Blue Jays In-Play Probability

The problem was to take pitch features and create a model that predicts the probability that the ball was in play. Based on the fact that it is a probability, I started with the most generic approach, being logistic regression. I then moved to random forests, boosted forests, and finally neural networks (multi-layer perceptron). I performed grid searches across the parameters of these models. I found that the logistic regression did just about as good as any of the other classifiers. Interpretability is important in this project since the results are going to a pitcher. They need to know what their approach should be going forward, and a black box model can make that difficult. Logistic regression has much higher interpretability than the other models, and since the performance was just as good, I chose this model. I used the best parameters found in grid search and made a new model. I looked at the resulting coefficients and was surprised to see that horizontal break had a positive effect on in-play probability. It occurred to me that this may be due to a relationship between vertical and horizontal break since logic says that they should likely have a negative relationship. After performing a quick linear regression, I found that they had a negative relationship with a strong R-squared. In order to try to mitigate the bias on the results from this collinearity, I incorporated an interaction term and ran the model again. This resulted in horizontal break having a negative effect on probability, as we'd expect. It also increased the negative effect of vertical break, as vertical break was probably also suffering from the collinearity. The positive coefficient for the interaction term also suggests that horizontal and vertical break have bigger individual effects when the other is larger. Lastly, it is assumed that it is important to have a big difference between the horizontal and vertical breaks. So, we can add this feature and rerun the model. This lessened the effects of horizontal and vertical break, but we can see that the difference does indeed have a negative effect. I ran a final grid search to find the best parameters with my new features.

Based on the results, if you increase any of velocity, spin rate, horizontal break, vertical break, or the difference between the breaks, the probability that the ball is put in play will reduce. Increasing the difference between horizontal and vertical break is more important than increasing either individually.

In future work on this problem, I would start by adding some additional features. I would start by adding higher order features to capture any nonlinearity and reduce bias.