2020 빅콘테스트 결과보고서

현풍데이타(Hyeonpung Data)
팀장 | 한현영 | hyhan@dgist.ac.kr
팀원 | 고낙헌 | skrgjsdl23@dgist.ac.kr
팀원 | 김주형 | kimmold@dgist.ac.kr
팀원 | 정회성 | anen0310@gmail.com
팀원 | 최세영 | cyclone989@dgist.ac.kr

Index

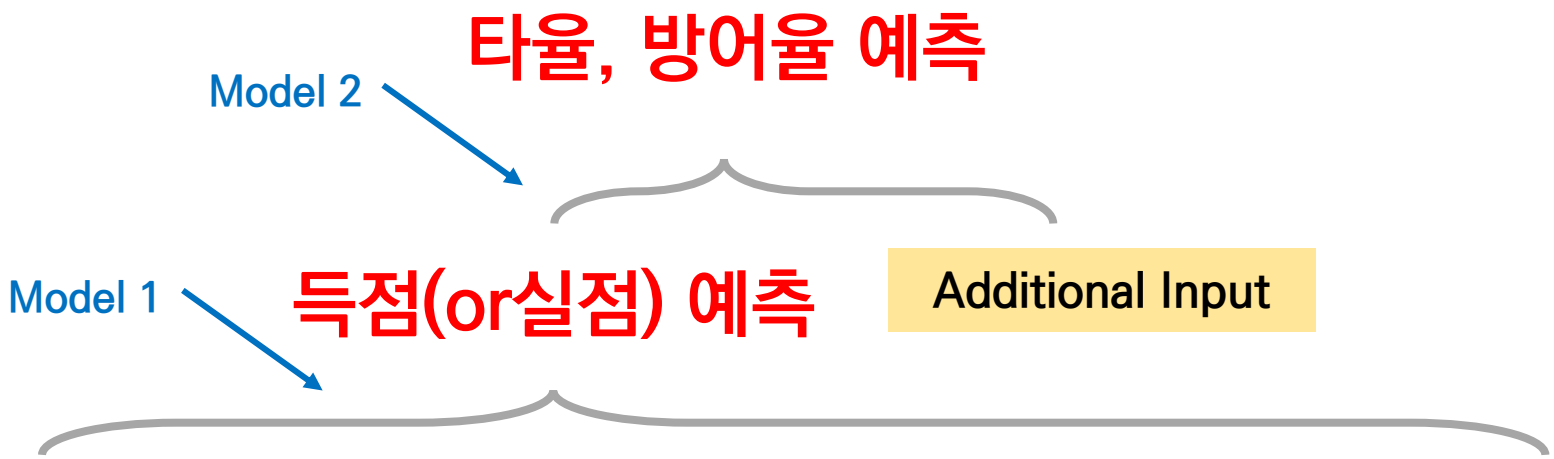
01

현풍데이타의 예측 아이디어

선수들의 누적데이터를 바탕으로 득점을 예측하자!

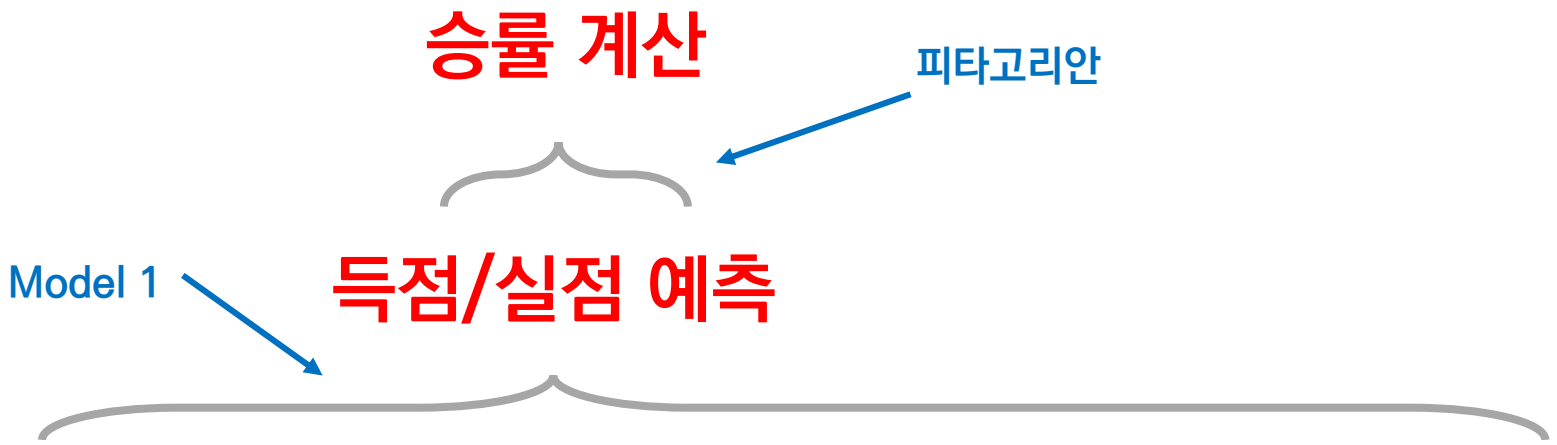
input1	input2-1	input2-2	input3
타자 누적데이터	선발투수 누적데이터	불펜투수 누적데이터	경기외 데이터
• 출루율(OBP) • 장타율(SLG) • OPS(OBP+SLG) • BABIP • wOBA • KK/BB	• 피장타율 • 방어율 • 삼진/볼넷 • 피출루율 • 피OPS • 잔루처리율	• 피장타율_ • 방어율_ • 삼진/볼넷_ • 피출루율_ • 피OPS_ • 잔루처리율_	• 의욕지수(최근경기전적) • Park Factor • BPF • PPF

