

유럽 주요 축구 리그 프로 선수들의 시장가치 결정 요인에 대한 연구

서강대학교 경영학과

200200977 한영태

<목차>

I. 서론	2-3
1. 연구의 동기 및 목적	
2. 연구 설계 및 가설 설정	
II. 모형 설정	3-10
1. 모형 설정	
2. 자료 설명	
III. 회귀분석 및 해석	11-12
IV. 해석 및 결론	12-13

I. 서론

1. 연구 동기 및 목적

축구는 전세계적으로 매우 인기 있는 스포츠이며, 특히 유럽의 주요 축구 리그들은 북아메리카의 여러 스포츠 리그들과 더불어 전세계에서 가장 규모가 큰 스포츠 시장 중 하나이다. 이러한 주요 리그에서 활동하는 선수들은 매우 높은 시장가치를 가지고 있으며, 그 가치는 다양한 요인에 의해 결정된다. 선수들의 시장 가치가 어떤 요인들에 의해 결정되는지에 대한 궁금증으로 “유럽 주요 축구 리그 프로 선수들의 시장가치 결정 요인에 대한 연구”를 연구의 주제로 선정하게 되었다.

해당 연구의 목적은 유럽 주요 축구 리그의 프로 선수들의 시장가치를 결정하는 주요 요인들을 파악하는 것이다. 이를 통해 선수들의 시장가치에 영향을 미치는 핵심적인 요소들을 도출하고, 이를 기반으로 한 정량적인 모델을 개발하고자 한다. 이러한 정량적인 모델은 선수들 개인의 가치를 넘어 각 구단 및 리그의 가치를 산정하는 데 이용할 수 있으며, 구단과 선수 간의 급여 협상에서도 활용될 수 있다. 추가로 축구 시장에서 선수의 가치를 정확하게 평가하고 결정함으로써, 이적 시장에서의 과도한 가격 거품이나 불공정한 거래를 방지하여 축구 산업의 건강한 발전에도 이바지할 수 있다.

2. 연구 설계 및 가설 설정

가. 연구 설계

각 선수의 시장 가치 (market value)를 종속변수로 설정하고 시장 가치에 영향을 줄 것으로 예상되는 feature들을 독립변수로 설정하여 다중회귀분석을 진행한다. 다중회귀분석의 결과를 토대로 독립변수를 제거 및 추가하는 과정을 거쳐 변수들이 유의미성을 판단한다. 독립변수로는 크게 신체적인 조건을 나타내는 변수, 인구통계학적 변수, 그리고 실제 경기에서의 performance를 나타내는 변수를 설정할 수 있다.

나. 가설 설정

선수들 개인의 시장가치에 영향을 줄 것으로 예상되는 독립변수를 선정하고 이 독립변수와 종속변수인 시장가치 사이의 관계를 고려해 가설을 설정할 수 있다.

- 1) 키가 클수록 경합에 유리하기 때문에 시장가치가 높을 것이다.
- 2) 오른발잡이보다는 왼발잡이가, 왼발잡이보다는 양발잡이가 시장가치가 높을 것이다.
- 3) 나이가 어릴수록 선수 생활을 오래 할 수 있기 때문에 시장가치가 높을 것이다.
- 4) 자국 선수보다는 해외에서 데려온 외국인 선수의 시장가치가 높을 것이다.
- 5) 득점이나 어시스트를 더 많이 기록한 선수가 시장가치가 높을 것이다.
- 6) 골 결정력이 높은 선수가 시장가치가 높을 것이다.
- 7) 패스나 태클, 또는 드리블의 성공률이 더 높은 선수가 시장가치가 높을 것이다.
- 8) 골키퍼의 경우 선방율이나 클린시트(무실점)를 기록한 선수가 시장가치가 높을 것이다.

II. 모형 설정

1. 모형 설정

앞서 연구설계에서 언급했듯이 독립변수로는 크게 1) 신체적인 조건을 나타내는 변수, 2) 인구통계학적 변수, 그리고 3) 실제 경기에서의 performance를 나타내는 변수를 설정할 수 있다. 신체적인 조건을 나타내는 변수로는 ① 신장 (HEIGHT) 변수와 ② 주발 (FOOT) 변수가 있다. 인구통계학적 변수로는 ③ 나이 (AGE)와 ④ 국적 (NATIONALITY) 변수가 있다. 마지막으로 실제 경기에서의 performance를 나타내는 변수로는 ⑤ 90분당 득점 (GOAL), ⑥ 90분당 어시스트 (ASSIST), ⑦ 총 출전 시간 (MINUTES), ⑧ 골 결정력 (SCORINGABILITY), ⑨ 패스 성공률 (PASS), ⑩ 태클 성공률 (TACKLE), ⑪ 드리블 성공률 (DRIBBLE), ⑫ 선방률(SAVE), ⑬ 클린시트 비율 (CLEANSHEET)이 있다.

