Expected Hits Model

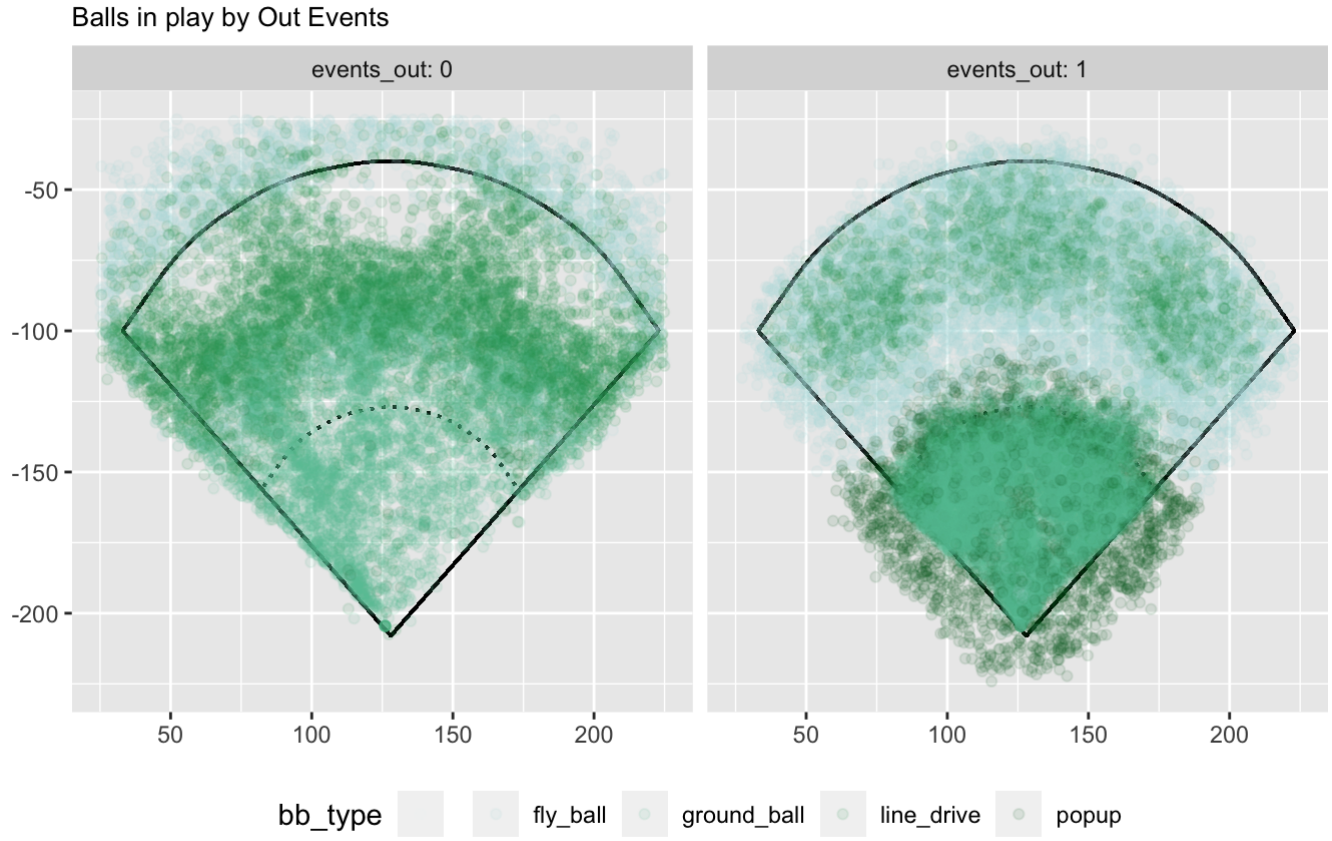
Quinn MacLean

Introduction

The purpose of this analysis is to model the probability of a hit given a ball in play. We will use this model to evaluate and find the batter that had the most hits added vs. expected. We will also see what variables contribute to a higher probability of hits as a way to compare players. We will have data from the current 2021 season through the end of May. The two months of data will be used to build our preliminary model.

