

# 미니 프로젝트 타자의 지표와 특점의 상관관계 분석

MLB BATTER  
DATA



**야구는 빠따조**

---

KIM MINSEOP  
RYU AIN  
LEE HWAN HEE  
PARK HONGGEUN

# INDEX

1. 팀원 소개
2. 주제 소개 및 선정 이유
3. 목표
4. WBS
5. 데이터 정의서
6. 머신 러닝 소스코드
7. 그래프
8. 웹 페이지
9. 결론
10. 아쉬운 점

## 팀원 소개



Mr. Kim



Mrs. Lee



Mr. Park



Mrs. Ryu

## 주제 소개 및 선정 이유

주제 :

타자의 지표와 득점의 상관관계 분석

선정 이유 :

통계의 대표적인 스포츠인 야구 분석을  
통해 다양한 예측을 해보고자 함.

## 목표

- 타자 지표와 승률과의 관계
- 투수 지표와 승률과의 관계
- **타자의 각 지표에 따른 득점의 총 합 예측**
- 투수의 각 지표에 따른 평균 방어율의 예측
- **올스타 팀 선정 및 해당 팀의 득점의 총 합 예측**

# WBS

[illegible]

## 데이터 소개

- MLB 2010~2019년 각 연도 별 타자 지표
- MLB 2010~2019년 각 연도 별 투수 지표
- MLB 2010~2019년 각 연도 별 시즌 팀 랭킹

## 데이터 정의서

- 투수 지표

투수 컬럼										
Player	Team	Age	G	GS	CG	SHO	IP	H	ER	K
선수명	팀	나이	게임수	선발출전	완투 수	완봉 수	이닝	피안타	자책점	삼진

BB	HR	W	L	SV	BS	HLD	ERA	WHIP
볼넷	피홈런	승	패	세이브	블론세이브	홀드	평균자책점	이닝당 출루허용률

- 타자 지표


타자 컬럼											
Player	Team	Pos	Age	G	AB	R	H	2B	3B	HR	RBI
선수명	팀	포지션	나이	게임수	타수	득점	안타	2루타	3루타	홈런	타점


SB	CS	BB	SO	SH	SF	HBP	AVG	OBP	SLG	OPS
도루성공횟수	도루실패횟수	볼넷	삼진	희생안타	희생플라이	데드볼	타율	출루율	장타율	(출루율+장타율)





## 데이터 소개


전체 폴더


 KBO\_data


 KBO\_wOBA


 MLB\_data


 MLB\_data\_edit

 MLB\_Season\_Ranking

 MLB\_TRADE

 MLB\_TRADE\_EDIT

 MLB\_wOBA

 ranking\_player

2010 ~ 2019년

KBO 투수, 타자 자료

 2010\_bat.csv

 2010\_pit.csv

 2011\_bat.csv

 2011\_pit.csv

 2011\_pit\_edit.csv

 2012\_bat.csv

 2012\_pit.csv

 2013\_bat.csv

 2013\_pit.csv

 2014\_bat.csv

 2014\_pit.csv

 2015\_bat.csv

 2015\_pit.csv

 2016\_bat.csv

 2016\_pit.csv

 2016\_pit\_edit.csv

 2017\_bat.csv

 2017\_pit.csv

 2017\_pit\_edit.csv

 2018\_bat.csv


 2018\_pit.csv



 2019\_bat.csv


 2019\_pit.csv


## 데이터 소개


전체 폴더


 KBO\_data


 KBO\_wOBA 


 MLB\_data


 MLB\_data\_edit

 MLB\_Season\_Ranking

 MLB\_TRADE

 MLB\_TRADE\_EDIT

 MLB\_wOBA

 ranking\_player

2010 ~ 2019년

KBO 투수 wOBA 자료

 KBO\_wOBA10.csv

 KBO\_wOBA11.csv

 KBO\_wOBA12.csv

 KBO\_wOBA13.csv

 KBO\_wOBA14.csv

 KBO\_wOBA15.csv

 KBO\_wOBA16.csv

 KBO\_wOBA17.csv

 KBO\_wOBA18.csv

 KBO\_wOBA19.csv


Total


KBO 투수 wOBA 자료


 **KBO\_wOBA\_total.csv**


## 데이터 소개


전체 폴더


 KBO\_data


 KBO\_wOBA


 MLB\_data


 MLB\_data\_edit

 MLB\_Season\_Ranking

 MLB\_TRADE

 MLB\_TRADE\_EDIT

 MLB\_wOBA

 ranking\_player

2010 ~ 2019년

MLB 투수, 타자 자료

 2010\_bat.csv

 2010\_pit.csv

 2011\_bat.csv

 2011\_pit.csv

 2011\_pit\_edit.csv

 2012\_bat.csv

 2012\_pit.csv

 2013\_bat.csv

 2013\_pit.csv

 2014\_bat.csv

 2014\_pit.csv

 2015\_bat.csv

 2015\_pit.csv

 2016\_bat.csv

 2016\_pit.csv

 2016\_pit\_edit.csv

 2017\_bat.csv

 2017\_pit.csv

 2017\_pit\_edit.csv

 2018\_bat.csv


 2018\_pit.csv


 2019\_bat.csv


 2019\_pit.csv


## 데이터 소개


전체 폴더


 KBO\_data


 KBO\_wOBA


 MLB\_data


 MLB\_data\_edit

 MLB\_Season\_Ranking

 MLB\_TRADE


 MLB\_TRADE\_EDIT


 MLB\_wOBA


 ranking\_player


2010 ~ 2019년


MLB 투수, 타자 자료를  
데이터 전처리 후 만들어진  
파일들


 All\_star\_final.csv


 b\_total\_no\_scale.csv


 bat\_r.csv


 bat\_total.csv


 bat\_total\_2010\_2018.csv


 bat\_total\_2019.csv


 p\_total\_no\_scale.csv


 pit\_total.csv


 Precision.csv

 r\_2010\_2018.csv

 r\_2019.csv

 rank\_good.csv

 real\_bat\_total.csv


 real\_pit\_total.csv


2010 ~ 2018년

MLB 타자 자료로 학습 시켜

2019년


MLB 타자 자료와 일치하는지 확인

 bat\_total\_2010\_2018.csv

 bat\_total\_2019.csv


2019년

MLB 타자 득점 수의 합  
예측한 값

 Precision.csv










각 포지션 별

최고 선수들

 All\_star\_final.csv











## 데이터 소개

전체 폴더

-  KBO\_data
-  KBO\_wOBA
-  MLB\_data
-  MLB\_data\_edit
-  MLB\_Season\_Ranking
-  MLB\_TRADE
-  MLB\_TRADE\_EDIT
-  MLB\_wOBA
-  ranking\_player

2010 ~ 2019년

MLB 시즌 랭킹

-  2010\_ranking.csv
-  2011\_ranking.csv
-  2012\_ranking.csv
-  2013\_ranking.csv
-  2014\_ranking.csv
-  2015\_ranking.csv
-  2016\_ranking.csv
-  2017\_ranking.csv
-  2018\_ranking.csv
-  2019\_ranking.csv

