

Part. 07 빅콘테스트

|2015빅콘테스트 야구경기 예측 알고리즘 개발

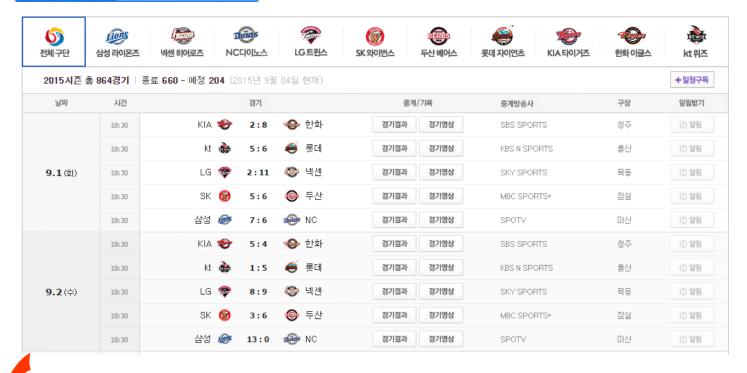
FASTCAMPUS ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택

■ 데이터수집

NAVER SPORTS



NAVER SPORTS 이용 – 팀별 일별 경기기록 수집(2013~2015)



■ 데이터수집



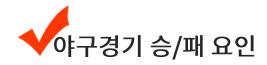
✔NAVER SPORTS 이용 - 경기 별 상세 기록 수집(선발 투수 방어율, 불펜 투수 방어율, 타자기록 등)

[KIA 데이터프레임]

	A	В	С	D	Е	F	G	Н	I J	K	L	М	N	0	Р	O	R	S	Т	U	V	W	Х	Υ	Z	AA	AB	AC
1	날짜	요일	득점1	득점2	팀1	팀2	지역	요일코	연속일주말.	원정.	결과	득점	실점	상대팀	home_	home_	away_	away_l	home_	home_	away_l	away_l	defen	bul_de	attack	attack	최근te	yester
2	9 14	(일)	6	10	KIA	한화	대전	7	1 주말	원정	패	6	10	한화	5.65	NA	8.08	6.788	12	0.288	13	0.288	8.08	6.788	13	0.288	0.3	승
3	9 13	(토)	3	1	KIA	한화	대전	6	1 주말	원정	승	3	1	한화	4.7	4.14	4.8	4.228	4	0.288	9	0.288	4.8	4.228	9	0.288	0.4	패
4	9 12	(금)	4	14	KIA	삼성	대구	5	1 주말	원정	ᅖ	4	14	삼성	4.28	6.35	4.27	5.914	19	0.304	11	0.288	4.27	5.914	11	0.288	0.3	패
5	9 11	(목)	4	5	KIA	삼성	대구	4	1 평일	원정	ᅖ	4	5	삼성	5.33	4.36	6.09	4.606	13	0.303	7	0.288	6.09	4.606	7	0.288	0.3	패
6	9 10	(수)	12	6	LG	KIA	광주	3	1 평일	홈	패	6	12	LG	6.06	6.4	5.17	5.117	10	0.289	10	0.277	6.06	6.4	10	0.289	0.3	승
7	9 9	(화)	7	10	LG	KIA	광주	2	2 평일	홈	승	10	7	LG	7.76	5.09	4.16	4.629	10	0.289	13	0.277	7.76	5.09	10	0.289	0.4	패
8	9 7	(일)	3	6	KIA	NC	마산	7	1 주말	원정	ᅖ	3	6	NC	3.75	4.685	5.7	5.168	11	0.287	6	0.289	5.7	5.168	6	0.289	0.3	승
9	9 6	(토)	9	0	KIA	NC	마산	6	6 주말	원정	승	9	0	NC	5.29	4.575	3.82	4.391	6	0.287	15	0.29	3.82	4.391	15	0.29	0.3	퍠
10	8 31	(일)	4	3	SK	KIA	광주	7	1 주말	홈	ᅖ	3	4	SK	7.5	4.457	9	6.024	10	0.289	11	0.29	7.5	4.457	10	0.289	0.2	패
11	8 30	(토)	7	2	SK	KIA	광주	6	1 주말	홈	ᅖ	2	7	SK	5.79	NA	4.18	5.713	7	0.289	12	0.289	5.79	NA	7	0.289	0.3	승
12	8 29	(금)	9	5	KIA	롯데	사직	5	1 주말	원정	승	9	5	롯데	4.26	10.95	3.98	6.573	8	0.289	12	0.289	3.98	6.573	12	0.289	0.4	패
13	8 28	(목)	5	6	KIA	롯데	사직	4	1 평일	원정	ᅖ	5	6	롯데	5.57	4.978	6.75	5.597	11	0.289	11	0.289	6.75	5.597	11	0.289	0.4	패
14	8 27	(수)	4	6	KIA	넥센	목동	3	2 평일	원정	ᅖ	4	6	넥센	3.57	2.565	4.03	4.436	8	0.298	8	0.289	4.03	4.436	8	0.289	0.5	패
15	8 25	(월)	9	0	한화	KIA	광주	1	1 평일	홈	ᅖ	0	9	한화	5.6	5.47	5.53	NA	3	0.289	13	0.29	5.6	5.47	3	0.289	0.5	승
16	8 23	(토)	5	6	한화	KIA	광주	6	1 주말	홈	승	6	5	한화	4.15	6.387	3.12	5.436	9	0.29	12	0.289	4.15	6.387	9	0.29	0.5	패
17	2177	(二)	2	2	KΤΔ	IG	자신	5	5 즈만	위저	πH	2	2	IG	2 75	3 10	2 57	2 227	Q	ი 279	2	n 291	3 57	2 227	Q	n 291	04	πH



