KBL 2021-2022

플레이오프 승부 예측

22년도 1학기 주제분석

범주형자료분석팀

박지성 박지민 서희나 윤경선 이지윤

INDEX

2주차 분석 흐름

2주차

플레이오프 진출에 **영향**을
파악

경기 결과와 **직접적인 영향**을 주는 요소 (예: 득점)는 고려항목에서 **제외**

플레이오프에 진출

→ 정규시즌에서 타 팀에 비해 많이 승리

= 어떤 요소가 정규 시즌에서의 각 경기 승리로 이어지는 파악

지난 4년간 정규시즌에 대한 경기데이터 / 선수데이터를 활용

4주차

직접 플레이오프 시즌
예측 해보기

반응변수(Y) → 경기데이터에서의 Win/Lose 변수

! 즉, 각 경기의 승패 여부!!

데이터 수집 범위 설정

분석 기간

2017년 기준 규칙 변경

① 언스포츠맨라이크 파울 1개 & 테크니컬 파울 1개
→ 퇴장

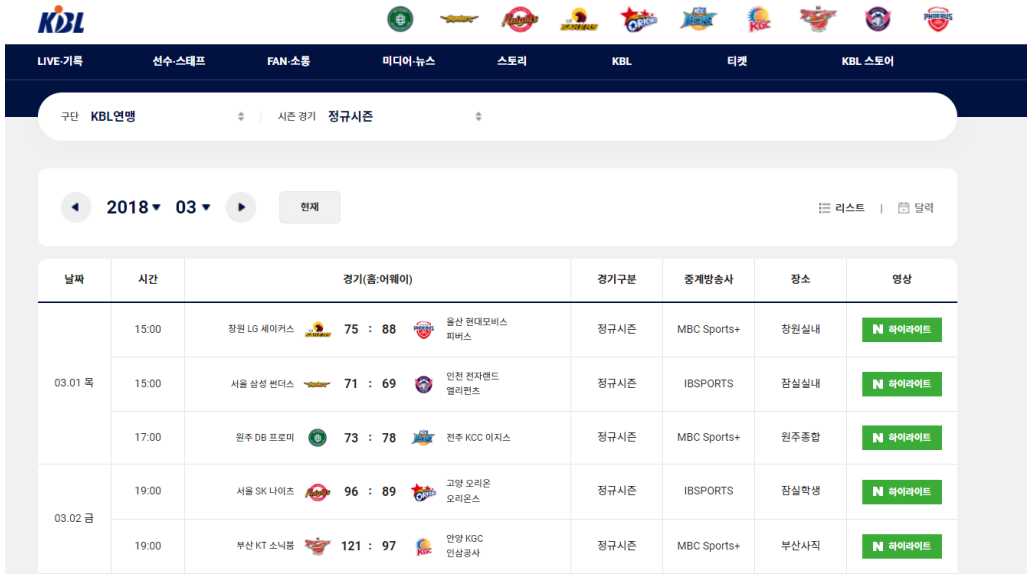
② 벤치 인원이 퇴장 → 감독에게 벤치 파울 부여
→ 상대팀에게 자유투 2개 주어짐

