Capstone 2 Baseball Report

By Thomas McMahon

Over the past decade sports have become much more data driven, none more than baseball. In sports to gain the upper hand over your opponents you either need to have deeper pockets than them, or you can be smarter than they are. This second idea is the option that the Oakland Athletics needed to choose in order to compete with bigger market teams such as the New York Yankees. How they accomplished that is through the use of data. Former professional baseball player turned manager Billy Bean after losing another star player to money decided to change the way a team could be put together. With the help of analysts, he determined that he was going to put the cheapest team on the field with the highest on base percentage. This is the idea that inspired this idea to be able to predict how a player well a player is going to bat in future seasons. This project will be able to be used by anybody that is trying to put gain a competitive advantage over opponents in fantasy baseball or even by MLB teams trying to decide which way to move forward with there franchises when it comes to who to sign and who to let go. This project will take public information on all players stats going back to 1970.

The first set of the project was to being the dataset into a notebook to clean the data. The data had columns for the players name, year, at bats, singles, doubles, triples, and home runs. What wasn’t included was the players batting average, which was added as a feature to the dataset as well as the players age.

The trends in the columns were that as all of the features in the dataset increased so did the batting average. Another observation with the at bats column was the widest range in batting averages came when players have fewer at bats. This makes sense if a player had one at bat their average will either be 100% or 0%. It was decided to removed all rows were the player had less than 50 at bats in the season to keep as much as the data as possible while still eliminating so of the outlining data.

With all of the features in the dataset having either a very weak correlation or a positive correlation feature engineering is going to be very important. The features that were created career bating average, last seasons batting average, last 3 seasons batting average, percent change of last season’s batting average vs career average, and percent change of last 3 seasons batting average vs career average. These new features gave the model more data to use to make predictions, but there was a problem with the format of the data. The fact that the model is going to be used to predict future performance all of the data needed to be shifted back by one season. This is to insure when training and testing the model the model does not have any access to data from the year that it is making the prediction for. With all of the new features added to the dataset as well as the data shifted it is time to build models.

The first step of building the model is to break up the dataset into a dataset of all the features and a series with the target variable. Scaling of the data is needed to make sure that the magnitude of any of the columns does not create bias in the data. This is achieved by changing the data to have each feature have a mean of 0 and a standard deviation of 1. The first 2 models build were linear regression models. The first model was split into training and testing data using train test split. The mean absolute error of the first linear regression model was .043. Not a very promising result so a new model was built. The second linear regression model had the data split into training and testing data based on the year of the season. Everything before 2005 was used as training and everything after was used as testing data. The mean absolute error of the second linear regression model was also .043.

After building the linear models it seemed that the data was too complicated for a linear model. For this reason, random forest models were built to try and make the models better. Random forest models will make better predictions because they models built from making individual models that learn from each simple model that is built. For this reason random forest models tend to do better with predictions vs linear regression models.

The first random forest model although better than linear regression still had a mean absolute error of .042. After building this model it was clear that tunning hyperparameters was needed to find the best possible random forest model. Using grid search while specifying different possible values for max depth of the model and min samples for a leaf I found the best parameters were a max depth of 10 per tree and min samples per leaf of 5. With these hyperparameters the mean absolute error dropped slightly to .041.

In conclusion although the model did not perform as well as I would have liked, before Billy Beans so called crazy money ball experiment this model would be more effective than how baseball scouts used to try to predict how a player was going to play. Some of these old metrics were how hot the girlfriend or wife of the player was with the rational being if they are hot then the player is more confident in themselves and therefore would preform better on the field. There may have been some features that I could have overlooked or data that I didn’t have, such as could the precent of righthanded vs lefthanded pitchers change the batting average, because statistically teams defensively want to have righthanded pitchers throwing to righthanded batters and vice versa. This would have been something that I would have liked to add to the model but I don’t have the information on what hand the players bat with so even knowing what hand the pitchers throw with it wouldn’t matter because I don’t have that information on the batters. I defiantly think this is a worth wild project although maybe not with the information we have now, but maybe down the line with more stats being tracked I wouldn’t be surprised if in a near future teams will be able to more accurately predict how a player will perform.