Cameron Poppen

The pitch types are divided into three categories: cutter, sinker, 4-seam, and other in the fastball group. Since I thought these pitches were more akin to fastballs, like a 2-seam or Skenes' "splinker," I added "other" into the fastball category. Curveball, slider, knuckle curve, sweeper, slurve, slow curve, and screwball were the next breaking ball pitches. And for the splitter, knuckleball, forkball, ephus, changeup, and off speed pitches. Since pitch outs are merely waste pitches and are not required to determine what the majority of pitches these players will see, I chose to leave them out of all the calculations.

In order to help switch hitters distinguish between the pitch mixes on both sides and gain a better idea of what they might see based on the side of the plate, I then divided the players by their batter ID and bat side. After that, I performed a multinomial linear regression model, accounting for the inning, balls, strikes, batting side, throwing side, and batter. Since we can better comprehend the data trend over the last three years and attempt to forecast it for the next, I think this provides a better insight of the players' habits.

I also decided to round the proportions to the nearest ten thousandths so there isn't an overwhelming amount of decimals, and helps it be reasonably readable. And from this, I was able to output it to the file called "predictions_filled."

For several reasons, I think this concept is perfectly logical. Some may have chosen to just average the ratios over the previous three years and predict that this is what will occur the next season, but clubs and players don't behave that way. In addition to changing from season to season, teams and tactics also alter during the season. Since multinomial linear regression may account for a wide range of variables and aid in the prediction of an excellent model, I believe it to be a reasonable model. Although I am aware that there are alternative models that are capable of doing better, I believe this model is the most appropriate for the task at hand.

I am also aware it takes some time for the model to calculate, but that is because of the amount of variables it is taking into account, which, if anything, helps provide a more accurate model in my opinion.

Overall, this model and code help to provide an accurate prediction of the data. Though the data may be very accurate, it is hard to give more advice and information to others since there is still so much more that can be done than just what type of pitches they might see. But, what this model does do is give us a general idea of the relative proportion of pitches these batters might see.

Examples of 1, 3, and 9 player comparisons: