Final Paper Group 15

Poojha Palle, Justin Geletko, Rose Houck, Sam Pell, Jacob Rogers

2022-11-20

Introduction

Shohei Ohtani has been a force in Major League Baseball ever since he arrived on the scene in 2018. As an outstanding pitcher for the Los Angeles Angels of Anaheim, Ohtani has produced impressive performances at a high rate, but that is not what makes him unique. Ohtani displays a natural talent at the plate and has proved to be a nuisance for opposing pitchers. He has drawn comparisons to the great Babe Ruth and has quickly become one of the faces of the MLB. The following analysis of Ohtani's game intends not only to explain what makes him a generational talent but also to predict what certain areas of his performance may look like in the future with looming rule changes.

In Major League Baseball, teams have started using a strategy called the infield shift, where they place players on the field in certain positions according to the tendencies of the opposing hitters. Teams have been doing this ever since the days of the great Ted Williams, but with the emergence of sabermetrics and data-driven decision-making, teams have been using the infield shift more and more to limit the product of opposing teams. This baseball practice has sparked debate amongst league officials as to whether it should still be allowed. This debate prompted our next question: If the infield shift were to be banned, what difference would we expect to see in Ohtani's performance as a hitter? What about as a pitcher? This is an important question to consider, as being a left-handed hitter, Ohtani is one of the more shifted players in the major leagues. Our models will predict whether Ohtani may see a rise in his already outstanding numbers or if they will trend downwards, uncovering a spot of concern not only in his game but also in pitchers around the league.

Ohtani has fanned batters left and right on his way to awards such as the American League Most Valuable Player and Associated Press Athlete of the Year, both of which he received in 2021. His fast start in the MLB has proved to not be just a one or two-season wonder and has quickly turned into something alarmingly

outstanding. While people gawk at his batting abilities, he has consistently been one of the best pitchers in each of the last three years. Our second question aims to decipher how Ohtani strikes out his opponents when he has a two-strike count. His ability to put batters away when they have two strikes is a large part of his success. Our tests will uncover what types of pitches he goes to most often when in a two-strike count and which pitch characteristics explain his success.

Data

2.1 Data Collection

To answer these questions, we needed data that spanned multiple years and allowed us to compare the performance of different players against Ohtani. We decided to use the popular Baseball Savant website (https://baseballsavant.mlb.com/savant-player/shohei-ohtani-660271?stats=statcast-r-hitting-mlb) to collect our data. Baseball Savant pulls their data from the official MLB site (MLB.com)—these files include daily games and probable pitchers for upcoming games. Savant uses statcast, a state-of-the-art tracking technology, to measure previously unquantifiable aspects of the game. This tracking is accomplished using high-resolution optical cameras set up at all thirty Major League ballparks, which track the location and movements of the balls and players on the fields.

2.2 Data Summary

Currently, statcast monitors ninety-two different variables, but the following are the most prominent in our analysis:

- events: a character variable describing the type of event that took place (i.e. home-run, strikeout, field-out, etc.)
- delta_home_win_exp: which provides the change in win expectancy for the home team based on the result of the pitch observed. This can be used to answer how Ohtani impacts the team's winning percentage
- if_fielding_alignment: a categorical variable that tells us whether the infield is shifted or playing straight up. This will help us compare Ohtani's results when he faces the shift versus when he doesn't
- pitch_name or pitch_type: this is the variable that gives us pitch type. There are multiple tags such as 4-seam fastball, sinker, slider, curveball, changeup, etc. that will allow us to analyze how Ohtani

performs against certain pitches as a hitter but also how he uses them as a pitcher. We can look into how he has adjusted to certain pitches as a hitter (i.e. is he swinging at fastballs more often, is he improving his exit velocity vs sliders)

- description: this is going to give us the play result of the pitch, can be as simple as a strike_swinging but it will be essential to know the results on the field that Ohtani is producing
- game_date: will be important as we will be attempting to make some comparisons year-over-year with Ohtani's performance
- release_speed: will tell us how fast the pitch was thrown to the plate
- strikes: this will give us information on the count, does Ohtani have two strikes during this observation
- outs_when_up & inning on_1b, on_2b, on_3b: this will help us look into the situational hitting of Ohtani, does he perform better later in games or when there are two outs? Does he do better with runners on base / in scoring position?
- launch_speed & launch_angle & hit_distance_sc: this gives us information on how hard Ohtani is hitting the ball, the trajectory at which he hits it, and the distance the ball travels, which are all important in understanding the results Ohtani produces as a hitter
- pfx_x (vertical break) & pfx_z: these variables will help us to understand the movement that Ohtani is generating on his pitches, which can make them more difficult to hit
- hc_x and hc_y: Hit coordinate X and Y of batted ball

Some of the main categorical variables used in our analysis were:

- game_dates: self-explanatory, it refers to the date a certain game took place
- player_type: pitcher or batter
- start and end_date: start and end dates of games
- playerid: used to identify players (Ohtani's id is 660271)

The categorical variables named above were used heavily in our initial data wrangling, as these descriptive factors were key in separating Ohtani's statistics from other players.

2.3 Data Collection

Using the baseballr library in R developed by Bill Petti and Saiem Gilani, data was divided based on player and date; Bringing the data into R was achieved by utilizing the scrape_statcast_savant() method. The data ranged from March 29th, 2018 to April 7th, 2022. Using the said method, data was scraped into two separate datasets—ohtani_bat and ohtani_pitch. As their names suggest, these sets of data encapsulated all ninety-two variables differentiated by when Ohtani pitches and bats.

The main variable analyzed for in-field shifting was, quite obviously, the in-field alignment variable. This variable is categorical and measures whether the infield has been shifted or not. The dataset used to answer Question 1 pulls all data where the in-field alignment is not strategic; the reasoning behind this is we wanted to deal with only in-field and standard shifts. After making this edit, the data frame contains the following variables:

- game_date
- events
- pitch_type
- p_throws
- release_speed
- release_spin_rate
- release_extension
- strikes
- outs_when_up
- hit_distance_sc
- plate_x and plate_z
- launch_speed
- hc_x and hc_y
- launch_angle
- hit_distance

- if_fielding_alignment

All these variables are crucial in developing a valid conclusion as to Ohtani's performance with and without the infield shift. After that, the main variable analyzed for infield shifting was a binary variable for base hits. This was created using the "events" variable in the data set and was re-coded as a 1 if a "single", "double", "triple" or "home_run" occurred, and a 0 otherwise. Also, to filter for only balls in play, we used the type == "X" feature, which gives us the observations that have data for the "events" variable.

pitch_id	if_fielding_alignment	basehit	hit_std	hit_shift	ab_std	ab_shift
1	Standard	0	0	0	1	0
2	Standard	0	0	0	1	0
3	Standard	1	1	0	1	0
4	Infield shift	0	0	0	0	1
5	Infield shift	0	0	0	0	1
6	Standard	0	0	0	1	0

Figure 1.

