KBL 2021-2022

플레이오프 승부 예측

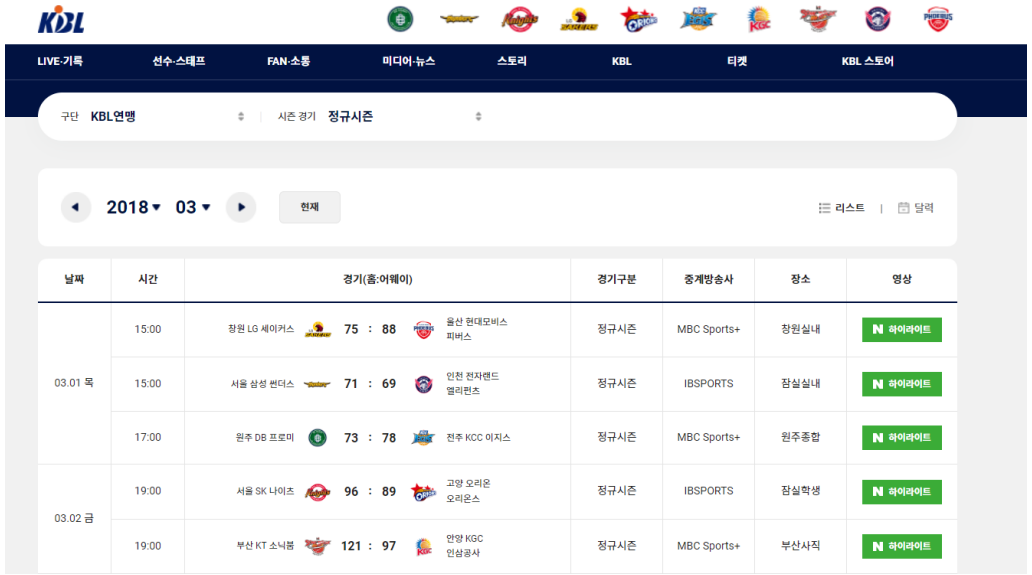
22년도 1학기 주제분석

범주형자료분석팀

박지성 박지민 서희나 윤경선 이지윤

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

후..크롤링..쉽지않아..error..NA..그만..멈춰..

시즌의 독립성 검정

Breslow-Day test

2*2*k 형태의 3차원 분할표의 독립성 검정

✓ 오즈비의 동질성 검정을 위해 고안된 카이제곱 검정법

⬇

즉, K개의 층에 대하여 동질연관성이 있는지 확인해보자

4 연관성 측도

오즈비 (Odds Ratio) | 3차원 분할표

부분분할표에서의 연관성

동질 연관성(homogeneous association)

조건부 오즈비가 모두 같은 값을 가지는 경우 $(\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)})$

- **대칭적** : XY에 동질 연관성 존재 → YZ와 XZ간에도 동질연관성이 존재

동질연관성에 대한 자세한 내용은
범주 1주차 클린업 참고~

시즌의 독립성 검정

Breslow-Day test

2*2*k 형태의 3차원 분할표의 독립성 검정

✓ 오즈비의 동질성 검정을 위해 고안된 카이제곱 검정법

즉, K개의 층에 대하여 동질연관성이 있는지 확인해보자

H_0: Homogeneity of odds ratio

각 k개의 층은 동일한 조건부 오즈비를 지닌다

(p-value <0.05 기각)

