Predictive Modeling with Sports Data

Homework 7 (Mini-Project)

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 `sc19.parquet` and `MLB_players19.csv`. Throughout, restrict to data from the regular season (where `game_type` is '`R`'). The column meanings of the data in the parquet file are explained at https://baseballsavant.mlb.com/csv-docs.

1. (Expanded Baseline Model) Using the data in the 2012-2018 seasons, fit baseline models for the two McCracken components `bbrate` and `krate` (as defined in the previous homework). We are only looking at two components to simplify our task.

   (a) Your model for `bbrate` should have the following form:

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

where `bbrate_prev` is computed using the year before `bbrate`. Restrict your fit to examples where the pitcher had at least 200 batters faced in the current year, and at least 200 batters faced in the previous year.

i. Report the coefficients of your fit.

ii. Use your model to make out-of-sample predictions for `bbrate` on 2019, and report the average square error. That is,

$$\frac{1}{P}\sum_{i=1}^{P}(\texttt{bbrate}_i - \widehat{\texttt{bbrate}}_i)^2,$$

where $i = 1, \ldots, P$ ranges over all players that pitched in 2019, $\texttt{bbrate}_i$ is the true value for pitcher i 2019, and $\widehat{\texttt{bbrate}}_i$ is the predicted value pitcher i in 2019. Restrict the test set to pitchers that faced at least 200 batters in each of 2019 and 2018 (yes, this introduces bias, but it also simplifies the assignment).

(b) Repeat the previous part for `krate`.

2. (Called Strike Model) In this problem we will compute the probability that a pitch is called as a ball or strike by the umpire.

Using the data in the 2012-2017 seasons, fit a model on pitches that to predict the probability that a pitch is called a strike by the umpire (`description` is `called_strike`). Restrict only to pitches where the batter did not swing. More precisely, restrict to pitches where `description` is `ball`, `blocked_ball`, or `called_strike`. You may use all of the features allowed for swinging strikes. Since the top and bottom of the strike zone depend on the height of the batter, the `sz_top` and `sz_bot` features may be useful.

Make sure to give separate answers to each of the following parts.

(a) State the type of model you used.

(b) List each of the features used in your model. For each feature, give a brief but clear explanation of what the feature is. If the feature is directly taken from the data with no transformation, you can say that.

(c) Explain your process for iterating on the model. What features were most significant? What ideas did you try?

(d) State your out-of-sample Brier score on the 2018 data (again restricting to pitches where the batter did not swing).

3. ($\star$ Swinging Strike Model: Extra Credit) In this problem we will engineer a feature that measures one aspect of pitch quality.

Using the data in the 2012-2017 seasons, fit a model to predict the probability that a pitch results in a swinging strike (`description` is `swinging_strike` or `swinging_strike_blocked`). You may use the following features: `pitch_type`, `pitch_name`,

batter, release_pos_x, release_pos_z, zone, stand, p_throws, balls, strikes, pfx_x, pfx_z, plate_x, plate_z, vx0, vy0, sz_top, sz_bot, vz0, ax, ay, az, outs_when_up, on_1b, on_2b, on_3b, release_speed, release_spin_rate, and release_extension.

Make sure to give separate answers to each of the following parts.

(a) State the type of model you used.

(b) List each of the features used in your model. For each feature, give a brief but clear explanation of what the feature is. If the feature is directly taken from the data with no transformation, you can say that.

(c) Explain your process for iterating on the model. What features were most significant? What ideas did you try?

(d) State your out-of-sample Brier score on the 2018 data.

4. (Improving the Baseline) In this problem, we improve the models from part 1.

   (a) Using the data from seasons 2012-2018, fit an improved model for bbrate. In particular, consider using features based on the called strike model we built earlier (and the swinging strike model, if you built one).

   Make sure to give separate answers to each of the following parts.

   i. State the type of model you used.

   ii. List each of the features used in your model. For each feature, give a brief but clear explanation of what the feature is. If the feature is directly taken from the data with no transformation, you can say that.

   iii. Explain your process for iterating on the model. What features were most significant? What ideas did you try?

   iv. Use your model to make out-of-sample predictions for bbrate on 2019, and report the average square error. Restrict the test set to pitchers that faced at least 200 batters in each of 2019 and 2018

   (b) Repeat the previous part for krate.