2020 데이터 사이언스 스쿨

축구선수의 시장 가치 예측에서 SNS 지표의 기여도

고건호, 김정섭, 이왕건

개요 - 주제 선정 이유?

너도나도 몸값 1000억원… "유럽 축구 시장 미쳤다"

조선일보 | 임경업 기자

입력 2017.07.08 03:04

루카쿠, 맨유 이적 성사된다면 몸값 1120억원으로 역대 5위
벨로티는 1200억 이적설 나와 "2~3년내 2600억 시대 올 수도"
유럽 구단 17곳, 빚 2000억 넘어 "이적료 인플레로 축구 산업 흔들"

소수의 메가 구단들의 수입 증대 및 중동(오일머니) 자본의 개입으로 몸값 인플레이션 현상 발생

당연해져버린 '거품잔뜩' 축구 이적시장

By 이상민 비평단 Posted 17-07-31 23:48 Comments 3건

+ −

1. 무엇을 기준으로 선수들의 **몸값이 결정**되는가?
2. 선수들이 경기장에서 보여주는 퍼포먼스만으로 그들의 **몸값 예측**이 가능할까?

선행연구 참고 : 국내외 30여편 선행연구 검토

1. 데이터 측면

- 해외 축구에 대한 국내 연구는 거의 전무한 상태. 더군다나 ML알고리즘을 활용한 연구는 현재까지 발견하지 못함
- 해외 축구에 대한 해외 연구는 회귀, ML알고리즘 활용한 연구가 다수 있어, 비교 연구하기에 적절.
  - (대상 데이터셋은 유사하나, 방법론 측면(선형회귀, ML알고리즘 각각 1건씩)에서 다른 논문 2건 발견)
- MARKET VALUE PREDICTION 시, SNS지표를 활용한 연구 전무

시사점 : SNS지표의 축구선수 MARKET VALUE와의 상관성 및 예측모델링 진행

2. 방법 측면

- 최근 연구에서 ML 방법론의 연구가 두드러지나, 동시에 크롤링 데이터를 토대로 한 회귀 연구도 함께 활발히 진행 중

시사점 : CRAWLING으로 데이터 수집 + 회귀 및 ML 연구로 발전

선수 데이터
API-Football
(200,000r x 51c)

몸값 데이터
Transfermarkt.com
(500 rows)

SNS 지표
Instagram Follower
(500 rows)

API Crawling

Selenium Crawling

Selenium Crawling

DATASET
Preprocessed and merged
(500 rows)

ALL DATA OLS
- Domain Base
- Selected features

ALL DATA OLS
- Domain Base
- All features

ALL DATA OLS
- PCA
- Backward Elimination

ATK DATA OLS
- PCA
- Backward Elimination

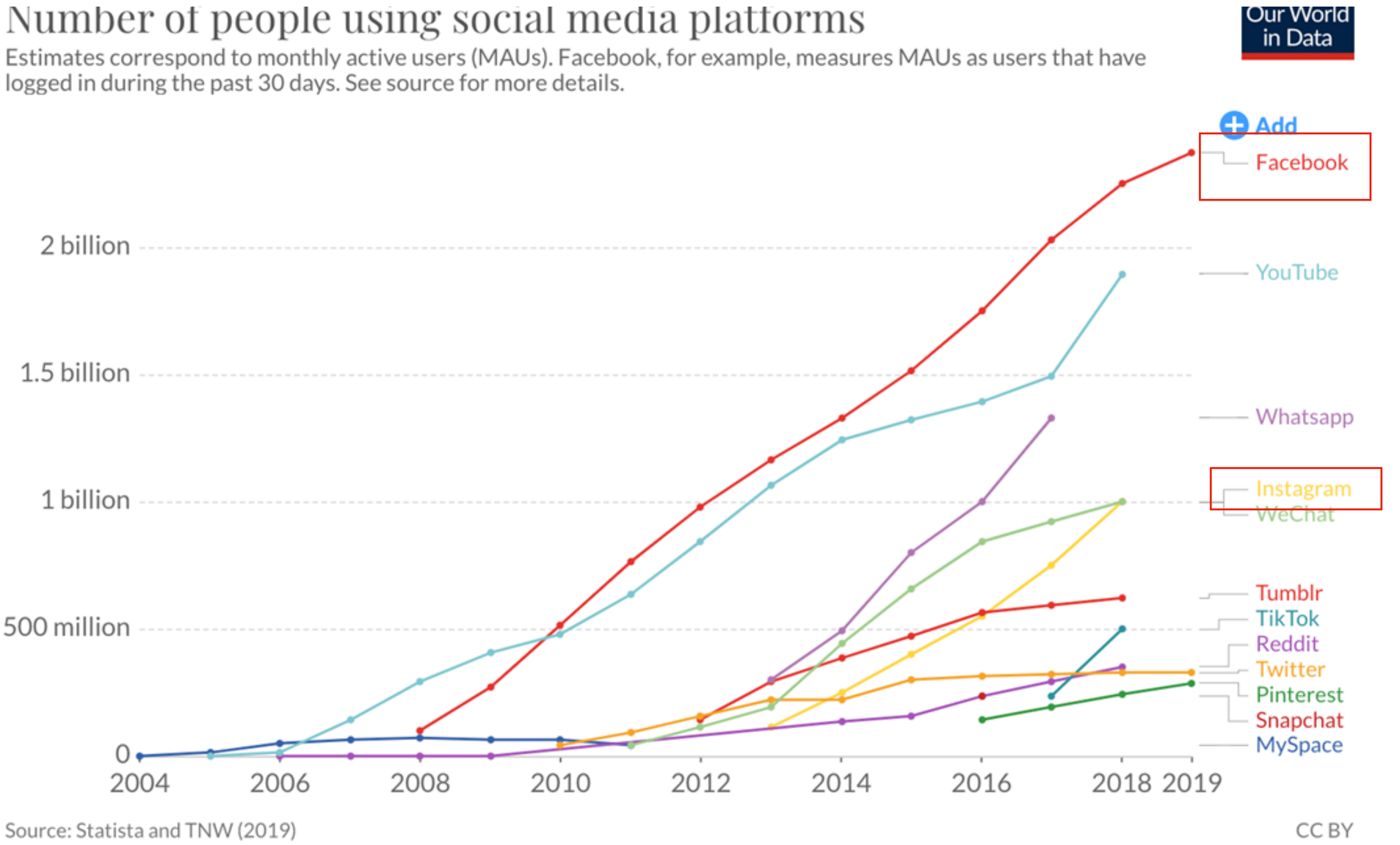
개요

추론 및 기술통계

1. 선수들의 몸값에 경기 결과 데이터가 영향을 미칠 것이다. (EX.득점, 키패스, 태클 .. 등)
2. 실력뿐만 아니라, 선수의 상품성도 고려 > 구단의 상품판매량 증대
3. 과거와 달리, 선수들과 팬들의 소통은 많아졌고, 특히 SNS를 활용한 소통이 매우 많음 즉 SNS의 FOLLOWER수가 선수의 인기를 대변해준다고 할 수 있다.

