# Predicting HRs

*Hakeem Yatim, Mitsuki Shimonishi*
*{Hakeem.Yatim,Mitsuki.Shimonishi}@trojans.dsu.edu*


Figure 1. Illustrate your task and/or approach

**1. Task**

Describe the task you are going to solve, e.g. generating natural images like those shown in Fig. 1. Say a few words about what is difficult about the task, e.g. preserving spatial relationships, etc.

We are trying to predict homeruns of different MLB players for the year of 2022 using a predictive model. What will be difficult about the task is cleaning the data up by removing useless data and improving the accuracy of the model itself after building and training the model using training and test datasets. Cleaning up the data will be tough because there are many different variables with datatypes that are not number based so we have to filter them out, while scaling the model to improve accuracy can be tricky also.

**2. Approach**

Outline the approach you took in detail, using diagrams, equations and defining all notation. Do not assume the reader is familiar with the paper you are implementing! If you re-used existing libraries or implementations, say which ones, and say what additional code you wrote for the project.

We found out that python has a python package called pybaseball that lets us download baseball stats from various websites, and we want to download stats from 2002 to 2022. We then want to focus on the HR stats alone, so we need to create an extra column in the data as “Next_HR” so that we can use it later to compare with the predictions that the model use.

After this, we then will clean up the data by dealing with missing values and deleting columns that will give our model some trouble later. We also do not want to use columns with data with a non-number datatype. We would also either delete it or categorized them and convert them into numbers.

Then, we decided to use a ridge regression model, sequential feature selector for feature selection, and time series split for model selection to build the model and use a minmax scaler to scale the data (and making sure we don’t scale our target data, which is HR’s). We then want to use backtest to generate our prediction for us (similar with cross validation but since we want to use past data only while cross validation uses future data too).

We then make predictions starting from the year 2007, using the 2002-2006 dataset as training data and predict 2007 data, then use 2002-2007 dataset as training data and predict 2008 data, and so on. We did 2002-2006 initially as a training dataset because we want to have a big enough historical dataset to improve the data’s accuracy. We then compare the predictions with the test dataset (which is our target data Next_HR). We then want to improve the accuracy of the model by considering different variables of why someone might have a significantly lower or higher HR’s the next season(shortened season, injury, etc.).

**3. Dataset**

Say which dataset you used, how many training and test examples, and describe any data pre-processing that was done. The dataset must contain sufficient labels if your project needs them, typical machine learning datasets have 10K-1M examples.

We used data from the python package pybaseball and removed columns that has null values and other columns that we don’t need for our prediction. As mentioned before, we wanted to focus on columns with number datatypes (float or double) so we had to find them and remove the other columns before using it for machine learning. We then used around 16 training data sets (since we started with the 2002-2006 season as training data all the way up to 2021) and also used 16 test datasets to compare with the predictions created with the training datasets.

## 4. Results

Describe your experiments evaluating your approach. Define the metric(s) you used, e.g. we used classification accuracy, defined as … State explicitly how you measured success, e.g. "Our results in Table 1 show that our new loss function improved accuracy on the test set of dataset X compared to the method in [4]," etc. Describe all hyper-parameter settings, and analyse the results in detail. What were the lessons learned?

Out[177]:

| | IDfg | Season | Name | HR | Next_HR | prediction | diff |
|---|---|---|---|---|---|---|---|
| **4292** | 3790 | 2009 | Jayson Nix | 12 | 14.0 | 13.998627 | 0.001373 |
| **5045** | 157 | 2009 | Melvin Mora | 8 | 7.0 | 6.998262 | 0.001738 |
| **5521** | 10 | 2009 | David Eckstein | 2 | 1.0 | 0.996646 | 0.003354 |
| **6596** | 4400 | 2014 | Chris Denorfia | 3 | 3.0 | 2.994689 | 0.005311 |
| **5171** | 12434 | 2017 | Kevin Pillar | 16 | 15.0 | 15.011800 | 0.011800 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **1584** | 14221 | 2018 | Jorge Soler | 9 | 48.0 | 17.364781 | 30.635219 |
| **5386** | 12533 | 2020 | Marcus Semien | 7 | 45.0 | 13.176496 | 31.823504 |
| **2156** | 4949 | 2016 | Giancarlo Stanton | 27 | 59.0 | 26.248720 | 32.751280 |
| **451** | 15640 | 2021 | Aaron Judge | 39 | 62.0 | 29.228941 | 32.771059 |
| **2596** | 1887 | 2009 | Jose Bautista | 13 | 54.0 | 17.037484 | 36.962516 |

4127 rows × 7 columns

The result of our model is above. One some players our prediction was very close to being accurate while some other players our results were way off the actual HR's. this is due to other variables that can affect the player's performance in general. Things like injuries, a shortened season due to COVID, or even a breakout year did impact these numbers since those variables can cause a drastic change within the players' numbers also. We also used the correlation between HR stats with the player season and compare the individual averages with the group averages per season to improve our model's accuracy but we still cannot take out the outliers of the dataset (and there are lots of them since there are many players who has drastic changes of HR numbers on certain periods of their career). We also used the mean squared error to measure accuracy.

The lesson learned here is that HR's are numbers that can have a linear correlation but can also have many outliers that makes the data less correlated. There are certain players that show trends and if looked into further we can determine why the trend exist (due to age for example) but there are other variables that affects the statistic that makes us not able to make very accurate predictions correctly. The next steps for this project if continued further would be to look at individual predictions and research why players perform better or worse on certain years and put that into consideration for the model. This way, we can deal with the outliers more and make a better prediction model overall.

## 5. Detailed Roles

Fill in the table below for each teammate.

| Task | File Names | Lead |
|---|---|---|
| Research baseball statistics with python | - | Mitsuki |
| Research predictive models for baseball statistics | - | Hakeem |
| Building model on jupyter notebook (making training dataset, predicting numbers, improving model's accuracy) | GroupProject.ipynb | Both |
| Report Section 1 and 3 | Project Final Report Template.docx | Mitsuki |
| Report Section 2 and 4 | Project Final Report Template.docx | Hakeem |