Baseball Player Investment

How to correctly import foreign hitters from MLB to NPB?

Presenter: Ying Wei Hung

Date: 25/10/21

Agenda

Introduction

Project Intention
League difference
Project goal

Data Collection

Method to collect the data

Approach

Clustering Analysis &
Statistical Prediction

Results

What do the models tell us?

Conclusion

Insights for the ball clubs

Limitations

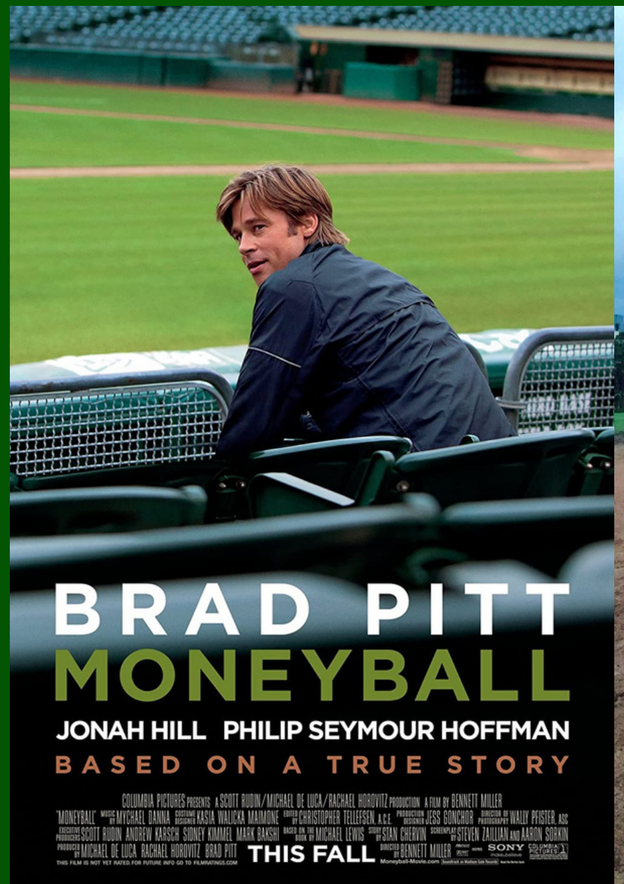
Constraints for the analysis

Introduction

- Intention
- Difference between MLB & NPB
- Goal

Intention

- Watching and playing baseball games for over 12 years
- Inspired by MoneyBall movie



MLB

30 teams
National League
American League
Power pitchers
Power hitters

NPB

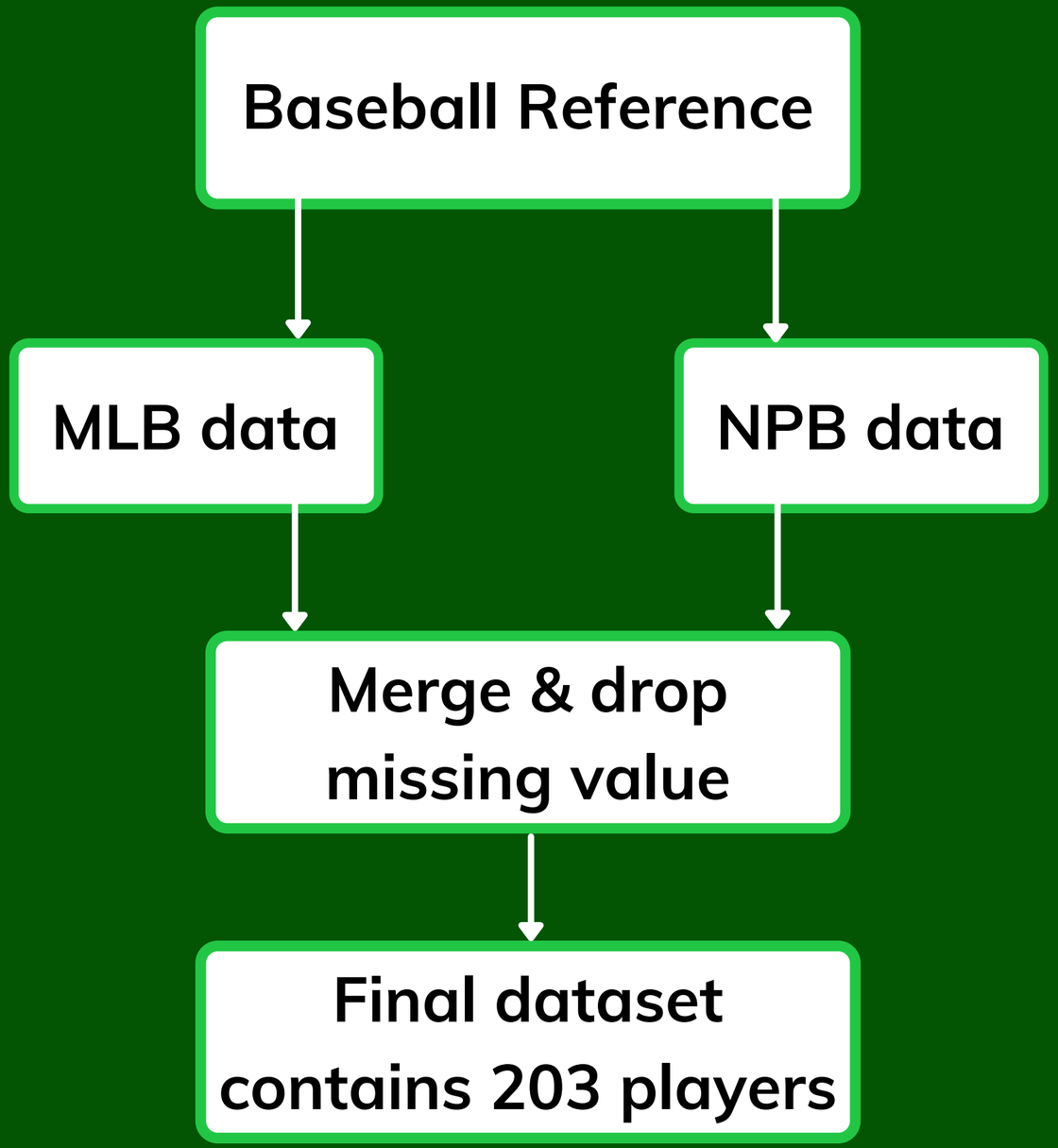
12 teams
Central League
Pacific League
Delicate pitchers
Efficient hitters

Project Goal

Assist NPB baseball club front office to invest the right foreign players

Data Collection

How do I collect my data
Scope of the data



Where I collect my data?

Baseball reference.com

Who I focus on?

Hitters who have played in MLB and NPB

What time range is included?

from 2000 to 2020

What statistics I choose?

30 variables for hitting performance
BA, OBP, SLG, OPS, SB%, BB%, K%.....

Approach

Clustering Analysis & Regression Prediction

Clustering Analysis (K Means)

Find trends in the data (NPB)

Linear Regression

Predict 16 response variables (numerical) by using multiple exploratory variables

---

Lasso Regression

Predict 16 response variables (numerical) by using multiple exploratory variables

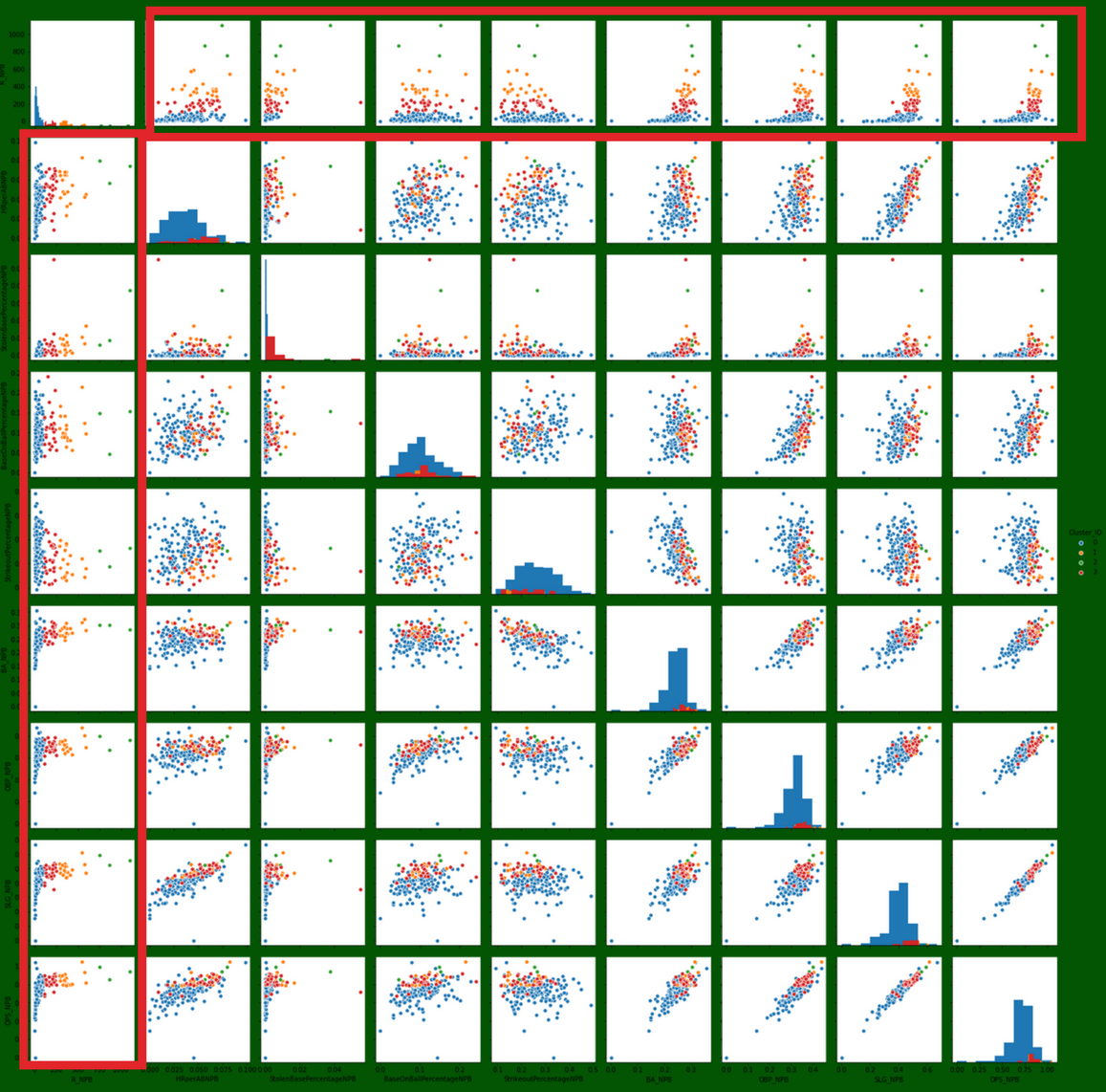
Why Lasso Regression

Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.

Results

Clustering Analysis for NPB

1. The majority of clusters mess up together
2. Only the first row and column have clear boundaries
3. Notice some linear relationships between variables



Results

Linear & Lasso Regression

1. The majority of the R-squared values are low
2. The highlighted parts are the interesting variables
3. Strikeout percentage is the highest value

R-squared	Linear Model	Quadratic Model	Lasso Model
BA	3.56%	0.29%	0.00%
OBP	4.31%	2.51%	0.00%
SLG	11.23%	2.26%	0.00%
OPS	10.27%	2.42%	0.00%
RBI	3.75%	0.49%	0.00%
SB%	4.95%	0.31%	0.68%
BB%	23.56%	3.76%	29.26%
K%	44.48%	4.65%	45.25%
2B per AB	3.88%	0.28%	0.00%
3B per AB	14.02%	0.11%	11%
HR per AB	22.65%	3.07%	25.71%
Game	5.02%	0.57%	0.00%
PA	4.61%	0.36%	0.00%
AB	4.96%	0.36%	0.00%
R	3.57%	0.35%	0.00%
CS	6.41%	0.08%	1.74%

The R-Squared values of each model are shown

Results

Linear & Lasso Regression Significant Coefficients

For K% NPB

Strikeouts + OPS +3B per AB + HR per AB + BB% + K% + GDP + SH

For BB% NPB

CS+ BB + HBP + HR per AB + BB% + K% + H + 2B + OBP + SH + IBB + 2B per AB + 3B per AB

For HR per AB NPB

HR + CS+ SF + HR per AB + BB% + K% + G + 2B + 3B + OBP + HBP +SH + IBB + 2B per AB + 3B per AB

Conclusion (Insights)

Insights

In NPB,
there is one group of players who can get more runs earned.
In general, they have high
HR per AB, Stolenbase% , BB%, BA, OBP, SLG, and OPS.

Suggestion:
Front office should investigate why these players can survive in NPB.
For example: Past experience, age, race, etc.

Insights

Lasso model is more likely to predict strikeout percentage in NPB based on MLB performance.

Based on the coefficients,
the higher the strikeouts, OPS, 3B per AB, HR per AB, BB%, and K%,
the more likely to get more strikeouts.
The type of player tends to be Power Hitter.

Suggestion: If the team is not good at scoring runs, the team can consider to invest these players.

Limitations

Limitations

01

League difference

1. Different playing style
2. Different coaching style
3. Different ballparks factor
   - pitcher or batter favor
   - weather factor

02

Lack of scout perspectives

1. Long term observations
2. Potential injuries prediction
3. and more...

03

Imbalanced data

1. Different game played
2. Different PA
3. Different AB


Thanks for listening
Feel free to give me
feedbacks