예측, 탐구

1. 유럽리그 소속 선수 대상 상위 500명 선수의 몸값 예측
2. 2가지 모델로 나누어 성능을 평가


Number of people using social media platforms
Estimates correspond to monthly active users (MAUs). Facebook, for example, measures MAUs as users that have logged in during the past 30 days. See source for more details.
Our World in Data
Add
Facebook
YouTube
Whatsapp
Instagram
WeChat
Tumblr
TikTok
Reddit
Twitter
Pinterest
Snapchat
MySpace
2 billion
1.5 billion
1 billion
500 million
0
2004
2006
2008
2010
2012
2014
2016
2018
2019
Source: Statista and TNW (2019)
CC BY

SNS 수입도 甲…호날두, 인스타그램 수입 1위

공유 0 댓글 0

HOME > 라이프

인스타그램 팔로워 늘리기 전문 '인스타터보', 팔로워 기반 마케팅 효과 높여

이다연 기자 | 승인 2020.07.04 09:00 | 댓글 0

SNS지표가 선수 개인의 인기도를 반영하고, 이것을 통해 구단은 팀을 홍보하는 효과가 매우 큼

데이터 출처 및 수집

TABLES
- api_football
- attacker
- defender
- goalkeeper
- market_instagram
- midfielder

Search: player_name = Filter

player_name	position	age	nationality	height	weight	rating	team_name	league	season	captain	shots_total	shots_on	goals_total	goals_conceded	goals_assists	passes_total
R. BÃ¼rki	Goalkeeper	30	Switzerland	187	85	7	Borussia Dortmund	Bundesliga	2019-2020	0	0	0	0	33	0	544
Ahmet Can Tekin	Midfielder	22	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
AnÄ±l SarÄ±oÄŸlu	Defender	23	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
AnÄ±l SarÄ±oÄŸlu	Defender	23	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2018-2019	0	0	0	0	0	0	0
AnÄ±l SarÄ±oÄŸlu	Defender	23	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2017-2018	0	0	0	0	0	0	0
BatÄ±nay Ak	Midfielder	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
Burak GenÃ§bay	Defender	24	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
Burak Kurttekin	Defender	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
Can DÃ¼ndar	Defender	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2018-2019	0	0	0	0	0	0	0
Can DÃ¼ndar	Defender	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2017-2018	0	0	0	0	0	0	0
Cihan Bal	Goalkeeper	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
Emre KaragÃ¼zel	Attacker	24	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
Eren Ayhan	Midfielder	19	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2019-2020	0	0	0	0	0	0	0
HÃ¼rkal Eren Turan	Defender	25	Turkey	0	0	0	1074 Ã‡ankÃ±rÃ±spor	Cup	2018-2019	0	0	0	0	0	0	0

Compact | Detailed | Gallery

#	Player	Age	Nat.	Club	Market value
1	Kylian Mbappé Centre-Forward	21			€180.00m
2	Raheem Sterling Left Winger	25			€128.00m
3	Neymar Left Winger	28			€128.00m
4	Sadio Mané Left Winger	28			€120.00m
5	Mohamed Salah Right Winger				
6	Harry Kane Centre-Forward				
7	Kevin De Bruyne Attacking Midfield				