The main variable that was analyzed for our second question—Ohtani and two strikes—was the type of pitch. To begin this investigation, an initial dataset of all ninety-two variables was created—the major difference was Ohtani's pitching data was filtered such that the number of strikes he performed equaled two. Since the primary variable of interest is pitch types, a frequency table is the best way to visualize how Ohtani's two-strike count differs based on the type of pitch, which we can see in Figure 2.

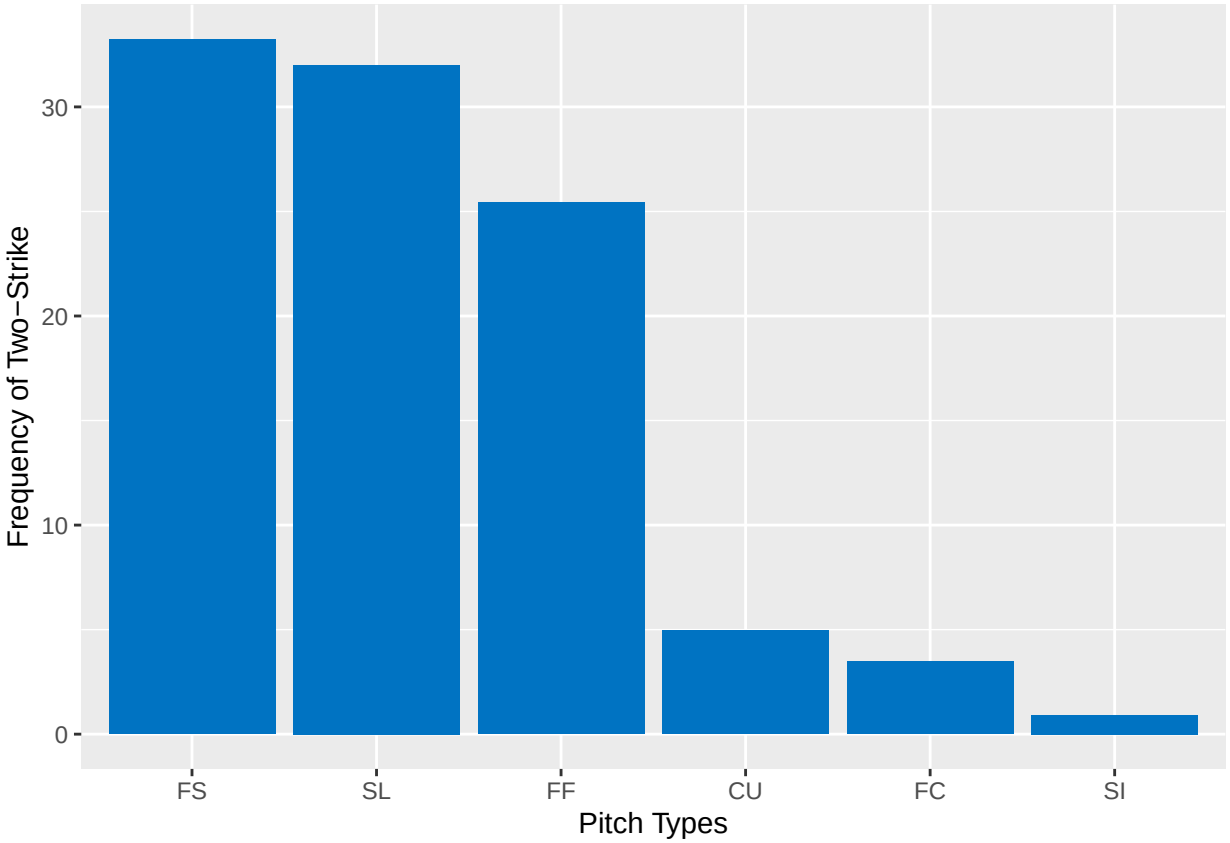


Figure 2. A frequency plot of Ohtani's two-strike count dependent on pitch type

To further simplify the data, a question was asked: within Ohtani's two-strike plays, how many were strikeouts? Figure 3 contains a visualization of Ohtani's frequency grouped by pitch type for both two-strike plays and two-strike plays that led to strikeouts. From this table alone, one can see that there is some type of correlation between the type of pitch and how it impacts Ohtani's strike count. Overall, different datasets had to be created to deal with these two unique questions.

PitchType	PitchDist	StrikeoutDist	Diff
FS	33.2315211	41.0022779	7.7707568
SL	32.0097740	33.4851936	1.4754196
FF	25.4123396	17.9954442	-7.4168955
CU	4.9480757	5.4669704	0.5188946
FC	3.4819792	1.5945330	-1.8874462
SI	0.9163103	0.4555809	-0.4607295

Figure 3. A table showcasing frequencies of Ohtani performing two-strikes and strikeouts depending on pitch type

Results

Question 1

To examine the differences in results that Ohtani has experienced against the shift versus standard defense, the first step was to examine summary statistics. Because infield shift is used more often on players with a large sample size of data, Ohtani has a decent sample size of at-bats against standard defense from when he first emerged in the league. Filtering for at-bats against the shift and at-bats against the standard defense, it was found that Ohtani has a .413 batting average against the standard defense and a .371 batting average against the shift, indicating that he has less success at getting base hits on batted balls when teams are shifted against him (Figure 4). Taking this into consideration, we then set out to predict Ohtani's results as a hitter if teams were no longer able to shift against him. First, we split the data, using the infield shift data to train the models and the standard defense data to test the models. We created a binary variable for a base hit and first used a decision tree to classify hits based on predictors such as hit location, hit distance, exit velocity, and launch angle (Figure 5). Initially, the decision tree yielded a test misclassification rate of 15.2%. After pruning the tree, we found that a size=11 tree (Figure 7) was optimal, which led to a slightly better tree with a test misclassification rate of 14.9% and a 29.67% false negative rate. These aren't the worst results, but we are still making false predictions, so we look to other methods for more accuracy.