위의 변수 중 주발(FOOT) 변수와 국적(NATIONALITY) 변수는 정성적 변수이기 때문에 다음과 같이 더미변수처리를 해준다.

	LEFT	RIGHT
왼발잡이	1	0
오른발잡이	0	1
양발잡이	0	0

[표 1] 주발에 따른 가변수

	NATIONALITY
자국 선수	0
외국인 선수	1

[표 2] 국적에 따른 가변수

[추정 모형]

$$\begin{aligned} \text{VALUE}_i = & \beta_{0i} + \beta_{1i}\text{HEIGHT}_i + \beta_{2i}\text{LEFT}_i + \beta_{3i}\text{RIGHT}_i + \beta_{4i}\text{AGE}_i + \beta_{5i}\text{NATIONALITY}_i + \beta_{6i}\text{GOAL}_i + \\ & \beta_{7i}\text{ASSIST}_i + \beta_{8i}\text{MINUTES}_i + \beta_{9i}\text{SCORINGABILITY}_i + \beta_{10i}\text{PASS}_i + \beta_{11i}\text{TACKLE}_i + \beta_{12i}\text{DRIBBLE}_i + \\ & \beta_{13i}\text{SAVE}_i + \beta_{14i}\text{CLEANSHEET}_i + \varepsilon_i \end{aligned}$$

축구 시장에서 수비수나 골키퍼보다는 공격수의 가치가 더 높게 평가되는 경향이 있다. 따라서 GOAL, ASSIST, SCORINGABILITY 등의 변수가 시장가치에 큰 영향을 줄 것이라고 예상해볼 수 있다. 반대로 수비수나 골키퍼를 평가할 때 많이 사용되는 변수인 TACKLE, SAVE 등의 변수는 시장가치에 주는 영향이 작을 것이라고 예상해볼 수 있다. 또한 GOAL변수와 SCORINGABILITY 변수, SAVE 변수와 CLEANSHEET 변수는 다중공선성 문제를 발생시킬 수 있는 가능성이 있다. 또한 앞서 설정했었던 가설들을 고려해보았을 때 LEFT와 RIGHT 변수를 제외한 나머지 변수들은 모두 종속변수와 양의 상관관계, 즉 회귀계수의 값이 양수일 것이라고 예측해볼 수 있다.

2. 자료 설명

가. 자료 설명 및 전처리 과정

본 연구에서는 2019-20시즌 유럽 5대 리그 (잉글리시 프리미어리그, 스페인 라 리가, 독일 분데스리가, 이탈리아 세리에 A, 프랑스 리그앙)의 선수 데이터를 활용했다. 해당 데이터는 kaggle에 올라와 있는 transfermarkt 자료를 기반으로 한다.¹

```
df=pd.read_csv('transfermarkt_fbref_201920.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2644 entries, 0 to 2643
Columns: 400 entries, Column1 to Season
dtypes: float64(239), int64(152), object(9)
memory usage: 8.1+ MB
```

[이미지 1] 원본 데이터 정보

[이미지 1]을 통해 원본 데이터는 400개의 feature를 가진 2644개의 횡단면 데이터임을 확인할 수 있다. 데이터의 크기가 아주 크지는 않은 관계로 특별한 샘플링 과정을 거치지 않고 모집단을 그대로 사용했다.

```
data=df[['height', 'foot', 'age', 'nationality', 'league', 'goals', 'goals_per90',
         'assists_per90', 'minutes', 'xg', 'passes_pct', 'tackles_won',
         'dribbles_completed_pct', 'save_pct', 'clean_sheets_pct', 'value']]
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2644 entries, 0 to 2643
Data columns (total 16 columns):
```

[이미지 2] 필요한 변수만 추출한 데이터

앞서 추정한 모형에서는 원본 데이터의 모든 feature가 필요한 것은 아니다. 따라서 [이미지 2]와

¹ https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csv

같이 원본 데이터에서 필요한 변수들만 추출하여 새로운 데이터를 만들어냈다.

