Soccer Players Machine Learning

Position Classifier | Player's value Regression

FIFA 2021 Soccer Players Dataset

선수들의 포지션 분류와 가치 예측

팀 내에서 효율적으로 포지션을 지정할 수 있고, 자신에 대한 가치를 확인할 수 있음.

FIFA 2021 Soccer Players Dataset

Classification

Target : position

Features

선수들의 주로 사용하는 발
기술 정확도
잠재력
체형
...

18944 rows
X
106 columns

Regression

Target : value_eur

Features

선수들의 주로 사용하는 발
기술 정확도
잠재력
체형
...

Data source : https://sofifa.com/player

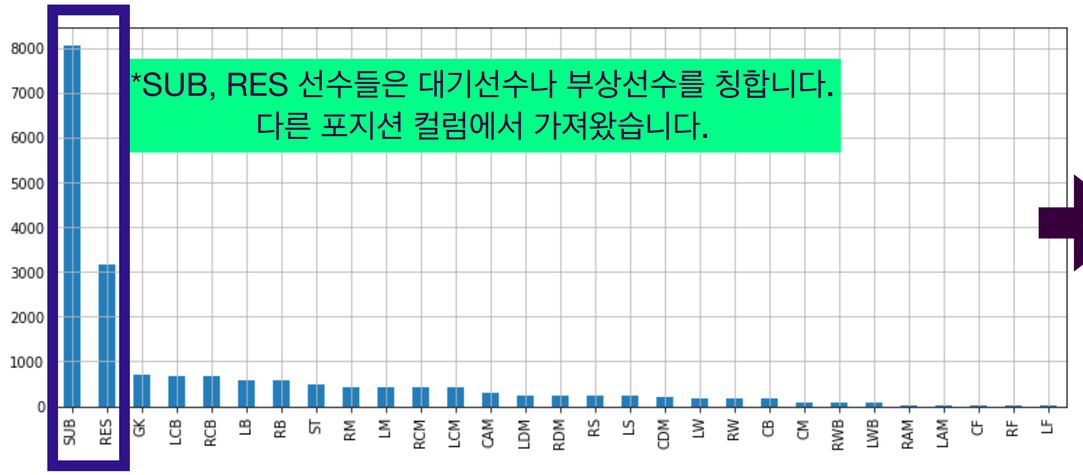
EDA

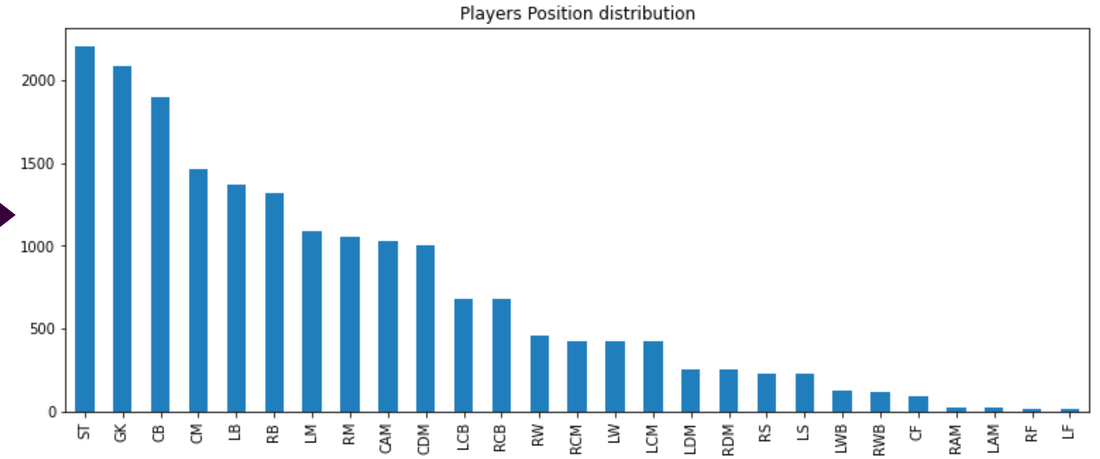
Classification

Nation position(국가전)
Missing : 17,817

Team position(리그전)
Missing : 255

Player position
Ex) 1 row => ST, LS…





EDA

Classification

position
포지션 갯수 : 27개

Broad_position
포지션 갯수 : 4개

Detail_position
포지션 갯수 : 11개

Broad_position	Detail_position		
공격수	중앙 공격수(CF)		
	윙 포워드(LWF)	세컨드 스트라이커(SS)	윙 포워드(RWF)
미드필더	공격형 미드필더(AM)		
	측면 미드필더(LM)	중앙 미드필더(CM)	측면 미드필더(RM)
	수비형 미드필더(DM)		
수비수	윙백(LWB)	센터백(CB)	윙백(RWB)
	풀백(LB)		풀백(RB)
	스위퍼(SW)		
골키퍼	골키퍼(GK)		

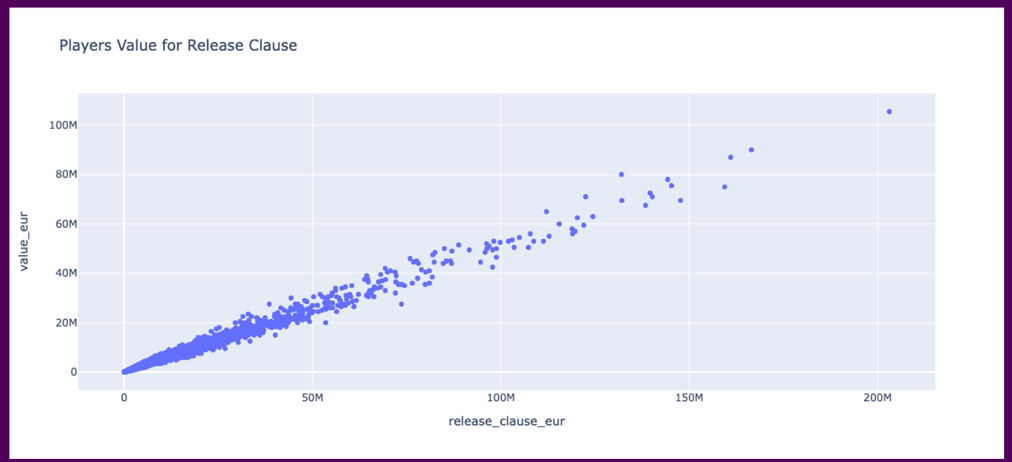
방향과 넓은 의미의 포지션을 고려하여 사진과 다르게 11개로 나누었습니다.

EDA

Regression

Value_euc

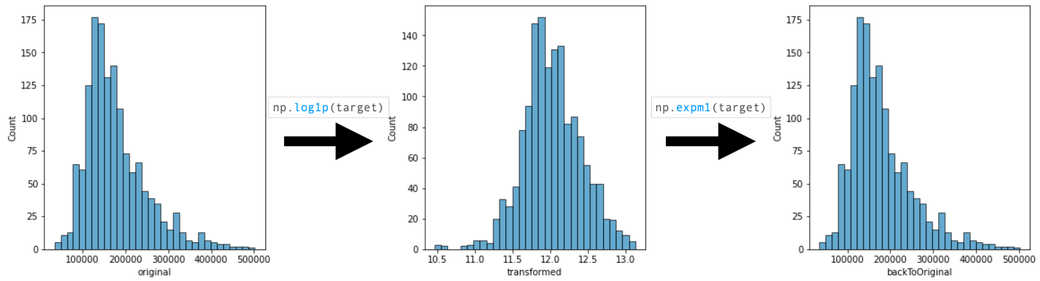
release_clause_eur
(릴리즈 조항)



선수들의 몸값이 증가할수록 릴리즈 조항 금액도 증가함을 알 수 있다.

Regression

일반적인 예시 그림입니다. 지금 분석과 관계 없습니다.



분포가 비대칭이면 그만큼 **모델로서 해석을 잘 못한다는 뜻**이기 때문에
로그 변환을 이용하여 이상적인 분포형태로 바꾸어 준다.