예측된 득점을 이용, 방어율·타율을 예측하자!



input1	input2-1	input2-2	input3
타자 누적데이터	선발투수 누적데이터	불펜투수 누적데이터	경기외 데이터
• 출루율(OBP) • 장타율(SLG) • OPS(OBP+SLG) • BABIP • wOBA • KK/BB	• 피장타율 • 방어율 • 삼진/볼넷 • 피출루율 • 피OPS • 잔루처리율	• 피장타율_ • 방어율_ • 삼진/볼넷_ • 피출루율_ • 피OPS_ • 잔루처리율_	• 의욕지수(최근경기전적) • Park Factor • BPF • PPF

예측된 득점을 이용, 승률을 예측하자!



input1	input2-1	input2-2	input3
타자 누적데이터	선발투수 누적데이터	불펜투수 누적데이터	경기외 데이터
• 출루율(OBP) • 장타율(SLG) • OPS(OBP+SLG) • BABIP • wOBA • KK/BB	• 피장타율 • 방어율 • 삼진/볼넷 • 피출루율 • 피OPS • 잔루처리율	• 피장타율_ • 방어율_ • 삼진/볼넷_ • 피출루율_ • 피OPS_ • 잔루처리율_	• 의욕지수(최근경기전적) • Park Factor • BPF • PPF

02

데이터 전처리

타자데이터 전처리

- 타자데이터의 모든 데이터는 상대 팀에 대한 데이터만 누적시켜 적용하였다.

  ex) LG vs HH 일 경우 HH를 상대했을 때의 경기들의 데이터만 누적

타자데이터 전처리

OBP	SLG	OPS
BABIP	KK/BB	wOBA

	G_ID	GDAY_DS	T_ID	VS_T_ID	cOBP	cSLG	cOPS	cBABIP	KK/BB	wOBA	RUN
0	20160401HHLG0	20160401	LG	HH	0.282609	0.333333	0.615942	0.266667	2.750000	0.270872	5
1	20160401HHLG0	20160401	HH	LG	0.326531	0.326087	0.652618	0.361111	3.333333	0.276538	4
2	20160401HTNC0	20160401	NC	HT	0.400000	0.566667	0.966667	0.368421	1.800000	0.406500	5
3	20160401HTNC0	20160401	HT	NC	0.315789	0.411765	0.727554	0.304348	3.333333	0.322579	4
4	20160401KTSK0	20160401	SK	KT	0.277778	0.500000	0.777778	0.321429	0.000000	0.333556	4
...	...	...	...	...	...	...	...	...	...	...	...
635	20200719LTSS0	20200719	LT	SS	0.293578	0.317568	0.611146	0.261411	1.962963	0.274064	2
636	20200719OBHT0	20200719	HT	OB	0.343195	0.381271	0.724466	0.302419	1.361111	0.316644	4
637	20200719OBHT0	20200719	OB	HT	0.364162	0.436306	0.800468	0.354962	2.550000	0.346083	8
638	20200719WOSK0	20200719	SK	WO	0.311765	0.349673	0.661438	0.288210	2.366667	0.295120	4
639	20200719WOSK0	20200719	WO	SK	0.381215	0.443366	0.824581	0.324675	1.511111	0.365380	3

투수데이터 전처리(선발&불펜)

- 투수데이터는 자기 팀을 대상으로 한 상대 팀 선발 및 불펜투수의 개인 기록을 누적시켜 적용하였다.

  ex) LG vs HH 일 경우 LG를 상대했을 때의 HH 선발/불펜 투수들의 데이터를 누적

투수데이터 전처리(선발&불펜)

