How Much Luck Is Involved in Getting a Hit?
A Study Using Decision Trees and Random Forests to Understand the Factors Influencing Batting Average on Balls In Play

Y KIM[1], *J LIM*[2] , *R LERCH*[3]

[1,2] Department of Statistics and Data Science, Yonsei University, Seoul, South Korea

[3] College of Education, Florida State University, Tallahassee, Florida, USA

Abbreviated abstract: Is BABIP truly an indicator of luck, or is it a skill that players can systematically improve through dedicated practice? In this project, we used decision trees and random forests to build a classification model predicting balls in play as either expected hits or expected outs. Since the test accuracy of this classification model hovered around 90%, we concluded that BABIP is mostly an explainable indicator, depending primarily on how the batter hits the ball.

Related publications:

– Hothorn, T. et al., Journal of Graph Statistics 15(3), 651-674(2006)
– Breiman, L. Machine Learning 45, 5-32 (2001)

Goal and Data

"If you see trash around, pick it up, and do lots of good. Then the BABIP gods will help you."

– KBO Lotte Giants Manager MH Heo

Does BABIP represent batter's luck?

- BABIP (Batting Average on Balls in Play): A player's batting average on balls that defenders can handle.

$$BABIP = \frac{H - HR}{AB \ - K \ - HR + SF}$$

Classification model - Whether balls in play are expected to become outs or hits.

- We would claim that BABIP was primarily a luck indicator if were to observe:
  ① Low model performance ② Field situation is the most important variable
- Predictions for the 2019 season were made using the model trained on the 2018 data.

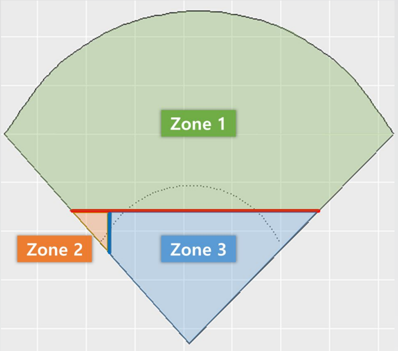
Data source and variables – Statcast (2018 season : train and test, 2019 season : test)

Predictor variables	Controllable	Batted ball characteristics	hc_x(hit location), hc_y(hit location), launch_angle, launch_speed
		Batter characteristics	sprint speed, stand(Left or Right-handed)
	Uncontrollable	Field situation	state (bases empty, runner on first, all others), if_shift (infield shift; 0 or 1)
Response variable			babip (out or hit; 0 or 1)

Methods (Trained with down-sampled data)

Decision Tree (ctree)

Launch angle	N	P(H)
25 ~	21,475	0.086
20 ~ 25	5,997	0.367
14 ~ 20	7,467	0.617
~ 14	13,579	0.918

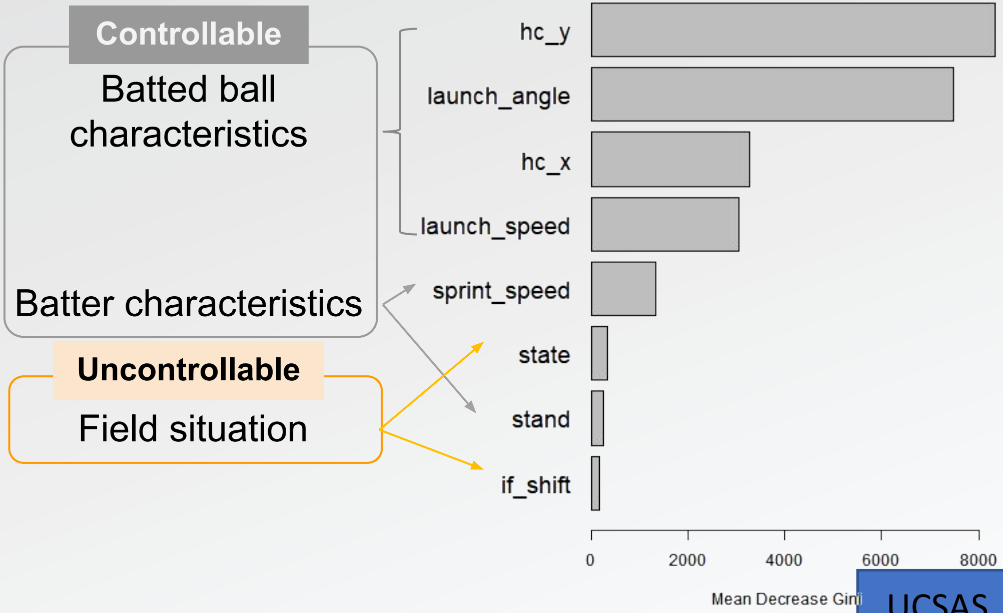


Launch angle	N	P(H)
64 ~	295	0
38 ~ 64	202	0.144
~ 38	221	0.959

Launch angle	Sprint speed	N	P(H)
53 ~ 90	-	4,046	0.002
~ 53	~ 27	13,374	0.071
	27 ~ 50	19,265	0.120

Random Forest (Test AUROC = 0.975)

Feature importance



Controllable

Batted ball characteristics

Batter characteristics

Uncontrollable

Field situation

Results and Conclusions

BABIP is explainable. Not a mystery.

- Can predict whether a batted ball will become a hit for every at-bat in a future season.
- About 90% accuracy with 2018 test data.
- Uncontrollable variables are not as influential.

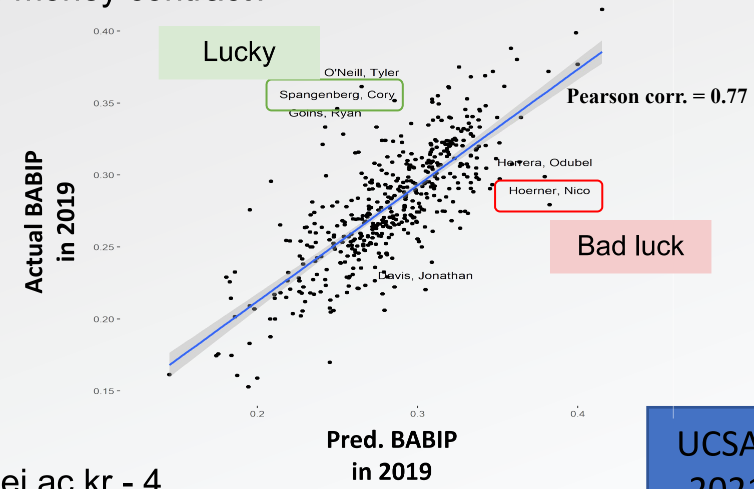
	Pred Out	Pred Hit
Actual Out	9,475	1,335
Actual Hit	860	9,181

Model failed to explain 10% of cases

- Undiscovered variables?
- Better methods for classification?
- Unexplained portion is regarded as luck.
- This is what makes baseball exciting.

Who was lucky in 2019? Who had bad luck?

- Pred. BABIP (average classification result by player) for all players with at least 50 balls in play
- Did Spangenberg have good luck by picking up all the trash that Hoerner threw around the field?
- Would you want to sign a lucky player to a big-money contract?



UCSAS 2021