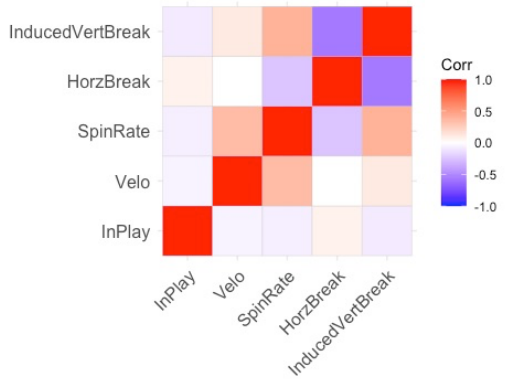
Toronto Blue Jays Baseball Research Analyst Assignment – Harrison Rubin

1. See the attached CSV file for my predictions.
2. I decided to use a random forest model to predict the probability of each pitch being put in play. I believe that a tree-based model, such as a random forest model, is optimal for this exercise because they tend to generate more accurate predictions for a dataset of this size. To construct this model, I first needed to handle the missing values for SpinRate, which was done by using a linear imputation. Next, I set the mtry argument to be a range from 1 to 3 in order to avoid model overfitting. After applying the hyperparameters to the random forest model, I evaluated the model's performance, which resulted in a maximum accuracy score of 0.73. After evaluating the model itself, I deployed the model on the "deploy" set, which generated the resulting probabilities of each pitch being put into play.
3. According to this correlation matrix, a pitch is slightly more likely to be put in play if it has greater horizontal break (weak positive correlation). However, a pitch is less likely to be put in play if it has higher velocity, spin rate, and induced vertical break (weak negative correlation).



4. If I were in the analyst role and had another week to work on the question posed by the pitcher, I would (1) build an ensemble model including components from random forest, neural network, MARS, SVM, and k-nearest neighbor models, and (2) explore ways to integrate other predictors into this ensemble model. From my own experience building similar models, I learned that various hit ball metrics (such as exit velocity, launch angle, and spray angle) are incredibly important in understanding the success of a pitch.