시즌의 독립성 검정

파울

2점 파울 수치형 변수 → **범주형**으로 변환

중간값보다 크면 1 작으면 0 부여

시즌(k)을 기준으로 **파울**(범주)과 **승패**에 관한 분할표 만들기

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

data:  foul
X-squared = 1.1448, df = 3, p-value = 0.7663
```

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 파울의 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

자유투 득점	페인트 득점	페인트존 외 2점슛 득점	3점슛 득점

0. 경기시간 가중평균

경기시간 가중평균

적은 경기시간을 뛴 선수의 기록보다

많은 경기 시간을 책임진 선수의 기록이 더 반영되어야 한다고 판단

파생변수의 [illegible]

적은 출전시간(Min)을 가진 선수와

많은 출전시간을 가진 선수가 **같은 조건**으로 반영

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

파생변수에 경기시간을 곱한 합을 총 경기시간으로 나누는 과정을 통해

파생변수의 경기시간 가중평균으로 경기데이터 구성

Game	Team	파생변수
1	범주	$\frac{10+3+2}{3}=5$

차원 축소 (Dimension Reduction)

Feature의 개수가 증가할수록 차원의 저주 발생, 모델의 예측 신뢰도 감소

차원의 저주에 대한 자세한 내용은 데마팀 1주차 클린업 참고~

Method

변수 선택

분석 목적에 부합하는

소수의 예측 변수만을 선택

Ex) Stepwise Selection

변수 추출

상관관계가 높은 변수끼리 묶어

변환을 통해 새로운 변수 추출

Ex) 주성분 분석, 요인분석

상관관계 – 범주형 vs 범주형

Cramer's V Test

카이제곱 검정 통계량을 이용

↓

Crammer 계수 계산

↓

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

'연장여부'와 'FB_p'에 대한 크래머 계수 값

상관관계 – 순서형 vs 연속형

순서형 변수 vs 연속형 변수

순서형 변수를 Ordinal encoding

앞선 두 연속형 변수의
상관관계 파악을 위해 활용한
Spearman 상관분석 진행!

TO_p
높음
…
낮음
보통

TO_p
3
…
1
2

상관관계 분석 – 명목형 vs 연속형

Point Biserial correlation coefficient

이분화된 명목형 변수 0과 1로 Encoding

↓

Encoding된 명목형 변수와
수치형 변수의 Pearson 상관계수로
상관관계 파악

명목형

	Home/Away	연장여부
TR	0.04	0.21
PF	-0.05	0.18
...	...	...
