EDUARDO VASQUEZ-VILLALPANDO - 11/18/2020

Bayesian Analysis in Major League Baseball

Predicting Prospect Trajectory

Research Question

- How can Bayesian Analysis help predict the career trajectory of a professional baseball player?

High School/College → Draft → Minor Leagues → Majors

Significance

- Baseball is an analytics driven sport.
- Better predicting a prospect's trajectory has obvious benefits for a team:
  - **Scouting**
  - **Game strategy**
  - **Business**

Data Used

- Data Courtesy:
    - Lahman's Baseball Archive - Major League player info
    - Michael Lee (https://www.mikelee.co/projects/) - Scraper for Minor League data
    - Baseball Almanac - Tables of draftees
    - Baseball-Reference - Baseball stats

Data Used

- Focused on non-pitchers
- Draftees from 2007-2010, first year of play.
- LOTS of data manipulation

1. Draft Data

Number	PlayerName	DraftedBy	POS	DraftedFrom
1	Bryce Harper	Washington Nationals	OF	College of Southern Nevada
1	Stephen Strasburg	Washington Nationals	RHP	San Diego State University
1	Tim Beckham	Tampa Bay Rays	SS	Griffin (Griffin,GA)
1	David Price	Tampa Bay Rays	P	Vanderbilt University
2	Jameson Taillon	Pittsburgh Pirates	RHP	The Woodlands High School (The Woodlands, TX)
2	Dustin Ackley	Seattle Mariners	CF	University of North Carolina
2	Pedro Alvarez	Pittsburgh Pirates	3B	Vanderbilt University
2	Mike Moustakas	Kansas City Royals	3B	Chatsworth High School (Chatsworth,CA)
3	Manny Machado	Baltimore Orioles	SS	Brito Miami Private School (Miami, FL)
3	Donavan Tate	San Diego Padres	CF	Cartersville High School (Cartersville, GA)

2. Nonpitchers only,

Primitive stats added

PlayerName	POS	major_league_player	G	PA	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB
Bryce Harper	OF	Yes	109	452	387	63	115	24	2	17	58	26	7	59
Manny Machado	SS	Yes	9	39	36	3	11	1	1	1	5	0	0	3
Christian Colon	SS	Yes	60	271	245	38	68	12	2	3	30	2	4	13
Delino DeShields	CF	Yes	18	83	76	14	22	6	1	0	8	5	1	6
Michael Choice	CF	Yes	30	130	109	21	29	10	2	7	26	6	1	17
Yasmani Grandal	C	Yes	8	33	28	4	8	1	0	0	1	0	1	4
Jake Skole	CF	No	65	261	229	36	59	11	2	2	32	9	4	28
Josh Sale	RF	No	60	239	214	24	45	11	3	4	15	4	3	23
Kolbrin Vitek	2B	No	68	286	244	37	66	16	4	4	33	17	3	33
Kellin Deglan	C	No	32	121	110	12	21	2	1	1	9	0	0	9

