축구선수의 시장 가치 예측과 SNS 지표의 기여도

고건호, 김정섭, 이왕건

개요 - 주제 선정 이유?

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

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"

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?

	데이터셋(선수 수)	모델	연구목적	연구 결과
우리연구	약 500명	선형회귀, ML	1. Salary prediction Regression analysis	
Predicting Market Value of Soccer Players Using Linear Modeling Techniques	357명	OLS, KNN, Ridge 회귀, PCR	1. predicting market value of top players using statistical modeling techniques	1. PCR(k=15)가 제일 좋은 성능
Football Player's Performance and Market Value	[transferred player] - 37명(market value train data : 실제 이적 선수) - 40명 (performance train data : 이적 시장에 올랐던 40명 데이터 활용)	Lasso regression	1. prediction for salary and performance 2. realation between salalry and performance - how market value and performance of La Liga (the Spanish League) players can be modeled using extensive public data sources. - develop regression models to predict the real market value and assess a player's performance	1. Performance, market 모델 확인 2. Market value와 performance간 관계 확인 (고평가 선수의 overpay는 marketing 비용으로 고려됨)
Machine Learning for Soccer Analytics	EPL 선수 전원	ML 알고리즘 다수	1. Performance - rating 2. Match outcome - performance 3. Match outcome - rating	
Assessing the market values of soccer players –a robust analysis of data from German 1. and 2.Bundesliga	493명	robust regression	1. extent a player's market value depends on his skills	1. performance와 club reputation이 market value에 큰 영향 2. 고/저 평가 선수 존재(marketing 등 영향 고려)
ANALYSIS OF STATISTICS IN MAJOR LEAGUE SOCCER	340게임(게임 stasts - 승점 간 관계 연구)	Linear regression(simple, multiple)	1. to analyze statistical data collected from all games (n = 340) in the 2016 Major League Soccer regular season and discover which statistics are correlated to total season points.	1. 승점 관련 경기 stats 확인
Application of Neural and Regression Models in Sports Results Prediction	Javelin 경기 대상	regression, Perceptron	1. comparing regression and neural models with respect to their accuracy of predicting sports results.	1. the investigation demonstrated a significantly greater accuracy of prediction for perceptron models.
A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position	105명	OLS regression	1. to examine if the independent variables are successful in predicting the outcome variable and which independent variables are significant predictors of the outcome	1. Market value의 주요 변수 확인
Modelling the transfer prices of football player	**424명**	**1. Forward stepwise selection 2. lasso 3. Ridgeregression**	**1. predicting transfer prices**	**1. LASSO모델의 예측성능이 제일 좋음 2. 이적료에 영향을 주는 요소들 확인**

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

ATK DATA OLS
- Domain Base
- Selected features

ALL DATA OLS
- PCA
- Backward Elimination

ATK DATA OLS
- PCA
- Backward Elimination

ATK+MDF DATA OLS
- Backward Elimination

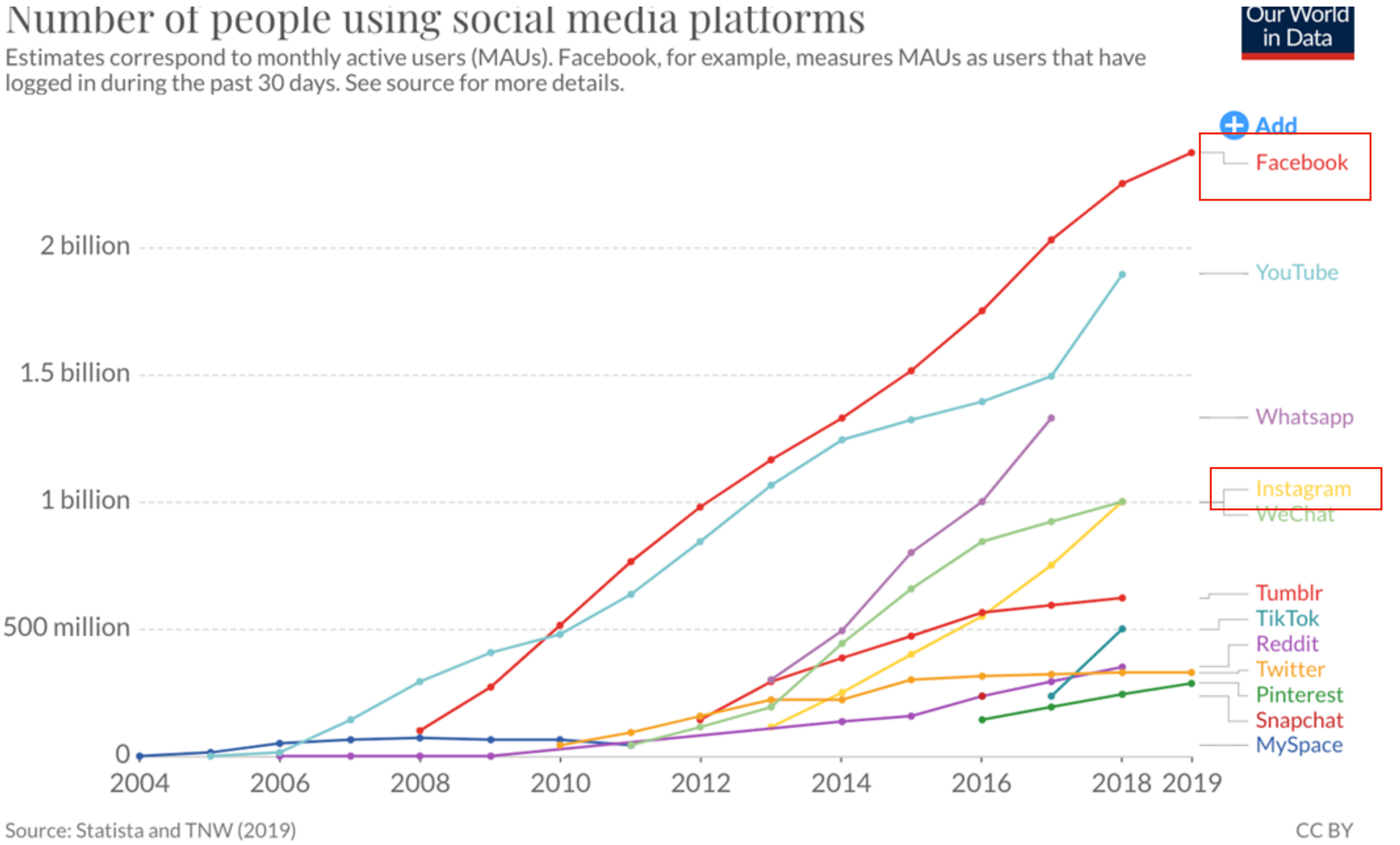
개요

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

1. 단순 사용자의 수는 페이스북이 더 많지만 참여도는 인스타그램이 훨씬 강함(페이스북은 기사당 10,000명의 반응 / 인스타그램은 150,000명의 반응)

2. 성장속도면에서 인스타그램이 2018년에 이어 2019년에도 20% 이상 성장을 거둠(사용자 수 증가)

3. 페이스북은 친구를 맺는 개념이 강했던 반면 (서로 맺어야 교류 가능)인스타그램은 자신이 좋아하면 얼마든지 팔로잉을 신청할 수 있음

Feb 1, 2018, 01:22pm EST

As Facebook Shifts, Instagram Emerges As A New Home For Brands

Ryan Holmes Contributor ⓘ
Entrepreneurs

Photographer: Krisztian Bocsi/Bloomberg

For big brands, the difference can be staggering. When Mercedes-Benz shared a post on Facebook recently about the premiere of its new A-Class, the update quickly garnered more than 10,000 Likes. Impressive ... until you consider that the very same image on Instagram generated more than 150,000 Likes—15 times the response!

