Data Science Capstone: Baseball Homerun Predictor

Andrew Hood

March 14, 2024

Contents

1 Introduction 3

2 Initial Analysis 3

2.1 Player ($playerID) . . . 4

2.2 Year ($yearID) . . . 4

2.3 Games ($G) . . . 5

2.4 At-bats ($AB) . . . 5

2.5 Walks ($BB) . . . 6

2.6 Strikeouts ($SO) . . . 6

2.7 Summary . . . 8

3 Methods 8

3.1 Preparing and Splitting the Data . . . 8

3.2 Random Forest Model . . . 8

3.3 Linear Regression Model . . . 9

4 Results 9

4.1 Random Forest Model . . . 9

4.2 Linear Regression Model . . . 9

4.3 Comparison . . . 10

5 Conclusion 11

1 Introduction

The Baseball Homerun Prediction project forms the second part of the HarvardX: PH125.9x Data Science: Capstone course; the final course in the Harvardx Data Science Professional Certificate series. This project was a 'choose your own' project. I decided to choose a dataset containing historical baseball batting data. The dataset was downloaded from Kaggle, and was created by Ed King who used data from Sean Lahman's baseball Databank website. The File used for this study is available at: (https://www.kaggle.com/datasets/open-source-sports/baseball-databank?select=Batting.csv).

The goal of this project was to use two different machine learning methods to predict a player's number of homeruns for a given year. My determination of success for this effort is a Root mean Square Error (RMSE) of less than 4, and an R-Squared value of greater than 0.7. I will also provide the Mean Square Error (MSE), however, I will not be analyzing that metric as RMSE is easier to interpret. MSE is simply provided for user comparison purposes. These benchmarks would allow me to predict home runs within 4 home runs, for 70% of the data. Given the large variation in the data, these cutoffs were deemed acceptable.

This report will begin by exploring the data and performing an initial analysis, and then outline the methods used for the two different models. After that, this report will present the findings from the two models, and then compare them. Finally, a conclusion section will be presented to recap the effort, and identify limitations.

2 Initial Analysis

First, I had to ensure the data was clean by checking for missing values, and then normalizing them. I chose to replace the missing rows with the column medians. I chose the median over the mean because the dataset contained over 100 years of data, and the game of baseball has changed over time, leading the mean to be skewed foe this specific use case.

When pulled from Kaggle, this dataset was much larger. I decided to trim the dataset down to include five predictors of homerun numbers. These predictors were; yearID, G, AB, BB and SO. I also kept playerID as an identifier. An explanation of each variable is as follows; playerID (A unique player), yearID (A unique year), G (The number of games a player played for a given year), AB (The number of at-bats a player had in a given year), BB (The number of walks a player received in a given year), SO (The number of times a player struck out in a given year).

Firstly, we see that the dataset contained 101,332 rows and the columns defined above, the 8th column is the target variable; HR, or the number of homeruns a player hit in a given year. Given the nature of sport, there is variation in the data for each player due to injuries, slumps, and length of career. Below is the first 5 rows of the batting dataset.

Table 1: First 5 Rows of 'batting' dataset

HR	playerID	yearID	G	AB	BB	SO
integer	character	numeric	numeric	integer	integer	integer
0	abercda01	1871	1	4	0	0
0	addybo01	1871	25	118	4	0

HR	playerID	yearID	G	AB	BB	SO
0	allisar01	1871	29	137	2	5
2	allisdo01	1871	27	133	0	2
0	ansonca01	1871	25	120	2	1
0	armstbo01	1871	12	49	0	1

2.1 Player ($playerID)

playerID was used as an identifier in this study, similar to how movie was used as an identifier in the first project of the capstone course. However, it is important to note that some players hit more or less homeruns than other players. A certain players homerun numbers in a vacuum could be influenced by strength, skill and quality of pitching faced— all things not accounted for in this study. Future studies could take these things into account.

2.2 Year ($yearID)

Analysis of year data revealed that there was a general increase in homeruns hit in any given year. However, as evidenced by the plot below, this increase was not entirely linear. The general trend is positive while still maintaining troughs and peaks over the years. The overall average of homeruns in the dataset was 2.8.

