Scott Matsubara
11/16/2023

Atlanta Braves Questionnaire

1. On 8/24/2021, the Cardinals trailed the Tigers 4-3 going into the top of the 9th. To begin this inning, Daz Cameron doubled, Akil Baddoo struck out, and Jonathan Schoop grounded out, moving Cameron to 3rd. The batter is now Robbie Grossman. Assume Luis Garcia will pitch through the 5th spot in the batting order, Jeimer Candelario. Should the Cardinals intentionally walk Grossman? Describe what your process would be to determine whether to pitch to him. The following link contains the box score information for this game: https://www.mlb.com/gameday/tigers-vs-cardinals/2021/08/24/632781#game_state=final,lock_state=final,game_tab=box,game=632781

    - The first factor I would look at is the situation. With a base open at first base and a runner in scoring position, any hesitation to pitch to a batter makes the decision to walk him even easier. After this, the question boils down to which player performs the worst in situations closest to this. First, I would look at historical matchups between Garcia and Grossman and Garcia and Cabrera. If there are enough at bats and a significant difference in batting average or slugging percentage between the two hitters, then I would choose one over the other based strictly on this. If the battings statistics aren't enough evidence, I would look how many times either batter has faced Garcia. Hitters will say that familiarity and being comfortable facing a pitcher is very important to their success against that pitcher. So if, for example, both players have batted .250 against Garcia but Grossman has 10 at bats while Cabrera has 4, I would choose to walk Grossman and face Cabrera. The next factor I would look at is splits against right handed pitchers. Usually same handedness is a bigger decision process, but I would rather look at the actual data instead of simply choosing the right handed hitter to face the right handed pitcher. In 2021, Grossman batted significantly worse against right handed pitchers, while Cabrera had a slightly higher batting average against lefties with more power against righties. Finally, if we cannot make a concrete decision from these numbers we look at a player's clutch ability. There is still not a solid metric to measure "clutch", but some statistics to look at would be WRC+ or BA with RISP and WRC+ in high leverage scenarios. This situation would classify as a very high leverage scenario since it is a one run game in the 9th inning with a runner in scoring position. At this moment, we do not want to face a player who has that xDAWG in him, so we choose to face the player with lower numbers in this scenario. Given all this data, we could make an educated decision on whether or not to walk Grossman. If after doing this analysis, I decide that walking Grossman is the right decision, I would then compare Grossman to essentially all players on the team's bench since Cabrera can easily be pinch hit for. Personally, I agree with the Cardinals decision to walk Grossman and if I were the Tigers, I would have left Miggy in to bat because who wouldn't want to see more Miggy home runs.

2. You are running a generic mid-market team and are exploring the idea of signing Cody Bellinger this offseason. What contract would you be willing to offer him? Please explain your thought process and discuss any important considerations.

    - In 2023, Cody Bellinger proved all his doubters wrong by having one of the best seasons of his career (I would like some credit for this because I met him and took a picture with him at the Chicago airport before the season started). In a season where he was slightly hampered by injury, he hit for power and average and continued to play gold glove caliber defense in the outfield. In order to figure out what contract to offer him, I would look at Carlos Correa who signed a 6 year $200M contract after the 2022 season. After an exceptional offensive and defensive season, he chose to become a free agent at the same age Cody Bellinger currently is. In their seasons before becoming a free agent their offensive numbers were very similar and they both play a premium offensive position. I believe that long term contracts are always a huge risk and this contract length and AAV are what my team should aim for when signing him. I might be willing to increase his AAV as well with his injury risk not being intimidating as Correa who had failed physicals. However, there are signs of some regression in his advanced statistics given that his BABIP was the highest it has been in his career at .319 and his hard hit % was the lowest in his career at 31.4%. This indicates a bit of luck factored into his numbers. I believe that these two factors should cancel each other out. Finally, considering all these factors and the fact that Bellinger has proven to have the work ethic to improve after disappointing seasons, I would be willing to offer him a contract somewhere in the range of 6 years $200M - $220M.

3. Pitcher A walks half the batters he faces and strikes out the other half. Pitcher B doesn't walk or strike out any of the batters he faces. Which pitcher would you prefer? What ratio of strikeouts to walks would make you indifferent between the two pitchers?

    - If the only two outcomes that occur when a pitcher pitches are a strikeout or walk, the only way a run will score is if they walk in a run. This occurs if 4 walks occur before 3 strikeouts happen. Let's do a simulation on one million innings and see the average number of runs allowed by this pitcher. According to this simulation, this pitcher would allow 0.4 runs per inning which equates to a 3.60 ERA.