5,532 views | Dec 26, 2019, 08:00am EST

Is Instagram The Social Media Service For Business In 2020?

Peter Suciu Contributor ⓘ
Social Media

With more than one billion monthly active users, Instagram ranks third after Facebook (with just over two billion active users) and YouTube (with 1.9 billion) in terms of the most popular social media network. While Facebook still has more total users, it is no secret that Instagram has a reach that shouldn't be ignored by business users.

According to the latest data compiled by UK-based service platform Superviral, Instagram could be a crucial part of a social media marketing strategy. The Facebook-owned service now ranks second as the most downloaded free app in the Apple App Store, and is the 10th most searched query on Google.

The service has seen serious growth in 2019, which could continue next year.

"With a 20% increase in users from June 18 to April 19, I believe Instagram's growth isn't slowing down any time soon," said Rabban Faruqui of Superviral.co.uk. "From a biological perspective, we can see why visual-based social media companies like Instagram and Snapchat will continue thriving as visuals are processed 60,000x faster than text in the brain, meaning consumers simply prefer images/videos over text."

Reaching Millennials And Generation Z

Instagram also has appeal to younger adults, far more than Facebook these days. According to data from SproutSocial, 64% of Instagram users are 18-29 years old – but more importantly nearly two out of every three adults in this age group use the service.

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

인스타그램 팔로워 수가 선수 개인의 인기도를 반영한다고 할 수 있고, 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				