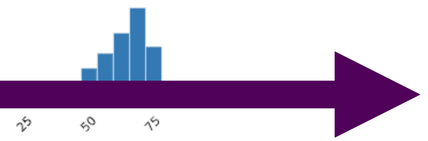
Feature Engineering

physic
Real number (ℝ≥0)

MISSING

Distinct count	63
Unique (%)	0.4%
Missing	2083
Missing (%)	11.0%
Infinite	0
Infinite (%)	0.0%

Mean	64.4589881976158
Minimum	28.0
Maximum	91.0
Zeros	0
Zeros (%)	0.0%
Memory size	148.1 KiB



Toggle details

SimpleImputer을 사용할 예정

gk_diving
Real number (ℝ≥0)

HIGH_CORRELATION
MISSING

Distinct count	46
Unique (%)	2.2%
Missing	16861
Missing (%)	89.0%
Infinite	0
Infinite (%)	0.0%

Mean	65.16418626980317
Minimum	45.0
Maximum	90.0
Zeros	0
Zeros (%)	0.0%
Memory size	148.1 KiB

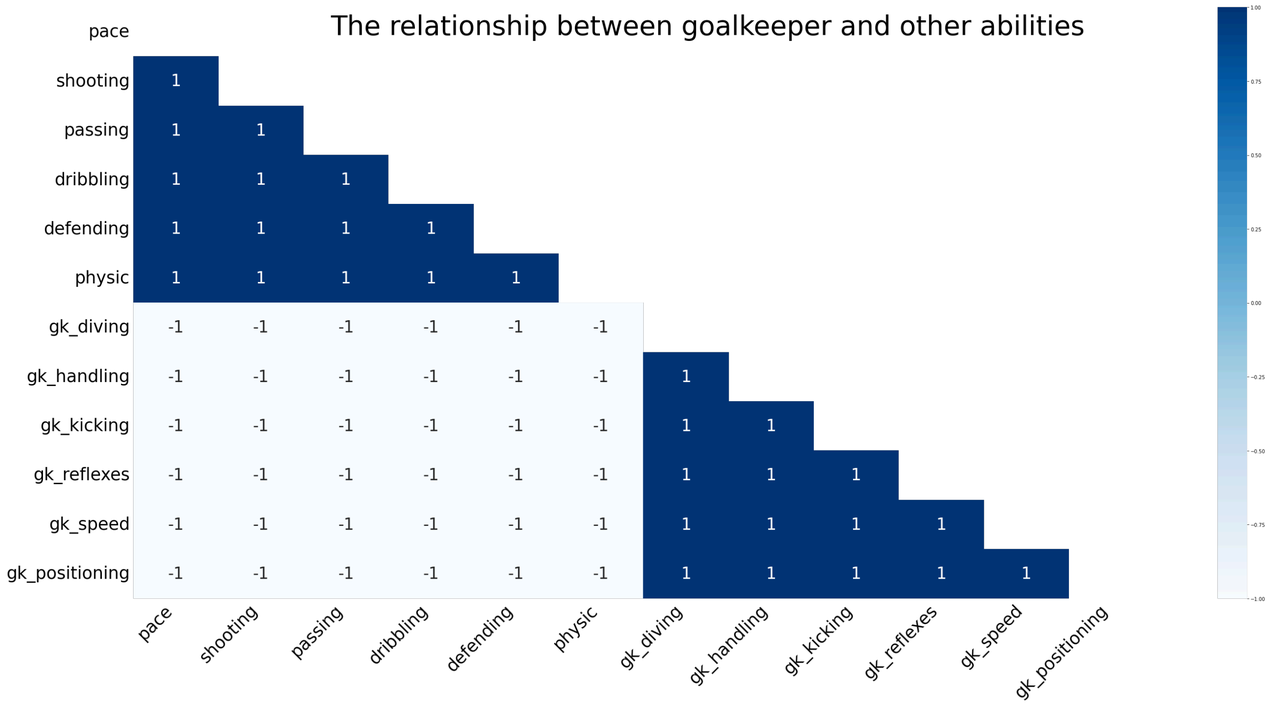
Toggle details

중복되는 골기퍼 능력의 컬럼이 있기 때문에 이 컬럼을 제거한다.
(중복되는 컬럼은 다른 포지션 선수들의 능력까지 입력되어 있다.)

Pandas_Profiling 을 이용하여 결측치 분석

몇몇 컬럼에서 골기퍼 능력만 해당하는 컬럼으로 결측치가 생기게 되었음.

Feature Engineering



	gk_diving	goalkeeping_diving
0	NaN	6
1	NaN	7
2	87.0	87
3	NaN	15
4	NaN	9
...	...	...
18939	NaN	10
18940	NaN	12
18941	NaN	12
18942	NaN	15
18943	NaN	8

골기퍼 능력과 다른 능력과의 관계성

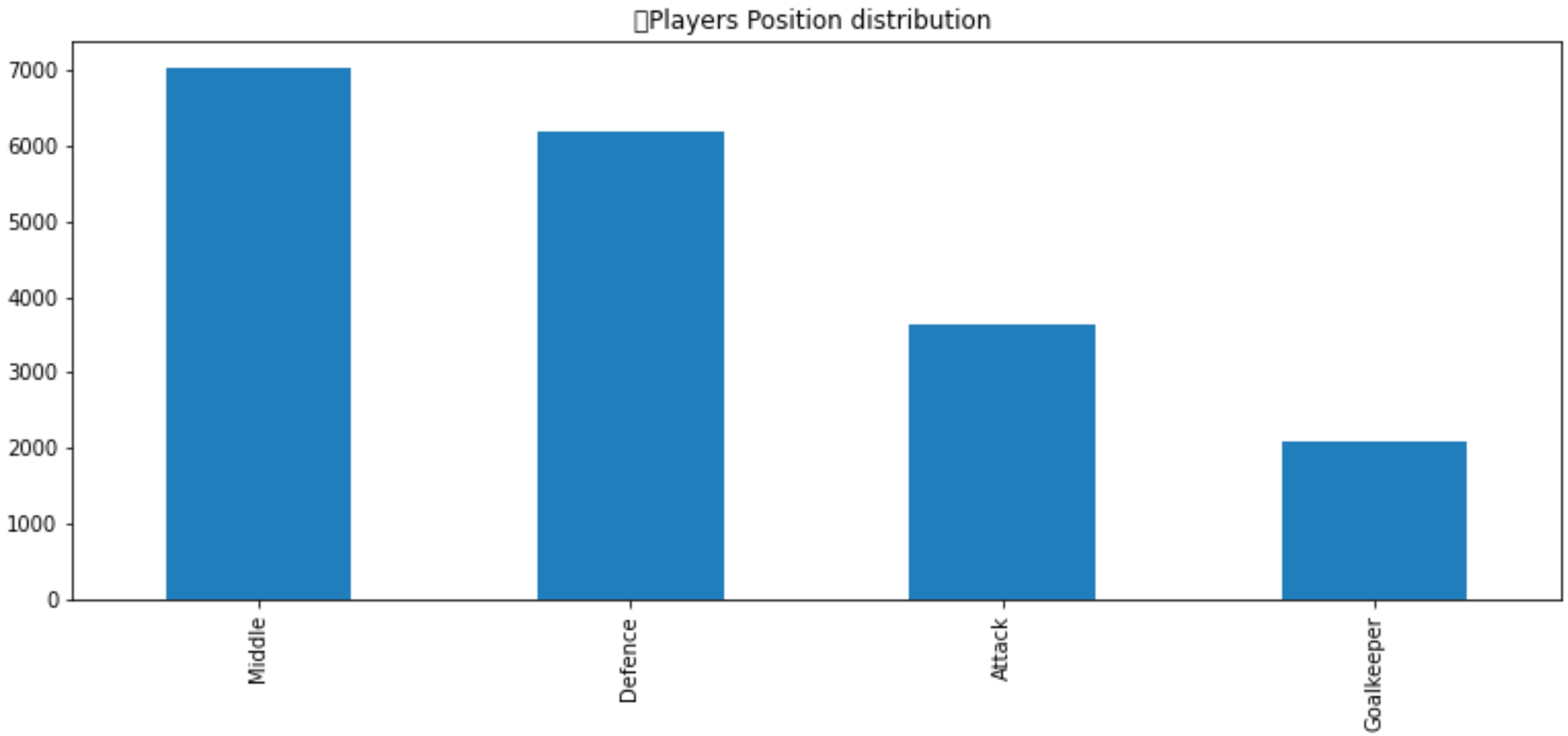
위 히트맵을 봐도 다른 선수들은 골기퍼 능력에 기입이 되어 있지 않은 상태였으며
오른 쪽 표 처럼 중복되는 열(컬럼)을 발견할 수 있었습니다.