2010 ~ 2019년











MLB 시즌 랭킹

Total

-  concat\_rank.csv











## 데이터 소개

전체 폴더

-  KBO\_data
-  KBO\_wOBA
-  MLB\_data
-  MLB\_data\_edit
-  MLB\_Season\_Ranking
-  MLB\_TRADE 
-  MLB\_TRADE\_EDIT
-  MLB\_wOBA
-  ranking\_player










2010 ~ 2019년

MLB FA 선수들 리스트

-  2010\_2011\_FA.csv
-  2011\_2012\_FA.csv
-  2012\_2013\_FA.csv
-  2013\_2014\_FA.csv
-  2014\_2015\_FA.csv
-  2015\_2016\_FA.csv
-  2016\_2017\_FA.csv
-  2017\_2018\_FA.csv
-  2018\_2019\_FA.csv
-  2019\_2020\_FA.csv

## 데이터 소개

전체 폴더

-  KBO\_data
-  KBO\_wOBA
-  MLB\_data
-  MLB\_data\_edit
-  MLB\_Season\_Ranking
-  MLB\_TRADE
-  MLB\_TRADE\_EDIT
-  MLB\_wOBA
-  ranking\_player







2010 ~ 2019년

MLB FA 선수들 리스트를

전 처리한 후 만들어진 데이터










선수 영입 데이터 전체 파일

선수 영입 데이터 NA값 없는 것











-  trade\_batter\_final.csv
-  trade\_batter\_notna.csv
-  trade\_pitcher\_final.csv
-  trade\_pitcher\_notna.csv
-  trade\_total\_FA.csv
-  trade\_total\_final.csv

## 데이터 소개




전체 폴더

-  KBO\_data
-  KBO\_wOBA
-  MLB\_data
-  MLB\_data\_edit
-  MLB\_Season\_Ranking
-  MLB\_TRADE
-  MLB\_TRADE\_EDIT
-  MLB\_wOBA
-  ranking\_player

2010 ~ 2019년  
MLB wOBA 데이터

-  wOBA\_10.csv
-  wOBA\_11.csv
-  wOBA\_12.csv
-  wOBA\_13.csv
-  wOBA\_14.csv
-  wOBA\_15.csv
-  wOBA\_16.csv
-  wOBA\_17.csv
-  wOBA\_18.csv
-  wOBA\_19.csv

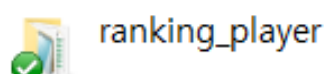
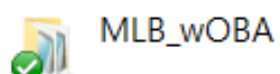
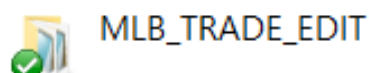
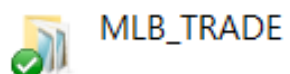
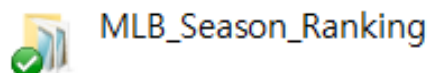
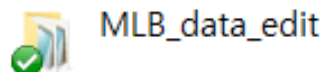
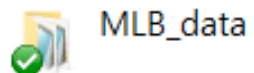
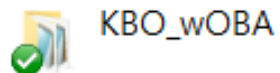
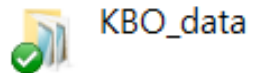
2010 ~ 2019년  
MLB wOBA 데이터 Total

-  MLB\_wOBA\_final.csv
-  MLB\_wOBA\_final\_csv
-  MLB\_wOBA\_total.csv



## 데이터 소개

전체 폴더



2010 ~ 2019년

MLB 선수 랭킹 별 자료



## 머신 러닝 소스코드

```
# 학습용 데이터
from sklearn import datasets
# 데이터를 학습용과 테스트용으로 나눌 수 있는 함수
from sklearn.model_selection import train_test_split
# 데이터 표준화
from sklearn.preprocessing import StandardScaler
# Perceptron 머신 러닝을 위한 클래스
from sklearn.linear_model import Perceptron
# 정확도 계산을 위한 함수
from sklearn.metrics import accuracy_score
# 파일 저장을 위해..
import pickle
import numpy as np
from sklearn import preprocessing
from sklearn import linear_model
```

## 머신 러닝 소스코드

```
#=====step1
# 전체 타자 데이터
bat_total=pd.read_csv('/Users/admin/Dropbox/Rank_Predict/Data/MLB_data_edit/real_bat_total.csv')
# 전체 투수 데이터
pit_total=pd.read_csv('/Users/admin/Dropbox/Rank_Predict/Data/MLB_data_edit/real_pit_total.csv')
# 전체 랭크 데이터
rank_total=pd.read_csv('/Users/admin/Dropbox/Rank_Predict/Data/MLB_data_edit/rank_good.csv')
# 10년치 팀 평균(타자)
# R, 2B, HR, RBI, BB, SF, OBP, SLG, OPS 사용
b_total=pd.read_csv('/Users/admin/Dropbox/Rank_Predict/Data/MLB_data_edit/b_total_no_scale.csv')
# 10년치 팀 평균(투수)

# 타자, 투수 지표
b_corr=b_total[['R','H','2B','HR','RBI','SF','OPS','BB']]

# 승률
rank_PCT=r[['R']]

# y = 승률, X = 투수지표
y=rank_PCT; X=b_corr
```

## 머신 러닝 소스코드

```
# =====step2

# train 데이터와 test 데이터 정의 및 분리
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,shuffle=False)

# 표준화 함수 정의
sc = preprocessing.StandardScaler()
# 데이터를 표준화
sc.fit(X_train)
X_train_std = sc.transform(X_train)
sc.fit(X_test)
X_test_std = sc.transform(X_test)

clf_ = linear_model.Ridge(alpha=0.5)

# 학습
print(X_train.R)
clf_.fit(X_train, y_train)
y_pred = clf_.predict(X_test)

# 학습정확도 확인
print('Mean Squared Error :',mean_squared_error(y_test,y_pred))
print('Mean Absolute Error :',mean_absolute_error(y_test,y_pred))
print("train 학습 정확도 :", clf_.score(X_train, y_train))
print("test 학습 정확도 :", clf_.score(X_test, y_test))
print(clf_.coef_)
print(clf_.intercept_)

with open('./bs.dat', 'wb') as fp:
    pickle.dump(sc, fp)
    pickle.dump(clf_, fp)
```

