

Predicting Home Runs

Predicting the 2026 Detroit Tigers Home Runs



Jonathan Dela Cruz

12.05.2025

CSE482 Big Data Analysis

ABSTRACT

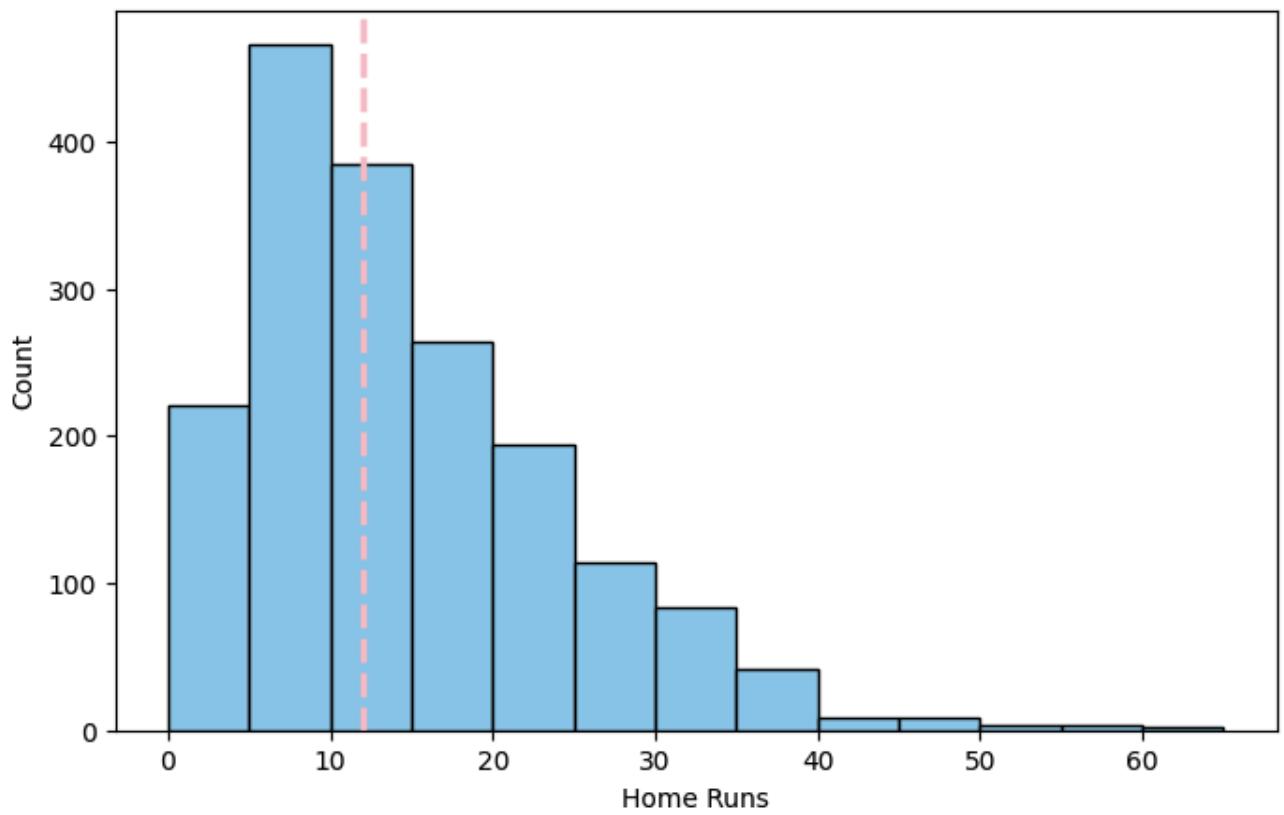
The 2025 Detroit Tigers demonstrated the potential of a young club by finishing with the 11th most home runs in franchise history and making it to the American League Divisional Series. In order to forecast each player's home run totals for the 2026 season, this project creates a machine learning model. Along with more advanced analytics like bat speed, launch angle, and exit velocity, the model also includes conventional statistics like home runs, batting average, and plate appearances. Predictions for 13 qualifying players on the 2026 roster were produced using these features, showing the value of integrating traditional and advanced analytics for player forecasting and offering insights into possible offensive performance.

