Can you Predict the MLB?

A report by John Fleissner and Brendan Doucette

Table of Contents

Introduction

For our final project we used an online github repository that featured the statistics of every MLB team in a bunch of different categories. We wanted to model a prediction of a certain statistic of a player to see how they could potentially perform in that statistic in the upcoming season. This github repository included all sorts of stats including pitching, team performances, and individual player performances. This repository also has stats from every season dating back to the first recorded MLB seasons.

Our motivation for this project is based on the movie Moneyball. In Moneyball, the manager was focused on trying to rebuild the Oakland A's with a small budget and not many players on the market. To overcome this obstacle, they manipulated certain stats to try and find players that could fill the void of their good players that left for more money. They used stats like on base percentage and other analytical approaches to find players that they could sign on their budget. In the movie they did not use any predictive measures to find players, but used stats given to them from these players from past seasons. Even though they did not use any predictive models, this movie still motivated us to manipulate statistics in a way to make a prediction on a player's stats. We wanted to see if our predictive models on a stat can be used to accurately represent how a player could perform in their upcoming season.

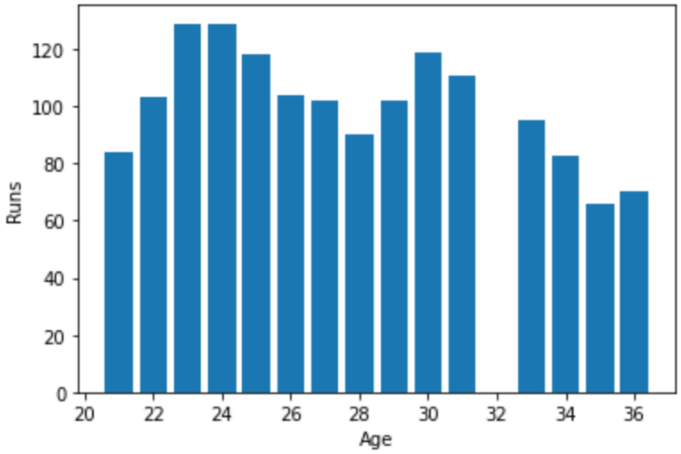
We are also big sports fans and enjoy betting on games, so we wanted to see if we could make accurate predictive models to help see if any players will get better over the coming years or begin their decline. We can also use this information to create bets that can have a strong basis for why we are choosing to bet in a certain way. Also, for people who participate in fantasy baseball they could eye players that might make a performance jump next year. If you had to pick between two people on your fantasy baseball team you could run our model and see which is expected to have better outcomes the upcoming season.

We decided to focus solely on a player's total runs stats. We were able to break up the data to only provide us with stats that can be used to help us predict future runs stats. For our sake, we only used data from 2017 and 2018 before covid started because we felt that was the last good season where players were most fully healthy and more committed to baseball. Furthermore, through covid there were a lot of players out at sporadic times so we wanted data that was the most normal.

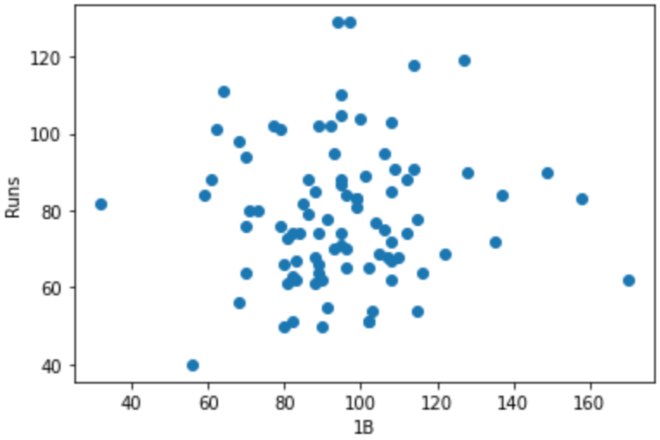
The data for batting stats has over 300+ different variables, but for our regression we only focused on 12 variables. The variables that we focused on were age, games played, at bats, plate appearances, hits, singles, doubles, triples, home runs, runs, runs batted in (RBI), and times walked.