분석 기간 선택

2017-2022년
정규시즌 선수 & 경기 데이터

데이터 수집

KBL 홈페이지에서 2017-2022년 **모든 경기의 경기/선수 데이터 크롤링**



	Date	team	Home/Away	entry	player_num	name	Min	Pts	2PT	3PT	...	DR	TOT	DK	AST	TO	Stl	BS	PF	FO	PP
1	2019.11.01	인천 전자랜드 엘리펀츠	Home		20	정영삼	4:25	2	1/1	0/0	...	0	0	0	0	0	0	0	1	0	1
2	2019.11.01	인천 전자랜드 엘리펀츠	Home		1	민성주	16:34	4	2/3	0/1	...	3	4	0	1	0	2	0	4	0	2
3	2019.11.01	인천 전자랜드 엘리펀츠	Home	*	30	박찬희	28:36	14	2/5	2/4	...	7	7	0	5	3	0	0	3	2	1
4	2019.11.01	인천 전자랜드 엘리펀츠	Home		11	홍경기	0:0	0	0/0	0/0	...	0	0	0	0	0	0	0	0	0	0
5	2019.11.01	인천 전자랜드 엘리펀츠	Home	*	6	자바위	27:30	18	1/2	5/5	...	0	2	0	0	1	0	0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8	2019.11.30	전주 KCC 이지스	Away	*	43	이대성	33:15	24	0/2	7/16	...	1	1	0	3	3	2	0	1	5	0
9	2019.11.30	전주 KCC 이지스	Away	*	2	송교창	35:58	7	2/6	0/3	...	8	9	0	11	1	2	0	4	4	2
10	2019.11.30	전주 KCC 이지스	Away		9	최승욱	5:30	0	0/0	0/0	...	0	0	0	0	0	0	0	0	0	0
11	2019.11.30	전주 KCC 이지스	Away		5	유현준	12:15	0	0/0	0/0	...	1	1	0	0	3	0	0	2	0	0

데이터 전처리

선수 데이터

정규시즌에 **참가 하지 않은 팀** 제거

(D리그 / 이벤트성 경기)

Team	
서울삼성	…
…	라건아드림팀
오리온	오세근매직팀
KCC	이정현드림팀
DB	…
전주KCC	서울 삼성
…	고양오리온

플레이 시간 **초 단위**로 변경

Min
00:00
35:30:00
36:49:00
1:50
…
13:05

→

Min
0
2130
2209
110
…
785

시즌의 독립성 검정

2점 성공률

2점 성공률 수치형 변수 → **범주형**으로 변환

↓

중간값보다 크면 1 작으면 0 부여

↓

시즌(k)을 기준으로 **2점 성공률**(범주)과 **승패**에 관한 분할표 만들기

```
        Breslow-Day test on Homogeneity of Odds Ratios

data:  per_point2
X-squared = 3.1854, df = 3, p-value = 0.3639
```

H_0: Homogeneity of odds ratio vs H_1: not H_0

(p-value <0.05 기각)

각 시즌별로의 승패에 대한 오즈값이 동질적이다

→ 각 시즌에서 승패에 대한

2점 성공률의 영향은 유사!

파생변수 - 기존 경기 데이터 활용

Q1) 왜 비율로 확인해보고자 하였나요?

단위가 득점인 경우 승/패에 대해 직접적인 지표로써 작용된다고 판단

승패 여부에 영향을 주는지 확인해보자!!

Q2) 왜 4가지의 경우로 더 세분화를 하였나요?

