Capstone 2 Project Report

The Problem

The problem that I explored was predicting a baseball player's on-base plus slugging percentage (OPS) which is a popular offensive metric to express not only how often a player gets on base, but also how many bases they get when they do (aka their ability to hit for "power" and get extra-base hits). Data has become such an integral part of baseball when it comes to player evaluation and a model that provides insight into the possible future production of players can be extremely valuable to an organization looking to sign a new player or extend the contract of one of their existing players. With some contracts fetching hundreds of millions of dollars, a predictive model has the potential to drive decisions to pay players who are most likely to succeed in future years.

Data Collection

My data came from the Baseball History dataset on Kaggle which has data on players, teams, ballparks and all things baseball. Over the entire dataset (6 different csv files that I merged together), there were 101,000 rows and 55 columns and it was surprisingly clean with no null values needing to be filled. Once the data was all pulled together, I deemed many of the initial columns useless and dropped them from the dataset. I also filtered out everything from before 1985, any seasons where a player had fewer than 200 at bats and any records where a player hadn't had two prior qualifying seasons. After cutting down my dataset, I had a framework of every player's yearly offensive output in the form of basic counting statistics (hits, runs, walks, etc.) along with some additional features like salary and the position they played. Although the data was clean, in order to be prepared for the next steps in the analysis, I had to clean up the data by doing a few things.

First, I had to ensure that each record represented a single season for each player. There were hundreds of instances where players had multiple records for a single year, largely due to situations like mid-season trades where an individual played for multiple teams. On top of that, I created some additional features to add to my data. The raw data had simple counting stats, but I calculated a number metrics to further contextualize the type of season a player had including my response variable, OPS. And finally, I had to shift the data to include the key data points from each of the previous 2 seasons so that I didn't encounter leakage of my response variable into my predictor variables. On top of these steps, I tried a number of feature engineering strategies that were ultimately unsuccessful. To name a few:

- I indicated whether a player was an all-star or not, but that added little value because if their OPS was high enough, the model assumed they are an all-star-caliber player.

- I assigned a 'luck factor' to players who had an above average 'batting average on balls in play' which would indicate the ball somehow dropped in safely more often for them than others
- I highlighted players who were coming off of a shortened season in an effort to capture an injury from the prior season but that ended up including players with a short prior season for a host of reasons (mostly due to young players and the nature of a career starting with some short seasons)

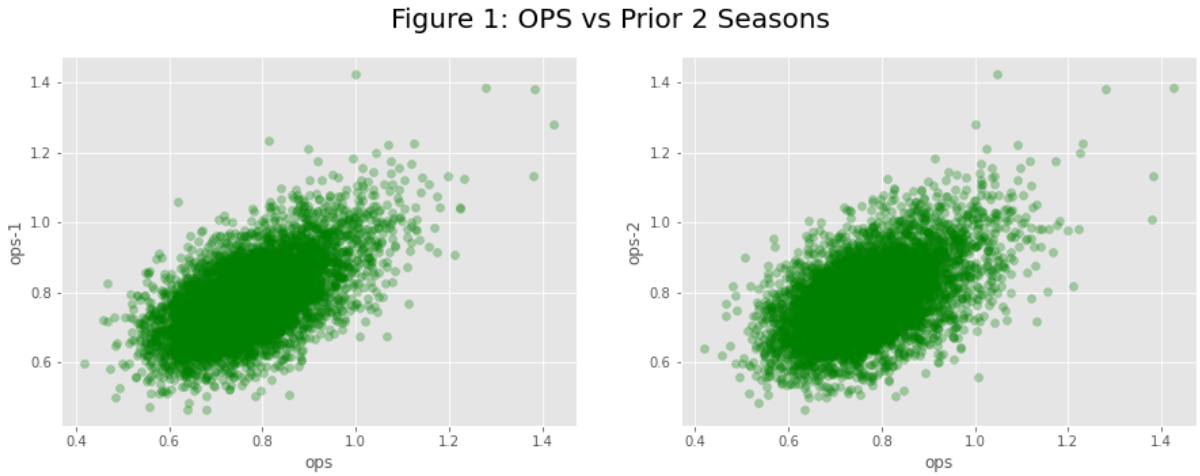
Using what I learned, I decided to create a number of ratios that are not only indicators of a successful season but also seen as better predictors of sustainable or unsustainable production in the future. An example of one of a more sustainable statistic would be Walk Rate or Extra Base Average. If a player has a patient approach at the plate or they tend to drive the ball for extra bases, their future production is likely to be more consistent and thus more predictable. On the flip side; an unsustainable statistic would be batting average on balls in play (BABIP). This relies heavily on the chance that the ball drops in for a hit instead of resulting in an out. Both types of statistics can assist with prediction by either indicating that a player's production is likely to continue or can serve as an early warning sign of worse seasons to come.