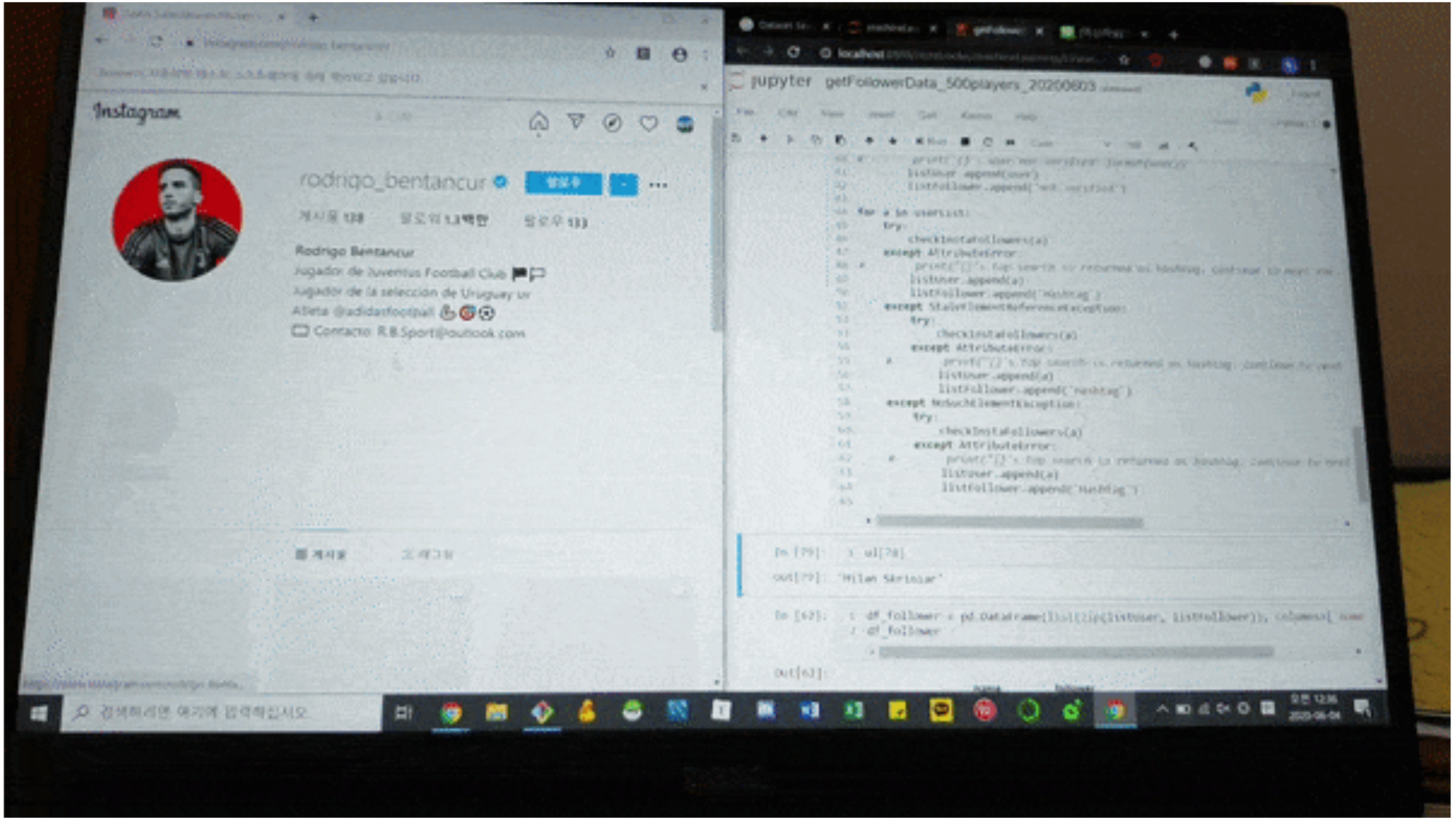
messi_messi10 Follow

512 posts 1.4m followers 36 following

Lionel Messi
Fanpage Of @leomessi
themessistore.com

예측 모델링 생성

데이터 출처 및 수집


Instagram
rodrigo_bentancur
Rodrigo Bentancur
Jugador de Juventus Football Club
Jugador de la selección de Uruguay
Atleta @adidasfootball
Contacto: R.B.Sport@outlook.com
jupyter getFollowerData_500players_20200603
In [79]: ul[78]
Out[79]: 'Milan Škriniar'
In [62]: df_follower = pd.DataFrame(list(zip(listUser, listFollower)), columns=[
df_follower
Out[62]:

OLS Regression Results

Dep. Variable:	value	R-squared:	0.479
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	6.484
Date:	Wed, 17 Jun 2020	Prob (F-statistic):	2.66e-18
Time:	17:21:07	Log-Likelihood:	-1082.9
No. Observations:	259	AIC:	2232.
Df Residuals:	226	BIC:	2349.
Df Model:	32		
Covariance Type:	nonrobust		

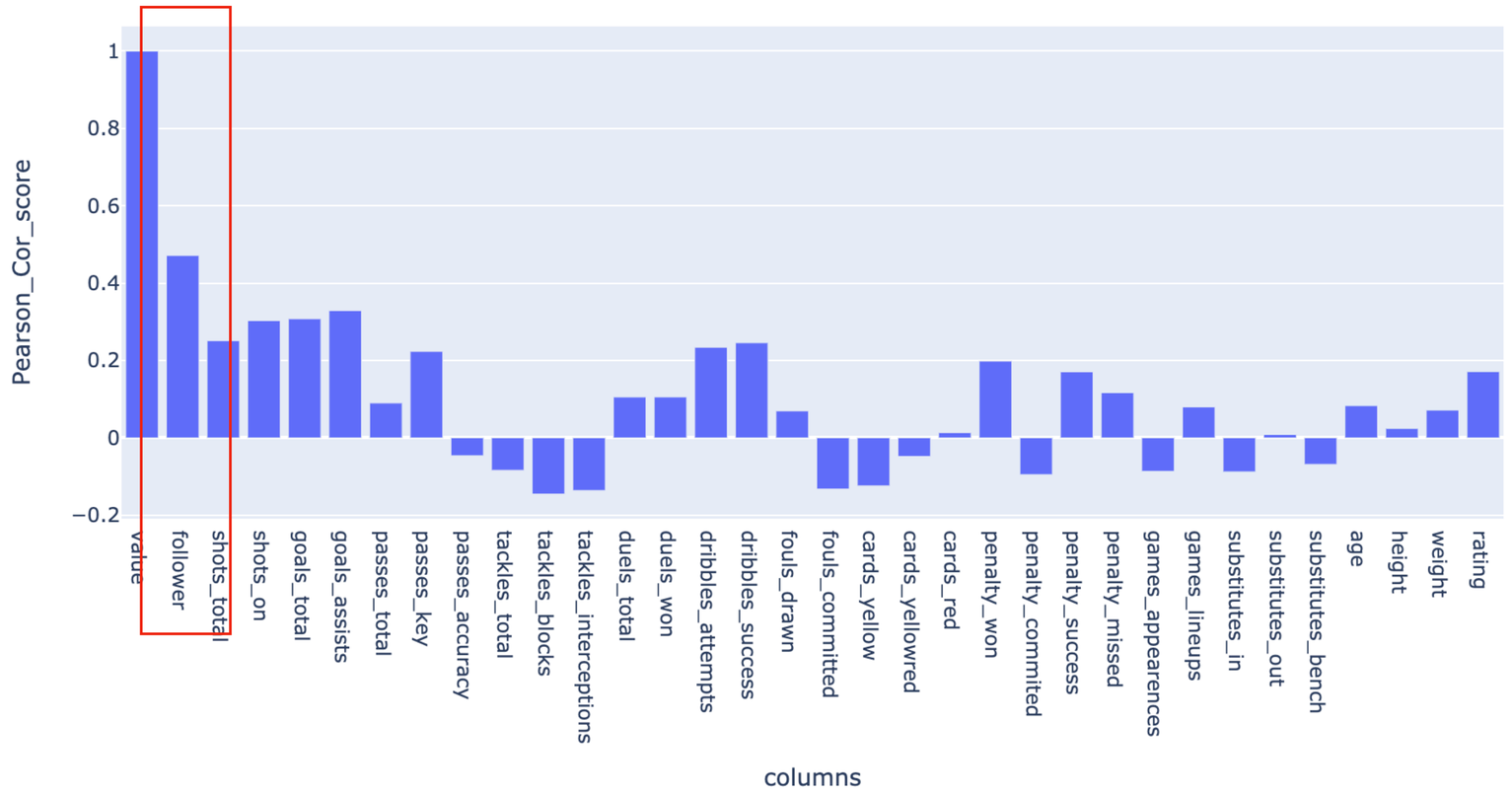
	coef	std err	t	P>|t|	[0.025	0.975]
const	128.3997	76.689	1.674	0.095	-22.717	279.517
shots_total	-5.0267	5.321	-0.945	0.346	-15.512	5.459
shots_on	-4.8542	16.065	-0.302	0.763	-36.511	26.803
goals_total	93.8629	22.586	4.156	0.000	49.357	138.369
goals_conceded	13.7809	11.252	1.225	0.222	-8.391	35.953
goals_assists	-2.8796	27.745	-0.104	0.917	-57.551	51.792
passes_total	0.2343	0.106	2.210	0.028	0.025	0.443
passes_key	1.6174	4.055	0.399	0.690	-6.373	9.608
passes_accuracy	1.3074	0.666	1.964	0.051	-0.004	2.619
tackles_total	1.3316	2.831	0.470	0.639	-4.247	6.910
tackles_blocks	-7.4574	9.786	-0.762	0.447	-26.742	11.827
tackles_interceptions	1.6870	3.939	0.428	0.669	-6.075	9.449
duels_total	1.2885	1.814	0.710	0.478	-2.286	4.863
duels_won	-2.2954	3.711	-0.619	0.537	-9.608	5.018
dribbles_attempts	-5.4973	5.339	-1.030	0.304	-16.017	5.023
dribbles_success	14.2732	8.422	1.695	0.091	-2.322	30.868
fouls_drawn	-3.1057	3.052	-1.018	0.310	-9.119	2.908
fouls_committed	-3.3257	4.120	-0.807	0.420	-11.444	4.793
cards_yellow	12.9325	18.614	0.695	0.488	-23.748	49.613
cards_yellowred	-111.6613	173.064	-0.645	0.519	-452.687	229.364
cards_red	340.8754	187.001	1.823	0.070	-27.614	709.364
penalty_won	140.2245	68.938	2.034	0.043	4.381	276.068
penalty_commited	-123.7150	109.580	-1.129	0.260	-339.645	92.215
penalty_success	-185.5997	48.819	-3.802	0.000	-281.798	-89.402
penalty_missed	-123.0722	148.052	-0.831	0.407	-414.811	168.666
penalty_saved	-114.3852	281.359	-0.407	0.685	-668.808	440.038
games_appearences	2.137e+04	2.43e+04	0.880	0.380	-2.65e+04	6.92e+04
games_played	0.2346	0.062	3.794	0.000	0.113	0.356
games_lineups	-2.151e+04	2.43e+04	-0.885	0.377	-6.94e+04	2.64e+04
substitutes_in	-2.141e+04	2.43e+04	-0.881	0.379	-6.93e+04	2.65e+04
substitutes_out	8.6243	16.689	0.517	0.606	-24.261	41.510
substitutes_bench	0.3071	11.275	0.027	0.978	-21.911	22.525
follower	2.378e-07	1.36e-07	1.747	0.082	-3.05e-08	5.06e-07

