Connor Phillips

Data Scientist

May 15th, 2023

General Assembly

Technical Summary

In this document, I would like to highlight my exact process for creating the Baseball Modeling and Forecasting application, in a more technical way. The first step was to identify a source of data, which ended up being Pybaseball. Pybaseball was able to provide me with career aggregate data as well as seasonal data for every player in a range of dates, which I selected to be from 2007 to 2022.

Career aggregate data was collected from Pybaseball, and then a feature named ‘Seasons’ was added to the data frame, which was simply the amount of seasons that player played in, from counting the season-by-season data in a for loop. Once data was split between pitchers and batters, I was then interested in creating dataframes of only batters and pitchers who played at least ten seasons. In order to make my app relevant, I made sure to only include players who played ten or more seasons and also played in 2022; this was done in order to ensure I could predict for 2023.

Following the finalization of the data, it was time to create an application for the user to interact and model with. To do this, I utilized streamlit. In my streamlit file, I first asked the user for a decision between a batter or a pitcher, which would then separate the data for the rest of the file. Modeling was the first course of action, and in order to do so, the user is prompted through

streamlit to provide as many variables as they would like, in addition to a response variable. The application takes this information into a list, and stores it to be used in linear regression.

For linear regression, I utilized a pipeline that scaled the data. Mean squared error and R squared are provided to the user as a quick model evaluation. The user is able to make adjustments to metrics used for X, or to the response variable Y, and the app will immediately make a new model, with new metrics to evaluate it.

Once the user is satisfied with modeling/predicting, they can move on to specific forecasting. The application asks for input of a specific player, which is now preset to their earlier choice of a batter or pitcher, and three variables from a list to model with in time series analysis. Given that information, the application creates three time series plots for the user to get an idea of how that player has changed overtime. Next, a while loop is used to get all data stationary before VAR modeling begins. Once stationary, a time series model is fit, and forecasts are generated. Finally, the user is provided with specific forecasts for the 2023 season, in the three respective types of baseball statistics chosen.