Please complete the following questions:

1. Predict the chance of a pitch being put in play. Please use this model to predict the chance of each pitch in the “deploy.csv” file being put in play and return a csv with your predictions.

See pdf code file in github link.

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

Since the “InPlay” column operated as a binary variable, the best course of action for completing this assignment was to build a logistic regression model to answer the question. Before building the model, I wanted to clean the csv of any non-numeric values within the rows or values that would be difficult for the model to produce an output which is why I stored both files as new data frames using the ‘Numpy.isfinite()’ function. The deploy dataset also does not contain an InPlay column so I created that column to add from the training.csv column. I could not complete this assignment without the use of that variable which is why it was necessary to pull that column from the training.csv. Once I cleaned both datasets I split the cleaned deploy.csv into a training and test set for the model. This allowed me to fit the model to the data set and then produce an unbiased result.

3. 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.

A pitch type can be the same, but the surrounding and voluntary effort of the pitch has a profound effect on the batter’s reaction and ability to act on the pitch.

4. In one of 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.

Next steps would be to further investigate possible tweaks or replacements to the model. The model returned a 0% probability of a ball being in play, but a 74% probability of a ball going out of play from the data set, which is alarming as there are values within the data set that show a ball was actually put into play.

5. Please include any code (R, Python, or other) you used to answer the questions. This code doesn’t need to be production quality or notated.

See pdf code file in github