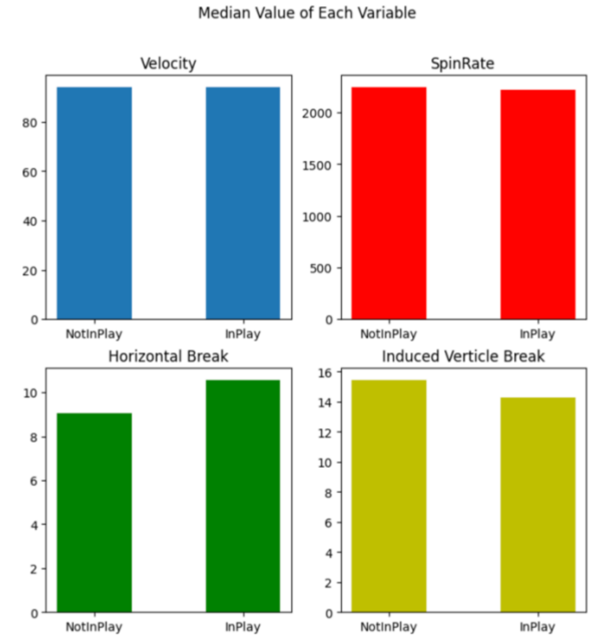
1. Predict the chance of a pitch being put in play. Please use this model to predict the chance of each pitch in the “deploy.csv” file being put in play and return a csv with your predictions.

2. In one paragraph, please explain your process and reasoning for any decisions you made in Question 1

Beginning with preprocessing, there were only six observations with missing values. I decided to impute the median value, as it is more stable than the mean. I decided to remove observations with a velocity less than 80; when I plotted the data these observations appeared to be outliers, and could be caused by faulty readings or mis-pitches. To be able to perform hyperparameter tuning and evaluate whether the model was overfitting or underfitting, I split the training data into training and validation datasets using a 70/30 split. I standardized each feature such that they had a mean of 0 and variance of 1 so that each feature has a common scale. I tested multiple models, including decision trees, random forests, support vector machine classifier, and xgboost. Each one performed similarly, but I decided to go with an xgboost model, hoping it would perform the best on the testing data due to the extra capabilities of the model from boosting and focusing on observations which are harder to classify. The training data contains approximately 73% observations with a label of 0, and the model could easily obtain 73% accuracy however it was predicting every observation as class 0. I used gridsearchCV to perform hyperparameter tuning, to reduce overfitting while trying to balance high accuracy and high f1 score. Unfortunately, as f1 score increased, the accuracy decreased, however I decided to prioritize f1 score over accuracy to avoid models which did not learn, which is how I chose a learning rate of 0.05 and scale_pos_weight of 2.7.

3. In one or two sentences, please describe to the pitcher how these 4 variables affect the batter’s ability to put the ball in play. You can also include one plot or table to show to the pitcher if you think it would help.

The velocity and spin rate have the least effect on the batter’s ability to put the ball in play, while the horizontal break and induced vertical break have the most affect. A batter is more likely to put the ball in play when it has a higher horizontal break and a lower induced vertical break.



4. In one or two sentences, please describe what you would see as the next steps with your model and/or results if you were in the analyst role and had another week to work on the question posed by the pitcher.

I would perform more detailed hyperparameter tuning and analyze/interpret the feature importances more thoroughly to be able to provide the pitcher with more detailed information about how the horizontal break and vertical break affect a batter’s chance of getting a ball in play.

5. Please include any code (R, Python, or other) you used to answer the questions. This code doesn’t need to be production quality or notated.