## 머신 러닝 소스코드

```
# =====step3
with open('./bs.dat', 'rb') as fp:
    sc = pickle.load(fp)
    clf_ = pickle.load(fp)

# X = []
X = [[97.100000, 151.100000, 32.300000, 35.500000, 101.000000, 5.100000, 0.965600, 75.900000]]

# R          97.100000
# H          151.100000
# 2B         32.300000
# HR         35.500000
# RBI        101.000000
# SF         5.100000
# OPS        0.965600
# BB         75.900000

X_std = sc.transform(X)
# 결과를 추출한다

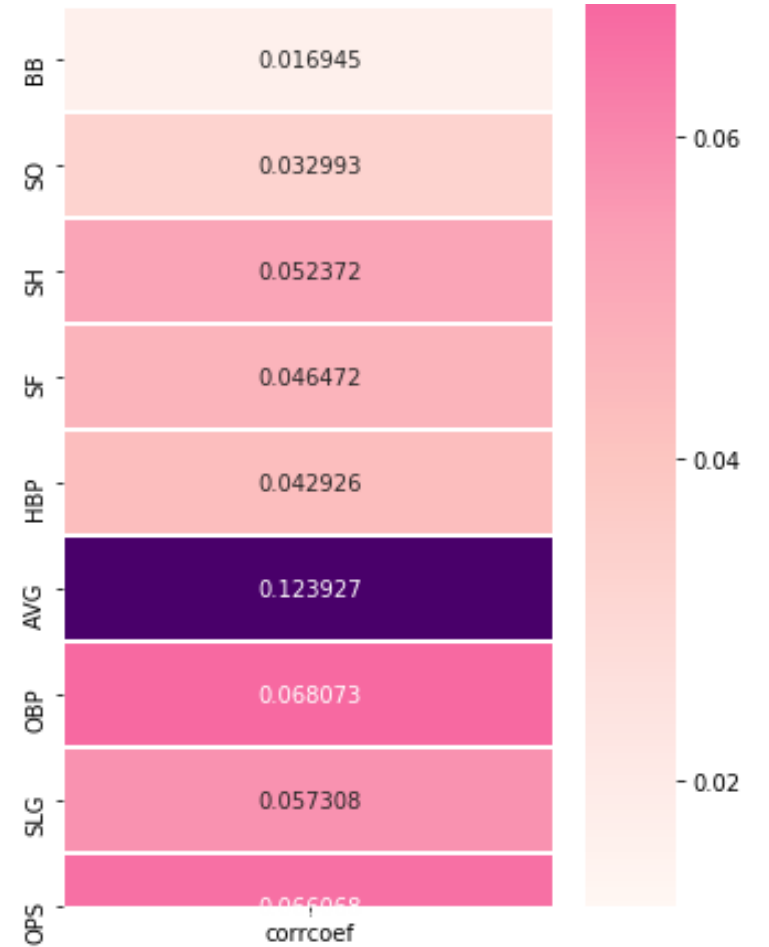
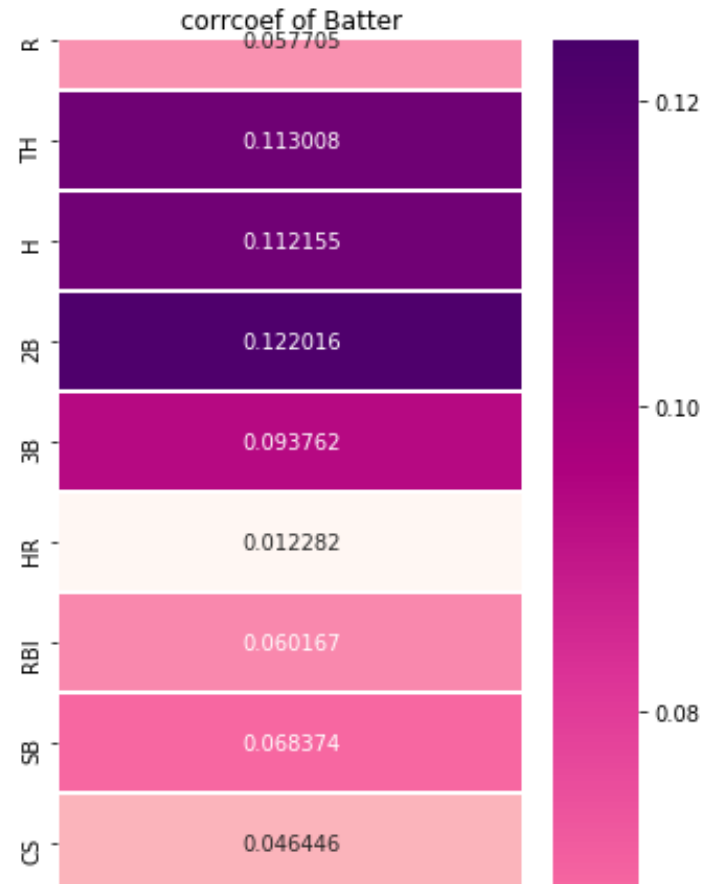
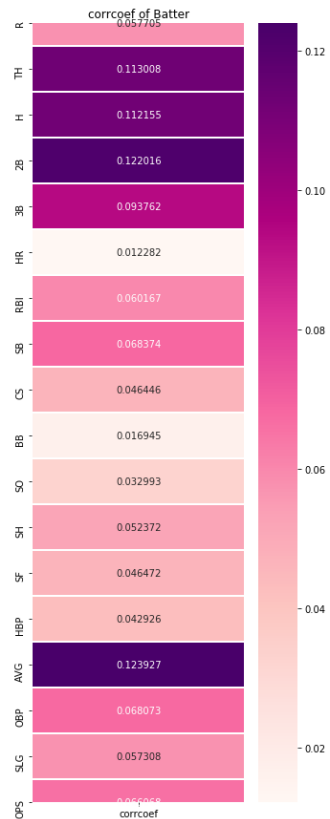
y_pred = clf_.predict(X_std)
print("★ 예상 총 득점수는 ★")
print(y_pred)
```