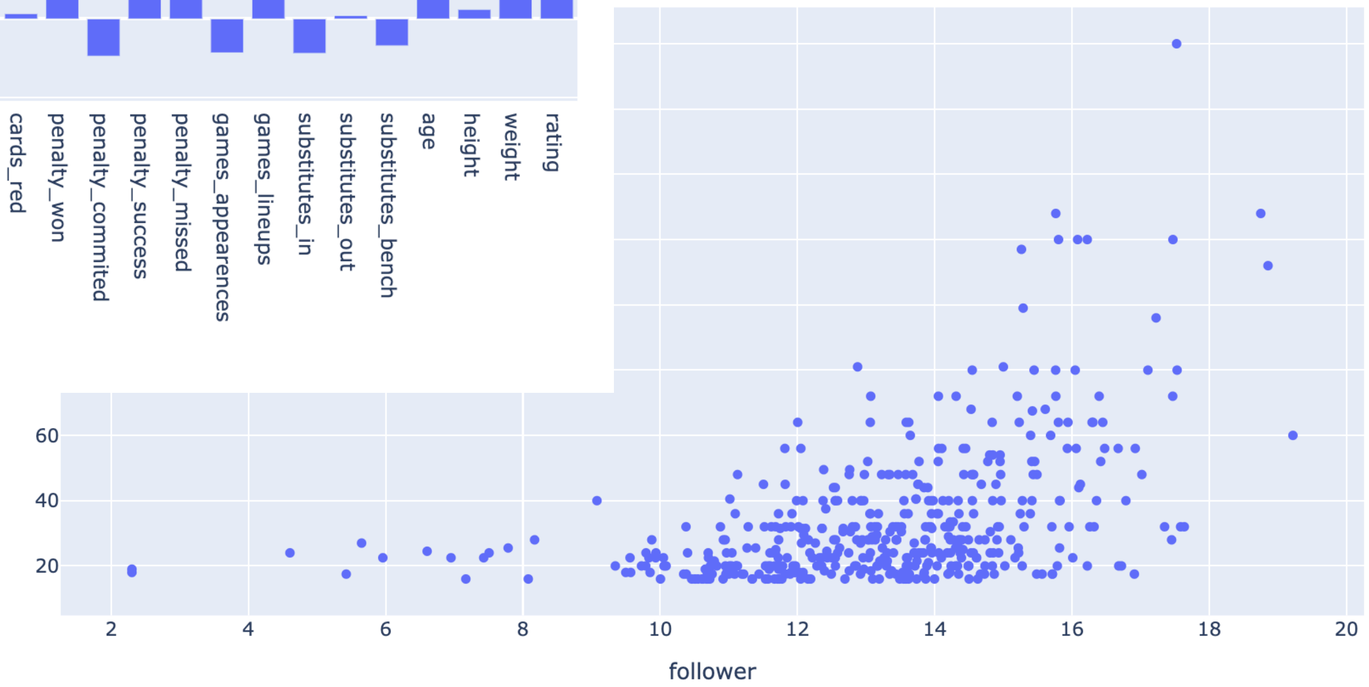
Omnibus:	69.959	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	171.789
Skew:	1.251	Prob(JB):	4.97e-38
Kurtosis:	6.108	Cond. No.	4.69e+11

```python
# 1. 상수항 결합

import statsmodels.api as sm

x_total= df_copy[['shots_total', 'shots_on', 'goals_total',
       'goals_conceded', 'goals_assists', 'passes_total', 'passes_key',
       'passes_accuracy', 'tackles_total', 'tackles_blocks',
       'tackles_interceptions', 'duels_total', 'duels_won',
       'dribbles_attempts', 'dribbles_success', 'fouls_drawn',
       'fouls_committed', 'cards_yellow', 'cards_yellowred', 'cards_red',
       'penalty_won', 'penalty_commited', 'penalty_success', 'penalty_missed',
       'penalty_saved', 'games_appearences', 'games_played',
       'games_lineups', 'substitutes_in', 'substitutes_out',
       'substitutes_bench','follower']]

X_total = sm.add_constant(x_total)
y_total = pd.DataFrame(df_copy['value'])
```

