Optimizing Infield Defensive Positioning from Balls in Play Data:

Ayush Arora

March 13, 2020

0.1 Abstract

The problem I want to solve is how to decrease the likelihood of an infield batted ball turning into a base hit, therefore simultaneously increasing its likelihood of turning into an out since those are the only 2 possible outcomes of an infield ball in play. I would like to solve this problem in order to minimize the number of runs, which are generated through base-hits, that the opposing team scores. This will help the Wake Forest Baseball team by decreasing their opponents' runs scored, therefore increasing their own chances of winning individual baseball games and eventually the College World Series. To tackle this problem, I implemented Unsupervised Learning, specifically K-Means Clustering, and Supervised Learning, specifically Logistic Regression. Based on my results, I was able to decrease the likelihood of an infield batted ball turning into a base-hit by strategically shifting the position of defenders in the infield.

0.2 Data Pre-Processing

The initial dataset includes all pitch by pitch data from the past 4 years (2014-2019) of Atlantic Coast Conference collegiate baseball. In order to tackle the problem stated, I pre-processed the data to only include pitch by pitch data that resulted in infield batted balls by excluding all rows whose play result was one of the following: undefined, homerun, sacrifice, fielders choice, or error.

I then converted all the features of the proper subset of data into numeric fields rather than object fields for future scaling and modeling purposes.

0.3 Feature Extraction

I evaluated our modified dataset to search for any columns containing NA values. With my prior baseball knowledge, I was aware that features such as launch angle, hang time, direction, last tracked distance, and exit speed are important features that determine the likelihood of a batted ball falling in for a base-hit. Therefore, I drop all rows that contain an NA value in any of these features to create the purest version of data. This eliminates all the NA values that may impact our results in the future.

To restrict our dataset to infield batted balls in play only, I subset our dataset to balls within the foul lines which is denoted by the direction value being within -45 degrees and +45 degrees and subset

that dataset to balls that are hit at a launch angle of 5 degrees or less and those that bounce less than 180 feet from home plate. These specific conditions were consulted by my colleagues who are part of the Wake Forest Baseball analytics program.

To hone in on the most important features that dictate play result, I ran a correlation matrix to evaluate the R-squared value of each feature of the dataset against the play result:

Variable	Correlation with Play Result (Base-Hit/Out)
PositionAt110Y	-0.454
Hang Time	-0.430
HitSpinRate	-0.285
ExitSpeed	-0.200
LaunchAngle	-0.136
PositionAt110X	-0.116
Direction	0.105
LastTrackedDistance	0.190

[1] **Direction: Left-right (horizontal) direction in which the ball leaves the bat, reported as an angle. A negative number represents a ball initially traveling toward the third base side of second base while a positive number represents a ball initially traveling toward the first base side.**

Launch Angle: How steeply up or down the ball leaves the bat, reported as an angle. A positive number means the ball is initially traveling upward, while a negative number means the ball is initially traveling downward.

Hang Time: Amount of time, measured in seconds, elapsed from when the ball hits the bat until the ball landed or would have landed had it not been caught or obstructed.

Last Tracked Distance: The distance, measured in feet, that the radar actually tracks the ball. In some cases, the radar tracks the ball only for a portion of the hit trajectory and in this case the remaining part of the trajectory is estimated by the software using a ball flight model.

Exit Speed: The speed of the ball, measured in miles per hour, as it comes off the bat at the moment of contact

[1] https://trackman.zendesk.com/hc/en-us/articles/115002776647-Radar-Measurement-Glossary-of-Terms

0.4 Feature Engineering

With the end goal in mind, I partition the balls in play into slots bound by the direction feature in order to assess current and hypothetical defensive placements. I create a slot for each conventional infielder position: Third Base, Shortstop, Second Baseman, and First Baseman. I then create a slot for each hole created by this conventional infield: the 5-6 hole, the up-the-middle hole, and the 3-4 hole.

I run a correlation matrix with these slots and the play result:

Slot	Correlation with Play Result (Base-Hit/Out)
3B	-0.071
5-6 Hole	0.107
SS	0.084
Up the Middle	0.103
2B	-0.079
3-4 Hole	0.057
1B	-0.10

These correlation coefficients align with my preconceived thoughts that a ball hit to a conventional fielder location will have a negative correlation coefficient, meaning more likely to be an out, and a ball hit into the gap would have a positive correlation coefficient, meaning more likely to be a base-hit.

0.5 Standardizing

I scale each feature in this dataset to within a range of 0 and 1 so that each feature contributes proportionally in order to create the most accurate models in the future.

