Dew Point Pitching

I was tasked to predict the probability that a pitch is affected by a dew point greater than 65 degrees F given a dataset with a variety of features and no target. My first step was selecting which programming language I should use to complete this task. I decided to go with Python because I am familiar with the language and have abundant experience in creating machine learning models with it. I organized my code using a Jupyter Notebook.

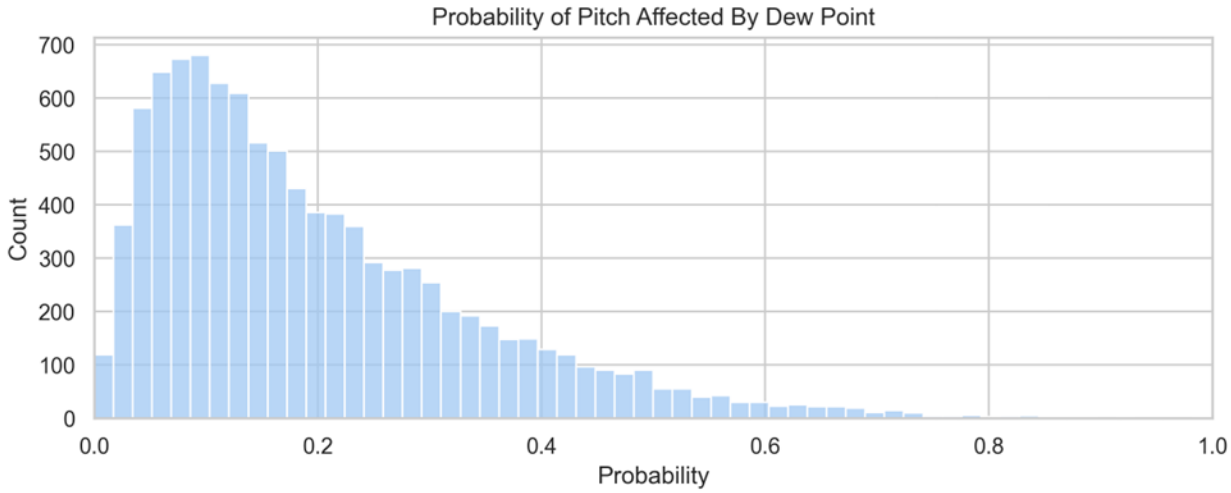
My next step was to determine which machine learning model would be used to complete this task. From the provided information, an anomaly detection model seemed like the best fit. Before I decided on a specific machine learning model, I needed to conduct research on humidity and its effect on pitchers and pitches. From my research, I learned that the primary impact of humidity is on a pitcher's delivery rather than the movement and shape of the pitch itself. Knowing this information, I selected features that describe the pitcher's delivery while also including features that can help distinguish pitches from one another. The features I selected were Release Extension, Release Side, Release Height, Horizontal Break, and Induced Vertical Break.

My next step was to determine which anomaly detection machine learning model I should train to complete this task. The model I chose was an Isolation Forest model. An important aspect of Isolation Forest models is that it works well on multi-modal data. The features of different pitch types are typically distinct from each other, which creates clusters within the data set. Isolation forest is designed to effectively handle these clusters and detect anomalies within them. Additionally, Isolation Forest outputs anomaly scores, which I can translate into the probability that a pitch was affected by a dew point greater than 65 degrees F.

