Billy Fryer

Blue Jays Technical Interview

1. CSV returned is called “Deploy-With-Predictions.csv”
2. Given that this question is binary in nature (1 if in play and 0 if not in play), I knew I needed to do some type of classification model. My initial thoughts were to try a logistic regression model as a baseline and see what was or was not working from there. The biggest issue I ended up having was a big class imbalance; there were a lot more balls not put into play than there were balls put into play. After trying to adjust for that by using a more stratified training set (as opposed to the random assignation as before), the logistic model was still not great. I advanced on to more complicated models such as a random forest and ultimately settled on the xgboost model as it had the highest test set accuracy and lowest log loss.
3. All four of Velo, Spin Rate, Horizontal Break and Induced Vertical Break are pretty much equally important in preventing balls from being put into play. Developmentally, this means that we should focus on whichever one is your weakest/the one that has the most room for growth and improve on that.
4. With any data at my exposure, I would ideally like to add a few more variables into my model such as location of the pitches and difference in velocity between fastball and off speed since these fastballs all have context in the game with pitches being thrown before and after.
5. Code is Included