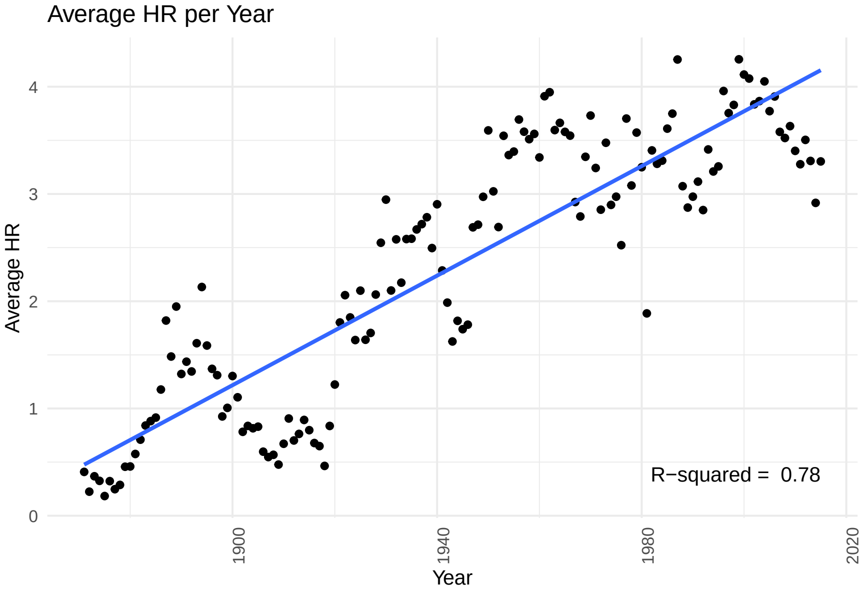

Figure 1: Distribution of Average Homeruns by Year

While not a super strong effect, we can see that year does have a slight positive linear correlation with the average number of homeruns hit (0.78).

2.3 Games ($G)

Next, we will examine the effect of the number of games played (G) on average number of homeruns hit. The average number of games played in the dataset was 51.4. One would assume that a higher number of games played would correlate to a higher amount of home runs hit. That is not always the case as evidenced below.

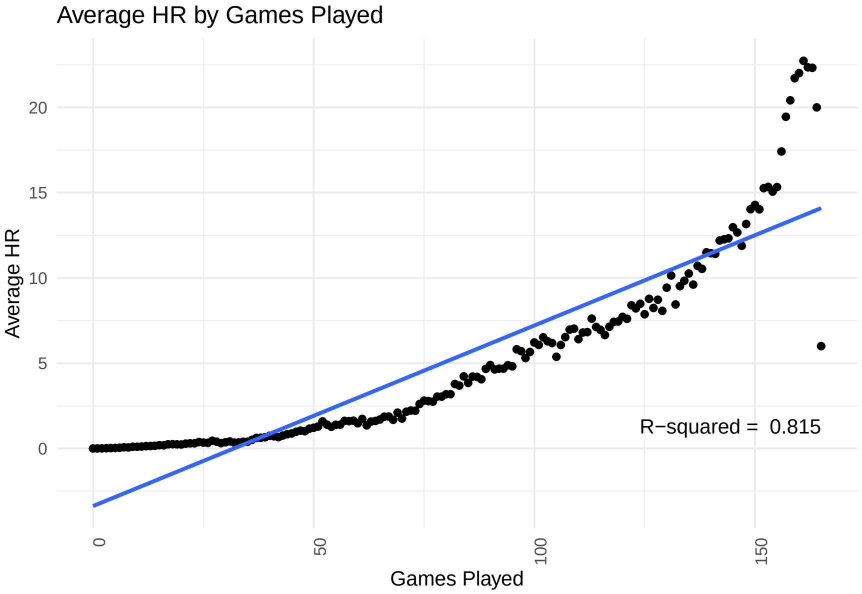

Figure 2: Distribution of Average Homeruns by Games Played

From the chart above we see that the average number of home runs hit is relatively flat until slightly more than 50 games are played. After 50 games played, the average number of homeruns follows a slight positive linear trend (0.815). Right around 125 games, the number of average home runs assumes a steeper positive trend. Overall, the trend is exponential in nature; it starts off slow and then rapidly rises further along the x-axis. Given the proximity of the data points to the line of best fit, we can say that games played and average home runs hit are correlated, making G a good predictor of home runs.