Before manipulating the data to create split data sets for our regression models, we wanted to plot different variables on their next year's runs. The plots that we looked at were runs based on age, singles, doubles, triples, RBI's, and lastly 2017 runs compared to 2018 runs.

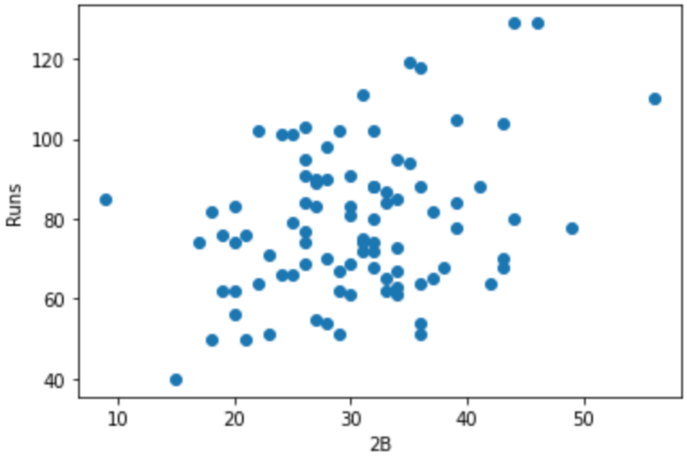
Initial Data Discovery



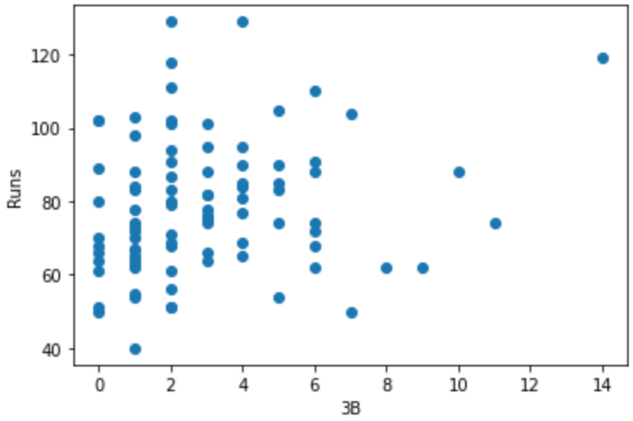
This graph does not show a strong correlation between runs and age. There does not seem to be a trend between the increase in runs and increase in age. The points are very sporadic. It is safe to assume that as age increases it will have a somewhat negative correlation to a player's runs. For players on the older side, age would definitely have a negative impact on their performance in upcoming seasons. For some players though, if they are on the younger side then aging could have a positive correlation on their performance in upcoming seasons.



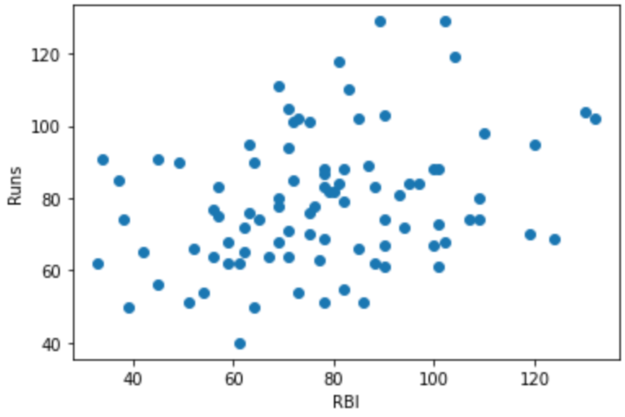
The amount of singles a player has seems to be a good representation on how many runs the player has. If a player has a lot of singles then we can presume overall that the player also has a lot of runs. This variable will most likely have a positive correlation in predicting how many runs a player will have in their upcoming season. There are a couple outliers noticeably on both ends of the scale, but most of the points lie in similar positions based on their number of singles. This makes a lot of sense that if a player has a lot of singles then they will have a lot of runs because a single is the easiest type of hit to get and clearly happens the most often.