	Standard Defense	Infield Shift
Batting Average	0.413	0.371
Hits	155.000	318.000
At-Bats	375.000	858.000

Figure 4. At-bats against Standard Defense vs. At-bats against an Infield Shift

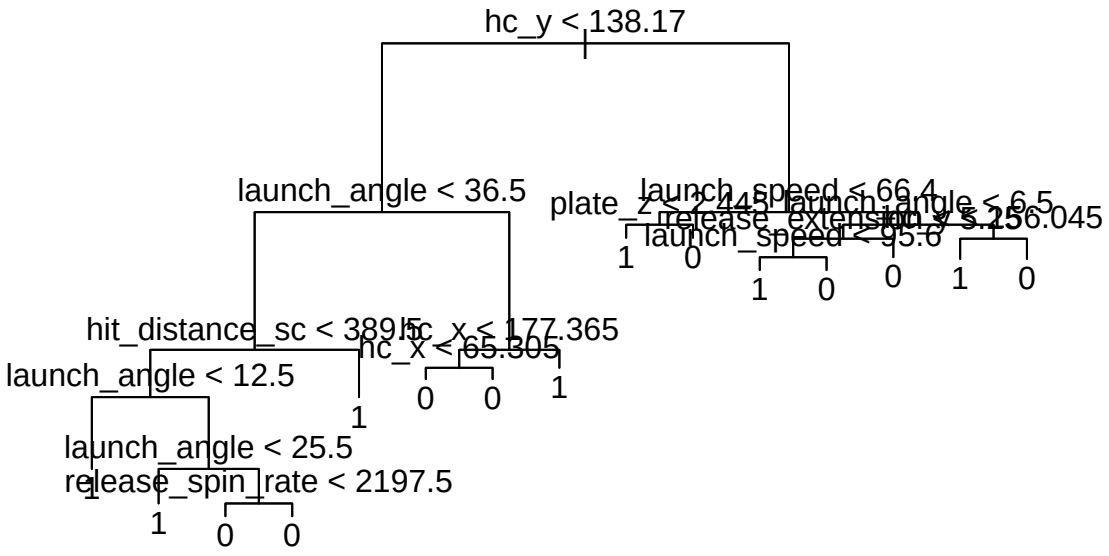


Figure 5. Decision Classification Tree without pruning

	0	1
0	400	65
1	140	253

Figure 6. Confusion Matrix

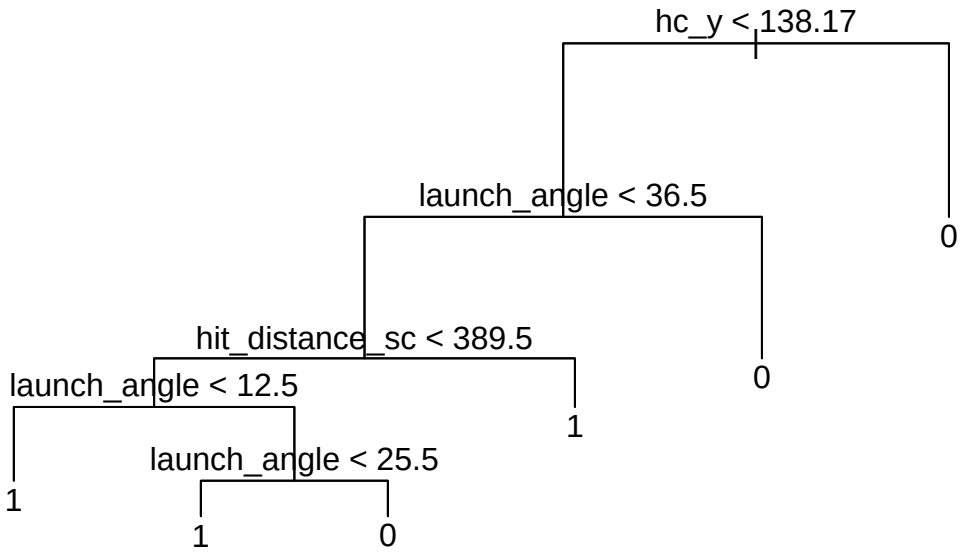


Figure 7. Pruned Decision Tree on predicted hits

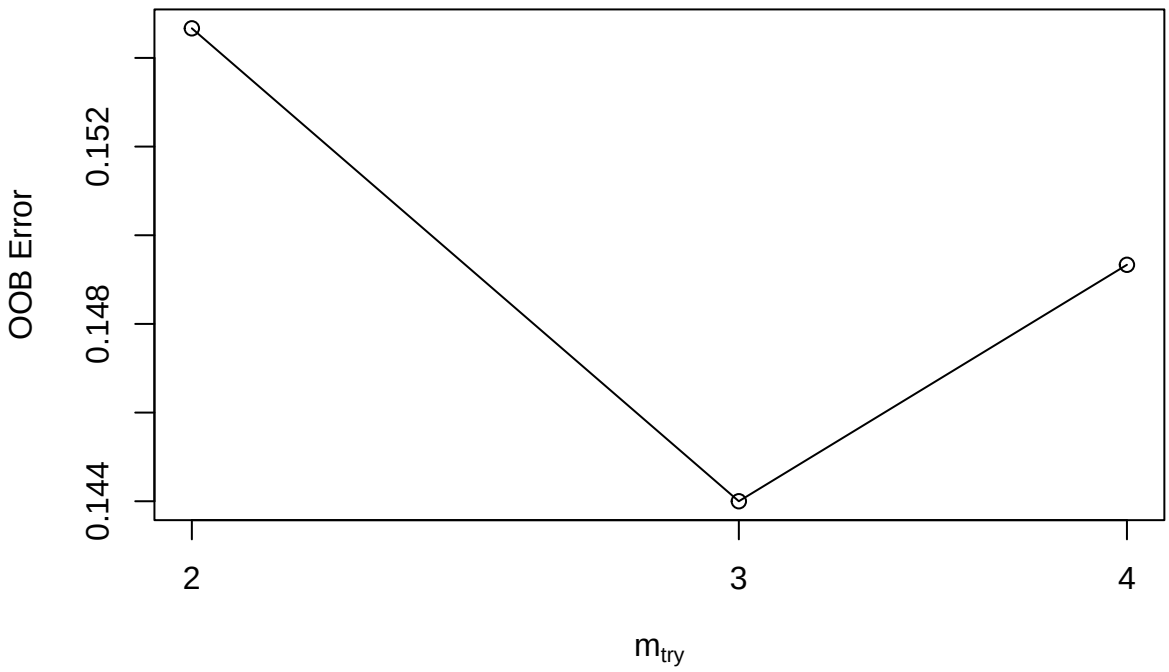
	0	1
0	453	74
1	87	244

Figure 8. Improved confusion matrix

Another model we considered for predicting Ohtani's performance is random forests. The random forest model was performed very similarly to the decision tree model, with the same predictors and the base hit binary variable as the outcome. The random forest models effectively carry out bagging, where several decision trees are built on bootstrapped training samples, but the trees are dissociated from each other. This occurs because each tree has a random number of predictors that are considered at each split, and only one of those randomly selected predictors is then used at the split. We used cross-validation to find the best number of variables randomly considered at each split (mtry=4). After finding the optimal tuning

parameter, we created a random forest with 500 trees and the OOB estimate of the error rate on the initial model returned a 12.94% rate, which improved upon the other classification tree model (Figure 9). We also plotted the mean decrease accuracy and mean decrease Gini for each variable to determine variable importance. The mean decrease accuracy measures how much accuracy the model loses when the predictor is excluded from the model. The higher the mean decrease accuracy is, the more important the variable is to the model. Similarly, variables with a high mean decrease Gini measure are more important to the model. This measures how each variable contributes to the homogeneity of the nodes in the model. Both plots noticeably show that launch_angle and launch_speed are important variables in the model, which is useful in informing our next method, the generalized additive model.

```
## mtry = 3  OOB error = 14.4%
## Searching left ...
## mtry = 2     OOB error = 15.47%
## -0.07407407 0.01
## Searching right ...
## mtry = 4     OOB error = 14.93%
## -0.03703704 0.01
```




```
## [1] 3
```