피장타율	방어율	피OPS
피출루율	삼진/볼넷	잔루처리율

G_ID	GDAY_DS	T_ID	VS_T_ID	투수 ID	선발불펜0	경기수	피장타율	방어율	삼진/볼넷	피출루율	피OPS	9이닝당 실	9이닝당 볼	잔루처리율
20160826L	20160826	WO	LG	62937	start	3	0.333333	3.9375	6.5	0.308824	0.642157	7.3125	1.125	0.666667
20160826L	20160826	LG	WO	63111	start	2	0.358974	4.21875	4.5	0.340909	0.699883	7.59375	1.6875	0.666667
20160826L	20160826	OB	LT	61240	start	2	0.487805	7.83871	2.142857	0.4	0.887805	13.06452	6.096774	0.639535
20160826L	20160826	LT	OB	64021	start	2	0.75	12.27273	3	0.463415	1.213415	11.04545	3.681818	0.608108
20160826N	20160826	NC	HH	65931	start	3	0.430556	4.5	2.142857	0.395062	0.825617	7.5	3.5	0.718954
20160826N	20160826	HH	NC	73750	start	1	0.411765	6.230769		0.263158	0.674923	6.230769	0	0.555556
20160826S	20160826	SK	KT	60841	start	3	0.365079	3.705882	1.3	0.368421	0.7335	6.882353	5.294118	0.634921
20160826S	20160826	KT	SK	67845	start	2	0.542857	7.56	2.666667	0.384615	0.927473	8.64	3.24	0.57377
20160827H	20160827	HH	SK	79764	start	3	0.285714	1.472727	2.333333	0.257143	0.542857	6.872727	2.945455	0.986842
20160827K	20160827	KT	LG	66050	start	1	0.636364	13.5	3.5	0.458333	1.094697	15.75	4.5	0.3125
20160827L	20160827	LT	SS	65543	start	1	0.4	3	0.4	0.4	0.8	3	7.5	0.8
20160827L	20160827	SS	LT	74454	start	1	0.44	2.571429	1	0.4	0.84	3.857143	3.857143	0.943396
20160827O	20160827	HT	OB	77637	start	2	0.510204	7.714286	2	0.339623	0.849827	4.628571	2.314286	0.460526
20160827O	20160827	OB	HT	79535	start	1	0.5625	9.818182	2	0.388889	0.951389	4.909091	2.454545	0.535714
20160828H	20160828	SK	HH	74838	start	1	0.473684	12.46154	0.5	0.538462	1.012146	6.230769	12.46154	0.571429
20160828K	20160828	LG	KT	62698	start	4	0.319149	4.207792	2.6	0.29703	0.616179	4.558442	1.753247	0.6
20160828O	20160828	OB	HT	74513	start	3	0.478261	5.09434	1.333333	0.392405	0.870666	6.113208	4.584906	0.783582
20160830H	20160830	HH	OB	60768	start	1	0.571429	8.1	1	0.375	0.946429	2.7	2.7	0.652174
20160830H	20160830	OB	HH	79229	start	2	0.23913	1.285714	1	0.283019	0.522149	3.857143	3.857143	0.866667
20160830L	20160830	LT	LG	65546	start	4	0.514286	5.684211	2	0.350877	0.865163	5.684211	2.842105	0.697674

경기외데이터 전처리

- 경기외데이터는 해당 팀의 의욕지수를 나타내는 최근 경기 전적, Park Factor, BPF, PPF를 포함한다.

  ex) BPF, PPF의 경우 팀별로 계산

경기외데이터 전처리

최근 경기 전적(WLD) | Park Factor | BPF | PPF

	G_ID	GDAY_DS	T_ID	VS_T_ID	PARK	WLD	BPF	PPF
0	20160401HHLG0	20160401	HH	LG	938	0.0	0.858404	0.879156
1	20160401HHLG0	20160401	LG	HH	938	0.0	0.827749	0.829172
2	20160401HTNC0	20160401	HT	NC	964	0.0	0.851834	0.841000
3	20160401HTNC0	20160401	NC	HT	964	0.0	0.871238	0.869567
4	20160401KTSK0	20160401	KT	SK	1019	0.0	0.866062	0.892822
...	...	...	...	...	...	...	...	...
6395	20200719LTSS0	20200719	SS	LT	1067	0.5	0.917426	0.927864
6397	20200719OBHT0	20200719	HT	OB	1013	0.6	0.851834	0.841000
6396	20200719OBHT0	20200719	OB	HT	1013	0.6	0.831685	0.798610
6399	20200719WOSK0	20200719	SK	WO	1019	0.5	0.907633	0.907751
6398	20200719WOSK0	20200719	WO	SK	1019	0.4	0.862672	0.849878

데이터 병합

타자데이터 | 투수데이터(선발) | 투수데이터(불펜) | 경기외데이터

T_ID	VS_T_ID	투수 ID	경기수	피장타율	방어율	삼진/볼넷	피출루율	...	cSLG	cOPS	cBABIP	KK/BB	wOBA	WLD	PARK	BPF	PPF	RUN
LG	WO	62937	3	0.333333	3.937500	6.500000	0.308824	...	0.402464	0.750612	0.329975	2.205128	0.329119	0.50	983	0.862672	0.849878	3
WO	LG	63111	2	0.358974	4.218750	4.500000	0.340909	...	0.457203	0.831068	0.337629	1.629630	0.354106	0.60	983	0.862672	0.849878	2
LT	OB	61240	2	0.487805	7.838710	2.142857	0.400000	...	0.510417	0.920905	0.385675	1.634921	0.392502	0.30	938	0.831685	0.798610	4
OB	LT	64021	2	0.750000	12.272727	3.000000	0.463415	...	0.479570	0.854570	0.319892	1.509804	0.371804	0.80	938	0.831685	0.798610	11
HH	NC	65931	3	0.430556	4.500000	2.142857	0.395062	...	0.386473	0.721256	0.332258	2.750000	0.315769	0.30	1000	0.858404	0.879156	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
SK	HH	69748	2	0.215686	0.627907	4.666667	0.236364	...	0.403346	0.756964	0.323601	2.053571	0.331771	0.30	1000	0.858404	0.879156	6
NC	OB	68240	3	0.278689	2.647059	3.000000	0.333333	...	0.421793	0.757805	0.313305	2.447368	0.327551	0.55	938	0.831685	0.798610	5
OB	NC	66920	1	0.266667	2.250000	0.250000	0.400000	...	0.398577	0.748499	0.293617	1.303030	0.329462	0.85	938	0.831685	0.798610	6
LT	WO	69343	3	0.275362	2.368421	3.000000	0.266667	...	0.286528	0.548176	0.284635	5.120000	0.244900	0.20	1038	0.877477	0.885673	1
WO	LT	77318	2	0.425000	9.000000	1.333333	0.404255	...	0.435252	0.802402	0.360000	2.160714	0.347722	0.60	1038	0.877477	0.885673	3