각 독립적인 득점 방법을 하나씩 고려해 봄으로써 결과에

강한 요인을 주는 요소가 있는지 파악하고자 함

팀의 총 득점 중 각 득점 비율 확인

자유투 득점	페인트 득점	페인트존 외 2점슛 득점	3점슛 득점

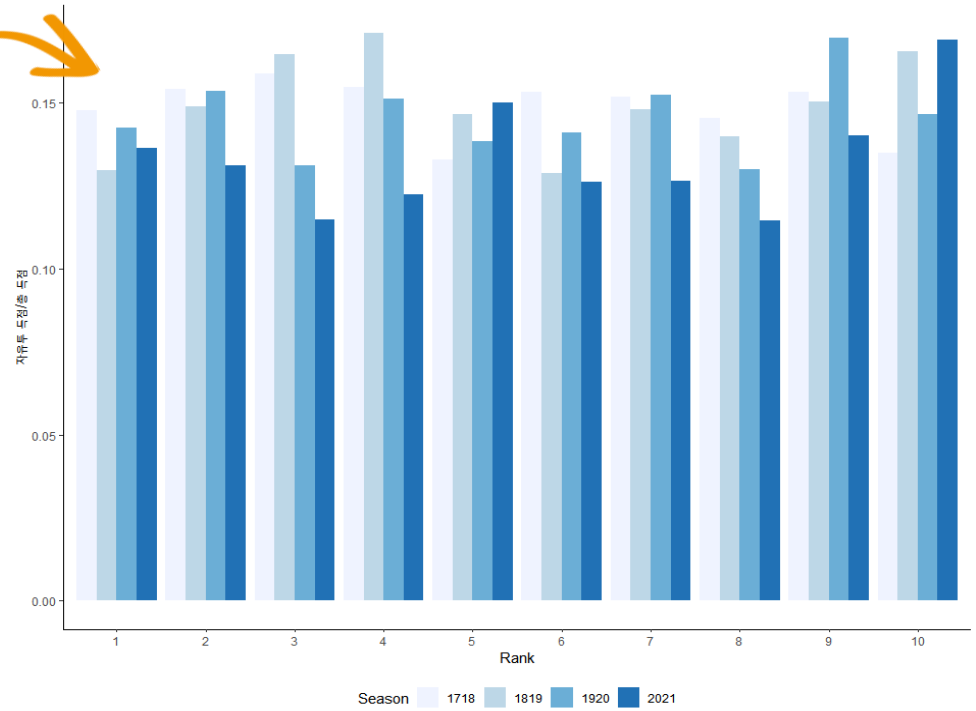
1. 총 득점 중 자유투 득점 비율

자유투 득점 비율

가정: 팀별 자유투 득점 비율이 높을 수록
정규리그 순위가 높을 것

유의미하지 않다고 판단

파생변수로 사용 X



경기시간 가중평균

경기시간 가중평균

선수 데이터 기반으로 여러가지 파생변수 생성

↓

우리가 예측하고자 하는 것은

팀 vs 팀의 **경기 승패 여부**

↓

경기별로 파생변수 할당 필요

Name	Team	Min	파생변수
경선	범주	2400	10
지윤	범주	100	3
희나	범주	500	2

↓

Game	Team	파생변수
1	범주	?

5. KBL Efficiency (가산점 항목)

Efficiency 공식

(득점+스틸+블록슛+수비리바운드)*1.0+(공격리바운드+굿디펜스 +어시스트)*1.5 + 출전시간(분)/4

선수 **개개인**을 위한 스탯

수정된 공식

[(득점+스틸+블록슛+수비리바운드)*1.0+(공격리바운드+어시스트)*1.5]의 **가중평균**

팀의 전체적인 능력을 위한 스탯

5. KBL Efficiency (가산점 항목)의 가중평균

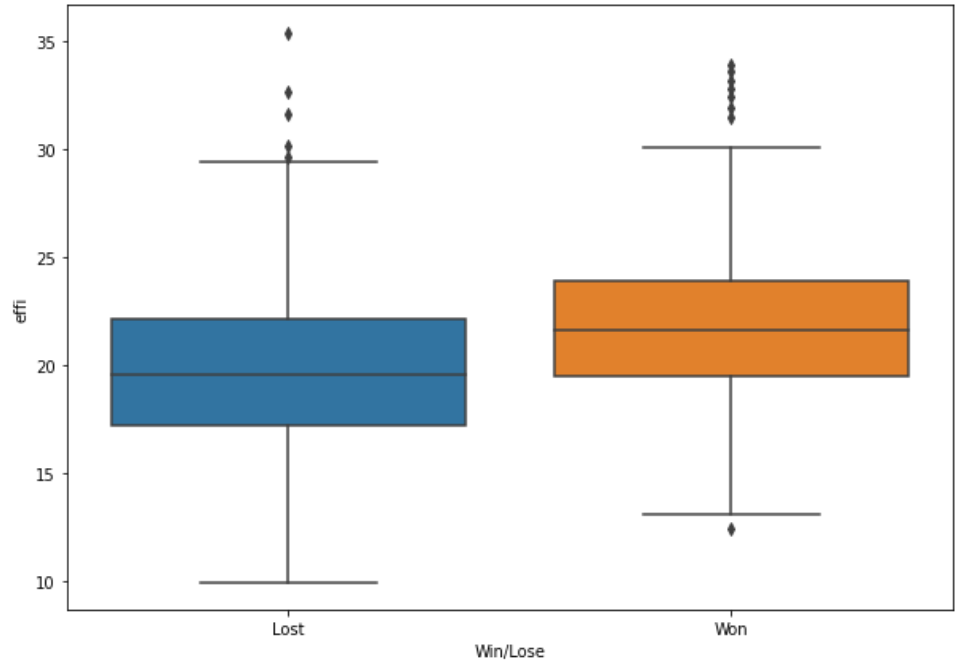
KBL Efficiency (가산점)

