Authors:

Aaron Cooper: aaroncooper2022@u.northwestern.edu

Isaiah Jones: isaiah.jones@u.northwestern.edu

Derek Wen: derekwen2022@u.northwestern.edu

CS396: Introduction to Data Science

Professor Huiling Hu

Introduction and Motivation

We wanted to examine how the introduction of statistics has impacted the game of baseball. More specifically, we wanted to focus on managers' decisions regarding which pitchers to start as well as pitchers' decisions when deciding what type of ball to pitch. These two questions can help us better understand how baseball has changed over time as more data was collected and subsequently introduced into the decision making processes.

Using the data we've collected, we wanted to predict which individual pitchers are likely to become All-Stars[1] in a given season. This will help both us and team managers decide which pitchers to keep on the team as team construction is an integral part of the sport.
Managers must make key decisions over the course of a baseball game and season regarding which pitchers to use.

Dataset

The data we are working with was found via a publicly available database which contains single season pitching data for individual baseball pitchers. This data includes virtually all relevant pitching statistics although there will be missing components as certain statistical recording methods have evolved.

This dataset has 48,399 entries and each entry corresponds to one season worth of data for an individual pitcher. The dataset contains the following statistics per entry:
playerID, yearID, stint, teamID, lgID, W, L, G, GS, CG, SHO, SV, IPouts, H, ER, HR, BB, SO, BAOpp, ERA, IBB, WP, HBP, BK, BFP, GF, R, SH, SF, GIDP

We do not plan to use all of them and instead will focus on the more standardized metrics.

Data Cleaning

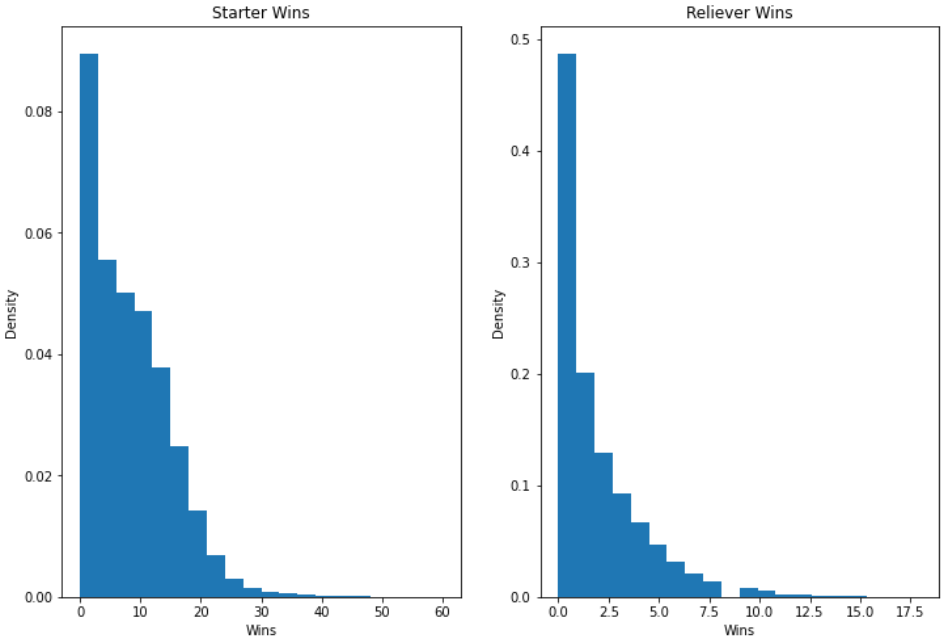
To begin our data cleaning process, we first needed to fill in all null values in the dataset. After filling in the missing values, we accounted for the fact that there were certain baseball seasons which were shorter than others, and removed the shortened seasons. This is because we did not want the shortened seasons to bias the rest of the data as each game would have been weighted differently.

We later noticed that there are a lot of baseball players who only play in a few games each season, and this represents another possible bias due to the small game sample size. We filtered out all players who played in what we deemed an insufficient amount of games (under 7 games per season).

After doing some more research, we realized that the number of teams have changed over the years, and decided to split our data into two categories: modern and classic eras of baseball. The modern era was defined as all seasons after 1982, which is when the current number of teams first was established. We used this classification for some of the EDA.