Feature Engineering

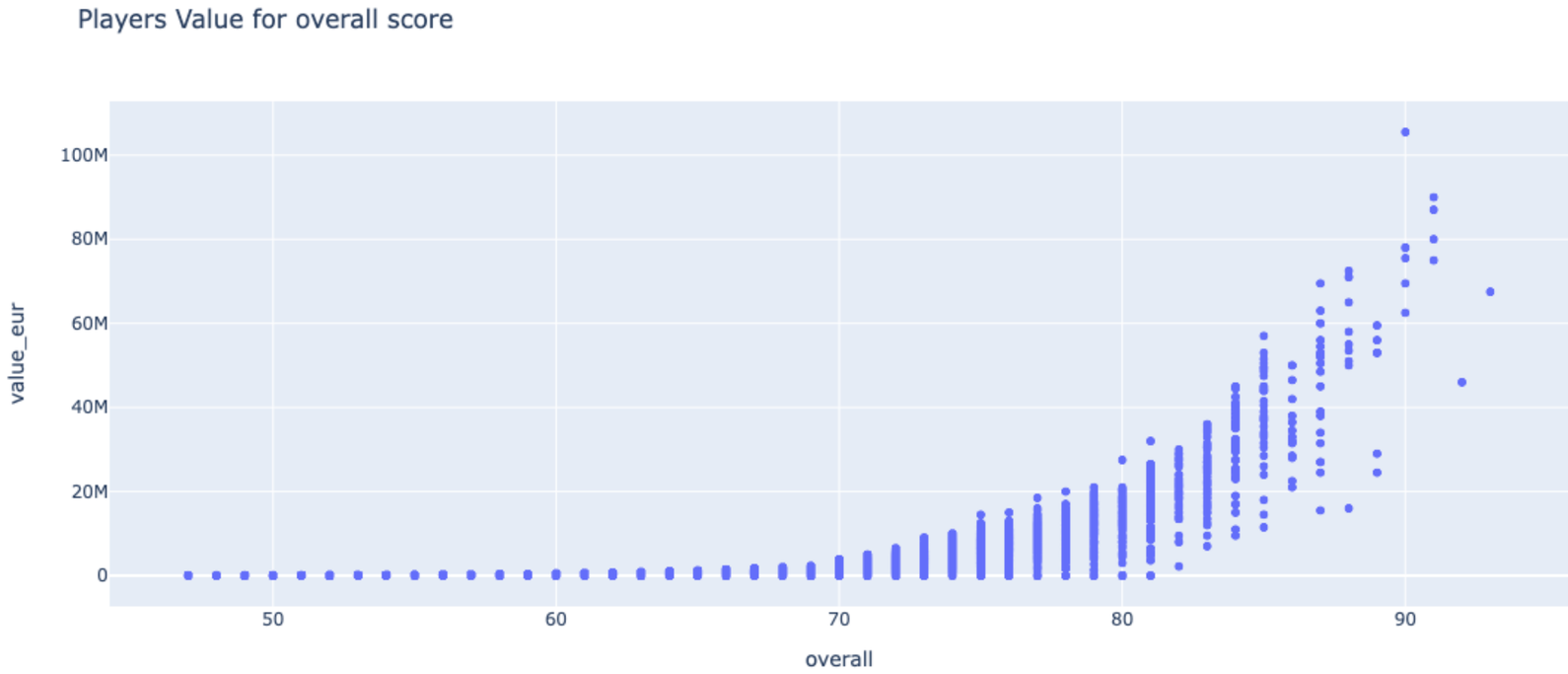
Player_tag	#Dribbler, #Distance Shooter, #FK Specialist, ...	다른 능력치로 충분히 짐작할 수 있는 특성이고 특수적인 선수들에게 만 가지고 있는 tag이기 때문에 제거를 한다.
work_rate	Medium/Low Medium/High Medium/Medium	작업속도를 의미하며 경기중에 선수들의 행동빈도를 의미한다. **/를 기준으로 두 개의 컬럼을 나눈다. (Attacking / Defense)** 순서형인 Category이므로 실수형으로 전환한다.
Body_Type	PLAYER_BODY_TYPE_259 PLAYER_BODY_TYPE_304 lean … Messi	**결측치와 범주들의 갯수를 확인하고 Lean / Normal / Stocky 타입으로 나눈다.** Player_body_type은 Normal에서 세분화된 것으로 Normal로 통일한다. 선수명으로 되어있는 값들은 Unique한 값이고 9개 밖에 없으므로 Unique로 통일했다. 총 4개의 범주

이외 다른 열들도 Groupby나 Split등의 기능을 활용하여 Feature Engineering을 진행하였습니다.
국가전 팀이나 리그 팀 같은 컬럼도 이번 모델을 만드는데 필요가 없으며 리그에는 있지만 국가팀에는 없는 선수들이 있기 때문에 제거하였습니다.

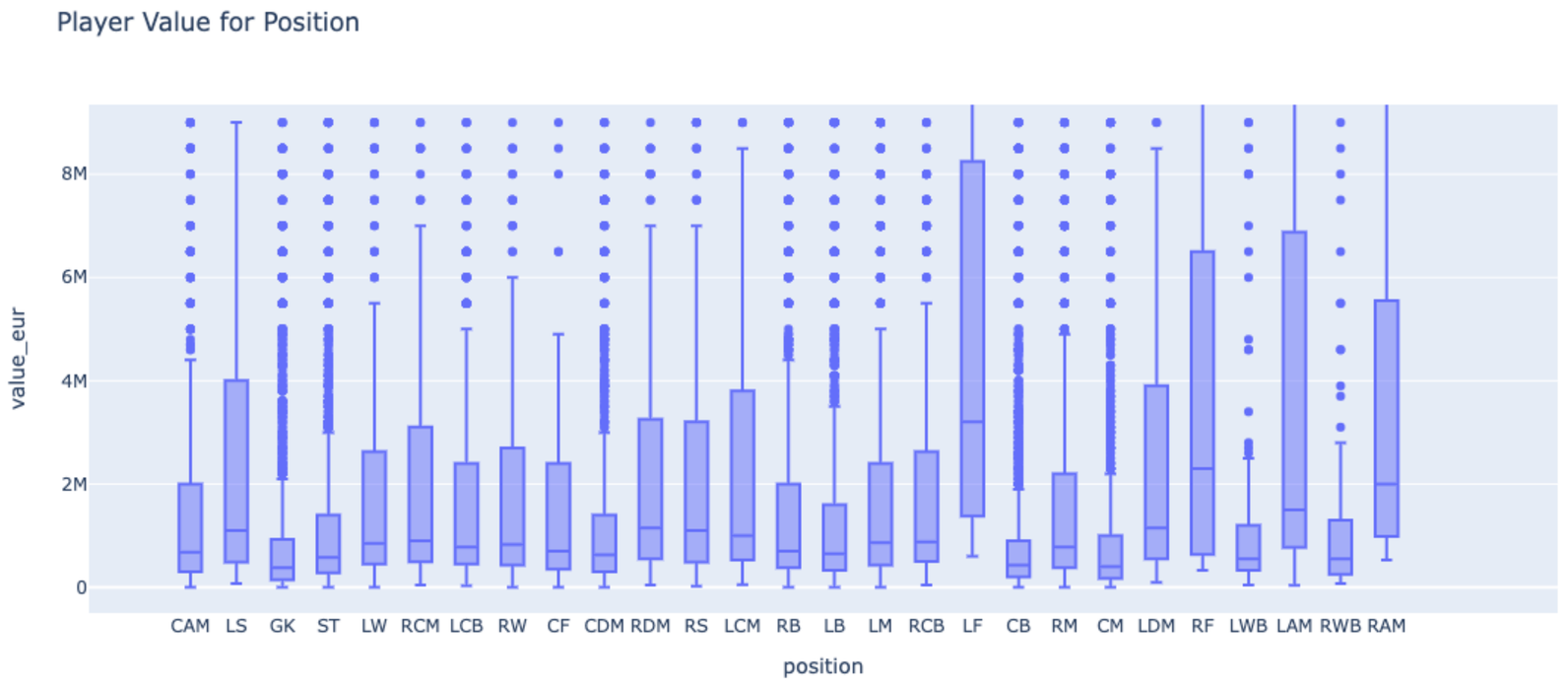
Visualization



Visualization



Visualization



Classification Modeling 4 Positions

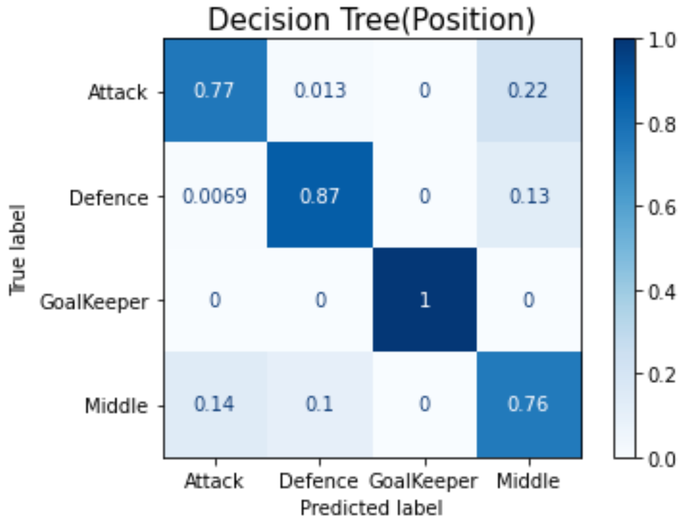
OneHotEncoder
0,1로 인코딩

SimpleImputer
결측치 보완

DecisionTree
기준모델(Baseline)