2.4 At-bats ($AB)

The at-bats (AB) variable had a mean of 145.2. One would think that as the number of at-bats increases, the average number of homeruns should also increase due to the higher amount of chances to hit a homerun. However, this is not always the case as shown below.

The chart above shows a positive linear correlation (0.861) similar to the G variable, however, as the number of at-bats increases, we see an increase in the number of outliers and the distance of the data points from the trend line. This indicates that AB is a good predictor of homeruns for players with less at-bats, but is not as good of a predictor for players with a higher number of at-bats. The at-bats variable will still be an important predictor, but it may not be as accurate as others as its level rises. This variation could potentially be explained by improved scouting reports on players as their number of at-bats rises, and their underlying skill level.

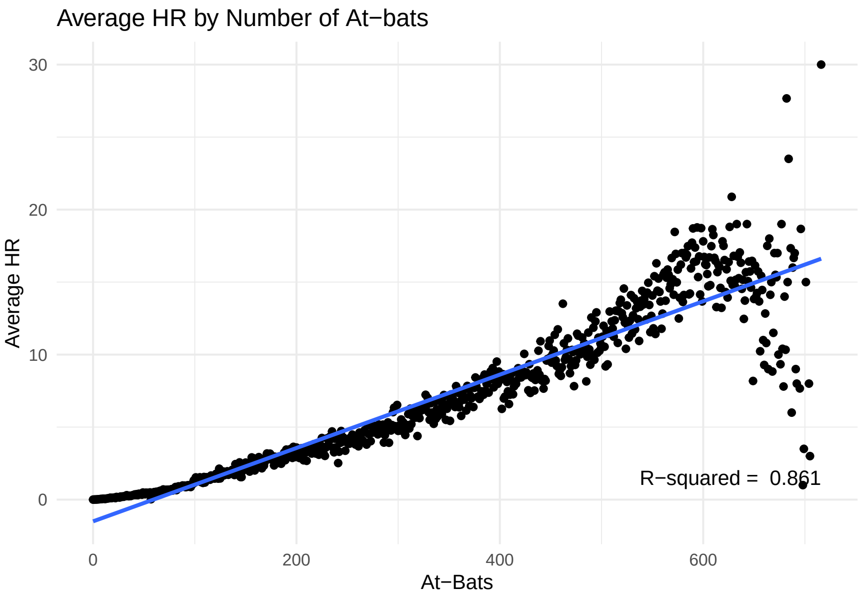

Figure 3: Distribution of Average Homeruns by At-bats

2.5 Walks ($BB)

The average number of walks in the dataset was 13.3. Walks were not expected to be a strong predictor of average home runs due to the fact that if you receive a walk, you can not hit a homerun in the same at-bat. However, it is commonplace in baseball to intentionally walk a player who is known to hit a lot of homeruns, or a player who is known to perform above average in high-stress situations. I did expect to see some correlation between walks and average homeruns due to the phenomenon mentioned, this correlation could also potentially account for previously unaccounted effects, such as skill level.

As evidenced above, average number of home runs and number of walks have a relatively strong positive linear correlation (0.829). We also see that as walks increase, the number of outliers increases. As mentioned prior, this could possibly be explained by the effects of skill level or stress level of a given at-bat. Lastly, walks will be an important predictor of average homeruns, due to the correlation shown here and the relation to other unmeasurable effects mentioned.

2.6 Strikeouts ($SO)

Another truth in baseball is that true 'power hitters' tend to strike out more than their counterparts who are 'contact hitters'. Using this truth, we can identify a relationship between the number of strikeouts and the average number of home runs hit. Below is a chart illustrating the distribution of the average number of home runs by number of strikeouts.