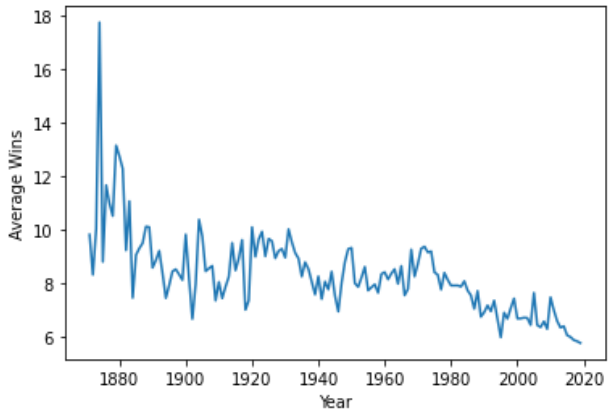
EDA

We decided to look into the difference in number of wins between starters (pitchers who start the game in the first inning) and relievers (pitchers who substitute in for the starter).



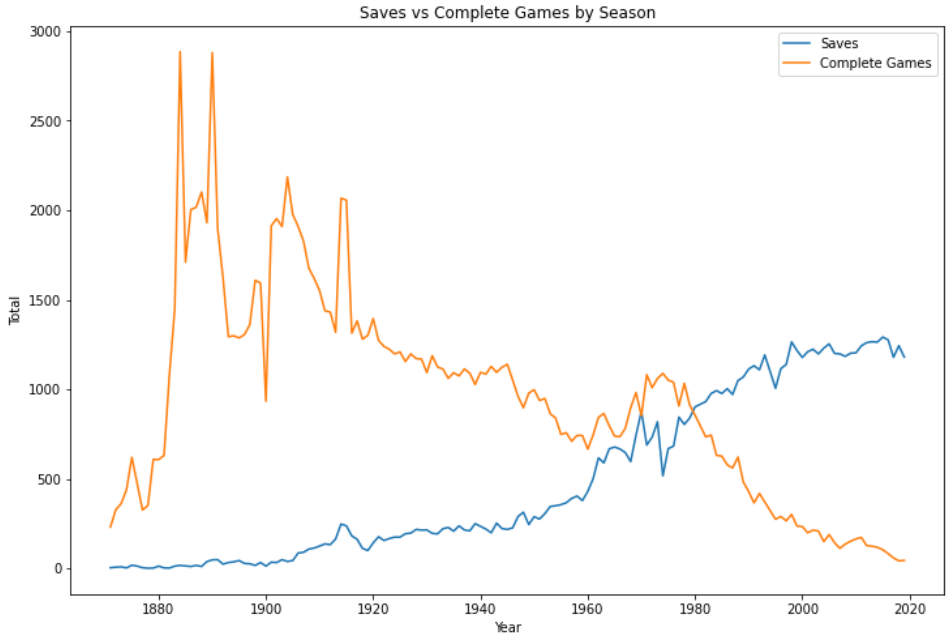
What this shows us is that the number of wins is only really applicable to starters, due to the above criteria to be awarded a ‘win’ in a game. Relievers on the other hand have their own metric called ‘saves’.

We also wanted to examine how the introduction of the relievers classification would impact the total number of wins in a given season:



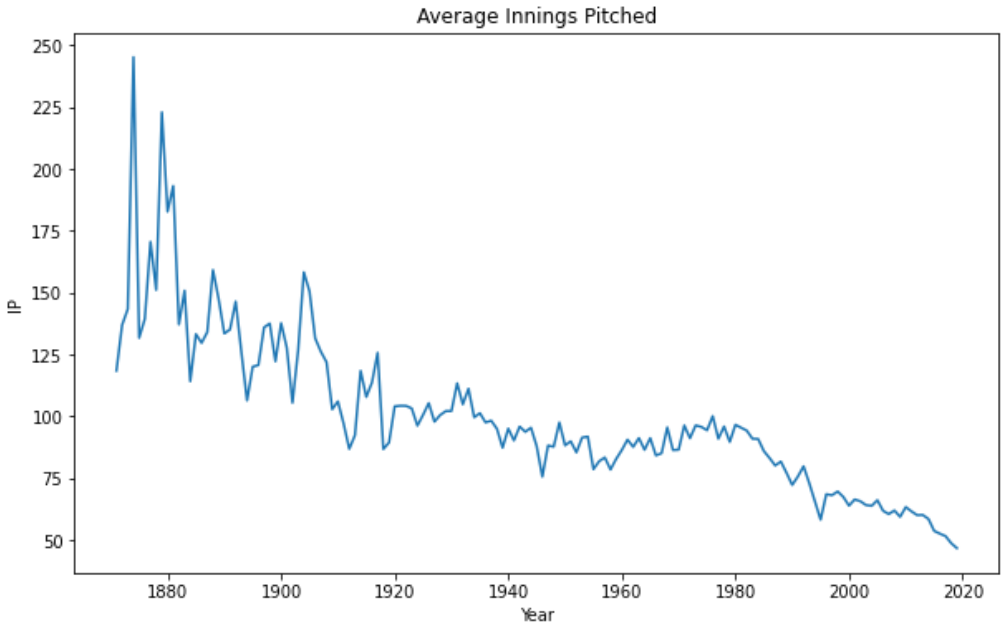
This chart shows us that with the introduction of more relievers into the game, the average number of wins has decreased. Managers are pulling out their starters in earlier innings and utilizing different relievers, which is a large factor in this decrease.

We wanted to further examine the impact that relievers had in recent years, so we decided to compare the number of saves against the number of complete games (where a pitcher pitches the entire game).



As we can see from the chart, as the number of saves, and subsequently relievers increases, there is a related decrease in the number of complete games. This shows how managers are shifting away from the one pitcher one game strategy and instead are more reliant upon a combination of multiple pitchers for a single game.

The last statistic we wanted to examine is how the average number of innings a pitcher pitches has changed over time.



There has been a clear downward trend in the average number of innings played by a pitcher. We can attribute this trend to the introduction of relievers, more pitchers, and manager decisions.

Data Modeling

First, before any models were the feature engineering steps. All features had N/A values filled in for 0 and infinity values filled with a high integer value. Then each feature was scaled and all categorical features were dummy coded to binary values. Then the data was split into training and testing data sets. Important in the split was the stratification option. Because the target feature is very imbalanced, it's important to have a good amount of positive occurrences in both training and testing sets. So stratification ensures that there's an even distribution of the classes of positive and negative in both training and testing sets.

We decided to test this over 4 different models and compare the scores to see which one would perform best. The models we chose were Logistic Regression, Random Forest, Boosted Tree and K-Nearest Neighbors. We chose these models for their aptitude at solving two class classification problems. Each model had a sample size as mentioned above of 48,399 and each used the scoring metric of balanced_accuracy. At first we used accuracy but realized that it was unsuited for a problem with a really unbalanced target feature. Balanced accuracy takes the average of recall on both classes of the target feature so was much better suited for our use case and doesn't check one class like recall.

For each model except KNN, we also looked at the feature importance of each model to conduct our analysis. This is essential for our project because we really want to see what statistics are being evaluated as important by these models to have insight on how pitchers perform data. Each model's parameters were tuned using Grid Search.

