



2022 KUBIG 머신러닝 분반 내 프로젝트

16기 유우혁 16기 이수찬 16기 하예은

NBA 선수 경력 분류 예측

Will NBA Rookies' Careers last for 5 years or not?





<분석 목표>

1년 차 NBA 선수들이
5년 이상 활약할 여부를 예측



CONTENT



01 EDA

02 데이터 전처리

03 모델 학습

04 모델 수정

05 한계 및 의의



01 EDA

(1) 데이터 구조 파악

RangeIndex: 1340 entries, 0 to 1339

Data columns (total 22 columns):

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1340 non-null	int64
1	name	1340 non-null	object
2	gp	1340 non-null	int64
3	min	1340 non-null	float64
4	pts	1340 non-null	float64
5	fgm	1340 non-null	float64
6	fga	1340 non-null	float64
7	fg	1340 non-null	float64
8	3p_made	1340 non-null	float64
9	3pa	1340 non-null	float64
10	3p	1340 non-null	float64
11	ftm	1340 non-null	float64
12	fta	1340 non-null	float64
13	ft	1340 non-null	float64
14	oreb	1340 non-null	float64
15	dreb	1340 non-null	float64
16	reb	1340 non-null	float64
17	ast	1340 non-null	float64
18	stl	1340 non-null	float64
19	blk	1340 non-null	float64
20	tov	1340 non-null	float64
21	target_5yrs	1340 non-null	int64

dtypes: float64(18), int64(3), object(1)

```
df.isna().sum()
```

```
Unnamed: 0    0
name          0
gp            0
min           0
pts           0
fgm           0
fga           0
fg            0
3p_made       0
3pa           0
3p            0
ftm           0
fta           0
ft            0
oreb          0
dreb          0
reb           0
ast           0
stl           0
blk           0
tov           0
target_5yrs   0
dtype: int64
```

Feature : 범주형 1개, 수치형 20개

Target : 범주형(target_5yrs)