Figure 9. Optimal number of predictors tried at each split based on minimizing OOB error

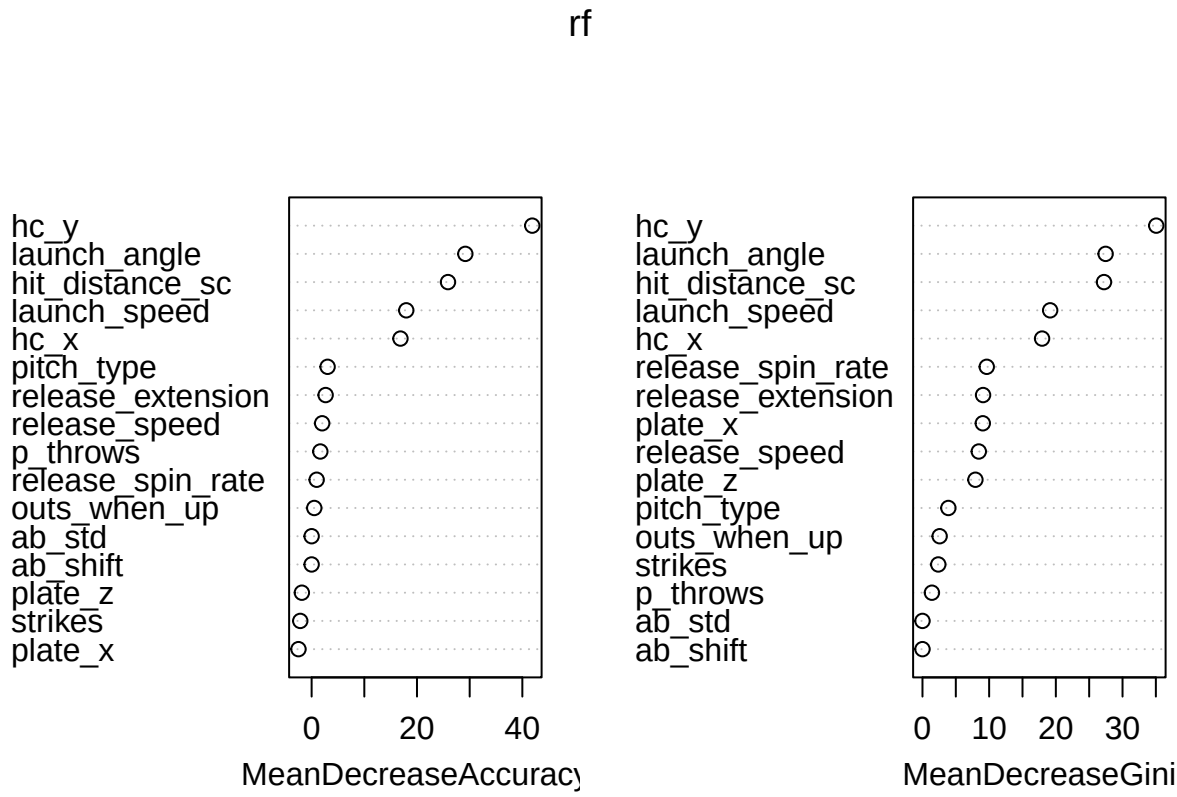


Figure 10. Variable Importance produced by random forests

The final model that was used to predict expected hits without the shift was the generalized additive model. This type of model allows for non-linear relationships to be captured by replacing beta coefficients on linear model predictors with splines. These splines are more complex, flexible functions that allow for on-linear relationships making different predictors. The two variables that were used as predictors were launch_angle and launch_speed, two of the more important variables highlighted in the random forests model. We fit the model on the training data, which was infield shifts, and made predictions on the test data to find the probabilities of a hit for each pitch. Our result gives us an estimate with a significant p-value and chi-square value of 202.7, but the deviance explained is only 35.9%. This model also resulted in the highest misclassification rate at 22.1%. Figure 11 shows the component effects on hit probability, which are non-linear relationships. Exit velocity seems to be better for getting a hit when it is higher, but it dips from 80 to 90 mph, whereas the launch angle is best when it is low or in the sweet spot between 0 and 25 degrees.