■ 분석방법론



야구팬들의 대화

- 1) 오늘 어디랑 해?
 - 약팀이냐 vs 강팀이냐
 - 먹고먹히는 앙숙 관계
- 2) 오늘 선발 누구야?
 - 야구는 투수 놀음
- 3) 최근 분위기 좋더라?
 - 야구는 분위기 싸움
- 4) 홈경기야?
 - 유난히 홈경기에 강한 팀



상대팀이 중요



선발ㆍ불펜 방어율 중요



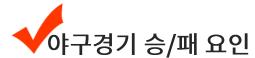
최근 경기 결과 중요

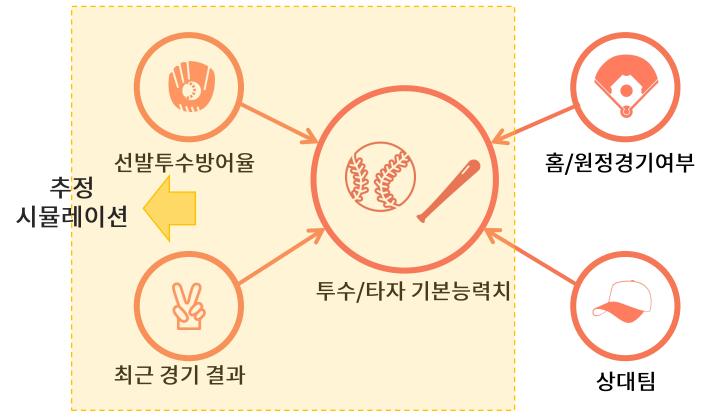


홈/원정 경기 중요



■ 분석방법론



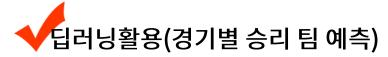


FAST CAMPUS ONLINE 이경택 강사.

예측구간 선수 능력치/ 그때 당시에 팀 최근 승률?



■ 분석방법론



1) 적용변수

원정여부, 연속경기일수, 주말/주중경기, 홈/원정경기, 이전경기 승리여부, 상대팀과의 연전수(ex 3연전중 1경기), 선발투수 방어율, 타자평균타율 등

- 2) 모형 설명력 향상을 위한 Bagging 기법적용
- Boostrap시 최근의 데이터에 가중치
- 각 데이터가 뽑힐 확률의 분포 형태 Y = log(x)
- 3) 1경기 예측 시, 2종류의 모델사용
- ① 해당 팀의 올해 전적 데이터 승리 예측 이길 확률
- ② 상대 전적 데이터(올해 + 작년) 이길 확률
 - ① 의 확률 + 가중치 ②



경기 승/패 예측

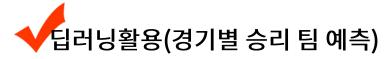
Ex) ① KIA 승리확률 0.82 (승리) / ② KIA승리 확률 0.31(패배) => 0.82 - 0.2 = 0.62 (승리 예측)