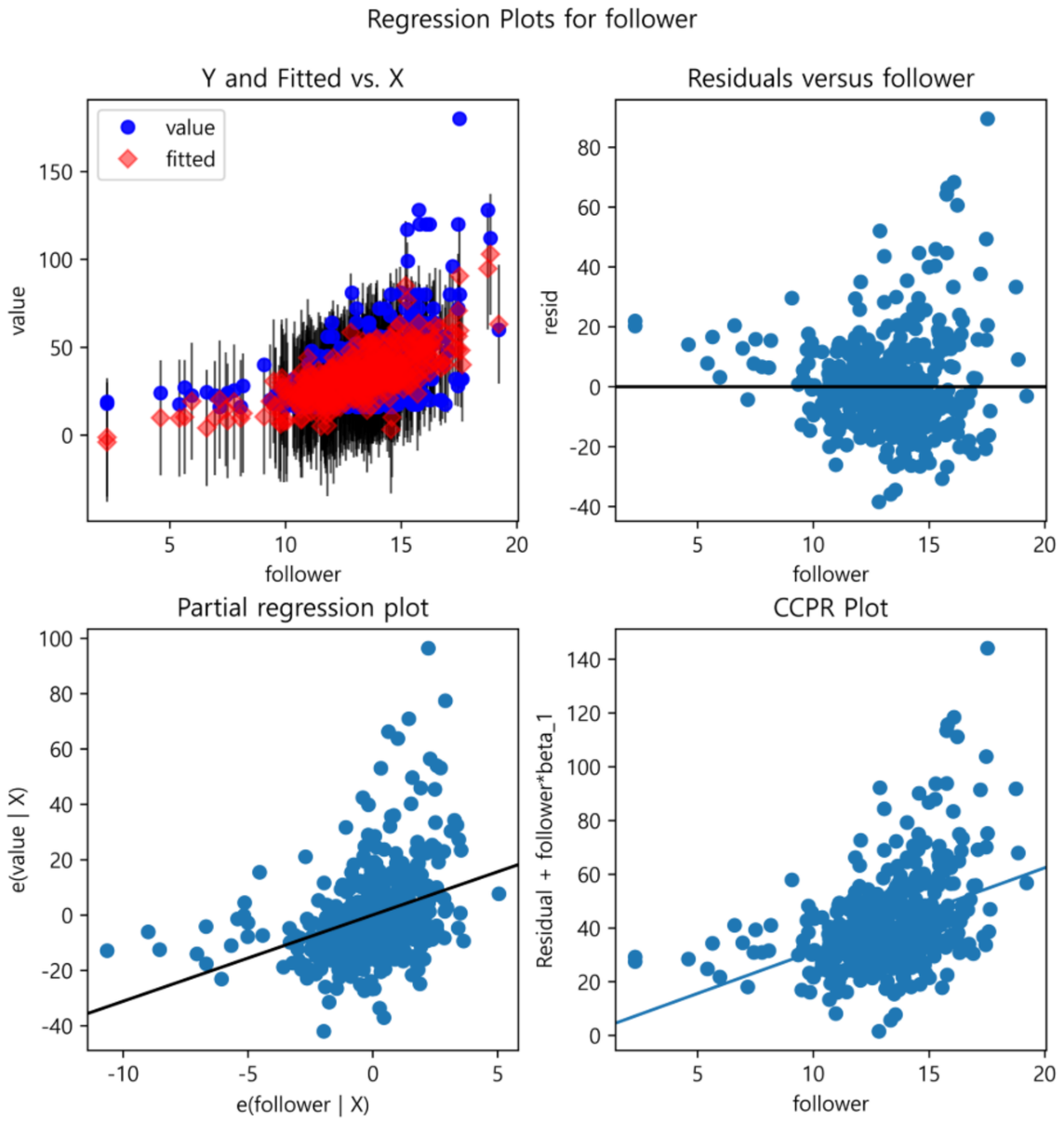
FOLLOWER 수의 P값은 0.082로 다른 변수에 비해 P값이 굉장히 낮게 나온 것을 확인

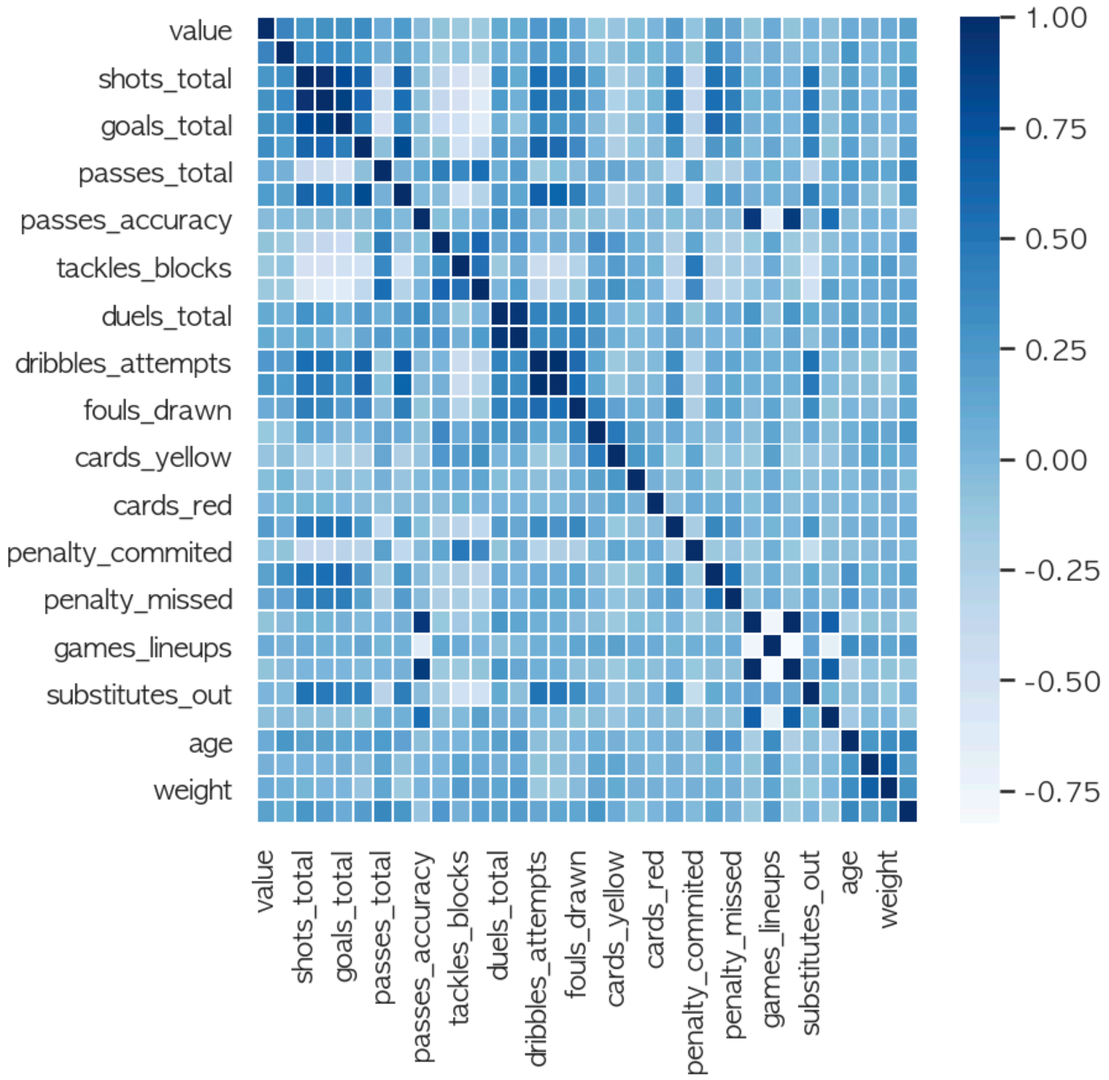
FOLLOWER수가 많고, 적음이 선수들의 몸값을 결정하는데 영향을 미칠 것이다