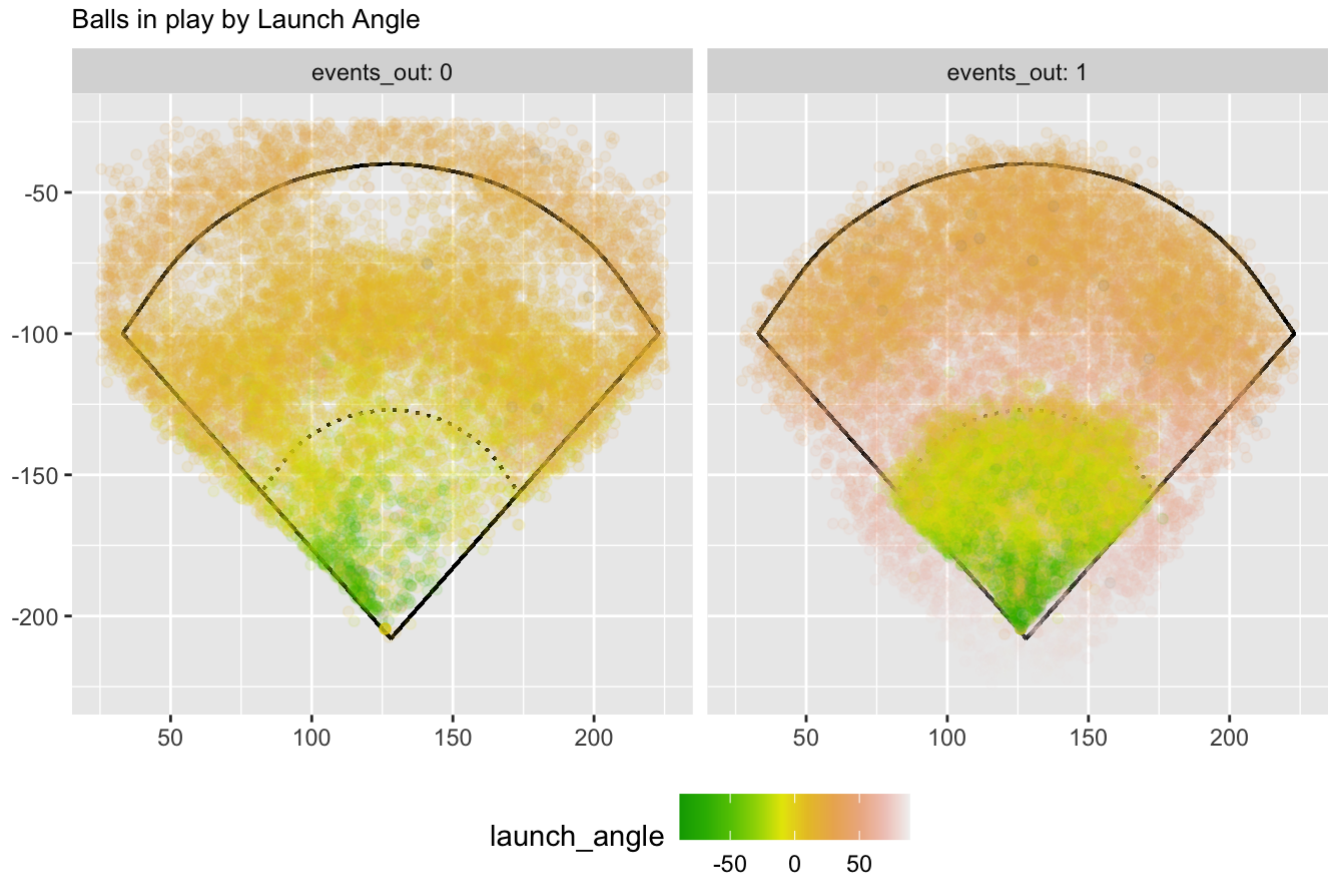
Exploratory Analysis

We see that most balls into place are shallow outfield and that those balls hit into the infield are likely outs. We can derive that line-drive hits are more likely to become hits.

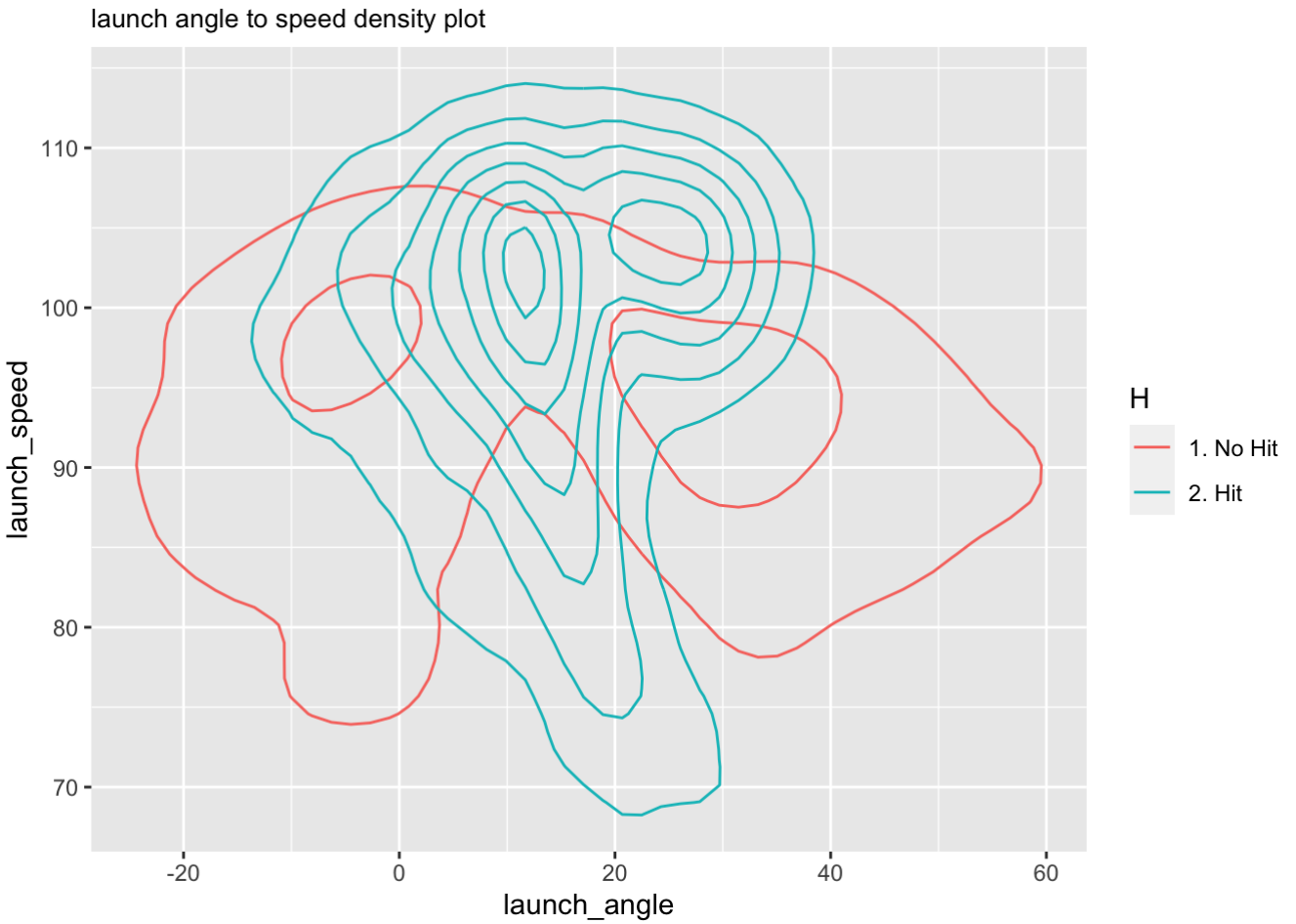


Line drives can be determined from a batter's launch angle and speed. Fangraphs derived that ground balls are less than 10 degrees, line drives are between 10-26 degrees, fly balls are 26-39 degrees and pop-ups are greater than 39 degrees. You can see the light orange below are those with the launch angle of 10 -26 degrees.

Link: https://fantasy.fangraphs.com/anglebbtypes/ (https://fantasy.fangraphs.com/anglebbtypes/)



Launch angle isn't the only variable as the exit velocity or launch speed can determine how hard the ball comes of the bat. MLB Statcast defines a hard hit is those with a launch speed of greater than 95. We can see there's a high concentration of batter's with hard hits overall.



Nick Castellanos & JD Martinez have a conistent launch angle for hard hit balls in play, which results in a higher percentage of hits as a result. Eric Hosmer has the lowest average launch angle resulting in a lower conversion of his hard hit balls resulting in hits. His average launch angle of 4 degrees is likely a hard hit ground ball, which has been fielded appropriately.