FAST CAMPUS
ONLINE

2014~2013년 시뮬레이션을 통해 선정



■ 분석방법론



- 4) 고려요건
 - -야구는 이전의 경기에 영향을 많이 받는다(분위기)
 - -하지만 예측 구간의 경기 승/패, 팀 경기력(방어율,타율)을 알수없음

▶예측시의 결과를 '맞다'고 가정 / 팀 경기력 -> 추정

- 방어율은 대부분 5선발체제 이기 때문에 어느 정도 추정가능
- 타율은 경기후반이기 때문에 거의 변동 없음
 - -> 최근경기 타율에 난수(Uniform, min=-0.05, max=0.05) 추출하여 타율에 +/- 적용

첫 예측이 어떻게 되냐에 따라 예측구간의 결과가 달라짐



시뮬레이션 필요(각 팀마다 시뮬레이션 10회)



■ 분석방법론



- 여러가지 모형을 결합해 안정적인 성능을 낼수 있도록 하는 기법
- 모형설명력 향상
- 독립변수에 난수추출 등 Random한 요소존재
- 시뮬레이션이 목적이기 때문에, 다수의 모형이 결합된 안정된 모형 사용
- Bootstrap시 최근의 데이터에 가중치 (야구는 분위기 싸움)



■ 시뮬레이션

9월 6일 KIA vs 삼성

KIA 전체 데이터프레임

상대팀	어제결과	최근승률	
LG	승	0.7	
두산	패	0.8	

KIA vs 삼성 데이터프레임

상대팀	어제결과	최근승률	
삼성	승	0.5	
삼성	패	0.6	

딥러닝 모형 10회반복(Bagging)

① KIA 승리 확률: 0.782



② KIA 승리 확률 : 0.412(가중치로 사용)

결과 예측: KIA 승리 확률: 0.782 - 0.2: 0.582 (KIA 승리)

9월 8일 KIA vs NC

KIA 전체 데이터프레임

상대팀	어제결과	최근승률	
삼성	승	0.8	
LG	승	0.7	

KIA vs NC데이터프레임

상대팀	어제결과	최근승률	
NC	승	0.7	
NC	패	0.6	

FAST CAMPUS ONLINE 이경택 강사.

Fast campus

■ 알고리즘

1) 팀데이터 추출(ex: KIA)

2) 상대팀 전적 데이터 추출(ex: 삼성)

3) 딥러닝 모델 생성 - 1)의 데이터를 통한 모델링 2)의 데이터를 통한 모델링



- (10회반복) Bagging 기법 적용

- 4) KIA가 이길 확률(1) + 가중치(2)
- 5) 4)의 결과가 맞다고 가정, 다음 변수(승률, 최근 경기 결과 등) 추정
- 5) 1) ~ 5) 반복

FAST CAMPUS ONLINE 이경택 강사. 시뮬레이션 10회씩 반복



■ 결과

예측

순위	팀	승	무	패	승률
1	삼성	85	0	56	0.602
2	NC	81	2	58	0.581
3	두산	78	0	63	0.553
4	넥센	77	1	63	0.549
5	한화	68	0	73	0.482
5	KIA	68	0	73	0.482
7	롯데	67	0	74	0.475
7	SK	66	2	73	0.475
9	LG	60	1	80	0.429
10	kt	52	0	89	0.3687

실제

순위	팀	승	무	패	승률
1	삼성	85	0	56	0.603
2	NC	82	2	56	0.594
3	넥센	76	1	64	0.543
4	두산	76	0	64	0.543
5	SK	68	2	71	0.489
5	한화	67	0	74	0.475
7	KIA	66	0	73	0.475
7	롯데	65	1	75	0.464
9	LG	62	2	76	0.449
10	KT	51	0	89	0.364



시뮬레이션

Deep Learning + Ensemble Method(Bagging)

최근 분위기 데로 갈 확률 + 징크스에 발목 잡힐 확률



경기 결과 예측

장점

단점

- 앙상블 기법과 시뮬레이션을 통한 안정성 증대
- 두 모형을 사용함으로써 예측력 강화

- 추가적인 변수 고려 필요성
- 긴 시간의 학습과정

FAST CAMPUS ONLINE

이경택 강사.





Part. 07 빅콘테스트

| 2017빅콘테스트 대출연체자 예측 알고리즘 개발

FASTCAMPUS ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택