Ricardo Trujillo

Project 1 Proposal:

The topic of choice for this project will be on pitcher performance across all teams using only pitcher stats in the Major League Baseball (MLB) division. I will be using the data across the given months of active baseball seasons as noted in the given RStudio code for the years 2018 and 2019. I have accessed this data using the "baseballr" package that includes functions for web-scraping data from different websites, in particular, Baseball-Reference.com was the main source of data I was looking at using.

The latest version of the package was released January 7th, 2020, so the information was updated recently. For more information about the package you can reference the following link: http://billpetti.github.io/baseballr/about/

Across both years there are 46 available predictors (some qualitative and some quantitative). For the 2018 season data, there are a total of 799 observations and for the 2019 season there are a total of 831 observations.

Overall, I will be constructing a regression model that focuses on the ERA, or Earned Run Average, as the response variable. The goal is to see which statistical variables lead to a lower ERA score for a pitcher. Looking a little into the glossary terms from the dataset I can point out a few predictors that I will hypothesize would be most statistically important. Those being: G (Games Pitched or played), IP(Innings Pitched), H (Hits/Hits allowed) , R (runs/ scores allowed) , ER(Earned Runs Allowed) , HR (Home Run Hints Allowed), SO (strikeouts), SHO (shutouts), and a few more.

Given that my data is all pitcher performance with 46 variables, I believe my approach for the project will be more data-centric. I would like to approach the project by doing forward and backward selection to see which variables are statistically significant overall to the response in a structured selection procedure. I also would like to see how certain regression models will work in comparison with each other so a few I had in mind were to compare Lasso Regression, Ridge Regression, possibly PCA, and K-Nearest Neighbors. In addition, if time permits, I would like to take a look at k-fold cross validation and/or bootstrapping in attempt to aid the overfitting issue.

On another note, as I was exploring the data, I did notice some potential obstacles such as some columns being fully empty/ N/A as well as some random N/A values for certain pitchers on columns that were mostly filled. I will have to look into it but I just assume that for that season the pitcher had little to no games to base the statistic off of. Another thing I ran into trouble with that may be a potential issue is that on baseball reference there is an asterisks next to a player to denote if they are left handed or right handed. I thought this would be nice to include as well as it

would be interesting and because it can be quantitative in the sense the 1 can be for right handed and 0 for left handed -- similar to a previous example in the book where it was just two options in a qualitative predictor.

I chose this topic because personally I like the Dodgers and wanted to focus on the Dodgers solely but it was limited in data so I expanded it to all pitchers. I personally was interested because I never knew the extent of what this topic can be applied to until I found a job title of Qualitative Analyst for the Dodgers. As I looked more into it, it intrigued me the extent of industries/types of data you can access and evaluate it to make predictions or evaluations. The success of a baseball game and the outcome of a win is not solely dependent on the amount of runs a team has to the amount of wins they have because all those numbers root from how well or how bad a pitcher is. Assessing a pitcher's performance will be able to see where a given flaw hinders the team or player's overall success, so when applying these regression models it will allow me to personally see and statistically prove what aspects of a pitchers tends to lead them to a lower ERA score -- and that will be the ultimate question.

Overall:

- I want to see what statistical variables have a positive or negative relationship with a pitchers ERA across all MLB teams
- I would like to do so with the best, most efficient of my data using a training, validation, and testing data split
- I would like to look at will my model be able accurately define this relationship and fit well to data-split sets through various metrics
- It will also be a change to compare and contrasts the strengths of each method and ultimately pick the best approach
- Possibly predict and compare to the 2020 season??

These are a glossary index for the variable terms as reference for outside the ones mentioned:

L -- Losses

W-L% -- **Win-Loss Percentage**
W / (W + L)
For players, leaders need one decision for every ten team games.
For managers, minimum to qualify for leading is 320 games.

ERA -- **9 * ER / IP**
For recent years, leaders need 1 IP
per team game played.
Bold indicates lowest ERA using current stats
Gold means awarded ERA title at end of year.

G -- Games Played or Pitched

GS -- Games Started

GF -- Games Finished

CG -- Complete Game

SHO -- Shutouts
No runs allowed and a complete game.

SV -- Saves

SV -- Saves

IP -- Innings Pitched

H -- Hits/Hits Allowed

R -- Runs Scored/Allowed

ER -- Earned Runs Allowed

HR -- Home Runs Hit/Allowed

BB -- Bases on Balls/Walks

IBB -- **Intentional Bases on Balls**
First tracked in 1955.

SO -- Strikeouts

HBP -- Times Hit by a Pitch.

BK -- Balks

WP -- Wild Pitches

BF -- Batters Faced

- **ERA+** -- **ERA+**
  100*[lgERA/ERA]
  Adjusted to the player's ballpark(s).

FIP -- **Fielding Independent Pitching**
this stat measures a pitcher's effectiveness at preventing HR, BB, HBP and causing SO
$(13*HR + 3*(BB+HBP) - 2*SO)/IP + Constant_{lg}$
The constant is set so that each season MLB average FIP is the same as the MLB avg ERA

WHIP -- **(BB + H)/IP**
For recent years, leaders need 1 IP
per team game played

H9 -- **9 x H / IP**
For recent years, leaders need 1 IP
per team game played

HR9 -- **9 x HR / IP**
For recent years, leaders need 1 IP
per team game played

BB9 -- **9 x BB / IP**
For recent years, leaders need 1 IP
per team game played

SO9 -- **9 x SO / IP**
For recent years, leaders need 1 IP
per team game played

SO/W -- **SO/W or SO/BB**
For recent years, pitching leaders need 1 IP
per team game played.
No batting leaders computed.

Rk -- **Rank**
This is a count of the rows from top to bottom.
It is recalculated following the sorting of a column.

Name -- **Player Name**
Bold can mean player is active for this team
or player has appeared in MLB
* means LHP or LHB,
# means switch hitter,
+ can mean HOFer.

Age -- Player's age at midnight of June 30th of that year

Lg -- **League**
AL - American League (1901-present)
NL - National League (1876-present)
AA - American Association (1882-1891)
UA - Union Association (1884)
PL - Players League (1890)
FL - Federal League (1914-1915)
NA - National Association (1871-1875)

W -- Wins