2023 Reds Data Scientist Problem Set

In the `data` folder, you will find three files:

- `train.csv` which is 100k rows of randomly selected pitch data from the 2021-2022 season, consisting of both MLB and triple-A level pitching
- `test.csv` which is 9295 rows of pitching data from five different pitchers, totaling one season-worth of pitches for each pitcher
- `sample_submission.csv` which is an example of what your question 1 solution format should look like

Using the provided data at your disposal, please answer the following two questions.

Question 1: Classify pitch types

Let's assume the data from `train.csv` is collected from a high tech scout. The scout is able to provide information on the speed of the pitch, the spin rate of the pitch, and the release point of the pitch (specifically the height of the pitch and the side of the pitch). The scout also provides information on pitch type for these particular pitches.

Using the information from our scout, please predict the pitch types thrown in `test.csv`. You can use any techniques you want; some statistical/machine learning techniques would be appropriate here. Comment your code thoroughly, so we can review your thought process on methodology for this question. The output file should have the same structure as `sample_solution.csv`.

Your solution will be evaluated by log loss.

Question 2: Rank the pitchers

In `test.csv` there is additional data that is not in `train.csv`. These data are pitch level data (whether a pitch is a ball, a strike, a swinging strike, or if the batter contact on the pitch was good quality or bad quality). Using any methods you would like, please rank the 5 pitchers from best to worst. Please present your results in a 2-page write up (including any graphs or tables). Please also include one data visualization and one table that helps supplement or emphasize your analysis. And please also include the code you used to create your analysis, commented throughout.

You will be evaluated on your methodology for ranking pitchers as well as your ability to communicate your findings effectively and succinctly.

Note: the `UID` column is also sequential, where the minimum `UID` for a given pitcher is their first pitch of the season and the maximum `UID` for a given pitcher is their last pitch of the season. For a given pitcher, if you sort by `UID` that is the order of pitches thrown throughout the season.

Data dictionary

`UID`: Unique row-level, pitch-level ID for the dataset. This ID is also sequential in nature, where the minimum `UID` for a given pitcher is their first pitch of the season and the maximum `UID` for a given pitcher is their last pitch of the season. For a given pitcher, if you sort by `UID` that is the order of pitches thrown throughout the season. For the `train.csv` dataset, due to the random sampling of pitches, not all pitchers will have a complete season represented.

`PITCHER_KEY`: Unique pitcher ID

`RELEASE_SPEED`: Speed of the pitch, measured in MPH

`SPIN_RATE_ABSOLUTE`: Spin of the pitch in flight, measured in revolutions per minute (RPM)

`RELEASE_HEIGHT`: The location of a pitch release, where 0 is the ground and 10 would be 10 feet above ground

`RELEASE_SIDE`: The location of a pitch release, where 0 is center of the rubber/center of home plate, -5 is 5 feet towards the 3rd base side, and 5 would be 5 feet towards the 1st base side

`PITCH_TYPE_TRACKED_KEY`: (only in `train.csv`) The type of pitch thrown as determined by our high tech scout

`B`: (only in `test.csv`) If the pitch resulted in a ball

`S`: (only in `test.csv`) If the pitch resulted in a strike

`SS`: (only in `test.csv`) If the pitch resulted in a swinging strike (will also be registered as `S`)

`CONTACT_QUALITY`: (only in `test.csv`) If the pitch resulted in good contact quality for the batter (positive run value; e.g. hit events, home runs, scoring plays, etc), bad contact quality for the batter (negative run value; e.g. outs, fouls, etc), or if no contact was made at all (represented by `NULL`).