Cincinnati Reds Technical Assessment: Dew Point Affected Pitches

Introduction: Determining if pitches are affected by a dew point of 65 degrees fahrenheit is extremely valuable in assisting with player performance and evaluation. If we are able to figure out which pitches are externally affected by humidity, we will be able to then figure out which pitchers specifically are most affected by humidity. From here, we can help with important decisions such as lining up the rotation so pitchers who do worse in humidity do not have to pitch in those conditions. We can also make sure that rookies who are affected by humidity are not called up to pitch in conditions where they are less likely to succeed since that will negatively impact their confidence. Another usage of this research would be to help pitchers who struggle with specific pitches in these conditions make necessary adjustments

Abstract: My hypothesis is that pitches that rely heavily on movement and spin such as curveballs and sliders will be the most affected by humidity. Because of this, I proceeded under the assumption that changes in spin rate and vertical and horizontal movement are the best indicators of if a pitch is affected by humidity. The first approach I tried is called the naive approach in my study. It finds the difference between spin rate and horizontal and vertical movement for each pitch compared to the average value of that specific pitch for a specific pitcher. This intends to capture how different that pitch is to a similar pitch thrown by that pitcher. I then normalized these three differences on a 0-1 scale and found the average. This was the first probability value I found.

Next, I created a binary response variable that is 1 if the spin rate is more than one standard deviation away from the mean for a pitch and player. I then used the rest of the data to predict this value. After using a Random Forest, Logistic Regression, and Linear Mixed Model I found that none of these models performed better than predicting 0 for all pitches. The rest of the data does not seem to capture information about humidity affected pitches and I did not want to use spin rate and movement as predictors since this would be a data leakage problem. Because of this, I chose to use the first probability values found by analyzing spin rate and movement for my submission and analysis.

Methodology:

Data Cleaning: The data was very clean with no NA values or duplicate rows. I did remove some rows for pitches thrown at less than 65 mph in the 9th inning since these are more than likely position players pitching. I also removed rows containing unknown pitch types and knuckleballs. For the rows I removed, I predicted 0 for the probability in the final submission.

Feature Engineering:

- Combined the IS_RUNNER_ON_1B, …2B, and …3B into a single flag that tells if there are any runners on base. We would just like to know if there is anyone on base so that we can know if the pitcher is pitching from the stretch or the wind up
- Categorize the balls and strikes into a count category: Full (3-2), Behind (2-0, 3-0, 3-1), 2-Strike (0-2, 1-2 ,2-2), and Neutral (All other counts)
- Normalize Horizontal Break so that a negative value means armside run and a positive value means glove side cut

- Create a new variable that is the difference between mean spin rate grouped by pitcher and pitch type and the actual spin rate value. Also create a variable for std based on the same groupings. Do this for spin rate, induced vertical break, and induced horizontal break. This will tell us how much different a specific pitch's spin rate and movement is compared to normal
- Define a pitch as DEWPOINT_AFFECTED if it has a spin rate that is more than one standard deviation less or greater than the mean. We include pitches one standard deviation less because less spin rate means the pitcher did not have a good grip on the ball. This loss of grip can be attributed to humidity/the ball being slippery. We include pitches one standard deviation greater than the mean because when the air is humid, the air density is lower and therefore it allows spin rates to be faster according to this article: https://sabr.org/journal/article/how-climate-change-will-affect-baseball/

Evaluation: I created two probability values as mentioned in the abstract. Once there is an actual evaluation dataset, it is important to determine which kind of accuracy measure we would like to use. In the case where we want to avoid having pitchers pitch in humid conditions it is better to be safe than sorry, so we would like as many pitches as possible that are actually affected by humidity to be classified as such. In this case Recall would be more important and we would not care about false positives as much. In this case where we want to help players make adjustments, we would want to keep false positives low so that we only try to help pitchers who actually need it since making too many adjustments to a pitcher may be problematic. In this case, we would prioritize Precision.