[이미지 2]의 필요한 변수만 추출한 데이터를 기반으로 추정 모형을 만들기 위해 첫 번째로 파생변수를 생성한 후 불필요한 변수들을 제거하는 과정을 거쳤다. 우선 SCORINGABILITY 변수를 실제 득점(goals)에서 기대 득점(xg)을 뺀 값으로 만들어 주었다. 이후, SCORINGABILITY 다중공선성 문제가 발생할 수 있기 때문에 goal 변수와 xg 변수를 제거해주었다. 두 번째로 정성적 데이터를 더미변수처리했다. NATIONALITY 변수는 선수와 리그의 국적이 같은 경우 값을 0으로, 아닌 경우 1로 처리했다. FOOT 변수 역시 LEFT와 RIGHT로 바꾸어 더미변수처리를 해주었다. 세 번째로 오기입된 데이터를 제거해주었다. 결측치는 없었으나 HEIGHT 변수 또는 AGE 변수가 0인 데이터가 12개 존재하여 이 데이터를 제거했다. 마지막으로 모든 변수를 앞서 설정한 추정 모형의 변수명으로 바꾸어주었다. 그 결과는 다음과 같다.

	VALUE	HEIGHT	LEFT	RIGHT	AGE	NATIONALITY	GOAL	ASSIST	MINUTES	SCORINGABILITY	PASS	TACKLE	DRIBBLE	SAVE	CLEANSHEET
0	4000000	178	0	1	23	0	0.04	0.04	2099	0.1	71.2	22	53.8	0.0	0.0
1	4000000	188	0	1	22	1	0.06	0.13	1429	-2.2	67.9	5	59.1	0.0	0.0
2	4000000	183	0	1	25	0	0.00	0.07	1293	-0.3	80.7	14	55.2	0.0	0.0
3	4000000	172	1	0	23	0	0.00	0.03	2663	-1.3	71.2	38	40.0	0.0	0.0
4	1000000	188	0	1	25	1	0.08	0.00	2121	1.3	79.5	16	100.0	0.0	0.0
...
2639	12000000	165	0	1	23	1	0.31	0.00	289	-0.6	70.5	2	69.2	0.0	0.0
2640	4000000	188	1	0	29	1	0.07	0.03	2705	-0.7	81.3	30	77.8	0.0	0.0
2641	25000000	178	0	1	23	1	0.14	0.31	2605	0.5	80.9	3	73.9	0.0	0.0
2642	6000000	184	0	1	22	1	0.00	0.00	163	0.0	77.9	3	100.0	0.0	0.0
2643	9000000	174	1	0	20	1	0.00	0.00	652	-0.3	74.7	13	43.9	0.0	0.0

2632 rows x 15 columns

[표 3] 전처리한 데이터

나. 자료의 기본 통계치

	VALUE	HEIGHT	AGE	GOAL	ASSIST	MINUTES	SCORINGABILITY	PASS	TACKLE	DRIBBLE	SAVE	CLEANSHEET
count	2.632000e+03	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00	2632.00
mean	9.612080e+06	182.28	25.37	0.12	0.08	1290.51	0.02	77.26	14.12	55.47	0.05	1.68
std	1.492289e+07	6.54	4.43	0.31	0.14	945.58	1.14	11.15	13.98	28.91	0.18	7.84
min	5.000000e+01	162.00	14.00	0.00	0.00	1.00	-5.10	0.00	0.00	0.00	0.00	0.00
25%	1.000000e+06	178.00	22.00	0.00	0.00	436.00	-0.40	72.00	2.00	45.50	0.00	0.00
50%	4.000000e+06	183.00	25.00	0.00	0.00	1186.50	0.00	78.20	10.00	60.00	0.00	0.00
75%	1.200000e+07	187.00	28.00	0.15	0.11	2051.25	0.20	84.50	22.00	73.30	0.00	0.00
max	1.800000e+08	202.00	41.00	10.00	1.76	3420.00	9.00	100.00	89.00	100.00	1.00	100.00

[표 4] 정량적 변수들의 기본 통계치

[표 4]는 각 정량적 변수들의 평균, 표준편차, 최소값, 최대값, 그리고 사분위수를 포함한 기본 통계치이다.