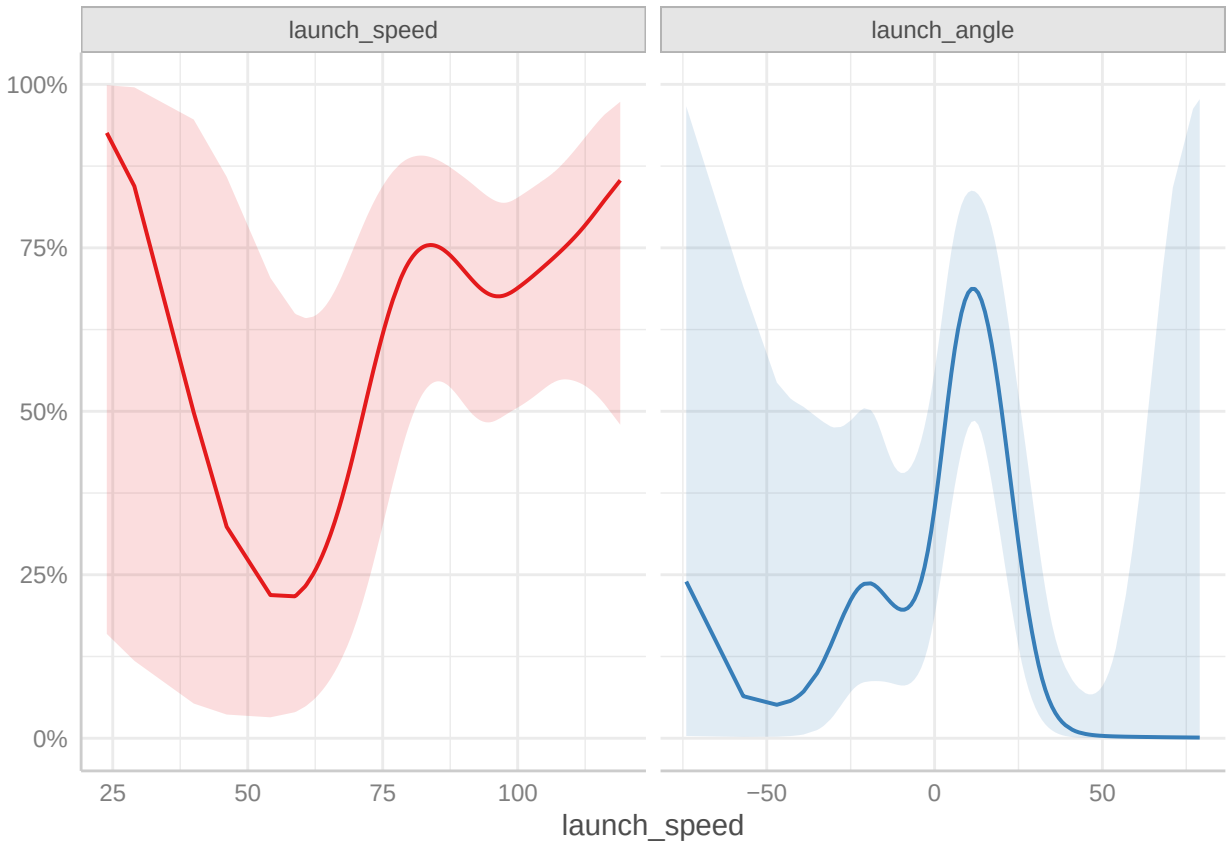


Figure 11. Component plots that illustrate the implied effects of the variables on predictions

Once the predictions for all three models were collected, we performed a two-sample proportions test to examine the differences between the predictions and the actual test set. After all, the main objective is to project if there will be differences between the predictions based on the infield shift and the actual production against the standard defense. The proportions test for all three models resulted in insignificant p-values, suggesting the difference we would expect without the infield shift is not major. The estimated batting average for the pruned decision trees, random forests, and generalized additive model was .386, .389, and .398, respectively. All of these suggest that there would be an improvement in Ohtani's batting average on balls in play if there was standard defense rather than a shift, but again, they are not statistically significant. We produced spray charts using the baseballr package to illustrate why this might be the case. These show that although he has a concentration of balls hit to the right side of the infield, most of those would not be a hit whether the defense is shifted or not. Also, it's evident that he hits the ball to all fields when he hits it in the air, so it makes sense that the shift wouldn't affect him as much.