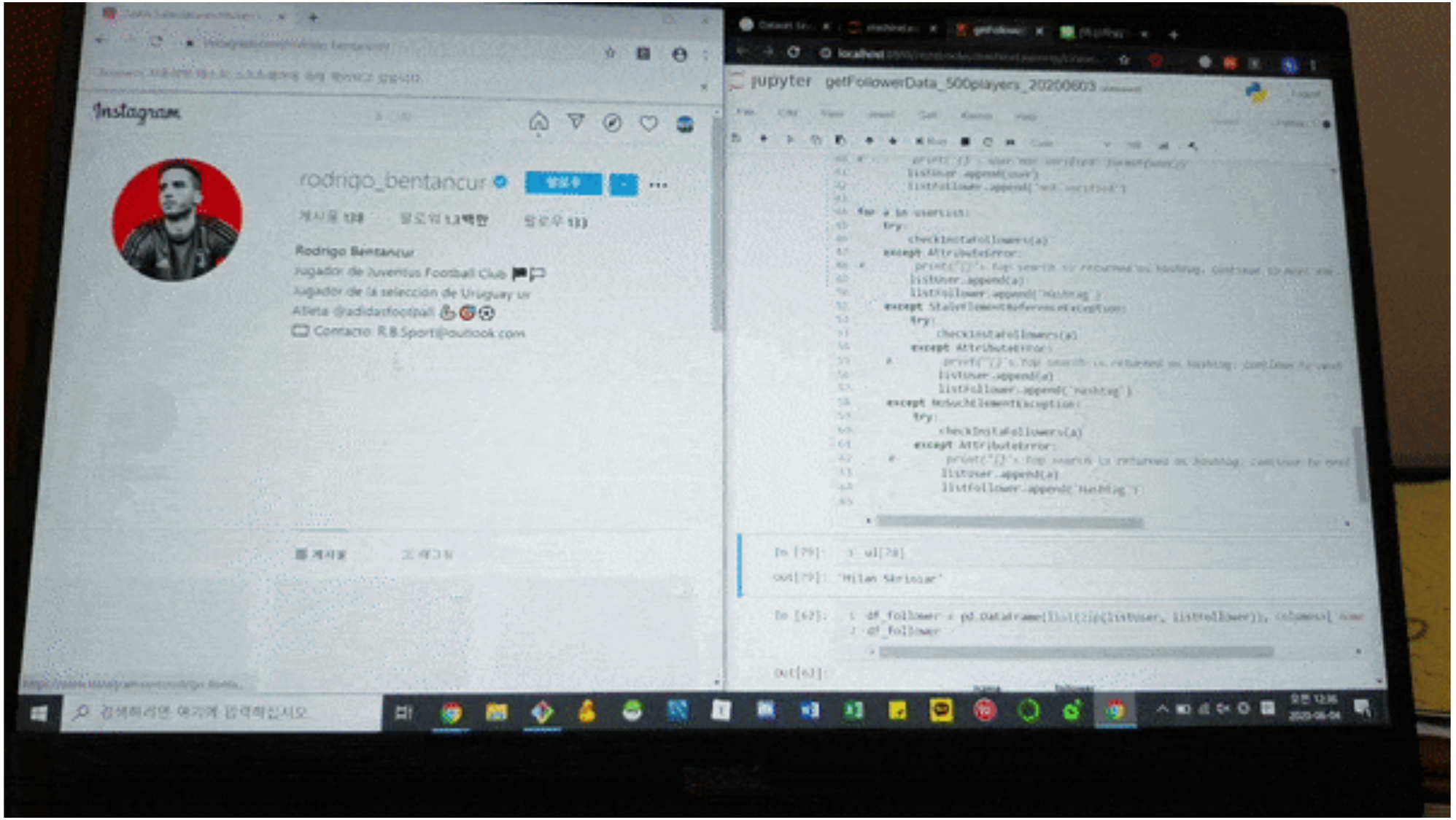
messi_messi10 Follow •••

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

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		

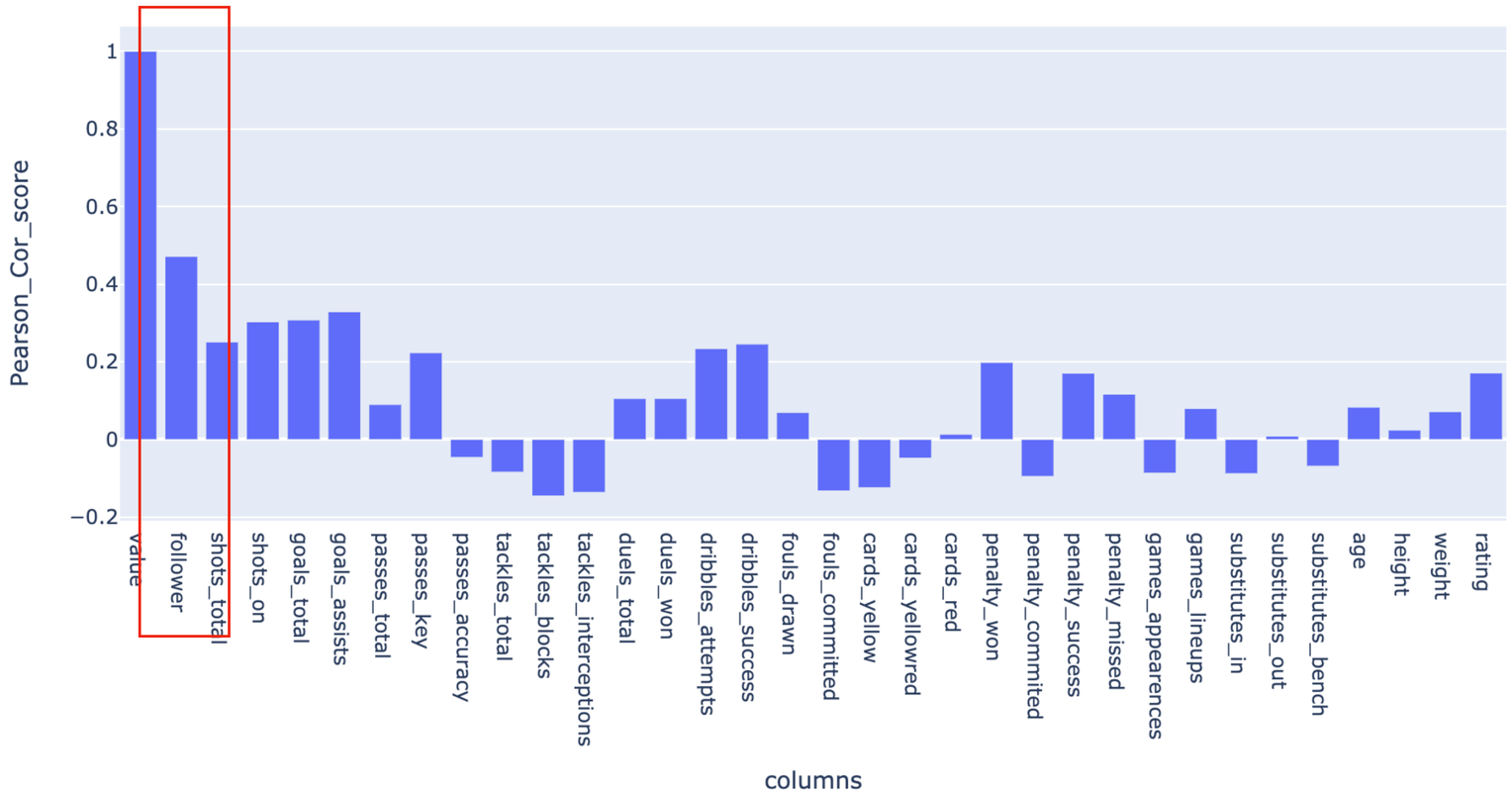
	coef	std err	t	P>|t|	[0.025	0.975]
const	128.3997	76.689	1.674	0.095	-22.717	279.517
shots_total	-5.0267	5.321	-0.945	0.346	-15.512	5.459
shots_on	-4.8542	16.065	-0.302	0.763	-36.511	26.803
goals_total	93.8629	22.586	4.156	0.000	49.357	138.369
goals_conceded	13.7809	11.252	1.225	0.222	-8.391	35.953
goals_assists	-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

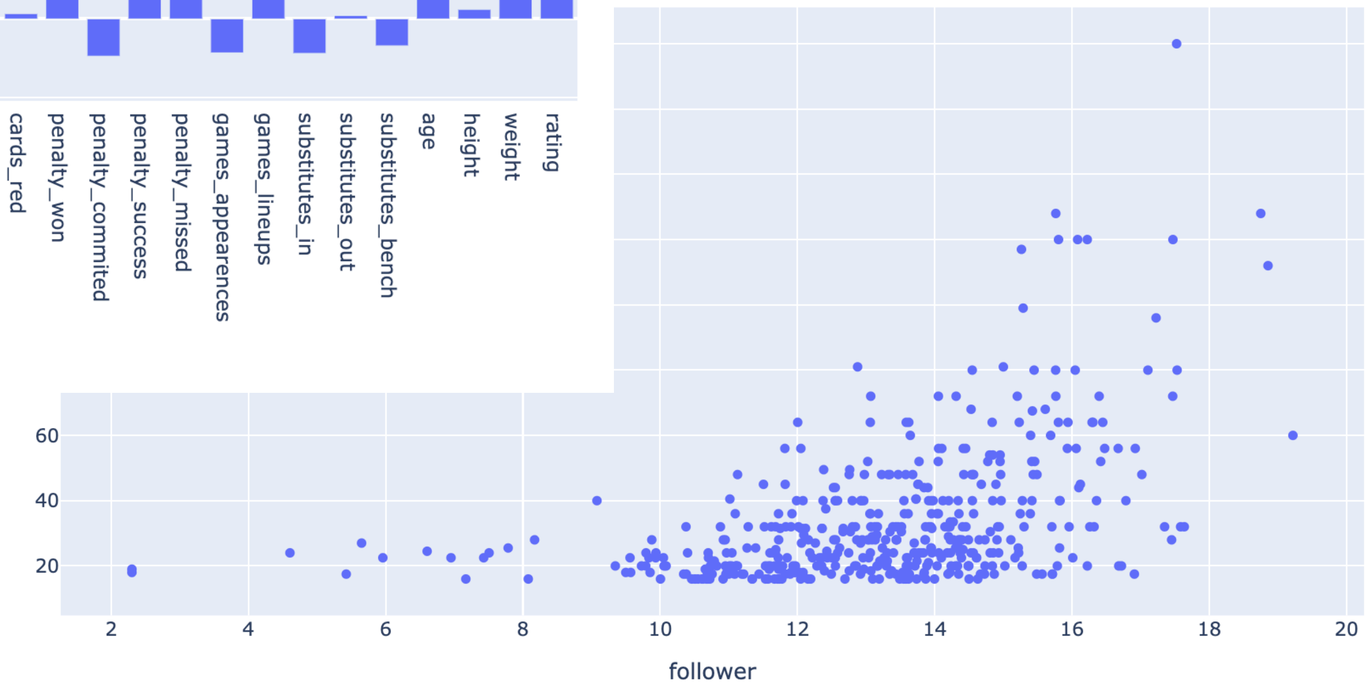
Omnibus:	69.959	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	171.789
Skew:	1.251	Prob(JB):	4.97e-38
Kurtosis:	6.108	Cond. No.	4.69e+11

```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'])
```

