Cincinnati Reds Hackathon

Liam Jennings, Adam Gehr, Josh Knapp, and Brogan Berkey

Robert Morris University

February 5, 2024

Introduction

As the years progress, the average number of innings pitched for a starting pitcher has declined, promoting a more fluid use of pitching depth. Our goal in this project is to determine key factors that classify a pitcher as either a starter or a reliever and find starters who may benefit from converting to a reliever, and vice versa. The code includes our step-by-step analytical approach, using different methods to come up with a sound conclusion. We first filtered the data to include useful cases and modeled to determine the most important variables. We then cleaned the data to ensure the names were consistent across both datasets. Once we cleaned the data, we defined what variables to use, used linear regression to weigh our variables, and used a decision tree to model the data. Once the decision tree was complete, we used k-fold cross-validation to check our final model.

Methodology

A key aspect of our analysis was determining which variables were necessary for classifying pitchers as either starters or relievers. We filtered our dataset by looking at summarized basic statistics such as ERA. To avoid low-workload pitchers, we cut the data at 19.1 innings, the average number of innings pitched per pitcher in the dataset. We ran a cluster analysis to group the pitchers based on the variables we selected. The batting average on balls in play (BABIP) finds batting averages excluding strikeouts and home runs, which isolates the pitcher by removing defensive influence. The next variable we chose was strikeout percentage (K%), which measures a pitcher's ability to strike batters out; this is important for relievers who may be entering a game in a high-leverage scenario. We also used walk percentage (BB%) because it is vital for pitchers to keep batters off the bases. Another variable that we deemed important was the Left On Base percentage (LOB%), which measures a pitcher’s ability to work

out of difficult situations with runners on base; this is useful for starters' pitch counts as well as high-leverage relievers. The rest of the variables we chose were batted-ball stats, which measure a pitcher's ability to generate ground balls and fly balls rather than home runs, as well as preventing hard contact and line drives. After multiple methods of analysis, we determined that batting averages against and WHIP were not great predictors of FIP because of their dependence on defense. As a result, we altered our approach to look for a statistic that isolated pitcher performance. We ended up deciding that Weighted On Base Average (wOBA) was a good replacement for these two variables. This offers a better understanding of an offensive player's added benefit in each plate appearance, which we used to show how successful pitchers are at getting outs. The Statcast dataset included wOBA per pitch, so we averaged each pitch for the individual pitchers and added that to our dataset. The issue that we had here was that the names did not match up across the Fangraphs dataset and the Statcast dataset; the Fangraphs dataset did not include any accents in player names, but the Statcast dataset did. This caused us to recode all the names with accents in the Fangraphs data so that we were able to bind our wOBA statistic from Statcast to Fangraphs. The data cleaning process was tedious, but worthwhile, as we feel that wOBA vastly improved our final analysis. Our next step was to use a linear regression model with a target variable of Fielding Independent Pitching (FIP), to see which variables were the most impactful based on what we had previously identified. We narrowed down the dataset to 6 meaningful variables.

Before utilizing our decision tree, we created linear regression models using stepwise regression. The base model, which is FIP explained by the twelve predictor variables, was created to be a basis for the stepwise regressions. All of the variables have approximately normal distributions. The stepwise regression models used AIC and BIC, which penalizes having more

predictors in the model. The first model, forward stepwise regression using AIC, had an F-statistic of 2854, and all variables are significant at alpha 0.01 except for GB/FB. After running a hypothesis test to determine the significance of GB/FB at alpha 0.01, we concluded that GB/FB was not significant and removed it. The F statistic increased to 3252 (Figure I). The forward and backward stepwise regressions using BIC resulted in the same model as the first one. The final variables that were deemed significant as K%, BB%, BABIP, LOB%, LA, Barrel%, and WOBA.

Analysis

The objective of the decision tree (Figure II) is to classify a pitcher as either a reliever or a starter using the variables deemed significant by the linear regression model with the target variable of FIP. According to the decision tree, relievers represent 63% of the data in the dataset, leaving the remaining 37% represented by starters. The decision tree's first classification step is determining whether or not a pitcher's Barrel% is above or below 6.1%. If the pitcher has a Barrel% less than 6.1%, the model identifies relievers correctly 83% of the time. However, if the pitcher's Barrel% is greater than 6.1%, the decision tree uses the other important variables identified by linear regression, which are BB%, BABIP, K%, wOBA, and LOB%.