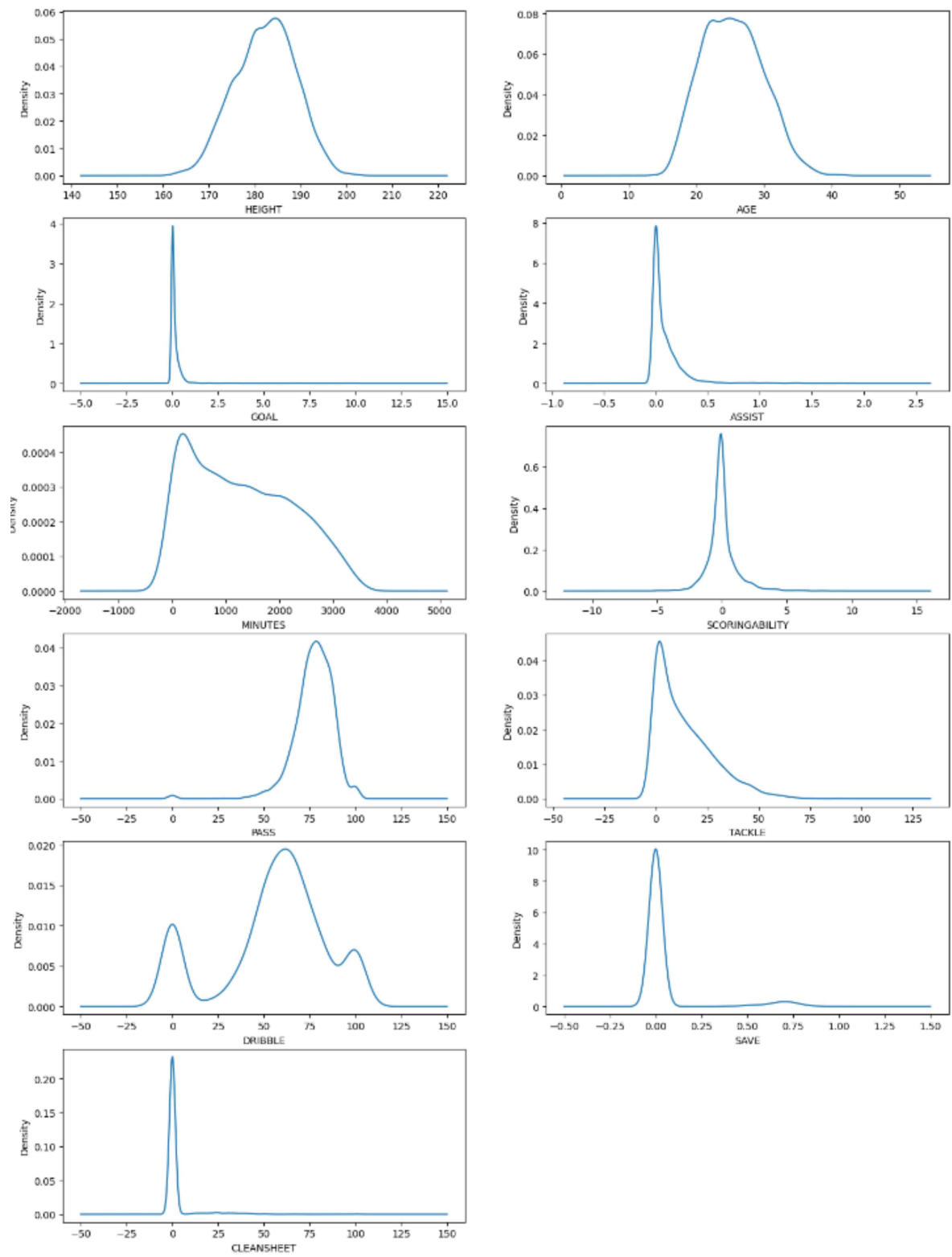
		VALUE							
		count	mean	std	min	25%	50%	75%	max
NATIONALITY									
0	1211.0	7297867.65	12802193.94	50.0	1000000.0	3000000.0	8500000.0	180000000.0	
1	1421.0	11584291.33	16263185.94	100.0	2000000.0	6000000.0	15000000.0	128000000.0	