    - Lets run a simulation to see how many runs a pitcher who only allows balls hit in play will allow. A quick google search tells us that the average BABIP is around .300. We will assume around ¼ of players hits are doubles and we will exclude triples in this analysis. So when a player get a a hit, which happens 30% of the time, there is a 25 % chance that this hit is a double. We also see from this article (https://theconversation.com/mlb-home-run-counts-are-rising-and-global-warming-is-playing-a-role-203226) that the average number of home runs hit per game in 2023 was 1.19, there are on average 33 at bats per game so this means there is a 0.036 chance of a home run every at-bat.. Using the simulation we see in the code, we can see that pitcher B allows nearly the same amount of runs per inning at 0.41. Since we are not including triples in this simulation, this expected number of runs should be slightly higher than this

- For these reasons, I would prefer pitcher A simply because he will allow less runs on average. Other factors to consider would be that pitcher A will throw many more pitches than pitcher B and will not pitch as deep into the game as pitcher B. In order for me to say that Pitcher A is definitely better, his strikeout rate would have to be at least 55% because this would drop his expected runs allowed to 0.28 runs per inning.

4. Briefly explain how you would go about estimating the effect of catcher framing at the major league level? Assume you only have access to the identities of the people involved, information about the pitch (location, characteristics, etc.), and information about the game (count, inning, score, etc.).

    - At an overall level, I would try to quantify how much a ball made a strike or strike made a ball impacts the outcome of the game. For example, a 3-2 pitch that is supposed to be a ball being called a strike would have much more weight than a 0-0 strike being called a ball. A value like 4 can be used to quantify the first outcome, 1 can be used to quantify the second, and 0 can be used for a correctly called pitch. The value would be negative if it is a strike that is called a ball. This would be at a team level where it is positive if it helps the home team and negative if it helps the away team. There would also be values between 1 to 4 and -1 to -4 based on the situation. I would call this variable something like “Framed Runs Created”. I would then find the average value per game to see how much framed runs impact a game. The problem here is separating catcher framing skill from poor umpire judgment. One way to address this would be to have the statistic adjusted based on a given umpire's overall accuracy. We could then do all sorts of analysis on this based on pitch type, pitcher, catcher, etc. For example, we could find that sliders are the best framed pitch and impact the game the most when framed correctly. We could then focus training on helping catchers frame sliders specifically.

    - At an individual catcher level, we would calculate how many pitches near the zone that could be called a strike are actually called a strike. MLB calls this zone the shadow zone and it is the area within one baseball inside the strike zone and one baseball outside the strike zone. So out of all pitches thrown in this area, how many were called strikes. This would give us an idea of how good individual catchers are. It also explains why players like Austin Hedges (World Series champion) still get jobs in the MLB.

Modeling Questions:

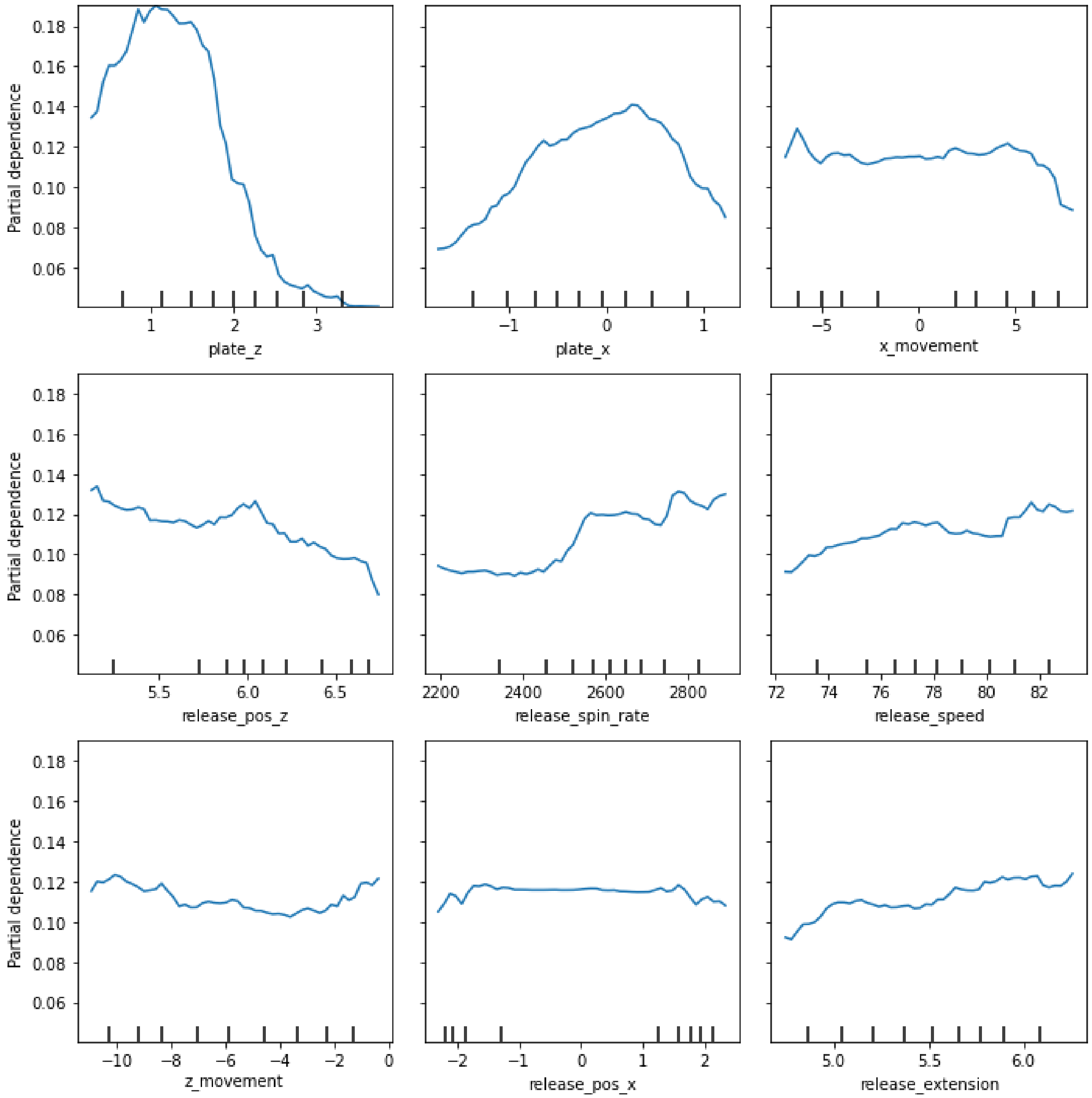
1. The two models I chose to analyze for this binary classification task are a Random Forest model and a Logistic Regression model. Both models are tasked with determining whether a curveball will induce a swing and a miss (1) or not (0).

   The first step in this process is to clean the data by filling NA values and removing outliers. Then I did some feature engineering such as: creating a variable that is the type of pitch thrown before the current pitch (this value is “first pitch” if its the first pitch to a batter) and count category (Full, Two Strike, Behind, and Neutral). None of these feature engineered categorical variables ended up being strong predictors to whether a curveball causes a swing and miss.

