**Question 2**

A random forest classifier was used to predict whether the pitch would be put in play or not depending on the characteristics of the pitch, including velocity (mph), spin rate (rpm), horizontal break (inches) and vertical break (inches). I chose a random forest classifier since it does not overtrain easily, is easy to implement, and is highly explainable. The first step was dropping rows that contained NaNs, as although some random forest models can natively handle NaNs, sci-kit learn’s implementation cannot. I used a Gini index to compute the error or quality of the split at each node as it has been shown to reduce error classification problems, specifically related to decision trees.

**Question 3**

After computing the importance of each of the parameters used to determine whether the ball was put in play or not, the model concludes that all variables have between a 20-30% impact on whether the ball was put in play, but maximizing vertical break appears to be most important to reduce the amount the ball is put in play. This is shown in Figure 1 below.

A graph with blue bars

Description automatically generated with medium confidence

Figure : Bar plot of feature importance showing the importance of different parameters on whether or not the ball was put in play.

**Question 4**

Provided with more time, I would try and conduct hyperparameter tuning of the random forest classifier I used, as this can greatly improve the model’s performance. I would also try multiple different models, including a gradient boosted decision tree, which I find can have a slightly higher performance than random forests, but can overtrain if careful feature engineering and hyperparameter tuning is not conducted, and an artificial neural network, as they tend to scale very well, and performance can greatly improve when large amounts of data is used in training. This makes it a good model for deployments where we are accessing the API frequently and new data is constantly being used to re-train the model.

Finally, I would conduct deeper analysis into the results of the model and how this might be used to gain an edge over the competition. This would require further research into each of these pitch parameters. Using any additional data I could to try and improve the model performance, such as the type of pitch, weather (temperature, humidity), opposing pitcher, the batter, and any other data I may have access to may provide further insights.