Feature & Target : 1340 non-null

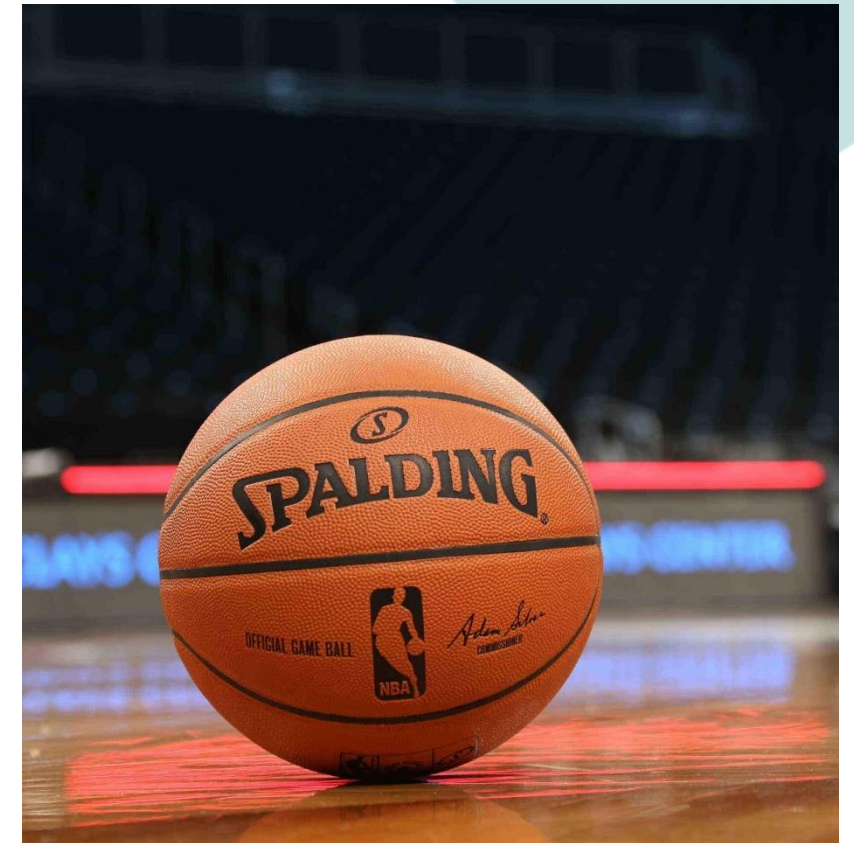
(2) 변수 설명



평균 스탯 vs 비율 스탯



공격 스탯 & 수비 스탯



출전 시간

(2) 변수 설명



득점(pts)

리바운드(reb)

어시스트(ast)

스틸(stl)

블락(blk)

→ 평균 스탯 ←

: 출전 시간에 비례해 증가

fg(야투 성공률)

3p(3점슛 성공률)

ft(자유투 성공률)

→ 비율 스탯 ←

name	gp	min	pts	fgm	fga	fg	
Michael Jordan*	82	38.3	28.2	10.2	19.8	51.5	
name	gp	min	pts	fgm	fga	fg	3p
Jelani McCoy	26	12.7	5.1	2.2	2.9	73.7	

평균 스탯 vs 비율 스탯

종합적으로 변수 고려!

(2) 변수 설명

공격

득점(pts)
어시스트(ast)



공격 스탯 & 수비 스탯

수비

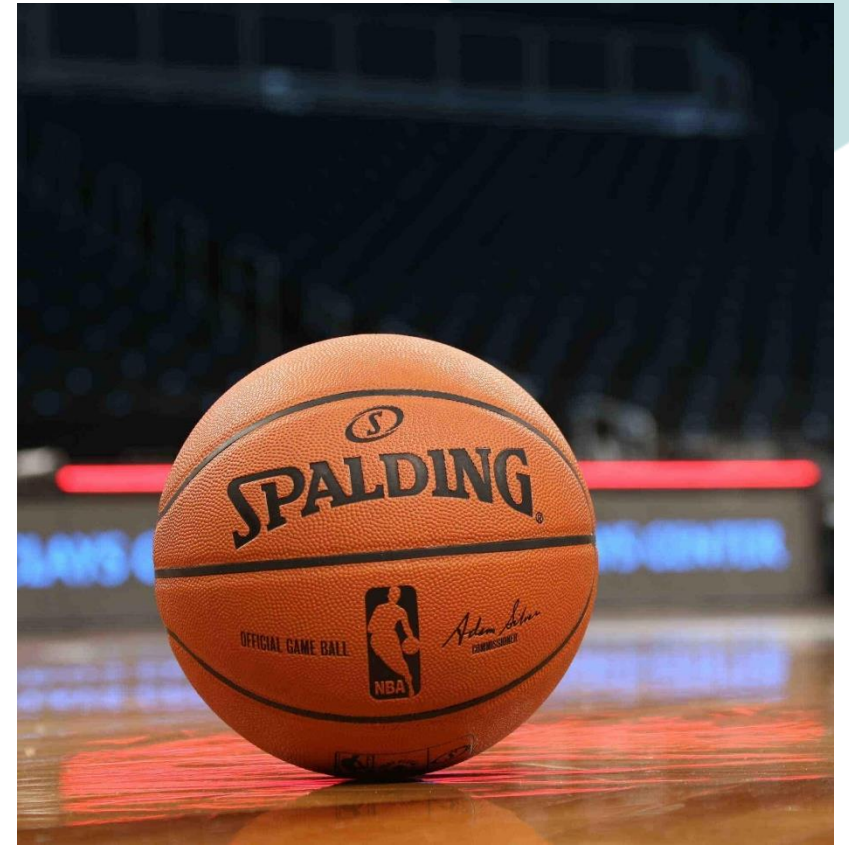
스틸(stl)
블락(blk)