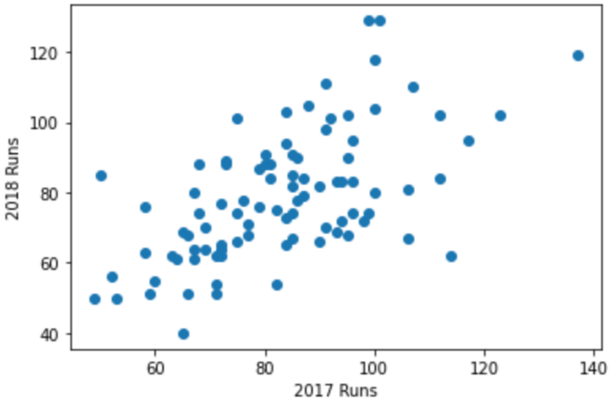
This graph is also similar to the singles graph where there is a correlation between the number of doubles and the number of runs. There are also some clear outliers on the lower and upper end, but most of the points fall in similar areas based upon the number of doubles. The number of doubles has decreased from the number of singles there which is expected because a double is a harder hit to accomplish than a single. The total number of actual runs has stayed the same and the graph is still showing that as the number of doubles increases then that means that the player is also hitting the ball more times too.



This graph displays triples on runs in 2018. This graph is not linear like the other graphs previously. There are multiple outliers for each number of triples and there is no linear correlation between the number of triples and runs. There are points with high triples and high runs, low triples and high runs, high triples and low runs, and high runs and low triples. The range of total triples is a lot smaller than doubles and singles. It only goes up to 14. This is more understanding because hitting a triple in baseball is a lot harder than hitting a single or double. Given this, triples will not be as strong of an indicator for runs as singles and doubles would be.



This plot shows RBIs (runs batted in) on total runs. This plot is a lot more sporadic and doesn't show any real strong correlation between total runs and amount of RBIs. This makes sense as well because RBIs consist of singles, doubles, triples, and homeruns. As long as the player's hit brings in runners then that is considered an RBI. Someone could be getting a lot of runs, but not bringing in batters. There are areas of points with high runs and low RBIs as long as high RBIs and low runs. This variable will have little significance to the runs prediction, but since the data is so sporadic it will not be as significant as singles and doubles. You can see a small trend between high RBIs and high runs, but there are a lot of outliers within the data.



This last plot shows 2017 runs on 2018 runs. There is a linear pattern between runs from 2017 to runs in 2018. Using runs from a prior season will be the best predictor for runs in a future season. Most of the points are in a linear pattern and show that players will most likely stay in the same relative area from prior seasons. There are outliers on the higher and lower end which can show that some players are making drastic improvements or decline. Improvements can be anything from offseason improvements, change in team, change in role, steroid use, etc. Players declining can be anything from injuries, aging, as well as changing teams or roles. This variable will also have a strong weight towards the prediction models. Using previous season runs will most likely have the strongest positive weight in the prediction models.

Potential Shortcomings and Upsides

Using this data it is easy to find how there can be many pros and cons to our prediction models. First for cons, there is human error involved in trying to predict a players future stats based off of their past seasons. Players can get hurt, get traded, take on different roles on a team, get out performed by other players, can drastically improve or worsen over the off season. Also, since covid a lot of players could drastically change in how they play or how they have improved because the league would have to adapt to new covid protocols. Since there is a covid drop-off and the data is being used from the last normal season, it would be hard to create an accurate prediction model. Also, there could be some variables from the selection that we chose that might not be as helpful in our models and there may be variables that we left out that could be more helpful.

