Predicting a Batter's Pitch Mix

Donald Stricklin

2024-10-21

Introduction

The goal of this project is to predict the proportion of pitches a Batter will see during the 2024 season. Understanding what pitches a batter will see during the course of the season can help encourage better and more focused training and hopefully produce better swing decisions at the plate. To accomplish this, we used pitch-by-pitch data from the 2021-2023 MLB regular seasons.

The Model

When working with a question and dataset as the ones we have, it's important to know that their exist many different relationships and correlations between the variables, while also understanding that our problem is not linear. That is why to create our predictions, I used a regression XGBoost model.

In order to predict the proportion of pitches a batter sees, I decided that I first needed to predict the number of pitches of one of the pitch groups the hitter would see. To do so, I decided to group and summarize the data in two ways:

- By **Game Year**, to show the batter is as a whole from year to year
- By **Game Year** and **Pitch Group**, to show how the batter fares to each pitch group from year to year

Across both groups, I created batter specific metrics including:

- Batter Info: Top and Bottom of the Strike Zone
- Standard Metrics: Hits (1Bs, 2Bs, 3Bs, HRs), Walks, Strikeouts, Total Bases
- Advanced Metrics: Swing%, Z-Swing%, Chase%, Whiff%, PutAway%, wOBA, xwOBA, Exit Velo, Exit Angle, SweetSpot%, HardHit%, Barrel%, WPA

Training and Tuning

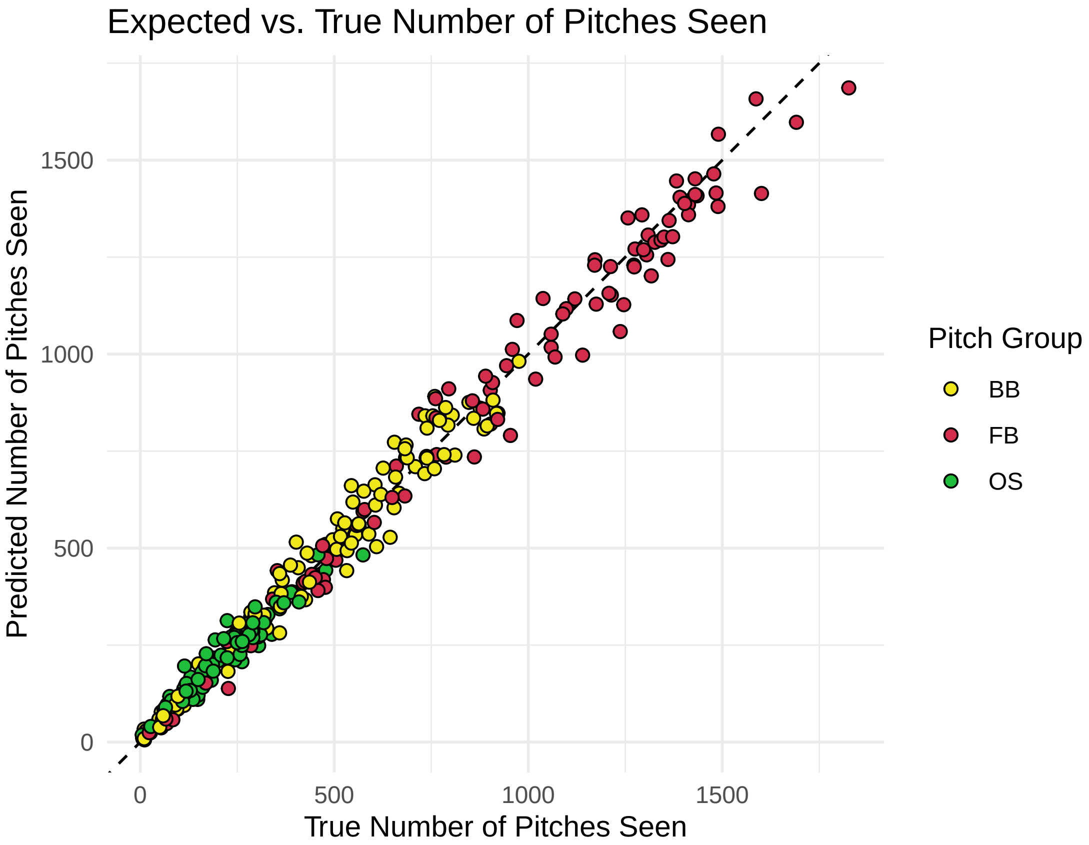
The model was trained on the 2021 and 2022 seasons data, where as the 2023 data was saved to predict the final 2024 proportions. I implemented K-fold CV to help the model learn the patterns in the data set and to see even more unseen data given the lack observations available. Furthermore, the XGBoost model was tuned on `min_n`, `mtry`, and `trees` in order to achieve the best possible predictive power. The final parameters are: `min_n` = 12, `mtry` = 44, `trees` = 100.

Metrics

After predicting on our test set, our model has a final RSQ of 0.985, RMSE of 52.3, and MAE of 38.7

Visual

We can see how the model predicted the number of pitches vs. the true number of pitches below



Limitations

Based on how I approached this problem, there will be gaps in our data and in our predictions. We are predicting year-to-year trends, and while the results are promising, the reality is that a players performance, along with the pitches he sees, changes throughout the year. There are times when a hitter may be swinging at a higher rate than usual, which in turn may influence the pitcher to throw more pitches out of the zone, which will likely be non-fastball pitches.

Additionally, the variables I used don't take into account what is going on in the game at that moment. There could be a scenario where a pitcher is unable to locate his breaking ball pitches that day, and thus means that fewer pitch types seen.

Furthermore, players spend an entire offseason and spring training to make better swing decisions and improve as a hitter. So, a player who may have had a high whiff and chase rate on offspeads or breaking balls the year prior, may have been able to cut those rates in half. The model, unfortunately, is unable to take those changes into effect unless the way the model predicts its values were to change as well.

Lastly, the model doesn't take into account the team and/or division the batter plays for, or where in the lineup he bats. The 3-5 hitters are typically going to see a higher number of non-fastballs than other hitters given their stereotypical power abilities. It also doesn't have any bat speed or swing length metrics which could also impact the proportion of pitches seen.