

이스포츠 데이터분석가 양성과정

국내 롤 프로게이머 평가

김혜린

배나경

오종용

임도현



1. 주제 선정 이유 및 목표

2. 분석과정

- 2-1. 데이터 수집 및 전처리
- 2-2. 데이터 모델링
- 2-3. 데이터 시각화

3. 결과

01

1. 주제 선정 이유 및 목표

1. 주제 선정 이유 및 목표

선정 이유

일반인의 시선에서 바라본 선수평가와 데이터가 말해주는 선수평가의 일치여부를 확인

목표

데이터를 활용한 LCK 출전 프로게이머 선수평가

사용 데이터 출처 : <https://oracleselixir.com/tools/downloads>



2. 분석

- 2-1. 데이터 수집 및 전처리
- 2-2. 데이터 모델링
- 2-3. 데이터 시각화

2-1. 데이터 수집 및 전처리

초기 데이터

5868 rows * 121 cols

	gameid	datacompleteness	url	league	year	split	playoffs	date	game	patch	participantid	side	position	playername
0	ESPORTSTMNT01/1690520	complete	http://matchhistory.na.leagueoflegends.com/en/...	LCK	2021	Spring	0	2021-01-13 07:00:34	1	11.01	1	Blue	top	Rascal
1	ESPORTSTMNT01/1690520	complete	http://matchhistory.na.leagueoflegends.com/en/...	LCK	2021	Spring	0	2021-01-13 07:00:34	1	11.01	2	Blue	jng	Clid
2	ESPORTSTMNT01/1690520	complete	http://matchhistory.na.leagueoflegends.com/en/...	LCK	2021	Spring	0	2021-01-13 07:00:34	1	11.01	3	Blue	mid	Bdd
3	ESPORTSTMNT01/1690520	complete	http://matchhistory.na.leagueoflegends.com/en/...	LCK	2021	Spring	0	2021-01-13 07:00:34	1	11.01	4	Blue	bot	Ruler
4	ESPORTSTMNT01/1690520	complete	http://matchhistory.na.leagueoflegends.com/en/...	LCK	2021	Spring	0	2021-01-13 07:00:34	1	11.01	5	Blue	sup	Life

5 rows x 121 columns

2-1. 데이터 수집 및 전처리

ex) top_df

	playername	golddiffat15	xpdiffat15	csdiffat15	killsdiffat15	assistsdiffat15
0	Rascal	1548.0	1464.0	-8.0	3.0	1.0
1	Doran	-1548.0	-1464.0	8.0	-3.0	-1.0
2	Doran	-725.0	-117.0	-4.0	-1.0	-2.0
3	Rascal	725.0	117.0	4.0	1.0	2.0
4	Canna	-152.0	420.0	-14.0	0.0	1.0
...
973	Morgan	-1317.0	-757.0	5.0	-3.0	1.0
974	Canna	8.0	-363.0	7.0	-1.0	0.0
975	Morgan	-8.0	363.0	-7.0	1.0	0.0
976	Morgan	-98.0	-8.0	-22.0	1.0	1.0
977	Canna	98.0	8.0	22.0	-1.0	-1.0

978 rows × 19 columns

전체 지표로 선수평가를 하는 것보다는 포지션 별로 선수 평가를 하는 것이 공정한 평가이며 또한 평가의 정확도를 상승시킬 것이라 기대되어 주어진 데이터프레임을 포지션 별로 분리
(총 5개의 데이터 프레임으로 분리)

또한 올해에 본인 주 포지션이 아닌 다른 포지션으로 플레이한 선수들이 있어 그 데이터는 제외함

2-1. 데이터 수집 및 전처리

ex) top_df

	playername	golddiffat15	xpdiffat15	csdiffat15	killsdiffat15	assistsdiffat15
0	10	0.722478	0.822894	0.437500	0.8	0.583333
1	3	0.277522	0.177106	0.562500	0.2	0.416667
2	3	0.395803	0.474195	0.468750	0.4	0.333333
3	10	0.604197	0.525805	0.531250	0.6	0.666667
4	1	0.478155	0.592633	0.390625	0.5	0.583333
...
973	9	0.310721	0.333039	0.539062	0.2	0.583333
974	1	0.501150	0.419938	0.554688	0.4	0.500000
975	9	0.498850	0.580062	0.445312	0.6	0.500000
976	9	0.485915	0.498236	0.328125	0.6	0.583333
977	1	0.514085	0.501764	0.671875	0.4	0.416667