Hard Hit Percentage

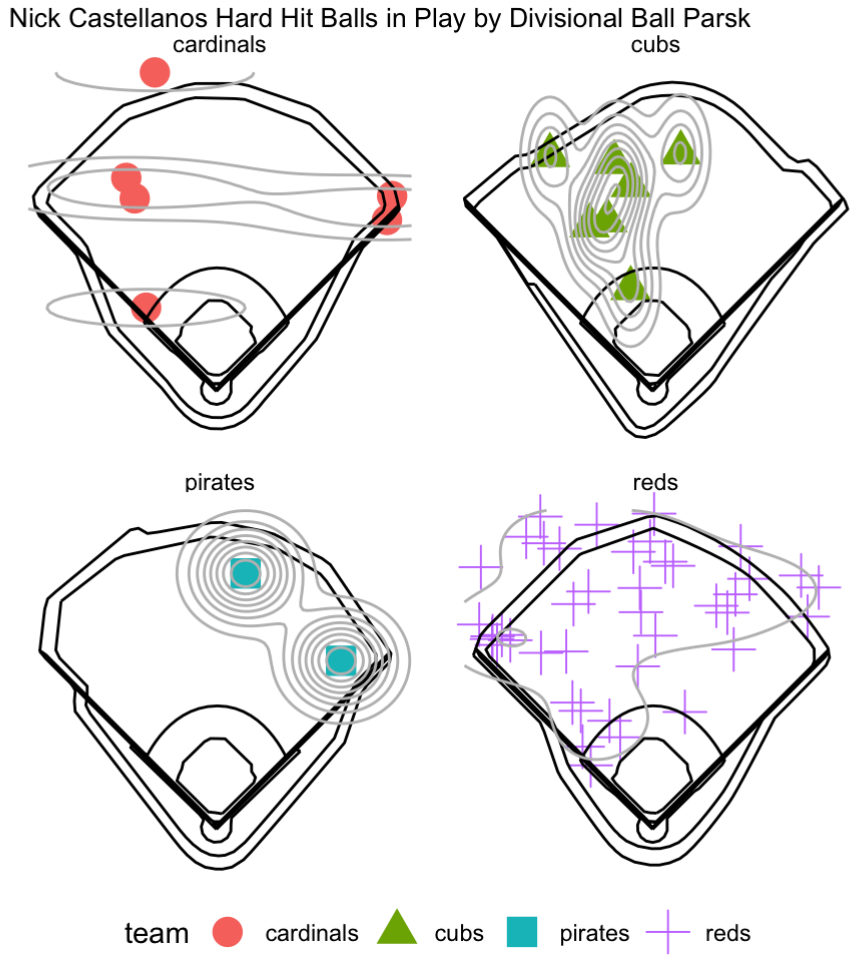
batter	PLAYERNAME	TEAM	ALLPOS	hard_hit_bip	hard_hit	hard_hit_pct	Launch Angle Median of Hard Hit BIP	Launch Angle Median of Hard Hit
592206	Nick Castellanos	CIN	OF	78	50	0.64	16.0	14.0
502110	J.D. Martinez	BOS	DH	71	44	0.62	17.0	16.0
493329	Yulieski Gurriel	HOU	1B	71	43	0.61	6.0	11.0
608385	Jesse Winker	CIN	OF/DH	78	44	0.56	13.0	13.0
656555	Rhys Hoskins	PHI	1B	71	39	0.55	16.0	16.0
521692	Salvador Perez	KC	C	85	44	0.52	16.0	17.0
571448	Nolan Arenado	STL	3B	71	37	0.52	16.0	16.0
571745	Mitch Haniger	SEA	OF	71	37	0.52	16.0	20.0
592450	Aaron Judge	NYY	OF	77	39	0.51	12.0	12.0
666182	Bo Bichette	TOR	SS	80	41	0.51	7.5	11.0
669242	Tommy Edman	STL	2B/3B/SS/OF	73	37	0.51	7.0	9.0

Note:
Filtered for at least 70 hard hits
[1] Green > 60% Hard Hit Percentage, Grey <45% Hard Hit Percenatage

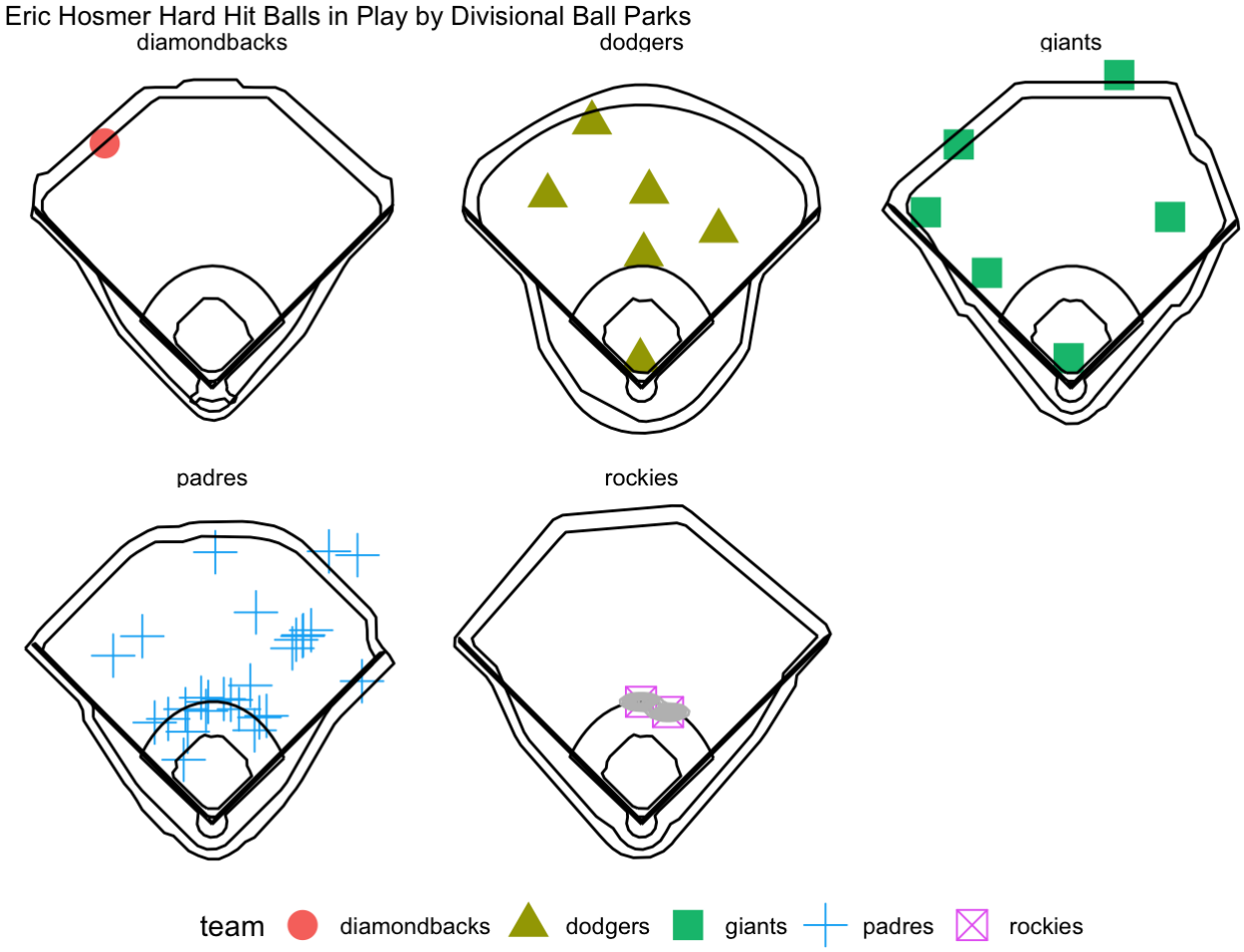
batter	PLAYERNAME	TEAM	ALLPOS	hard_hit_bip	hard_hit	hard_hit_pct	Launch Angle Median of Hard Hit BIP	Launch Angle Median of Hard Hit
446334	Evan Longoria	SF	3B	70	34	0.49	16.5	17.0
457759	Justin Turner	LAD	3B/DH	73	36	0.49	16.0	13.5
656976	Pavin Smith	ARI	1B	79	38	0.48	12.0	11.0
641857	Ryan McMahon	COL	1B/2B/3B	76	35	0.46	19.5	20.0
502671	Paul Goldschmidt	STL	1B	81	36	0.44	15.0	10.0
543333	Eric Hosmer	SD	1B	72	32	0.44	4.0	6.0
621043	Carlos Correa	HOU	SS	70	31	0.44	7.5	12.0
592518	Manny Machado	SD	3B	79	34	0.43	8.0	9.0
621566	Matt Olson	OAK	1B	76	33	0.43	17.0	17.0
572122	Kyle Seager	SEA	3B	71	30	0.42	23.0	23.5
663656	Kyle Tucker	HOU	OF/DH	81	34	0.42	14.0	14.0