# 그래프

지표 별 승률과의 상관계수 그래프

〈 상관 계수가 높은 지표들을 이용해 머신 러닝에 활용함 〉

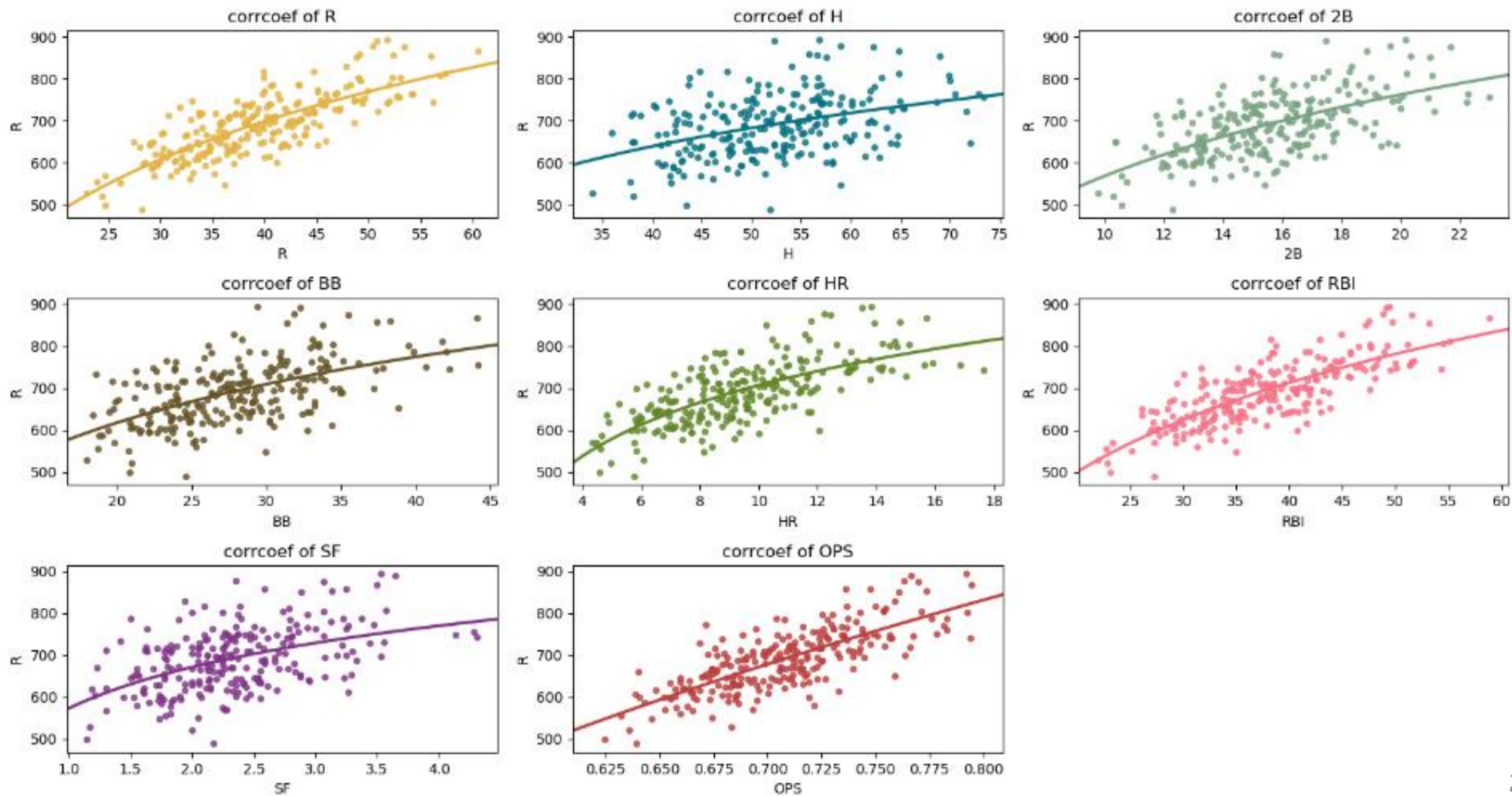
〈 원본 〉





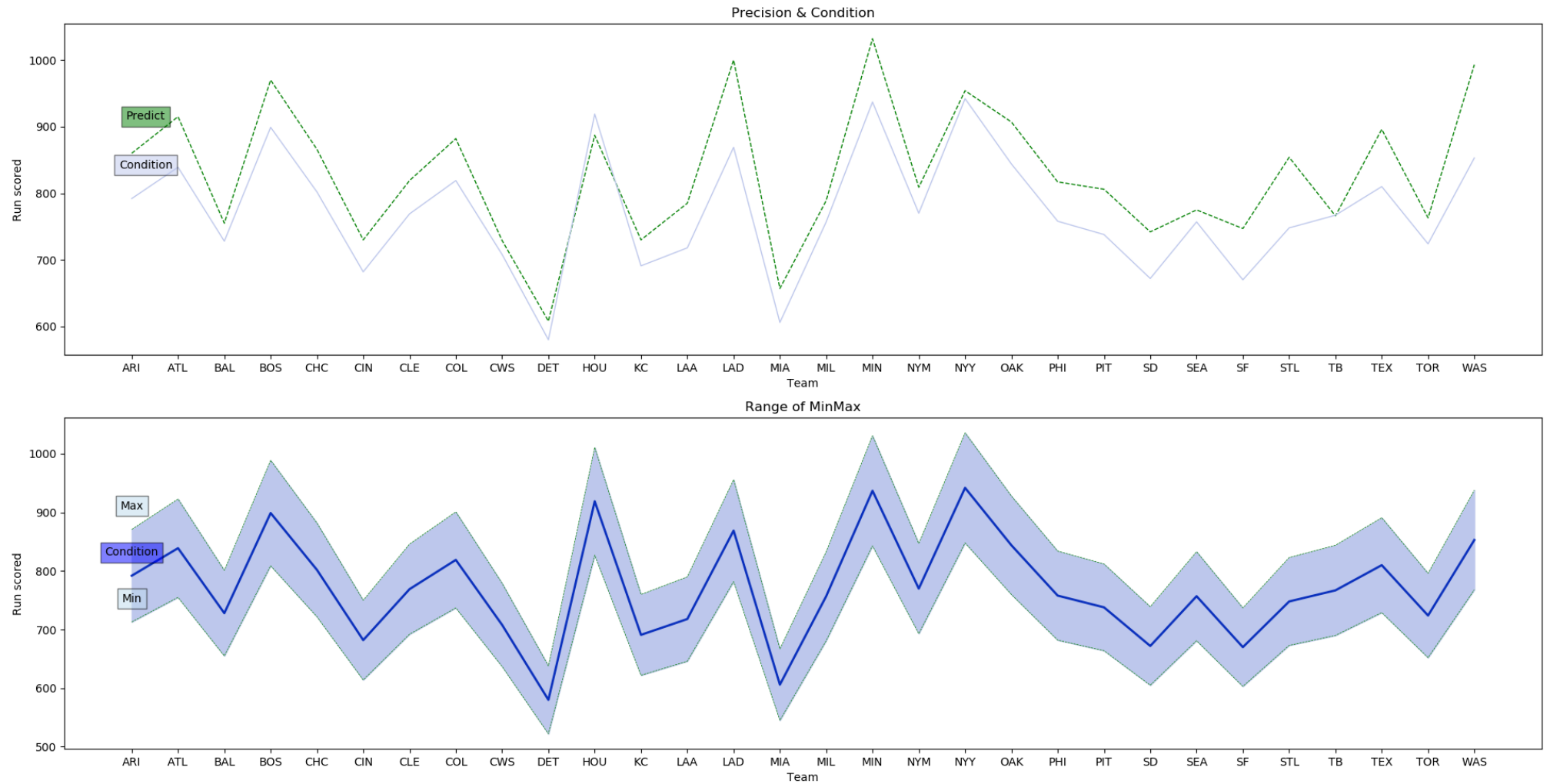