INTRO

With the 11th-highest home run total in team history and a trip to the American League Divisional Series, the 2025 Detroit Tigers had an incredible season. The performance demonstrated the skill and potential of one of the league's youngest rosters, even though the team did not win the World Series. Accurately estimating individual contributions can assist in predicting team success and guide strategic choices, as many of the same players will be back in 2026.

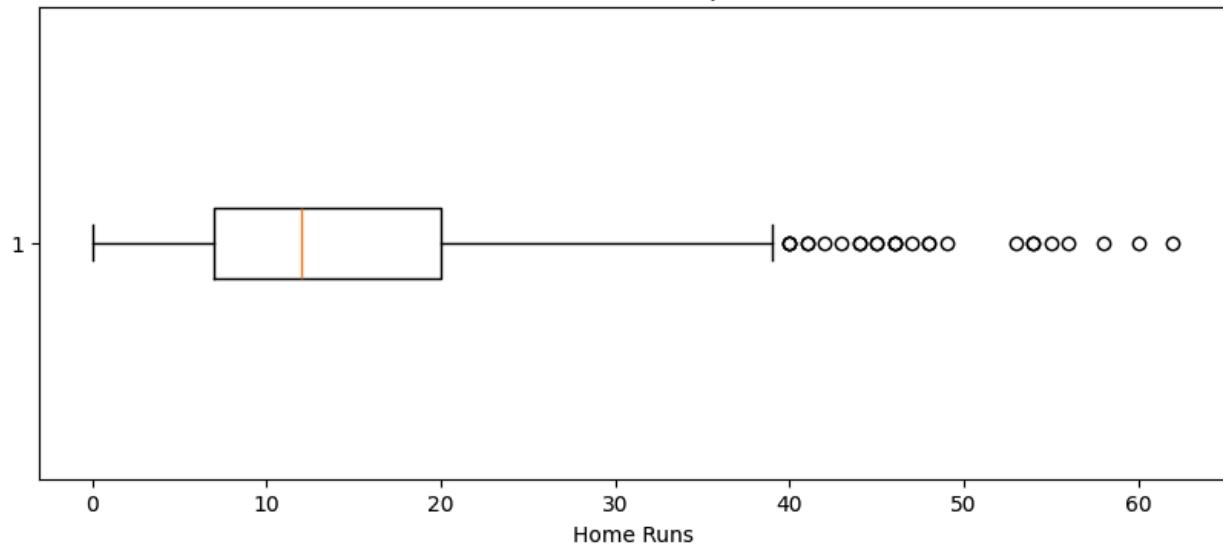
The goal of this research is to apply machine learning to forecast the Detroit Tigers' home run totals in 2026. The dataset I used came from the MLB's Baseball Savant website that contains a wide variety of every individual player's batting analytics. The model I built combines advanced analytics like bat speed, launch angle, and exit velocity with conventional hitting statistics like home runs, batting averages, and plate appearances. The algorithm attempts to generate more precise predictions of individual offensive production by fusing player-specific metrics with previous performance data. Thirteen qualifying players on the 2026 roster were predicted, providing information about the team's possible performance and demonstrating the importance of data-driven methods in sports analytics.

Home Run Distribution



Histogram of Home Runs Over the Last 5 Years (Median - Pink Line)

Home Runs Boxplot



Boxplot of Home Runs Over the Last 5 Years

METHOD

Imputation

During my Exploratory Data Analysis, I discovered that a few of my features had missing data for the 2021 and 2022 seasons. But since some of the missing variables had high correlation with the home runs variable, on top of the fact that the data from the 2023, 2024, and 2025 season did not have missing data, I decided that it would be best to impute the missing values of some of the features. I implemented a Random Forest Regressor that imputed the missing values present for the 2021 and 2022 seasons by taking into account the relationship between the missing variables and the known variables during the 2023, 2024, and 2025 seasons. I also included an imputed column into the dataframe, representing a boolean that indicates whether or not the data sample contains imputed values. The features that ended up being imputed were ‘average swing speed’, ‘fast swing rate’, ‘blasts contact’, and ‘blasts swing’.

