**Q1.**

**PredictionsChance\_log.csv is the output file for logistic model.**

**PredictionsChance\_random.csv.is the output file for Random Forest model.**

**PredictionsChance\_knn.csv.is the output file for KNN model.**

**I write it in python.**

**Q2.**

**I used logistic regression since it’s a simple binary classification model. I considered the four variables as features to predict the likelihood of a pitch being put in play. First, I cleaned the data, removed the nan value and check the distributions of variables. Second, I established the model and tested it on the deploy dataset. Third, I return the possibilities of class 1 instead of 0/1. Since we have not figure out the thresholder. Additionally, I use random forest model and K-NN model to see if they improve the prediction. I use the grid search with cross validation to find the best hyperparameters.**

**Q3.**

**Higher velocity and IndecedVerBreak make it slightly harder for batter to put the ball in play. While increased HorzBreak make it slightly easier. Spin rate has a negligible effect since the coefficient is too small.**

**coefficient from my model: -0.00217874 -0.00023124 0.01507367 -0.02962609**

**Correlation matrix:**

**InPlay Velo SpinRate HorzBreak InducedVertBreak**

**InPlay 1.000000 -0.052441 -0.065993 0.072993 -0.093791**

**Q4.**

**If I had another week, I would evaluate the logistic regression model and figure out the appropriate threshold using confusion matrix. Secondly, I will consider using some other modes: beta-binomial model, neural network, Decision Trees to see if they improve the prediction accuracy. Additionally, I’d like to include more features such as batter’s statistics.**