The confusion matrix shows that our model improved the accuracy of classifying pitcher types by 4.07% (Figure III). This is proven through the accuracy level of 71.47%, compared to the no-information rate of 67.40%. This means that our model provides enough information to improve the prediction of classifying pitcher type by 4.07% compared to a prediction given no statistical pitcher information. To validate the model, we used k-fold cross-validation, which randomly splits data into *k* groups and then creates *k* models, each tested against one of the groups. It utilizes and randomizes all of the training data to build and assess the final model. The

accuracy of the k-fold CV is approximately 71%, which is nearly identical to the confusion matrix and shows that the model is good (Figure IV).

After validating our model, we used the test data to filter pitchers from 2023 who had incorrectly predicted roles. There were 33 occurrences, with 18 starters and 15 relievers. We calculated the averages of the significant variables to compare to the new data. Using the computed averages and subjective reasoning, we selected three pitchers who could thrive in a role change.

Player Selection and Conclusion

Since his 2022 arrival in St. Louis, the plan was for Steven Matz to be a back-end starter. Since then, the Cardinals experimented with bringing Matz in from the bullpen; 2023 saw a relatively high level of success. In 15 innings out of the bullpen in 2023, Matz had an FIP of just 2.69, with a minuscule average Barrel% of 2.27% and a launch angle of 5.94^0. Although his sample size out of the bullpen is small, his full-season averages are very good when compared to other relievers. His BB%, LOB%, and Barrel% are all better than the average for relievers, although his K% is just barely below average (Figure V). His high BABIP is not a cause for concern, as more consistent innings should help this regress towards the mean (Figure VI). Transitioning into the bullpen could allow Matz to extend his career while providing valuable innings.

Under normal circumstances, Blake Snell would not be considered to make a switch from starter to reliever. However, he has had consistency and durability concerns that have prevented him from pitching deep into games. However, it would not be unprecedented to see a high-level starter such as Snell be converted to the bullpen. This occurred with John Smoltz after he suffered arm injuries, and he was able to find success in the bullpen. If a team transitions Blake

Snell into a reliever role, it could lead to similar results. Snell's numbers also suggest that he could find success as a high-leverage reliever, as he has an above-average strikeout rate (with a high walk rate), a high LOB%, and a low wOBA (Figure VI). If these numbers hold up, he could be a dominant back-end reliever.

Víctor Gonzalez comes into a Yankees organization with many options in the rotation. Gonzalez could be useful considering the injury concerns at the top of the rotation. Gonzalez has proven to be an effective ground-ball pitcher, which will be important in Yankee Stadium. His 2023 saw a very low launch angle of 1.36^0 and an above-average K% at 22.73% (Figure VI). He also has a low BB% and wOBA compared to the averages of the classified starters (Figure V). Víctor Gonzalez's 2023 statistics look to present the Yankees with the potential to utilize him as a contributor in the rotation.

Appendix

Figure I

```
Call:
lm(formula = FIP ~ wOBA + BABIP + LOB_pct + BB_pct + Barrel_pct +
    K_pct + LA, data = linFan)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50688 -0.16370 -0.00095  0.16466  0.96483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.833469   0.123763  -6.734 2.29e-11 ***
wOBA         27.183508   0.374505  72.585  < 2e-16 ***
BABIP       -15.517390   0.278121 -55.794  < 2e-16 ***
LOB_pct       0.864654   0.093575   9.240  < 2e-16 ***
BB_pct        2.568125   0.257168   9.986  < 2e-16 ***
Barrel_pct    2.521551   0.354390   7.115 1.69e-12 ***
K_pct        -1.423241   0.180405  -7.889 5.62e-15 ***
LA            0.004228   0.001318   3.207  0.00137 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.26 on 1587 degrees of freedom
Multiple R-squared:  0.9348,    Adjusted R-squared:  0.9345
F-statistic:  3252 on 7 and 1587 DF,  p-value: < 2.2e-16
```

