Jason Eubank

08/11/2025

Using Statistical Models to Identify Value in MLB Data: Handout

1. Background

This project focuses on identifying players who present value opportunities based on their recent and season long performance metrics. The goal is to use statistical analysis to highlight high potential outcomes. The example used will be MLB Home Runs, but the process is not limited to that.

2. Preliminary Definitions

- HR/PA - Home Runs per **Plate Appearance**. How often a batter hits home runs to opportunities at the plate.
- League HR/PA - Average home run rate per plate appearance across all MLB batters for a given period.
- HR/IP - Home Runs allowed per **Inning Pitched**. How often a pitcher allows home runs per inning.
- League HR/IP - Average home runs allowed per inning pitched across all MLB pitchers.
- Barrel % - Percentage of batted balls allowed by a pitcher that are classified as **barrels** (optimal exit velocity + launch angle).
- Hard Hit % – Percentage of a batter's batted balls hit at 95+ mph.
- Index - The player's HR/PA (or HR/IP) divided by the league average for that stat. An **index** above 1.0 indicates above-average performance (batters) or above-average susceptibility (pitchers).

3. Cross-Matching to Identify Value

The following code represents the algorithm that was used to gather the full season and recent data, calculate individual player stats, compare them to league average, then identify which batters were above average and which pitchers were extremely susceptible:

```
# Read in data

library(readr)
batters <- read_csv("~/Downloads/stats.csv")
batters_recent <- read_csv("~/Downloads/savant_data.csv")
pitchers <- read_csv("~/Downloads/stats (1).csv")

# Filter out players with missing data
library(tidyverse)
batters_recent <- batters_recent %>%
   filter(player_id != 660644) %>%
   mutate(hr_per_pa = hrs / pa)

# Create new columns
batters <- batters %>%
```

```
   mutate(hr_per_pa = home_run / pa)

pitchers <- pitchers %>%
   mutate(hr_per_ip = home_run / p_formatted_ip)

batters_recent <- batters_recent %>%
   mutate(avg = sum(hrs) / sum(pa),
          index = hr_per_pa / avg)

batters <- batters %>%
   mutate(avg = sum(home_run) / sum(pa),
          index = hr_per_pa / avg)

pitchers <- pitchers %>%
   mutate(avg = sum(home_run) / sum(p_formatted_ip),
          index = hr_per_ip / avg)

# Filter out players with low index
batters_recent_filtered <- batters_recent %>%
   filter(index > 1)

batters_filtered <- batters %>%
   filter(index > 1)

pitchers_filtered <- pitchers %>%
   filter(index > 1)
```

Example: Suppose Player A has 25 HR in 500 plate appearances: HR/PA = 0.05. League HR/PA = 0.035. Batter Index = 0.05 ÷ 0.035 = 1.43. Pitcher B has allowed 20 HR in 150 innings pitched: HR/IP = 0.133. League HR/IP = 0.09. Pitcher Index = 0.133 ÷ 0.09 = 1.48. Since both index are above 1.0, this batter-pitcher matchup is flagged as high value.

4. Home Run Predictions

```
# Combine data tables

good_batters <- inner_join(batters_filtered, batters_recent_filtered, by = "player_id")

pitchers_today <- read_csv("~/Downloads/todays_pitchers_names_only.csv")

bad_pitchers <- inner_join(pitchers_filtered, pitchers_today, by = "last_name, first_name")


# Select relevant columns

good_batters <- good_batters %>%

   select(c('last_name', 'first_name', player_id, ab, pa.x, pa.y, hit, home_run, hard_hit_percent))

bad_pitchers <- bad_pitchers %>%

   select(c('last_name', 'first_name', player_id, p_formatted_ip, home_run, hr_per_ip,
        barrel_batted_rate, exit_velocity_avg, batting_avg))
```

After getting these two table and inserting to R, you can use any MLB schedule to highlight good batters that are versus bad pitchers. If you have trouble determining which players to choose even after the list is filtered, you can use other metrics of your choice. For my example, hard hit percentage would be used. Below are the results from running the model on 08/09/2025.

A	B	C	D	E	F	G	H	I	J
	last_nam	player_id	ab	pa.x	pa.y	hit	home_ru	hard_hit_percent	
1	Soto, Juar	665742	402	496	24	100	26	55.8	
2	O'Hearn,	656811	324	375	13	90	14	48.3	
3	Volpe, An	683011	408	456	23	90	17	44.2	
4	RodrÃgue	677594	477	518	27	121	21	45.5	
5	Soderstro	691016	418	467	20	108	21	48.7	
6	Diaz, Yan	650490	438	486	24	122	20	52.3	
7	Contreras	575929	399	456	18	103	16	48.8	
8	Larnach,	663616	370	417	13	91	15	44.9	
9	Devers, R	646240	432	525	23	110	20	56.1	
10	Raleigh, C	663728	424	500	19	105	42	50.7	
11	Goodman	696100	377	405	19	106	22	50.6	
12	Ward, Tay	621493	432	488	20	100	26	42.2	
13	Adell, Jo	666176	363	402	22	83	23	49.4	
14	Story, Tre	596115	429	458	20	112	18	48.2	
15	Grisham,	663757	327	382	22	80	20	44.9	
16	Perez, Sal	521692	428	457	24	108	20	46.9	
17	Muncy, M	571970	264	325	9	68	15	53.3	
18	Bader, Ha	664056	285	322	12	72	13	40.1	
19	Moniak, M	666160	292	314	23	78	17	46.8	
20	Arozarena	668227	424	495	25	104	23	52.4	
21	Ramirez, J	608070	419	477	24	124	23	37.9	
22	Yelich, Ch	592885	412	464	21	107	21	47.4	
23	Abreu, W	677800	320	358	19	80	20	45.8	
24	Harper, Br	547180	323	379	23	86	17	49.2	
25	HernÃ¡nd	606192	355	377	22	90	18	46.2	
26	Happ, Ian	664023	405	468	23	93	16	41.4	
27	Ozuna, M	542303	352	434	10	84	16	46.9	
28	Polanco, J	593871	325	365	16	83	18	44.4	
29	Altuve, Jo	514888	428	475	24	120	19	30.3	
30	Schwarbe	656941	422	506	23	108	40	60.8	
31	Sosa, Len	672820	350	368	24	97	13	42.6	
32	Alonso, P	624413	428	497	23	113	25	52.3	
33	Trout, Mik	545361	308	376	14	74	20	49.2	
34	Chisholm	665862	304	350	20	73	19	45.7	
35	Ramirez, A	682663	359	385	27	87	17	49.3	
36	Adames,	642715	426	489	24	99	18	44	
37	Torkelson	679529	400	460	21	97	24	44.5	
38	Langelier	669127	320	350	31	87	22	44.1	
39	Lowe, Bra	664040	350	379	16	96	20	48.2	
40	Ohtani, Sh	660271	442	521	22	122	39	57.9	
41	SuÃ¡rez, E	553993	415	467	23	99	37	50.5	
42	Neto, Zach	687263	382	416	24	102	16	46.4	
43	Seager, C	608369	310	365	23	81	16	53.4	
44	Carpente	681481	292	308	18	79	20	46.9	
45	Stowers, K	669065	372	424	32	109	25	52.3	
46	Suzuki, Se	673548	427	478	18	107	27	50.3	
47	Guerrero	665489	432	508	27	128	18	51.4	
48	Walker, C	572233	416	459	26	99	16	45.3	
49	Carroll, C	682998	392	440	21	97	23	49.5	
50	Swanson,	621020	428	464	20	105	18	47.9	
51	Marte, Ke	606466	306	359	18	87	21	48.4	
52	Barger, A	680718	324	351	25	86	17	54.2	
53	Lindor, Fr	596019	458	512	19	114	21	42.9	
54	Caminero	691406	430	464	26	109	30	49.1	

Results vs Outcome:.

