For this process, the first step was to create the “is\_swing” variable in the training data set in order to use it as the outcome variable for a binary classification model. After this, the predictor variables were chosen based on what my baseball knowledge tells me are important factors in a batter’s swing decision: level, pitcher and batter handedness, outs, count, pitch type and velocity, horizontal and vertical break, and location of the pitch. The data was then cleaned to remove any extreme values or impossible values that imply a mis-read or incorrect entry.

Once the data was prepared for modelling, it was randomly sampled into a 10,000 row data frame which was then partitioned into a train and test set in order to fit and test preliminary models. Four different models were tested on the subset of data: a logistic regression model, gradient boosted machine, a neural network, and a random forest. Each model was used to predict whether a pitch was swung at or not using the subsetted test data, and confusion matrices were generated to evaluate the performance of each. It was evident that the gradient boosted machine and random forest were the two best models, at 76.36% and 76.39% accuracy respectively. After comparing other values in the confusion matrix and looking at ROC curves, as well as considering the ease of tuning and risk of overfitting of the two models, the random forest model was chosen as the best option.

After choosing the random forest, it was tuned to provide maximum accuracy and mitigate risk of overfitting. The model with tuned parameters was then applied to a larger subset of the training data (50,000 rows). After fitting the final model, it was used to predict the “is\_swing” variable in the test data.

The biggest area for improving this model would be having access to a couple new variables in the data. In baseball, there are very complicated circumstances that go into a batter’s decision to swing and how aggressive they will be. Where and who the baserunners are is an example of this. Additionally, the hitter’s strengths and tendencies play a big role in the probability of a swing at a given pitch. For example, an average changeup thrown over the middle of the plate might be likely to be swung at with a power hitter at the plate and no baserunners in a 1-0 count. However, that same changeup with a less productive hitter and a good base stealer on first base becomes much less likely to be swung at. So, having data about the individual hitter’s swing rate, as well information about baserunners could help create a better model.