLB ready talent, Eric Filia, Tyler O’Neill and Gianfranco Wawoe should merit additional consideration as they are already in AAA

Dashboard & Mobile Application
Dashboard

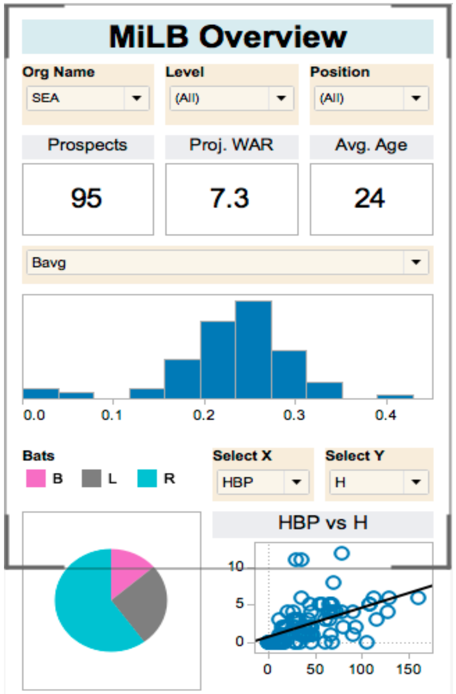
To aid in the management's use of these models, we have developed a dashboard and mobile application

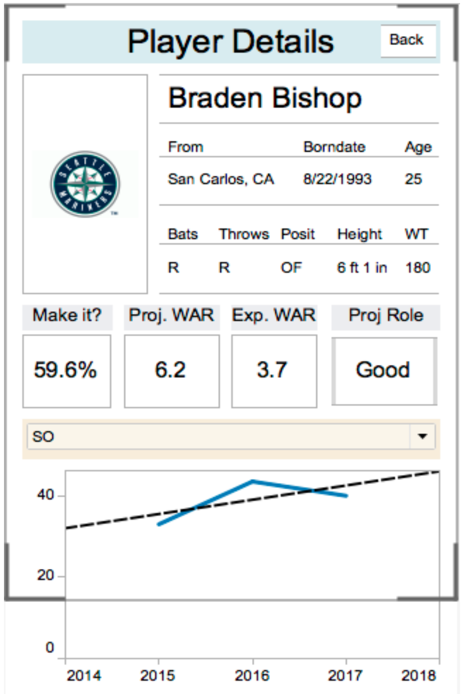




Dashboard & Mobile Application
Mobile Application

To aid in the management's use of these models, we have developed a dashboard and mobile application





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

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

Q&A