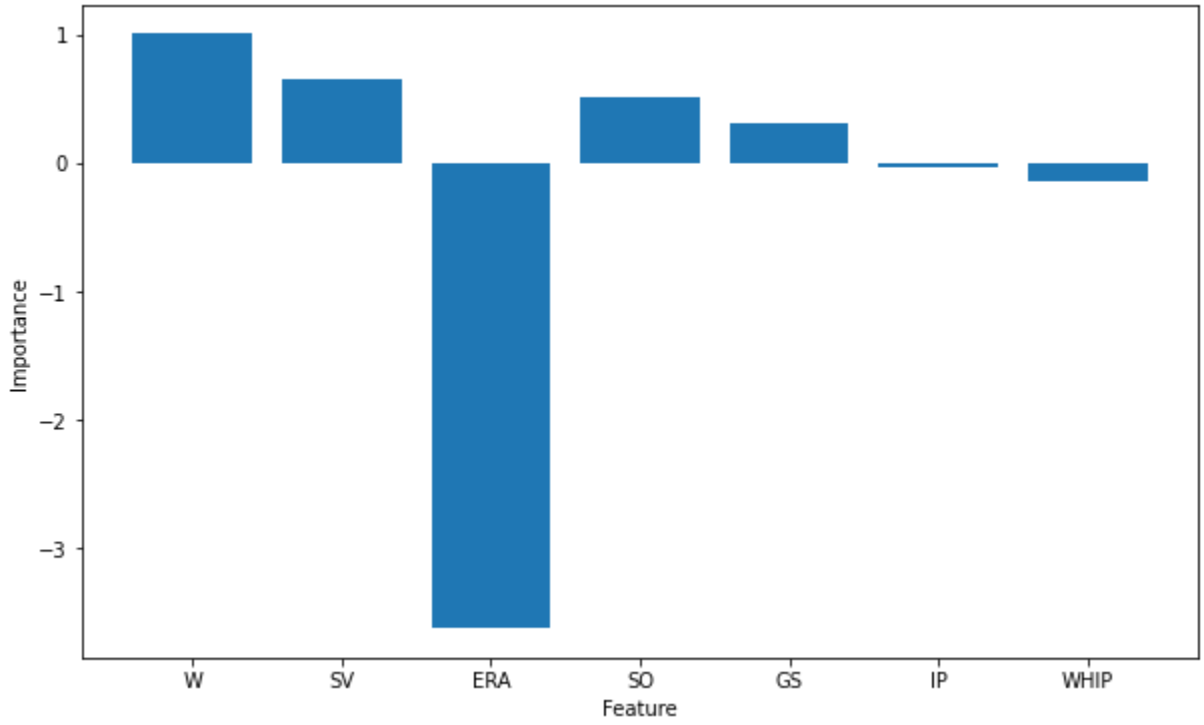
Model 1: Logistic Regression

Parameters (Grid Search): `{'l1_ratio': 0.5, 'penalty': 'none'}`

Performance on test set: `0.6808125696013754`

Feature Importance: Note that in this graph specifically, negative values indicate importance towards predicting the negative category and positive values indicate predicting the positive

category, “Yes”.