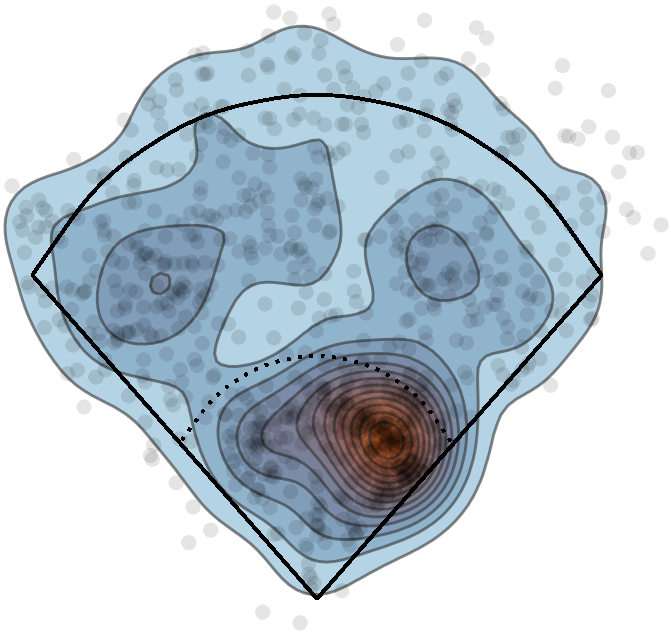


Figure 12. Spray Chart for Infield Shift

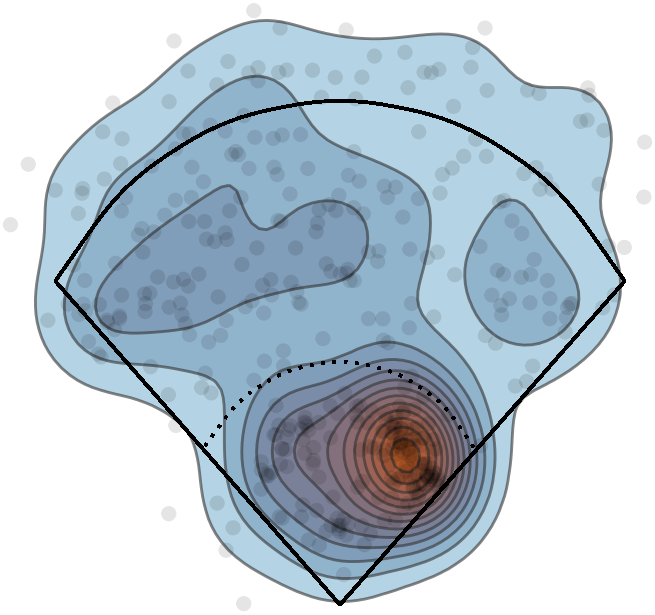


Figure 13. Spray chart for standard defense

Question 2

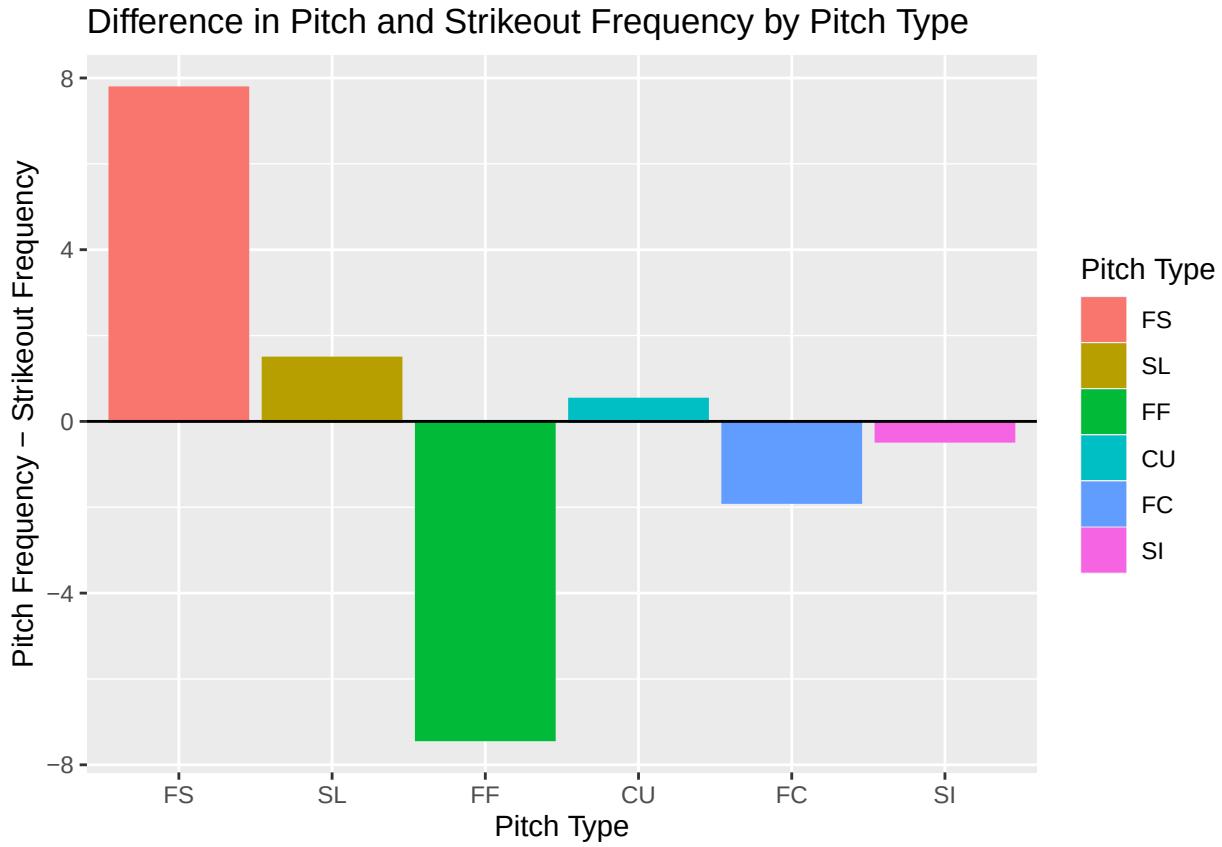


Figure 14. Difference in Pitch and Strikeout Frequency by Pitch Type

We compared the differences in percent distributions for each pitch type in the 2-Strike Pitches and 2-Strike Count Strikeouts observed in the two datasets. We then found how often Ohtani uses his three top pitch types in combination. From here, we observed that Ohtani uses three pitch types over 90% of the time: the Splitter, the Slider, and the 4-Seamer. We, therefore, filtered the data to look only at statistics from data of these 3 pitch types.