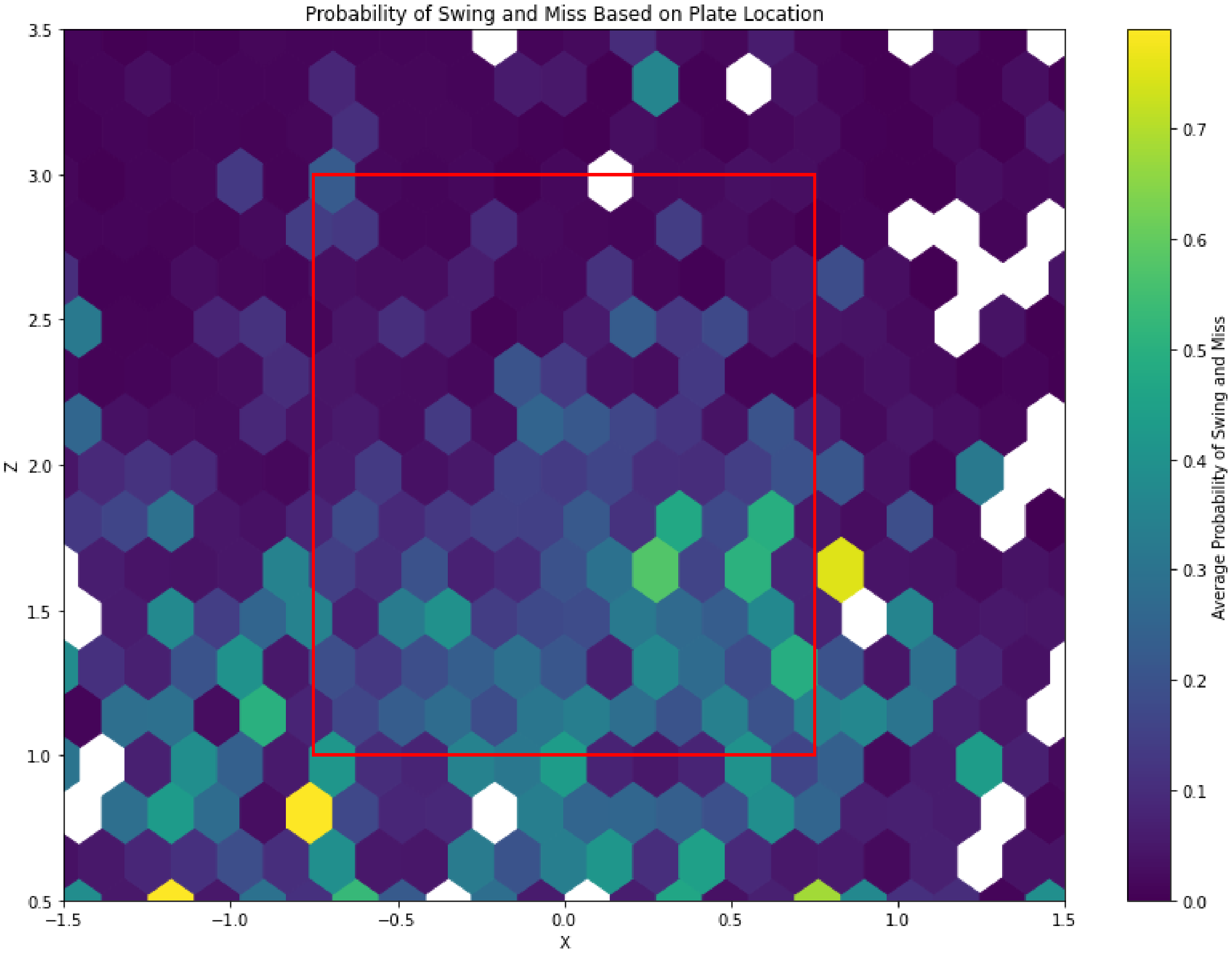
   I then subsetted the data to only include curveballs and created the 1 or 0 flag to indicate swinging strike. This dataset is imbalanced, so we have to take this into account when splitting the data into train and test by stratifying the data based on the response variable. This means that we make sure the same ratio of 0 and 1 are in both the training and testing data. Additionally, I decided to weigh the observations that result in a 1 more heavily when modeling. The last step before modeling was to One-Hot encode the categorical variables.

   For the decision tree, I decided not to standardize the data since ensemble methods like decision trees are not sensitive to the scale of features. I then used 5-fold cross validation to get the feature importance for all the variables in the model. Then for the logistic regression model I used the 5-fold cross validation again with the features that were found to be important by the random forest. For the purpose of this exercise, we want to know what features cause swings and misses in curveballs. For that reason, we want to prioritize precision. If our model classifies a pitch as a swing and miss pitch, we want to be sure that this pitch did result in a swing and miss. Then we can analyze the values of important features of this pitch to accomplish this task. Because of this, we prefer the Random Forest model for our final model which has a precision of 1.0 compared to 0.18 on the test data. The Random Forest model only predicted 1 pitches as swing and miss pitches, but given the imbalanced data and small amount of data, that is okay. Plus, we could look at using a different threshold for classifying the pitch as a swing and miss.

2. Visualizations



- The above visualization is a grid of the Partial Dependency Plots for each of the important features. Each plot shows the relationship of each predictor to the probability of a swing and miss if all other features are held constant. The probability values on the y-axis only go up to 0.18 because this is an average value based on each predictor value. Given the dataset is unbalanced, these values will naturally be closer to 0. After using the random forest to get feature importance, we found that the most important predictor is the height of the pitch (plate_z) and the second most important is the horizontal location of the pitch (plate_x). The top 2 most important variables are essentially the location of the pitch. We see for both of these variables, the line is a U-shape indicating an optimal location for the curveball to be thrown. For height, it is anywhere from 1 to 1.75 feet off the ground and for horizontal location it is around 4 inches outside of the middle of the plate. We can also see some interesting relationships between the predictors and the response. One that we would have expected is the higher spin rate, the higher the likelihood of a swing and miss. I would have thought that lower velocity of a curveball would equal more success but maybe that is only the case for changeups. A low height for release position means more deception, which makes a curveball harder to hit and this is shown in the above plot. For z_movement we can see that either a lot of vertical movement or no movement can maximize swing and miss potential which is also a bit peculiar.



- Since the location of the pitch is the biggest factor in curveball swing and miss, I wanted to show a visualization of the exact zones the random forest model predicts that swing and misses will occur. This strike zone may not be completely accurate, but we can get an idea of where a curveball should be thrown. Pitches in the lower part of the strike zone and outside (since the horizontal location is normalized by handedness) will get the highest percent chance of being missed after swinging.