Lead Author: Andrew Willmore

Co-authors: David Tu, Albert Zhong, Kaveya Sivaprakasam

Title: **Projecting Major League Baseball Pitcher Success for the 2024 Season**

Research Question

Major League Baseball teams spend hundreds of millions of dollars each season to put the best team that they can on the field. The main question every team faces is 1) how to evaluate talent and 2) how to build a cost efficient roster. Every team in the league now employs an analytics/research & development department that utilize data to assist in making these kinds of decisions. In baseball, the major concerns when building a roster are figuring out how to prevent runs (pitching and defense) and how to score runs (hitting and baserunning). For this project, we will be focusing on projecting how pitchers will fare in the 2024 season. There is no single metric for evaluating how good a pitcher is. Popular methods include looking at the average number of runs a pitcher gives up to the opposing team per 9 innings – earned run average (ERA) or a calculated score of how many wins a player contributes above that of a fringe-caliber major league player – wins above replacement (WAR). We will be focusing on projecting a metric wOBA – weighted on-base average, which weights the outcomes of each at bat (for example giving up a double is given more weight than giving up a single). The lower a pitchers' wOBA, the better they typically are. A "successful" model would be able to outperform simply using the previous years' expected wOBA (xwOBA) to project wOBA for the current season.

Data/Materials

Publicly available pitcher data can be found on the baseball savant website: https://baseballsavant.mlb.com/leaderboard/custom?year=2024&type=pitcher&filter=&min=q&selections=pa%2Ck_percent%2Cbb_percent%2Cwoba%2Cxwoba%2Csweet_spot_percent%2Cbarrel_batted_rate%2Chard_hit_percent%2Cavg_best_speed%2Cavg_hyper_speed%2Cwhiff_percent%2Cswing_percent&chart=false&x=pa&y=pa&r=no&chartType=beeswarm&sort=xwoba&sortDir=asc

From this website you can access past years data and export CSV files containing metrics that might be relevant for projecting future wOBA. Baseball analytics have evolved over time and statcast data is available starting in 2015. Statcast metrics such as barrel% (percentage of batted balls with the perfect combination of exit velocity and launch angle) have become more popular in recent years and are thought to have higher predictive value than outcome data such as whether a player got a hit or not (much of which is out of a player's control once the ball is in play).

Course impact/relevance

Because the measure of success we are choosing to predict (wOBA) is continuous, this will be a regression problem focused on prediction. At the same time, the model predictions are meaningless unless you can convince higher ups that the model makes sense and can properly

identify a pitcher worth investing millions of dollars into. Inference is therefore also an important consideration. Other aspects from the class that the project will address are model selection and cross validation.

Outcomes

The expected outcome will be a projected value of 2024 wOBA for pitchers based on their previous season(s) data. The projected values will be selected from the modeling technique that provides the best combination of accuracy (based on results of a test dataset) and interpretability – to be able to explain which variables were important in the prediction to higher ups/final decision makers.

Ideally, the predictions will outperform a simple linear regression with predictor of the previous year's expected wOBA (xwOBA). If the xwOBA simple linear regression performs better, it wouldn't make sense to use a different model. One consideration to make when creating the training and test datasets will be which pitchers to include. Many pitchers will have only thrown a small number of innings in a season due either to their role with the team or only having pitched in a few games. A pitcher who has only thrown a small number of innings will likely not have the same predictive validity as one who throws more innings.