## 그래프

〈 상관 계수가 높은 지표들을 이용해 머신 러닝에 활용한 것을 스캐터 그래프로 표현 〉



# 그래프

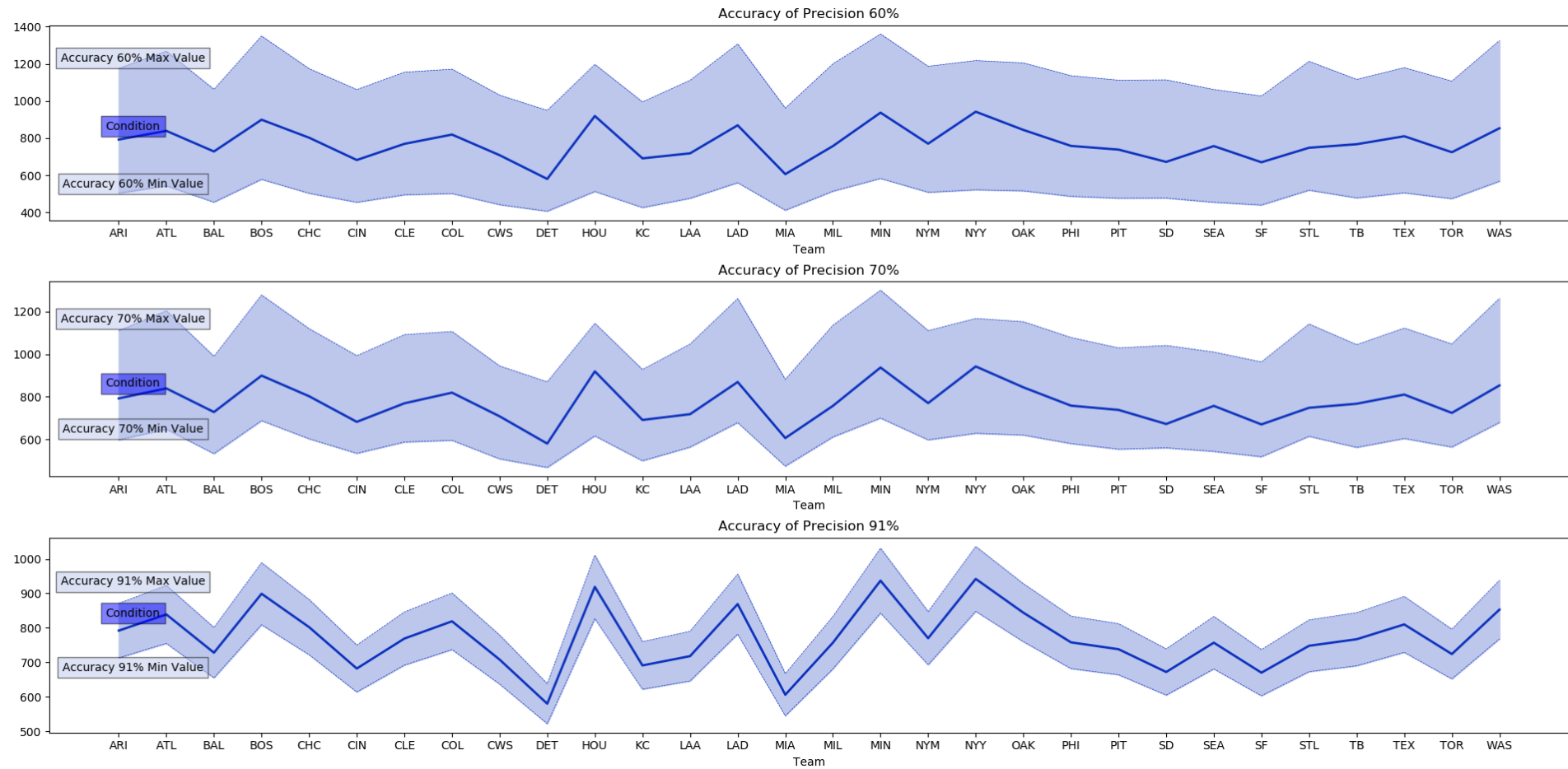
〈 상관관계가 있는 지표들로 머신 러닝을 돌려서 나온 값과 실제 2019년 득점 수의 관계(득점 수 예측) 〉