머신러닝 알고리즘을 적용시키기 위하여 Encoding 작업을 진행

One-hot encoding 방식을 채택하려 하였으나 'playername' 컬럼의 unique값이 상당히 많아 컬럼이 과생성될것이 우려되어 Label encoding 방식을 채택

또한 머신러닝의 정확도 상승을 위해 연속형(숫자형) 데이터에 Scaling 작업을 진행하고 프로리그이므로 이상치 또한 의미있는 데이터라고 생각하였기 때문에 MinMaxScaling 방식을 채택

2-2. 데이터 모델링

LightGBM Classifier 교차검증 평균 정확도 : 0.8763

선수의 평가에 대한 feature가 없으므로, 그와 높은 상관 계수를 가질것이라 예상되는 승패('result')컬럼을 종속변수로 채택하여 머신러닝 분류 알고리즘 적용

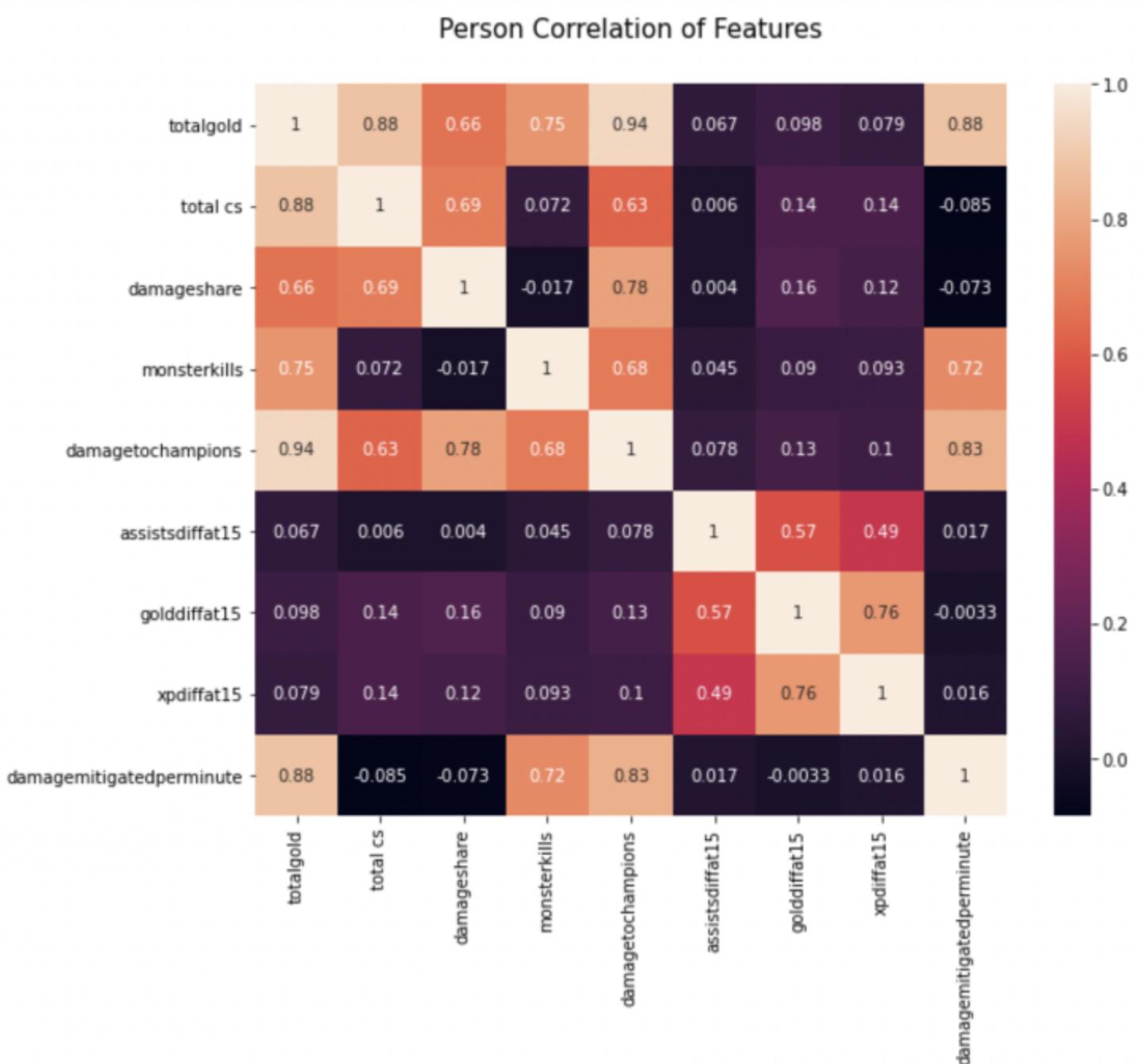
CatBoost Classifier 교차검증 평균 정확도 : 0.7414