As shown above, there is a strong positive linear relationship (0.908) between number of strikeouts and average number of homeruns. This illustrates that strikeouts will be an important factor for the model to accurately predict homeruns.

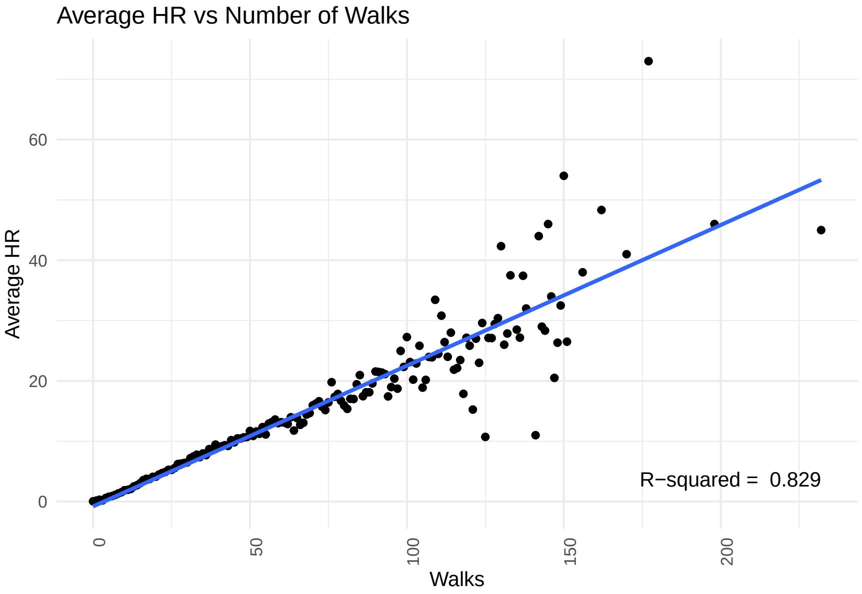

Figure 4: Distribution of Average Homeruns by Number of Walks (BB)

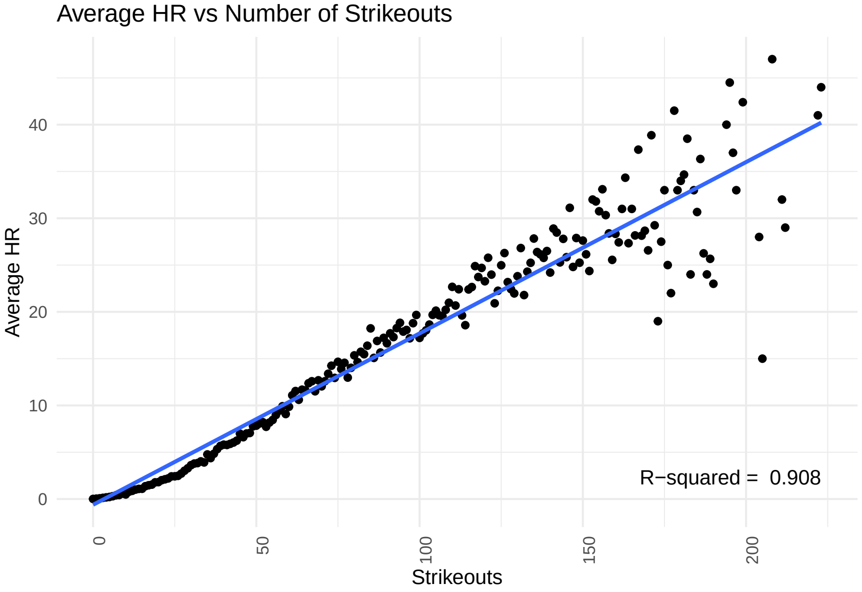

Figure 5: Distribution of Average Homeruns by Number of Strikeouts

2.7 Summary

While the five predictors chosen for the model vary in their correlation to average number of homeruns, they will all be useful in the prediction. Higher correlation shown in the graphs above indicates a more direct relationship of a variable on average number of homeruns, whereas, a slightly lower correlation indicates a less direct relationship, but allows us to account for certain 'immeasurables' in the game of baseball described above. The next section will dive deeper into the two machine learning algorithms used for prediction accounting for the variables outlined above.

