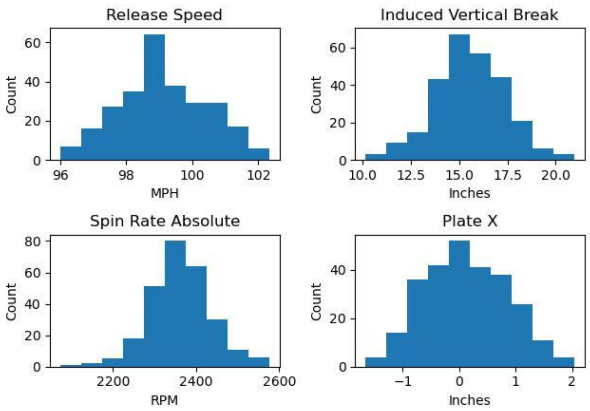
Reds Take Home Assessment – Selah Dean

Approach

The data provided did not have any information about the dew point, so an unsupervised model is necessary to predict the probability that the pitch was affected by the dew point. Two common unsupervised models are clustering methods and anomaly detection. Pitches that are affected by the dew point will be more likely to appear as outliers; therefore, using an anomaly detection model would be effective in predicting the probability that a pitch was affected by the dew point.

The features of the data provide information about the pitch itself as well as the situational data when the pitch was thrown. When looking at how the dew point affected a pitch, the situational data holds less meaning than the actual movement of the pitch. Similar situations can arise independent of the dew point, but the statistics that explain the movement and speed of the pitch are more susceptible to effects of the dew point. The statistics of the pitch movement and speed are approximately normally distributed for each player and each pitch type. This can be seen in Figure 1 which shows a few of the aspects of all the fastballs thrown by the pitch with ID 66881. Since each pitcher and pitch type have their own distributions when looking for anomalies it is better to look at the model for each pitcher and pitch type instead of the data as a whole.

Figure 1: Feature Distribution for Fastballs for Pitcher 66881



Methods

I chose to use an Isolation Forest Algorithm to perform the anomaly detection on the data set. I used Python and the sklearn library to perform the analysis on the data. I fit an isolation forest model for each pitcher and each pitch type. The model gave the anomaly score for each pitch. I used the anomaly score to compute a probability for each pitch type. The anomaly score is a number between -1 and 1 where scores closer to 1 means the pitch is more likely to be an inlier and scores closer to -1 means the pitch is more likely to be an outlier. I scaled the scores by the minimum and maximum to get a range of values from zero to one that represented the probability that the pitch was not affected by a dew point of greater than 65 degrees Fahrenheit. I found the probability that the pitch was affected by a dew point of greater than 65 degrees Fahrenheit by taking one minus the probability calculated using the scaler.