Figure II

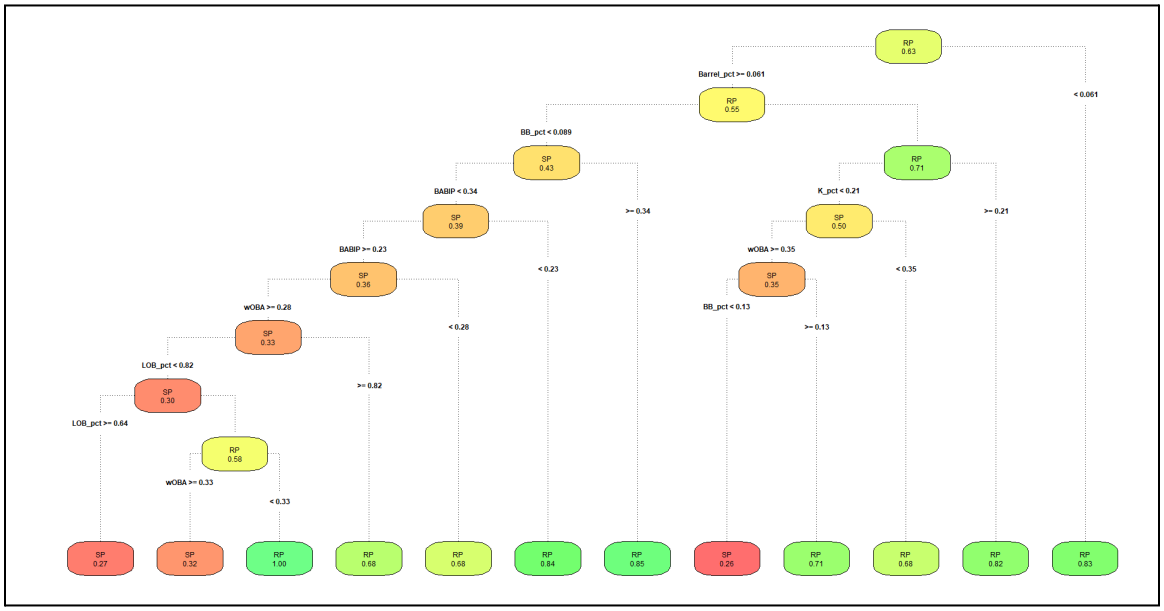

Figure III

```
Confusion Matrix and Statistics

          Reference
Prediction  SP  RP
        SP  63  50
        RP  41 165

               Accuracy : 0.7147
                 95% CI : (0.6618, 0.7637)
    No Information Rate : 0.674
    P-Value [Acc > NIR] : 0.0665

                  Kappa : 0.3651

 Mcnemar's Test P-Value : 0.4017

            Sensitivity : 0.6058
            Specificity : 0.7674
         Pos Pred Value : 0.5575
         Neg Pred Value : 0.8010
             Prevalence : 0.3260
         Detection Rate : 0.1975
   Detection Prevalence : 0.3542
      Balanced Accuracy : 0.6866

       'Positive' Class : SP
```

Figure IV

```
CART

1276 samples
   7 predictor
   2 classes: 'SP', 'RP'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 2 times)
Summary of sample sizes: 1149, 1149, 1148, 1148, 1148, 1148, ...
Resampling results:

  Accuracy   Kappa
  0.7013872  0.3482068

Tuning parameter 'cp' was held constant at a value of 0.001
```

Figure V

```
[1] "Averages From 2021-23"
      K_pct      BB_pct       BABIP     LOB_pct          LA  Barrel_pct        wOBA
 0.23309197  0.08764771  0.28832106  0.72842502 12.72288842  0.07557160  0.31872888

[1] "Starting Pitcher Averages From 2021-23"
      K_pct      BB_pct       BABIP     LOB_pct          LA  Barrel_pct        wOBA
 0.21873379  0.07772693  0.29074426  0.72871407 13.02349293  0.08203636  0.32807455

[1] "Relief Pitcher Averages From 2021-23"
      K_pct      BB_pct       BABIP     LOB_pct          LA  Barrel_pct        wOBA
 0.24136354  0.09336294  0.28692508  0.72825850 12.54971408  0.07184733  0.31334496
```

Figure VI

NAME	K% ↓	BB%	BABIP	LOB%	LAUNCH ANGLE	BARREL%	WOBA
Snell..Blake_2023	0.3154	0.1334	0.2557545	0.8673	10.60075	0.0739	0.2718121
González..Victor_2023	0.2273	0.0682	0.2840909	0.663	1.362011	0.0667	0.279771
Matz..Steven_2023	0.2167	0.0705	0.3218391	0.7528	12.74572	0.0515	0.3336815