03

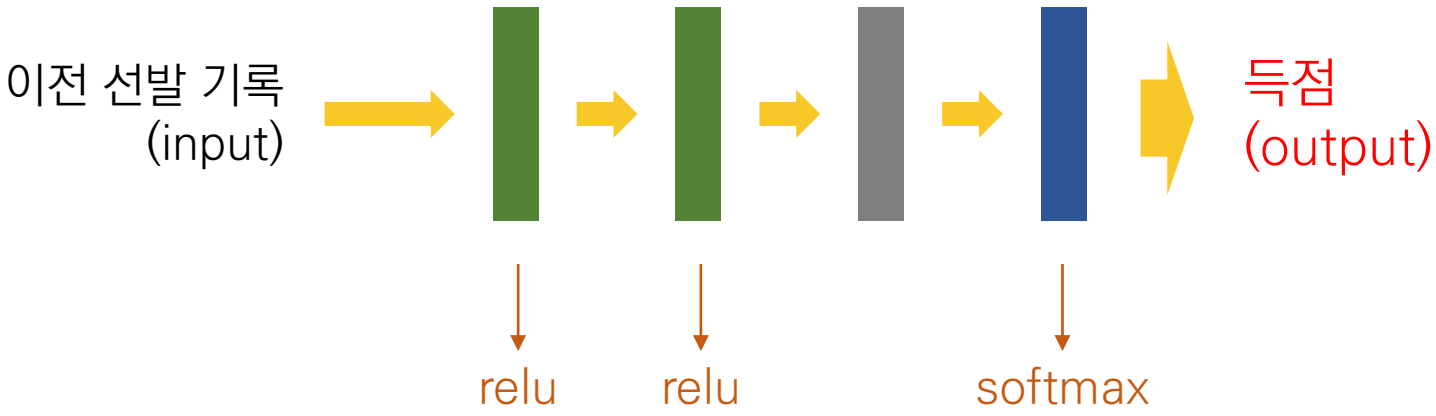
예측 모델 설계

2020년 후반기 선발 투수 예측 모델

- 득점 예측 모델에 2020년 후반기 경기 데이터를 input하여 득점을 predict해야 하므로,
  해당 경기에 어떤 투수가 선발될 지 예측해야 한다.
- 예측 방법: 4경기 전부터의 선발(4경기 전, 3경기 전, 2경기 전, 1경기 전) 기록을 통해 5경기 째마다의 선발 경향성을 보기 위한 모델.
- 사용 모델: RNN
- 선정 이유: 이전 선발 기록이 현재 예측에 영향을 끼치기 때문