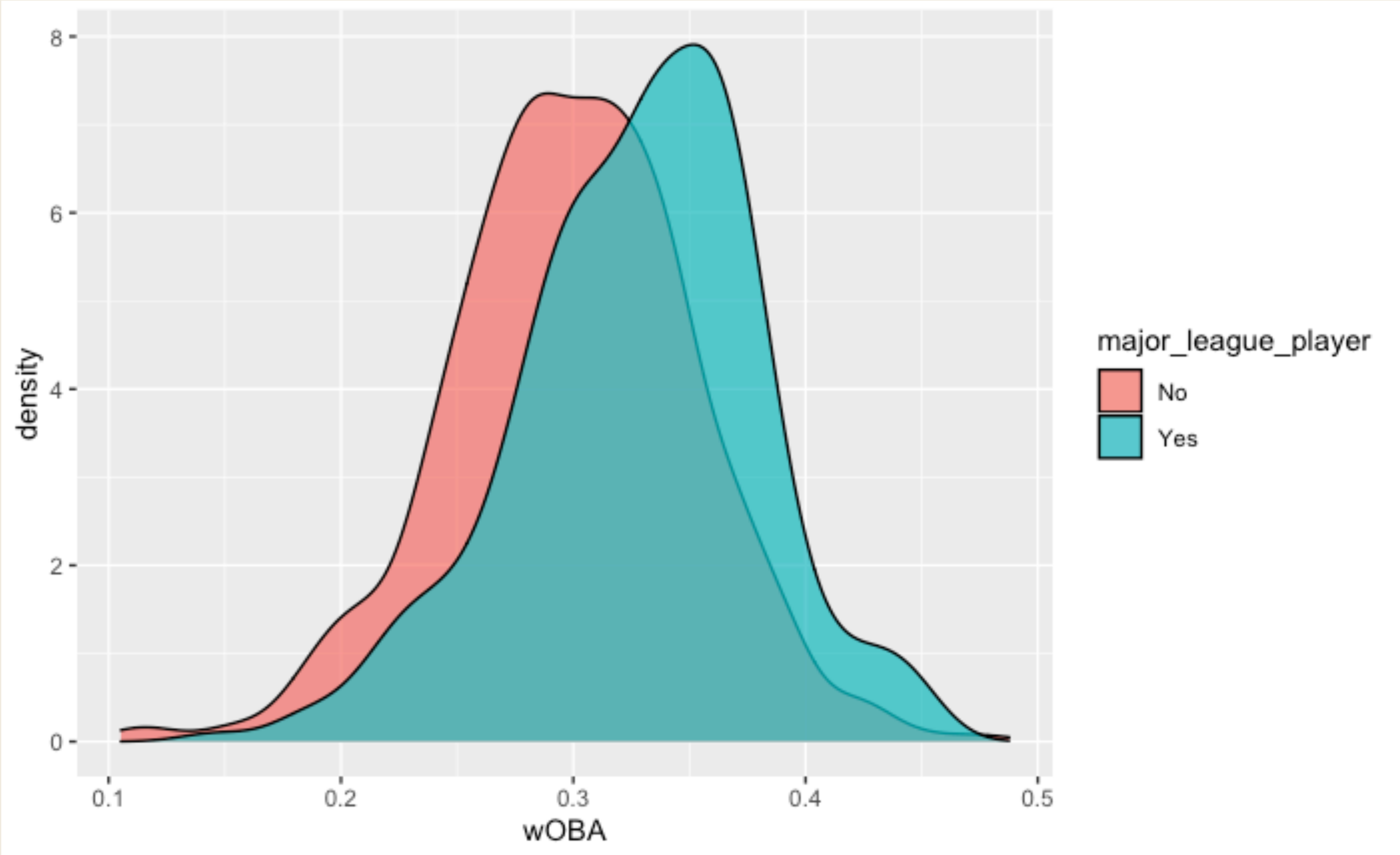
3. Converted to Advanced Stats

(Generally more accepted predictors)

```
advanced_stats<- complete %>%
  filter(AB >= 20) %>%
  mutate(StrikePercentage = round(SO / AB, 4),
         ISO = SLG - BA,
         wOBA = ((.69*(BB-IBB)) + (.719*HBP) + (.87*(H - X2B - X3B - HR)) + (1.217*X2B) + (1.529*X3B) + (1.94*HR))
         / (AB + BB - IBB + SF + HBP)) %>%
  select(c(PlayerName, major_league_player, POS, ISO, BA, wOBA, StrikePercentage))
```

	PlayerName	major_league_player	POS	ISO	BA	wOBA	StrikePercentage
1	Bryce Harper	Yes	OF	0.204	0.297	0.3737005	0.2248
2	Manny Machado	Yes	SS	0.166	0.306	0.3516923	0.0833
3	Christian Colon	Yes	SS	0.102	0.278	0.3024286	0.1347
4	Delino DeShields	Yes	CF	0.106	0.289	0.3135060	0.2632
5	Michael Choice	Yes	CF	0.321	0.266	0.3930620	0.4128
6	Yasmani Grandal	Yes	C	0.035	0.286	0.3268485	0.1429
7	Jake Skole	No	CF	0.091	0.258	0.2993605	0.2489
8	Josh Sale	No	RF	0.136	0.210	0.2718692	0.1916
9	Kolbrin Vitek	No	2B	0.148	0.270	0.3365490	0.3033
10	Kellin Deglan	No	C	0.064	0.191	0.2282810	0.2545

Exploratory Visualization



Weighted On-Base Average (wOBA)

The Model

- Naive Bayes Classifier, or *"idiot bayes"*
- **Two Key Assumptions:**
  1. **The predictors are "conditionally independent" of other predictors**
  2. **Continuous variables are normally distributed**

```
advanced_stats %>%
  filter(major_league_player == "Yes") %>%
  select_if(is.numeric) %>%
  cor() %>%
  corrplot::corrplot()
```

Running the correlation function across all numeric variables

The Model

Why use Bayes in the first place?

1. Simple, computationally inexpensive
2. Despite violating its assumptions, works reasonably well

Model in Action

- Used the “caret” package for naive Bayes model
- Split the data sets into two: testing and training
- Model using Primitive Stats vs Advanced Stats

PlayerName	POS	major_league_player	G	PA	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB
Bryce Harper	OF	Yes	109	452	387	63	115	24	2	17	58	26	7	59
Manny Machado	SS	Yes	9	39	36	3	11	1	1	1	5	0	0	3
Christian Colon	SS	Yes	60	271	245	38	68	12	2	3	30	2	4	13
Delino DeShields	CF	Yes	18	83	76	14	22	6	1	0	8	5	1	6
Michael Choice	CF	Yes	30	130	109	21	29	10	2	7	26	6	1	17
Yasmani Grandal	C	Yes	8	33	28	4	8	1	0	0	1	0	1	4
Jake Skole	CF	No	65	261	229	36	59	11	2	2	32	9	4	28
Josh Sale	RF	No	60	239	214	24	45	11	3	4	15	4	3	23
Kolbrin Vitek	2B	No	68	286	244	37	66	16	4	4	33	17	3	33
Kellin Deglan	C	No	32	121	110	12	21	2	1	1	9	0	0	9

Primitive Stats Table

Bad News

- The frequentists win this time
- The model's accuracy is *worse* than if we just predicted "No" on every single player!
- We'd be correct 80.82% of the time

Percentage of draftees that make it to the majors within 10 years <dbl>
0.19179

Our priori probability of making it to the majors is 19.18%

Conclusion

- Neat, simple idea
- Lost to the null (Priori > predicted)
- Might be useful for other data sets
- Could use better, more advanced baseball stats as predictors