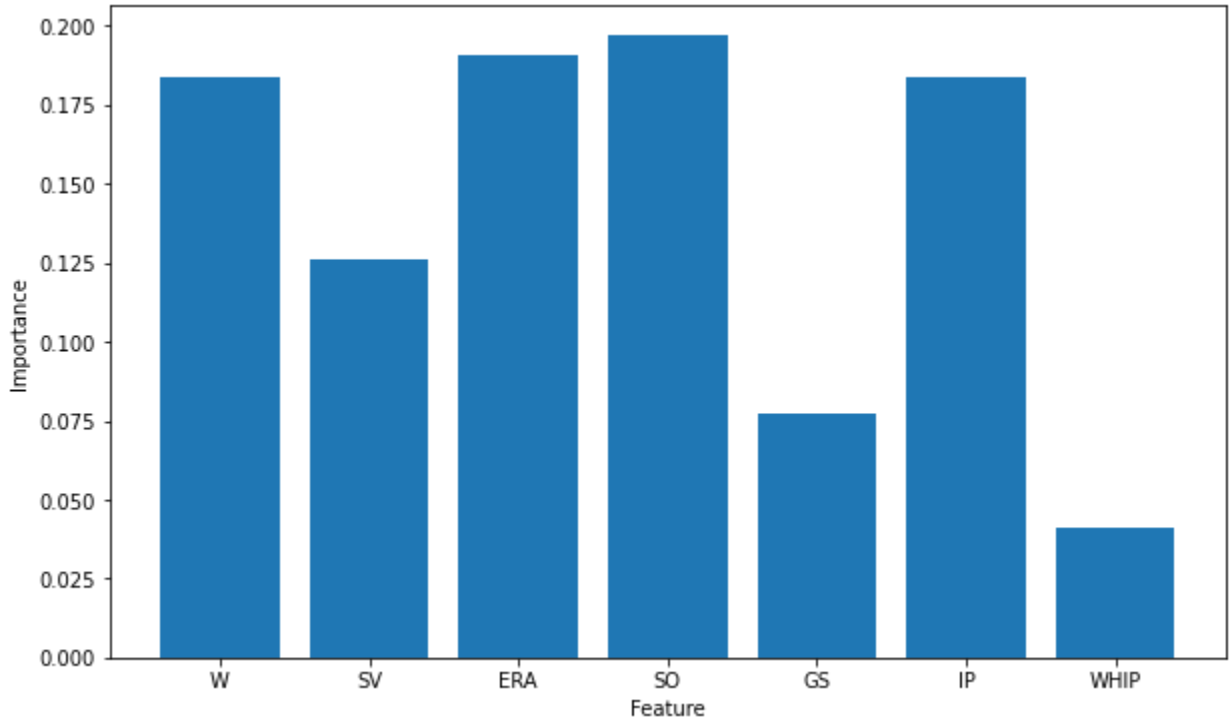


Model 2: Random Forest

Parameters (Grid Search): `{'min_samples_split': 6, 'n_estimators': 50}`

Performance on test set: `0.6817759991664116`

Feature Importance:



Model 3: Boosted Tree

Parameters (Grid Search): `{'learning_rate': 0.4,`

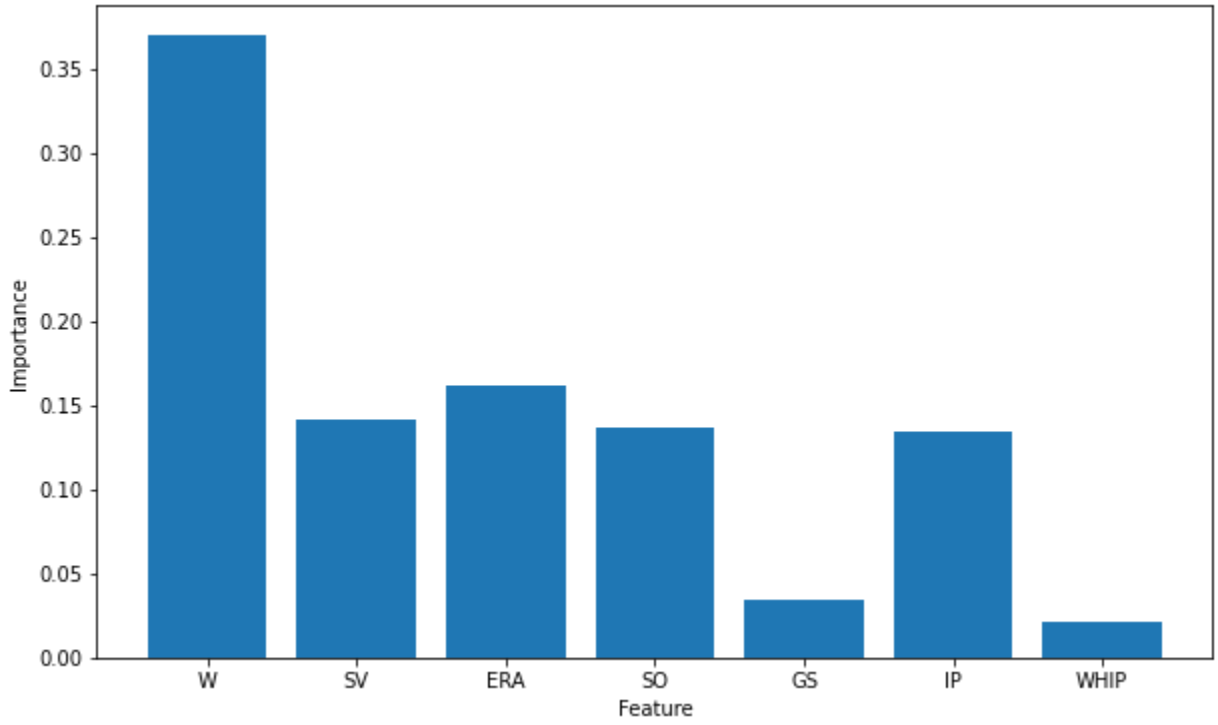
`'max_depth': 5,`

`'min_samples_split': 2,`

`'n_estimators': 50}`

Performance on test set: `0.6886600229236811`

Feature Importance:



