

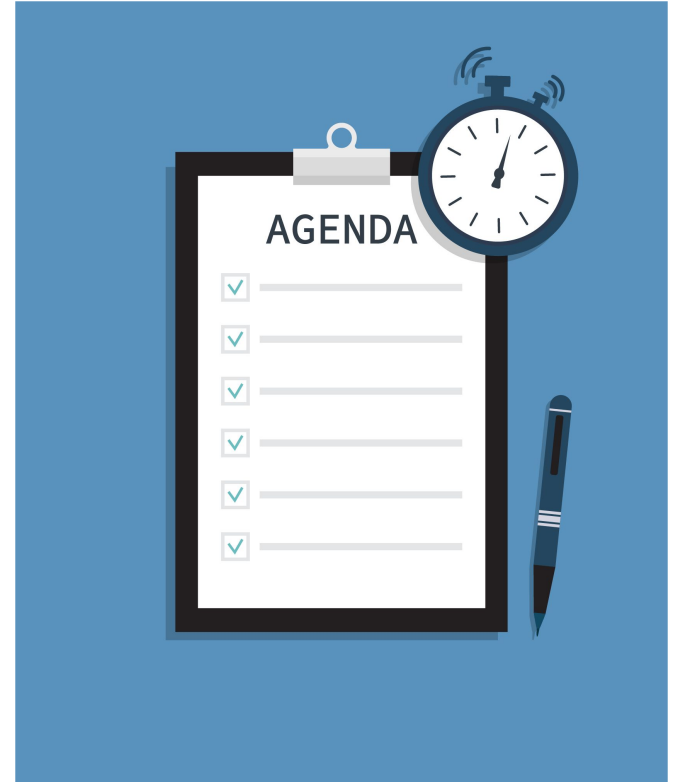
Baseball Modeling And Forecasting (GA)



Data Scientist - Connor Phillips

Agenda

- ❏ Problem Statement
- ❏ Data Collection
- ❏ Modeling
- ❏ Forecasting
- ❏ Demo!
- ❏ Application/Limitation
- ❏ Future Plans



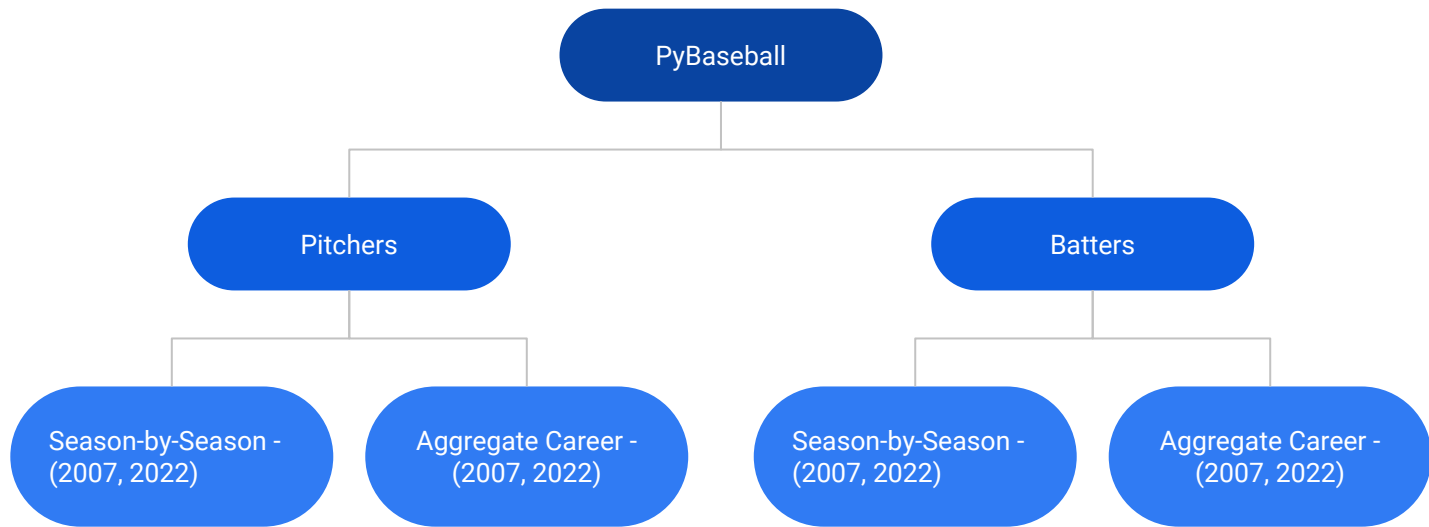
Problem Statement

This project was driven by a focus to create an interactive user experience with Baseball Player metrics. Using pybaseball, I obtained aggregate career and season-by-season data for baseball players since 2007.

The application offers two main features, the ability to make predictions and forecast specifically for individual players based on variables selected by the user.