3 Methods

Given the variation of the data and the instructions for the project, the first method I chose to use was the Random Forest model. Random Forest allows for a high-accuracy approach using decision trees that will also prevent overfitting the model to the dataset. The second method I chose to employ was the Linear Regression method. I chose this method due to its simplicity and lightweight, fast computation time.

3.1 Preparing and Splitting the Data

Next, I had to split the batting dataset into train and test datasets for use in training the models, and then using them to make predictions. I used the createDataPartition function to achieve the split, and I chose an (80/20) training to test split. I chose 80/20 rather than 50/50 because I wanted to maximize the models' ability to learn patterns and relationships present in the data. I felt that a 50/50 split would have been yielded a large enough training dataset to accurately predict the HR values in the test dataset.

3.2 Random Forest Model

The main model used for this study was a Random Forest model with 100 decision trees. The idea of Random Forest modeling is that the combination of multiple decision trees each trained on random subsets of the data will provide an accurate model to be used for predicting a target variable, in this case HR. The individual predictions are made by averaging the results of each individual decision tree in the model. In this case, the results of 100 trees were averaged to make the predictions.

While simply adding more trees, or tuning the model for the optimal number of trees can improve accuracy, that method was not employed in this study. I chose to not tune the model, and stick with 100 trees due to the limitations of the machine this model was ran on, and runtime constraints. Additionally, 100 trees performed well enough to achieve my stated goal.

To utilize the Random Forest model, I simply ran a Random Forest with 100 trees on the train dataset, and then used this model to predict homeruns in the test dataset. After that I used postResample to calculate the key evaluation metrics.

3.3 Linear Regression Model

Next, after analyzing the five variables presented prior, I felt that a linear regression model accounting for these variables would be a good model to compare to my Random Forest model. Since every variable used for prediction had a linear relationship greater than 0.7 with the average number of homeruns hit, Linear Regression showed promise in its prediction abilities. Since earlier i illustrated the correlation of each variable on the target variable (HR), the next step for that model was to simultaneously account for all five variables, and then use that model to predict the number of homeruns per year in the test dataset.

First, I ran a Linear Regression model using the five predictors on the train dataset. This established the model for predictions. I then used the model along with the 'predict' function to predict the number of homeruns per year in the test dataset. I then calculated three metrics for evaluation of the model. These three metrics were, Mean Square Error (MSE), Root Mean Square Error (RMSE) and R^2. In the next section, we will dive deeper into the results of the model.

4 Results

4.1 Random Forest Model

The Random Forest model was accurate enough in the predictions of actual homeruns in the test dataset to achieve both goals stated at the beginning of this report. The key evaluation metrics for the Random Forest model are presented below.

Table 2: Random Forest Model Results

Method	MSE	RMSE	R2
RF	10.7808	3.283413	0.7241027

The Random Forest model achieved an RMSE of 3.28 and an R^2 of 0.72. Both of these numbers surpassed the goal stated at the beginning of this report. Random Forest modeling is powerful enough to be a more accurate predictor, however, it was limited in it's development for this study. The conclusion section will dive deeper into the limitations of this model.

4.2 Linear Regression Model

The Linear Regression model was moderately accurate in predicting homeruns per year. This disproved my hypothesis that when accounting for all five variables, the correlation would increase. Instead, the table below illustrates that the R^2 value actually decreased when accounting for all five variables.

Table 3: Linear Regression Model Results

Method	MSE	RMSE	R2
LR	12.97595	3.602215	0.6815353

Despite the R^2 value being relatively low, the RMSE was ~3.6, indicating that the Linear Regression model was within ~3.6 homeruns of the actual values in the dataset. Given the simplicity of a Linear Regression model, and the complexity of baseball, this was an acceptable margin error. The chart below illustrates the prediction accuracy.

