Capstone Project II Proposal

Ian Nielsen

Problem Statement: What possible improvements could this MLB organization make to optimize its 40-man roster (rotation, and lineup) in order to win the World Series by determining if the team's typical lineup is satisfactory enough to beat the opposing team in a seven game series before the series begins?

Context: This MLB team has made it to the World Series. They will compete against the opposing team in a seven game series. This wish to determine what metrics they could improve on their 40-man roster before the series begins. Once possible metrics to improve are determined the team will make subsequent changes to their 40-man roster.

Criteria for Success: Will your team win the world series? If not, did the proposed improvements allow you to win?

Scope of Solution Space: Data from regular season play as well as the playoffs will be used from both teams. This includes batting and pitching statistics, along with fielding statistics.

Constraints: No 1994 and 1904 data due to MLB lockout. Historical data goes back to 1871. Need to determine where to truncate data as league structure has changed since league conception. Historical data is available only until 2015.

Stakeholders: Head of Baseball Operations. General Manager. Head Manager. Club owners.

Data Sources: https://www.kaggle.com/datasets/open-source-sports/baseball-databank in form of csv files. Data will include a list of world series winners and regular season/playoff data for all teams in all applicable seasons.

Possible Method of Solving Problem:

- Regular season data is available as an aggregation for all teams for each season. Playoff data is only available at the batting and pitching level. Playoff data for all series for all applicable years will need to be joined and aggregated. For all past World Series and for each team data will need to be aggregated for each playoff run to summarize playoff data. A table of World Series winners will also be used, which includes opponents, year, etc.
- Notable features to be used from regular season and playoff data: Wins, Loses, Team Batting Avg, Team ERA, Runs, Hits, RBIs, 2Bs, HRs, Walks, SO's, OBP, SLG%.
- A record for each World Series will be created. This will include notable features for each team (shown above), American League first, National League second. A separate

column will be created which will signify who won the series (1 for AL, 0 for NL). This will be done by joining the WS winners/opponents data with the aggregated data from the regular season and playoffs.

- A classification algorithm will be used to predict the winner of the 2022 world series. A larger test set could be used by incorporating more seasons.