Final Project Memo

Sunrise Gao

2022-10-02

R Markdown

I will be doing a baseball hitting(or on-base) probability prediction for my final project.

An overview of your dataset What does it include?
* It will include players main hitting statistics such as plate appearances, run batted-in, hits, and more.
Where and how will you be obtaining it? Include the link and source.
* I will be selecting most useful statistics which can indicate a player's ability to hit and select the 2021 and 2022 (as the 2022 regular season is about to end).
* source: https://www.baseball-reference.com/ https://baseballsavant.mlb.com/ http://fangraphs.com http://www.brooksbaseball.net/pfxVB/zoneTrack.php.

About how many observations? How many predictors?
* It will be about 100 players stats for 2 regular seasons, and to predict these 100 player's stat for 2023 season.

What types of variables will you be working with?
* It will be mostly numberic variables.

Is there any missing data? About how much? Do you have an idea for how to handle it?
* All players' stats are free and accessible online, so probably won't have miss data I think, at least for now.

An overview of your research question(s) What variable(s) are you interested in predicting? What question(s) are you interested in answering?
* Either player's hitting probability (the probability of player to hit the ball) or player's on-base probability (the probability of players to get on bat including hitting the ball, taking walks, or hit by pitch).

Name your response/outcome variable(s) and briefly describe it/them.
* hitting probability, it is the probability of a player to hit the ball in-play.

Will these questions be best answered with a classification or regression approach?
* classification.

Which predictors do you think will be especially useful?
* I haven't finalize all the variables to use for prediction, but they will mostly be players' offensive stats as they are the most useful stats that can indicate a player's hiting ability.

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.
* It is predictive as hitting probability is "uncertain", "non-linear" that could be impact by some other factors.

Your proposed project timeline When do you plan on having your data set loaded, beginning your exploratory data analysis, etc? Provide a general timeline for the rest of the quarter.
* Week 2: doing research for selecting the best statistics which can indicate player's hitting ability the best: Then start picking players (may just the top rank 100 players as they may have a more steady performance).
* Week 3: wrapping up dataset selection, checking selected dataset, and prepare to load them.
* Week 4: Start to load dataset.
* Week 5: Checking the dataset before start running the code and to get the predictions out.
* Week 6: May just repeat week 5 and checking the code, debugging, asking for help if needed etc.
* Week 7: Repeating week 6, and check with professor to make sure the project is in the right direction.
* Week 8: Repeating week 7. * Week 9: Finishing up the whole document make it easy to read and more organized.

Any questions or concerns Are there any problems or difficult aspects of the project you anticipate?
* Selecting players' stat, modeling, and definitely the coding and debugging. but I will see how it goes with the project goes on.

Any specific questions you have for me/the instructional team?
* So far so good, Thank you professor!.