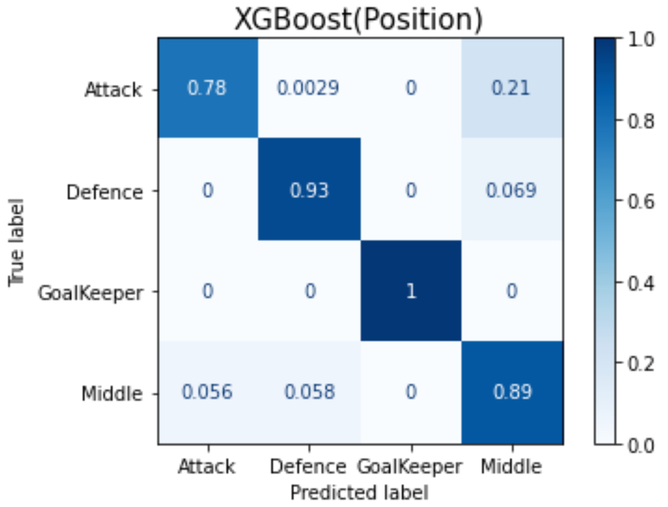
DecisionTree

	precision	recall	f1-score	support
Attack	0.73	0.77	0.75	684
Defence	0.87	0.87	0.87	1160
Goalkeeper	1.00	1.00	1.00	391
Middle	0.77	0.76	0.76	1317
accuracy			0.82	3552
macro avg	0.84	0.85	0.84	3552
weighted avg	0.82	0.82	0.82	3552



XGBoost

	precision	recall	f1-score	support
Attack	0.88	0.78	0.83	684
Defence	0.93	0.93	0.93	1160
Goalkeeper	1.00	1.00	1.00	391
Middle	0.84	0.89	0.86	1317
accuracy			0.89	3552
macro avg	0.91	0.90	0.90	3552
weighted avg	0.89	0.89	0.89	3552



RandomForest, LogisticRegression등 모델 중에서 가장 최고의 성능을 발휘한 XGBoost를 사용할 것 입니다.

Classification Modeling 11 Positions

DecisionTree

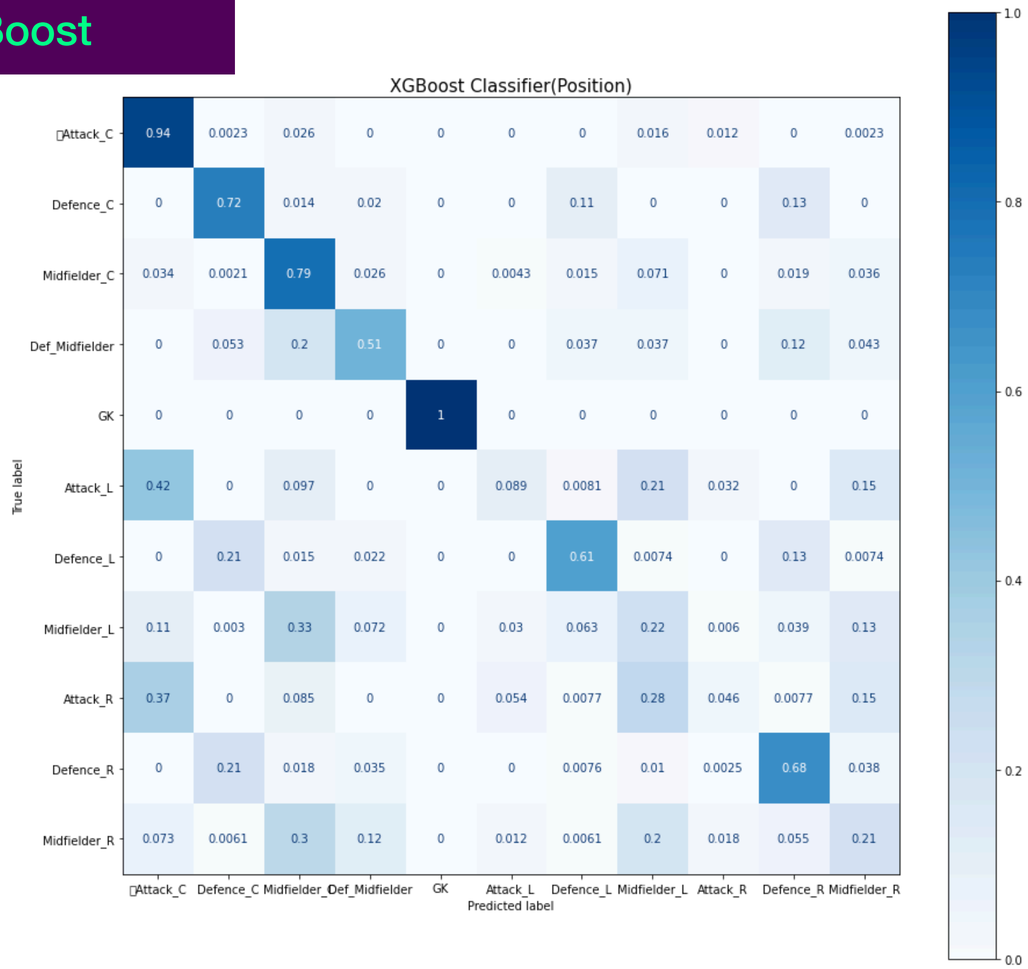
	precision	recall	f1-score	support
central_attack	0.69	0.66	0.67	430
central_defence	0.55	0.53	0.54	356
central_middle	0.47	0.45	0.46	466
defence_middle	0.33	0.37	0.35	188
goalkeeper	1.00	1.00	1.00	391
left_attack	0.12	0.14	0.13	124
left_defence	0.55	0.55	0.55	408
left_middle	0.19	0.18	0.18	334
right_attack	0.10	0.12	0.11	130
right_defence	0.55	0.54	0.55	397
right_middle	0.20	0.20	0.20	328
accuracy			0.49	3552
macro avg	0.43	0.43	0.43	3552
weighted avg	0.50	0.49	0.49	3552

RandomForest, LogisticRegression등 모델 중에서 가장 최고의 성능을 발휘한 XGBoost를 사용할 것 입니다.

Classification Modeling 11 Positions

XGBoost

	precision	recall	f1-score	support
central_attack	0.70	0.94	0.80	430
central_defence	0.58	0.72	0.64	356
central_middle	0.55	0.79	0.65	466
defence_middle	0.48	0.51	0.49	188
goalkeeper	1.00	1.00	1.00	391
left_attack	0.32	0.09	0.14	124
left_defence	0.76	0.61	0.67	408
left_middle	0.28	0.22	0.24	334
right_attack	0.25	0.05	0.08	130
right_defence	0.62	0.68	0.65	397
right_middle	0.35	0.21	0.26	328
accuracy			0.62	3552
macro avg	0.54	0.53	0.51	3552
weighted avg	0.59	0.62	0.59	3552



RandomForest, LogisticRegression등 모델 중에서 가장 최고의 성능을 발휘한 XGBoost를 사용할 것 입니다.