가정: 팀의 전체적인 능력이 좋을 것으로 예상

특히, 공격 리바운드가 좋을 것으로 기대

유의미한 차이를 보여

파생 변수로 활용하기로 결정



7. 변수 범주화

속공에 대한 득점

1718, 1819시즌은 (8 , 15)

1920시즌 이후는 (6 , 12) 기준

각각 **'낮음' , '보통', '높음' 으로 범주화**

FB_p
4
...
18
10

FB_p
낮음
...
높음
보통

변수선택 – (1) eFG vs TS

eFG vs TS

Shapiro-Wilk Test를 통해 변수 간 정규성 검정

정규성을 따르지 않는 변수 존재

비모수적 방법 중 하나인

Spearman 상관분석 이용

변수	P-value
TR	1.981..
2P%	0.27..
AST_Rating	0.0029..
...	...

KBL 2021-2022

플레이오프 승부 예측

22년도 1학기 주제분석

범주형자료분석팀

박지성 박지민 서희나 윤경선 이지윤

INDEX

차원 축소 (Dimension Reduction)

설명변수 간 상관관계 확인

변수 선택

분석 목적에 부합하는

소수의 예측 변수만을 선택

변수 간 상관관계 고려 어려움

변수 추출

상관관계가 높은 변수끼리 묶어

변환을 통해 새로운 변수 추출

변수 간 상관관계 고려

데이터의 수를 많이 줄일 수 있음

변수들의 조합들 중 상관관계가 높은 조합의

비율이 얼마나 되는지 확인해보자!

이때, 변수들의 type에 따라 상이한 상관분석 기법 적용

상관관계 – 범주형 vs 범주형

Cramer's V Test

카이제곱 검정 통계량을 이용

Crammer 계수가 0.25미만이면

두 변수간 상관성이

거의 존재하지 않는다고 판단

범주형 변수끼리의 **연관성 측도** 파악

	Home/Away	FB_p	TO_p	SC_p	BC_p	연장여부
Home/Away	0.999	0.020	0.025	0.032	0.028	0.000
FB_p	0.020	1.000	0.181	0.045	0.074	0.018
TO_p	0.025	0.181	1.000	0.058	0.069	0.079
SC_p	0.032	0.045	0.058	1.000	0.057	0.078
BC_p	0.028	0.074	0.069	0.057	1.000	0.073
연장여부	0.000	0.018	0.079	0.078	0.073	0.996

해당 계수의 최대값이 0.181

→ **범주형 변수들 간 상관성이 거의 존재하지 않음**

상관관계 분석

Point biserial correlation coefficient

이분화된 명목형 변수 0과 [illegible]

Pearson 상관계수 계산

최종적으로 전체 설명변수 중

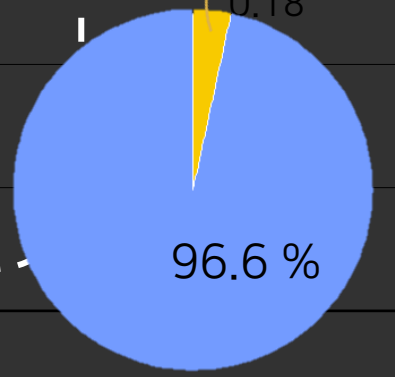
강한 상관관계를 가지는 변수 조합 4% 남짓

변수 선택을 통해 예측을 위한 변수를 구성하고자 함

	Home/Away	연장여부
[illegible]	0.04	0.21
PF	-0.05	0.18
...	...	
4Q득점비	-0.05	

강한 상관관계가 有

조합의 비율 : 3.4 %



변수 선택

부제 : 변수듀스 101

두 가지의 기준에 대해 공통적으로 중요하다고 판단된 변수 최종적으로 선택

① Select K Best 를 통한 타겟 변수와의 상관관계 확인

② KS test , 카이 제곱 검정을 통한 승/패에 따른 이질성 확인

선택된 변수

득점우위시간, TS, eFG, Effi, 최다연속득점,

3P%, 2P%, AST_Rating, TR

추가 데이터 수집

데이터 수집기간

0910 시즌 경기 규칙 변경

① Traveling 규칙 변경

(패스, 정지, 슈팅 전 2 걸음 더 움직일 수 있음 →Foul 기준 달라짐)