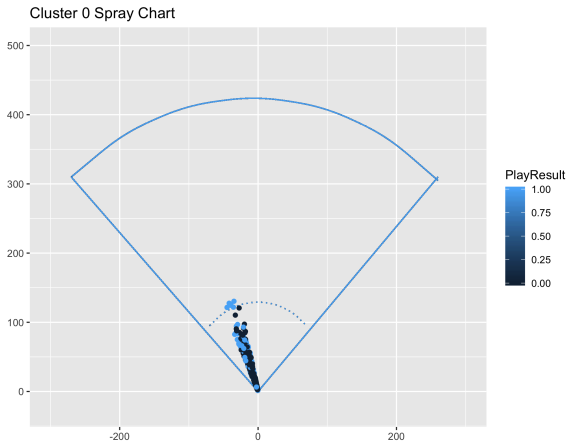
0.6 Approach

With the newly generated 7 directional slots in mind, I assign the K-Means clustering model to divide the data into 7 clusters. The algorithm agrees, via the elbow method, that assigning 7 clusters to the dataset will minimize the loss function the most and appropriately categorize balls for further research.

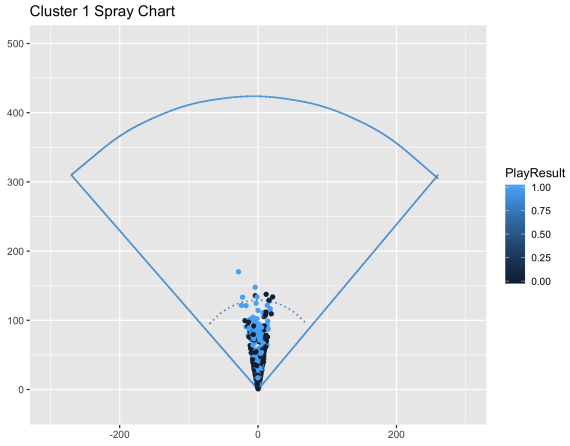
After generating 7 clusters, I deploy 7 Logistic Regression models, each training on a specific cluster, to compare and contrast the expected number of base-hits in each slot and the overall infield based on variations in defensive positioning.

0.7 Results

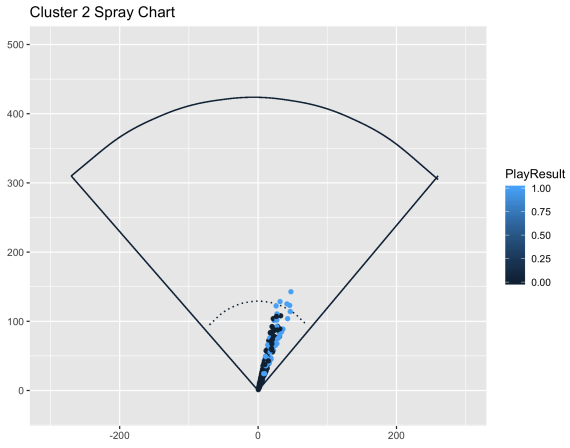
I run K-Means clustering with 7 clusters and create spray charts for each cluster on a baseball field with Direction on the X-axis, Last Tracked Distance on the Y-axis, and Play Result denoting the probability of a base-hit:



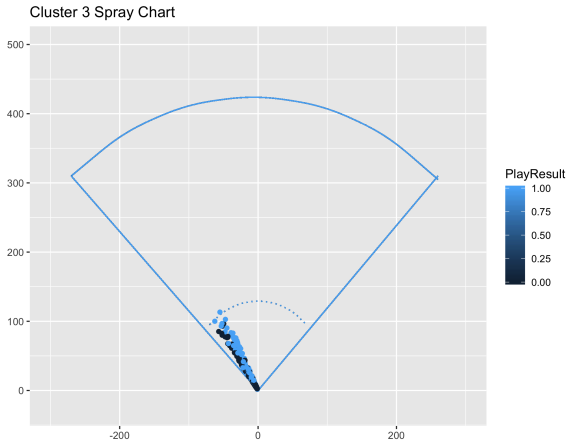
Clusters all batted balls to the the conventional shortstop



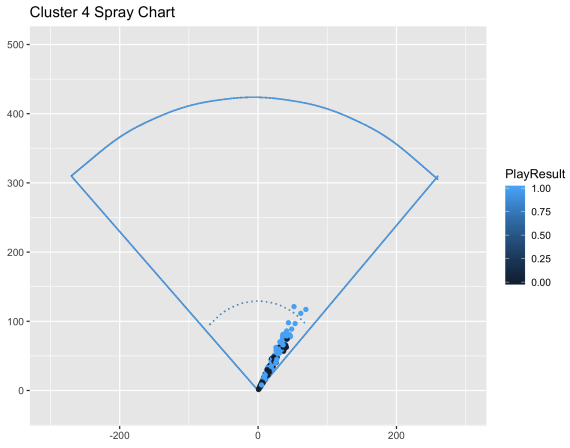
Clusters all batted balls in the up-the-middle hole



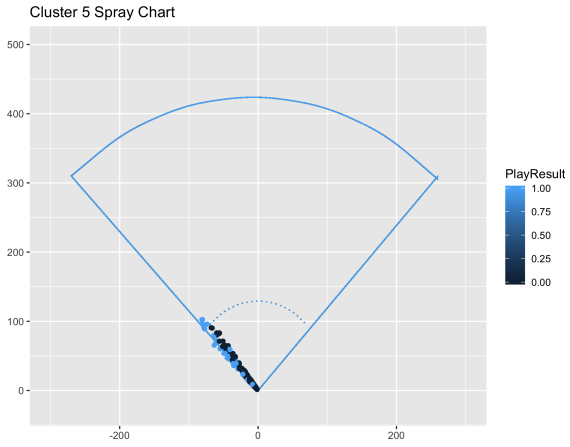
Clusters all batted balls to the conventional second basemen