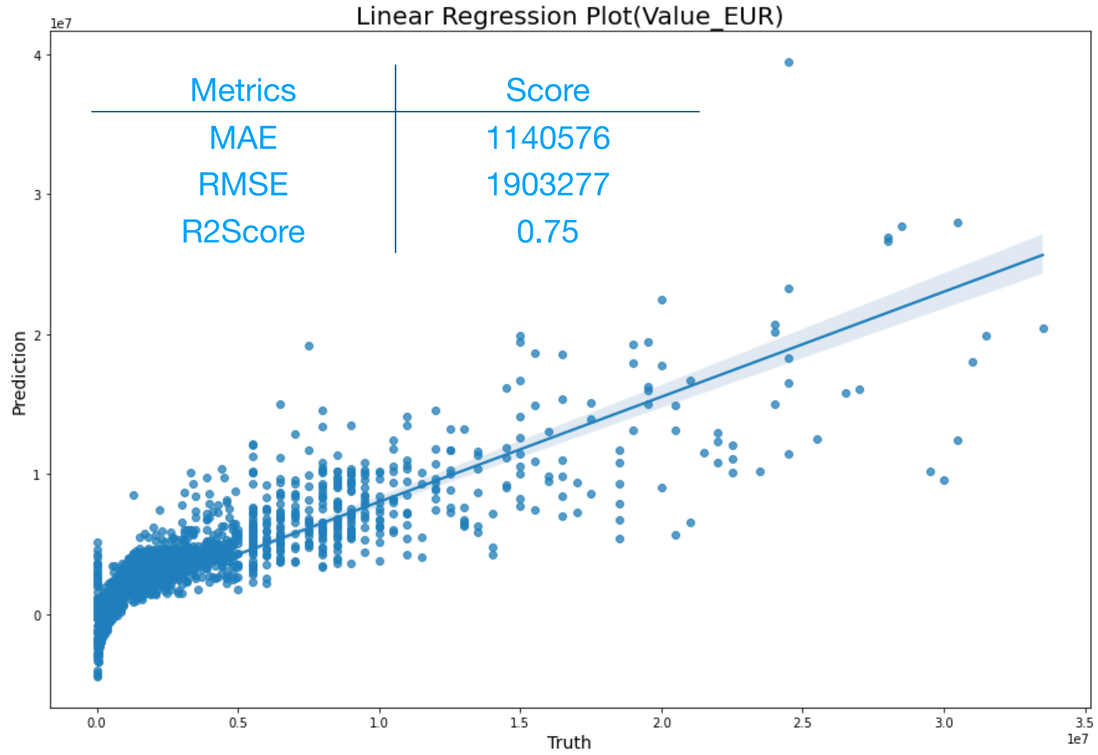
Regression for Players Value

OneHotEncoder
0,1로 인코딩

SimpleImputer
결측치 보완

MinmaxScaler
0과 1사이로 정규화

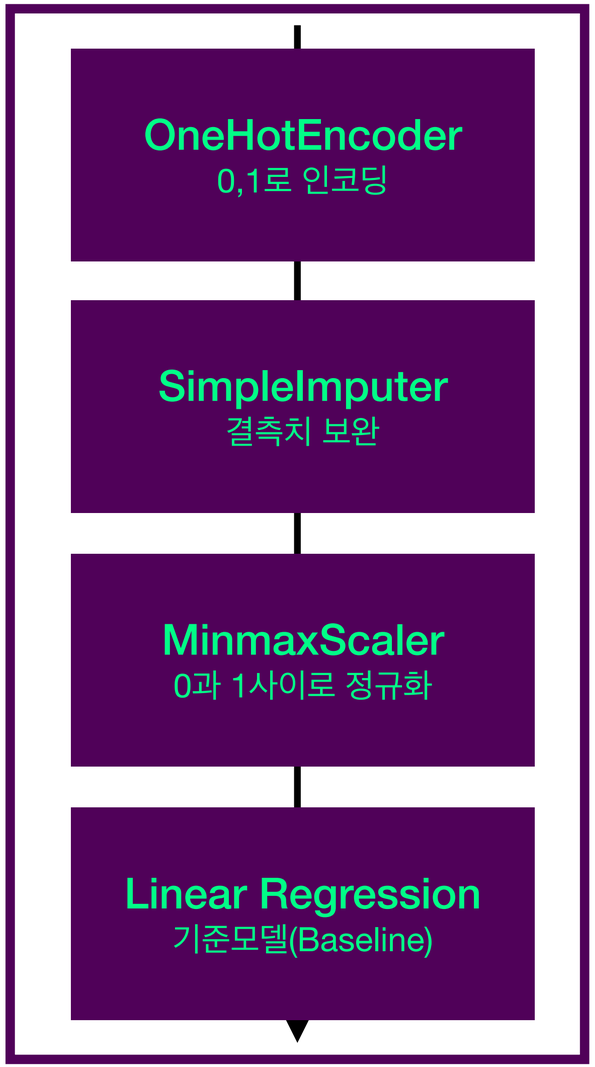
Linear Regression
기준모델(Baseline)



Metrics	Score
MAE	1140576
RMSE	1903277
R2Score	0.75

RandomForest, Ridge등 모델 중에서 가장 최고의 성능을 발휘한 XGBoost를 사용할 것 입니다.

Regression for Players Value



Linear Regression

Metrics	Score
MAE	1140576
RMSE	1903277
R2Score	0.75

RandomForest

Metrics	Score
MAE	328832
RMSE	1379733
R2Score	0.94

XGBoost

Metrics	Score
MAE	427093
RMSE	1251733
R2Score	0.95

RandomForest, Ridge등 모델 중에서 가장 최고의 성능을 발휘한 XGBoost를 사용할 것 입니다.

SelectKBest & Cross Validation

Classifier

body_type_unique	1.821248
...	...
skill_long_passing	1154.116164
skill_moves	1168.590638
skill_curve	1264.985323
power_stamina	1346.867540
defending	1448.544773
attacking_crossing	1651.284685
mentality_penalties	1756.804420
power_long_shots	1793.971681

Regression

body_type_Stocky	0.935981
...	...
RCB_value	526.577499
mentality_penalties	559.658817
power_stamina	602.184292
attacking_finishing	652.657127
physic	693.818333
LB_grow	728.413688
RB_grow	728.413688
skill_fk_accuracy	750.318943

Select K Best

가장 효과적인 특성 K 개를 선택하는 방법으로, 수작업이 필요없는 유용한 기능이다.

SelectKBest & Cross Validation

XGBoost_reg(k=85)

Metrics	Score
MAE	330578
RMSE	1205942
R2Score	0.95

XGBoost_reg

Metrics	Score
MAE	427093
RMSE	1251733
R2Score	0.95

XGBoost_clf(k=85)

	precision	recall	f1-score	support
central_attack	0.71	0.94	0.81	430
central_defence	0.57	0.74	0.64	356
central_middle	0.53	0.72	0.61	466
defence_middle	0.43	0.54	0.48	188
goalkeeper	1.00	1.00	1.00	391
left_attack	0.33	0.07	0.12	124
left_defence	0.79	0.60	0.68	408
left_middle	0.27	0.21	0.24	334
right_attack	0.24	0.04	0.07	130
right_defence	0.64	0.69	0.66	397
right_middle	0.31	0.20	0.25	328
accuracy			0.61	3552
macro avg	0.53	0.52	0.50	3552
weighted avg	0.58	0.61	0.58	3552

XGBoost_clf

	precision	recall	f1-score	support
central_attack	0.70	0.94	0.80	430
central_defence	0.58	0.72	0.64	356
central_middle	0.55	0.79	0.65	466
defence_middle	0.48	0.51	0.49	188
goalkeeper	1.00	1.00	1.00	391
left_attack	0.32	0.09	0.14	124
left_defence	0.76	0.61	0.67	408
left_middle	0.28	0.22	0.24	334
right_attack	0.25	0.05	0.08	130
right_defence	0.62	0.68	0.65	397
right_middle	0.35	0.21	0.26	328
accuracy			0.62	3552
macro avg	0.54	0.53	0.51	3552
weighted avg	0.59	0.62	0.59	3552

k갯수가 85일때 미세하게 좋은 성능이 있었으나 k개수를 줄일수록 성능은 감소했습니다.
모든 특성을 가지고 있을때가 좋은 성능을 나타내는 것 같습니다.

SelectKBest & Cross Validation

```
Accuracy (20 folds): [0.61726079 0.60787992 0.64165103 0.65103189 0.62101313 0.62664165
 0.62288931 0.60412758 0.61538462 0.62664165 0.63789869 0.62851782
 0.61538462 0.63414634 0.63602251 0.61913696 0.61654135 0.64849624
 0.61090226 0.60902256]
```

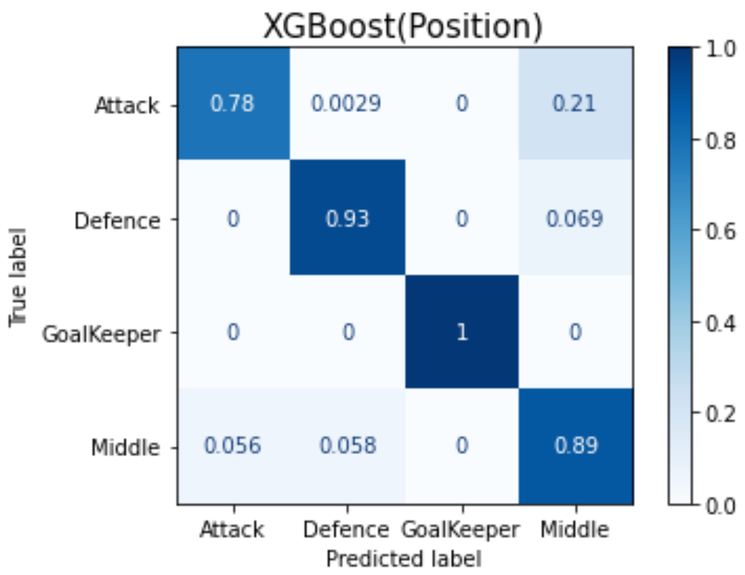
Accuracy

```
    (20 folds): [0.98968697 0.97864485 0.99140931 0.99542995 0.98194478 0.97934179
 0.99114425 0.96686929 0.97676287 0.98347295 0.94917021 0.99211371
 0.97285289 0.99137888 0.99052507 0.99611971 0.9867705  0.989142
 0.99264297 0.97045908]
```