[표 5] NATIONALITY 변수의 VALUE 변수에 대한 통계치

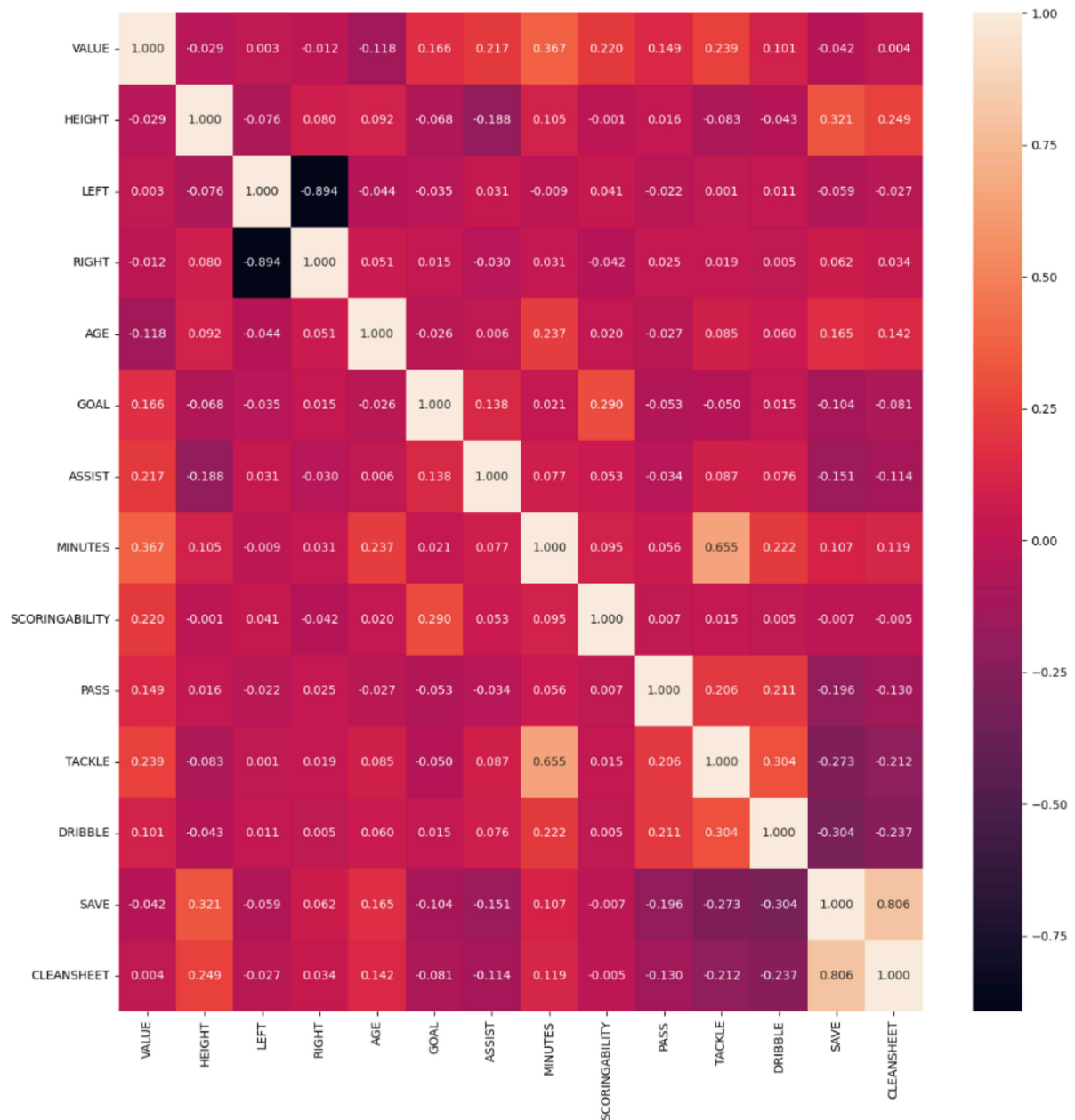
	value							
	count	mean	std	min	25%	50%	75%	max
foot								
both	102.0	12453382.35	19540261.00	10000.0	1000000.0	3500000.0	14500000.0	81000000.0
left	656.0	9628934.83	14370160.95	250.0	1000000.0	4000000.0	12000000.0	120000000.0
right	1870.0	9473400.78	14825772.12	50.0	1000000.0	4000000.0	12000000.0	180000000.0

[표 6] FOOT 변수의 VALUE 변수에 대한 통계치

[표 5]와 [표 6]은 각각 NATIONALITY 변수와 FOOT 변수의 VALUE 변수에 대한 통계치이다. 우선 NATIONALITY 변수는 앞서 설정한 자국 선수보다 외국인 선수의 시장 가치가 더 높을 것이라는 가설과 같이 외국인 선수의 시장가치가 자국 선수의 시장가치보다 약 1.5배 높다는 결과가 나타났다. FOOT 변수의 경우 역시 양발잡이 선수, 왼발잡이 선수, 오른발잡이 순으로 시장가치가 높을 것이라는 가설과 일치하는 결과가 나타났다. 다만, 왼발잡이 선수와 오른발잡이 선수의 시장가치 차이는 아주 크게 나타나지는 않았다.



