Cincinnati Reds Dew Point and Pitching

Daniel Mueller

Introduction

In the initial prompt, it is stated that high humidity can have an affect on both the flight of a pitch, as well as the comfort of the pitcher. In my write up, I decided to make sure I thoroughly explored both of these aspects, using pitch control (ball percentage, or one minus strike percentage) as a measurement to gauge whether or not a pitch was effected by a dew point greater than 65 degrees F. Specifically, the metrics from the data I chose to investigate were Induced Vertical Break, Horizontal Break, Release Speed, Release Side, Release Height, and Release Extension. The approach angle statistics, as well as Plate X and Plate Z were not used due to their strong correlations with ball/strike ratios.

Pre-Processing

I began by filtering out data points where the events were stolen bases, pick-offs, or catcher's interference. These events either don't have pitch data or don't provide ball/strike information. These entries are also a small subset of the overall data and removing them did not significantly impact my results. I then converted the Horizontal Break and Release Side metrics to their respective absolute values, aligning data from pitchers of different handedness to the same domain. To end the pre-processing, I created binary variables "Strike" and "Ball" to provide for easier calculations later on.

Analysis

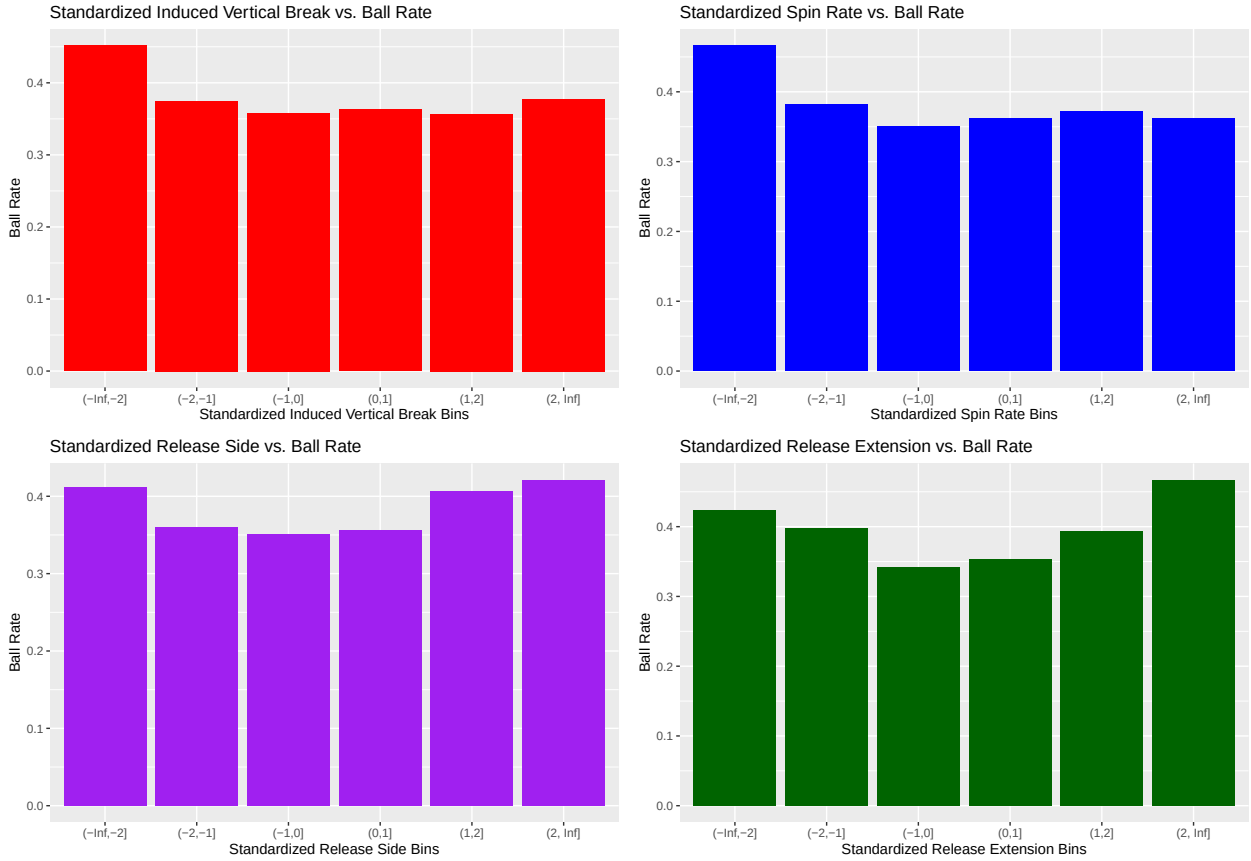
Since no weather data was provided, I was required to use the pitch data itself to determine if humidity had an influence on a pitch. I did this by grouping the data by pitcher and pitch type and creating standardized normal distributions for every pitch in each pitcher's arsenal. From there, I went through each metric and marked if a pitch had a value that was statistically significant with an alpha value of $\alpha = 0.05$. I then placed each pitch into p-value bins of $(-\inf, 0.05]$ and $(0.05, \inf]$ and compared the two bins' ball percentages for each metric. A table comparing these percentages is attached here:

Metric	Ball Rate for P $<$ 0.05 Bin	Ball Rate for P $>$ 0.05 Bin
Induced Vertical Break	0.4194260	0.3616409
Horizontal Break	0.3795455	0.3635880
Spin Rate	0.4187082	0.3616999
Release Speed	0.3705263	0.3639859
Release Side	0.4179487	0.3620872
Release Height	0.3774403	0.3636558
Release Extension	0.4487472	0.3603575

From the table, Induced Vertical Break, Spin Rate, Release Side, and Release Extension all have large differences in ball rate between the two bins, while the other three metrics have smaller differences. This makes sense, as lower-than-usual values of movement, as well as spin rate, could be the result of increased

water vapor in the air. Additionally, discrepancies in release metrics could be a consequence of a pitcher feeling uncomfortable due to a high dew point.

I then visualized the four metrics with the largest differences between bins, plotting bins of standardized values against ball rate:



Afterwards, I created another binary variable named "AFFECTED", which contains a 1 if any of Induced Vertical Break, Spin Rate, Release Side, or Release Extension are statistically significant for a pitch, and 0 otherwise. From there, I split the data into training and testing set and built a logistic regression model predicting the "AFFECTED" variable using the four metrics. Applying the model to the testing set using a probability of 0.2 as the cutoff between being affected or not affected by humidity resulted in a test accuracy of 84.5%. The model was then applied to the entire data set and the output was written to "submission.csv".

Conclusion and Limitations

The results of this investigation imply that it may be possible to determine if a pitch was affected by a high dew point. This is demonstrated through pitch flight metrics such as movement or spin rate, as well as pitcher mechanics such as release side and extension. However, it is important to acknowledge some limitations of the research. This analysis is built on the assumption that humidity is responsible for certain outliers and trends in the data, but it is entirely possible that large deviations in pitch data could be the result of other external factors. Additionally, due to the lack of weather information and additional data sets, this model was developed, trained, and testing on the same set of data. This could raise concern for issues such as overfitting. Thus, while this investigation suggests that humidity may impact pitch performance, it is important to be aware of potential confounding variables and limitations in the data.