(2) 변수 설명

출전 시간(min) → 팀 내 비중

출전 경기 수(gp) → 꾸준함

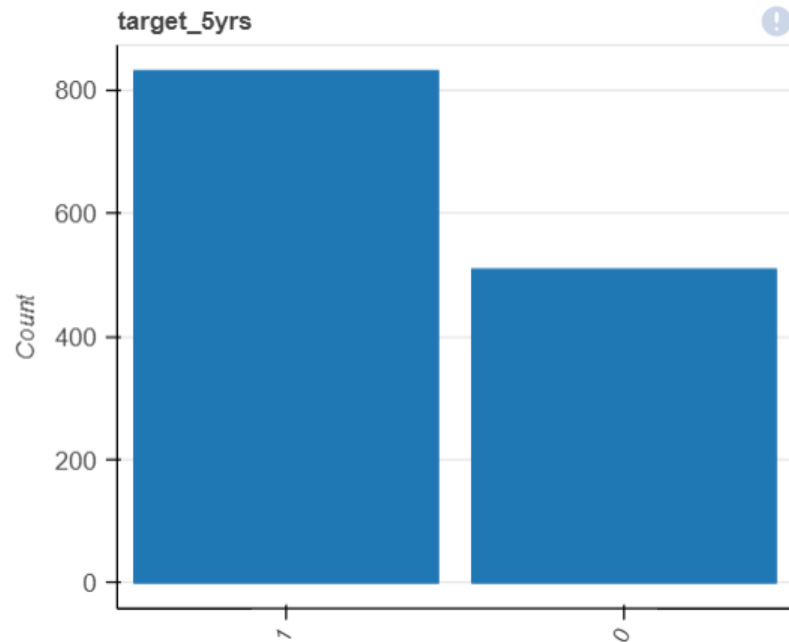
팀 내에서의 인정, 부상 정도 파악!



