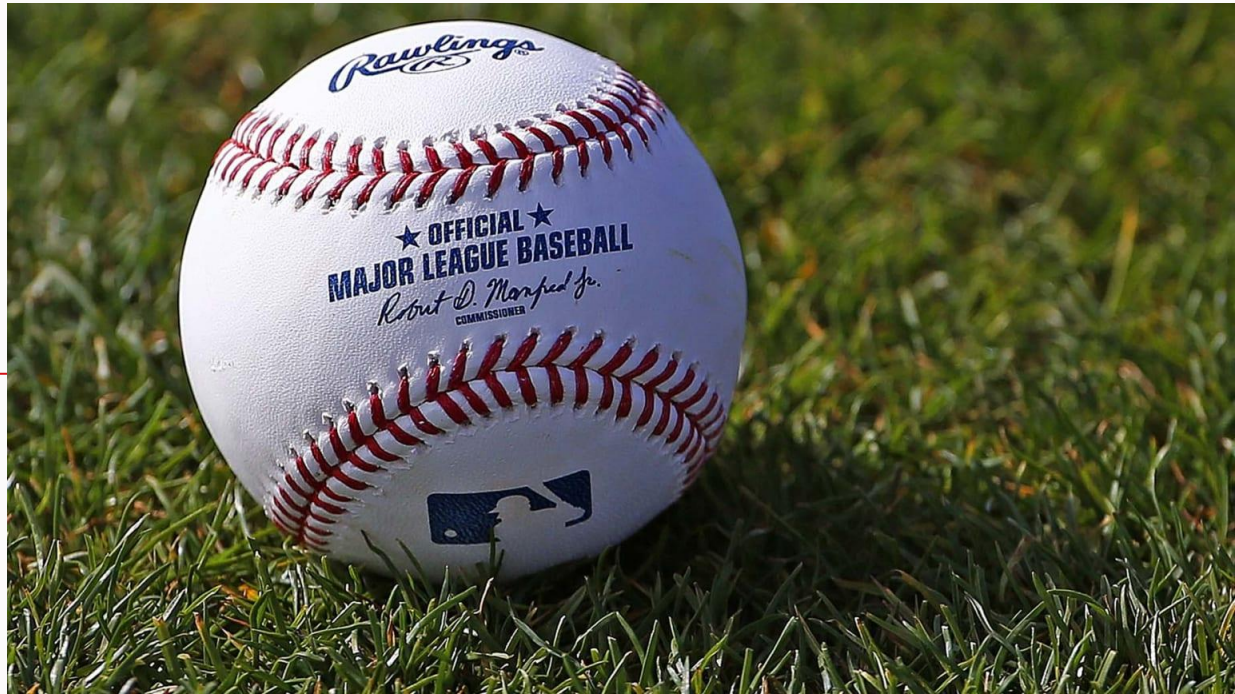
STAT155 Final Presentation



Sui Sakai

06/06/2025

Agenda

01 Introduction

02 Project I: Data Wrangling

03 Project II: Exploratory Data Analysis

04 Project III: Modeling

05 Project IV: Monte Carlo Simulation

06 Summary and Reflection

Introduction

Research Question:

Can we predict an MLB (Major League Baseball) player's performance/worth in a given season based on their previous performance stats and other metrics?

Context

We are predicting WAR (Wins Above Replacement) which essentially quantifies a player's value by measuring how many additional wins they contribute to their team compared to the average replacement-level player.

We are looking at only batter data (no pitchers)

Data Wrangling

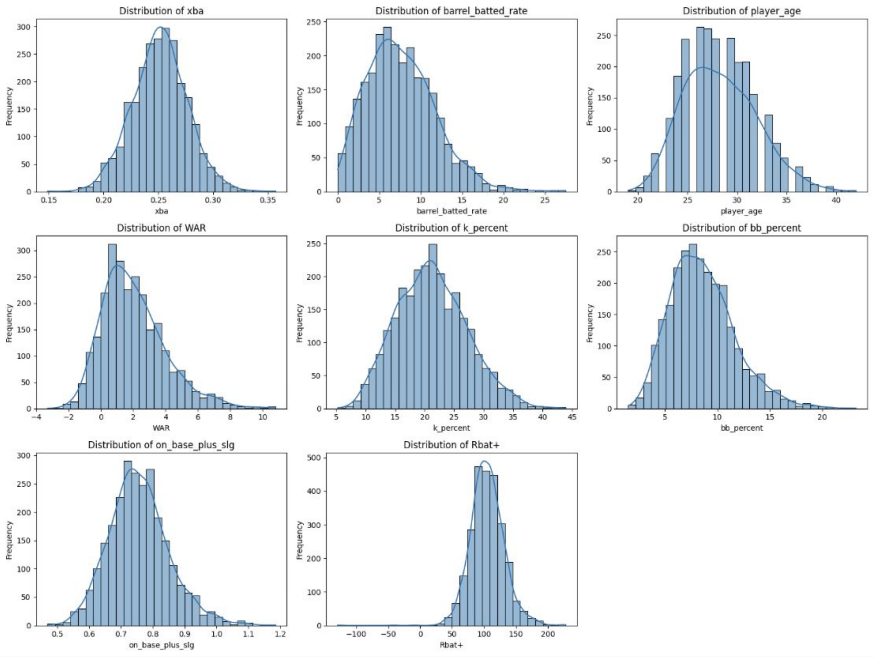
Sources:

- Baseball Reference https://www.baseball-reference.com/
- Statcast https://baseballsavant.mlb.com/statcast_search

Data Wrangling:

Some of the statistical features I wanted to use were only in one of the datasets while others were in the other so I had to merge the two. Furthermore, Baseball Reference only gave me a csv file per year while Statcast could give me all the combined data at once so I had to merge by player name + season year.

Exploratory Data Analysis

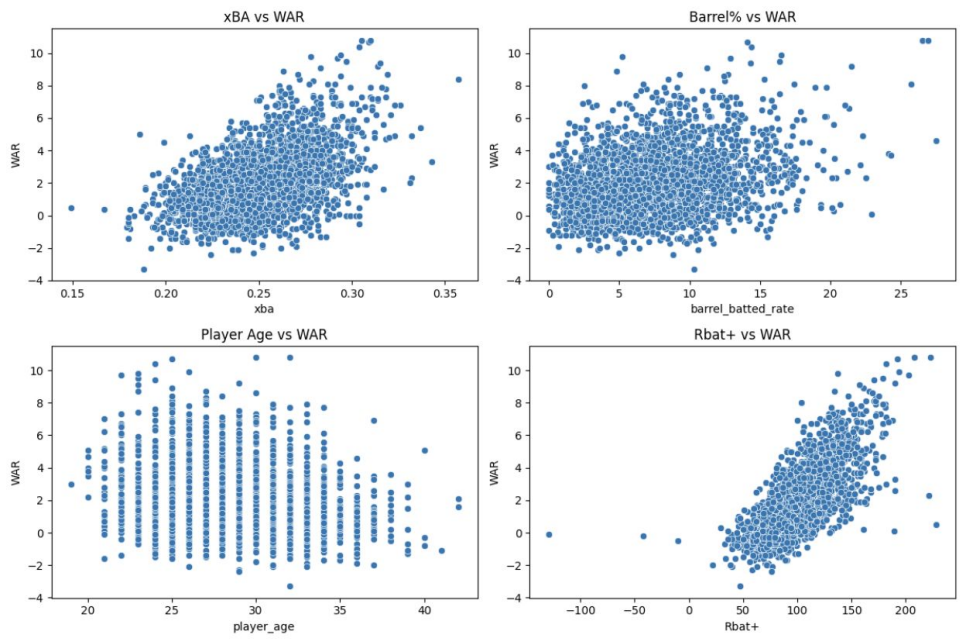


Univariate Analysis:

Out of the 30+ columns I had, I hand-picked 8 that I thought were the most important for WAR.

We can see that all the variables follow a somewhat normal distribution.

Continued



Bivariate Analysis:

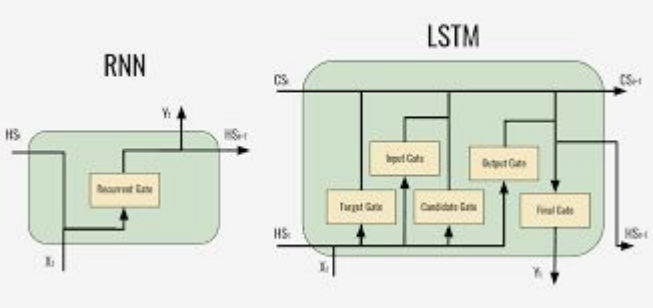
Putting some key features vs a player's WAR, we can see that Rbat+ has the most correlation with WAR , which makes sense because Rbat+ measures offensive run contributions, which directly affects a player's value.

Player age has very little to no correlation with WAR. This is likely due to baseball being a sport where longevity is much more of a norm compared to other sports.

Modeling

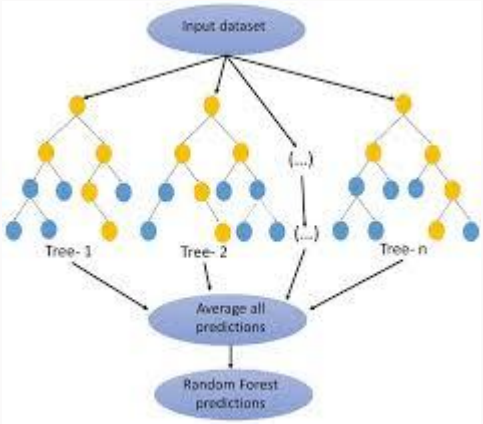
Recurrent Neural Network (LSTM):

The first model I considered was the Long Short-Term Memory (LSTM) model. This takes into account previous data points to learn sequential data. I thought this might be a good model for my data because each player plays through multiple seasons, representing a sequence.