Note:
Filtered for at least 70 hard hits
[1] Green > 60% Hard Hit Percentage, Grey <45% Hard Hit Percenatage

If we view Nick Castellanos hard hits solely in both home and NL Central ballparks during the 2021 season (thru end of May). We can see how much he stretches the hit at home. Even at Wrigley, he's hit hard line drives between Left & Center. His aim & velocity has contributed to his early season success.



On the contrary, we can see Eric Hosmer's hard hit ground balls at Petco. If he were to increase his launch angle by nearly ~4 to 6 degrees at least, he'd have a lot more hits.



When we view the percentage of balls in play that result to hits, not surprisingly that those hit to 7,8,9 location (outfielders) result in more hits. A defensive & athletic outfielder can do numbers to reduce hits against.

Hit Percenage to Position Location

hit_location	N	Hit	HitPct
NA	1989	1989	1.00
7	5986	3220	0.54
9	6142	3100	0.50
8	6613	2865	0.43
1	1617	208	0.13
5	4568	462	0.10
6	5266	455	0.09
4	4490	319	0.07
2	425	19	0.04
3	2829	121	0.04

Note:
Filtered for balls in play

Not surprising infield shifts work against batters who pull the ball well. In fact, if a player pulls the ball more and infield shift is more effective than that player than one who hits it opposite. Although, those who hit it opposite

would indicate they were late on the swing than those who were on top of it.

Hit Position by Infield Fielding Alignment

hit_position	if_fielding_alignment	BIP	Hit	Pct
pulled	Standard	6531	3166	0.48
pulled	Strategic	812	378	0.47
pulled	Infield shift	4292	1813	0.42
opposite	Infield shift	2799	1060	0.38
opposite	Standard	5458	2004	0.37
opposite	Strategic	711	253	0.36
center	Infield shift	5551	1140	0.21
center	Standard	11521	2398	0.21
center	Strategic	1686	354	0.21

Note:
Filtered for balls in play

Model Building

In our EDA, we can determine that hit coordinates (hc_x, hc_y), launch angle, speed and speed angle can be determined to model probability of a hit. Our first model included “batter stand”, which we can see was statistically significant and was removed from the final variables selected for our model.