다른 변수들보다 SNS지표가 가장 높은 상관관계를 보이고 있음


Regression Plots for follower
Y and Fitted vs. X
value
fitted
value
follower
Residuals versus follower
resid
follower
Partial regression plot
e(value | X)
e(follower | X)
CCPR Plot
Residual + follower*beta_1
follower



조건수와 상관관계 분석 결과 독립변수간 강한 다중공선성이 의심이 되어 주성분 분석(PCA)를 실시함

SNS지표의 CONTRIBUTION 분석을 위한 모델링

SNS지표 O

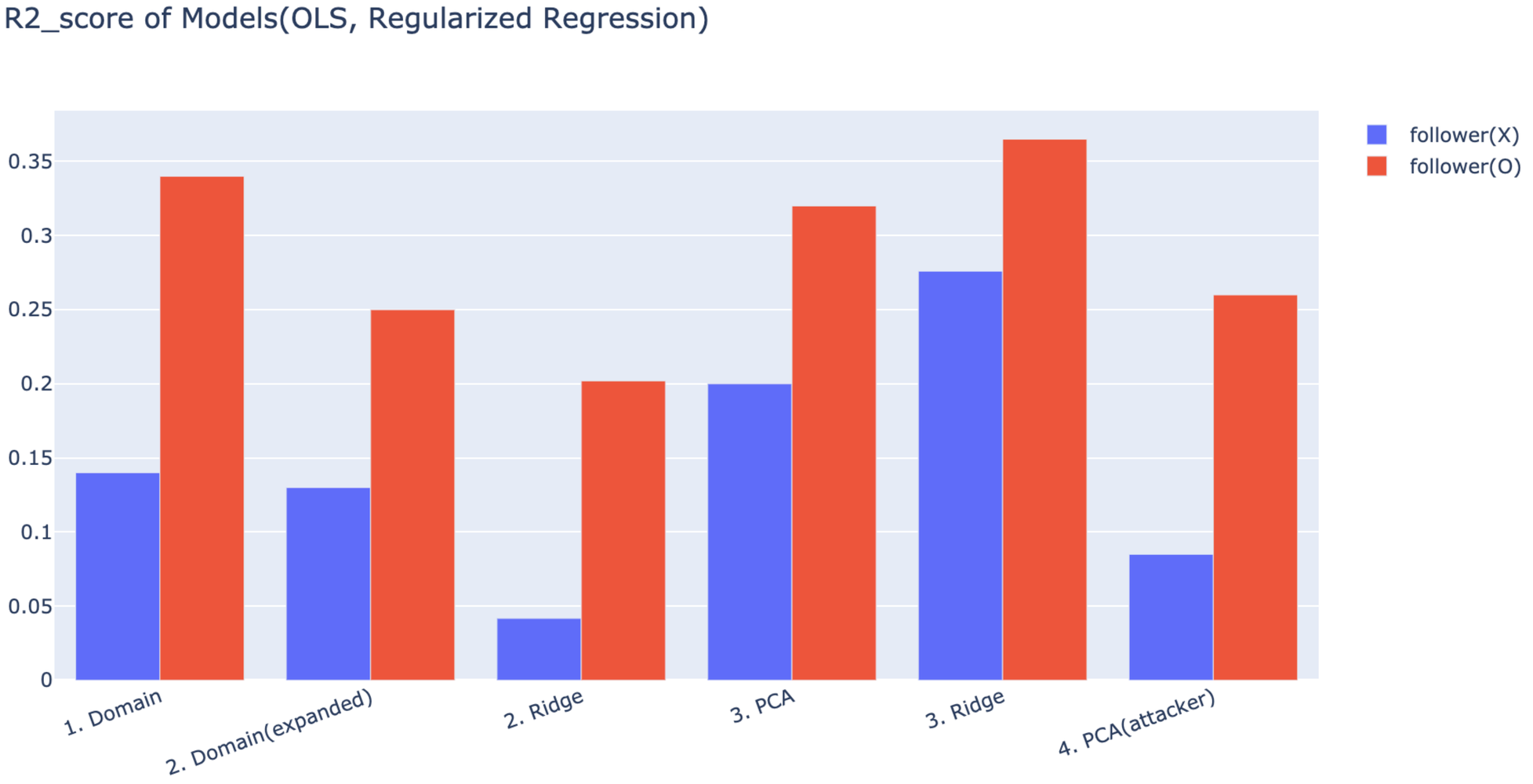
SNS지표 X

DOMAIN BASE FEATURE SELECTION 1

DOMAIN BASE FEATURE SELECTION 2

PCA BASE FEATURE SELECTION

PCA BASE FEATURE SELECTION (공격수)


R2_score of Models(OLS, Regularized Regression)
follower(X)
follower(O)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1. Domain
2. Domain(expanded)
2. Ridge
3. PCA
3. Ridge
4. PCA(attacker)

WITHOUT FOLLOWERS

1.

OLS Regression Results

Dep. Variable:	value	R-squared:	0.174
Model:	OLS	Adj. R-squared:	0.160
Method:	Least Squares	F-statistic:	12.57
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	7.11e-13
Time:	16:57:13	Log-Likelihood:	-1560.0
No. Observations:	365	AIC:	3134.
Df Residuals:	358	BIC:	3161.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.8712	0.918	34.710	0.000	30.065	33.677
scale(age)	-4.5290	8.968	-0.505	0.614	-22.166	13.108
scale(I(age ** 2))	19.9525	27.058	0.737	0.461	-33.259	73.164
scale(I(age ** 3))	-15.1705	19.031	-0.797	0.426	-52.597	22.256
scale(shots_on)	-3.5131	2.131	-1.649	0.100	-7.703	0.677
scale(goals_total)	7.2687	1.897	3.833	0.000	3.539	10.998
scale(goals_assists)	5.5425	1.181	4.694	0.000	3.220	7.865

Omnibus:	139.396	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	426.346
Skew:	1.783	Prob(JB):	2.63e-93
Kurtosis:	6.914	Cond. No.	67.0

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
모델성능: 0.1389196313721379

WITH FOLLOWERS

1.