[이미지 1] 각 독립변수별 분포 밀도 함수



[이미지 3] 독립변수 간 상관계수 표

[이미지 1]은 각 독립변수별 분포 밀도함수, [이미지 2]는 종속변수에 대한 독립변수별 막대 그래프, [이미지 3]은 독립변수 간 상관계수 표이다. [이미지 2]를 보면 GOAL, TACKLE 등의 변수는 종속변수와 양의 상관관계, AGE 변수는 음의 상관관계를 가지는 경향성을 볼 수 있다. [이미지 3]을 보면 SAVE 변수와 CLEANSHEET 변수가 아주 높은 다중공선성을 가짐을 알 수 있다.

III. 회귀분석

Model 1				
	coefficient	std. error	p-value	$\overline{R^2}$
const	1.66844e+07	7.67232e+06	0.0297**	0.282300
HEIGHT	-66362.9	40841.4	0.1043	
LEFT	-1.96605e+06	1.27902e+06	0.1244	
RIGHT	-1.64159e+06	1.21516e+06	0.1768	
AGE	-718262	58201.7	4.74e-034***	
NATIONALITY	3.31573e+06	500215	4.10e-011***	
GOAL	3.86505e+06	841955	4.63e-06***	
ASSIST	1.84571e+07	1.89269e+06	4.29e-022***	
MINUTES	6692.04	394.930	3.42e-061***	
SCORINGABILITY	1.96671e+06	226641	6.97e-018***	
PASS	179201	23341.1	2.28e-014***	
TACKLE	-71030.8	27061.7	0.0087***	
DRIBBLE	-5828.24	9442.64	0.5371	
SAVE	-3.74680e+06	2.55857e+06	0.1432	
CLEANSHEET	108936	53429.6	0.041**	

[표 7] Model 1 회귀분석 결과

[Model 1]

$$\text{VALUE}_i = \beta_{0i} + \beta_{1i}\text{HEIGHT}_i + \beta_{2i}\text{LEFT}_i + \beta_{3i}\text{RIGHT}_i + \beta_{4i}\text{AGE}_i + \beta_{5i}\text{NATIONALITY}_i + \beta_{6i}\text{GOAL}_i + \beta_{7i}\text{ASSIST}_i + \beta_{8i}\text{MINUTES}_i + \beta_{9i}\text{SCORINGABILITY}_i + \beta_{10i}\text{PASS}_i + \beta_{11i}\text{TACKLE}_i + \beta_{12i}\text{DRIBBLE}_i + \beta_{13i}\text{SAVE}_i + \beta_{14i}\text{CLEANSHEET}_i + \varepsilon_i$$

처음 모형을 설정했던대로 모든 변수를 넣은 상태로 회귀분석을 진행했다. 예상과 다르게 HEIGHT, DIRBBLE, SAVE 등의 변수가 유의성이 낮은 것으로 나타나 유의성이 낮은 것으로 나타난 변수를 제거하고 다시 회귀분석을 진행했다.

Model 2				
	coefficient	std. error	p-value	$\overline{R^2}$
const	2.84660e+06	2.29595e+06	0.2151	0.281420
AGE	-728949	57926.2	2.65e-035***	
NATIONALITY	3.31128e+06	498451	3.72e-011***	
GOAL	4.10488e+06	836438	9.79e-07***	
ASSIST	1.91776e+07	1.86151e+06	2.00e-024***	
MINUTES	6447.48	381.629	7.28e-06***	
SCORINGABILITY	1.95870e+06	226379	8.64e-018***	
PASS	179262	22898.3	7.10e-015***	
TACKLE	-57209.2	26062.7	0.0282**	
CLEANSHEET	43099.2	35044.4	0.2189	

