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

I began by performing exploratory data analysis, investigating the distributions of each variable as well as the relationships between them. In doing so, I learned that about 27% of the fastballs in this dataset were put in play and that, on average, these pitches had a lower velocity, spin rate, and vertical break (and a higher horizontal break). From there, I began the model building process after changing the variable types as needed and removing missing values. Because the dataset was so large, I evaluated the performance of each model on a single validation set with area under the curve (common for binary classification problems) as the performance metric. After hyperparameter tuning, the logistic regression model outperformed both the random forest and gradient boosted tree models, so I decided to use this model to generate the predicted chances of each pitch being put in play in Question 1.