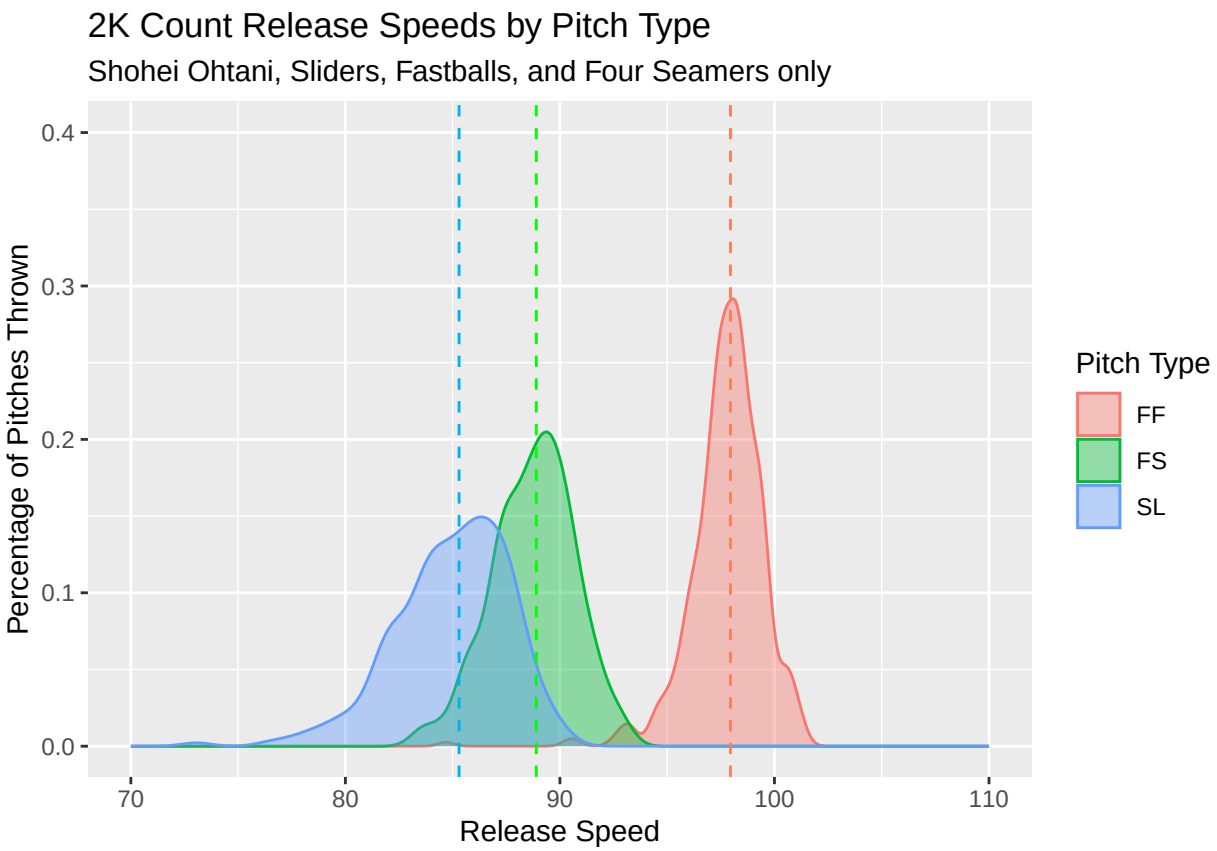


Figure 15. 2K Count Release Speeds by Pitch Type

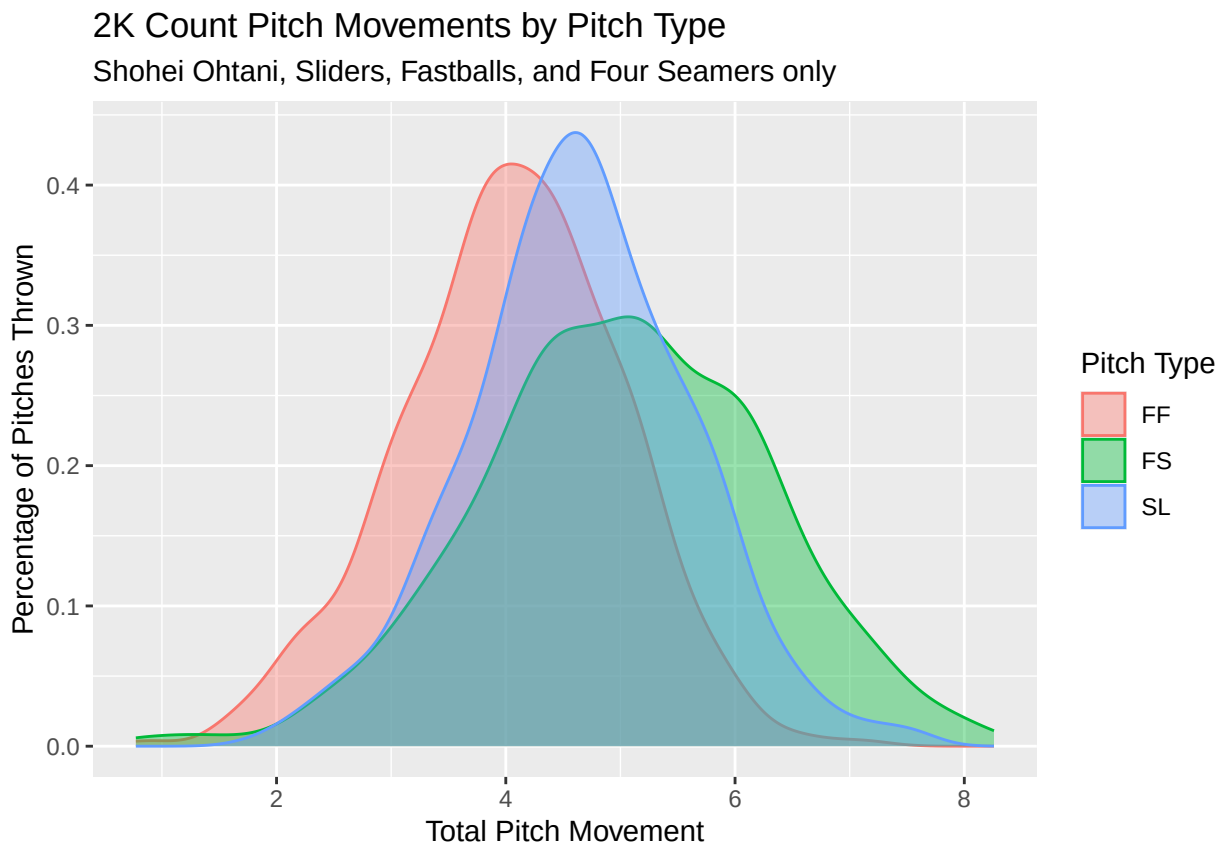


Figure 16. 2K Count Pitch Movements by Pitch Type

We found that each pitch differed in speed and movement, and since each of the three pitch types has a substantial dataset size we were able to compute regressions to find which factors lead to strikeouts. To complete the stepwise regressions, we first created a strikeout dummy variable and removed non-factor/non-numeric columns to simplify the 3 datasets for each pitch type we will use in our analysis. From here, we performed a stepwise logistic regression by AIC with 95% significance to see which variables predict the indicator variable best. Our regression output for the Splitter pitch type found that the horizontal position of the ball when it crosses home plate from the catcher's perspective is a significant indicator for a strikeout.

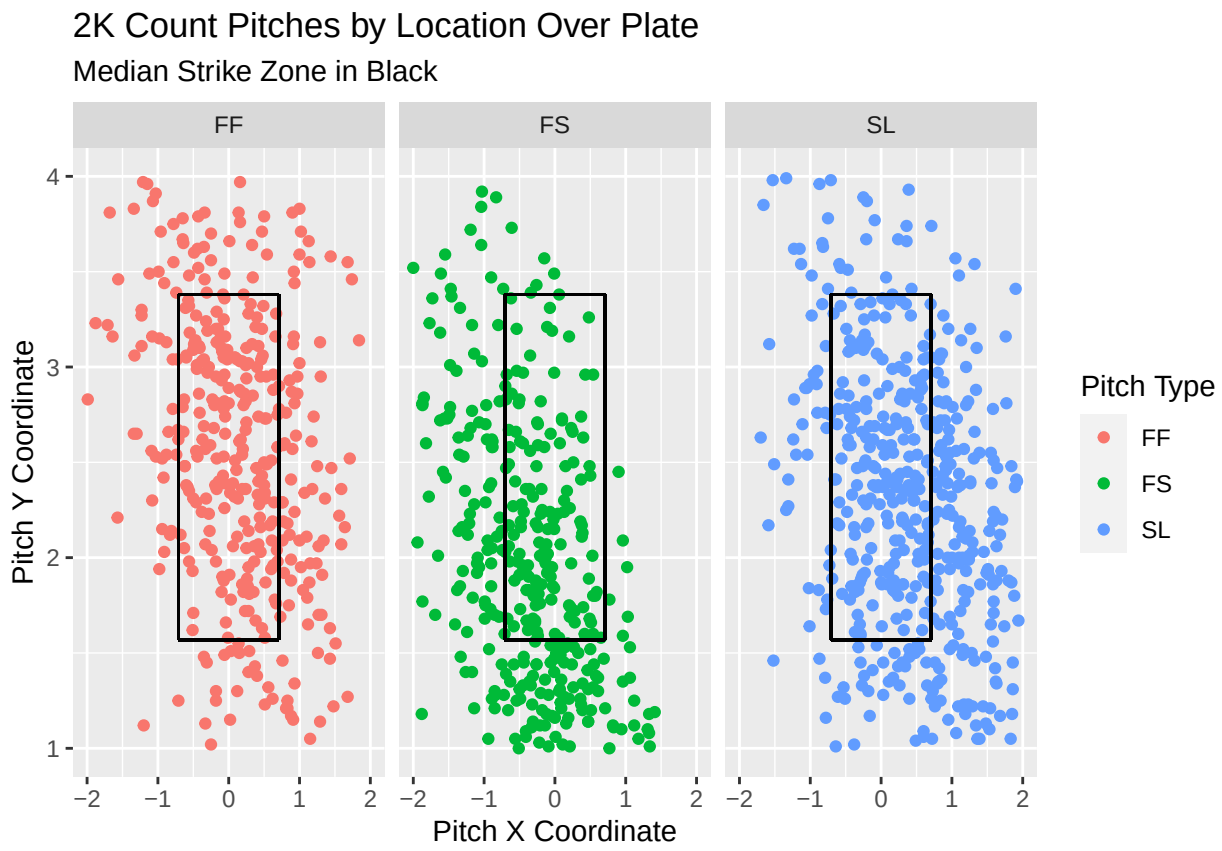


Figure 17. 2K Count pitches by Location Over Plate

The next regression we ran was for the slider dataset. This regression found multiple significant variables that could predict strikeouts. At 95% significance, the horizontal release position of the ball (ft) from the catcher's perspective correlated as an indicator for a strikeout. At 99% significance, the horizontal movement (ft) from the catcher's perspective and the derived speed based on the extension of the pitcher's release correlated as an indicator for a strikeout.