Train/Test Split

Due to the project specifications, notably the fact that the model needed to be trained on one individual player’s sequential data, I could not use sklearn’s train/test split. Instead, I created a for loop, looping over all of the names in the data, that saved all of the indices of each player in an array, and appending all of the individual arrays into one large array containing the indices of each player. However, some players had missing indices for a variety of reasons; these reasons include injuries, insufficient data, and rookies in their first season. To combat this, I created another for loop that checked if a player had a year missing in its data. If a player was missing data for a given year, for whatever reason, then a placeholder index would be added for that year and the corresponding data would be all zeros. This way, the input for the model would be the same regardless of how many seasons a player played.

Modeling

In terms of the model, I decided to build a sequential model from the Keras and Tensorflow library with a LSTM layer, and two dense layers. I compiled the model with the mean squared error loss function and the Adam’s optimizer. I fit the model with my scaled training data across 15 epochs, using my validation data as the performance metric.

RESULTS

Validation Data			Predictions		
	Actual Home Runs	Predicted Home Runs	Detroit Tiger		Predicted Home Runs
0	15.0	21.0	0	Javier Báez	14.0
1	24.0	20.0	1	Kerry Carpenter	24.0
2	12.0	22.0	2	Dillon Dingler	16.0
3	12.0	27.0	3	Riley Greene	31.0
4	7.0	17.0	4	Colt Keith	16.0
5	6.0	14.0	5	Justyn Henry Malloy	12.0
6	10.0	13.0	6	Zach Mckinstry	13.0
7	7.0	11.0	7	Parker Meadows	11.0
8	21.0	32.0	8	Wenceel Pérez	14.0
9	6.0	19.0	9	Jake Rogers	11.0
10	17.0	17.0	10	Spencer Torkelson	31.0
11	11.0	14.0	11	Gleyber Torres	20.0
12	7.0	14.0	12	Matt Vierling	10.0
13	15.0	15.0			
14	3.0	17.0			
15	19.0	15.0			
16	0.0	12.0			
17	23.0	29.0			
18	11.0	13.0			
19	17.0	26.0			
20	7.0	12.0			
21	4.0	26.0			
22	11.0	13.0			
23	16.0	13.0			
24	5.0	13.0			
25	9.0	25.0			
26	7.0	12.0			
27	17.0	32.0			
28	18.0	13.0			
29	3.0	10.0			
30	11.0	27.0			
31	16.0	18.0			
32	2.0	11.0			
33	16.0	27.0			
34	27.0	19.0			
35	9.0	15.0			

RESULTS ANALYSIS

The model performed fairly well when predicting the home runs for the validation set. In its best run, the model only incurred about an 85 mean squared error across 36 samples. However, the model failed to predict any single digit home run totals for anyone, leading to a very biased model. Since the project is predicting the upcoming home runs totals for the Detroit Tigers, there is not an accurate way to measure the quality of the model in terms of the testing data. But from the perspective of a fan who has been following the Detroit Tigers for the past decade, I would conclude that these predictions are optimistic, but overall reasonable.

LESSONS LEARNED

1. Imputing missing values is computationally expensive, so weighing the cost to reward of imputing these values should be taken into consideration.
2. Using sequential data poses a challenge to model training, where you have to organize the data in such a way that the model is taking in data in the correct order.
3. There are a ton of variables that contribute to an individual player's success outside of the conventional and advanced statistics. While I tried to account for these variables, it is extremely difficult to account for all of these variables in an efficient manner.

FUTURE IMPROVEMENT

There are a ton of different avenues I would like to explore to see improvements in my predictions. Some of the first methods that come to mind are changing the model parameters to optimize the results. While I experimented with the different number of epochs and neurons per layer, I did not see a drastic change in performance. Iterating across a range of different values for these parameters, while computationally expensive, could find the optimal values to achieve the best results. I would also like to see how adding features indicating different events that pop up during the course of a season would impact the model predictions. Events such as injuries or insufficient data due to performance-related reasons that could only be observed from the human eye could assist the model in its predictions.