5-fold cross validation 결과 LightGBM Classifier 모델의 평균 정확도가 가장 높게 측정되었으므로 LightGBM Classifier 모델에서 종속변수(승패)를 결정하는데 영향을 준 feature들을 추출하여 선수평가 지표로 사용하면 선수 평가의 신뢰도가 높을 것이라 예상

RandomForest Classifier 교차검증 평균 정확도 : 0.8517

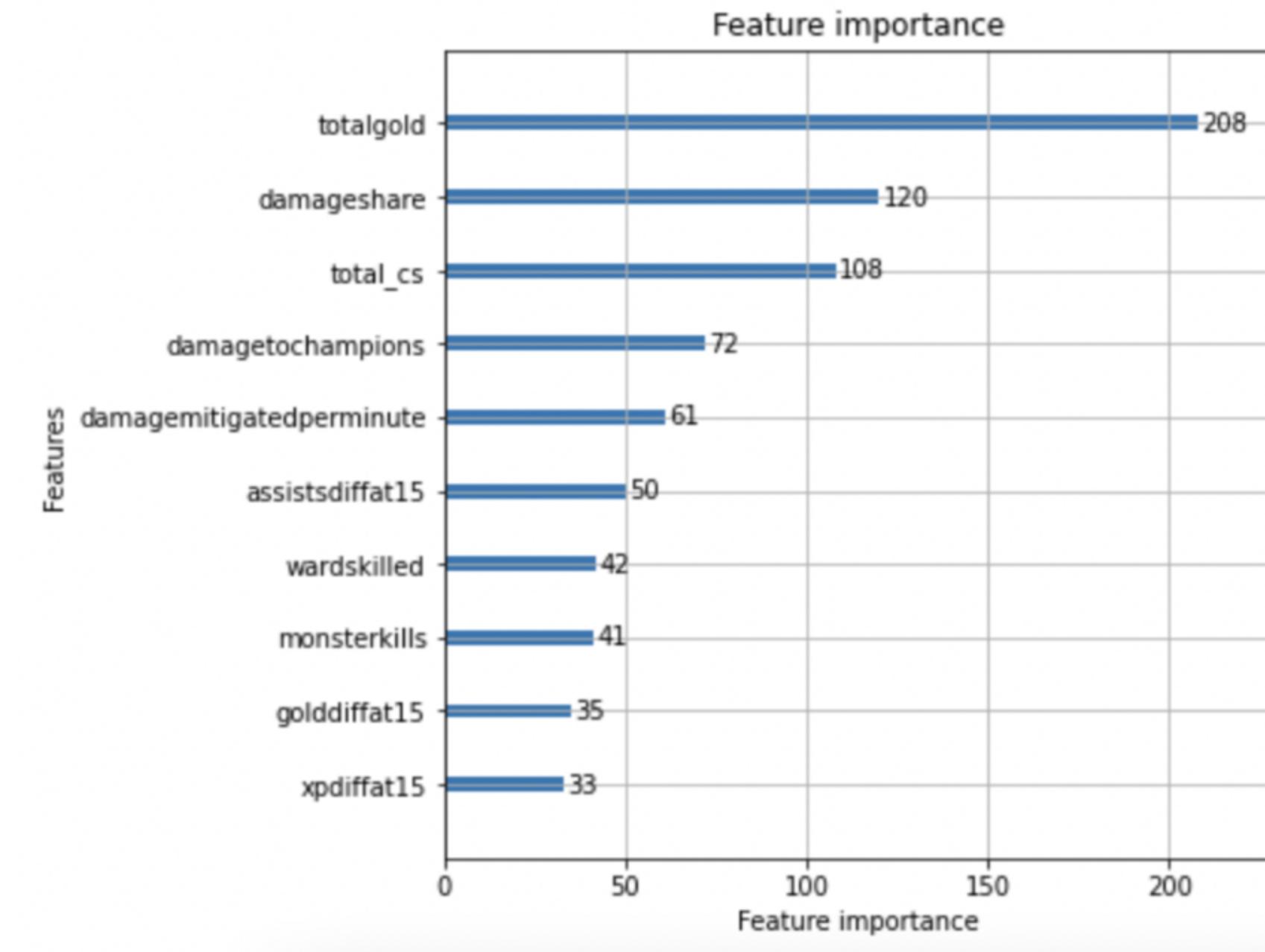
종속변수를 결정하는데 영향을 준 feature들을 조사하기 위해, Feature importance / SHAP values feature importance / Pearson Correlation of Features 등을 시각화

