2024 Reds Take Home Assessment: Determining the Probability a Pitch was Affected by Dew Point

The dataset provided contains information on 9,889 pitches thrown at the Great American Ballpark during a particular season. The primary objective of this analysis is to determine the probability that a pitch was influenced by a dew point exceeding 65 degrees Fahrenheit.

Initially, I considered employing either a logistic regression model or a tree-based model to calculate probabilities based on the given explanatory features. However, I soon realized that the data was unsupervised and did not explicitly contain information on humidity levels, making these methods unfeasible.

To address this challenge, I pondered how to establish a baseline for what a 'normal' or unaffected pitch should resemble. I conducted an Exploratory Data Analysis (EDA) with this concept in mind. My goal was to identify the most relevant features while assessing the variation in pitch flight and data across pitchers and pitch types. I specifically focused on features that I believed would significantly impact pitch flight and pitcher comfort. For each of these features, I computed the average values for each pitcher and pitch type and constructed histograms of these averages. This enabled me to visualize the range of "average pitches" for each pitcher and pitch type and assess for any notable distinctions.

Interestingly, there were considerable variations in the averages across all features for both pitchers and pitch types, particularly in terms of induced vertical break, spin rate, release speed, and release extension. Consequently, I decided to evaluate pitches based on both the pitcher and the pitch type.

I also utilized a correlation matrix and intuitive reasoning to identify five features that would be valuable for the analysis: induced vertical break, horizontal break, spin rate, release

speed, and release extension. These features were selected based on their correlations with other variables and their presumed relevance to pitch flight and pitcher performance.

To evaluate the pitches, I grouped them based on the combination of pitcher and pitch type. My objective remained to define what a typical, unaffected pitch should look like. Within each of these groups, I computed the averages of the selected features mentioned above. Subsequently, for each pitch, I calculated its Euclidean distance from its respective 'group average.' I then applied a Gaussian probability distribution to all the computed distances, which allowed me to determine the final probability that a pitch had been influenced by the dew point.

The distances of the pitches from the average effectively represented how dissimilar a pitch was from the norm. Thus, the calculated probability effectively represented the likelihood that the pitch deviated from the norm due to external factors. These "outside sources" might include various factors, not limited to dew point. However, by selecting features that are generally expected to be influenced by dew point and recognizing the unsupervised nature of the data, I was reasonably satisfied with this methodology for identifying the probability that a pitch was affected by a dew point exceeding 65 degrees Fahrenheit.