② 무조건 5 대 5로 경기

(2008년 12월 30일Portland Trailblazers와 Boston Celtics 경기 참고)

2009-2021년

P/O 선수 & 경기 데이터

추가 데이터 전처리

필드 골 (FGA, FGM)과 3PT로 2PT, 2P% 생성

FGM	FG3M	FGA	FG3A
4	2	13	3
…	…	…	…
8	3	13	4

2PT성공	2PT시도	2P%
2	10	0.2
…	…	…
5	9	0.556

2PT 성공 = 필드골 횟수 (FGM) - 3점슛 횟수 (FG3M)

2PT 시도 = 필드골 시도(FGA) – 3점슛 시도 (FG3A)

2P% = 2PT성공 / 2PT 시도

샘플링 (Sampling)

① 2122 KBL 팀 별 7개 설명변수에 대한 샘플링 진행

Sampling이란?

무작위로 표본을 추출하는 방법

진행방식 (목표: 경기 데이터(X)값 생성)

① 후보 분포 정의

'norm','t', 'f', 'chi', 'cosine', 'alpha', 'beta', 'gamma', 'dgamma', 'dweibull', 'maxwell', 'pareto', 'fisk'

② Fitter 함수를 사용하여 각 설명변수 별로 SSE **최소화**하는 후보 분포 찾기

③ 각 설명 변수에 적합 된 분포에서 필요한 경기 수 만큼 **random하게 하나의 값 선택**

예측 흐름

1

DATA 1718 ~ 2021 KBL 정규시즌

KBL의 농구경기에서 승패에 영향을 주는 중요 **7개 변수** 추출

2

DATA 0910 ~ 2021 NBA P/O

P/O 예측을 위한 모델 형성

NBA와 KBL의 **분포가 다를 수 있음**

본래 모델을 적합하기 위해선 train set과 Test set이 같은 분포에서 유래 해야함

Why?

NBA P/O는 0910~2021 시즌 데이터

각 팀의 이번 시즌 설명변수를 샘플링을 위해 활용한 KBL P/O는 2021 오직 한 시즌,,

예측 흐름

전체적 예측 흐름

시뮬레이터를 이용

대진표를 토대로 N번의 시뮬레이터 결과로

플레이오프 결과 예측

1번의 시뮬레이션마다 주어진 대진표에 따라

6강부터 4강, 결승까지 진출팀 예측

예측 흐름 - 평가지표

평가지표

실제 결승에 진출한 두 팀의

4강 진출 시 결승 진출 조건부 확률의 **조화 평균**

↓

한 쪽 팀의 결승 진출 조건부 확률만 잘 예측하는 모델

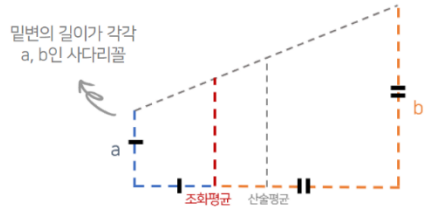
패널티 부여

P(결승 진출 | 4강 진출)

1 혼동행렬

⑥ F1-Score | 조화평균

불균형한 데이터가 주어졌을 때도 보다 정확한 성능 파악



조화평균을 기하학적으로 접근해보면..

밑변의 길이와 동일한 거리에 떨어진 지점에서 빗변으로의 높이가 곧 조화평균!

조화 평균의 자세한 내용은 범주팀 3주차 클린업 참고!

예측 흐름

1718~2021
KBL 정규시즌

경기 승리에 영향을 주는 7개 변수 선택→ 고정 !