[표 8] Model 2 회귀분석 결과

[Model 2]

$$\text{VALUE}_i = \beta_{0i} + \beta_{4i}\text{AGE}_i + \beta_{5i}\text{NATIONALITY}_i + \beta_{6i}\text{GOAL}_i + \beta_{7i}\text{ASSIST}_i + \beta_{8i}\text{MINUTES}_i + \beta_{9i}\text{SCORINGABILITY}_i + \beta_{10i}\text{PASS}_i + \beta_{11i}\text{TACKLE}_i + \beta_{14i}\text{CLEANSHEET}_i + \varepsilon_i$$

유의성이 낮은 변수들을 제거한 후 다시 회귀분석을 진행했을 때 CLEANSHEET 변수가 유의하지 않은 것으로 나타났다. SAVE 변수와 다중공선성 문제가 있었던 것의 영향으로 파악된다. 이후 CLEANSHEET 변수를 제거하고 다시 회귀분석을 진행했다.

Model 3				
	coefficient	std. error	p-value	$\overline{R^2}$
const	2.90901e+06	2.29561e+06	0.2152	0.281280
AGE	-723019	57730.8	5.42e-035***	
NATIONALITY	3.30885e+06	498496	3.86e-011***	
GOAL	3.99648e+06	831863	1.64e-06***	
ASSIST	1.89445e+07	1.85202e+06	4.14e-024***	
MINUTES	6602.48	360.251	1.07e-070***	
SCORINGABILITY	1.95687e+06	226396	9.31e-018***	
PASS	177428	22852.0	1.17e-014***	
TACKLE	-68989.7	24240.9	0.0045***	

[표 9] Model 3 회귀분석 결과

[Model 3]

$$\text{VALUE}_i = \beta_{0i} + \beta_{4i}\text{AGE}_i + \beta_{5i}\text{NATIONALITY}_i + \beta_{6i}\text{GOAL}_i + \beta_{7i}\text{ASSIST}_i + \beta_{8i}\text{MINUTES}_i + \beta_{9i}\text{SCORINGABILITY}_i + \beta_{10i}\text{PASS}_i + \beta_{11i}\text{TACKLE}_i + \varepsilon_i$$

CLEANSHEET 변수를 제거한 후 다시 회귀분석을 진행했을 때 모든 변수가 유의함을 확인할 수 있었다.

IV. 해석 및 결론

회귀분석을 진행한 결과 최초에 설정했던 가설과 일치했던 것도 있었고 그렇지 않을 것도 있었다. 우선 신체적인 조건인 키와 주발의 경우 시장가치와 유의하지 않은 것으로 나타났다. 반면, 인구통계학적 변수였던 나이와 국적의 경우 유의한 변수인 것으로 나타났다. AGE 변수의 회귀계

수가 음수로 나온 것을 보아 나이가 들수록 시장가치가 떨어진다는 가설이 참임을 알 수 있다. 또한 NATIONALITY 변수가 양수로 나온 것으로 보아 자국 선수보다 외국인 선수의 시장가치가 더 높은 것을 알 수 있다. 실제 경기에서의 performance와 관련된 변수 중에서는 유의한 변수도, 유의하지 않은 변수도 존재했다. 특히 SAVE 변수와 CLEANSHEET 변수가 유의하지 않은 것으로 밝혀졌다. 처음 가설에서 골키퍼의 경우 이 두 변수와 시장가치가 양의 상관관계를 가질 것으로 예상했다. 다만, 표본에서 골키퍼가 차지하는 비중이 극히 적었던 영향으로 두 변수의 유의성이 떨어진 것으로 보인다. TACKLE 변수의 경우 모델에서 유의한 변수임과는 별개로 회귀계수가 음수로 나오면서 태클 성공률이 높을수록 시장가치가 높을 것이라는 가설과는 반대되는 결과가 나왔다. 수비수보다는 공격수와 미드필더의 시장가치가 더 높은 축구 산업의 특성이 반영된 것으로 보인다. 만약 회귀 분석을 포지션별로 따로 진행했다면 수비수 표본에서 TACKLE 변수가, 골키퍼 표본에서 SAVE와 CLEANSHEET 변수가 기존의 가설과 일치하는 결과가 나타날 것으로 예상된다.

해당 연구를 통해 유럽 주요 축구 리그 프로 선수들의 시장가치에 어떤 요인들이 영향을 미치는지 알아보았다. 이 모델을 활용해 선수와 구단 간 급여 협상을 진행하기 용이할 것이며, 축구 시장이 과열되지 않고 정확한 시장 가치를 평가하는 데 도움이 될 것으로 보인다.