Timothy Clay
tjc0426@gmail.com
Dew Point Project Write-up
October 23rd, 2023

Introduction

Throwing a baseball is a very fine-tuned science, with thousands of different components determining the exact flight path of the ball. One of those components and the focus of this project is humidity. Primarily, humidity affects pitches due to changes in atmospheric conditions. In a high-humidity environment, there is a high percentage of water vapor in air, which is what makes that weather feel so sticky and soupy. These conditions likely have countless unquantifiable impacts on how a pitch is thrown, but I will consider two quantifiable metrics that we might expect to change under different humidities: spin rate and movement.

For the purposes of this project, I will assume that, in a high-humidity environment, both spin rate and pitch movement will decrease. I assume spin rate decreases since the extra water vapor in the atmosphere and sweat from the uncomfortable conditions might make it harder for pitchers to comfortably grip the seams of the baseball. When pitching the ball, a worse grip means they will be able to spin the ball less, resulting in a lower spin rate. I assume movement will also decrease, in part because of the assumed drop in spin rate, but also due to the atmospheric changes. Since water vapor is much less dense than dry air, when there's a lot of water vapor in the air (i.e. when it's humid), the air as a whole will be less dense than low-humidity air. This, in turn, impacts the flight path of a pitched baseball because there are less molecules in the air for the seams of the ball to interact with as it travels. As a result, we would expect the magnitude of the horizontal and vertical break of the pitch to decrease in high-humidity environments.

To predict which pitches in the dataset might have been impacted by a high dew point, I used these characteristics (spin rate and movement) to preliminarily classify a subset of pitches as being high-humidity. With these identified pitches, I then used a cross-validated classification model on the entire data set to predict the probability that each pitch was one of these humidity-affected pitches. Finally, I visually tested a few examples to make sure my results logically made sense. I used Python for all of the steps.

Data Processing

The purpose of the data processing step was to choose a subset of pitches to classify as humidity-affected. I did this by comparing the spin rate, vertical break, and horizontal break for each pitch with other comparable pitches (same pitch type, thrown by the same pitcher). For ease-of-comparison, I calculated the z-score of each feature among the other comparable pitches. This way, each feature has a mean of 0 and standard deviation of 1, so no one feature outpowers any other feature (e.g. a spin rate of 2000 would have overpowered a horizontal break of 20). I then calculated the average of all three z-scores to compile these three features in one number. Interpreting this average, a high average means that the spin rate and movement for that pitch was better than the mean pitch among comparable pitches, and a low average means that the spin rate and movement would be worse.

With this average, I was then able to select a percentage of the pitches, which I considered to have been humidity-affected. Using historical data, I found that, over the course of the MLB season, about 25% of games at Great American Ballpark are played with a dew-point greater than 65℉. For this reason, I used the bottom quarter of the data and called those pitches humidity-affected. With these pitches having been chosen, I was then able to apply machine learning methods.

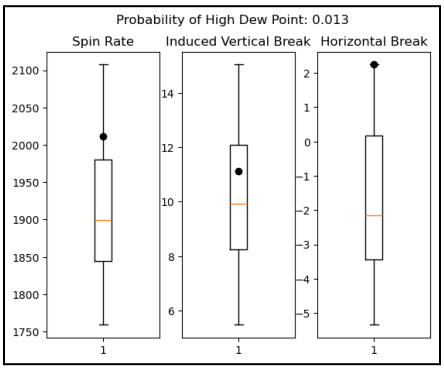
Machine Learning Methods

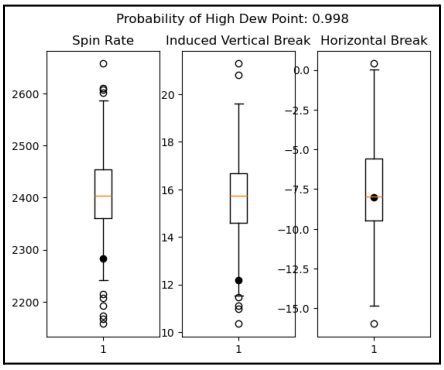
For each pitcher, I trained a separate XGBoost classifier model[1] to predict which of their pitches might have been humidity-affected, since different pitchers have different average spin rates and average pitch movement. Each model, however, used the same feature columns (SPIN_RATE_ABSOLUTE, INDUCED_VERTICAL_MOVEMENT, and HORIZONTAL_MOVEMENT) and predicted the same target column (IS_HIGH_HUMIDITY). For each model, I cross-validated the results, chose the best model (evaluated by f1 score), and saved it in a dictionary, with that pitcher's ID as the key.

Finally, for each pitch in the original dataset, I calculated the probability that it was a high-humidity pitch. I did this by taking the model for that pitcher (which was saved in a dictionary previously) and using the predict_proba() function, which, given the relevant stats for that pitch, returns a probability between 0 and 1 that the target variable (IS_HIGH_HUMIDITY) is 1. These probabilities were all appended together in a dictionary, which I then turned into a data frame and exported as a CSV.

Results

Finally, I double checked the results by visualizing a few samples of the data. I made a series of boxplots showing the distribution of spin rate and pitch movement for all pitches by the sample's pitcher. Overlaid on each plot is a point showing where the sample lies within these distributions, and at the top is the model's probability that that pitch was humidity-affected. Looking at a few samples, my model seemed to make sense based on the assumptions I made previously about how a humidity-affected pitch would behave. On the right are two graphs showing this visualization. The top graph shows a pitch with a low probability that it was affected by humidity, which makes sense since the spin rate and both vertical and horizontal breaks are above average for that pitcher. Meanwhile the graph on the bottom is the opposite, and, as a result, has a much higher probability of that pitch being humidity-affected.





Overall, for the assumptions that I made, I believe my model is fairly successful at predicting which pitches are humidity-affected. However, I am not 100% certain that my assumptions are always true and there could be several other factors that are near impossible to consider. For instance, I could see a reality where spin rate is instead improved by higher humidity (due to the resulting stickiness of mixing sweat and rosin). Humidity also doesn't take into account other weather factors. For instance, I would expect that wind would have just as significant an impact on pitch movement as humidity does, though wind is not considered at all in my methodology, which could be skewing results. Cases like these would be nearly impossible to determine without conducting a full-fledged experiment. Some assumptions must be made for an analysis like this to be realistic, but it is still worth considering whether those assumptions need to be questioned any further.

[1] Though it is not reflected in my notebook, I did try other models (e.g. KNearestNeighbors), but found the XGBoost classifier to be the most accurate.