0910~2021
NBA PLAYOFF

플레이오프 승/패 모델 형성

분포가 동일하지 않음

2122
KBL 정규시즌

6개 팀에 대한 설명 변수 sampling

강건한 모델 형성

↓

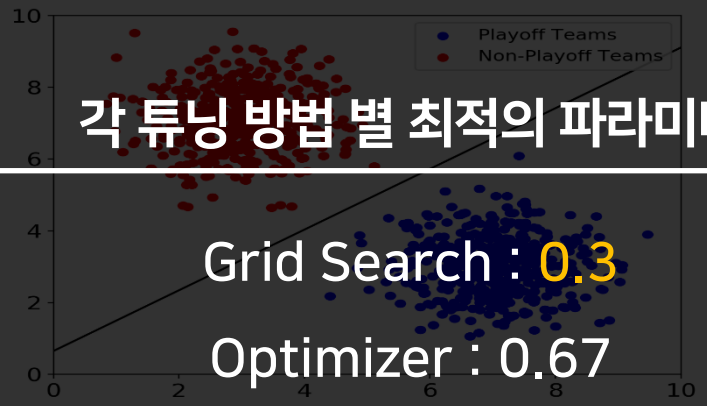
시뮬레이션 → 각 라운드 통과 확률 추출

SVC w. linear kernel

SVC with linear kernel

데이터셋의 구분이 **선형적인 관계**를 보일 경우

최적의 결정 경계로 **선형적인 경계**를 생성



각 튜닝 방법 별 최적의 파라미터

Grid Search : 0.3

Optimizer : 0.67

하이퍼 파라미터 튜닝

C값을 낮게 튜닝하여 이상치를

일부 허용한다면을 통해 Auxiliary Dataset에 대한 문제를 해결 기대

(Why? 선형적인 SVM모델)

C : 경계를 결정할 때 **이상치를 허용하는 정도**

C가 낮은 값을 가짐

정확도 간 차이가 크지 않아 과적합을

이상치를 많이 허용 **과적합** 방지

방지하기 위해 C=0.3으로 채택

SVM(Support Vector Machine) – 예측 결과

SVM _ Linear

우승 확률 부분에서
실제 순위와 유사하게 예측

실제 결과
1등 서울 SK / 2등 안양 KGC

	Team	4강진출진출	결승진출확률	우승 확률
0	안양KGC	0.675	0.418	0.310
1	울산현대모비스	0.653	0.124	0.023
2	서울SK	1.000	0.816	0.328
3	대구한국가스공사	0.325	0.105	0.048
4	고양오리온	0.347	0.060	0.007
5	수원KT	1.000	0.477	0.284

SVM(Support Vector Machine) – 결과 해석

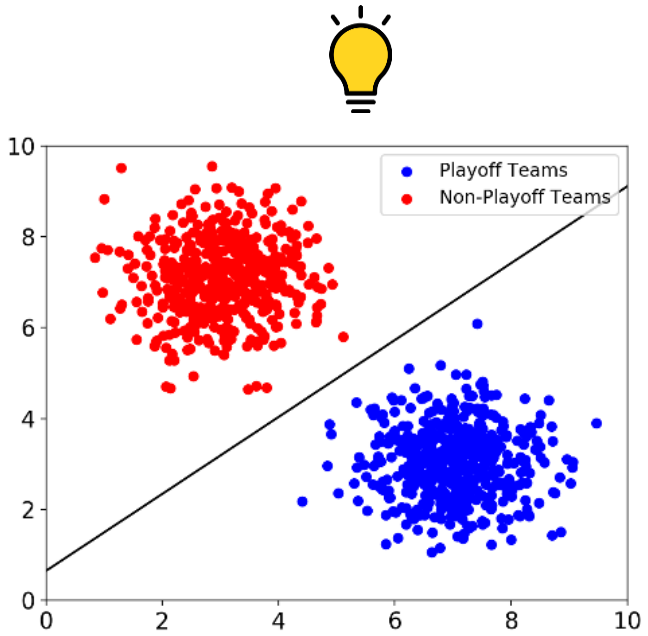
SVM – 결과 해석

반응변수와 설명변수의 관계가 **선형적인 형태**를 띔

Ex) 3P%가 높을 수록 승리와 가까워짐

SVM에서의 비선형 커널 함수를 사용하는 것보다

선형적인 경계를 생성하는 것이 효과적



시각화 및 해석

SVM Linear (C : 0.3)

안양이 6강전에서 승리한다면 수원과 4강에서 대결

안양은 0과 1 사이에서 증가, 수원은 증가 없음

안양과 수원의 대결은 안양이 승리할 확률 높음

(실제 경기결과와 동일)