OLS Regression Results

Dep. Variable:	value	R-squared:	0.430
Model:	OLS	Adj. R-squared:	0.417
Method:	Least Squares	F-statistic:	33.51
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	3.18e-39
Time:	16:56:33	Log-Likelihood:	-1492.4
No. Observations:	365	AIC:	3003.
Df Residuals:	356	BIC:	3038.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.8712	0.765	41.649	0.000	30.366	33.376
scale(follower)	32.7144	3.276	9.985	0.000	26.271	39.157
scale(I(follower ** 2))	-23.8157	3.251	-7.325	0.000	-30.210	-17.422
scale(age)	-12.3164	7.899	-1.559	0.120	-27.850	3.218
scale(I(age ** 2))	54.9110	24.473	2.244	0.025	6.782	103.040
scale(I(age ** 3))	-45.3498	17.436	-2.601	0.010	-79.640	-11.059
scale(shots_on)	-3.9166	1.818	-2.154	0.032	-7.492	-0.341
scale(goals_total)	4.6906	1.608	2.917	0.004	1.528	7.853
scale(goals_assists)	3.0974	1.020	3.036	0.003	1.091	5.104

Omnibus:	177.558	Durbin-Watson:	1.913
Prob(Omnibus):	0.000	Jarque-Bera (JB):	852.787
Skew:	2.102	Prob(JB):	6.60e-186
Kurtosis:	9.197	Cond. No.	79.2

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델성능: 0.33731893544051345

OLS Regression Results

Dep. Variable:	value	R-squared:	0.252
Model:	OLS	Adj. R-squared:	0.223
Method:	Least Squares	F-statistic:	8.594
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	4.34e-14
Time:	18:14:51	Log-Likelihood:	-1343.3
No. Observations:	319	AIC:	2713.
Df Residuals:	306	BIC:	2762.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.7335	0.933	34.024	0.000	29.898	33.569
scale(goals_total)	9.1139	2.335	3.903	0.000	4.519	13.709
scale(goals_assists)	5.3102	1.670	3.180	0.002	2.024	8.596
scale(tackles_blocks)	2.0157	1.243	1.621	0.106	-0.431	4.462
scale(tackles_interceptions)	2.2986	1.352	1.701	0.090	-0.361	4.958
scale(duels_won)	2.0613	1.176	1.753	0.081	-0.252	4.375
scale(dribbles_success)	4.1152	1.434	2.870	0.004	1.294	6.936
scale(fouls_drawn)	0.0117	1.318	0.009	0.993	-2.582	2.606
scale(shots_on)	-2.2837	2.647	-0.863	0.389	-7.493	2.926
scale(passes_key)	-3.2609	1.836	-1.776	0.077	-6.873	0.351
scale(fouls_committed)	-2.9923	1.076	-2.781	0.006	-5.109	-0.875
scale(cards_red)	1.8509	0.940	1.968	0.050	0.000	3.701
scale(age)	-1.0169	1.047	-0.971	0.332	-3.077	1.043

Omnibus:	175.842	Durbin-Watson:	1.839
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1358.137
Skew:	2.176	Prob(JB):	1.21e-295
Kurtosis:	12.124	Cond. No.	7.63

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.12877957185958908

OLS Regression Results

Dep. Variable:	value	R-squared:	0.357
Model:	OLS	Adj. R-squared:	0.332
Method:	Least Squares	F-statistic:	14.18
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	2.01e-23
Time:	18:20:23	Log-Likelihood:	-1319.1
No. Observations:	319	AIC:	2664.
Df Residuals:	306	BIC:	2713.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-10.5764	5.791	-1.826	0.069	-21.971	0.819
scale(goals_total)	7.5436	2.167	3.480	0.001	3.279	11.809
scale(goals_assists)	4.7965	1.549	3.097	0.002	1.749	7.844
scale(tackles_blocks)	1.5172	1.154	1.315	0.190	-0.753	3.788
scale(tackles_interceptions)	2.1347	1.253	1.704	0.089	-0.330	4.600
scale(duels_won)	1.7345	1.089	1.593	0.112	-0.409	3.878
scale(dribbles_success)	3.2252	1.335	2.416	0.016	0.599	5.852
scale(fouls_drawn)	0.0172	1.221	0.014	0.989	-2.386	2.421
scale(shots_on)	-2.1637	2.452	-0.883	0.378	-6.988	2.660
scale(passes_key)	-3.6267	1.694	-2.141	0.033	-6.961	-0.293
scale(fouls_committed)	-2.5600	0.999	-2.562	0.011	-4.526	-0.594
scale(age)	-2.6073	0.996	-2.618	0.009	-4.567	-0.647
follower	3.2233	0.436	7.389	0.000	2.365	4.082

Omnibus:	178.406	Durbin-Watson:	1.893
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1631.181
Skew:	2.144	Prob(JB):	0.00
Kurtosis:	13.214	Cond. No.	89.8

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
검증 성능 : 0.24882780031153712

3번 모델 : PCA를 활용한 feature selection

OLS Regression Results

Dep. Variable:	value	R-squared:	0.302
Model:	OLS	Adj. R-squared:	0.282
Method:	Least Squares	F-statistic:	14.85
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	5.00e-20
Time:	16:48:37	Log-Likelihood:	-1332.3
No. Observations:	319	AIC:	2685.
Df Residuals:	309	BIC:	2722.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.7335	0.897	35.391	0.000	29.969	33.498
scale(passes_total)	5.3880	1.146	4.702	0.000	3.133	7.643
scale(fouls_committed)	-3.4955	0.944	-3.702	0.000	-5.353	-1.638
scale(cards_red)	2.1194	0.907	2.336	0.020	0.334	3.905
scale(games_lineups)	3.1723	1.011	3.138	0.002	1.183	5.161
scale(substitutes_out)	-3.5894	1.191	-3.013	0.003	-5.934	-1.245
scale(age_x)	-3.2800	1.091	-3.007	0.003	-5.427	-1.133
scale(shotsOnTotal_goalsTotal)	6.4967	1.676	3.877	0.000	3.199	9.794
scale(I(shotsOnTotal_goalsTotal ** 2))	4.1835	1.247	3.356	0.001	1.730	6.637
scale(dribblesAtmptsSuc)	4.4932	1.166	3.855	0.000	2.200	6.787

Omnibus:	188.557	Durbin-Watson:	1.820
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1699.573
Skew:	2.317	Prob(JB):	0.00
Kurtosis:	13.315	Cond. No.	3.63

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.205373973383963

OLS Regression Results

Dep. Variable:	value	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	20.71
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	3.16e-31
Time:	16:48:34	Log-Likelihood:	-1301.1
No. Observations:	319	AIC:	2626.
Df Residuals:	307	BIC:	2671.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.7335	0.816	38.900	0.000	30.128	33.339
scale(follower)	22.1857	2.581	8.597	0.000	17.108	27.264
scale(I(follower ** 2))	-18.2431	2.433	-7.499	0.000	-23.030	-13.456
scale(passes_total)	3.8487	1.060	3.632	0.000	1.764	5.934
scale(fouls_committed)	-2.8872	0.880	-3.283	0.001	-4.618	-1.157
scale(cards_red)	1.5941	0.829	1.924	0.055	-0.036	3.225
scale(games_lineups)	2.8849	0.918	3.143	0.002	1.079	4.691
scale(substitutes_out)	-2.7010	1.094	-2.468	0.014	-4.855	-0.547
scale(age_x)	-4.9357	1.037	-4.760	0.000	-6.976	-2.895
scale(weight_x)	2.1031	0.903	2.329	0.020	0.327	3.880
scale(shotsOnTotal_goalsTotal)	6.3015	1.264	4.985	0.000	3.814	8.789
scale(dribblesAtmptsSuc)	1.8289	1.058	1.729	0.085	-0.253	3.910