2-2. 데이터 모델링



<그림 1>

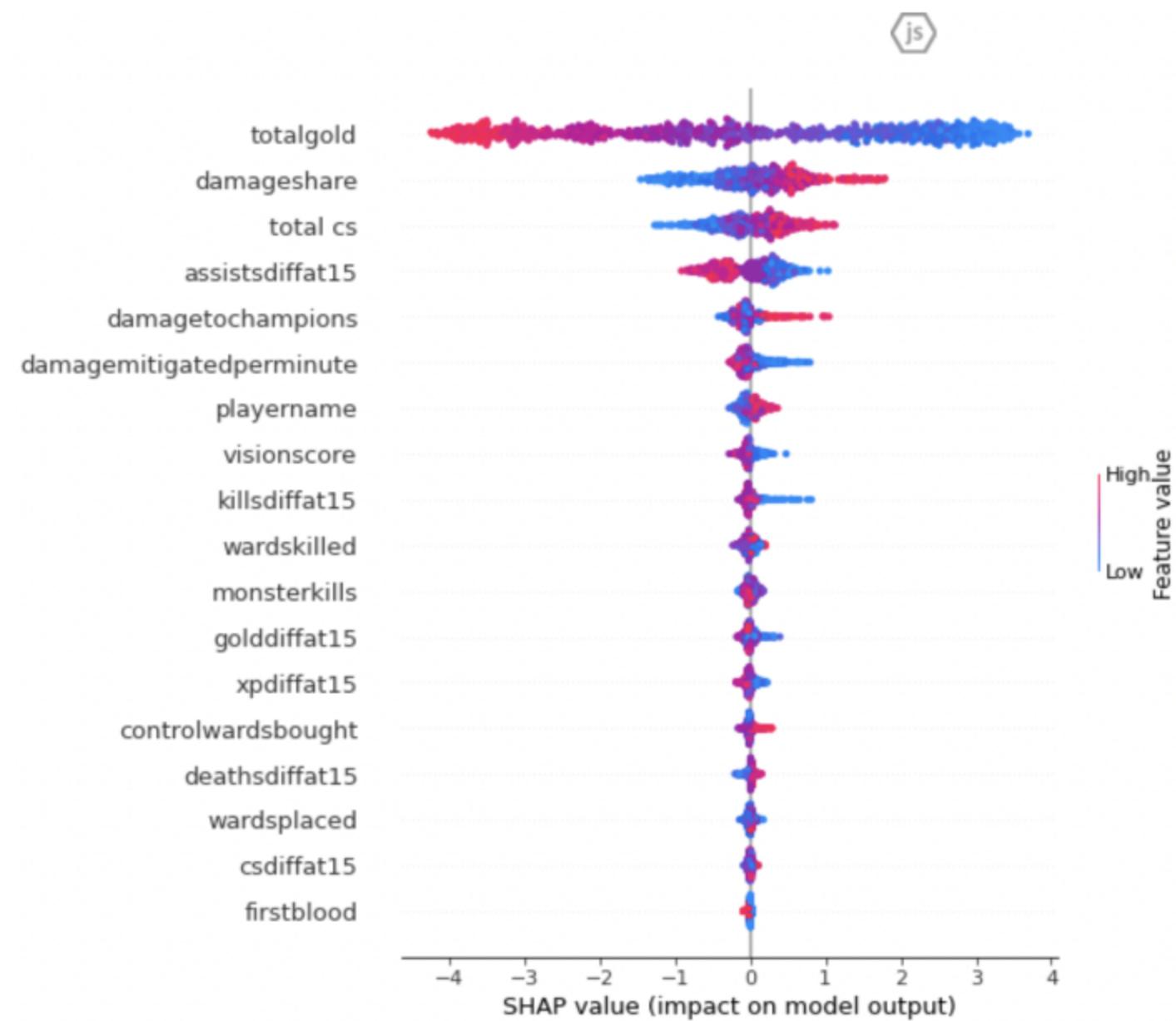
<그림1>에서 다중공선성문제를 가지는 feature들을 선수평가지표에서 제외