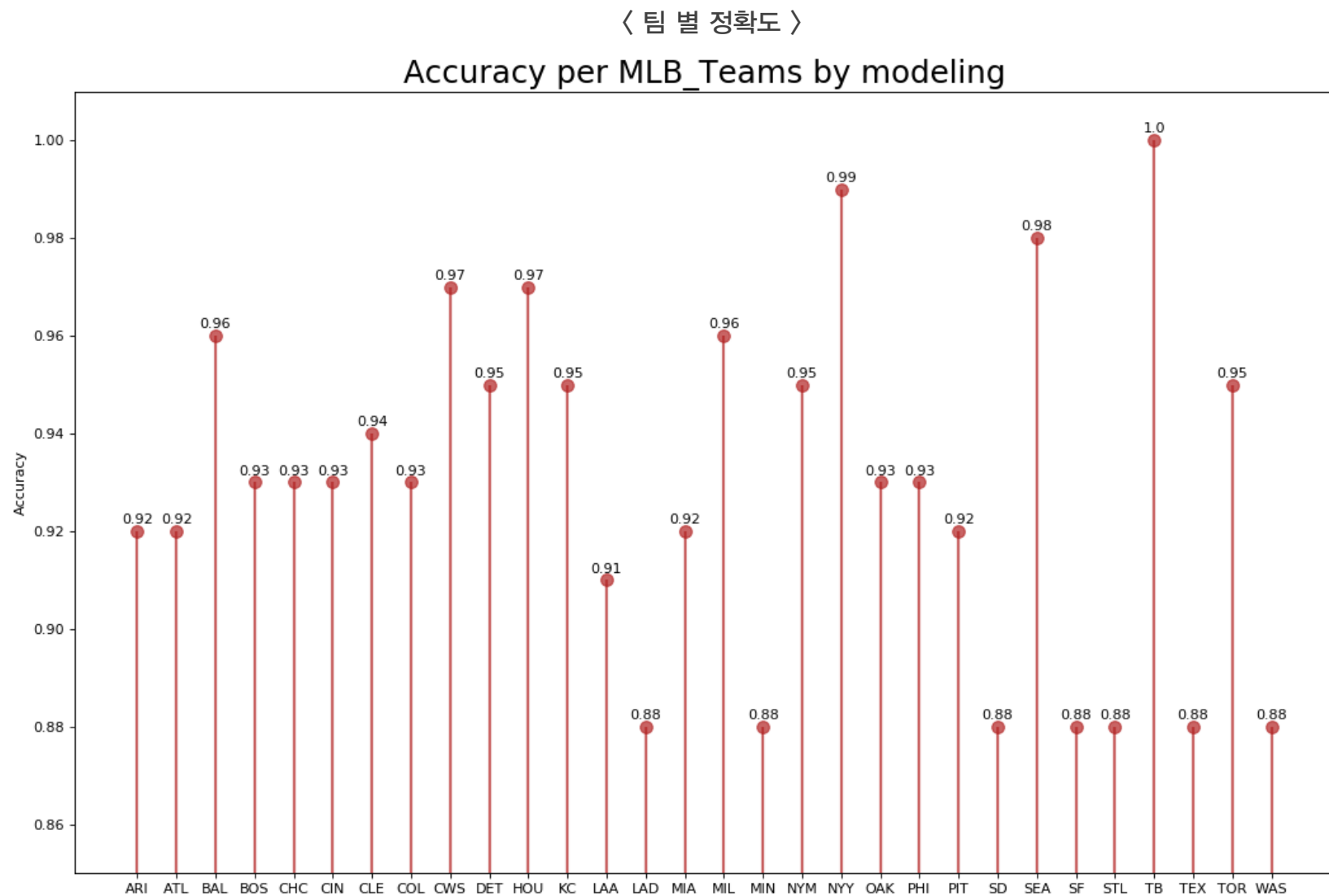


# 그래프

〈 60% 70% 90%에 따른 머신 러닝 정확도 상승 〉

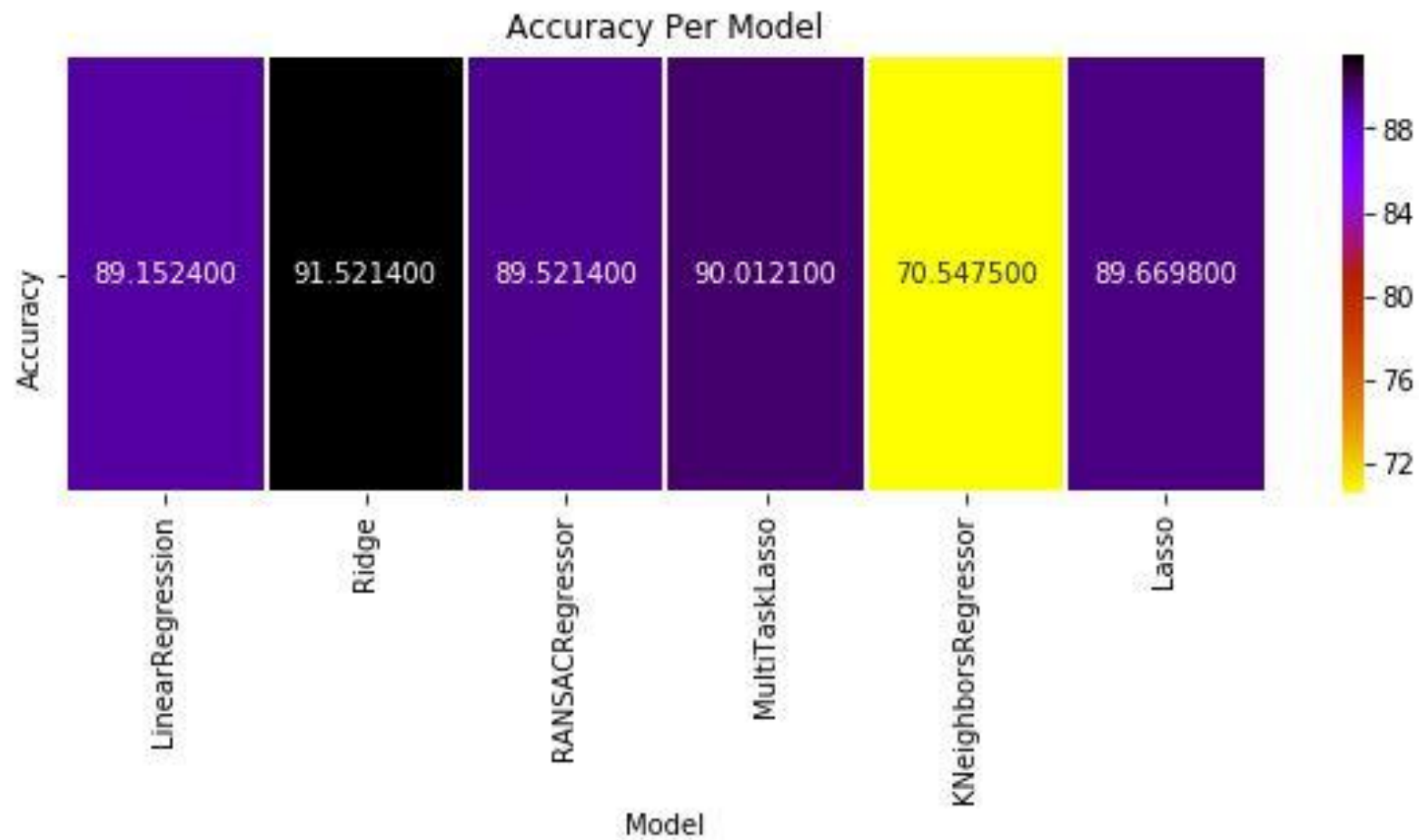


## 그래프



## 그래프

〈 회귀분석 종류에 따른 학습 정확도 비교 〉





당신이 응원하는 팀의 내년 총 득점수가 궁금하신가요?

지금 바로 투자하세요!

내년 총 득점수 확인

시즌 ALL STAR 확인



1. 타자의 지표와 득점수의 상관관계

2. 2019년 머신러닝 결과와 실제 득점수의 관계

3. 학습정확도에 따른 머신러닝 정확도

4. 팀별 정확도

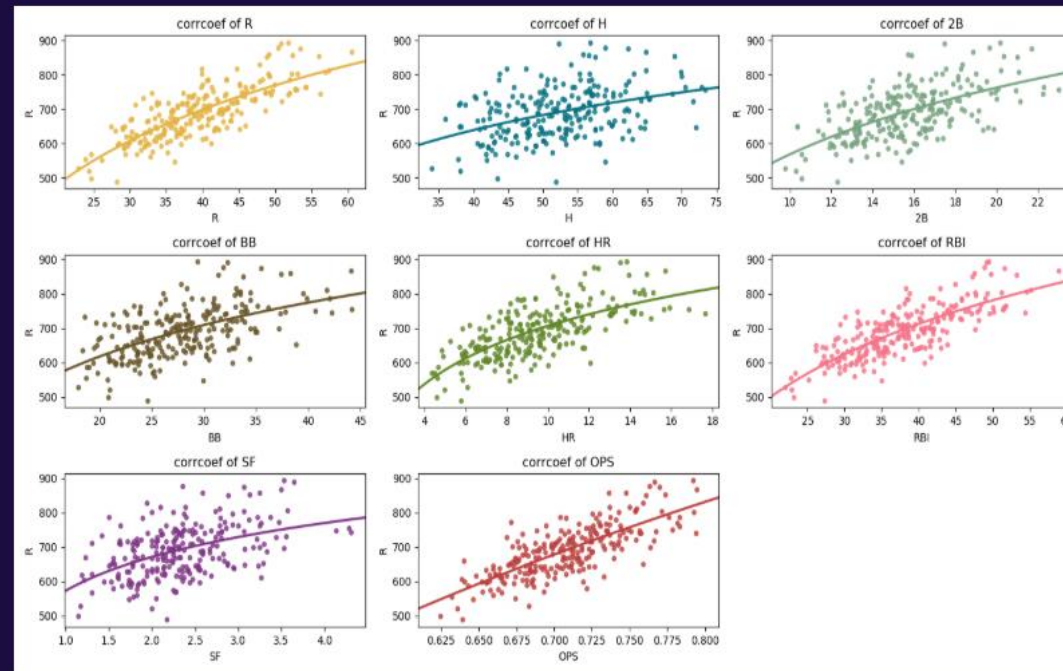
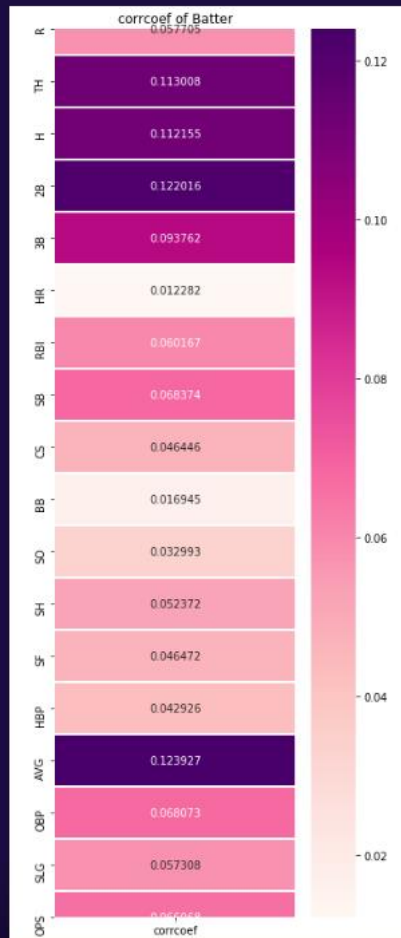
5. 회귀분석 종류에 따른 학습 정확도 비교