Random Forest:

The Random Forest Regressor creates a bunch of decision trees and outputs the average of the predictions of them.



Continued

Recurrent Neural Network (LSTM):

For LSTM, I got a RMSE of around 1.7 and r^2 of around 0.2. These are not the best numbers but considering that a player's WAR can be very unpredictable it is not the worst either..

Random Forest:

For Random Forest, I got a RMSE of around 1.2 and r^2 of around 0.6. These numbers are significantly better than the previous 1.7 and 0.2 of the LSTM, suggesting that Random Forest is the better model for this project than RNN/LSTM. 1.2 is still fairly high considering that WAR is usually within the range [-2, 10] but with how real world sports can be, 1.2 is not a bad error.

Observation: We can increase r^2 to up to around 0.65 and as high as 0.7 just by changing the max_depth to 12 instead of 10.

Monte Carlo Simulation

	Model	Correlation	Mean_MSE	SD_MSE	Mean_R2	SD_R2
0	Linear	No Corr	1.266972	0.291985	0.833440	0.067262
1	Linear	Mild Corr	1.219451	0.300025	0.871650	0.041114
2	Linear	High Corr	1.246117	0.296410	0.897093	0.035182
3	RandomForest	No Corr	2.133040	0.740327	0.732203	0.101382
4	RandomForest	Mild Corr	1.873765	0.460347	0.806304	0.040188
5	RandomForest	High Corr	1.498897	0.803587	0.882930	0.042562
6	Ridge	No Corr	1.266027	0.290804	0.833592	0.067104
7	Ridge	Mild Corr	1.218260	0.299633	0.871765	0.041087
8	Ridge	High Corr	1.196411	0.279257	0.901034	0.034459

Data Generation:

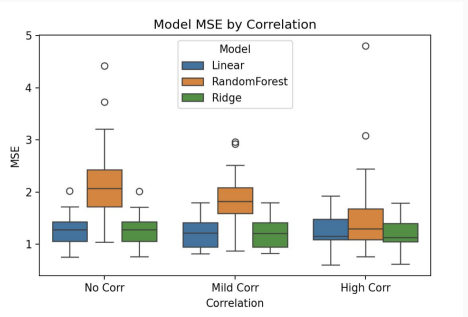
Using dataset_name.describe(), I got the means and standard deviations for all my data features which I then used to generate synthetic data from normal distributions with those numbers. 30 datasets are generated, 10 for each correlation level (0, 0.5, 0.99), with 100 samples in each dataset.

Modeling:

Each dataset is then split into training and test data (70% and 30% respectively) and standardized. Then we train and test each model (Random Forest, Linear Regression, and Ridge Regression) and compute the MSE and R^2

Findings:

Ridge Regression performs slightly better than Linear Regression due to its regularization. Random Forest performs the worst but drastically increases accuracy as correlation increases.



Summary

Revisit:

Can we predict an MLB (Major League Baseball) player's performance/worth in a given season based on their previous performance stats and other metrics?

Final Verdict:

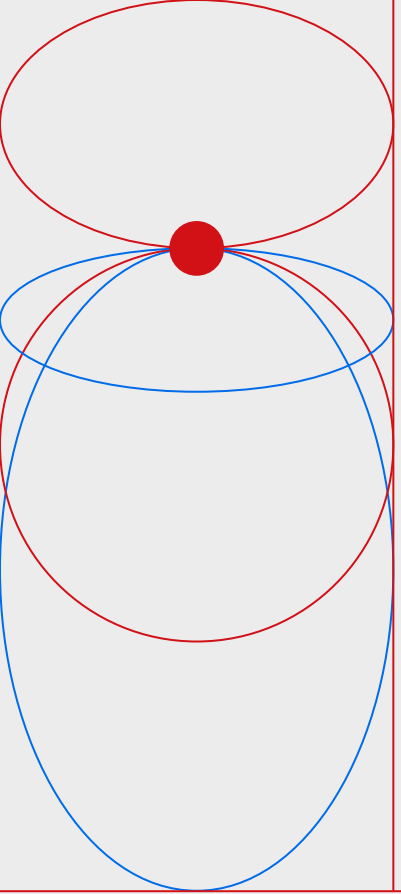
It is possible to get a fairly accurate prediction of a player's WAR given their previous stats but in real life there are so many factors including but not limited to: injuries, off-the-field matters, nerves, slumps, etc. which make it hard to accurately predict how a player would perform.

Reflection

Through this project, I was able to learn how to clean real-world data and test models to determine what the best model is for the situation. I also learned the importance of making any work reproducible and how much of a hassle that can be, both from my own work and from looking at the work of others.

Continued

The use of data in sports has been on the rise in the last few decades, and there is no slowing down. The rise is in part due to the success that teams have had using data to their advantage. *Moneyball* (2011) is one example of this where the 2002 Oakland Athletics used data to construct their roster with underrated players and had lots of success with it. Data can change how teams evaluate players and sometimes even how the entire sport is played.



Thank you