<그림 2>

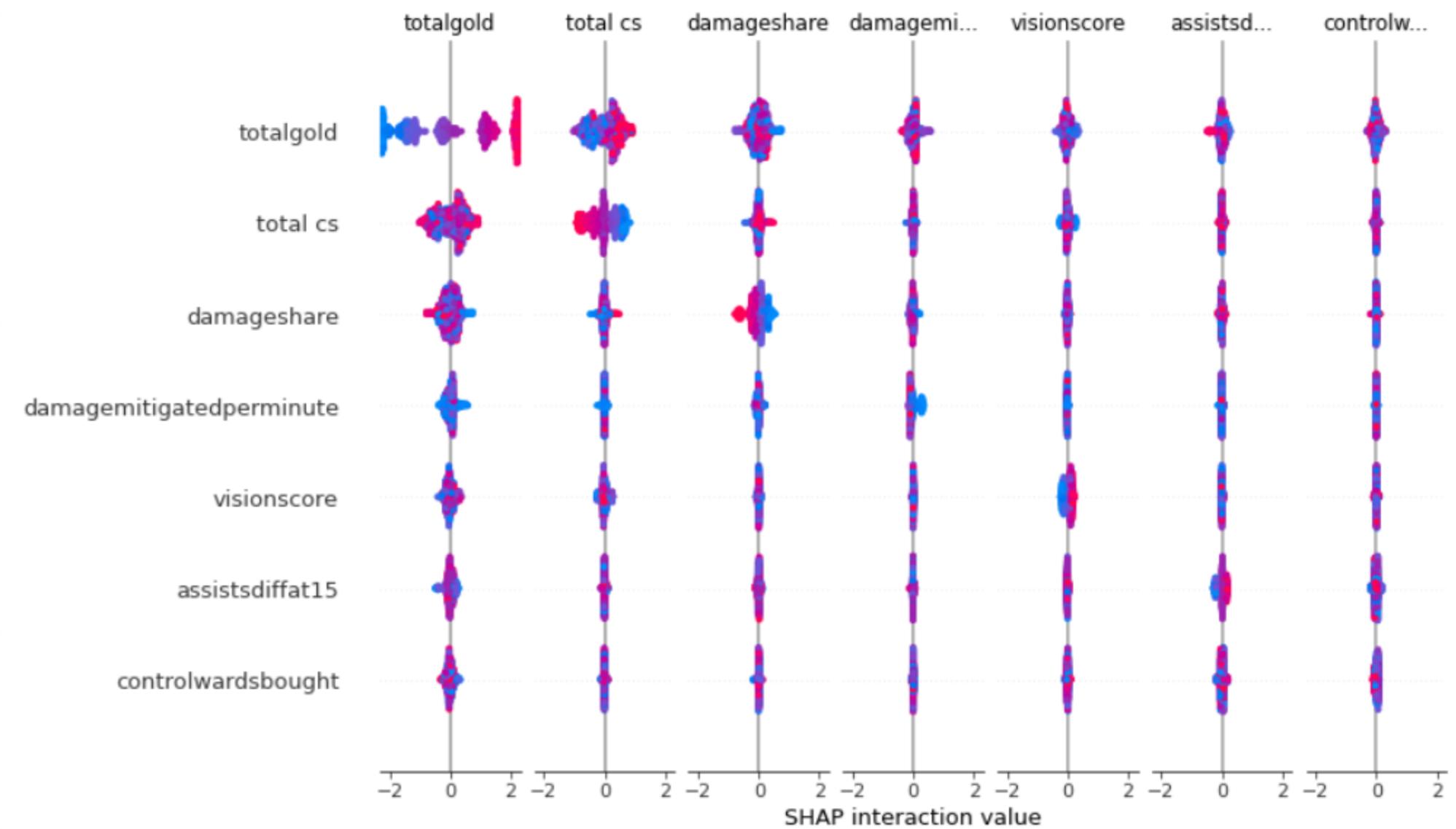
<그림2>에서 Feature importance가 낮은 feature들을 선수평가지표에서 제외