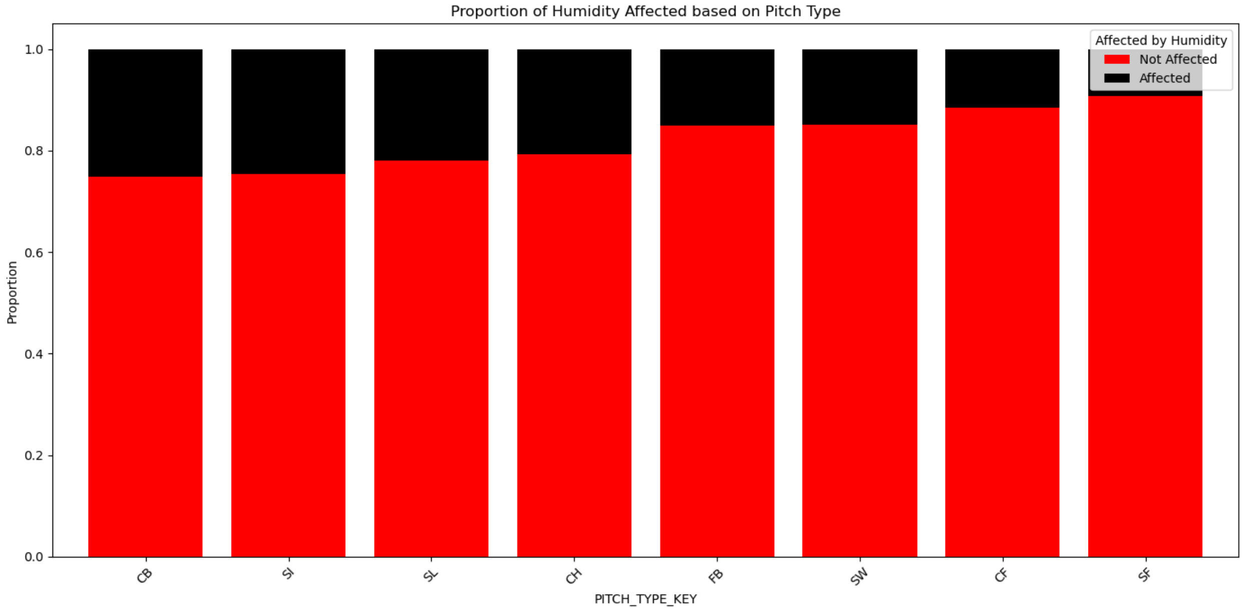
Analysis and Future Work: There is no way to tell how accurate these probabilities are, but looking at the below graph that shows the proportion of pitch types affected by humidity, we can confirm the initial hypothesis that pitches which require heavy spin rate and movement will be affected by humidity the most. This gives us some indication that the predicted probabilities are somewhat accurate.

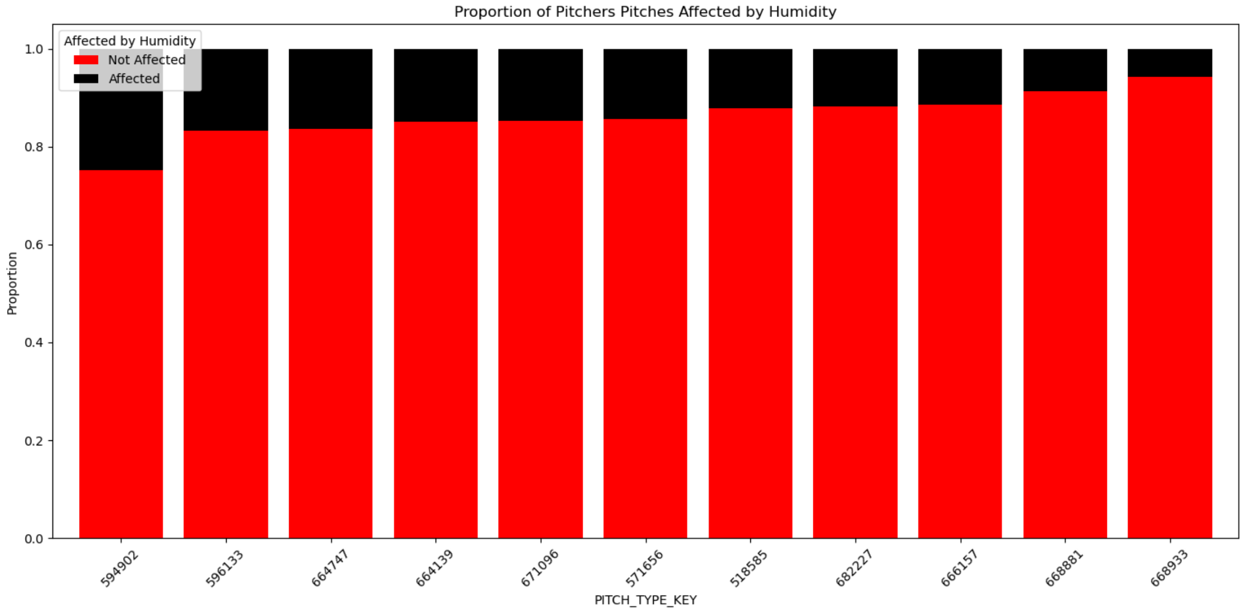
Next, we take a look at the pitchers with the highest proportion of humidity affected pitches. We can see that pitchers 594902, 596133 ,and 664747 have the highest proportion of humidity affected pitches. These would be the pitchers that we would either try to avoid having pitch in humid environments or work with them to improve in these environments.