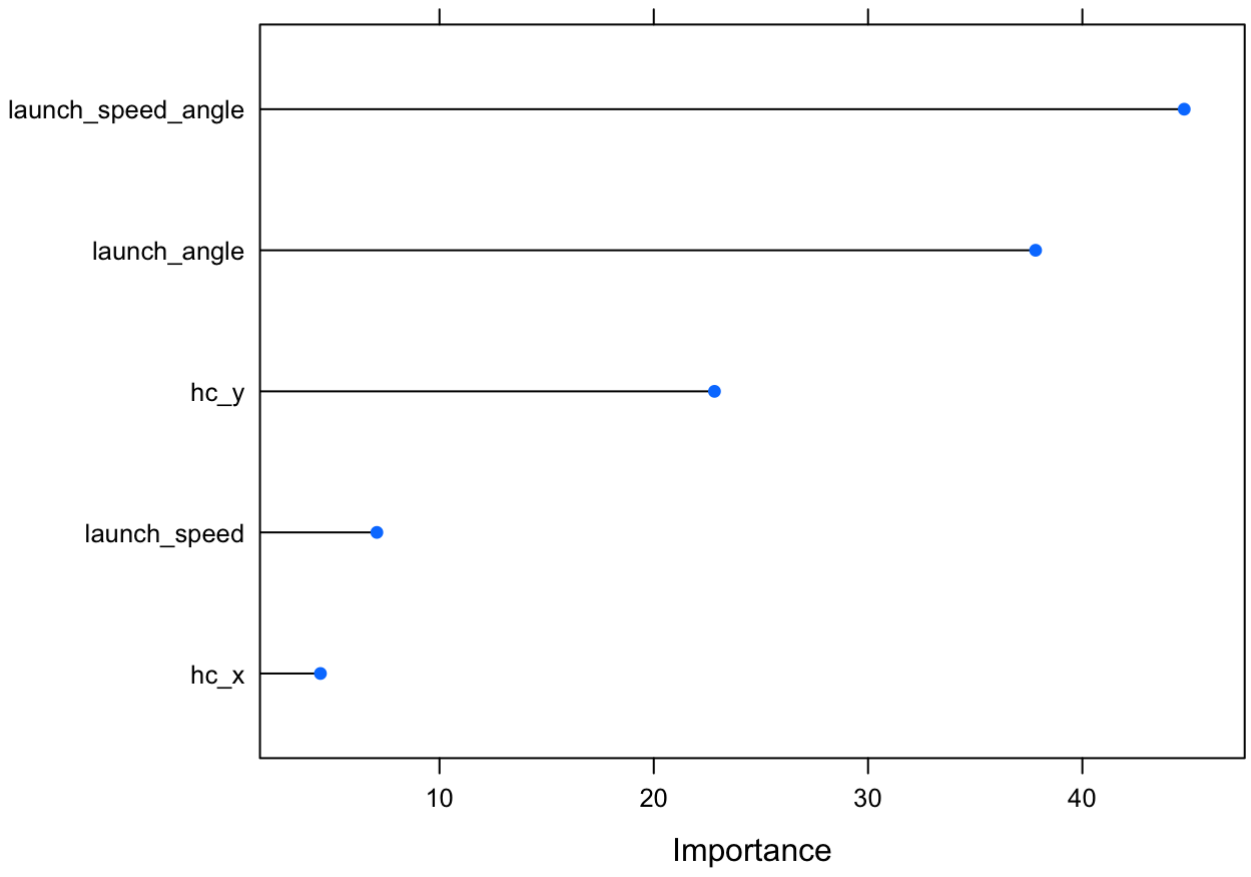
```
##
## Call:
## glm(formula = Hit ~ hc_x + hc_y + stand + launch_angle + launch_speed +
##     launch_speed_angle, family = "binomial", data = df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1281  -0.7388  -0.5221   0.8567   3.0395
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.5309803  0.1598712  -3.321 0.000896 ***
## hc_x               -0.0015648  0.0003625  -4.316 1.59e-05 ***
## hc_y               -0.0135600  0.0005938 -22.837  < 2e-16 ***
## standR              0.0159694  0.0302610   0.528 0.597693
## launch_angle       -0.0288844  0.0007639 -37.812  < 2e-16 ***
## launch_speed       -0.0090263  0.0012752  -7.078 1.46e-12 ***
## launch_speed_angle  0.8068036  0.0180302  44.747  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35212  on 28048  degrees of freedom
## Residual deviance: 28178  on 28042  degrees of freedom
## AIC: 28192
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:
## glm(formula = Hit ~ hc_x + hc_y + launch_angle + launch_speed +
##     launch_speed_angle, family = "binomial", data = df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1257  -0.7385  -0.5220   0.8561   3.0365
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.5185601  0.1581148  -3.280  0.00104 **
## hc_x               -0.0015933  0.0003585  -4.445 8.80e-06 ***
## hc_y               -0.0135573  0.0005937 -22.833  < 2e-16 ***
## launch_angle       -0.0288879  0.0007639 -37.817  < 2e-16 ***
## launch_speed       -0.0090251  0.0012752  -7.077 1.47e-12 ***
## launch_speed_angle  0.8068803  0.0180298  44.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35212  on 28048  degrees of freedom
## Residual deviance: 28178  on 28043  degrees of freedom
## AIC: 28190
##
## Number of Fisher Scoring iterations: 5
```

Now that we have a core formula, we will re-fit the models using 5 cross-fold to make sure to resample appropriately. We will first fit a GLM model to and check it's results. We see that launch speed angle & launch angle are the most important variables in the model. Secondarily, when we run a VIF on the coefficients we see no coefficient is adding extra or inflated weight, which is a good sign. If anything, launch speed angle has the most inflation at 2.34, which may contribute to its overall variable importance but we will keep in the final model.

```
## + Fold1: parameter=none
## - Fold1: parameter=none
## + Fold2: parameter=none
## - Fold2: parameter=none
## + Fold3: parameter=none
## - Fold3: parameter=none
## + Fold4: parameter=none
## - Fold4: parameter=none
## + Fold5: parameter=none
## - Fold5: parameter=none
## Aggregating results
## Fitting final model on full training set
```

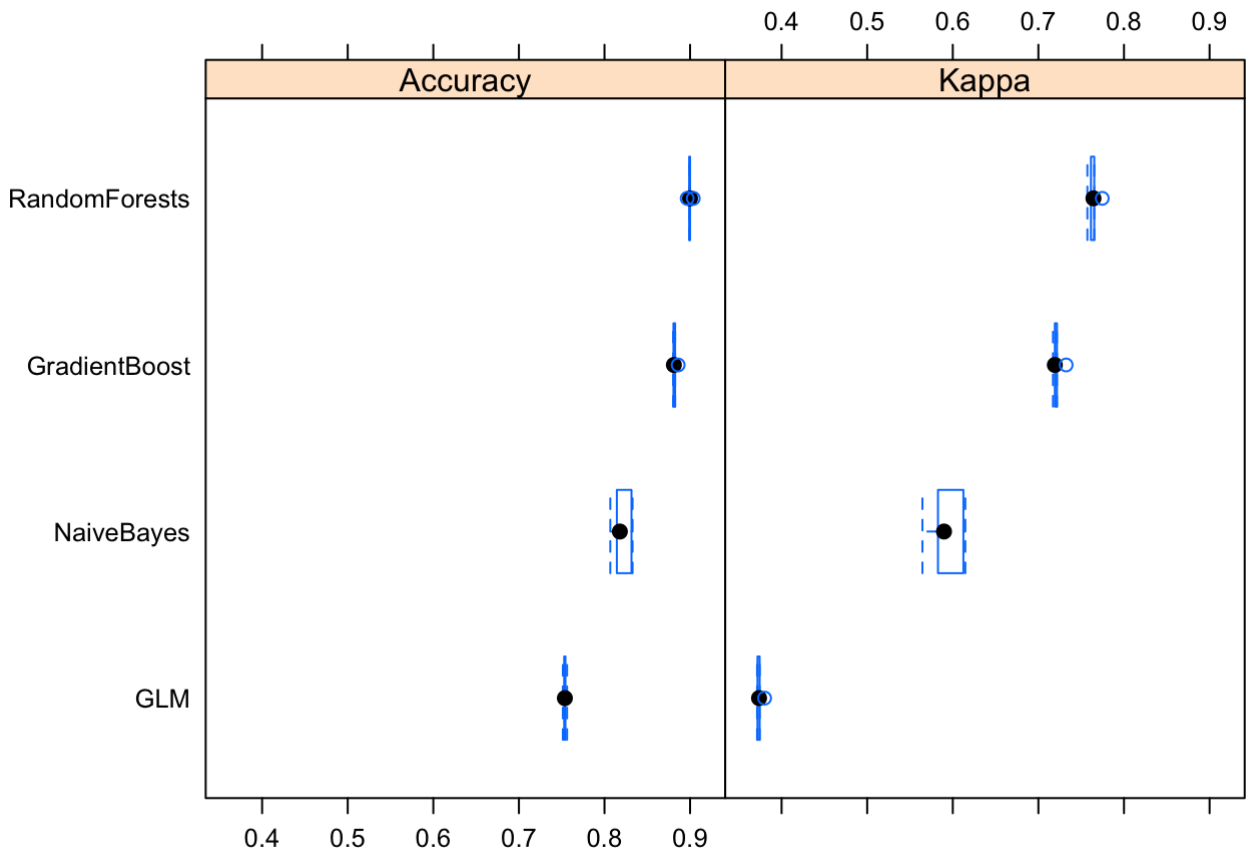
```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1257  -0.7385  -0.5220   0.8561   3.0365
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.5185601  0.1581148  -3.280  0.00104 **
## hc_x               -0.0015933  0.0003585  -4.445 8.80e-06 ***
## hc_y               -0.0135573  0.0005937 -22.833  < 2e-16 ***
## launch_angle       -0.0288879  0.0007639 -37.817  < 2e-16 ***
## launch_speed       -0.0090251  0.0012752  -7.077 1.47e-12 ***
## launch_speed_angle  0.8068803  0.0180298  44.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35212  on 28048  degrees of freedom
## Residual deviance: 28178  on 28043  degrees of freedom
## AIC: 28190
##
## Number of Fisher Scoring iterations: 5
```



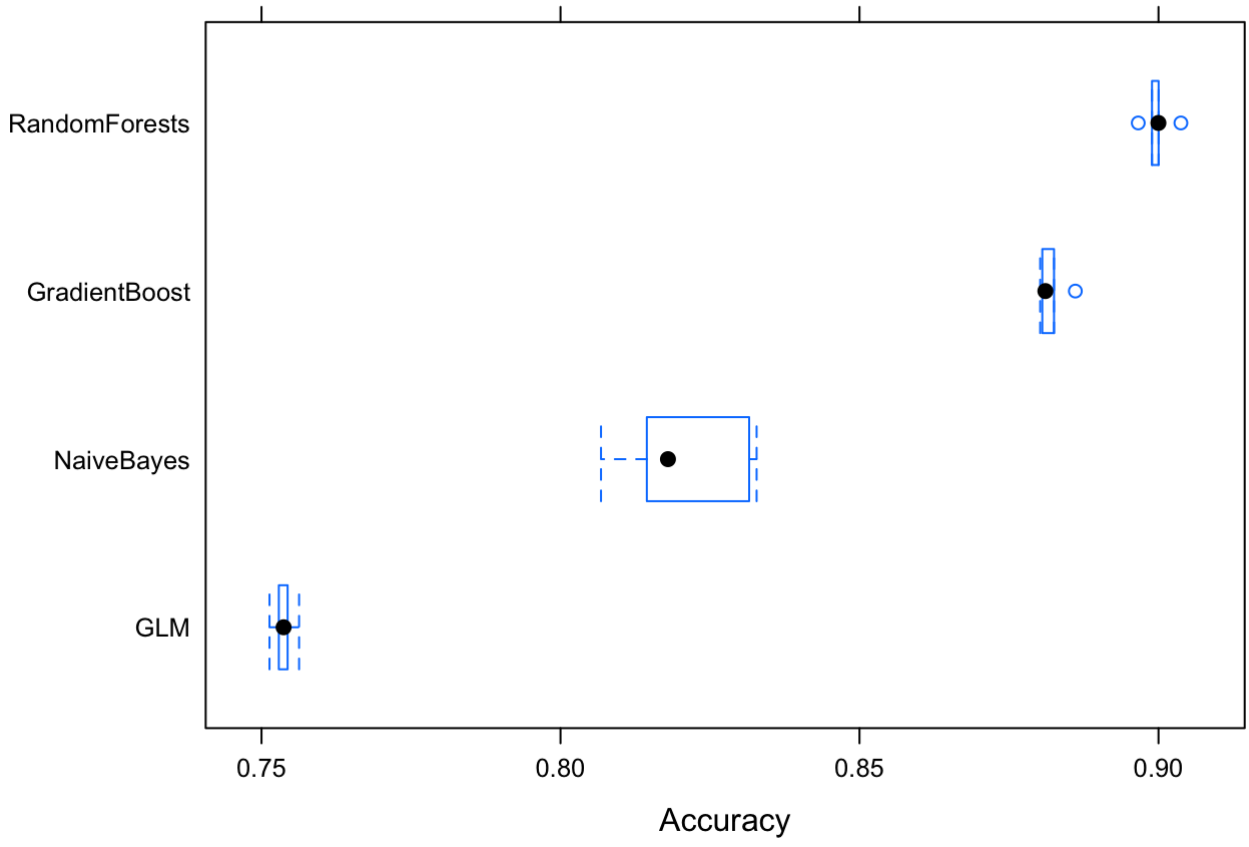
```
##               hc_x               hc_y       launch_angle       launch_speed
##           1.001911           2.476216           1.875033           1.578231
## launch_speed_angle
##           2.358369
```

Next we aim to look at other model fitting to see if we get improved accuracy. To do so, we will fit a Naive Bayes model, Gradient Boosted Model, and a Random Forest model. These help to test against simple “linear” fashion of a GLM model. Naive Bayes is a bayesian model that assumes the data set we have is the entire population of data. It is called “Naive” because it assumes all the varaibles are independent of each other. Our VIF calculations help to somewhat confirm that. Gradient Boosting uses an ensemble approach to model fitting and training, which helps to optimize the final prediction coefficients. Random Forests uses a similar appraoch in learning.

In evaluating our 4 models, we see random Forest is the most accurate in the binary classification of a hit or not at nearly ~90% accuracy. Not surprisingly, Gradient Boost was not far behind. We will be using the random forest model for our model.



