MAT 536 Final Project

Michael Pena

May 17, 2024

Summary

We have been asked to assess whether or not the number of balls impacts the the probability of an umpire calling a strike during a baseball game. After some analysis, the box plot below (Fig. 1) shows how the probability of the umpire calling strike differs from when there has been no balls, to when there have been three.

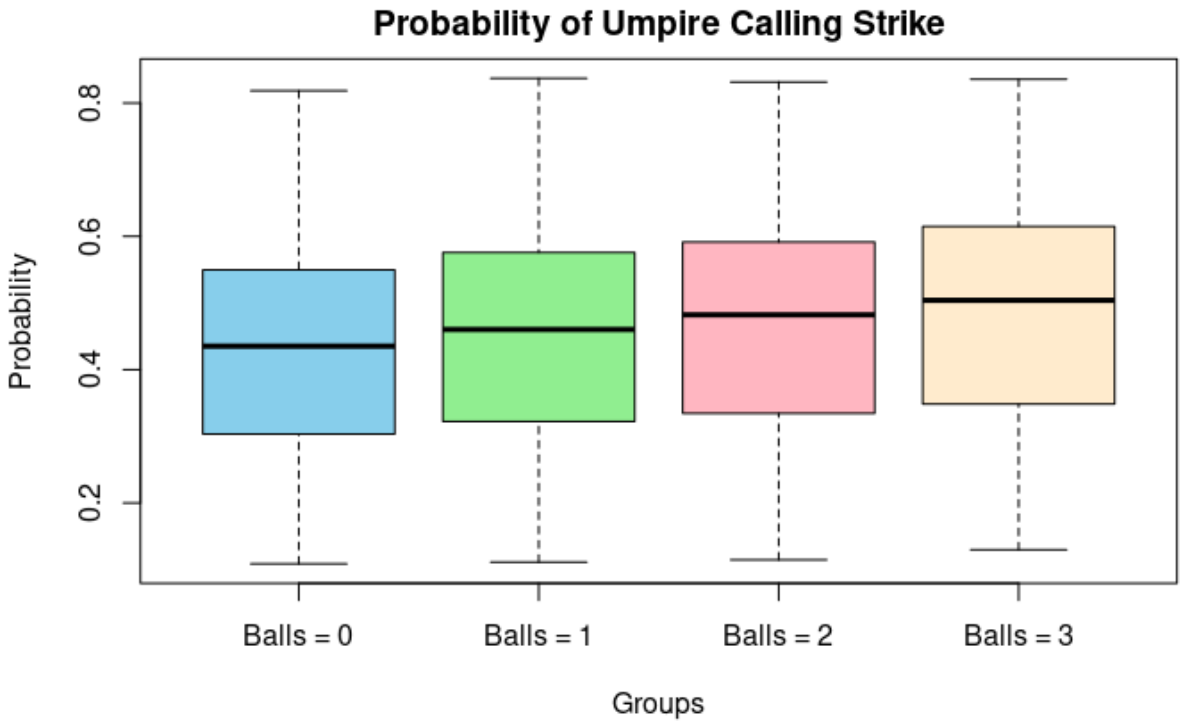

Figure 1: Enter Caption

We can see that there the probability of calling "strike" does, on average, rise as after each ball. It can also be noted that the spread is very similar throughout and that the average probability rises by 0.02 per ball (shown in Fig. 2).

mean	balls
0.4293092	0
0.4514809	1
0.4676672	2
0.487266	3

Figure 2: Average probability umpire calls "strike"

Methods

Our dataset includes features such as whether "ball" or "strike" was called by umpire, release speed of pitch, pitch type, the two-dimensional position of the pitched ball over home plate (as seen by the umpire), the current strike and ball counts, and the scores of both teams playing. The spatial data, describing the ball flew over the home plate, was a key factor in determining which data observations were most relevant for our analysis. The "fringe zone" is described as a spatial area over the home plate where the border of the strike zone in unclear. From the data, we were able to graphically represent these points and used only this observations for our analysis (see Fig . 3).

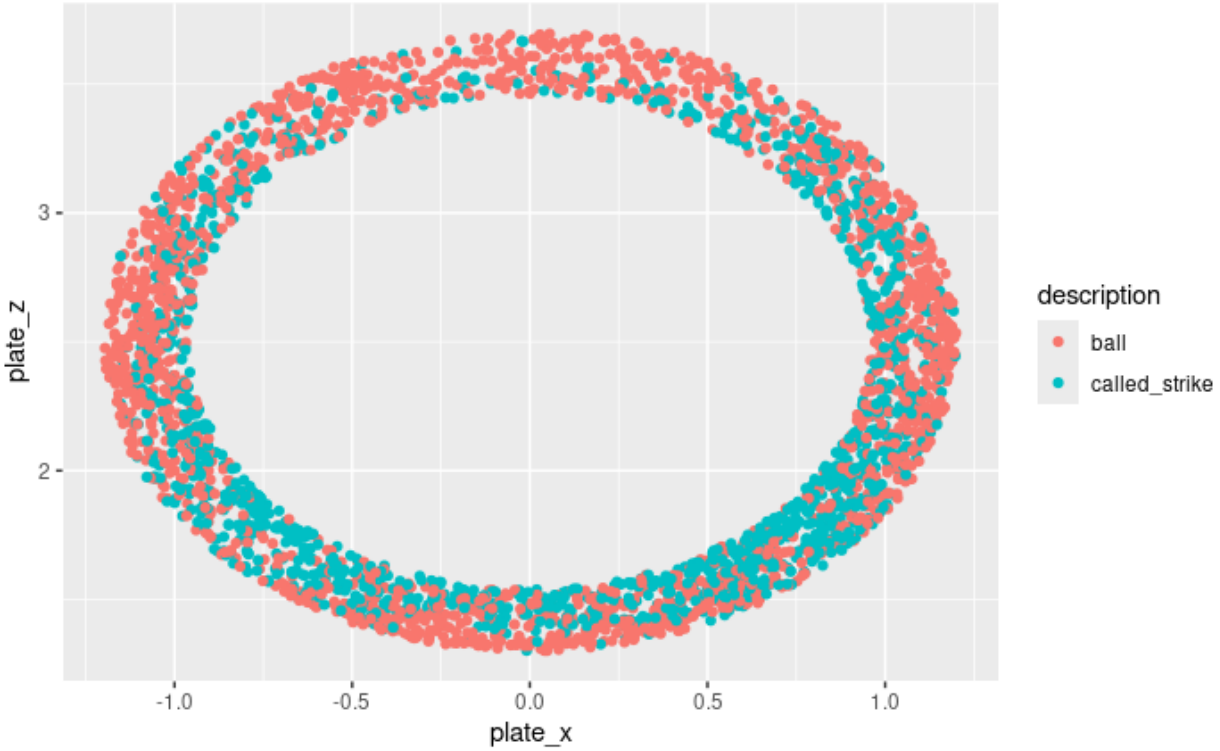

Figure 3: Plate z indicates the height at which the ball flew over the home plate. Plate x indicates the horizontal position at which the ball flew over home plate.

We found that between a Random Forest and Adaptive Boosting model, the Adaptive Boosting model (with 500 trees, cp of 0.01, and a minimum split of 10) we found ideal by using a with 5-fold cross Validation. 5-fold was chosen because we need something that was computationally inexpensive.