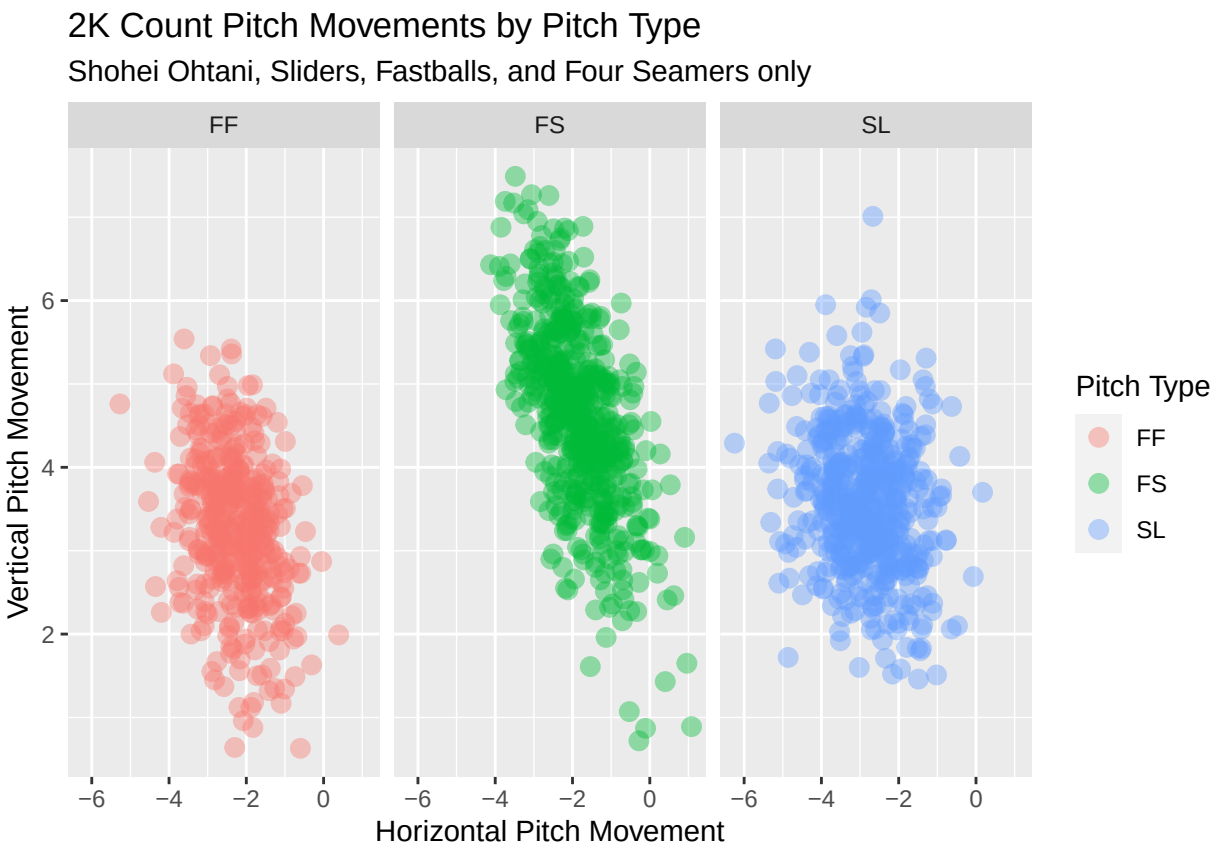


Figure 18. 2K Count Pitch Movements by Pitch Type

The final regression we ran was for the 4-seamer pitch dataset. This regression also found multiple significant variables that could predict strikeouts. At 95% significance, the pitch variables calculating the velocities of out-of-hand release point, vertical movement in feet from the catcher's perspective, and derived speed based on the extension of the pitcher's release are indicators for a strikeout. These outputs show that Ohtani's odds of pitching a strikeout can be improved by certain factors, which differ between pitch types. This helps answer our question about how to predict Ohtani's odds of pitching a strikeout because it develops logistic regressions based on pitch types, giving log odds. From these, we can derive the probability of throwing a strikeout on any pitch, given certain reduced parameters. These parameters are what we sought to uncover originally, and give a sense of what makes Ohtani a great pitcher, as opposed to a decent one.

Conclusion

In this analysis, we set out to answer two questions involving one of professional baseball's biggest stars, Shohei Ohtani. Our first questions involved the infield shift, where we questioned whether new rule changes

that restrict shifts will impact Ohtani's performance at the plate and on the mound. Our second question dealt more with Ohtani's success in striking people out and understanding how he can generate so many strikeouts. After examining our first question, we found that there was a difference between Ohtani's batting average against the shift versus against normal defense, but this difference was not significant. This was found through several methods, including decision trees, random forests, generalized additive models, and proportion tests. In examining our second question, our takeaways are that while Ohtani may switch up his putaway pitches, they tend to be predicted with relative ease. There are only a handful of variables that go into predicting the Ohtani strikeout, depending on the pitch type. This was uncovered via stepwise linear regression using AIC and ended up producing a minute amount of statistically significant variables to consider.

A potential limitation of our method was that our test and training sets were not randomized and could be subject to other effects such as year-to-year changes and ballpark effects. To avoid these issues, we could expand the research to include more players to have a larger sample size. It would also be important to control for weather conditions and ballpark effects, which affect the outcome of a ball in play. Models with more predictive power, such as boosting or support vector machines, could be better models to use, as they should provide more accurate predictions. Our analysis of this question could've been improved with better data on fielder positioning before the pitch and accurate data on ball flight trajectory to properly predict the number of hits that were either lost or gained by the infield shift.

Potential limitations when dealing with our second question, as to the efficacy of different statistics in predicting Ohtani's strikeout rates, would involve overfitting to our dataset, which includes a lot of early-career and covid year data. This could potentially skew the results externally. Internal skewing could be caused by the use of AIC as the lone criterion for logistic stepwise regression. This could be supplemented by BIC or p-value selection, where we would only retain variables if they had a low enough p-value to be considered significant. This is especially relevant to our concerns on predicting strikeout rates as certain variables returned by the AIC stepwise regression were not statistically significant, so they could be eliminated in other methods. Further limitations are reached in situational baseball, wherein Ohtani may have ulterior motives, not aiming to produce a typical splitter, 4-seamer, or other pitch. This could lead to skewing in our dataset. Our analysis could be improved by emphasizing outlier removal since this was not a significant issue with our current datasets.

All in all, our analysis brought to light several implications about Shohei Ohtani and the game of baseball today. The first question will be particularly relevant for teams to answer accurately because of recent rule changes that force teams to field four infielders in the dirt, two on each side of second base. Teams will want

to predict how these rule changes will impact players' performance when they sign them to contracts and build their roster for the season. There may be players that are undervalued because the shift prevented them from getting hits more than other players, so their performance next season should be elevated. According to the results from our analysis, teams should not anticipate star players such as Shohei Ohtani to be severely impacted by these rule changes. The second question is important to consider if teams are facing Shohei Ohtani and need a scouting report on his pitching tendencies. It also highlights important pitch characteristics, like velocity and horizontal movement, which are important qualities that lead to strikeouts. Pitchers who want to improve their performance should work on adding these elements to their pitches to generate swings and misses as Ohtani can. The questions posed in this investigation could be expanded upon on a broader scale to find interesting developments in baseball.