2-2. 데이터 모델링



<그림 3>

<그림3>에서 SHAP values가 낮은 feature들을 선수평가지표에서 제외



<그림 4>

<그림4>에서 SHAP values가 낮은 feature들을 선수평가지표에서 제외

2-3. 데이터 시각화

ex) top_df

	playername	KDA	total cs	damageshare	controlwardsbought	monsterkills
0	Canna	4.844591	283.586957	0.247202	7.565217	25.543478
1	Doran	4.609027	284.648352	0.274653	8.197802	15.120879
2	DuDu	3.794253	266.137931	0.224460	7.413793	17.379310
3	Hoya	3.757622	248.275862	0.226632	9.057471	10.000000
4	Khan	5.081987	251.565657	0.246029	7.656566	18.595960
5	Kiin	4.442236	291.641304	0.271148	7.195652	20.641304
6	Kingen	3.606477	268.021505	0.256408	5.978495	9.763441
7	Morgan	4.680128	252.397436	0.194151	8.192308	13.179487
8	Rascal	4.974721	287.648936	0.226897	7.904255	23.095745
9	Rich	4.197145	251.961165	0.256366	5.106796	11.466019
10	Summit	3.358578	279.329787	0.274101	7.106383	15.446809

지금까지 종합하여, 선수평가지표를 각 포지션별로 5개로 제한해 (이는 포지션마다 달라질 수 있음) 포지션 별 5개의 데이터 프레임을 재구성한 후, 선수 별로 groupby method를 적용시키고 선수의 모든 경기의 평균값으로 데이터프레임 구성

2-3. 데이터 시각화

ex) top_df

	playername	KDA	total cs	damageshare	controlwardsbought	monsterkills
0	Canna	8.898019	8.514144	7.272019	6.978230	10.000000
1	Doran	7.804537	8.709949	10.000000	8.259196	4.716059
2	DuDu	4.022386	5.295171	5.012055	6.671600	5.861015
3	Hoya	3.852346	2.000000	5.227882	10.000000	2.119928
4	Khan	10.000000	2.606897	7.155441	7.163208	6.477819
5	Kiin	7.030301	10.000000	9.651659	6.229872	7.514747
6	Kingen	3.150738	5.642650	8.186892	3.765164	2.000000
7	Morgan	8.134588	2.760343	2.000000	8.248069	3.731832
8	Rascal	9.502079	9.263493	5.254171	7.664772	8.759073
9	Rich	5.892598	2.679860	8.182752	2.000000	2.863156
10	Summit	2.000000	7.728787	9.945162	6.049104	4.881295