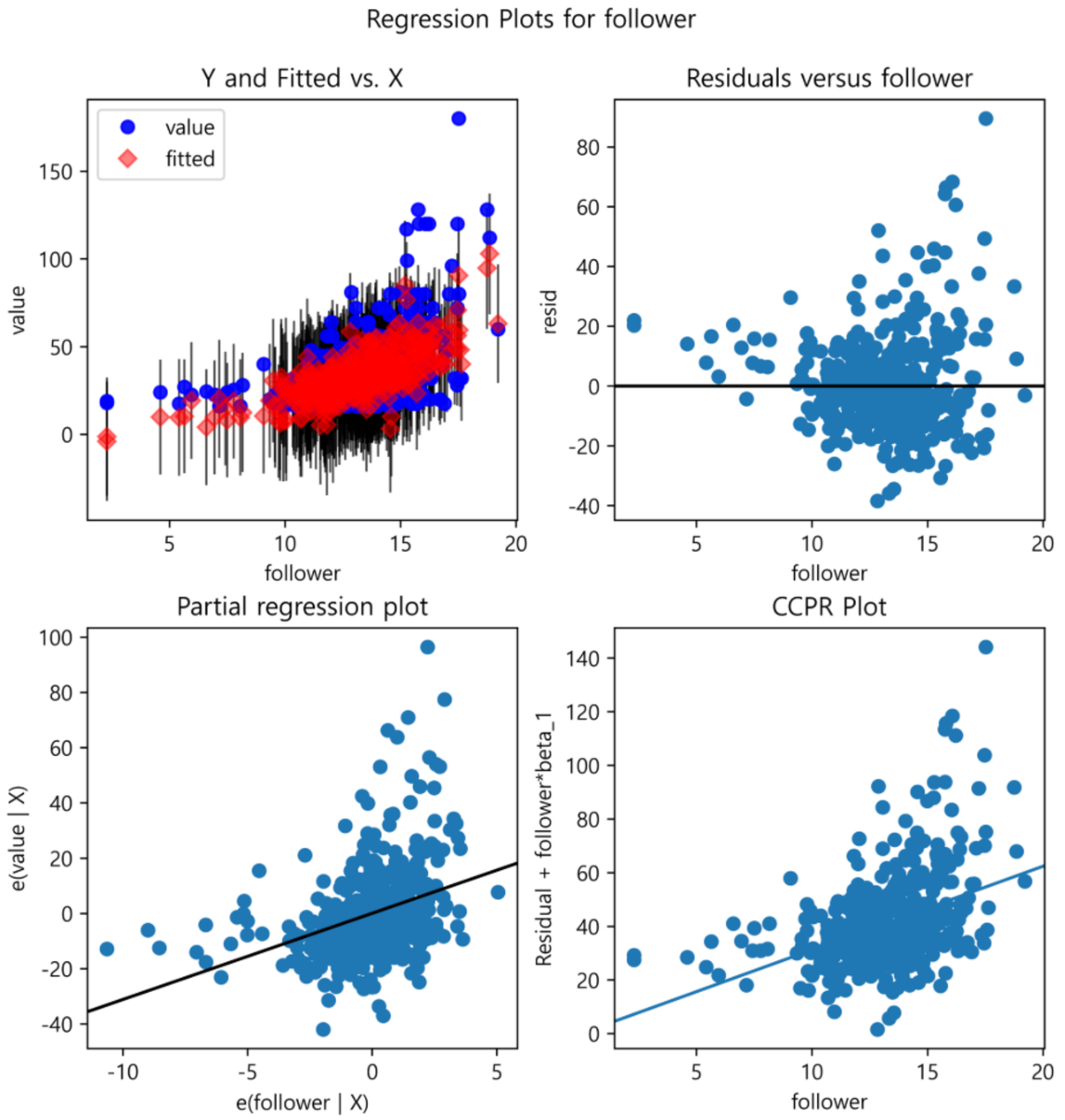
FOLLOWER 수의 P값은 0.082로 다른 변수에 비해 P값이 굉장히 낮게 나온 것을 확인

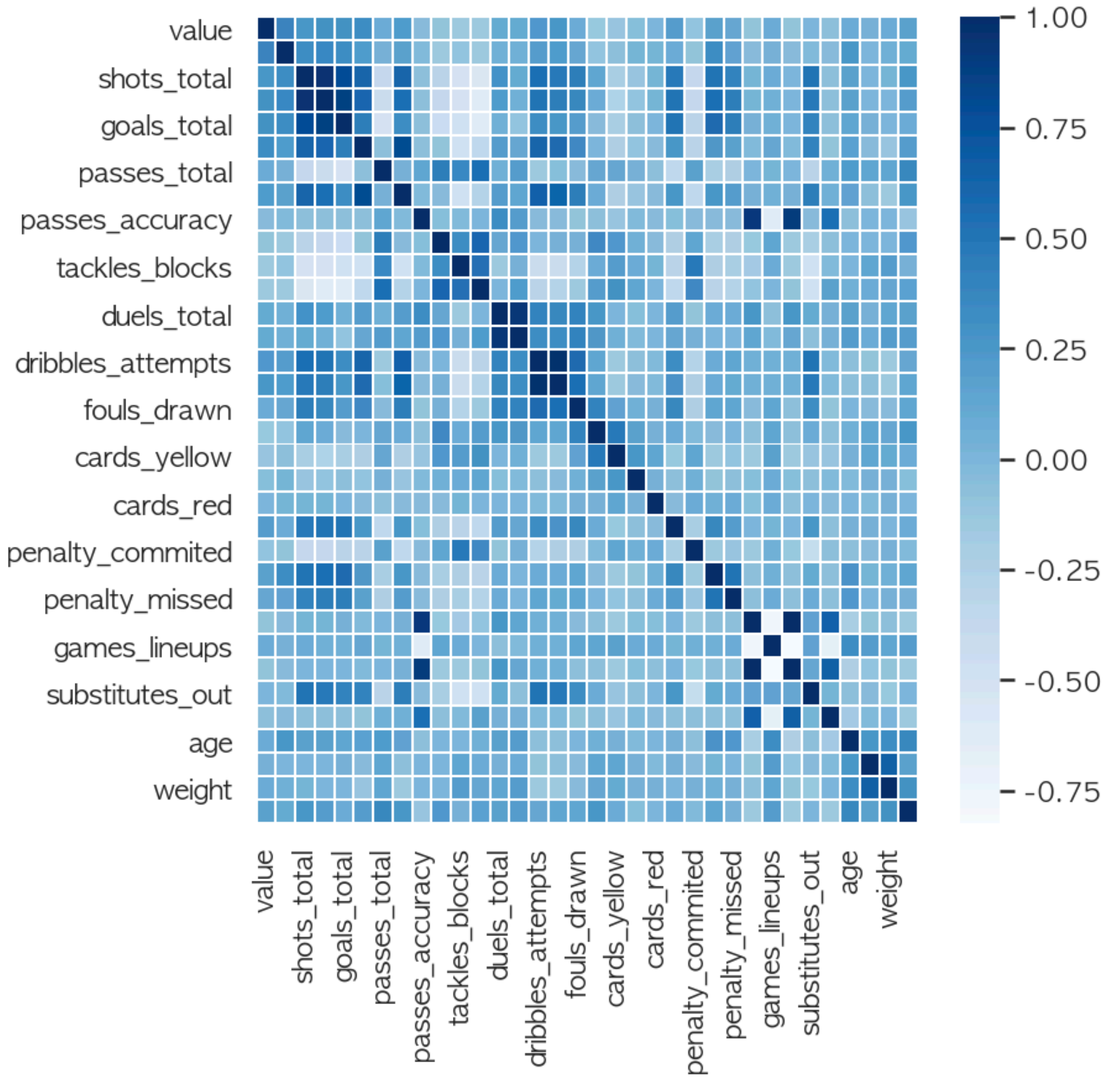
FOLLOWER수가 많고, 적음이 선수들의 몸값을 결정하는데 영향을 미칠 것이다