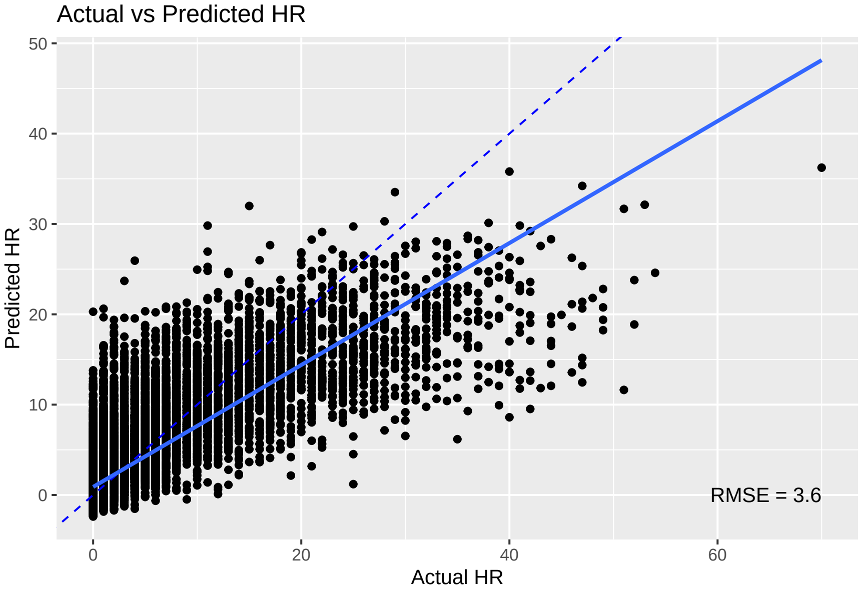

Figure 6: Linear Regression Predicted Values vs. Actual Values

As evidenced above, the Linear Regression Model was a decent predictor for lower numbers of homeruns, but the spread increase the higher the number of homeruns became. The dashed line indicates the Linear Regression model's fit to the data. Given the table and chart presented above, the Linear Regression model partially met the stated goal with an RMSE below the threshold, but an R^2 lower than desired.

4.3 Comparison

Finally, we compare the results of the two models against on another. The table below details the key evaluation metrics for both models.

Table 4: Comparison of the two Methods

Method	MSE	RMSE	R2
RF	10.78080	3.283413	0.7241027
LR	12.97595	3.602215	0.6815353

As evidenced by the table above, the Random Forest model achieved better performance than the Linear Regression model. The Random Forest model achieved and RMSE of 3.28 and an R^2 of 0.72. Indicating that the model was within 3.28 homeruns for roughly 72% of the data. Both of these numbers met the goal stated at the beginning of this report. In contrast, the Linear Regression model achieved an RMSE of 3.6 and an R^2 of 0.68. Indicating that the model was within 3.6 homeruns for roughly 68% of the data. The RMSE met the goal, but the R^2 did not.

These results were not surprising, as Random Forest is generally accepted to be a more accurate predictor than basic Linear Regression. The next section will explain the limitations of this study, and potential avenues to improve it.

5 Conclusion

In summary, the goal of this project was to explore new data and create meaningful machine learning insights from it. I chose to use historical baseball homerun data for my analysis. I used five predictors to predict the number of homeruns a player would hit in a year. I employed two different methods of prediction in this study, Random Forest modeling, and Linear Regression. My goal was to accurately predict homerun numbers for more than 70% of the dataset, with a Root Mean Square Error (RMSE) of less than 4. This ensures my methods were accurate, and had a reasonable margin of error.

The prior section of this report illustrated that Random Forest Modeling yielded a slight increase in accuracy over Linear Regression. However, the Random Forest model was constrained in development due to hardware limitations and the computationally intensive nature of the algorithm. Future efforts on this study should consider using a larger amount of trees, or tuning the model to find the optimal number of trees to maximize prediction accuracy. Future work may also include the use of parallel processing to speed up computation time on the Random Forest model, or adding additional predictors to enhance accuracy.