Eventually, I also found that using an average of the previous two seasons was more effective in order to smooth potential outlier seasons. In many cases, typically at the beginning and end of a player's career, there is a large disparity between their OPS in the season I want to predict and their prior seasons. The largest variances in my predictions were due to these breakouts and declines in player's careers. Upon completing the data cleaning and creating new features, I was left with a dataset of 6,062 rows and 32 columns.

Exploratory Data Analysis

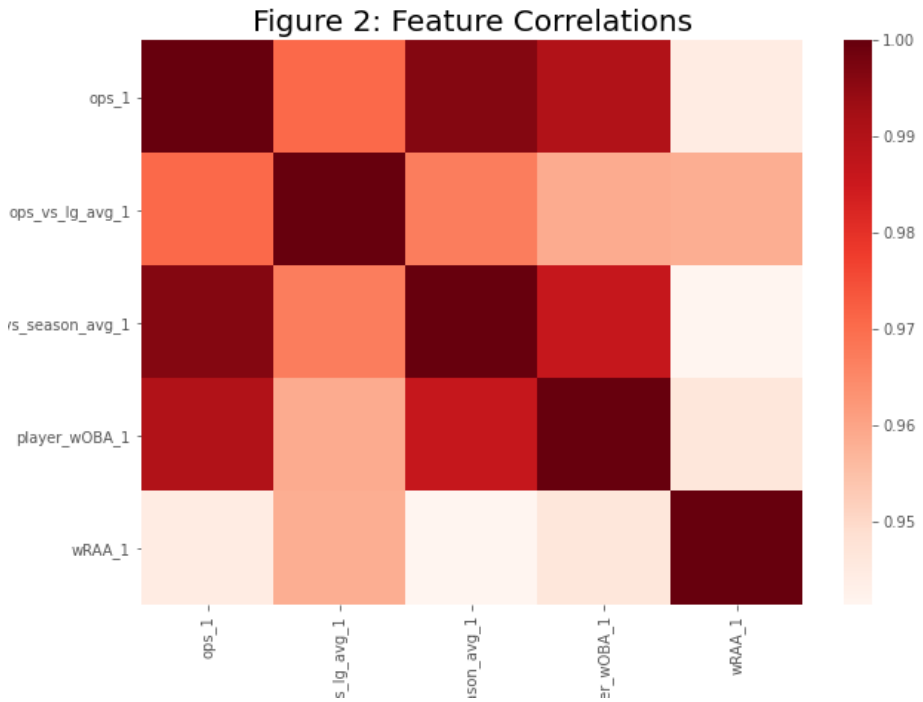
Upon my initial exploration of my features and their relationships, comparing OPS with prior years was an obvious place to start. As seen below in Figure 1, I compared the prior two year's OPS to the season I was trying to predict, and the relationship was clear.

Figure 1:



From there, I made an effort to create new ways to show the OPS in relation to not only an individual's career, but also the standard of the league. The issue became high correlation between a number of predictor variables that ultimately weakened my model, shown in Figure 2 below. Of course, most of these features were not included in the final model and I had to be more creative to avoid situations like these.

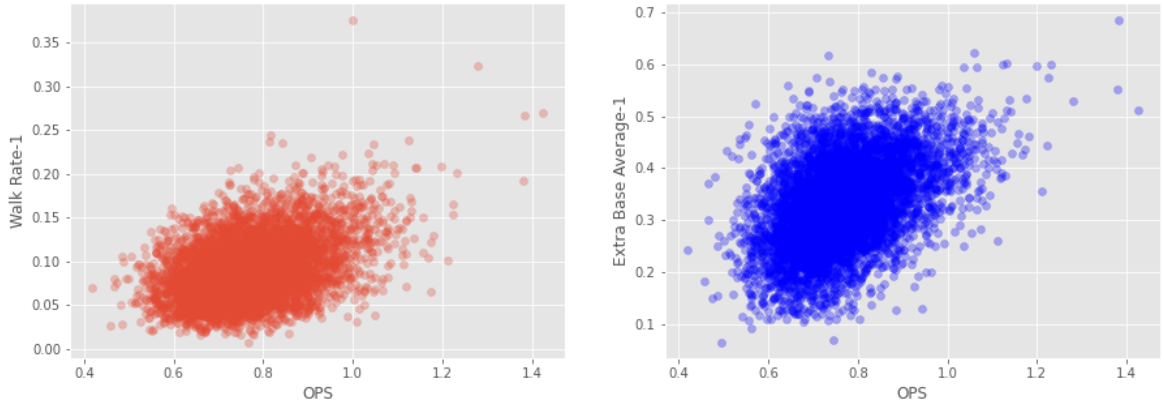
Figure 2:



The challenge of creating new features that would aid in the predictive power of the model without presenting a collinearity issue took some time to solve. Once I created the ratios mentioned earlier (Walk Rate, Extra Base Average, etc.), I had new features that would aid in predicting OPS without the issue of high correlation between my independent variables as demonstrated below in Figure 3 showing OPS against Walk Rate and Extra Base Average in the previous season.

Figure 3:

Figure 3: OPS vs Walk Rate & Extra Base Avg



With the addition of these features, I was prepared to predict the OPS

Modeling

Attempting to generate accurate predictive results from strictly historical data is a difficult task, especially when the variable you are trying to predict fluctuates quite a bit from player to player and year to year. When I started testing out some models, the results were disappointing. My errors were high with the average error of 0.1 being worse than if you were to just use the OPS from the prior year as the prediction (Relationship shown above in figure 1. This yields an MAE of 0.076. The MAE of OPS vs the mean of the prior 2 years is 0.070). As I created the new features mentioned above, the models results were improved significantly. I chose to use the below regression models since I was predicting a continuous variable:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor

To ensure I wasn't overfitting my model to the training set, I used k-fold cross validation for the Linear Regression model and GridSearchCV for the Random Forest and the Gradient Boosting Regressors. Although the Grid Search did not greatly improve the performance of my model, it did help me tune my parameters and ensure I was not overfitting.

My Findings

After running the models above, I found the most effective to be the Linear Regression model. Not only did it have the highest R-squared and the lowest mean absolute error, it also was not nearly as computationally expensive as the Random Forest or Gradient Boosting models (see metrics below in Figure 4). With the final model, I achieved a mean absolute error of 0.063, a 20% improvement over simply using last year's OPS as a predictor.

Figure 4:

Model	MAE	R-squared	fit_time	test_time
Linear Regression	0.0632	0.4634	0.04	0.02
Random Forest Regression	0.0635	0.4511	10.48	0.24
Gradient Boosting Regression	0.0638	0.4480	9.73	0.10

As mentioned above, the biggest misses for the model were predicting breakout years for younger players and the beginning of the end for older players. In figure 5 below, the first highlighted points are ***. Predicting a player's breakout or career decline is out of the scope of this project These are both age-old questions for player evaluators and unfortunately, due to human nature, it is impossible to predict exactly what OPS a player will have in the future. Fortunately, with the application of the Data Science method and Machine Learning models, we can make predictions with almost 20% more accuracy than an educated guess (prior year OPS).

Explore Further

While working through this project, several ideas jumped out at me that I would be interested to explore further. These ideas mostly come down to more robust data, but I am intrigued by the below:

1. For this project, I was able to see a player's basic offensive output for prior years but did not have some of the underlying advanced stats like exit velocity (how hard a player hits the ball) or launch angle (the angle at which the ball is hit). Access to stats like this would allow me to better predict when a player is about to break out or when their career is nearing the end which was the driver of the greatest variance in this model.
2. The true value of my project is evaluating players who are several years into their careers before committing to high-dollar, long-term contracts (players don't hit free agency until 6 years into their careers in baseball). However, looking into a player's college and minor league stats, including the more advanced stats mentioned above, would give me the opportunity to create a model that would help find "diamonds in the rough", or lesser known players who are undervalued that possess high upside potential.
3. A more in-depth project would be predicting a player's Wins Above Replacement (WAR). This seems similar to my current project, but beyond predicting an offensive metric like OPS, WAR includes defense and baserunning. When evaluating a player's value, these are also important parts of the game that should be considered before committing to a player long-term.

General Observations and Notes

As I worked through this project, it became clear that the steroid era in baseball had a major impact on my dataset. When I researched the most apparent outliers, it always seemed to be one of the players who all but certainly used steroids and put up staggering numbers (see the 4 points in the top right of Figure 1 if you're curious if Barry Bonds was the best player in baseball). Although it was the most exciting era in baseball, it did not do me any favors in my efforts to predict OPS. In addition; it is simply a very difficult thing to predict. Using past performance in the form of simple statistics does not allow for precise predictions. However, when evaluating players and making decisions, any indication of what the future may hold can be extremely valuable.