Preliminary Research

After I read through the Dew Point pdf, I started looking through the data to see what interesting patterns I could find or identify. This was a bit of a peculiar project, because there is no weather, date, or time information for each pitch, only the information about the pitch. Obviously, this leaves a bit to be desired. Common knowledge suggests that there are two main ways dew point can affect a pitch: flight path and pitcher fatigue.

Upon digging into dew points and flight paths, I found that indeed, the flight path is affected, if only slightly. Baseball VMI suggests that, because humid air is more moist, it's actually easier for the ball to travel through it. This can express itself as a deviation to the left or the right from the expected flight path, all other things equal. Fatigue is similar. Pitchers do in fact tend to get hotter faster, and as a result they get tired faster. When pitchers get tired, pitch location and speed can become issues and the ball can be sent flying. Compounding the fatigue, the ball tends to fly further as well.

Exploration

This project was a little tough to start for at least one obvious reason: there is no ground truth. And, in baseball, there are likely many contributing factors to any event that it can be hard to see through the noise. Further, there aren't even any weather related labels to start with. So, I started by trying to find some basic facts about what I was looking at. This data spans 65 games, all played in Cincinnati. There are 37 unique pitchers and they throw 10 unique pitches. The average pitcher throws for only 1.07 innings and throws 27.5 pitches, but many throw for longer than that-- starters average 4 innings per game and average 82 pitches.

After that, I turned to less concrete analysis, to try and see if I could find a difference in pitches. Given that there are no ground labels, this was a little harder, but it was immediately promising. I first limited my search space to just one pitcher: 668881. I wanted to get a visual idea of how pitches might cluster, so I visualized it using T-SNE. The result is in figure 1.

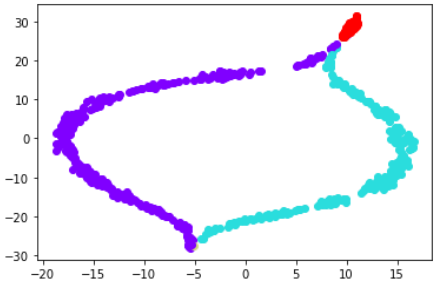

Figure 1: Fastballs, Sliders, and Changeups for 66881

Upon further inspection, 668881 actually throws another pitch UN (unknown?), but that was thrown only 4 times, so I removed it. One thing I was interested in was that the two more popular pitches for this pitcher, fastballs and sliders, had far more variability than changeups. This is most likely due to the fact that they just throw fastballs and sliders so much more than changeups (268, 218, and 29 times, respectively). The variability, however, got me thinking: if I know the starting conditions of the pitch, and know all the information as a ball approaches the plate, then I should be able to detect where it lands on the x axis. In the cases where the model is wrong, something changed the expected flight path: humidity!

Methods and Results

In order to explore this further, I first found the features that were most correlated with PLATE_X. Horizontal Approach Angle (HAA) was very highly correlated, greater than 95% in many cases. After that,

I found that Horizontal Break was also very highly correlated and did not give rise to a multicollinearity issue. I ran my linear regression on a pitcher by pitcher basis, so I created 37 models. Additionally, I also supply the pitch type as a dummy variable so that the model is able to differentiate pitch type. On average, the adjusted R^2 value is 0.94, with an average F-Statistic pvalue of 1.0331e-17. These both suggest that the model is highly able to capture the variability of the pitches and accurately predict PLATE_X. Next, I looked at the distribution of the residuals– however, here I chose to use a folded normal distribution because I am concerned with misses to either side of the predicted value. The output of this for a specific pitcher is shown in figure 2.

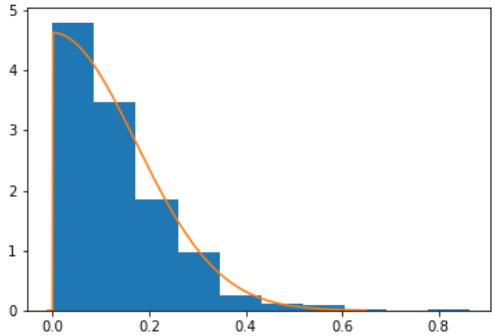

Figure 2: Output of residuals for a pitcher

The idea here is that the only thing affecting a pitch as it approaches is the humidity. Thus, I turn this distance into a probability, by computing the area under the curve for the folded normal distribution that is generated by the residuals. This does have the effect of “flattening” out the output space– have the pitches have probabilities above 0.5, which is probably not ideal, but without weather data, it would be hard to figure out where it’s missing and correct the model.

Conclusion

This was a very interesting and challenging project. If I had more time (and information), I wanted to blend a pitcher's fatigue probability into the model, but it was hard for me to get a sense of “faster” or “slower” pitcher degradation without knowing any label information. I did look to see if the residuals in the model get worse over time, perhaps indicating a degree of tiredness, but the trend seems to be fairly stable:

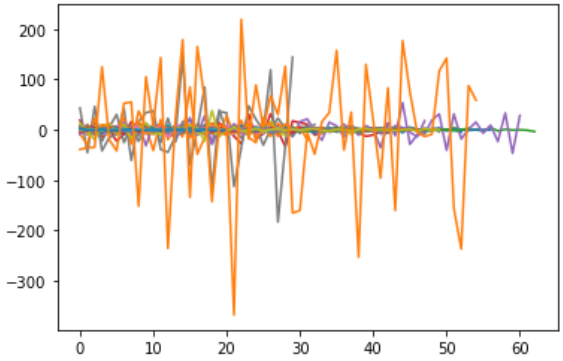

Figure 3: Pitch volatility over time

By the end of the research, I was able to generate a probability that a pitch was affected by dew point using a linear regression model on a pitcher by pitcher basis and assigning the probability that the pitch was affected by dew point to be the area under the curve between the pitch residual and 0.