Blue Jays Baseball Analyst Questionnaire

Jackson Thomas

Modeling Thought Process

After analyzing the shape of the data and the prompt at hand, I chose to use a logistic regression model. I selected a logistic regression model over a Naive Bayes model because a logistic regression model has less bias and the data's features are not independent. A Naive Bayes model assumes all features are independent; if they aren't, it greatly impacts the classification. Pitch movement, velocity, and spin have a clear relationship, and a change in one will affect a change in the others. Additionally, the dataset was large enough for a logistic regression model, as it has been known to underperform on small datasets. Given the data shape, prompt, and low-bias appeal of a logistic regression model, it is the best model for estimating ball-in-play probabilities.

Model Breakdown

An increase of velocity, spin rate, and induced vertical break all decrease the probability a ball gets put in play. However an increase in horizontal break increases the probability a ball gets put in play. (see visuals in code)

Next Steps

The next steps of the model would take into account pitch locations. Further analysis would need to be done on where to throw fastballs to get the lowest in play probability based on its velocity, movement, and spin characteristics.