출전 시간

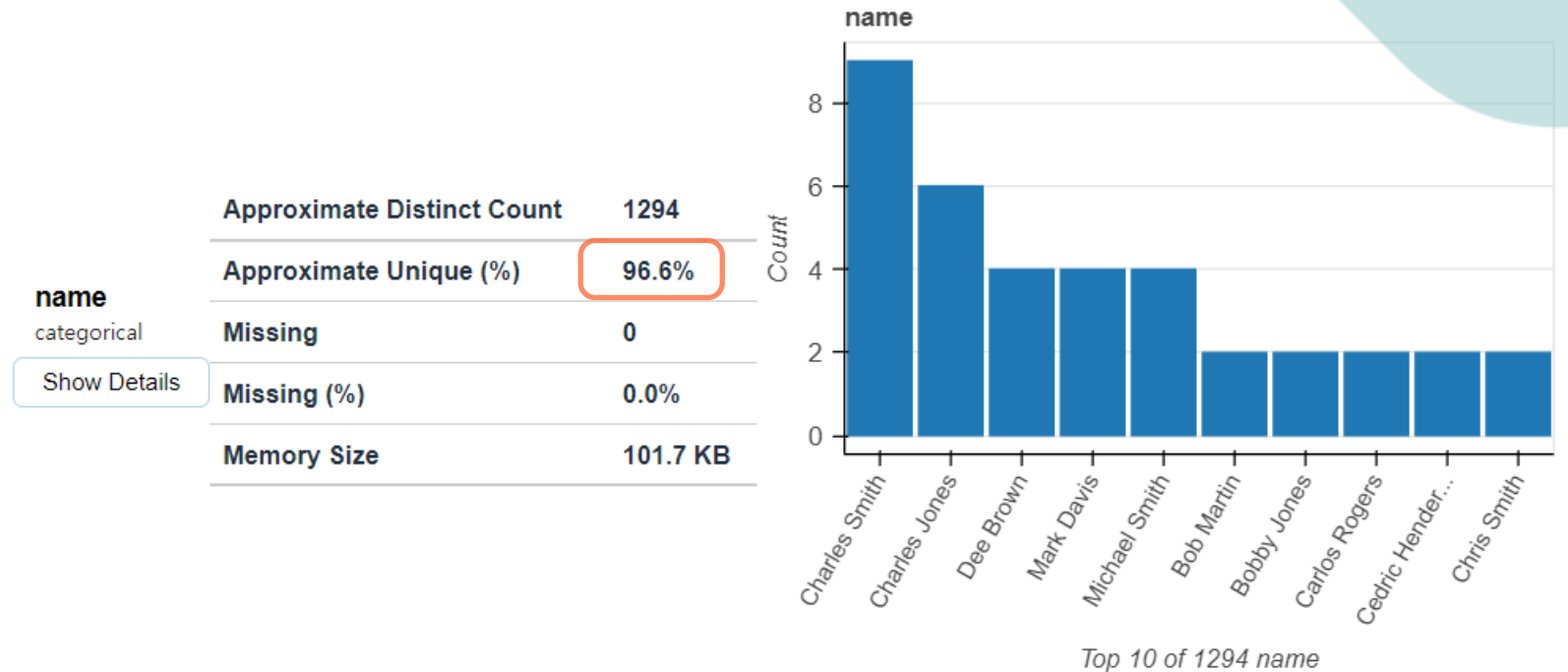
(3) 데이터 시각화

target



0 : career duration < 5 years
1 : career duration > 5 years

name

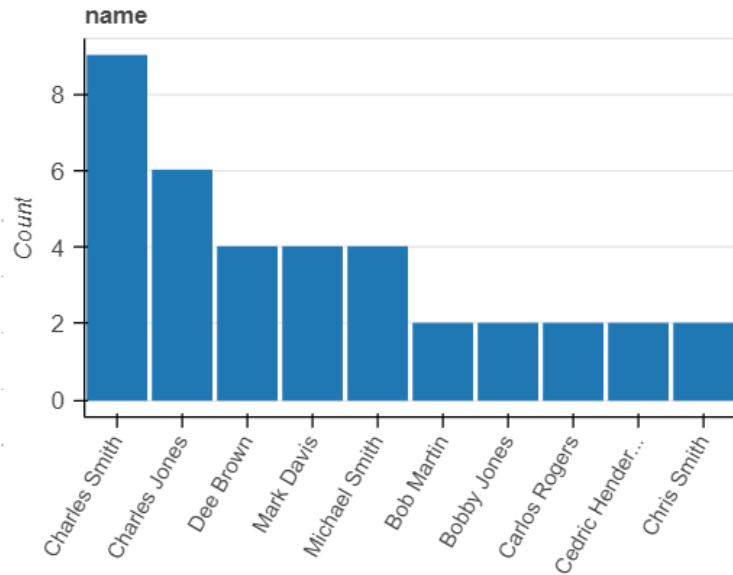


→ 같은 선수에 대해서 중복 입력된 값 존재



02 데이터 전처리

(1) 행 제거



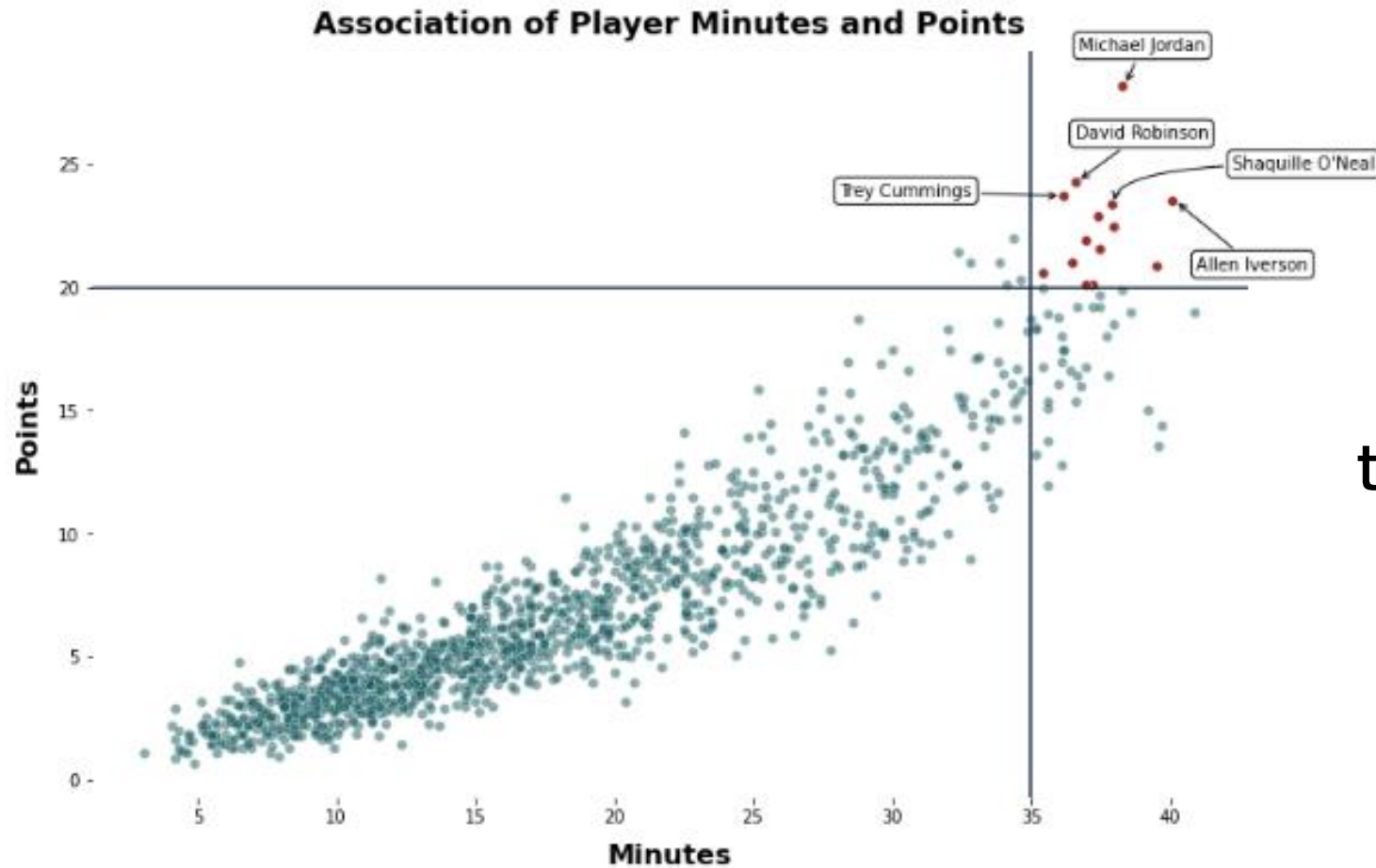
같은 선수에 대해서 target 값만 다르게 중복 입력된 값 존재
→ 실제 기록과 대조하여 **중복 입력된 33개의 행 제거**

(2) Train data / Test data 분리

(3) Feature Selection

‘name’ 변수 & index를 나타내는 ‘Unnamed: 0’ 변수 **drop**

(4) 이상치 제거



출처 : <https://www.kaggle.com/code/paytonfisher/nba-players-notebook>

! 이상치 !
제거해야 할 데이터 X
target을 확실히 예측할 수 있는 데이터 O
→ 이상치 제거 X

(5) 파생 변수 생성

: 포지션 별 데이터의 차이를 줄이고 평가를 용이하게 하는 변수 생성

total_min : gp(출전 경기) * min(출전 시간)

→ 꾸준함 & 팀 내 비중 표현

ts(유효 슈팅 성공률)

→ 3점슛에 가중치를 가하여 슛 효율성 표현

efg(3점 슛 보정 슈팅 효율성 수치)

efficiency(선수 공헌도)