2020년 후반기 선발 투수 예측 모델

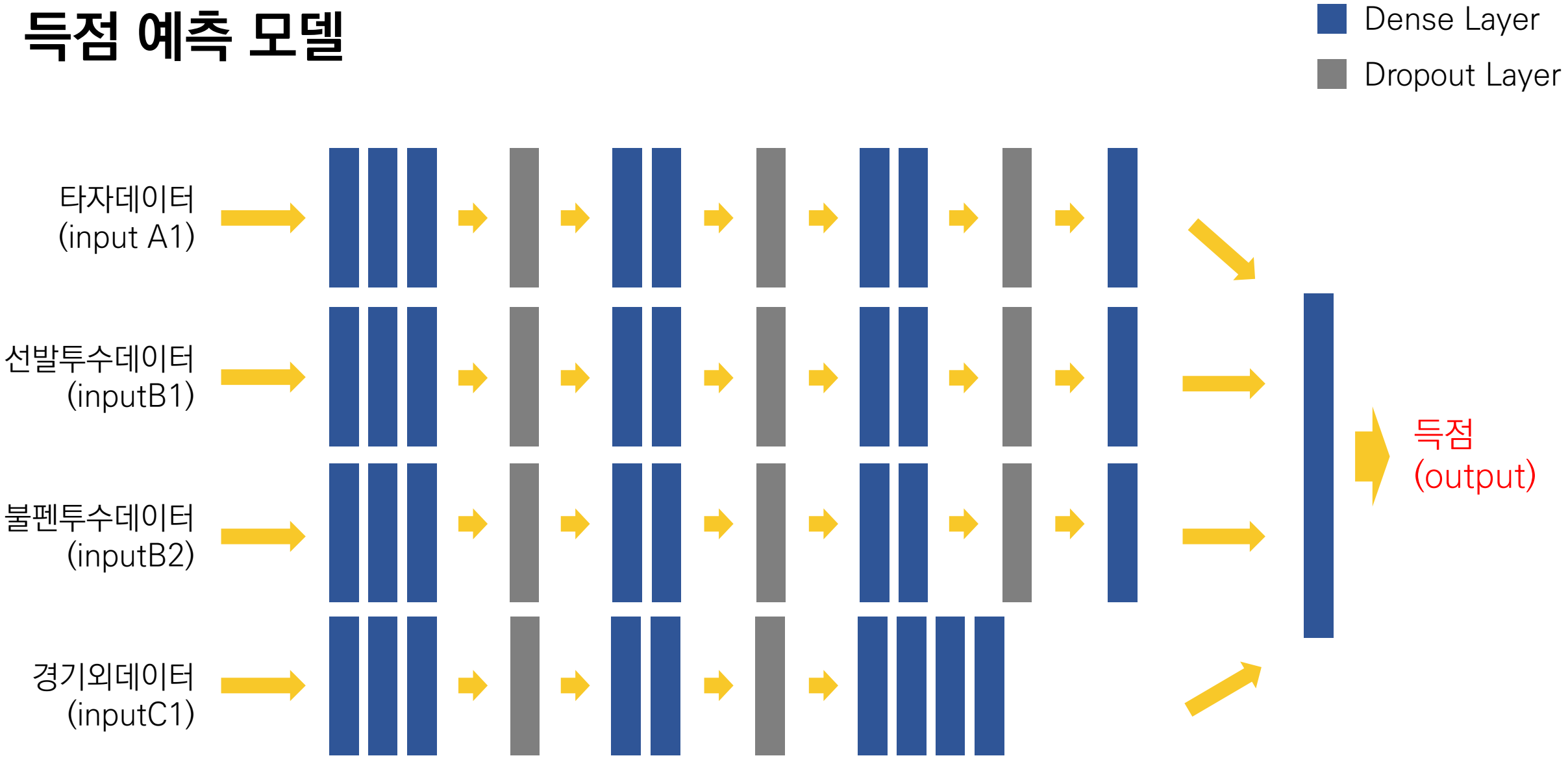
Dense Layer
LSTM Layer
Dropout Layer



득점 예측 모델

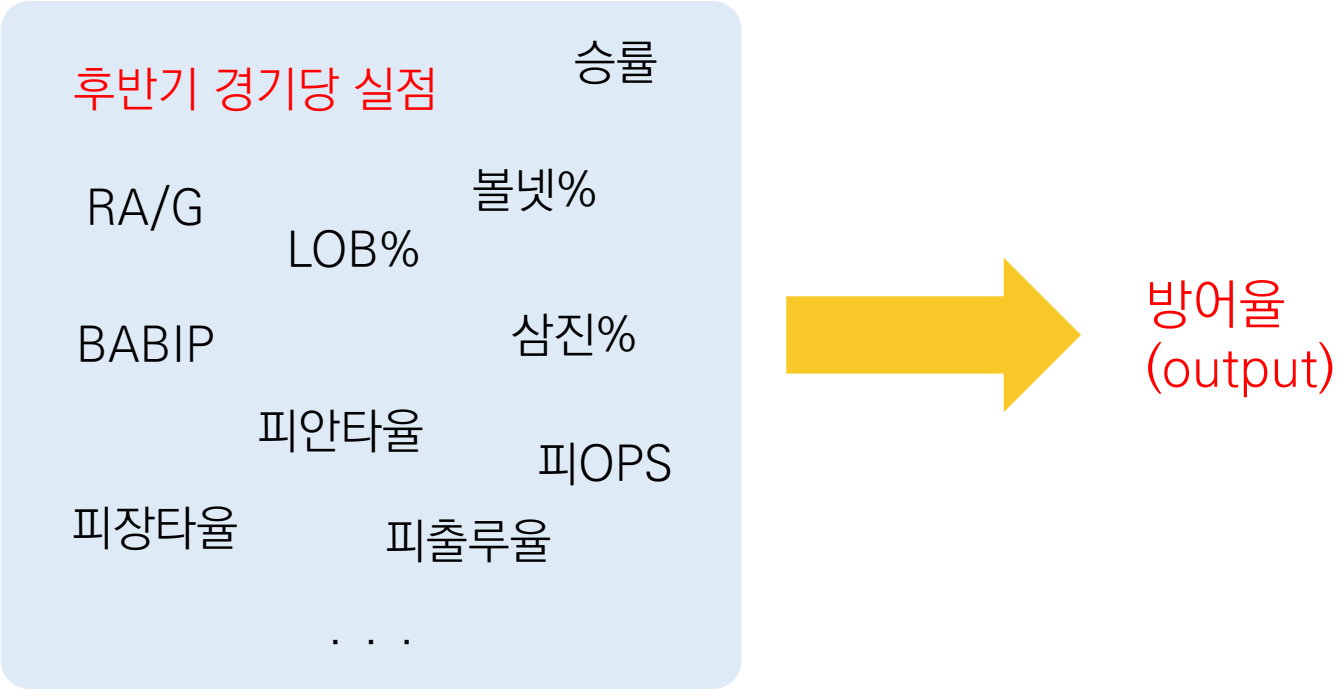
- 앞에서 가공한 데이터를 학습하는 득점 예측 모델을 형성한다.
- 예측 방법: 누적 타자데이터, 누적 투수데이터, 경기외데이터를 종합하여 경향성 분석 → 득점 예측
- 사용 모델: DNN
- Activation Function: mish

