1. See file “deploy\_predictions.csv”
2. I assessed several different classification models – random forest, boosted trees, KNN, and logistic regression. The performance threshold to beat in the training dataset was 73% accuracy, which is what you would get by assuming every pitch was not put in play. This meant that the outcomes were highly unbalanced and so required an adjustment of class probabilities. Because of this I split data into 70% training, 10% evaluation for deciding probability cut points, and 20% test. I tuned the random forest, boosted, and KNN models using k-folds cross validation and then assessed the best cut point using ROC and closest to top left criteria. I then assessed the AUC for each model on the test data and found that all were roughly the same around .55, but that logistic regression was the highest at .56. Since they were all similar and logistic regression is the most interpretable of the models, I proceeded with logistic regression to make predictions on the deploy dataset. I excluded spin rate from the model because the coefficient was not significant and added a horizontal/vertical movement interaction term because the coefficient was highly significant. The exclusion of the spin rate variable also solved the problem of the NULL values in the training and deploy datasets. The NULLs were accounted for in the training of the other models by simply removing the 6 observations with a missing spin rate.
3. To start, you cannot reliably predict if a ball will be put in play based on these variables and the models I evaluated. The combination of horizontal and vertical movement appears to have the largest impact on whether a batter puts the ball in play: less horizontal break and greater vertical break appears to be the optimal combination for preventing a ball from being put in play from this dataset. The plots below show the differences between ball in play vs ball not in play across the predictors.

Chart, box and whisker chart

Description automatically generated

1. For next steps, I would like to add the additional 2 variables to the dataset: whether the batter swung at the ball, and whether the batter made good contact (exit velocity). These variables would help provide a better idea of whether a pitch is hittable when an attempt is made and whether contact is solid.
2. See file “Blue Jays Analyst Questionnaire.R”