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



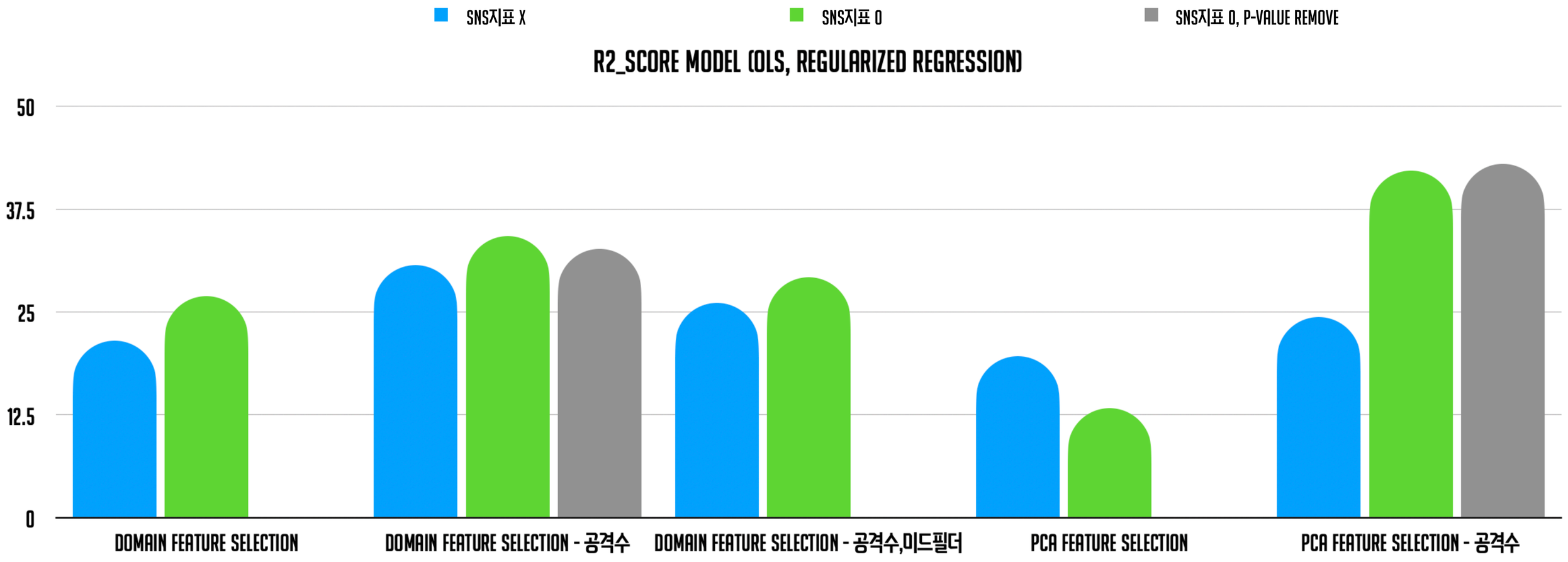
조건수와 상관관계 분석 결과 독립변수간 강한 다중공선성이 의심이 되어 주성분 분석(PCA)를 실시함

SNS지표의 CONTRIBUTION 분석을 위한 모델링

SNS지표 O

SNS지표 X

- DOMAIN BASE FEATURE SELECTION
- DOMAIN BASE FEATURE SELECTION (공격수)
- DOMAIN BASE FEATURE SELECTION (공격수+미드필더
- PCA BASE FEATURE SELECTION
- PCA BASE FEATURE SELECTION (공격수)


SNS지표 X
SNS지표 O
SNS지표 O, P-VALUE REMOVE
R2_SCORE MODEL (OLS, REGULARIZED REGRESSION)
50
37.5
25
12.5
0
DOMAIN FEATURE SELECTION
DOMAIN FEATURE SELECTION - 공격수
DOMAIN FEATURE SELECTION - 공격수,미드필더
PCA FEATURE SELECTION
PCA FEATURE SELECTION - 공격수

F 검정을 이용한 모형 비교 & 변수 중요도 비교

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	451.00	159931.88	0.00	nan	nan	nan
1	450.00	148124.19	1.00	11807.69	35.87	0.00

	sum_sq	df	F	PR(>F)
scale(follower)	11807.69	1.00	35.87	0.00
scale(goals_total)	260.74	1.00	0.79	0.37
scale(I(goals_total ** 2))	1985.36	1.00	6.03	0.01
scale(goals_assists)	5328.12	1.00	16.19	0.00
scale(duels_won)	1176.66	1.00	3.57	0.06
scale(dribbles_success)	53.06	1.00	0.16	0.69
Residual	148124.19	450.00	nan	nan

전체 선수 대상

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	84.00	50516.85	0.00	nan	nan	nan
1	83.00	46230.68	1.00	4286.17	7.70	0.01

	sum_sq	df	F	PR(>F)
scale(follower)	4286.17	1.00	7.70	0.01
scale(I(goals_total * goals_assists))	16418.36	1.00	29.48	0.00
scale(dribbles_success)	567.32	1.00	1.02	0.32
scale(age)	2746.18	1.00	4.93	0.03
scale(I(age ** 2))	2948.58	1.00	5.29	0.02
Residual	46230.68	83.00	nan	nan

