Zach Landry

Dew Point Pitching Project Write-Up

Introduction

This problem asks for a predictive machine-learning model to estimate the effects of dew point on pitched-ball flight using the provided data. The problem assumes that weather impacts several aspects of the game – namely the physical effects on ball flight and the pitcher's comfortability. For example, the pitcher and fielders playing in humid conditions might struggle to get a good grip on the ball caused by increased perspiration, possibly affecting the pitcher's command or velocity. Other weather-related factors such as temperature, air pressure, and wind can also influence ball flight. Initial research indicates that higher humidity decreases the air density due to the additional hydrogen molecules replacing the heavier nitrogen molecules. Therefore, the inference can be made that increased humidity causes less air resistance on the baseball. In terms of pitch attributes, the decreased drag could lead to less movement on certain pitches. Since the data is specific to only pitches thrown at Great American Ballpark, altitude remains constant and will not be accounted for in this model. Considering the above assumptions and factors, the model's output is a dew-point-affected probability for each pitch in the data.

Exploratory Data Analysis

As mentioned above, the data is limited to pitches thrown in Cincinnati and consists of 37 total pitchers who are anonymized using a PITCHER_KEY number. There are 9,889 total observations – each associated with a unique PID – and 26 categorical and numerical variables that describe the situation, result, and characteristic for each pitch. It is important to note the absence of any weather-related variables in the data. Ball-flight variables include INDUCED_VERTICAL_BREAK, HORIZONTAL_BREAK, SPIN_RATE_ABSOLUTE, RELEASE_SPEED, HORIZONTAL_APPROACH_ANGLE, and VERTICAL_APPROACH_ANGLE. The variables PLATE_X and PLATE_Z show the location of each pitch and are thus associated with command. Looking at the box plots of the ball-flight variables, each one appears to have outliers except HORIZONTAL_BREAK. The skewness values indicate that all except for HORIZONTAL_BREAK are negatively skewed, but since they are all between -1 and 1, their distributions can be considered approximately normal. The two most negatively skewed variables, SPIN_RATE_ABSOLUTE and RELEASE_SPEED, are likely a result of situations where position players are brought in to essentially throw glorified batting practice during a blowout. Nevertheless, the problem wants a probability for every pitch in the dataset so these observations will remain in the dataset. Lastly, there are no missing values in the data except for EVENT_RESULT_KEY, which is to be expected given that not every pitch results in an event.

Methodology

Since the initial research indicates that ball flight is affected by dew point, the effect will most likely reveal itself by deviations in the associated variables mentioned above. Therefore, the model could theoretically estimate the probability of dew-point effects by comparing an

individual pitch's metrics to the pitcher's typical profile for each pitch in his arsenal, which is calculated using the mean of each metric. The model will focus specifically on deviations in vertical (IVB) and horizontal (HB) movement since they are a byproduct of spin and velocity. It follows that the difference from the mean, the higher the likelihood of dew-point effects. Whether the dew point results in a positive or negative effect on movement is determined by the following intuition:

- Horizontal/vertical movement should decrease from a higher dew point due to the reduced air resistance working against the spin.
- Air resistance has negligible physical effects on velocity and spin rate since they contribute to the drag and Magnus force, respectively, which affect the trajectory, i.e. movement, of the pitch. Hence, any changes in velocity or spin rate will inherently affect the movement.
- The effects will be the same regardless of pitch type.

To control for the inherent variability among different pitchers, individual pitches will be compared relative to each pitcher's own set of metrics rather than comparing all pitches across the entire dataset. Hence, a new dataset is created containing each pitcher's average movement and merged with the original data before calculating the difference between the observed and mean value, DIFF_IVB and DIFF_HB. Absolute value is used to standardize these differences for righties and lefties. The distributions of average differences in movement for fastballs are shown in the box plots. Next, the model needs a target variable. This is based on the sum of the differences in IVB and HB to incorporate how much the movement differs from the pitcher's average. For cases in which this sum is positive, it is assumed that the dew point does not affect the pitch, resulting in a DEWPOINT_AFFECTED probability of 0. These movements are then normalized and essentially converted into probabilities by dividing each value by the minimum value, i.e. the biggest difference in movement. This approach is one way to convert the differences to a probability-like scale, but it assumes that this minimum value represents the most likely case of a pitch being affected by dew point, which may or may not be true.
There appears to be moderate negative associations between DEWPOINT_AFFECTED and DIFF_IVB/DIFF_HB, so a multiple linear regression (MLR) model will first be used to fit the data. After splitting the merged data into a 70-30 train/test split and training the MLR model, it produced an adjusted r-squared of 0.65, explaining approximately 65% of the variance. Plotting the predicted against the actual probabilities in the test set reveals that the MLR model tended to underestimate the dew point effects and generated some negative probabilities, which is not the goal. Thus, a random forests model is used to capture any non-linear relationships and stay within the target variable constraints. This model's predictions are much more accurate compared to the actual values with an r-squared of 0.991, and these final DEWPOINT_AFFECTED probability predictions are written to the 'submission.csv' file.

Conclusion

Ultimately, the random forests model was the most effective in predicting the dew-point-affected probabilities. Comparing the movement of each pitch relative to its pitcher is not the only way to estimate these effects, as other variables such as location could have also been considered. Future research could benefit from including weather-related data from multiple stadiums for a more comprehensive understanding of weather effects on pitch dynamics.