WILDCARD WEIGHTING

PLACE NAMES OF GROUP MEMBERS HERE

Abstract. Place abstract here. The abstract summarizes in one paragraph the main question and conclusions draw from your investigation.

1. Research question and overview of the data

The quality of the research question(s) you are asking plays a big role in how good the entire project is. Make a clear case for why your question is interesting, well thought out, precisely formulated, and answerable, at least in principle, with adequate data and the techniques of machine learning.

Briefly review what is already known about your research questions and what techniques others have used to study these questions. The best written reports include references to prior work. Explain the data set, before analysis. Form a thoughtful hypothesis or hypotheses about the data. Answer the following questions and any others that may be relevant to your question and your data set:

- What weaknesses or problems does the data set have?
- Why is this a good choice of data set to answer your research questions (as opposed to other similar data sets)?
- What do you expect your analysis to reveal?
- What other interesting questions will analyzing this data answer?

Also reference articles and sources [**?**] that are relevant or that you used when learning and/or thinking about your project. You should also reference prior work that has considered similar questions.

2. Data Cleaning / Feature Engineering

Tell us what you did when you were cleaning your data and engineering features. Why did you make the choices that you did? What are the consequences of those choices?

(1) **FIP** stands for Fielding Independent Pitching, and is a metric designed to capture how well a pitcher prevents runs independently of the rest of the defense. It is a better metric of a pitcher's skill than Earned Run Average because it disregards events outside of the pitcher's control.

Date: 4 October 2024.

(2) **Bullpen** Because pitching strains a pitcher's arm quickly, one pitcher will start and throw about 5-7 innings before being switched out for a *relief pitcher*. The bank of pitchers ready to relieve are called the Bullpen.

(3) **OOA** stands for Outs Above Average and signifies how many more outs a player makes than an average fielder at the same position.

(4) **WOBA** stands for Weighted On-Base Average and signifies how often players make it on base, weighted by how likely they are to score based on the manner in which they made it on base.

We gathered the data we used in three main pulls. Our final dataset incorporates data from 1) The pitcher logs for each game for each pitcher during the regular season from 2014-2015, 2) the pitcher logs for each game for each pitcher during the post season from 2014-2015, and 3) team level statistics (such as outs above average or weighted on-base average) for each team for each year from 2014 - 2025.

Early on, we decided that we wanted our match-up model to include the following features for both teams: starting pitcher's FIP, starting pitcher's freshness, an aggregated value for the FIP and freshness of the pitchers in the bullpen, average OOA (as an indication of the efficacy of the team's defense), average WOBA (as an indication of the efficacy of the team's offense), and whether the team is a wildcard in the playoffs. We also decided to include a feature indicating which team played at home.

The pitcher's logs (collectively referred to as df) included the following pertinent features: Date, Team, a unique Pitcher ID, FIP, Innings played, and the result of the game. In order to assemble it into a useable format for our problem, we made a few tweaks and added some new columns. We first converted the Date column to pandas.DateTime. After sorting the df by Pitcher Id and Date, we grouped the df by Pitcher ID and shifted the Date column to create a Previous Game Date column, and subtracted the Previous Game Date column from the Date column to create a Freshness column for each pitcher in each game.

Additional changes to make the pitcher logs more useful included adding a Starter feature that one-hot-encodes whether the pitcher started that game, and a Game ID to uniquely identify each game, based on the date and the two teams playing.

The dataset of team level statistics by year included the following pertinent features: Team, Year, OOA (averaged over the season), WOBA (averaged over the season), and Is-Wildcard. This dataset was nearly immediately useable after cleaning the Team names to match those in the pitcher logs dataset. We merged it with the pitcher logs on Year and Team.

Equipped with a newly-merged datset (with features Date, Team, Pitcher ID, FIP, Innings played, Game Result, Freshness, Starter, Game ID, Team OOA, Team WOBA, Is-Wildcard for each pitcher in each game) we were ready to assemble a final games dataset. To do so, we grouped the newly-merged dataset by Game ID and then by team. For each team in each game, we assembled a dictionary containing the starter's FIP and freshness, the meaned FIP and freshness of the relief pitchers, whether the team was at home, team OOA, team FIP, whether it was a post-season game, and if so, whether the team was a wildcard during the game. After shuffling the order of the teams, we concatenated the two team's stats into a single row, with a results feature signifying whether the first team in the row won or lost. There were no ties in our dataset.

With our data in a tabular dataset with a binary label, we were ready to train!

3. Data Visualization and Basic Analysis

Analyze the data, draw conclusions, and effectively communicate your main observations and results.

- Calculate appropriate summary statistics.
- Use appropriate plotting techniques, visualizations, and other tools and techniques you have learned, to thoughtfully identify and evaluate what the data are telling you,how well suited the data are to answering your problem,
- Reference figures and plots, like Figure **??**.

4. Learning Algorithms and In-depth Analysis

Analyze the data using the machine learning techniques discussed in class. Explain what research questions you can answer using the machine learning techniques presented this semester, and if applicable, what you think you may be able to answer next semester.

Be able to explain the results of your analysis, whether the results are meaningful, and why you chose the tools that you used.

5. Ethical Implications and Conclusions

Thoughtfully analyze the ethical implications of your research questions, the data you gathered, and the analysis that was performed. Are there privacy or other implications from the collection or use of the data? Could your results and methods be misused or misunderstood? What can and should be done to prevent misuse and misunderstanding? Could your algorithms and methods result in a destructive self-fulfilling feedback loop? How could that be prevented or controlled? What other ethical implications does your work have?

This part should all be done before you get to *page 5.* The bibliography can spill on to page 6, but we won't read text that goes past page 5.