6. 머신러닝 객체에 ALL STAR 데이터 대입하여  
득점수 예측

Main으로

# 1. 타자의 지표와 득점수의 상관관계

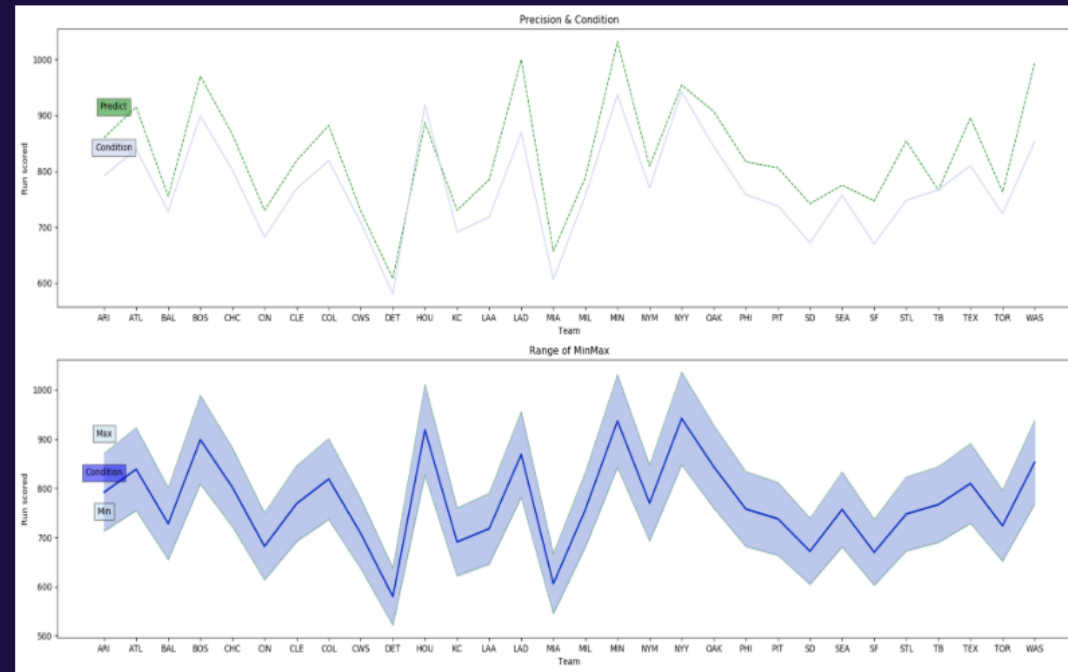
## ★ Hitmap ★



Go Back

## 2. 2019년 머신러닝 결과와 실제 득점수의 관계

### ★ Precision & Condition ★

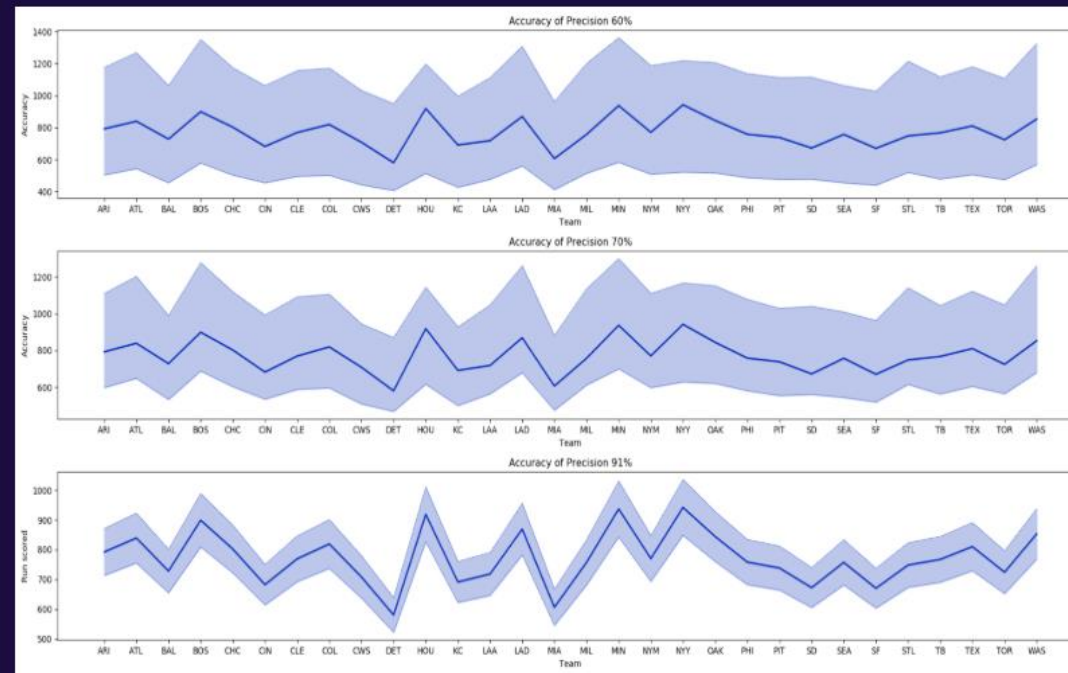


Go Back



### 3. 학습정확도에 따른 머신러닝 정확도

★ 60% 70% 90%에 따른 머신러닝 정확도 ★

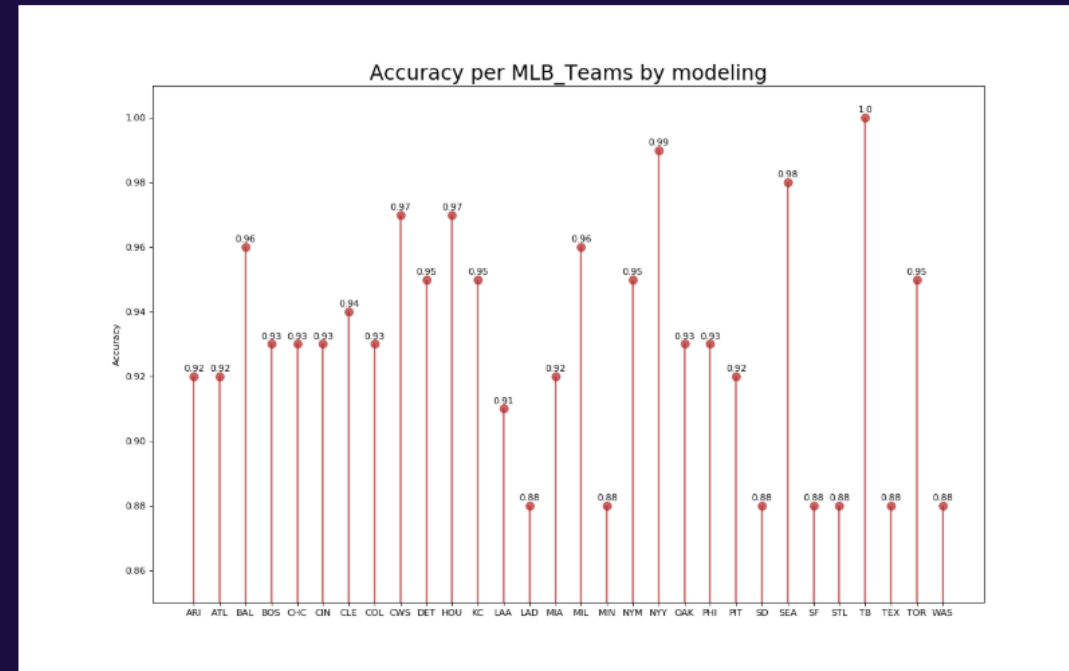


Go Back



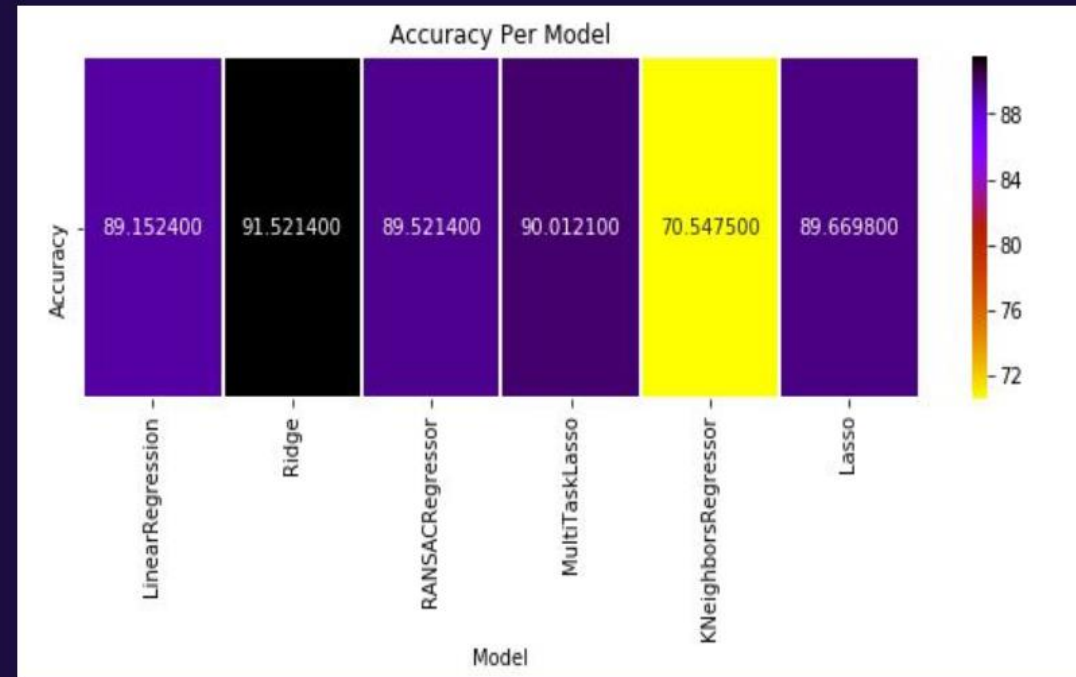
## 4. 팀별 정확도

### ★ LollipopBar Plot ★

[Go Back](#)

## 5.회귀분석 종류에 따른 학습 정확도 비교

★ Hitmap ★



Go Back

## 6. 머신러닝 객체에 ALL STAR데이터 대입하여 득점수 예측

ALL STAR TEAM의 지표를 입력해주세요.^\_^

R

97.100000

H

151.100000

2B

32.300000

HR

35.500000

RBI

101.000000

SF

5.100000

OPS

0.965600

BB

75.900000

득점수 예측 GO!

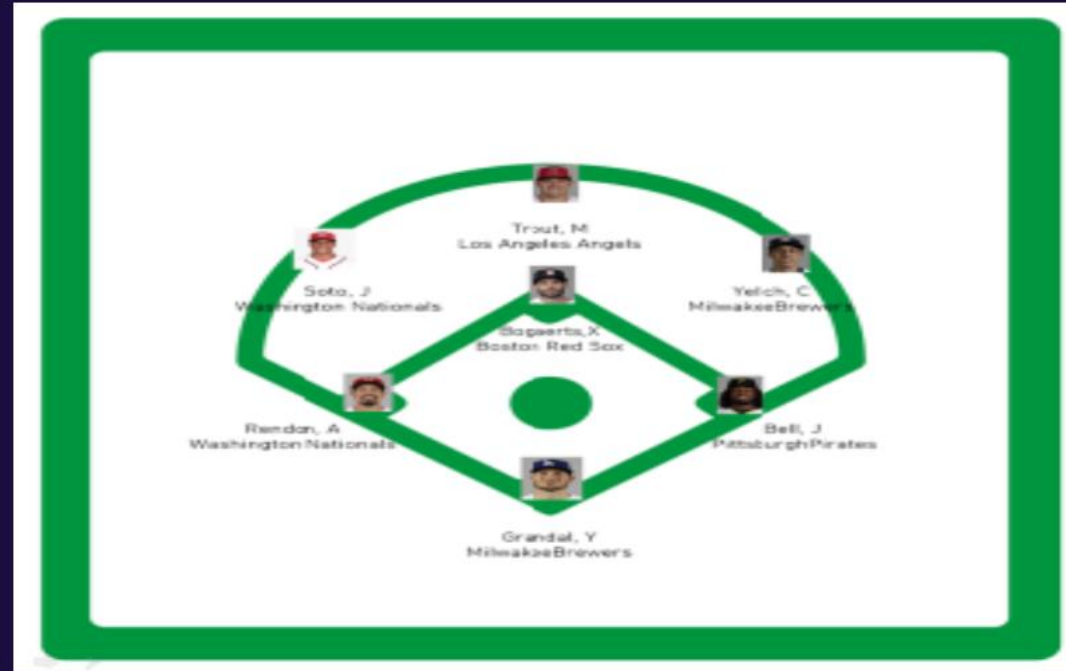