Model 4: KNN

Parameters (Grid Search): `{'n_neighbors': 4, 'weights': 'distance'}`

Performance on test set: `0.6913569744648428`

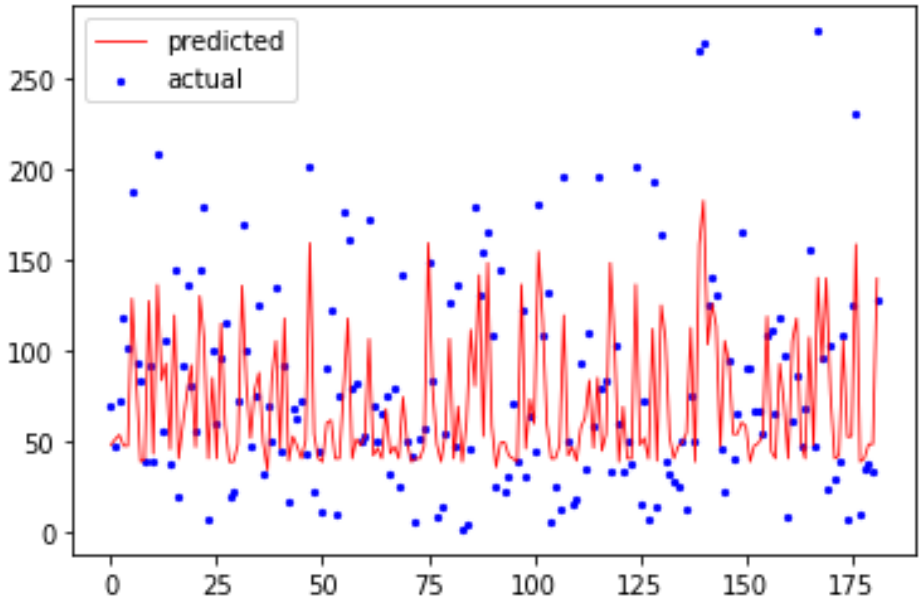
We settled on these features selected after analyzing feature importance of more of the metric based statistics which the models seemed to discard in favor of more cumulative statistics which indicate that the player played for most of the games available in the season. This makes sense because a player who was injured or simply wasn't a consistent factor throughout the season won't win an award for the season. So it makes sense that the models seem to favor Wins and Strikeouts which are often talked about stats and are cumulative. The emphasis on Strikeouts and not WHIP which measured walks and hits let up per inning pitched also backs up the emphasis on power pitching as opposed to to contact pitching (meaning balls put in play which leads to

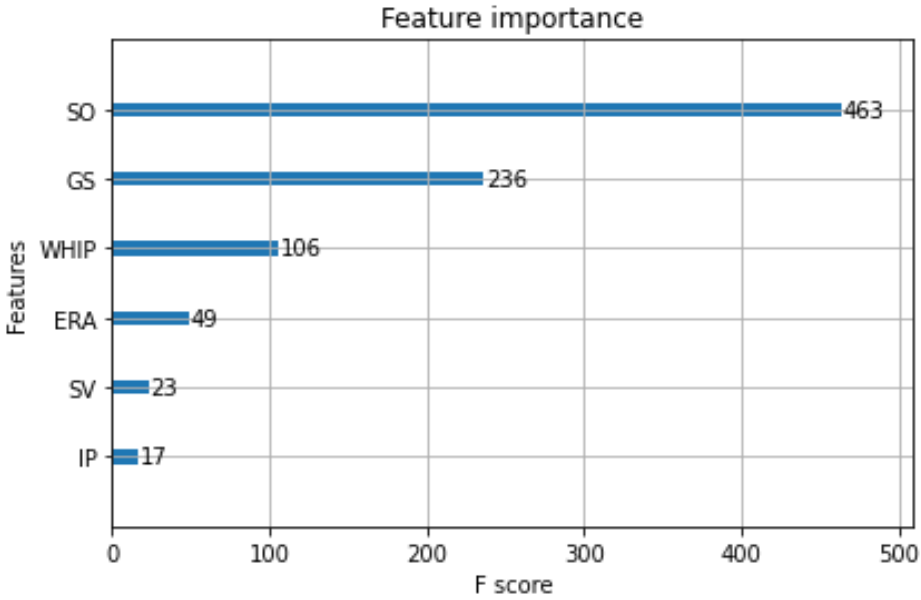
more hits). ERA is also a constant factor in the model's analysis, while it is a metric statistic letting up the fewest runs is a pitcher's chief goal and is perhaps the most looked at statistic when evaluating a pitcher's performance. So it was good to see the models pick up on that, especially that an above average ERA is usually guaranteed not to be an All-star candidate. This is as opposed to a stat like FIP, which tries to emulate ERA while only focusing on Home Runs, walks and strikeouts which are essentially results all within the pitcher's control and not other features. The models didn't deem FIP important so it was taken out. While the overall performance of the models wasn't exceptional, which isn't that surprising considering the abstract and subjective nature of an All-Star, the feature importance was very interesting and useful which is what we were looking for.