4Q득점비	-0.05	-0.05

연속형

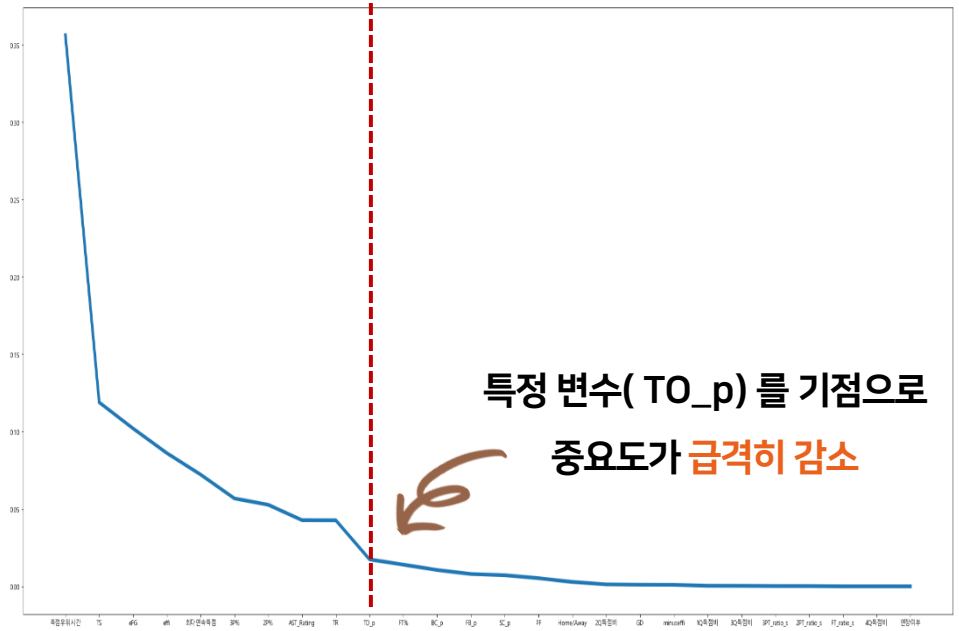
변수 선택 기준 (1) – Select K Best

모든 변수의 Feature Importance 파악

Select K best에서 K=27개로 설정하여 적합한 후 score점수 비교

9개

Feature	Score
득점우위시간	694.61
TS	231.86
eFG	199.00
…	…
TR	83.27
TO_p	**33.89**



변수 선택 기준 (2) – KS 검정

Kolmogorov Smirnov 검정

주어진 **두 표본의 분포가 일치**하는지의 여부를 검정할 때 이용

H_0 : 비교하는 두 분포가 동질적이다

P – value <0.05 → 비교하는 분포 간 이질성 존재

만약 요인 별로 승/패에 대한 두 분포가 **이질적임**을 확인

→ 승/패를 결정짓는 중요한 요인으로 작용함을 알 수 있음

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 , 카이제곱 검정을 통한 승/패에 따른 이질성 확인

선택된 변수

득점우위시간, TS, eFG, Effi, 최다연속득점,

3P%, 2P%, AST_Rating, TR

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

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

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 - 시뮬레이션

시뮬레이션

팀/라운드 별 전체 승리 대비 팀 승리 지표

각 **팀 별 라운드 통과 확률** 계산 가능!

	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

예측 흐름 - 평가지표

평가지표

실제 결승에 진출한 두 팀의

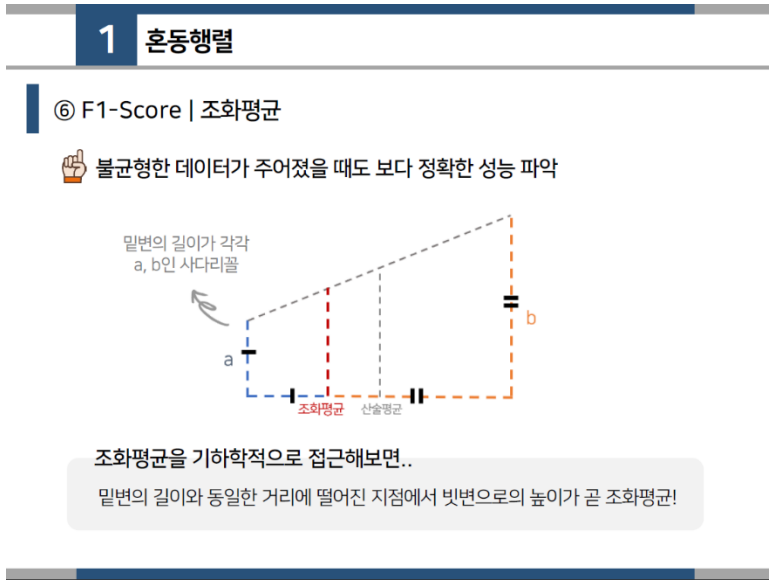
4강 진출 시 결승 진출 조건부 확률의 **조화 평균**

↓

한 쪽 팀의 결승 진출 조건부 확률만 잘 예측하는 모델

패널티 부여

P(결승 진출 | 4강 진출)



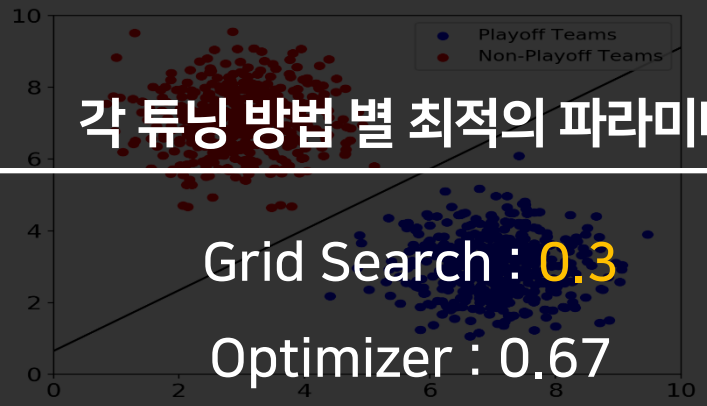
조화 평균의 자세한 내용은 범주팀 3주차 클린업 참고!

SVC w. linear kernel

SVC with linear kernel

데이터셋의 구분이 **선형적인 관계**를 보이므로

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

SVC(Support Vector Classifier) – 결과 해석

SVC Linear (C : 0.3)

2122KBL 정규리그 1,2위 서울, 수원

2122KBL PO 결승 진출팀 서울, 안양

↓

전반적으로 해당 세 팀의 승리지분 높음

실제 경기와 유사하게 잘 예측