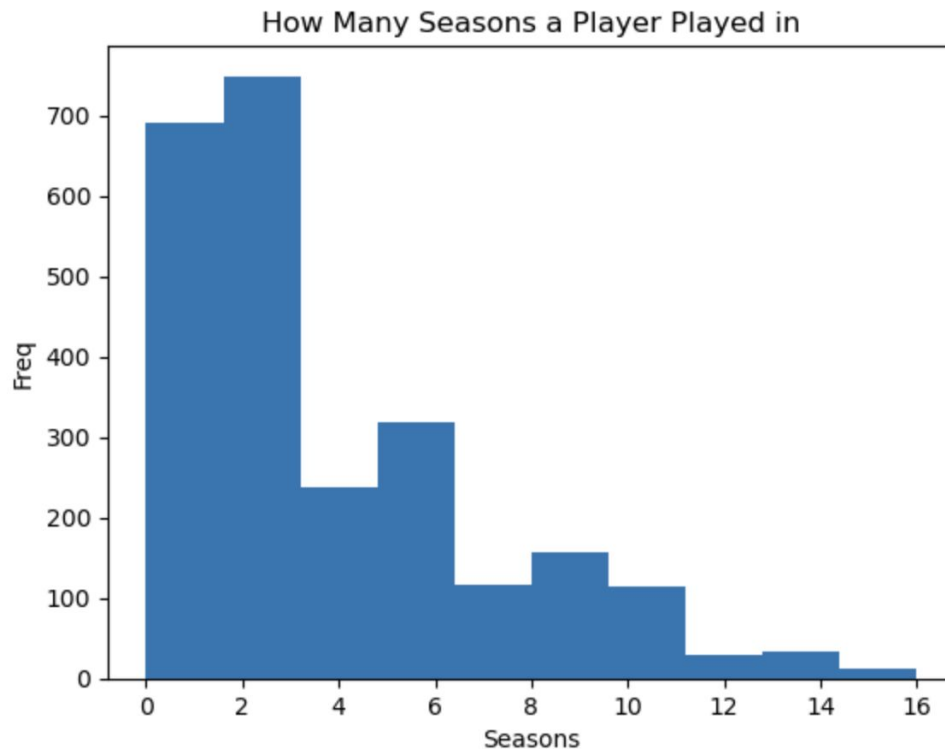
For the project to be a success, the modeling and forecasting has to be deeply customizable for the user, while still being informative. The user will be able to experiment with an array of different features, and learn about how baseball player statistics influence one another.

Data Collection



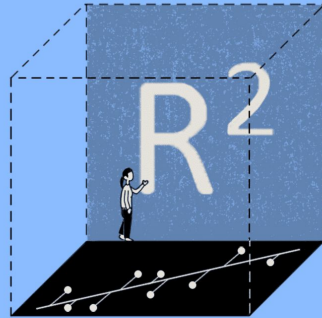
EDA - Seasons Played

❏ Few can be forecasted



Modeling

- ❑ First, user selects Batter or Hitter
- ❑ Linear Regression model used
 - ❑ User can select up to ten different features
 - ❑ One feature to predict
- ❑ The model will inform the user of the mean squared error, as well as R^2 value that the model obtained




The illustration shows a 3D wireframe cube on a blue background. Inside the cube, a person is standing next to a large, white R^2 symbol. The base of the cube is a black plane with a white line graph showing several data points and a fitted regression line. The text 'R-Squared' is written in a large, bold, black font to the right of the cube. Below it, the phonetic transcription ['är 'skwerd] is shown in a smaller font. Further down, a paragraph explains that R-Squared is a statistical measure representing the proportion of variance explained by a regression model.

R-Squared

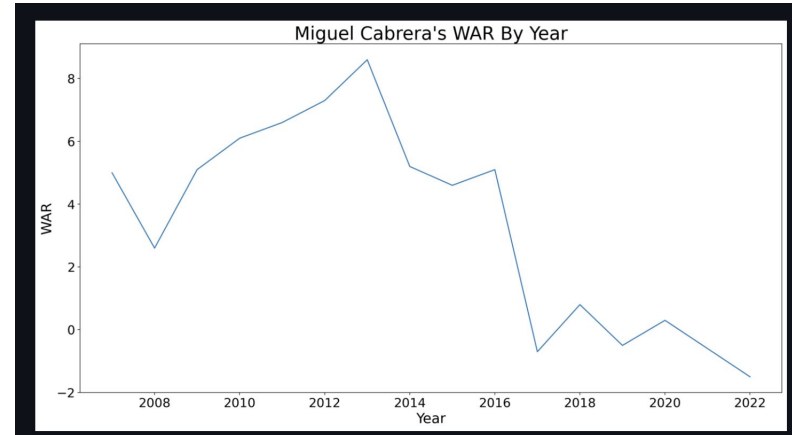
['är 'skwerd]

A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

 Investopedia

Forecasting

- ❑ App provides user choice of batter or pitcher
 - ❑ All players from 2007 that have player 10 seasons or more
 - ❑ And also played in 2022
- ❑ App allows for choice of metrics to use for VAR modeling
 - ❑ Shows time series plots for all metrics chosen
- ❑ Provides forecasted metrics for chosen player



Demo

Click [here](#) for a link to the product!

Application of Data

- ❑ Allow for MLB teams to discover which metrics are better indicators of others
 - ❑ Example. Average and BABIP being strong indicators of Hits
- ❑ Forecast totals for an aging player, try to predict if they can continue production

Limitations -

- ❑ Players who experienced a long term injury - Forecasts suffer
- ❑ Small amount of data in general for forecasts

Future Plans

- ❑ Increase the amount of available metrics for hitters and batters
- ❑ Try a monthly forecast of baseball players
 - ❑ Would allow for a prediction of their play as weather changes
- ❑ Expand data by position group
 - ❑ Ex. Starting Pitchers, Middle Relievers, Closers etc.

References

Data Collected from - <https://github.com/jldbc/pybaseball>

Images used -

https://assets-global.website-files.com/62196607bf1b46c300301846/62879c706fca5bc1fde3e0020_team%20meeting%20agenda%20%5BConverted%5D%201.png

[https://www.investopedia.com/thmb/lvCmuo4ls4GX_i7KWlc0S7G5iT8=/1500x0/filters:no_upscale\(\):max_bytes\(150000\):strip_icc\(\)/R-Squared-final-cc82c183ea7743538fdeed1986bd00c3.png](https://www.investopedia.com/thmb/lvCmuo4ls4GX_i7KWlc0S7G5iT8=/1500x0/filters:no_upscale():max_bytes(150000):strip_icc()/R-Squared-final-cc82c183ea7743538fdeed1986bd00c3.png)