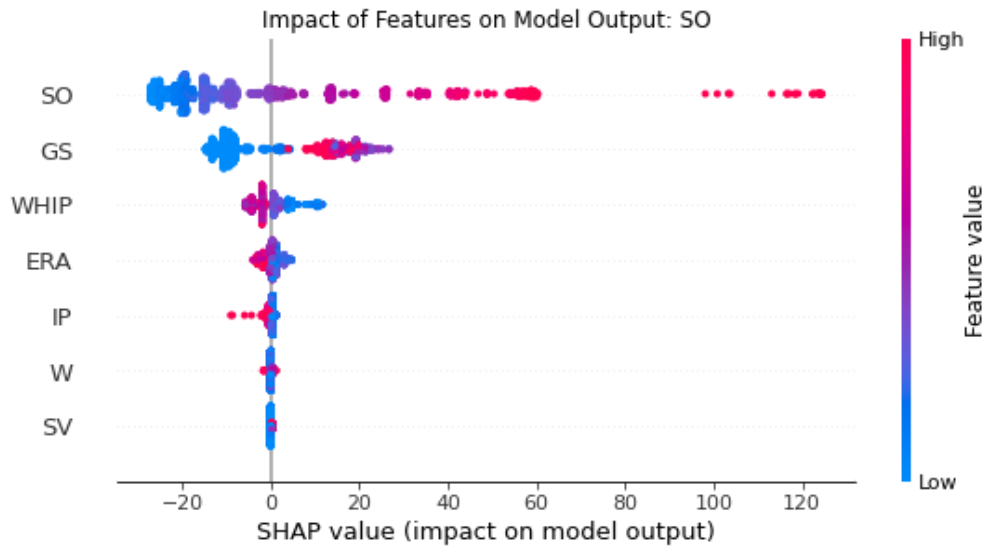
Model 5: XGBoost

For our final model, we utilized XGBoost regression to conduct an analysis of the importance of each feature in predicting next season statistics. We created a predictive model for each of the major pitching categories with tuned parameters for each model to minimize root mean square error with the given training data. We plotted feature importance as well as an accuracy comparison chart for each feature utilizing our models. Below are the charts generated for predicting strikeouts (SO). Features were often their own most important feature, as we can see below where strikeouts were taken as the most important feature for predicting future strikeouts, which is to be expected. It appears from the accuracy chart that our model succeeds best on predicting values within the interquartile range, while less successful for the largest or smallest outlying values. This suggests that our models reduce error by predicting conservatively. Our models also only utilize a single year of statistics to predict the statistics for a pitcher's following season, and so a model that accounted for consistency or trend over time could result in more

accurate predictions. Nonetheless, these models still gave us quite a large insight into how our models utilize each feature in making its regression predictions for future statistics. So much of the value that sports and baseball statisticians add hinges on predicting how players will perform in future seasons given a history of performance. With that in mind, the value of an analysis of the predictive importance of various features is clear.







Summary of findings/Potential implications

What we found out during the course of our project is that this is a very difficult ML problem, choosing All-Stars is subjective of course, and it's a problem that the criteria for an All-Star changes year to year. It's difficult to evaluate everyone at once when All-Stars are selected in comparison to only the players of that specific season, so the margin for an All-Star varies season to season.

Nonetheless the models were able to do a decent job evaluating the features and making sensible predictions.

The models seemed to favor cumulative statistics like Wins and Strikeouts (SO) as opposed to Metric stats like WHIP and FIP but what was encouraging is that it thought that ERA is very important which we know is a big factor for All-Star selection each year.

The importance of SO over WHIP is also indicative of the evolution of what stats are emphasized in the modern era which was picked up by the models. The game has evolved away from the traditional limiting hits and walks approach to a power approach focusing on strikeouts.