## 6. 머신러닝 객체에 ALL STAR데이터 대입하여 득점수 예측

★ vscode 실행 결과 ★

```
235    32.173913
236    28.875000
237    38.000000
238    25.592593
239    32.080000
Name: R, Length: 240, dtype: float64
Mean Squared Error : 707.5428677963616
Mean Absolute Error : 22.306872901063283
train 학습 정확도 : 0.8570435886194582
test 학습 정확도 : 0.9073314559151874
[[ 29.52573392 -11.75656111 -10.31359328 -19.23654464  3.31681039
  12.17001768  92.19807092  -8.18082397]]
[524.112942251]
★예상 총 득점수는 ★
[[1007.82593631]]
```

Go Back

## ★ 시즌 ALL STAR ★

[Go Back](#)

## 결론

야구에서 왜 Saber metrics를 쓰는 이유를 알게 되었고,  
투수의 지표와 합쳐보면 시즌 랭킹 예측이나 선수 별 stats  
예측이 충분이 가능 할 것이라고 생각함.

## 아쉬운 점

- 타자 데이터로만 작업한 것
- 더 많은 연도의 자료를 포함시켜 작업하지 못한 것 – 데이터 양이 작음
- 10년 간 팀 이적을 고려하지 못함

## Q & A

질문 해주세요~ ^^\*