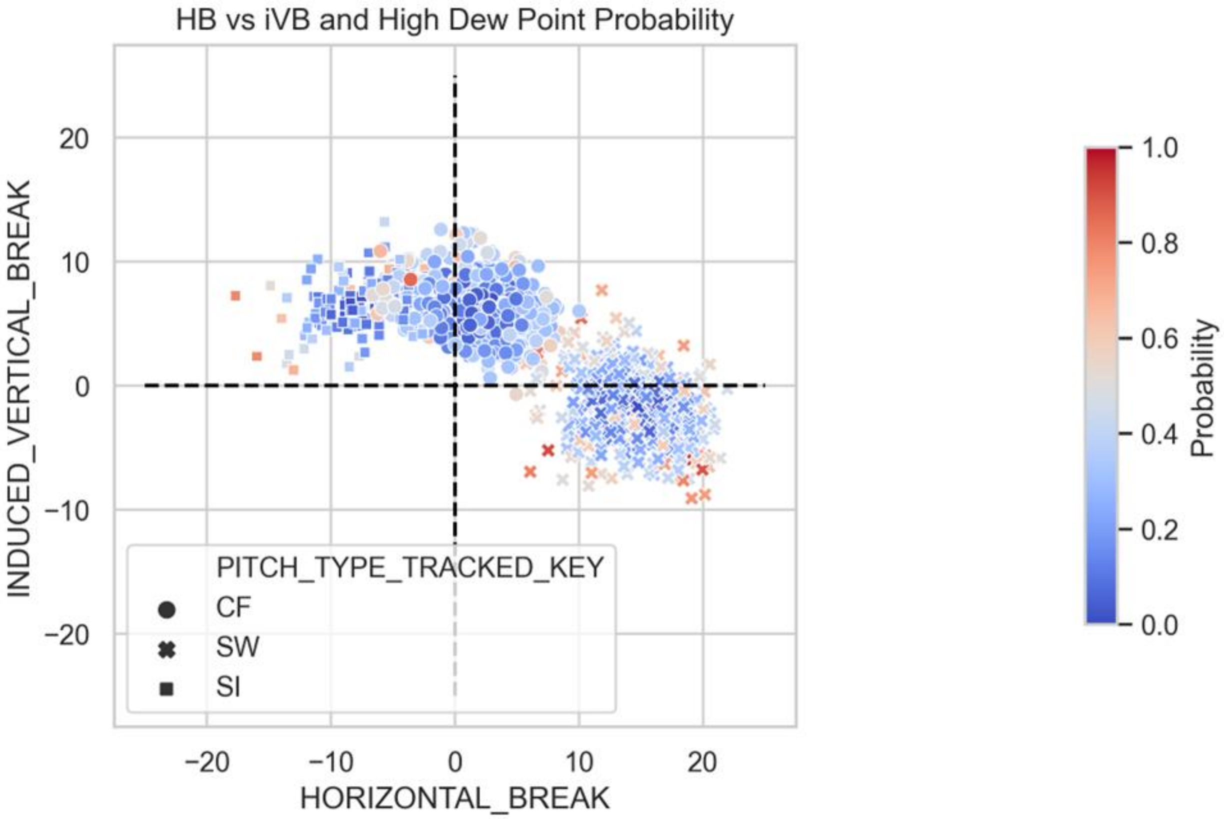
After opting to use an Isolation Forest model, the next step was preparing the data for training. Initially I thought that training a model for each pitcher was an effective approach to completing the task, but it was highly inefficient. Since I was dealing with an anomaly detection model, I was mostly concerned with determining outliers within the dataset. Knowing that humidity can alter a pitcher's delivery, I worked under the assumption that any deviations from their typical delivery could be attributed to an increase in humidity, no matter how the features changed. For example, if a pitcher has a change in release height, it would be considered a function of dew point regardless of whether the height was increased or decreased.

Using this assumption, I transformed the selected features from their raw values to the difference from the mean for each pitch, grouped by the pitcher and their specific pitch type. For example, while considering all sliders from a specific pitcher, the average metrics from that pitcher's slider were subtracted from the values of each individual slider to create a dataset of differences from the mean. This was applied to all pitches in the dataset. As the data set does not have classifications that could provide alternate explanations for anomalies, I assumed that all differences from the mean are caused by an increase in humidity.

Once the model was trained, I used it to determine the anomaly score of each pitch. Using these anomaly scores, I can convert them into probabilities of a pitch being an outlier. Since I assumed that all anomalies are a result of an increase in humidity, these output probabilities are the solution to this problem. The following figure is a histogram which illustrates the distribution of probabilities that a given pitch in this dataset was affected by dew point.



The following figure is a scatter plot of horizontal vs induced vertical break for the pitcher “668933” and the probability that a pitch was affected by a high dew point. This pitcher was chosen because he has the most total pitches in the dataset. As we can tell, different pitches have distinct breaks which helps illustrate different clusters. Additionally, we can see that pitches that sit away from the cluster’s centre are treated as outliers and assumed to be affected by high dew point.


HB vs iVB and High Dew Point Probability
INDUCED_VERTICAL_BREAK
HORIZONTAL_BREAK
PITCH_TYPE_TRACKED_KEY
● CF
✖ SW
■ SI
Probability