TBJ Interview Questions

Evan Howard

2023-09-07

Question 1

Predict the chance of a pitch being put in play.

This first question asks me to predict the chances of a pitch being put into play by using the training and deploy data sets that were given to me. You can see the predictions in the file titled Evan-Howard_PredictedHits.csv. Below are the top 15 rows of data from that file. Predicted_Hits are represented as binary with zeroes for not in play, and ones for in play.

```
##     X  Velo SpinRate HorzBreak InducedVertBreak Predicted_Hits
## 1   1 94.72     2375      3.10            18.15              0
## 2   2 95.25     2033     11.26            14.50              0
## 3   3 92.61     2389     11.00            21.93              0
## 4   4 94.94     2360      6.84            18.11              1
## 5   5 97.42     2214     16.70            13.38              0
## 6   6 95.98     2495     11.25            17.12              0
## 7   7 94.88     1998     15.13            15.22              0
## 8   8 92.73     2049      1.55            18.47              0
## 9   9 92.39     1955     18.15             7.25              0
## 10 10 95.77     1976     10.04            14.56              0
## 11 11 93.27     2076     11.45            11.50              0
## 12 12 91.95     2465      9.12            21.00              0
## 13 13 92.24     2391      0.56            13.23              0
## 14 14 93.10     2130      3.97            16.75              0
## 15 15 95.38     2262      7.44            17.64              1
```

Question 2

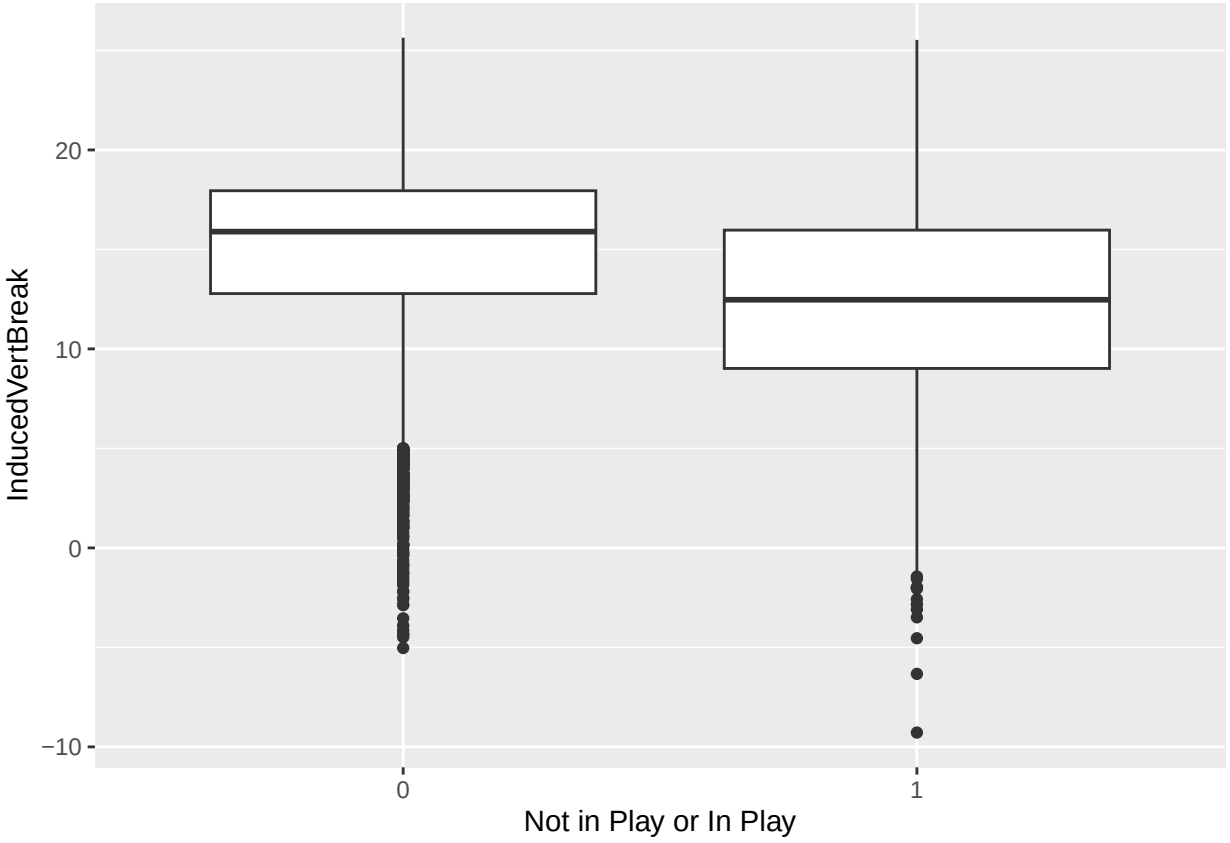
Explain your process and reasoning for any decisions made in Question 1.

I chose to employ a Random Forest algorithm because of how robust it is, and its ability to handle numerical and categorical features. I initially saw that the SpinRate column had character values so I decided to switch that to numerical for both the training and deploy data sets. This ultimately created a problem as it created a couple NA values in both data sets. The training data set had seven more rows than the deploy data set. I removed these seven rows randomly in order to stay objective. My model ended up having an accuracy of 60%, which I know is not very good. Ultimately, I would like the model to be in the 80-85% accuracy.

Question 3

Please describe to the pitcher how these four variables affect the batter's ability to put the ball in play.

Spin rate will tell you how effectively you are throwing a certain type of pitch, while induced vertical break will tell you how well your fastball rides upwards thus making it harder to hit. Horizontal break tells you how much your ball moves in, or out, to a hitter and velocity tells you how hard you're throwing the pitch. When taken in combination we can try to increase or decrease these four variables to find the right combination for you that makes your fastball, or any pitch in your arsenal, the most effective in getting outs.



As you can see in the boxplot above, fastballs with a higher induced vertical break play better in getting more not in play events.

Question 4

Please describe what you would see as the next steps with your model.

The future steps I would like to take with my model is, first and foremost, to increase the accuracy of the model. Next, I would like to add other variables such as release height, pitch location, as well as other pitch types.