득점 예측 모델


Dense Layer
Dropout Layer
타자데이터
(input A1)
선발투수데이터
(inputB1)
불펜투수데이터
(inputB2)
경기외데이터
(inputC1)
득점
(output)

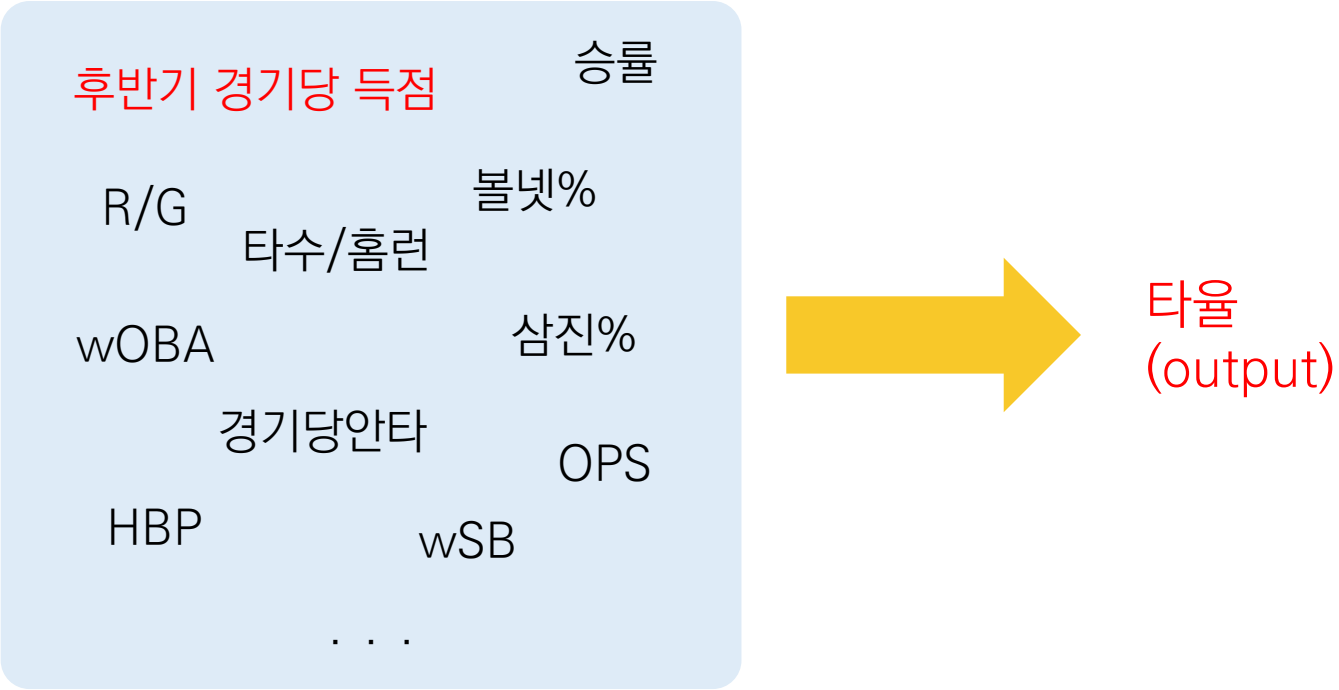
방어율 예측 모델

- 예측된 득점 → 후반기 경기당 실점 → 그 외 다른 지표들로 방어율을 예측한다.
- 사용 모델: Linear Regression
- 교차 검증(Cross Validation)을 통한 모델 검증 과정 추가



타율 예측 모델

- 예측된 득점 → 후반기 경기당 득점 → 그 외 다른 지표들로 방어율을 예측한다.
- 사용 모델: Bagging Regressor, Gradient Boosting Regressor, Voting Regressor
- 세 모델 예측값의 평균값으로 예측. 후에 교차 검증(Cross Validation)을 통한 모델 검증



승률 예측 모델

- 예측된 득점 → 후반기 득점/실점 예측 → 승률 예측
- 예측 방법: 피타고리안 승률 공식

04

예측 결과

타자_Predict데이터 생성

- 득점 예측 모델에 input하여 predict하기 위한 2020년 후반기 타자데이터 생성
- 2020년 후반기 기록을 KBReport에서 크롤링하여 csv 데이터 형성

T_ID	VS_T_ID	cOBP	cSLG	cOPS	cBABIP	KK/BB	wOBA
LG	HT	0.387	0.51	0.897	0.371	1.989796	0.395
LG	KT	0.343	0.376	0.719	0.292	1.3	0.327
LG	OB	0.338	0.432	0.77	0.336	2.880597	0.342
LG	WO	0.308	0.37	0.678	0.295	2.685714	0.304
LG	SS	0.302	0.405	0.707	0.249	2.445946	0.314
LG	LT	0.348	0.429	0.777	0.344	2.507042	0.344
LG	NC	0.377	0.501	0.878	0.331	1.777778	0.386
LG	HH	0.365	0.459	0.824	0.338	1.988506	0.367
LG	SK	0.396	0.485	0.881	0.342	1.371681	0.391
HT	LG	0.323	0.382	0.705	0.297	2.204545	0.318
HT	KT	0.363	0.42	0.783	0.324	1.65625	0.353
HT	OB	0.358	0.392	0.75	0.307	1.238938	0.342

