Questions

In one paragraph, please explain your process or reasoning for any decisions made in question 1.

Since the pitcher asked for insights into how speed, movement, and spin affect the outcome of his pitches, I chose to build a logistic regression model with scaled inputs to make the model more transparent and interpretable.

I filled the null values in SpinRate with the median of the column. I saved that value and used it to fill the null values in the deploy_df dataset with the same number (zero leakage)

Ultimately I chose to remove SpinRate from the logit model because: a) the p-values indicated it wasn't a significant feature, and; b) I suspect it affects the outcome primarily through the other features, so eliminating it would allow the other coefficients in the model to provide more insight.

In addition to the linear model I build a gradient boosting tree with a max_depth of 2 to minimize overfitting. I used the out of fold predictions to train a meta-model to combine them before training both models on the entire dataset and using them as inputs to the meta-model.

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 SpinRate primarily affects the InPlay outcome through the other variables.

As velocity increases the probability of a ball being hit InPlay drops. Specifically, for about every 2.8 additional mph the probability of a ball InPlay drops by about 2%

As HorzBreak increases the probability of a ball InPlay goes up.

InducedVerticalBreak has the largest effect on the outcome of the pitch and its impact gets even stronger as it increases.

If you could change on feature of your pitches to maximize the change in the InPlay outcomes it would be InducedVerticalBreak. The InPlay probability would drop by 5%+ for many pitches in the dataset if you could hypothetically increases it by ~4.5 inches.

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”d like to relate the features from this dataset with other relevant features such as ball placement in the strike zone and the ball-strike count.