공격수 대상

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	239.0	69996.876955	0.0	NaN	NaN	NaN
1	238.0	64620.287223	1.0	5376.589733	19.80227	0.000013

	sum_sq	df	F	PR(>F)
scale(age)	8213.323543	1.0	30.250113	9.778616e-08
scale(shots_on)	2632.990978	1.0	9.697448	2.071349e-03
scale(goals_total)	8696.530975	1.0	32.029792	4.347585e-08
scale(goals_assists)	4422.609018	1.0	16.288707	7.334183e-05
scale(passes_accuracy)	8622.839592	1.0	31.758383	4.917546e-08
scale(dribbles_attempts)	2415.148009	1.0	8.895120	3.156625e-03
scale(dribbles_success)	3736.710301	1.0	13.762505	2.582367e-04
scale(fouls_drawn)	1311.184541	1.0	4.829163	2.894626e-02
scale(cards_yellow)	443.908535	1.0	1.634939	2.022667e-01
scale(penalty_won)	3119.491869	1.0	11.489257	8.192536e-04
scale(penalty_success)	2848.270865	1.0	10.490335	1.371110e-03
scale(games_appearences)	2714.736128	1.0	9.998519	1.770242e-03
scale(games_played)	14007.504025	1.0	51.590392	8.712919e-12
scale(follower)	5376.589733	1.0	19.802270	1.319688e-05
Residual	64620.287223	238.0	NaN	NaN

공격수+미드필더 대상

F 검정을 이용한 모형 비교 & 변수 중요도 비교

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	448.0	146087.030421	0.0	NaN	NaN	NaN
1	447.0	139449.663696	1.0	6637.366724	21.275798	0.000005

	df	sum_sq	mean_sq	F	PR(>F)
scale(passes_total)	1.0	1574.263746	1574.263746	5.046236	2.516697e-02
scale(fouls_committed)	1.0	4071.562691	4071.562691	13.051222	3.374840e-04
scale(cards_red)	1.0	129.636305	129.636305	0.415544	5.194997e-01
scale(games_lineups)	1.0	1802.915122	1802.915122	5.779168	1.662329e-02
scale(substitutes_out)	1.0	335.865539	335.865539	1.076603	3.000197e-01
scale(age_x)	1.0	285.019110	285.019110	0.913617	3.396733e-01
scale(shotsOnTotal_goalsTotal)	1.0	36313.283741	36313.283741	116.400695	2.827373e-24
scale(dribblesAtmptsSuc)	1.0	3910.344551	3910.344551	12.534444	4.414277e-04
scale(follower)	1.0	6637.366724	6637.366724	21.275798	5.200915e-06
Residual	447.0	139449.663696	311.967928	NaN	NaN

전체 선수 대상

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	89.0	43724.431358	0.0	NaN	NaN	NaN
1	88.0	32469.346632	1.0	11255.084726	30.504077	3.335454e-07

	sum_sq	df	F	PR(>F)
scale(age)	7717.478051	1.0	20.963246	1.513244e-05
scale(I(age ** 2))	7949.782481	1.0	21.594263	1.160426e-05
scale(I(age ** 3))	8292.315533	1.0	22.524697	7.871220e-06
scale(passes_accuracy)	4921.616602	1.0	13.368754	4.325279e-04
scale(games_played)	15759.787362	1.0	42.808843	3.725049e-09
scale(shotsOnTotal_goalsTotal)	2820.611119	1.0	7.661721	6.861665e-03
scale(follower)	12658.317019	1.0	34.384214	7.562831e-08
Residual	32764.750805	89.0	NaN	NaN

공격수 대상

DOMAIN BASE FEATURE SELECTION (+ SNS지표) - 공격수

OLS Regression Results

Dep. Variable:	value	R-squared:	0.251
Model:	OLS	Adj. R-squared:	0.240
Method:	Least Squares	F-statistic:	24.03
Date:	Tue, 07 Jul 2020	Prob (F-statistic):	7.29e-21
Time:	19:16:23	Log-Likelihood:	-1542.2
No. Observations:	365	AIC:	3096.
Df Residuals:	359	BIC:	3120.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025
Intercept	31.8712	0.873	36.495	0.000	30.154
scale(goals_total)	-4.8075	2.196	-2.189	0.029	-9.127
scale(I(goals_total ** 2))	10.1332	2.068	4.900	0.000	6.067
scale(goals_assists)	3.8263	1.154	3.317	0.001	1.558
scale(duels_won)	2.0657	0.953	2.168	0.031	0.192
scale(dribbles_success)	3.1612	1.112	2.843	0.005	0.975

Omnibus:	141.565	Durbin-Watson:	1.860
Prob(Omnibus):	0.000	Jarque-Bera (JB):	463.318
Skew:	1.776	Prob(JB):	2.46e-101
Kurtosis:	7.224	Cond. No.	5.15

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

검증모델성능: 0.2150124360657613

OLS Regression Results

Dep. Variable:	value	R-squared:	0.320
Model:	OLS	Adj. R-squared:	0.308
Method:	Least Squares	F-statistic:	28.06
Date:	Tue, 07 Jul 2020	Prob (F-statistic):	1.85e-27
Time:	19:17:53	Log-Likelihood:	-1524.5
No. Observations:	365	AIC:	3063.
Df Residuals:	358	BIC:	3090.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.8712	0.833	38.250	0.000	30.233	33.510
scale(follower)	5.8099	0.964	6.029	0.000	3.915	7.705
scale(goals_total)	-3.7573	2.103	-1.787	0.075	-7.893	0.378
scale(I(goals_total ** 2))	6.8819	2.045	3.365	0.001	2.860	10.904
scale(goals_assists)	3.3618	1.103	3.047	0.002	1.192	5.532
scale(duels_won)	1.5554	0.913	1.704	0.089	-0.240	3.351
scale(dribbles_success)	2.1554	1.074	2.007	0.045	0.043	4.267

Omnibus:	158.797	Durbin-Watson:	1.853
Prob(Omnibus):	0.000	Jarque-Bera (JB):	648.569
Skew:	1.910	Prob(JB):	1.46e-141
Kurtosis:	8.297	Cond. No.	5.58

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

검증모델성능: 0.26999744253074326

DOMAIN BASE FEATURE SELECTION (+ SNS지표) - 공격수

OLS Regression Results

Dep. Variable:	value	R-squared:	0.328
Model:	OLS	Adj. R-squared:	0.298
Method:	Least Squares	F-statistic:	10.74
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	3.90e-07
Time:	14:25:35	Log-Likelihood:	-430.60
No. Observations:	93	AIC:	871.2
Df Residuals:	88	BIC:	883.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
Intercept	40.2903	2.645	15.235	0.000
scale(I(goals_total * goals_assists))	14.9395	3.535	4.227	0.000
scale(dribbles_success)	5.7061	3.142	1.816	0.073
scale(age)	63.4559	34.540	1.837	0.070
scale(I(age ** 2))	-67.5710	34.826	-1.940	0.056