평가지표를 점수화 하기 위한 방법으로 MinMaxScaling을 채택
이때, Min값이 0이기 때문에 시각화하는데 있어서 장애가 생길
꺼라 판단하여, MinMaxScaling 후 0~1범위로 만들어진 데이
터들에 8을 곱하고, default값으로 2를 더해주면 평가지표들은
2~10점으로 점수화되며 시각화 가능

2-3. 데이터 시각화

ex) top_df

	playername	KDA	total cs	damageshare	controlwardsbought	monsterkills	score	rank
0	Rich	5.892598	2.679860	8.182752	2.000000	2.863156	21.618365	C
1	Kingen	3.150738	5.642650	8.186892	3.765164	2.000000	22.745445	C
2	Hoya	3.852346	2.000000	5.227882	10.000000	2.119928	23.200156	C
3	Morgan	8.134588	2.760343	2.000000	8.248069	3.731832	24.874832	B
4	DuDu	4.022386	5.295171	5.012055	6.671600	5.861015	26.862227	B
5	Summit	2.000000	7.728787	9.945162	6.049104	4.881295	30.604349	B
6	Khan	10.000000	2.606897	7.155441	7.163208	6.477819	33.403365	A
7	Doran	7.804537	8.709949	10.000000	8.259196	4.716059	39.489741	A
8	Kiin	7.030301	10.000000	9.651659	6.229872	7.514747	40.426578	S
9	Rascal	9.502079	9.263493	5.254171	7.664772	8.759073	40.443589	S
10	Canna	8.898019	8.514144	7.272019	6.978230	10.000000	41.662413	S
11	Mean	5.857299	5.433441	6.490670	6.085768	4.910410	28.777588	NaN

5개의 평가지표를 총 합산하여 'Score'를 산출

'Score'컬럼을 사분위수로 나누어 S, A, B, C 등급으로 선수들을 등급화하여 'Rank'컬럼 생성

또한 시각화 하는 과정에서 해당 포지션의 선수들의 평균이 어떠한지 파악하기 위하여 해당 포지션 선수들의 평가지표를 평균으로 산출한 'Mean' 데이터를 기존 데이터프레임에 append

