Predictive Modeling with Sports Data

Homework 6

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem. Do not share any code.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission.

Make sure your answers to each problem are clearly stated in the submitted PDF. Do not include code in your submitted PDF. Any important figures, results, plots, and tables generated by your code should be extracted and inserted into the PDF in a visually appealing way. Think of the PDF as a presentation you are making based on the results you uncovered in your analysis. Your submitted PDF should be self-contained: the graders should not have to look through your code to find the solution.

Any code (Jupyter Notebooks and other source files) used to compute your answers should also be uploaded to Gradescope. If your code needs any special configuration in order to run, please include a `readme`. If the problem requires you to train and test a model, the training, validation, and testing code should all be submitted. If necessary, please indicate in your `readme` file if running your code requires special hardware (like a GPU, or 64GB+ of ram), a compilation step (if you decide to use Cython), or a signficant amount of time (longer than 5 minutes).

All of the questions in this homework will use the data in `pitchers17.csv`. You may not use any data other than what is given. Below we explain what each column denotes:

- `y` : Season
- `bf` : Number of batters faced (plate appearances) that season
- `bb` : Number of walks that season
- `k` : Number of strikeouts that season
- `hr` : Number of homeruns that season
- `h` : Number of hits that season

- `lo` : Number of lineouts that season (balls hit hard but relatively low, and caught in the air by the defense)
- `po` : Number of popouts that season (balls hit very high, but not far, and caught by the defense)
- `fo` : Number of flyouts that season (balls hit high and far, but caught by the defense)
- `go` : Number of groundouts that season (balls hit on the ground that are fielded by the defense, and lead to an out)

1. Read the McCracken article:

   https://www.baseballthinkfactory.org/fraser/articles/dips.htm

2. (Pitching Statistics) In this question we will become familiar with the data by analyzing some standard pitching statistics.

   (a) Estimate each of the following probabilities. Each is a single value computed using all of the seasons in the dataset.

      i. Probability of a plate appearance ending with a walk.
      ii. Probability of a plate appearance ending with a strikeout.
      iii. Conditional probability of a plate appearance ending with a strikeout given a walk did not occur.
      iv. Average number of home runs per plate appearance.
      v. Conditional probability of a plate appearance ending with a home run given neither a walk nor a strikeout occurred.
      vi. Conditional probability of a plate appearance ending with a non-HR hit given that the plate appearance didn't end with a walk, strikeout, or homerun.

   (b) Compute for each season and pitcher, the average number of strikeouts per plate appearance that did not end in a walk (the second McCracken component). Display the top 10 players in seasons 2016 and 2017 according to this statistic (two tables, one for each season). Include pitcher, `krate` (your statistic), and the number of batters faced. Restrict your analysis to pitchers with at least 500 batters faced, and sort your table in decreasing order by `krate`.

   (c) Repeat the previous problem with `hrate` (in place of `krate`), the fourth McCracken component. Recall that `hrate` is defined as the average number of non-HR hits that did not end in a walk, strikeout, or home run.

3. (Predicting McCracken Components) In this problem we forecast the McCracken components for a given season using data from the preceding season. Restrict to rows with at least 200 batters faced.

(a) Fit the following linear regression model:

$$\texttt{bbrate} \sim \texttt{bbrate\_prev}.$$

This is a single regression fit on the entire dataset. The response `bbrate` (average walks per plate appearance, the first McCracken component) varies over pitchers and seasons (except 2012), and the covariate `bbrate_prev` is average walks per plate appearance for the same pitcher over the preceding season. A row should be excluded if the pitcher had strictly fewer than 200 batters faced in either the current season, or the preceding season.

i. Report the two coefficients from your model.

ii. By using additional features from the preceding season, try to improve your fit of `bbrate`.

A. List some of the features you tried, whether they had significant coefficients, and what the value of the corresponding coefficient was.

B. For each feature listed above, give a brief explanation of your findings.

Your features can be other McCracken components, as well as other features you create using the columns in the given data.

(b) Repeat the previous part using `krate`, the second McCracken component (recall `krate` measures the average number of strikeouts per plate appearance that didn't end in a walk).

(c) Repeat the previous part using `hrrate`, the third McCracken component (recall `hrrate` measures the average number of home runs per plate appearance that didn't end in a walk or strikeout).

(d) Repeat the previous part using `hrate`, the fourth McCracken component (recall `hrate` measures the average number of non-HR hits per plate appearance that didn't end in a walk, strikeout, or home run).