Omnibus:	48.660	Durbin-Watson:	2.038
Prob(Omnibus):	0.000	Jarque-Bera (JB):	127.479
Skew:	1.928	Prob(JB):	2.08e-28
Kurtosis:	7.247	Cond. No.	28.1

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

검증모델성능: 0.3075672138353605

OLS Regression Results

Dep. Variable:	value	R-squared:	0.361
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	9.834
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	1.78e-07
Time:	14:26:23	Log-Likelihood:	-428.26
No. Observations:	93	AIC:	868.5
Df Residuals:	87	BIC:	883.7
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	40.2903	2.594	15.535	0.000	35.135	45.445
scale(follower)	9.9817	4.707	2.121	0.037	0.626	19.337
scale(I(goals_total * goals_assists))	8.1731	4.711	1.735	0.086	-1.191	17.537
scale(dribbles_success)	2.9291	3.348	0.875	0.384	-3.725	9.583
scale(age)	63.5421	33.873	1.876	0.064	-3.785	130.869
scale(I(age ** 2))	-67.4018	34.154	-1.973	0.052	-135.287	0.484

Omnibus:	55.157	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	172.457
Skew:	2.116	Prob(JB):	3.56e-38
Kurtosis:	8.157	Cond. No.	30.6

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

검증모델성능: 0.3424021825159443

DOMAIN BASE FEATURE SELECTION (+ SNS지표) - 공격수

OLS Regression Results

Dep. Variable:	value	R-squared:	0.355
Model:	OLS	Adj. R-squared:	0.326
Method:	Least Squares	F-statistic:	12.13
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	6.73e-08
Time:	14:30:52	Log-Likelihood:	-428.66
No. Observations:	93	AIC:	867.3
Df Residuals:	88	BIC:	880.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	40.2903	2.590	15.556	0.000	35.143	45.438
scale(follower)	11.5927	4.326	2.680	0.009	2.995	20.190
scale(I(goals_total * goals_assists))	8.5811	4.682	1.833	0.070	-0.723	17.886
scale(age)	65.0209	33.786	1.924	0.058	-2.122	132.164
scale(I(age ** 2))	-69.6170	34.015	-2.047	0.044	-137.215	-2.019

Omnibus:	53.438	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	158.844
Skew:	2.068	Prob(JB):	3.22e-35
Kurtosis:	7.887	Cond. No.	29.8

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

검증모델성능: 0.32708151699307325

DOMAIN BASE FEATURE SELECTION (+ SNS지표) - 공격수, 미드필더

OLS Regression Results

Dep. Variable:	value	R-squared:	0.597
Model:	OLS	Adj. R-squared:	0.565
Method:	Least Squares	F-statistic:	18.61
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	5.34e-26
Time:	16:55:31	Log-Likelihood:	-736.86
No. Observations:	177	AIC:	1502.
Df Residuals:	163	BIC:	1546.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	37.1497	1.218	30.501	0.000	34.745	39.555
scale(age)	-7.4497	1.611	-4.623	0.000	-10.631	-4.268
scale(shots_on)	-10.0646	3.603	-2.793	0.006	-17.180	-2.949
scale(goals_total)	21.9696	3.540	6.206	0.000	14.980	28.960
scale(goals_assists)	8.3245	1.841	4.523	0.000	4.690	11.959
scale(passes_accuracy)	10.1486	1.799	5.640	0.000	6.595	13.702
scale(dribbles_attempts)	-19.7844	6.240	-3.171	0.002	-32.106	-7.463
scale(dribbles_success)	23.9350	5.759	4.156	0.000	12.564	35.306
scale(fouls_drawn)	-4.4728	1.594	-2.806	0.006	-7.621	-1.325
scale(cards_yellow)	4.7496	1.482	3.205	0.002	1.823	7.676
scale(penalty_won)	4.1775	1.547	2.700	0.008	1.122	7.233
scale(penalty_success)	-5.1043	1.765	-2.893	0.004	-8.589	-1.620
scale(games_appearences)	-8.9745	2.095	-4.283	0.000	-13.112	-4.837
scale(games_played)	8.9945	1.744	5.157	0.000	5.551	12.438

Omnibus:	14.480	Durbin-Watson:	2.054
Prob(Omnibus):	0.001	Jarque-Bera (JB):	18.608
Skew:	0.539	Prob(JB):	9.11e-05
Kurtosis:	4.166	Cond. No.	14.3

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.2616066091426518

OLS Regression Results

Dep. Variable:	value	R-squared:	0.621
Model:	OLS	Adj. R-squared:	0.589
Method:	Least Squares	F-statistic:	18.99
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	2.06e-27
Time:	16:55:35	Log-Likelihood:	-731.42
No. Observations:	177	AIC:	1493.
Df Residuals:	162	BIC:	1540.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	37.1497	1.185	31.355	0.000	34.810	39.489
scale(age)	-8.1350	1.582	-5.142	0.000	-11.259	-5.011
scale(shots_on)	-11.3379	3.528	-3.214	0.002	-18.304	-4.372
scale(goals_total)	20.5038	3.474	5.903	0.000	13.644	27.363
scale(goals_assists)	8.3042	1.790	4.638	0.000	4.769	11.840
scale(passes_accuracy)	9.6587	1.757	5.497	0.000	6.189	13.129
scale(dribbles_attempts)	-16.1821	6.173	-2.621	0.010	-28.372	-3.992
scale(dribbles_success)	19.0881	5.802	3.290	0.001	7.630	30.546
scale(fouls_drawn)	-4.1767	1.553	-2.689	0.008	-7.244	-1.109
scale(cards_yellow)	3.9405	1.464	2.692	0.008	1.050	6.831
scale(penalty_won)	4.8166	1.518	3.172	0.002	1.818	7.815
scale(penalty_success)	-5.3745	1.719	-3.127	0.002	-8.768	-1.981
scale(games_appearences)	-7.7771	2.072	-3.753	0.000	-11.869	-3.685
scale(games_played)	9.2835	1.699	5.464	0.000	5.929	12.638
scale(follower)	4.9819	1.555	3.203	0.002	1.910	8.053

Omnibus:	18.254	Durbin-Watson:	2.094
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.557
Skew:	0.602	Prob(JB):	1.71e-06
Kurtosis:	4.467	Cond. No.	15.1

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.29228707508441715

PCA BASE FEATURE SELECTION (+ SNS지표)

