

Project Luther Proposal

How many strikeouts will a MLB pitcher record?

Jason Salazer-Adams

Scope

Consider an MLB pitcher. The pitcher has control of the ball and then releases the ball to the batter. The batter then has a chance to hit the ball. If the ball is hit in play then the rest of the team has a chance to make an out. If the pitcher strikes out the batter, then the pitcher is responsible for the out. The only out a MLB pitcher has complete control over in a game is a strikeout. If an opposing team can predict the number of strikeouts based on their lineup, then it can estimate how many outs they have to generate runs. I am going to attempt to predict the strikeouts given pitcher characteristics and team batting characteristics. My goal is to use the analysis to influence how a pitching coach or hitting coach can work with the pitcher or batter to either increase or reduce strikeouts, respectively.

A MVP would be to calculate an average number of strikeouts using historical boxscore data for 2018. The MVP could be improved by incorporating StatCast data or weather data.

Data

I am planning to scrape box score data for the 2018 season. Utilize the first half of the season (or some subset) as training data, and the second half as testing data.

Pitching Data

STAT	Description
Throws	Left (L) or Right (R)
BF	Batters Faced YTD (total or avg)
H	Hits allowed YTD (total or avg)
R	Runs allowed YTD (total or avg)
ER	Earned runs allowed YTD (total or avg)
HR	Home runs allowed YTD (total or avg)

Batting Data

STAT	Description
Bats	Left (L) or Right (R)
AB	At bats YTD (total or avg)
H	Hits YTD (total or avg)

STAT	Description
R	Runs YTD (total or avg)
RBI	Runs batted in YTD (total or avg)
HR	Home runs YTD (total or avg)
BB	Walks YTD (total or avg)
PA	Plate appearances (total or avg)
SO	Strikeouts (total or avg)
Starting Pitcher Throws	L or R of Starting pitcher*
Pit	Pitches seen YTD (total or avg)

Known Unknowns

- Is 2018 enough data?
- Impact of weather or time-off between appearances.
- Impact of injuries and variance of starting lineups.
- "Starting Pitcher Throws" is more categorical than numerical.