R2score

Classification_4 Hyper parameter Tuning

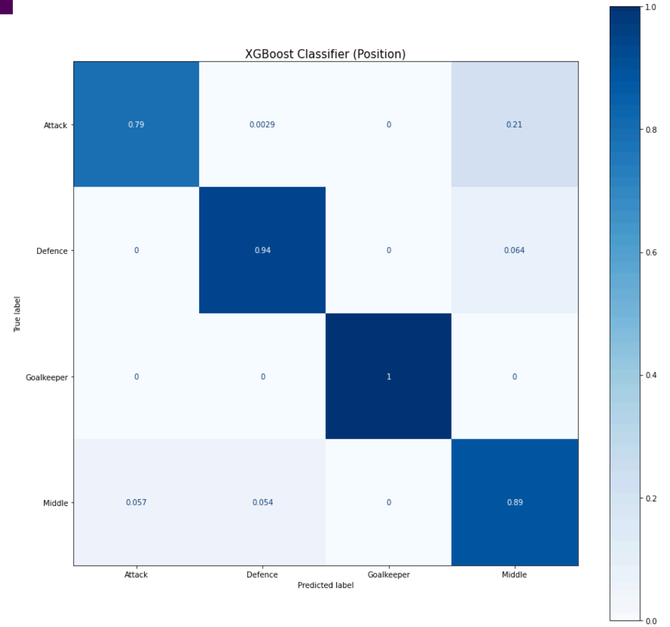
GridSearchCV + RandomizedSearchCV



	precision	recall	f1-score	support
Attack	0.88	0.78	0.83	684
Defence	0.93	0.93	0.93	1160
Goalkeeper	1.00	1.00	1.00	391
Middle	0.84	0.89	0.86	1317
accuracy			0.89	3552
macro avg	0.91	0.90	0.90	3552
weighted avg	0.89	0.89	0.89	3552

	precision	recall	f1-score	support
Attack	0.85	0.78	0.82	912
Defence	0.93	0.94	0.93	1547
Goalkeeper	1.00	1.00	1.00	521
Middle	0.84	0.87	0.86	1756
accuracy			0.89	4736
macro avg	0.91	0.90	0.90	4736
weighted avg	0.89	0.89	0.89	4736

훈련 f1 score (micro): 0.9240803303303303
검증 f1 score (micro): 0.8966779279279279
훈련 f1 score (macro): 0.9304575081587423
검증 f1 score (macro): 0.9076196750866233

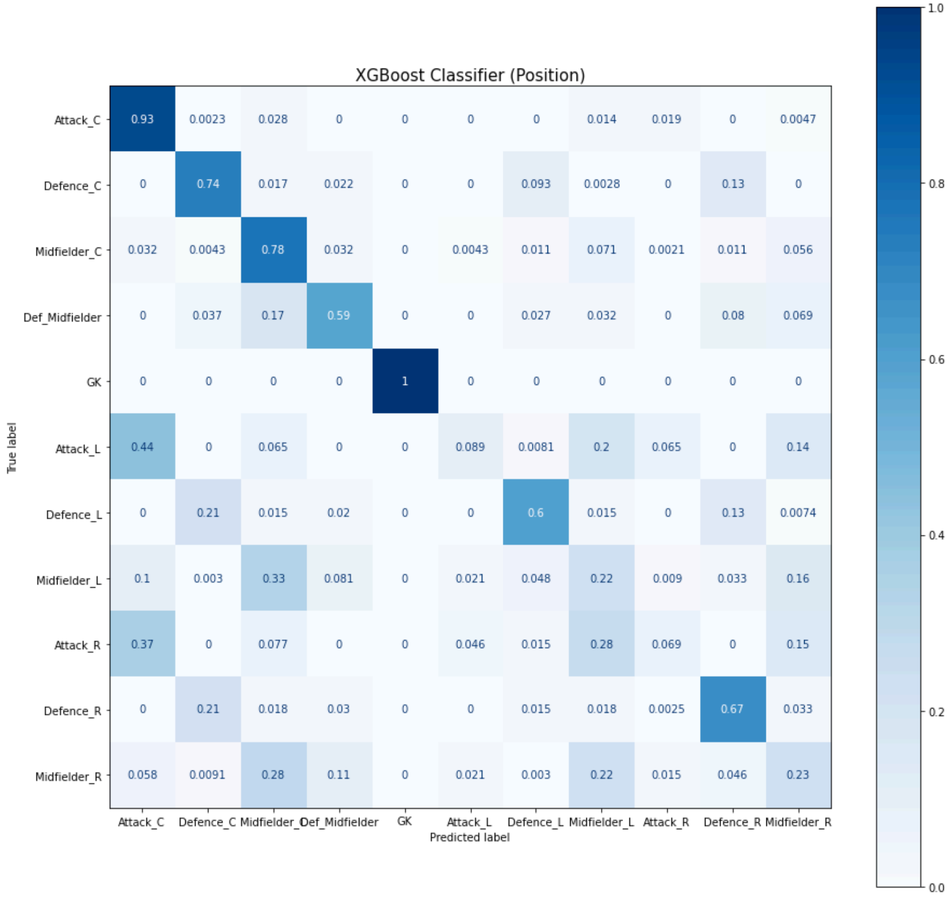


	precision	recall	f1-score	support
Attack	0.89	0.79	0.83	684
Defence	0.93	0.93	0.93	1160
Goalkeeper	1.00	1.00	1.00	391
Middle	0.84	0.89	0.86	1317
accuracy			0.90	3552
macro avg	0.91	0.90	0.91	3552
weighted avg	0.90	0.90	0.90	3552

훈련 f1 score (micro): 0.9049361861861862
검증 f1 score (micro): 0.8952702702702703
훈련 f1 score (macro): 0.9138712078596019
검증 f1 score (macro): 0.9071477578970077

Classification_11 Hyper parameter Tuning

GridSearchCV + RandomizedSearchCV

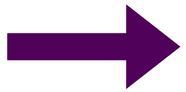


Classification_11 Hyper parameter Tuning

GridSearchCV + RandomizedSearchCV

```
훈련 f1 score (micro):  0.7317942942942943
검증 f1 score (micro):  0.6196509009009009
훈련 f1 score (macro):  0.6665750748768692
검증 f1 score (macro):  0.5213994516962651
```

	precision	recall	f1-score	support
central_attack	0.70	0.93	0.80	430
central_defence	0.59	0.73	0.65	356
central_middle	0.56	0.78	0.65	466
defence_middle	0.50	0.57	0.53	188
goalkeeper	1.00	1.00	1.00	391
left_attack	0.33	0.09	0.14	124
left_defence	0.77	0.60	0.67	408
left_middle	0.27	0.21	0.24	334
right_attack	0.27	0.07	0.11	130
right_defence	0.64	0.67	0.66	397
right_middle	0.34	0.23	0.28	328
accuracy			0.62	3552
macro avg	0.54	0.54	0.52	3552
weighted avg	0.59	0.62	0.59	3552

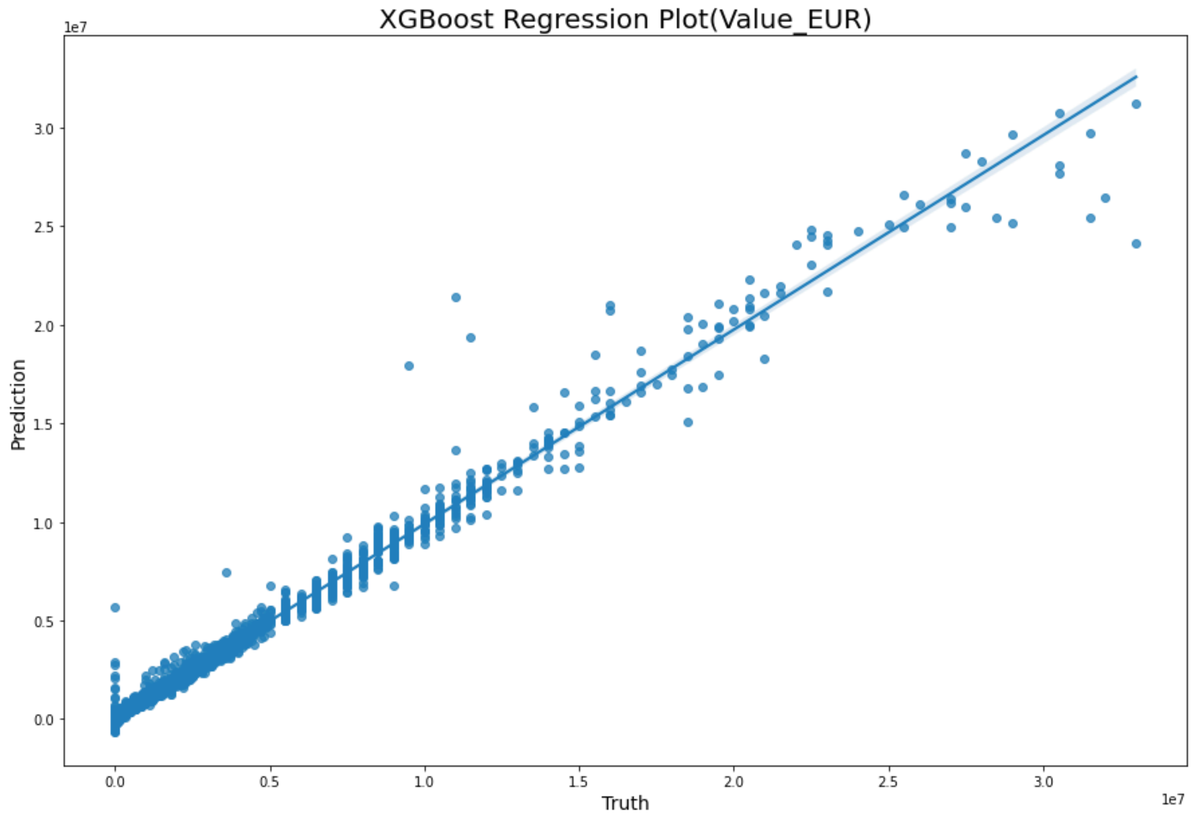


```
훈련 f1 score (micro):  0.7329204204204204
검증 f1 score (micro):  0.620777027027027
훈련 f1 score (macro):  0.6676560379390932
검증 f1 score (macro):  0.522790862828751
```