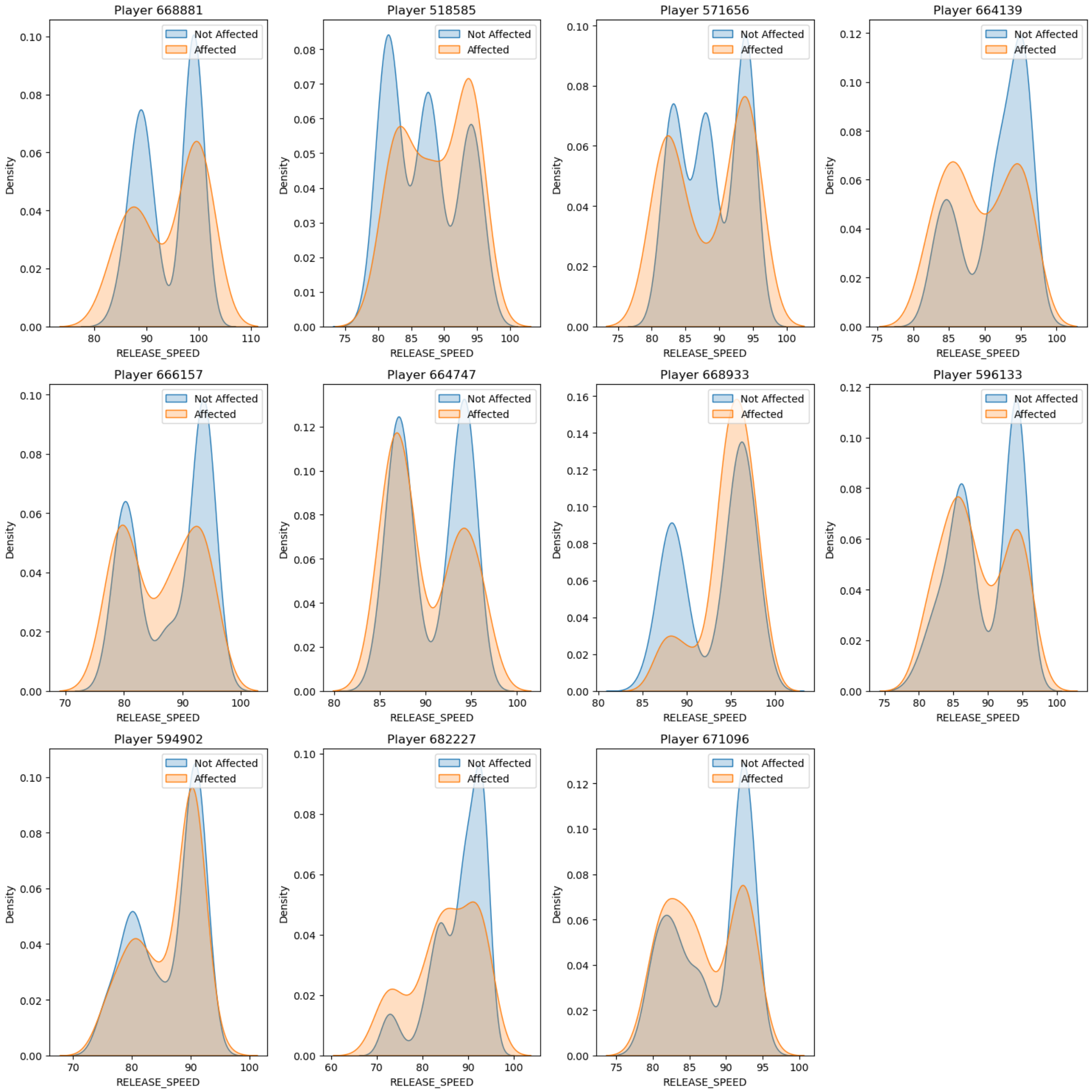
Finally, we can see that some players have similar distributions for dew point affected pitches and non affected pitches based on pitch velocity. We can say that these players are not affected by humidity. The ones that we would like to flag for further investigation are the ones that have different distributions. An example of this would be pitcher 6643 who has a greater proportion of offspeed pitches that are affected by humidity and pitcher 518585 who has a greater proportion of fastballs that are affected by humidity. This would give us specific pitches to work on with specific players to help improve their performance. We could also tell these specific players to stop throwing as many fastballs/offspeed pitches during these starts.

Conclusions: This is a fascinating project and one that has a ton of implications within baseball that can help improve team performance. More data is likely needed to create accurate predictions of if a pitch is affected by probability, but I like the insights I was able to generate. I am grateful I got a chance to work on this and given more data and time, I would love to continue exploring this topic.

Visualizations:


Proportion of Humidity Affected based on Pitch Type


Proportion of Pitchers Pitches Affected by Humidity



Player 668881
Player 518585
Player 571656
Player 664139
Player 666157
Player 664747
Player 668933
Player 596133
Player 594902
Player 682227
Player 671096
Not Affected
Affected
Density
RELEASE_SPEED
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.01
0.03
0.05
0.07
60
70
75
80
85
90
95
100
110