선발투수_Predict데이터 생성

- 득점 예측 모델에 input하여 predict하기 위한 2020년 후반기 투수데이터 생성
- 2020년 후반기 기록을 KBReport에서 크롤링하여 csv 데이터 형성한 후, 선발투수 예측 모델로부터 예측된 향후 선발투수 예상 데이터 형성

T_ID	VS_T_ID	P_ID	피장타율	방어율	삼진/볼넷	피출루율	피OPS	잔루처리율
LG	HT	69103	0.292	1.38	3.67	0.255	0.547	86.2
LG	HT	68135	0.434	3.75	3	0.339	0.773	68.2
LG	HT	50126	0.36	6	4	0.346	0.706	55.6
LG	HT	78148	0.318	2.25	0.83	0.333	0.652	82.4
LG	HT	76455	0.417	1.5	-	0.25	0.667	108.7
LG	HT	61101	0.403	6.11	1.06	0.411	0.814	61.4
LG	KT	61101	0.28	6	8	0.269	0.549	42.9
LG	KT	50157	0.225	0.71	1	0.255	0.48	91.7
LG	KT	68135	0.35	4.5	3	0.267	0.617	65.2
LG	KT	50126	0.316	1.8	0.8	0.44	0.756	90.9
LG	KT	69103	0.443	4.26	1.29	0.367	0.81	80.6
LG	KT	76455	0.644	9	3.5	0.375	1.019	58
LG	OB	69103	0.402	4.12	6.33	0.299	0.701	68.7
LG	OB	76455	0.54	11.25	1.29	0.459	0.999	45.1
LG	OB	78148	0.667	6.75	-	0.368	1.035	35.7

GDAY	T_ID	VS_T_ID	P_ID	피장타율	방어율
09월 19일	HT	HH	50640	0.262	3.09
08월 09일	HT	NC	77637	1.048	16.62
08월 02일	HT	LT	50636	0.216	1.25
09월 10일	HT	OB	62754	0.375	2
08월 26일	HT	OB	65616	0.397	5.19
07월 31일	HT	LT	50640	0.283	3.63
08월 13일	HT	LG	77637	0.341	3.86
08월 30일	HT	KT	50636	0.167	0
08월 12일	HT	LG	62754		
07월 30일	HT	KT	65616	0.565	10.12
09월 13일	HT	NC	50640	0.383	4.18
08월 23일	HT	WO	77637	0.387	3.67
09월 15일	HT	SK	50636	0.435	4.05
08월 22일	HT	WO	62754	0.771	13.5
08월 01일	HT	LT	65616	0.429	12.15
08월 21일	HT	NC	50640	0.383	4.18
09월 12일	HT	NC	77637	1.048	16.62
09월 16일	HT	SK	50636	0.435	4.05

불펜투수_Predict데이터 생성

- 득점 예측 모델에 input하여 predict하기 위한 2020년 후반기 투수데이터 생성
- 2020년 후반기 기록을 KBReport에서 크롤링하여 csv 데이터 형성

T_ID	VS_T_ID	방어율_	피출루율_	피OPS_	삼진/볼넷_	피장타율_	잔루처리율
HH	SK	6.45	0.4	0.89	1.297297	0.49	0.617647
WO	SK	6.06	0.336	0.746	1.68	0.41	0.592705
LT	SK	5.09	0.32	0.703	1.933333	0.383	0.652174
HT	SK	5	0.396	0.841	1.142857	0.445	0.666667
SS	SK	3.31	0.296	0.624	3.6	0.328	0.767196
OB	SK	3.2	0.335	0.668	2.125	0.333	0.783699
LG	SK	3.05	0.308	0.648	2.692308	0.34	0.79918
KT	SK	2.96	0.332	0.694	1.238095	0.362	0.794702
NC	SK	2.47	0.321	0.611	1.608696	0.29	0.821678
WO	OB	3.04	0.332	0.635	1.346154	0.303	0.740181
HH	OB	3.12	0.327	0.706	1.588235	0.379	0.829493
KT	OB	4.07	0.346	0.722	1.037037	0.376	0.687135
SK	OB	4.37	0.39	0.849	1.035714	0.459	0.725191
LG	OB	4.9	0.369	0.765	1.16129	0.396	0.674256
SS	OB	5.52	0.381	0.762	1	0.381	0.641026

경기외_Predict데이터 생성

- 탐색적 데이터 분석 과정에서의 득점과의 무관성 → 데이터 삭제 결정

득점 예측

- **2020년 데이터 input하여 predict**

득점 예측

- 2020년 데이터 input하여 predict

팀	경기당득점	경기당실점
HH	5.089861	4.604349
HT	4.403873	5.309298
KT	5.109405	4.7127
LG	5.076563	4.992215
LT	4.605638	4.992411
NC	4.95137	5.178043
OB	4.932087	5.166027
SK	5.253654	4.581849
SS	5.2436	4.955448
WO	5.005384	5.145807

