Malcolm Newell

Technical Report

October 14, 2024

Take Home Assessment Technical Report

Hello Director of Analytics,

This report will highlight my process of building a model to predict the pitch mixes all batters faced in 2024 across fastballs, breaking balls, and off-speed pitches. I will highlight the data, my model, and the limitations I faced.

Data:

I made two adjustments to the data I was given. I sorted the pitch types into three pitch groups, fastballs, breaking balls, and off-speed pitches. I excluded the Ephus, Knuckleball, and Other. The second adjustment I made was to find the usage percentage for each pitch group based on pitcher handedness. This was essential for predicting results in the model.

As I explored the dataset that I was given, I wanted to make sure that I chose a predictor variable that allowed me to include the most data possible to improve my prediction results. I was reluctant to use data that was just based on balls put in play or at the end of a plate appearance such as wOBA value or estimated wOBA because of how much it would limit my sample size.

I chose DELTA_RUN_EXP because each pitch has a value. Although it cannot perfectly predict good swing decisions or bat to ball skills, events such as taking a pitch for a strike or ball, swinging and missing, and many others should be included and evaluated when predicting pitch mixes.

Model:

```
model <- gam(usage_percent ~
               s(DELTA_RUN_EXP, BALLS) +
               s(DELTA_RUN_EXP, STRIKES) +
               as.factor(BAT_SIDE) + as.factor(THROW_SIDE),
                    data = train_data, method = "REML")
```

Above, I included the structure of the model that I created. I made three separate models for each of the different pitch groups. The change in run expectancy can be largely based on the count, so I created an interaction term with balls and strikes with DELTA_RUN_EXP to account

for those changes. I also used batter side and pitcher side to incorporate the platoon effects of each pitcher and batter matchup.

Limitations:

The largest limitation that I faced was not having pitch level data to capture the effects that pitch movement has on the run value. Although pitches are tagged the same, the movement profiles could be completely different, and that will impact a batter's performance. For instance, left-handed batters tend to perform better against a right-handed pitcher's slider if it has more horizontal movement and worse against a gyro slider profile. Both pitches from the right-handed pitcher would be included in the breaking ball pitch group.

Another limitation I found with my model was that there was not a large variation in predicted fastball usage percentage.

In the main predictions CSV, I also did not include the pitcher throwing-side usage split, but it is talked about in the presentation to the coaching staff.