M. Ozuna		J. Rodriguez		T. Ward		S. Langeliers	
							
H/AB	2/4	H/AB	2/5	H/AB	3/4	H/AB	3/4
HR	2	HR	2	HR	1	HR	1
RBI	4	RBI	3	RBI	3	RBI	3
R	2	R	2	R	2	R	2

MLB Batting Leaders

RK	PLAYER	TEAM	OPP	SCORE	H/AB	HR	R	RBI	BB
1	Marcell Ozuna	ATL	vs MIA	8-6 W	2/4	2	2	4	0
	Julio Rodriguez	SEA	vs TB	7-4 W	2/5	2	2	3	0
3	Taylor Ward	LAA	@ DET	7-4 W	3/4	1	2	3	0
	Shea Langeliers	ATH	@ BAL	11-3 W	3/4	1	2	3	1
	Matt Shaw	CHC	@ STL	9-1 W	2/3	1	2	2	1
	Michael Harris II	ATL	vs MIA	7-1 W	2/4	1	2	3	0
	Brent Rooker	ATH	@ BAL	11-3 W	2/5	1	1	4	0
	Michael Busch	CHC	@ STL	9-1 W	2/5	1	2	3	0
	Xander Bogaerts	SD	vs BOS	5-4 W	2/4	1	2	2	1
	Trent Grisham	NYY	vs HOU	5-4 W	3/4	1	1	1	0
	Luis Rengifo	LAA	@ DET	7-4 W	2/4	1	1	1	0
	Junior Caminero	TB	@ SEA	4-7 L	2/4	1	1	3	0
	Josh Bell	WSH	@ SF	4-2 W	2/3	1	1	1	1
	Max Muncy	LAD	vs TOR	9-1 W	2/3	1	1	2	1
	Rafael Devers	SF	vs WSH	2-4 L	2/2	1	1	1	2
	Shohei Ohtani	LAD	vs TOR	9-1 W	2/4	1	2	1	1
	Brenton Doyle	COL	@ ARI	5-6 L	2/4	1	1	2	0
	James Wood	WSH	@ SF	4-2 W	2/5	1	1	2	0
	Brice Turang	MIL	vs NYM	7-4 W	2/4	1	2	1	0
	Michael Taylor	CHW	vs CLE	1-3 L	2/3	1	1	1	1
	Corey Seager	TEX	vs PHI	2-3 L	2/4	1	2	1	0
	Paul DeJong	WSH	@ SF	4-2 W	2/5	1	1	1	0
	Gunnar Henderson	BAL	vs ATH	3-11 L	1/3	1	1	3	0
	Jo Adell	LAA	@ DET	7-4 W	1/4	1	1	3	0
	Ernie Clement	TOR	@ LAD	1-9 L	2/4	1	1	1	0
	Cal Raleigh	SEA	vs TB	7-4 W	1/4	1	1	3	1
	William Contreras	MIL	vs NYM	7-4 W	1/3	1	1	2	1
	Corbin Carroll	ARI	vs COL	6-5 W	1/4	1	2	1	1
	CJ Kayfus	CLE	@ CHW	3-1 W	1/3	1	2	1	1
	Pete Alonso	NYM	@ MIL	4-7 L	1/4	1	1	1	0

The model correctly predicted 3/13 High Value Batters and 18/52 Good Batters

5. **Conclusion**

   The model did pretty well with identifying matchups that had a statistical edge. Even though I used home runs as the example, the same setup could work for things like stolen bases, strikeouts, or on base percentage. The main drawbacks are that player performance can change fast and small sample sizes, especially with recent data, can throw things off if you do not read them carefully. In the future I would like to look at park factors to see how different stadiums impact results, give more weight to recent stats than season averages, build models that use multiple stats together, and test out some machine learning to make the predictions sharper.