Before writing any code, an exploratory analysis dataset was conducted to identify columns that related to weather or humidity to create a binary variable for a regression model. Unfortunately, there weren’t any columns relating to weather, dew point, or temperature, so the manual creation of a continuous target variable was necessary to represent the effect of high dew point.

After doing outside research, with a high dew point, the ball is more likely to travel further due to air resistance changes or spin less due to pitchers having a harder time gripping the ball. After careful consideration of the research and the dataset, the decision was made to highlight columns that related to the pitch being released or the flight of the ball. These columns included Induced Vertical Break, Horizontal Break, Spin Rate Absolute, Release Speed, Horizontal Approach Angle, Vertical Approach Angle, Plate X, and Plate Z.

The code writing process began by loading relevant libraries of the following: ‘tidyverse’ , ‘readxl’, ‘xgboost’, ‘caret’, ‘foreach’, and ‘doParallel’ packages to facilitate data manipulation and machine learning tasks. The dataset was then loaded, and the relevant columns were transformed and renamed to streamline code writing. To ensure consistency in data analysis, adjustments were made to the Horizontal Break and Horizontal Approach Angle columns to disregard pitcher throwing side information and to ensure positive values regardless of pitcher throwing side.

The next step included outlier creation as a statistical method to identify high dew point likelihood. An outlier threshold, or Z-score, of 1.28 was used, allowing for the top and bottom 10% of values to be used as potential outliers. The dataset was grouped by pitcher key and pitch type to ensure the comparison of pitches only related to the same pitcher and of the same pitch type. For each relevant column, one new column was created to hold a binary variable indicating whether that pitch was an outlier for that column. An iteration process occurred to populate the 8 new columns with the relevant data. Afterward, an outlier percentage column was added to determine the proportion of outlier columns for each pitch.

The outlier percentage was identified as the target variable, and the 8 relevant columns were identified as features. A 70-30 train-test split was created to prepare the data for employing a XGBoost regression model. To optimize model performance, the number of boosting rounds, maximum depth, and learning rate were all used as hyperparameters in the model.

Model performance was evaluated by calculating the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-Squared (R2). While the initial model resulted in an R2 of .432, further adjustments were made to improve the model performance to better predict high dew point likelihood.

A K-fold cross validation with k = 5 was performed to gauge the model’s robustness. However, after completing the cross validation, the MSE results, .0185, showed only minimal difference from the initial XGBoost model, .0181, so the results weren’t used to optimize the model.

A grid of hyperparameters was created, including the number of boosting rounds, max depth, and learning rate. An iterative process was used to highlight the optimal combination of hyperparameters for use in the model. From this process, the RMSE was lowered from .127 to .122.

With the best hyperparameters identified, a final XGBoost model was trained. Evaluation of the model revealed an improved R2 of .467. All 9889 pitches then had high dew point probabilities created from the model, showing an R2 of .8.

Overall, this project led to the creation of a predictive model for determining the likelihood that a pitch was affected by a high dew point. This model’s ability to explain 80% of the variance offers insights into the impact of dew point and humidity on baseball pitch flight performance.