Omnibus:	145.899	Durbin-Watson:	1.832
Prob(Omnibus):	0.000	Jarque-Bera (JB):	868.854
Skew:	1.813	Prob(JB):	2.14e-189
Kurtosis:	10.226	Cond. No.	6.74

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.3186820894643888

4번 모델 : PCA를 활용한 feature selection(공격수)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  value   R-squared:                       0.636
Model:                            OLS   Adj. R-squared:                  0.571
Method:                 Least Squares   F-statistic:                     9.786
Date:                Mon, 29 Jun 2020   Prob (F-statistic):           3.27e-09
Time:                        17:09:14   Log-Likelihood:                -291.63
No. Observations:                  67   AIC:                             605.3
Df Residuals:                      56   BIC:                             629.5
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       42.6493      2.512     16.978      0.000      37.617      47.681
scale(age)                     -16.4055      3.421     -4.796      0.000     -23.258      -9.553
scale(tackles_total)            12.7699      3.290      3.881      0.000       6.179      19.361
scale(tackles_interceptions)    -9.1911      3.180     -2.890      0.005     -15.561      -2.821
scale(fouls_drawn)              10.0402      3.247      3.092      0.003       3.535      16.545
scale(fouls_committed)         -11.7239      3.223     -3.637      0.001     -18.181      -5.267
scale(games_lineups)           -13.2792      5.312     -2.500      0.015     -23.921      -2.637
scale(substitutes_out)          15.1434      5.081      2.981      0.004       4.966      25.321
scale(games_played)              9.7119      3.562      2.726      0.009       2.576      16.848
scale(shotsOnTotal_goalsTotal)  21.9015      3.461      6.328      0.000      14.969      28.834
scale(gamesAppearance_sub)     -16.6716      6.078     -2.743      0.008     -28.848      -4.496
==============================================================================
Omnibus:                       28.893   Durbin-Watson:                   2.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               75.161
Skew:                           1.298   Prob(JB):                     4.78e-17
Kurtosis:                       7.493   Cond. No.                         6.02
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.08506576350759402
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  value   R-squared:                       0.653
Model:                            OLS   Adj. R-squared:                  0.599
Method:                 Least Squares   F-statistic:                     11.94
Date:                Wed, 24 Jun 2020   Prob (F-statistic):           2.57e-10
Time:                        19:46:14   Log-Likelihood:                -290.00
No. Observations:                  67   AIC:                             600.0
Df Residuals:                      57   BIC:                             622.0
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       42.6493      2.430     17.551      0.000      37.783      47.515
scale(age)                     -16.1354      3.342     -4.828      0.000     -22.828      -9.443
scale(follower)                 11.0422      3.285      3.362      0.001       4.465      17.619
scale(passes_accuracy)          11.9719      3.369      3.554      0.001       5.226      18.717
scale(penalty_won)               6.0666      2.552      2.377      0.021       0.957      11.177
scale(games_lineups)           -15.5574      5.116     -3.041      0.004     -25.803      -5.312
scale(substitutes_out)          15.0171      4.873      3.082      0.003       5.260      24.775
scale(games_played)             12.7842      3.485      3.668      0.001       5.805      19.764
scale(shotsOnTotal_goalsTotal)  10.3727      3.806      2.725      0.009       2.750      17.995
scale(gamesAppearance_sub)     -23.3770      6.168     -3.790      0.000     -35.729      -11.025
==============================================================================
Omnibus:                       24.845   Durbin-Watson:                   2.158
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.071
Skew:                           1.148   Prob(JB):                     4.05e-13
Kurtosis:                       6.895   Cond. No.                         6.19
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.2608856168050746
```

제한점

1. 수집된 데이터 양의 한계
   - MARKET VALUE 데이터가 500명으로 한정되어 있었음
   - 공격수를 제외한 나머지 포지션에 대한 FEATURE 부족
   - 데이터 세분화의 아쉬움 (EX. 패스)

2. 데이터 보완의 필요성
   - 해당 웹사이트에 데이터 요청을 해놓은 상태
   - A리그와 B리그의 수준차를 고려한 가중치 데이터의 필요성(가중치)
   - 개인 수상실적 및 팀 우승에 대한 정량화 데이터 필요(득점왕, 월드컵 우승 등)

받는사람 <info@transfermarkt.de>

Hello,

I am a student studying data science at an academic institute based in South Korea.
Our team is currently conducting research on “Prediction of market values for football players through machine-learning/deep-learning models”.

A brief summary of our research is as follows:
- Purpose :
To find correlation between market values, performance, social data of players and create a prediction model
- Data:
Performance data from https://www.api-football.com/
Market value data from https://www.transfermarkt.com/

We have been conducting our research through various models with 500 market values from your website, however, came to meet a limitation of too few data.
Transfermarkt is a recognized institution not only in Korea, but worldwide, **we decided to sincerely ask your team for access to the market data of the players.**
Your contribution will not only help us to get more accurate result, but also, through the citation of your institution, strengthen the credibility of our research.
So, kindly review our proposal and our team will sincerely hope to hear from you soon.

Best regards,

Jeongseob Kim

3. 선수들의 MARKET VALUE 데이터 자체의 심한 유동성 (시장 자체의 VALUE 인플레이션 현상 심화)
   - 정해진 규칙이 없이 돈이 많은 구단이 월할 시 얼마든지 오버페이가 가능한 구조 (EX. 네이마르 등)

* 개선 방향 :
   - 데이터의 추가 수집(요청 상태) 및 종속 변수의 변화 (몸값 > 연봉)
     - 웹 사이트 : HTTPS://WWW.CAPOLOGY.COM/
   - 리그 가중치 데이터 추가 수집
     - 웹 사이트 : HTTPS://WWW.UEFA.COM/
   - 개인 및 팀 실적에 대한 데이터 추가 수집
     - 웹 사이트 : HTTP://WHOSCORED.COM/

선수 몸값 예측 모델에 있어서 SNS지표의 기여도 및 상관성 확인

BUT

몸값의 유동성때문에 현재 FEATURE로는 적절한 선형예측모델링의 한계가 뚜렷함

THANK YOU

DATA HAS A BETTER IDEA