**Cincinnati Reds Technical Challenge Report:**

The goal of this report is to explain the process, the data used, and the data obtained to answer the question of predicting future MLB player pitch proportions received. To begin this assessment, I utilized 3 years of batting data for players who faced a large sample of pitches in 2024, around 1000. While there are many pitches in the MLB, this task focused on using the 3 main categories: fastball, breaking ball, and off-speed pitches. To prepare the data for prediction, I made a feature set, consisting of 13 features for each batter. I then ran this feature set through a simple neural network, using historical pitch proportions as weights, to obtain predictions for the 2024 season.

While many pitches clearly fall into a certain category, there are others that could fit in multiple. I want to specifically address the rationale for these fringe pitches. I chose to include the split finger in the fastball category, since it often has similar movement. Although curveballs are a slower pitch, I opted to place it with breaking balls due to the extreme movement on the pitch. A similar choice was made for the slider, I wanted to separate the pitches that tend to have a lot of movement, even if they could technically be considered off-speed for many pitchers.

To establish my feature set, I used data frame operations in the pandas library. The dataset lacked many helpful statistics like batting average, on base percentage, etc. I began grouping by player to get some standard values such as hits, pitches faced, and at bats. This aided me to calculate more meaningful features for my machine learning model. I really wanted to focus on providing features that differentiated the strengths and weaknesses of players. The features that are very telling are the whiff and chase rates and average exit velocity for specific pitch categories. Without getting into zone specific hot/cold spots and zones where hitters struggle with each pitch, these stats provide a good baseline. Looking at the pitches players chase, whiff, or hit weakly can give insight into the pitches they struggle with. My aim was to feed these to the neural network so it can learn based on the tendencies of the data.

To make the predictions, I used a simple neural network, with 3 linear layers, 2 hidden layers, 1 dropout layer, and the softmax function on the output to generate the proportions. This structure provides a quick model to run predictions for my data. To ensure that the data was avoiding equal proportions for each pitch type, I added weights to the loss function. The weights were the historical proportions of pitch types across all pitches thrown from 2021-2023. I ran the neural network across 1000 epochs to give it enough time to train.

I would like to quickly talk about how I would enhance this analysis. Due to the nature of this challenge, I did not spend more time creating or researching a more elaborate neural network. I used a basic structure that would get me predictions based on my feature set. With more time, I would research improvements to the neural network and the structure of the training to gain more accurate predictions. I also would spend more time on the feature set. There were a lot of statistics that I left out of this model, which could aid the predictions. Another aspect I would test would be running each player through the neural network separately, replacing the current feature set with a pitch-by-pitch feature set for each player. This would help reduce the risk of learning tendencies across players, which is not a factor in baseball. Each hitter is unique and there are not necessarily tendencies across batters. These future changes would allow me to develop better predictions.