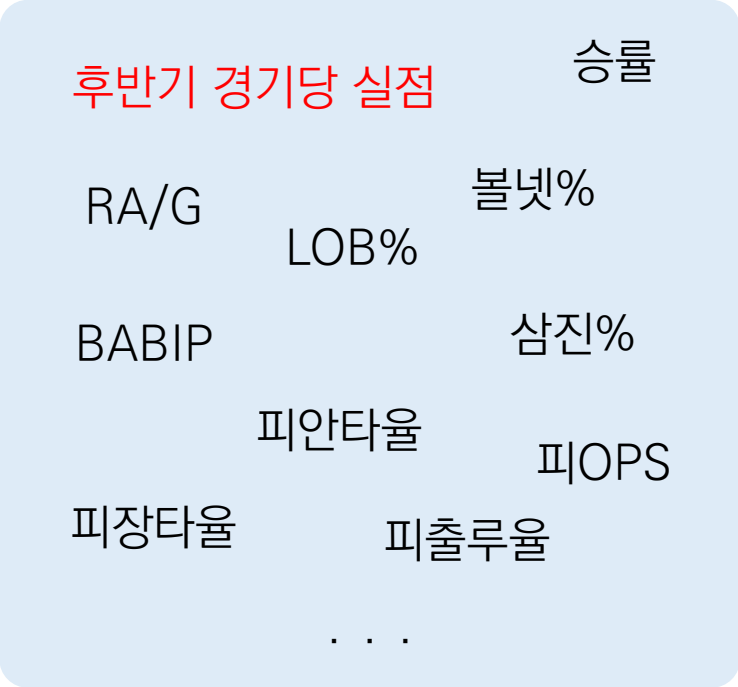
승률 예측

- 피타고리안 승률 예측 방식

팀	경기당득점	경기당실점	승률
HH	5.089861	4.604349	0.545736
HT	4.403873	5.309298	0.415284
KT	5.109405	4.7127	0.536909
LG	5.076563	4.992215	0.507665
LT	4.605638	4.992411	0.463175
NC	4.95137	5.178043	0.479532
OB	4.932087	5.166027	0.478811
SK	5.253654	4.581849	0.562271
SS	5.2436	4.955448	0.525835
WO	5.005384	5.145807	0.487345

방어율 예측

- 2020년 데이터 input하여 predict



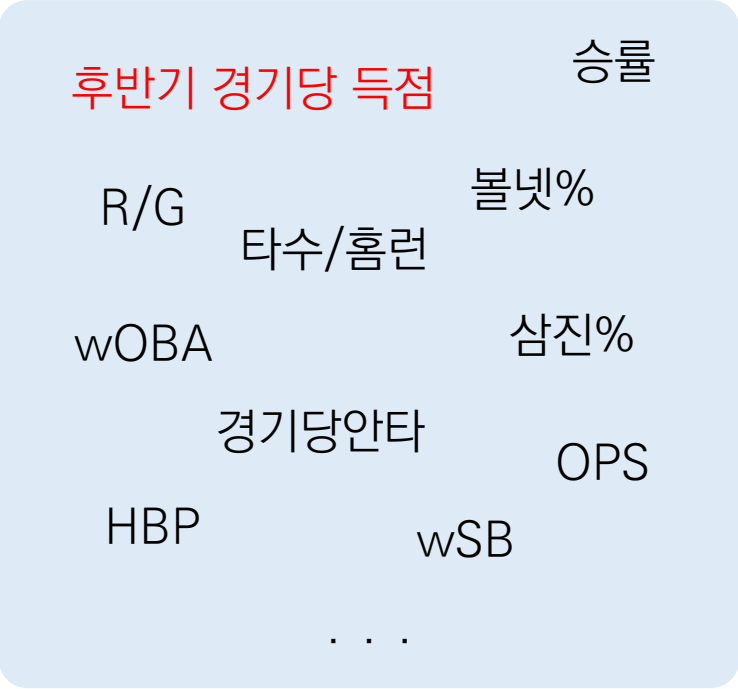
2020년 후반 데이터

2020 팀별 방어율 (output)

Hero	1.91766043,
KIA	5.79907324,
KT	2.65726899,
LG	4.04668223,
NC	4.09901947,
SK	5.63961309,
두산	4.70403755,
롯데	1.91489937,
삼성	3.88771714,
한화	5.14223079

타율 예측

- 2020년 데이터 input하여 predict



2020년 후반 데이터

2020 팀별 타율 (output)

팀	경기당득점	경기당실점	승률	타율
Hero	5.005384	5.145807	0.487345	[0.289704612203669 5,
KIA	4.403873	5.309298	0.415284	0.259487636263584 5,
KT	5.109405	4.7127	0.536909	0.296487811010219 4,
LG	5.076563	4.992215	0.507665	0.293477637595536 4,
NC	4.95137	5.178043	0.479532	0.285981744771542 5,
SK	5.253654	4.581849	0.562271	0.303603658281106 7,
두산	4.932087	5.166027	0.478811	0.288312290431390 45,
롯데	4.605638	4.992411	0.463175	0.267811089277700 8,
삼성	5.2436	4.955448	0.525835	0.300823536087591 66,
한화	5.089861	4.604349	0.545736	0.280786998192555 4]

05

개선할 점

개선할 점

- 탐색적 데이터 분석 과정의 미숙
- 딥러닝 모델 구축에서 구성에 대한 부분이나, 하이퍼파라미터 최적화에 대한 어려움
- 경험 부족으로 인한 전반적인 진행 과정에서의 크고 작은 애로사항 존재

THANK YOU

현풍데이타(Hyeonpung Data)