Accuracy Among 5 fold CV



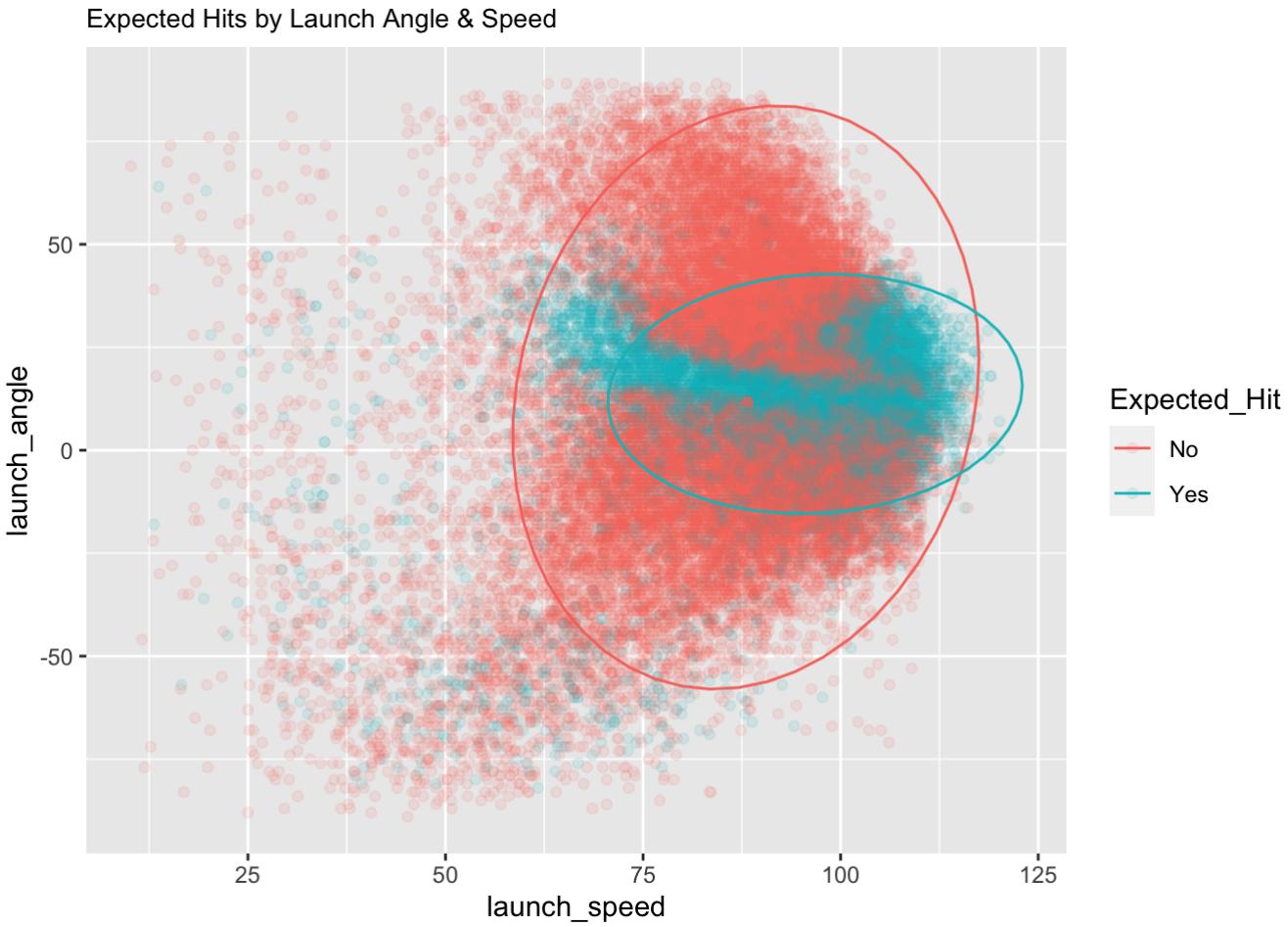
Model Application: Hits Against a shift

In apply our expected hits model, we thus validate that those hit to shallow outfield are more likely to drop for hits even against the shift. Sometimes as we can see the Strategic shift is much more successful in preventing a hit.

Similar to Infield shift we see that a strategic shift can help to limit some of the line drive hits but generally the effectiveness is about the same, which may indicate and outfield strategic shift complements and infield shift.

Model Application: Optimizing Launch Angle & Speed

This chart is likely the most important in showing range of launch angle and launch speed or exit velocity. It shows line drives (10-26) and hard hit balls (>95) result in more expected hits for balls in play.



A median Launch Angle of 14 & Launch Speed of 98 would classify it as a hard hit line drive just above the fielders heads given 10 degrees is the starting value for line drive classifications. We would suggest that batters target above 14-20 in launch angle while maintaining a hard hit ball. Easier said than done of course.

Expected Hit Metrics

			Launch Angle	Launch Speed	Launch Speed Angle
Expected_Hit	**Expected Hits**	**Hits**	**LA Median**	**LS Median**	**LSA Median**
No	27481	732	9	88.5	2.8
Yes	12444	12026	14	98.3	4.0