Some pros from our models and research is that we are using the most accurate way to make predictive models on the players. Using past stats is the only way that we could possibly make some future prediction for what their stats might look like. Also, these models give feedback on our tests so that we can find how accurate the tests are and the weight of each variable on the test. We can use the R-squared value to let us know which models are the more accurate models. Also, the mean absolute error can show us the magnitude of the errors and can let us know if there are outliers that are playing a factor into the accuracy of our results. Also, even if the predictions are not accurate, the models can set a good basis as to what to potentially expect from a player in their upcoming season. You could create a relative range from the data based on what the models predicted. Lastly, if someone is comparing two players, their stats were predicted based on the same models so you can make a smart assumption on which player would perform better in the upcoming season.

Models

To conduct our models, we partitioned our data into training and testing data frames. We used their 2018 runs as the y training and testing data frames, and the rest of the 2017 data as the x training and testing dataframes. The models that we conducted were a lasso cross validation, a ridge cross validation, an elastic net regression, a linear regression, and random forest classifier.

Our model that performed the worst was the lasso cross validation model. This model produced an r-squared value of 0.508. This is not bad accuracy for trying to predict an MLB's player's runs. There are a lot of outside factors that can play into a players performance that it is very hard to predict a very accurate prediction. 0.508 is still a really good r-squared value. The mean absolute error value was 11.361. This is the magnitude of the errors. This number is very good for this model.

The next model that performed a little bit better was the ridge cross validation. This model has an r-squared value of 0.512. This is also another good r-squared value, but not the best one. The mean absolute error value was 11.628. This is also a good number and very similar to the lasso model.

The Elastic model was the next model that performed a little bit better than the ridge model. This model had an r-squared value of 0.534 and an MAE value of 11.048. Both these values are really good.

The next model that performed similarly to the elastic model was the linear regression. This model had an r-squared value of 0.535 and an MAE value of 10.806. This model also performed very well and was similar to the other models.

The model that did the best and outperformed all of the other models was the random forest classifier. The r-squared value was 0.82 which was by far the best value compared to the rest of the other models. The MAE value was 16.741 which is a little higher than the other models, but is still a good value for these types of statistics.

Conclusion

After running these 5 models, we were satisfied with the results of all of them. There was a good correlation between the stats that we choose to help determine the runs of the players in the future. The ridge cross validation, lasso cross validation, linear regression, and elastic net regression models all performed very similarly to each other. All of their r-squared and MAE values were very close to each other. The model that performed the best compared to these was the random forest classifier. The random forest classifier had an r-squared value of 0.82 which performed a lot stronger than the rest of the models. Overall, all of the models produced promising strong results.

The Elastic net model was able to provide feedback on the weights of the variables in producing these predictive measures on a player's runs. Age and games played had the stronger negative weight towards our prediction. Age had a weight of -6.96 and games played had a weight of -4.07. These variables having a negative weight make sense. As a player ages they could either improve or decline which would make it harder to try and predict how their runs would be affected by their age. The players are not going to follow the same trend in their runs based on their age which makes it a negative variable in trying to predict their future stats. The same for games played. Every player is not playing the same amount of games as each other which would make it hard to use that variable as a strong predictor in their runs. Plate appearances and doubles had a positive weight towards our models. Both variables held a weight of about 3.8. This makes sense because the more plate appearances you have the higher chance you have to get on base which then ultimately leads to the player potentially scoring a run. As plate appearances increase you can assume their runs will increase as well. For doubles, This also makes sense because doubles put the player in a good position to score a run. A double is harder than a single, but puts the player in a better position to score. A triple puts a player in better position to score, but those hits are so much rarer than doubles that a double would be a better predictor. The variable that had the best positive weight was a player's runs from the prior year. Runs from a prior year had a weight of 9.55. This makes the most sense that a player’s past runs statistic would have the greatest weight since that is the variable we are trying to predict. It is the variable with the strongest basis in trying to predict future runs.

Although our models may not be the best in trying to bet on a specific amount of runs a player will get, it does a good job of showing which players will outperform others. This is helpful to people who play fantasy baseball and want an edge in their draft or recruiters in the MLB looking for new prospects. If college statistics were easily available it would be interesting to see if you could predict how a player would play in the MLB.