	precision	recall	f1-score	support
central_attack	0.71	0.92	0.80	573
central_defence	0.62	0.76	0.68	474
central_middle	0.55	0.80	0.65	622
defence_middle	0.50	0.51	0.51	250
goalkeeper	1.00	1.00	1.00	521
left_attack	0.24	0.05	0.09	166
left_defence	0.79	0.70	0.74	544
left_middle	0.28	0.23	0.25	446
right_attack	0.20	0.05	0.08	173
right_defence	0.67	0.69	0.68	529
right_middle	0.28	0.17	0.21	438
accuracy			0.63	4736
macro avg	0.53	0.53	0.52	4736
weighted avg	0.59	0.63	0.60	4736

Regression Hyper parameter Tuning

GridSearchCV + RandomizedSearchCV



Metrics	Score
RMSE	1251733
R2Score	0.95

Metrics	Score
RMSE	1205942
R2Score	0.95

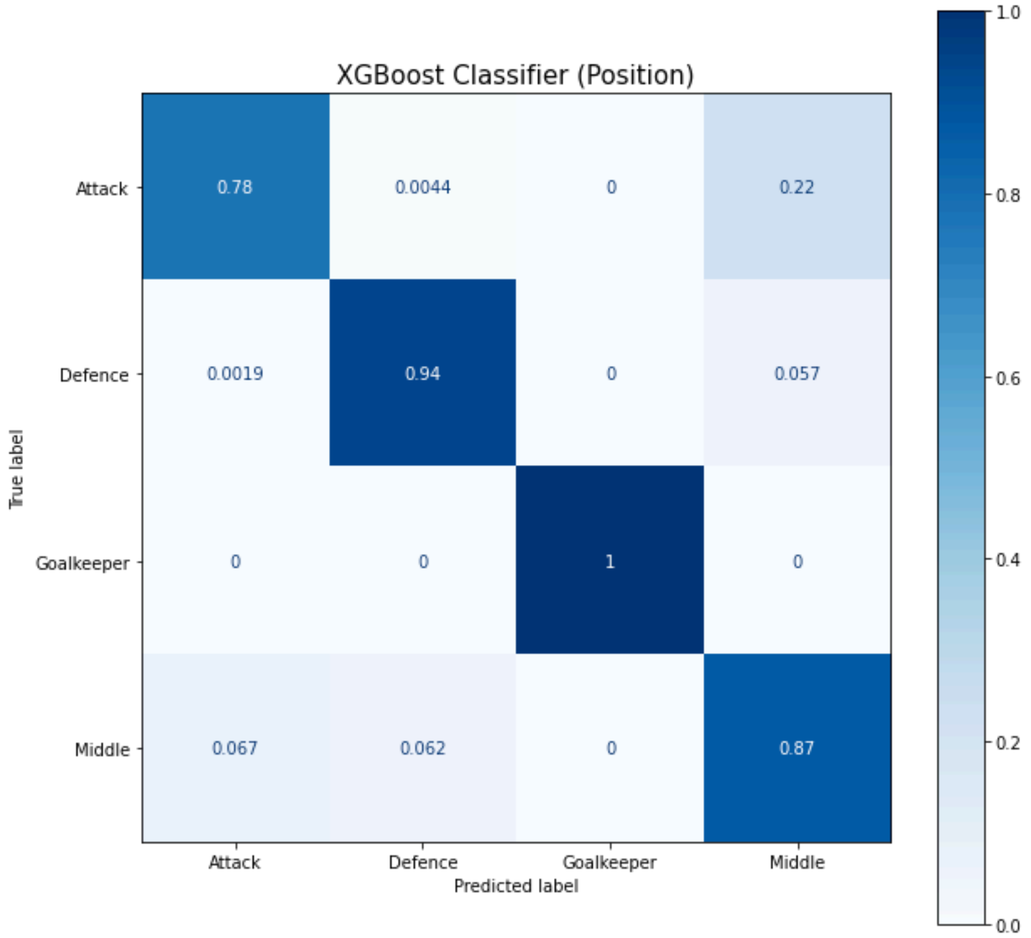
Metrics	Score
RMSE	435631
R2Score	0.985

Classification_4 Interpreting

	precision	recall	f1-score	support
Attack	0.86	0.78	0.82	912
Defence	0.93	0.94	0.93	1547
Goalkeeper	1.00	1.00	1.00	521
Middle	0.84	0.87	0.86	1756
accuracy			0.89	4736
macro avg	0.91	0.90	0.90	4736
weighted avg	0.89	0.89	0.89	4736

훈련 f1 score (micro): 0.9049361861861862
테스트 f1 score (micro): 0.8904138513513513
훈련 f1 score (macro): 0.9138712078596019
테스트 f1 score (macro): 0.9015873446919267

훈련 accuracy (micro): 0.9049361861861862
테스트 accuracy (micro): 0.8904138513513513
훈련 accuracy (macro): 0.9049361861861862
테스트 accuracy (macro): 0.8904138513513513

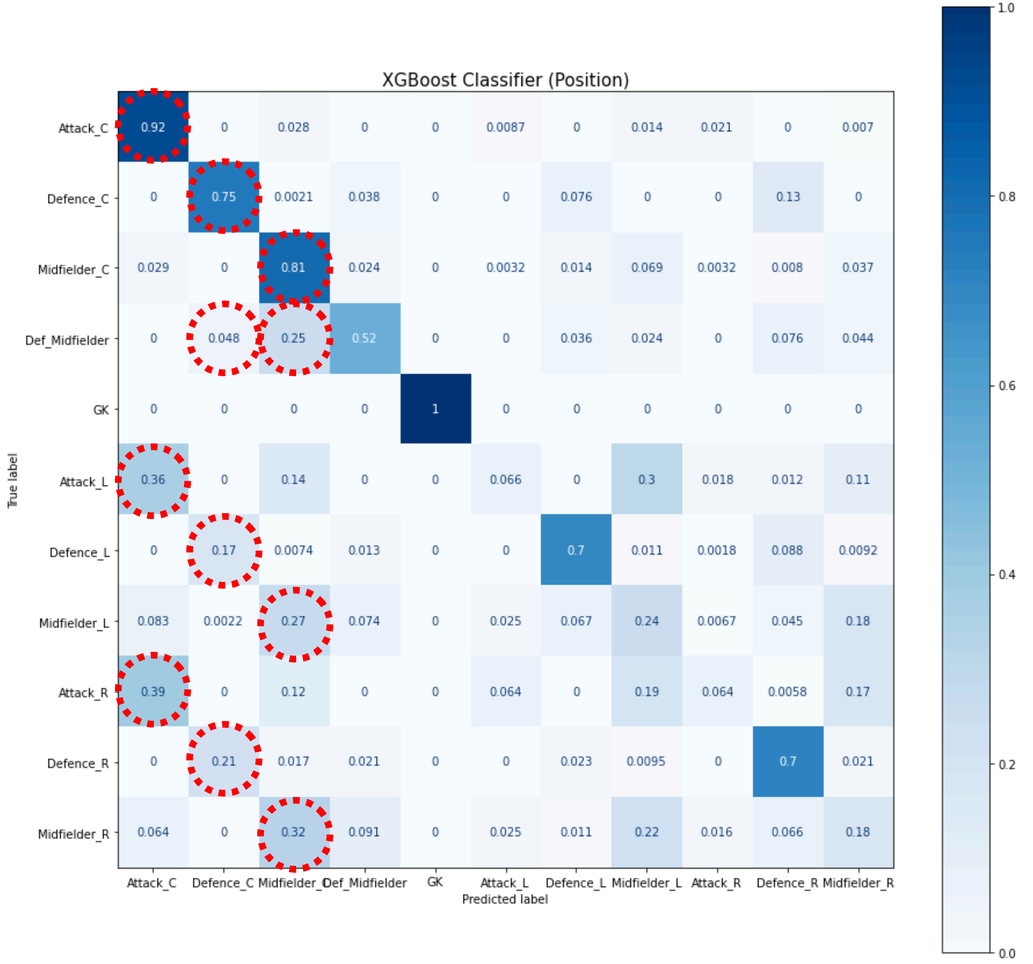


Classification_11 Interpreting

	precision	recall	f1-score	support
central_attack	0.72	0.92	0.81	573
central_defence	0.62	0.75	0.68	474
central_middle	0.56	0.81	0.66	622
defence_middle	0.51	0.52	0.52	250
goalkeeper	1.00	1.00	1.00	521
left_attack	0.22	0.07	0.10	166
left_defence	0.79	0.70	0.74	544
left_middle	0.31	0.24	0.27	446
right_attack	0.28	0.06	0.10	173
right_defence	0.67	0.70	0.68	529
right_middle	0.31	0.18	0.23	438
accuracy			0.63	4736
macro avg	0.54	0.54	0.53	4736
weighted avg	0.60	0.63	0.61	4736

