Nick Arrivo

Toronto Blue Jays

Baseball Research Analyst

Technical Interview

For data preprocessing, I removed the NaN values from SpinRate, where Trackman or Hawkeye didn’t read the pitch. I also removed some outlier values in Induced Vertical Break that could have been misreads as well. Since Neural Nets are very sensitive to training set class imbalance, we can expect a neural net model to have a strong bias towards predicting a pitch as not in play since the training set is about 80% balls not in play. My Neural Net model was very hesitant to predict balls put in play. One could change the binary functions threshold to lower than 0.5 when turning the outcome data into predicted data, but that would not be best practice. So, I opted to use a Random Forest Model which can also discern feature importance from the set of predictors. I would use the results of feature importance to give insight into what characteristic of fastballs should be optimized. My Random Forest Model was trained to about 70% accuracy on a partition of the training set, which was by design as I wanted to avoid overfitting. My model was still more likely to predict a ball would not be in play, which I why I would have preferred this model to be trained on pitches that were swung at. I think that my model scaled well to a new data set, and the list of feature importance is as follows. Horizontal Break is surprisingly the most important predictor of whiffs by a small margin, then IVB, then Velo, then Spin Rate. This list of importance for the model’s features are valuable insight that can have on-field applications to Blue Jays pitchers.

A graph of blue rectangular bars with white text

Description automatically generated

When working with a pitcher who wants to know about Spin Rate, Horz Break, and IVB, I would show them this chart, which is from the results of the model. The model predicts that increases in IVB, spin rate, and velo increase the chance of a whiff. Depending on this pitcher’s fastball profile, I would show how less horizontal break increases the chances of a hitter whiff. Horizontal break is going to help if you have a cutride fastball, you are more likely to get whiffs than a straighter profile. A cut-ride fastball is associated with a negative horizontal break. I would explain that for a pitcher with a straighter fastball, IVB is the name of the game and maximizing it is going to give you that “ride” effect that many of the top arms have. Velo is obviously going to help get more swings and misses. Spin rate can be tricky and is not as strong of a predictor of whiffs, but in certain cases, a pitcher with a inefficient spinning fastball can increase their IVB and Horizontal break by improving their spin efficiency.

A group of diagrams with different colored squares

Description automatically generated with medium confidence

If I had another week to work on this data set as an Analyst for the Blue Jays, I would delve further into the outcomes we can predict based on fastball type. I would like to find out what fastball characteristic is the most important for *them* to get more whiffs. This would involve more focused data sets that contained data from one pitcher. I would also spend time discerning a swing and miss from a take so that this data is more tailored to be used in a model that predicts whiffs. Furthermore, I would like to look at how their fastballs perform against different types of hitters and tailor their pitch calling based on the hitter. I ran a simple k-means cluster, but according to the feature importance model, it was not a strong indicator of predicting whiffs.