Blue Jays Assessment

Question 1

The file “deploy_new.csv” includes the model predictions for the probability of a ball being put into play given the features provided.

Please review the Jupyter Notebook “technical_assessment_blue_jays.ipynb” for code.

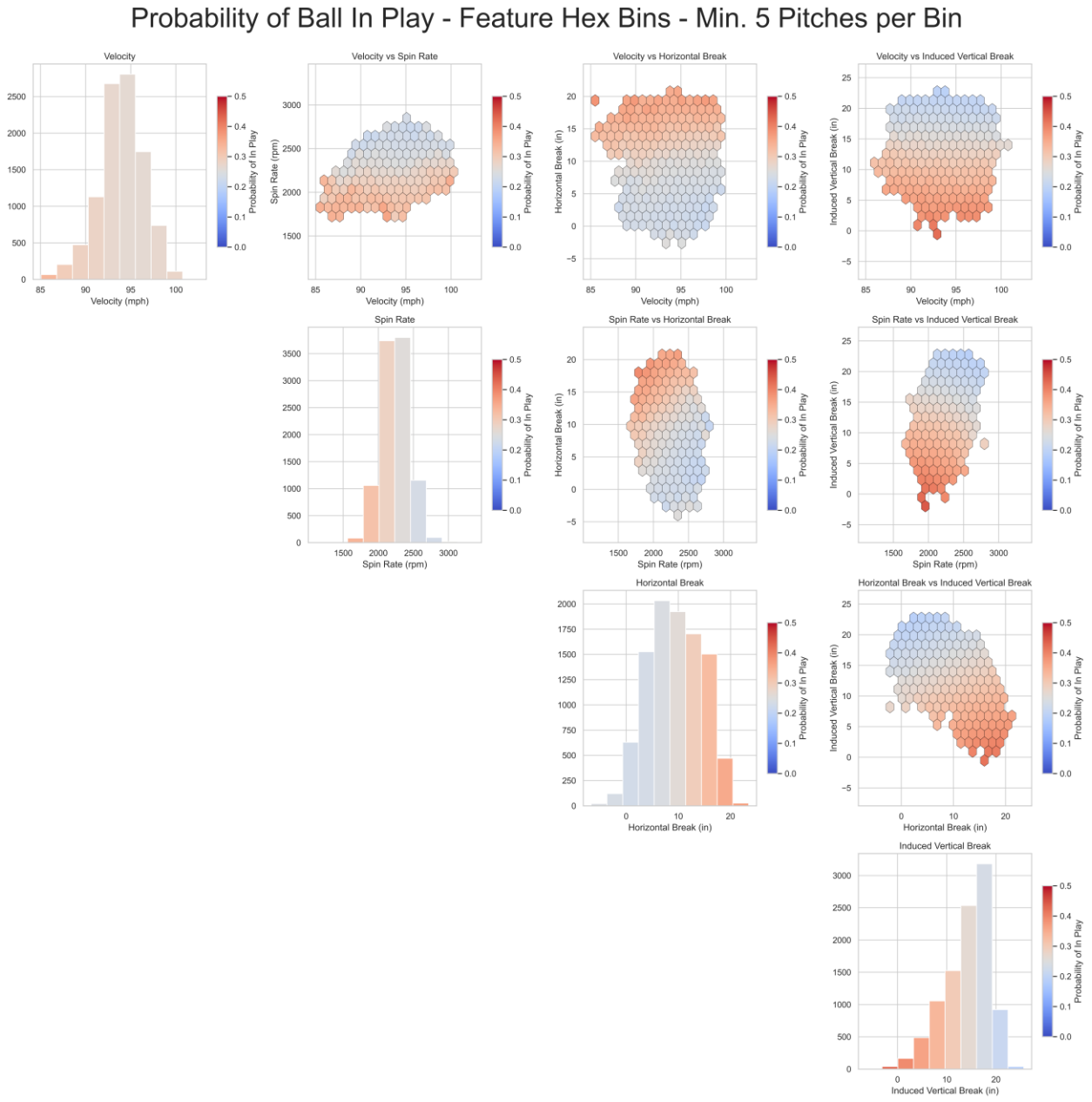
Question 2

I was tasked to predict the chance of a pitch being put into play given a dataset with a variety of features and a binary target. My first step was selecting which programming language I should use to complete this task. I decided to go with Python because I am familiar with the language and have abundant experience in creating machine learning models with it. I organized my code using a Jupyter Notebook. My next step was to prepare the data for training. I removed any rows which contained NULL data. An initial analysis showed that approximately 0.2% percent of the pitches thrown were below 85 mph; as I am analyzing fastballs, these pitches were considered outliers and filtered from the dataset. Following that, I split my data for the purpose of training and testing. The most integral step in the process was considering which machine learning model I should utilize. I decided to use a classification model because I was dealing with a binary target, "InPlay". Since the task was to predict the probability of a pitch being put into play, I decided to use a logistic regression model. This type of model can estimate the probability that a pitch will be put into play given the inputted features and will always provide an output between 0 and 1. Following the training and testing of the model, I was content with the results, so I proceeded with predicting the probability of the pitches in "deploy.csv" being put into play.

Question 3

The plot below uses the data in “deploy.csv” and illustrates the impact each feature has on the predicted probability of a batter putting the ball into play (Red represents more likely, blue represents less likely). Additionally, the relationship between each of the features is presented.

From these plots, we can infer the effect each feature has on likelihood of a ball being put into play. Fastballs with higher spin rate, less absolute horizontal break, and greater induced vertical break are pitches that result in fewer balls in play, while pitch velocity has less of an impact.



Question 4

My next steps with my model would be adding in other features such as pitch location, whether the pitch was swung at, and At-Bat count as independent variables to see how they affect the 4 initial variables. Additionally, I would be interested in diversifying the dataset to look at other types of pitches, not just fastballs.