2-3. 데이터 시각화

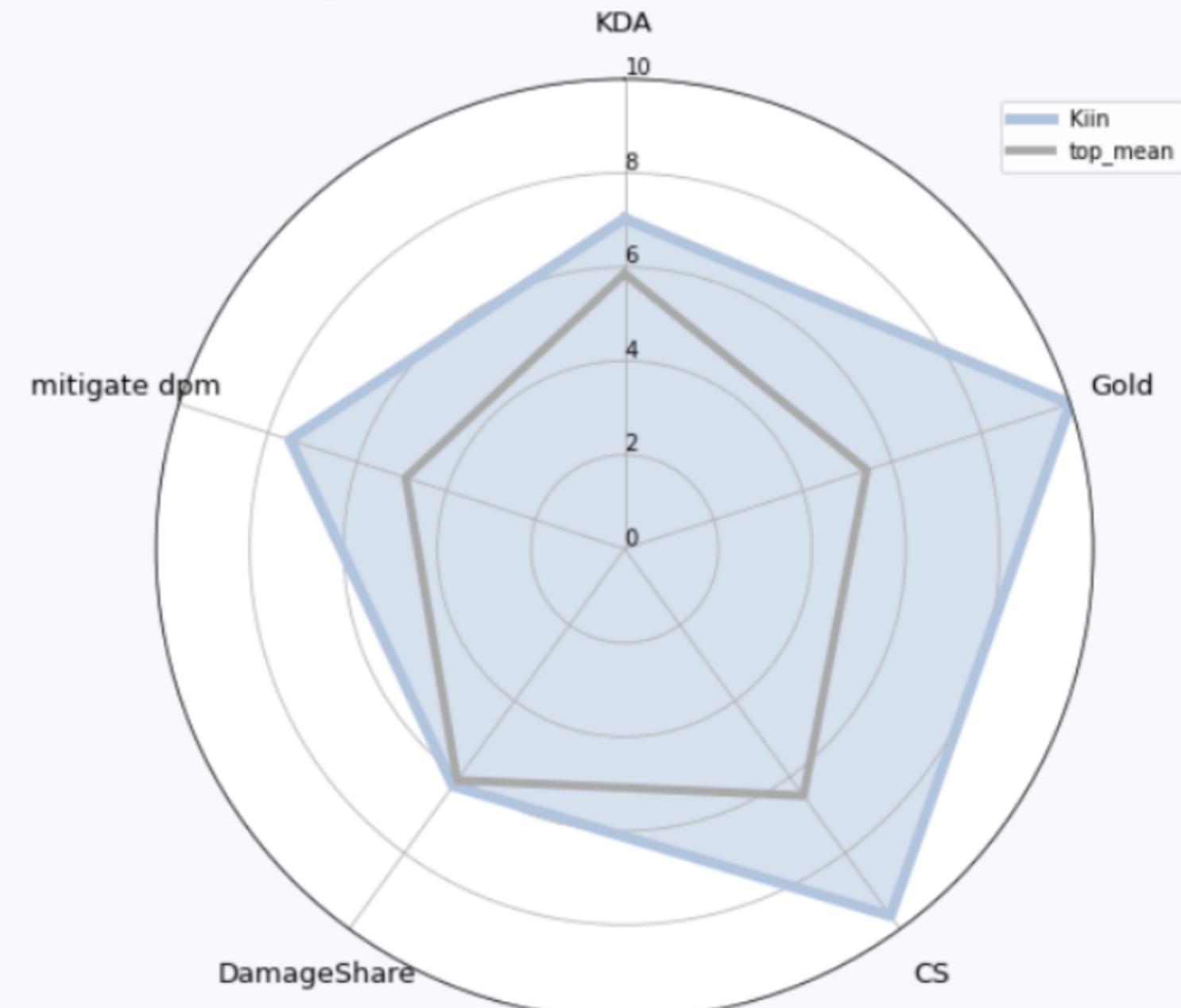
input문을 활용하여 원하는 라인, 플레이어의 선수평가 그래프 확인 가능

- 좌측 상단을 확인하면 선수 이름, 선수 RANK를 확인 가능
- 중앙에 있는 오각형은 해당 포지션의 모든 선수평가의 평균 그래프
- 오각형의 각 꼭짓점에서 앞서 추출한 5개의 평가지표 확인 가능

what lane?:

what name?:

Kiin
RANK : S



3. 결과

3. 결과

	playername	KDA	total cs	damageshare	controlwardsbought	monsterkills	score	rank
0	Rich	5.892598	2.679860	8.182752	2.000000	2.863156	21.618365	C
1	Kingen	3.150738	5.642650	8.186892	3.765164	2.000000	22.745445	C
2	Hoya	3.852346	2.000000	5.227882	10.000000	2.119928	23.200156	C
3	Morgan	8.134588	2.760343	2.000000	8.248069	3.731832	24.874832	B
4	DuDu	4.022386	5.295171	5.012055	6.671600	5.861015	26.862227	B
5	Summit	2.000000	7.728787	9.945162	6.049104	4.881295	30.604349	B
6	Khan	10.000000	2.606897	7.155441	7.163208	6.477819	33.403365	A
7	Doran	7.804537	8.709949	10.000000	8.259196	4.716059	39.489741	A
8	Kiin	7.030301	10.000000	9.651659	6.229872	7.514747	40.426578	S
9	Rascal	9.502079	9.263493	5.254171	7.664772	8.759073	40.443589	S
10	Canna	8.898019	8.514144	7.272019	6.978230	10.000000	41.662413	S

탑 데이터프레임에서 볼 때 GenG 라스
칼, T1 칸나, 아프리카 기인 선수가 S 랭
크에 위치되었다.

올해 비교적 팀 순위가 높고 좋은 평가를
받았던 선수들이 높게 랭크된 것으로 보
아 분류가 잘 되었다고 판단하였다.

감사합니다.