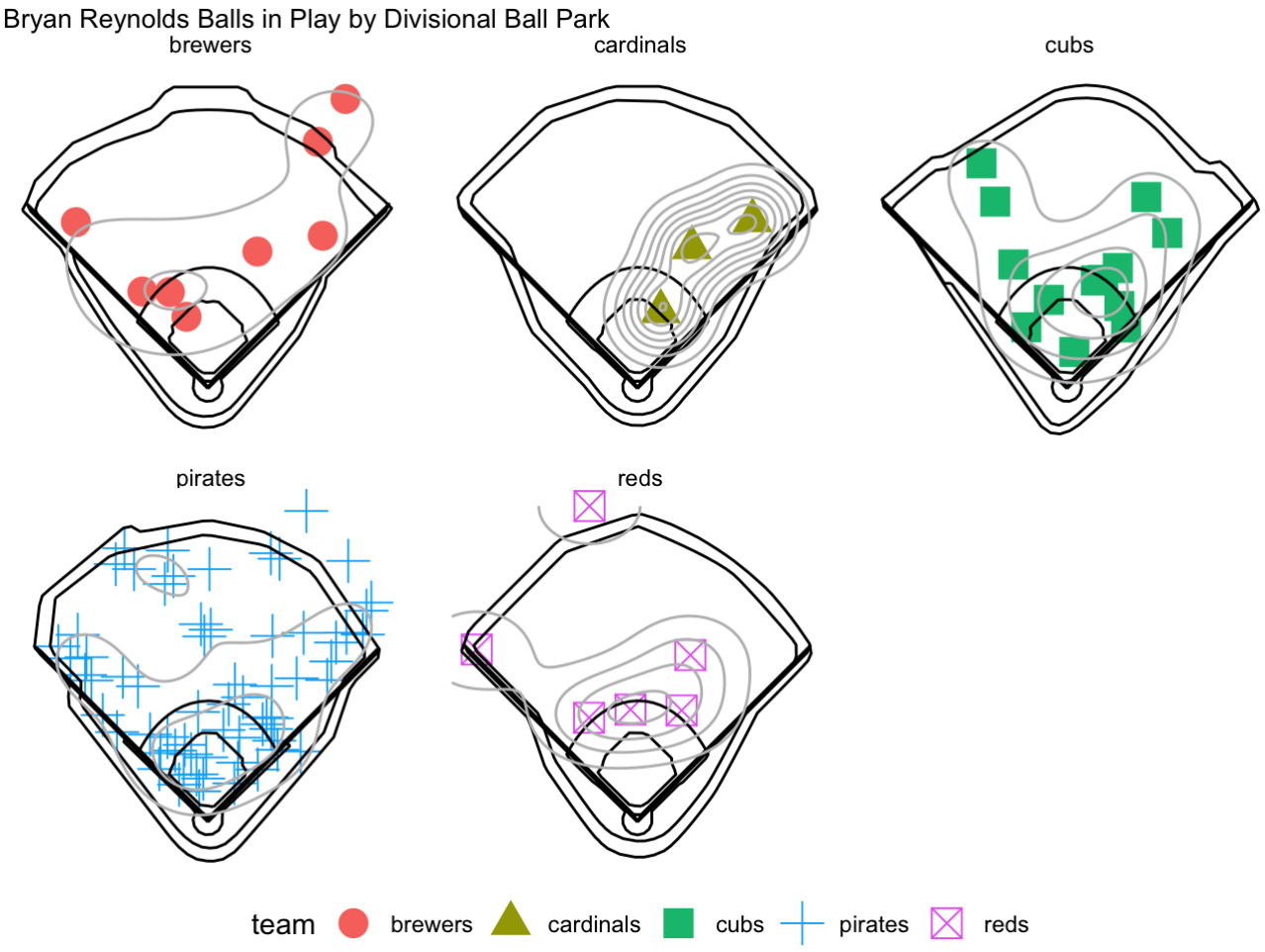
Model Application: Hits Added

Bryan Reynolds has added the most hits this year above expected at 9. This means he had a hit when the model expected a non hit.

Top Player with most Hits Added vs. Expected

batter	PLAYERNAME	TEAM	ALLPOS	HitsAdded
668804	Bryan Reynolds	PIT	OF	6

Majority of his hits are to shallow outfield, proving our theory that you can add hits even against the shift with proper launch angles. He also likely had lower launch angles and lower launch speed that may have dropped in given its specific location on the field.



Model Application: Evaluating Upcoming Free Agents

In Evaluating 2022 FAs, we see Nick Castellanos & JD Martinez would be great consistent hitters to add to the lineup w/ >60% of hard hits falling into play. Nick Castellanos is one of the youngest players on this list indicating him as a prime target to insert into the lineup. It's surprising to Corey Seager, 2020 World Series MVP so low on this list. A core reason may be opposing pitchers ability to shift and turn his hard hits into ground balls.

Hard Hit Percentage for Upcoming 2022 FAs

batter	PLAYERNAME	TEAM	ALLPOS	BIRTHDATE	hard_hit_bip	hard_hit	hard_hit_pct	Launch Angle Median of Hard Hit BIP	Launch Angle Median of Hard Hit
592206	Nick Castellanos	CIN	OF	3/4/92	78	50	0.64	16.0	14.0
502110	J.D. Martinez	BOS	DH	8/21/87	71	44	0.62	17.0	16.0
500871	Eduardo Escobar	ARI	3B	1/5/89	60	32	0.53	17.0	20.5
592192	Mark Canha	OAK	OF/DH	2/15/89	58	31	0.53	18.5	18.0
521692	Salvador Perez	KC	C	5/10/90	85	44	0.52	16.0	17.0

Note:
Filtered for at least 50 hard hits
[1] Green > 60% Hard Hit Percentage, Orange <45% Hard Hit Percenatage

batter	PLAYERNAME	TEAM	ALLPOS	BIRTHDATE	hard_hit_bip	hard_hit	hard_hit_pct	Launch Angle Median of Hard Hit BIP	Launch Angle Median of Hard Hit
571448	Nolan Arenado	STL	3B	4/16/91	71	37	0.52	16.0	16.0
592178	Kris Bryant	CHC	3B	1/4/92	61	32	0.52	18.0	21.0
500743	Miguel Rojas	MIA	SS	2/24/89	54	27	0.50	6.5	9.0
656941	Kyle Schwarber	WAS	OF/DH	3/5/93	59	28	0.47	14.0	14.0
596748	Maikel Franco	BAL	3B/DH	8/26/92	53	24	0.45	4.0	9.5
453568	Charlie Blackmon	COL	OF/DH	7/1/86	61	27	0.44	10.0	13.0
541645	Avisail Garcia	MIL	OF	6/12/91	61	27	0.44	7.0	16.0
621043	Carlos Correa	HOU	SS	9/22/94	70	31	0.44	7.5	12.0
502054	Tommy Pham	SD	OF/DH	3/8/88	51	22	0.43	16.0	12.5
519203	Anthony Rizzo	CHC	1B	8/8/89	57	24	0.42	10.0	14.5
572122	Kyle Seager	SEA	3B	11/3/87	71	30	0.42	23.0	23.5
516770	Starlin Castro	WAS	2B	3/24/90	61	25	0.41	5.0	5.0
624585	Jorge Soler	KC	OF/DH	2/25/92	57	23	0.40	11.0	13.0
608369	Corey Seager	LAD	SS/DH	4/27/94	61	24	0.39	8.0	12.0

Note:
Filtered for at least 50 hard hits
[1] Green > 60% Hard Hit Percentage, Orange <45% Hard Hit Percenatage

Application for Operations

Hard Hit balls generally result in more hits, this statistic will start to show players with great mechanics and can be used to evaluate current college baseball players ahead of the draft. We would have to look at summer leagues or college baseball data to see if this is readily available to find players with diamond in the rough mechanics. We can also likely find players who were sporatic in hard hit ability as a way to develop their tendancies in the minors.