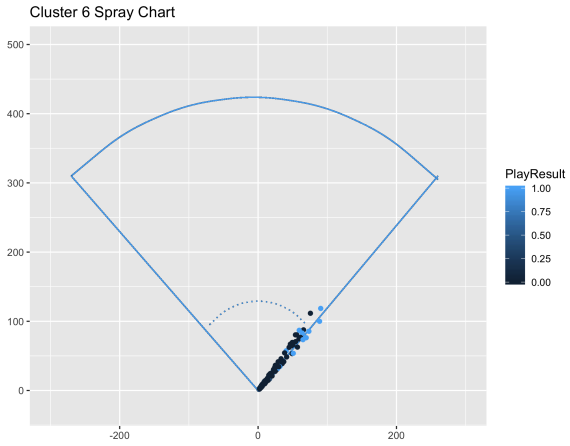
Clusters all batted balls to the 5-6 hole



Clusters all batted balls to the 3-4 hole



Clusters all batted balls to the conventional third basemen



Clusters all batted balls to the conventional first basemen

Slot Cluster and Base-Hit Probability for each:

Cluster (Slot)	Base-Hit Probabilty
3B	20%
5-6 Hole	43%
SS	21%
Up the Middle	37%
2B	21%
3-4 Hole	38%
1B	14%

The hole slots tend to have a higher base-hit probability than the conventional fielder slots

I run Logistic Regression on each slot cluster to assign a base-hit probability of either 0 (out) or 1 (base-hit) to each batted ball within each cluster:

Cluster (Slot)	Base-Hit Probabilty
3B	20%
5-6 Hole	43%
SS	21%
Up the Middle Hole	37%
2B	21%
3-4 Hole	38%
1B	14%

To Analyse my research, I pick the batter with the most number of balls in play from my dataset and separated his balls in play by slot:

Cluster (Slot)	Base-Hit% of balls in each slot %	Base Hit % for each slot
3B	9/127	4/9
5-6 Hole	13/127	6/13
SS	20/127	3/20
Up the Middle Hole	54/127	22/54
2B	19/127	8/19
3-4 Hole	5/127	3/5
1B	8/127	3/8

I calculate the probability of a batted ball being hit to a certain slot and being a base hit by multiplying the base hit % by the slot %:

Cluster (Slot)	Base-Hit% * Slot%
3B	3%
5-6 Hole	5%
SS	2%
Up the Middle Hole	17%
2B	6%
3-4 Hole	2%
1B	2%

Logistic Regression results for this specific batter only:

Cluster (Slot)	Expected Base-Hits/Total Batted Balls in Slot	Base-Hit Probabilty
3B	0/9	0%
5-6 Hole	4/13	31%
SS	2/20	10%
Up the Middle Hole	14/54	26%
2B	4/19	21%
3-4 Hole	3/5	60%
1B	1/8	13%

Expected Base-Hits from Conventional Infield: 28 Base-Hits

Since the 4 highest base-hit * slot % probabilities are in the up-the-middle hole, 2B, 5-6 hole, and 3B, I hypothetically create a shift where the shortstop will move from the SS slot to the 5-6 hole slot, the second basemen will move from the 2B slot to the up the middle hole, and the first basemen will move from the 1B slot to the 2B slot.

To see the expected Base-Hit probability for this hypothetical shift, I train the 3B logistic regression model the same, since the third basemen remains in the 3B slot. Since the shortstop has moved from the SS slot to the 5-6 hole slot, I will run the SS logistic regression model on the 5-6 hole slot and the 5-6 hole logistic regression model on the SS slot. Since the second basemen has moved from the 2B slot to the up the middle hole slot, I will run the 2B logistic regression model on the up-the-middle hole cluster and since the first basemen has moved to the 2B slot, I will run the 1B regression model on the 2B cluster. Since the first basemen has moved from the 1B slot to the 2B slot, I will run the 3-4 hole logistic regression model on the 3-4 hole cluster as well as the 1B cluster.

Cluster (Slot)	Expected Base-Hits/Total Batted Balls in Slot	Base-Hit Probabilty
3B	0/9	0%
5-6 Hole	3/13	23%
SS	2/20	10%
Up the Middle Hole	2/54	10%
2B	0/19	0%
3-4 Hole	3/5	60%
1B	0/8	0%

Expected Base-Hits from Conventional Infield: 10 Base-Hits

0.8 Conclusion

By implementing K-Means clustering and logistic regression modeling of batted balls in the infield, a strategic placement of infielder proves to be advantageous in terms of minimizing base hits on balls in play.

Full Code: https://github.com/ayusharora99/WakeForest