훈련 f1 score (micro): 0.6831831831831832
테스트 f1 score (micro): 0.6347128378378378
훈련 f1 score (macro): 0.5952402574769328
테스트 f1 score (macro): 0.5276295998506032

훈련 accuracy (micro): 0.6831831831831832
테스트 accuracy (micro): 0.6347128378378378
훈련 accuracy (macro): 0.6831831831831832
테스트 accuracy (macro): 0.6347128378378378

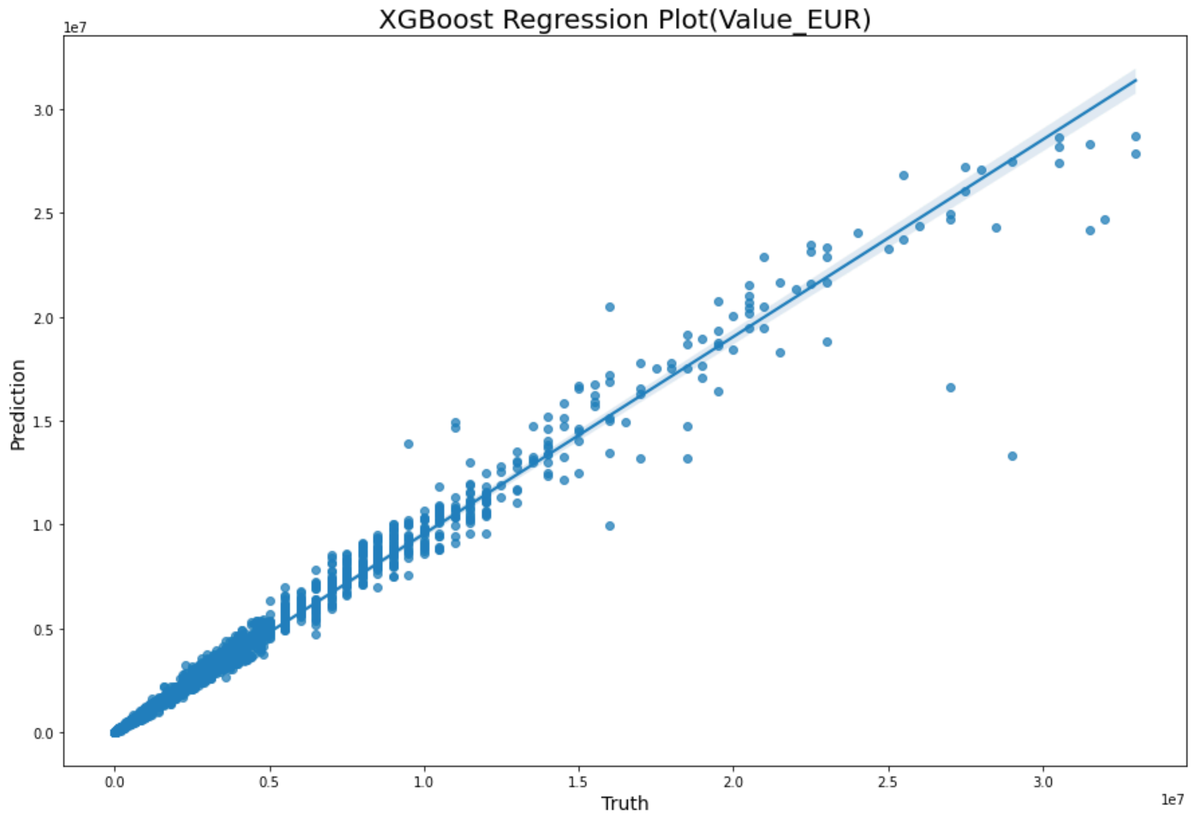


Regression Interpreting

Metrics	Score
MAE	330578
RMSE	1205942
R2Score	0.95

Metrics	Score
RMSE	486755
R2Score	0.98

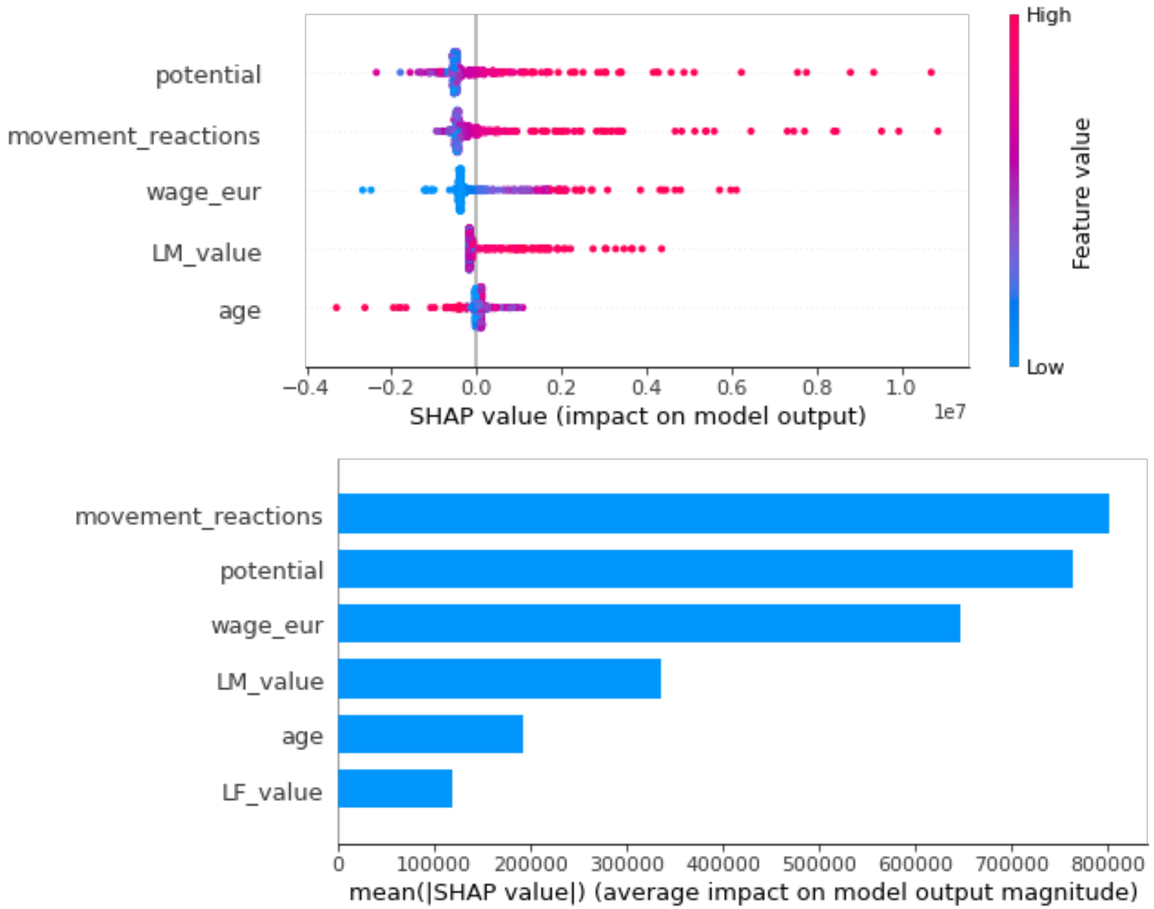
```
RMSE : 486755.9700681536
-------------
테스트 세트 점수 : 0.982199324886018
훈련 세트 점수 : 0.9898494987053915
```



Regression Interpreting

Weight	Feature
0.3338 ± 0.0105	potential
0.1274 ± 0.0134	movement_reactions
0.0634 ± 0.0096	wage_eur
0.0313 ± 0.0011	LM_value
0.0268 ± 0.0045	age
0.0100 ± 0.0002	LF_value

Permutation Importances



SHAP

감사합니다.