OLS Regression Results

Dep. Variable:	value	R-squared:	0.264
Model:	OLS	Adj. R-squared:	0.247
Method:	Least Squares	F-statistic:	15.91
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	7.41e-18
Time:	17:09:35	Log-Likelihood:	-1340.8
No. Observations:	319	AIC:	2698.
Df Residuals:	311	BIC:	2728.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.7335	0.918	34.570	0.000	29.927	33.540
scale(passes_total)	5.6890	1.158	4.912	0.000	3.410	7.968
scale(fouls_committed)	-3.8471	0.962	-4.000	0.000	-5.740	-1.955
scale(games_lineups)	3.1674	1.029	3.079	0.002	1.143	5.192
scale(substitutes_out)	-4.4599	1.184	-3.766	0.000	-6.790	-2.130
scale(age_x)	-3.3037	1.115	-2.964	0.003	-5.497	-1.110
scale(shotsOnTotal_goalsTotal)	9.9777	1.307	7.632	0.000	7.405	12.550
scale(dribblesAtmptsSuc)	3.5649	1.162	3.067	0.002	1.278	5.852

Omnibus:	200.765	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2171.320
Skew:	2.442	Prob(JB):	0.00
Kurtosis:	14.811	Cond. No.	2.83

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.1960951865791673

OLS Regression Results

Dep. Variable:	value	R-squared:	0.302
Model:	OLS	Adj. R-squared:	0.284
Method:	Least Squares	F-statistic:	16.76
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	1.20e-20
Time:	17:09:40	Log-Likelihood:	-1332.3
No. Observations:	319	AIC:	2683.
Df Residuals:	310	BIC:	2717.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.7335	0.895	35.446	0.000	29.972	33.495
scale(passes_total)	5.1076	1.138	4.487	0.000	2.868	7.348
scale(fouls_committed)	-3.3312	0.946	-3.520	0.000	-5.193	-1.469
scale(games_lineups)	3.1317	1.003	3.121	0.002	1.157	5.106
scale(substitutes_out)	-3.2432	1.192	-2.720	0.007	-5.589	-0.897
scale(age_x)	-3.7631	1.093	-3.443	0.001	-5.913	-1.613
scale(shotsOnTotal_goalsTotal)	8.1020	1.354	5.984	0.000	5.438	10.766
scale(dribblesAtmptsSuc)	2.7998	1.149	2.437	0.015	0.539	5.060
scale(follower)	4.1939	1.018	4.118	0.000	2.190	6.198

Omnibus:	192.681	Durbin-Watson:	1.869
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2110.933
Skew:	2.301	Prob(JB):	0.00
Kurtosis:	14.732	Cond. No.	2.94

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.1334134872466904

PCA BASE FEATURE SELECTION (+ SNS지표) - 공격수

OLS Regression Results

Dep. Variable:	value	R-squared:	0.552
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	10.38
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	2.08e-08
Time:	14:26:33	Log-Likelihood:	-298.61
No. Observations:	67	AIC:	613.2
Df Residuals:	59	BIC:	630.8
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.6493	2.716	15.704	0.000	37.215	48.084
scale(age)	-1089.7460	457.284	-2.383	0.020	-2004.768	-174.724
scale(I(age ** 2))	2162.6000	909.310	2.378	0.021	343.075	3982.125
scale(I(age ** 3))	-1094.3785	455.514	-2.403	0.019	-2005.860	-182.897
scale(passes_accuracy)	9.4791	3.576	2.650	0.010	2.323	16.636
scale(games_played)	12.5509	3.898	3.220	0.002	4.751	20.351
scale(shotsOnTotal_goalsTotal)	19.0195	3.572	5.324	0.000	11.871	26.168
scale(gamesAppearance_sub)	-8.4193	3.664	-2.298	0.025	-15.752	-1.087

Omnibus:	18.007	Durbin-Watson:	2.095
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.781
Skew:	1.014	Prob(JB):	1.53e-06
Kurtosis:	5.342	Cond. No.	810.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.24486857950072105

OLS Regression Results

Dep. Variable:	value	R-squared:	0.694
Model:	OLS	Adj. R-squared:	0.652
Method:	Least Squares	F-statistic:	16.43
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	2.15e-12
Time:	14:30:54	Log-Likelihood:	-285.84
No. Observations:	67	AIC:	589.7
Df Residuals:	58	BIC:	609.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.6493	2.264	18.839	0.000	38.118	47.181
scale(age)	-1900.7382	412.000	-4.613	0.000	-2725.445	-1076.031
scale(I(age ** 2))	3861.0495	825.680	4.676	0.000	2208.272	5513.827
scale(I(age ** 3))	-1988.2317	416.979	-4.768	0.000	-2822.905	-1153.558
scale(passes_accuracy)	8.5854	2.986	2.875	0.006	2.608	14.563
scale(games_played)	15.1871	3.289	4.618	0.000	8.604	21.770
scale(shotsOnTotal_goalsTotal)	10.2836	3.421	3.006	0.004	3.435	17.132
scale(gamesAppearance_sub)	-4.0112	3.171	-1.265	0.211	-10.358	2.335
scale(follower)	18.1669	3.502	5.187	0.000	11.157	25.177

Omnibus:	26.907	Durbin-Watson:	2.400
Prob(Omnibus):	0.000	Jarque-Bera (JB):	67.780
Skew:	1.209	Prob(JB):	1.91e-15
Kurtosis:	7.293	Cond. No.	914.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.4228051255765283

PCA BASE FEATURE SELECTION (+ SNS지표) - 공격수

OLS Regression Results

Dep. Variable:	value	R-squared:	0.685
Model:	OLS	Adj. R-squared:	0.648
Method:	Least Squares	F-statistic:	18.36
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	1.02e-12
Time:	14:26:01	Log-Likelihood:	-286.75
No. Observations:	67	AIC:	589.5
Df Residuals:	59	BIC:	607.1
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.6493	2.275	18.744	0.000	38.096	47.202
scale(age)	-2000.7546	406.397	-4.923	0.000	-2813.953	-1187.556
scale(I(age ** 2))	4066.0233	813.740	4.997	0.000	2437.734	5694.312
scale(I(age ** 3))	-2093.2343	410.710	-5.097	0.000	-2915.064	-1271.405
scale(passes_accuracy)	6.7495	2.623	2.573	0.013	1.501	11.999
scale(games_played)	16.5140	3.133	5.271	0.000	10.245	22.783
scale(shotsOnTotal_goalsTotal)	9.7725	3.414	2.862	0.006	2.940	16.605
scale(follower)	19.3545	3.391	5.707	0.000	12.569	26.140

Omnibus:	28.242	Durbin-Watson:	2.427
Prob(Omnibus):	0.000	Jarque-Bera (JB):	70.415
Skew:	1.286	Prob(JB):	5.12e-16
Kurtosis:	7.314	Cond. No.	881.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
모델 성능 : 0.4302133004566702

선수 몸값 예측 모델에 있어서 SNS지표의 기여도 및 상관성 확인

BUT

몸값의 유동성때문에 현재 FEATURE로는 적절한 선형예측모델링의 한계가 뚜렷함

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)

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

* 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/

THANK YOU

DATA HAS A BETTER IDEA