→ 포지션 상관없이 효율성 계산

ftr(야투 시도 대비 자유투 시도 비율)

hollinger(그 선수의 손에서 파생된 공격 중 어시스트가 차지하는 비율)

tovp(개별 선수에게 파생된 공격 중 턴오버가 차지하는 비중)

volume_tov(실책 빈도)

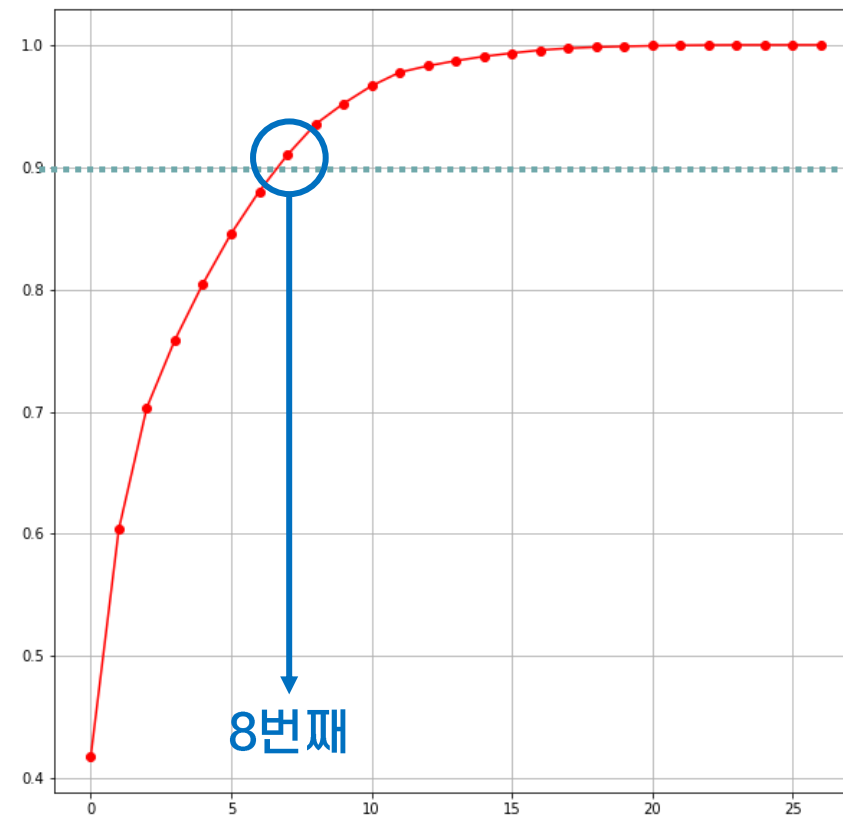
(6) 표준화

: 단위 차이로 발생하는 문제 방지

(7) PCA

VIF Factor	features	82.174635	min
11587.465187	pts	78.216224	fg
6939.014331	fgm	76.222400	ft
5197.437477	reb	65.297962	3pa
2327.020205	dreb	30.303054	tov
634.180252	oreb	23.697590	gp
604.471096	ftm	11.554527	stl
312.954537	fga	<u>11.109736</u>	ast
174.761312	fta	4.361659	3p
86.062842	3p_made	3.778004	blk

다중공선성 발견



8개의 주성분이 전체 분산의 90% 이상 설명

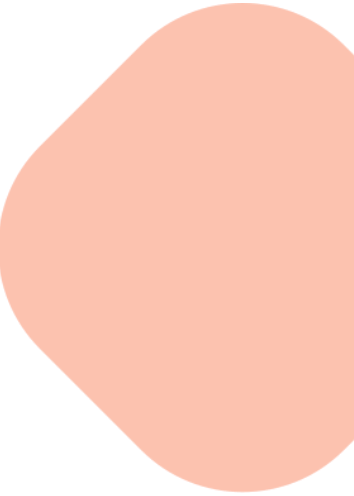
→ 주성분 8개로 결정
→ 다중공선성 해결



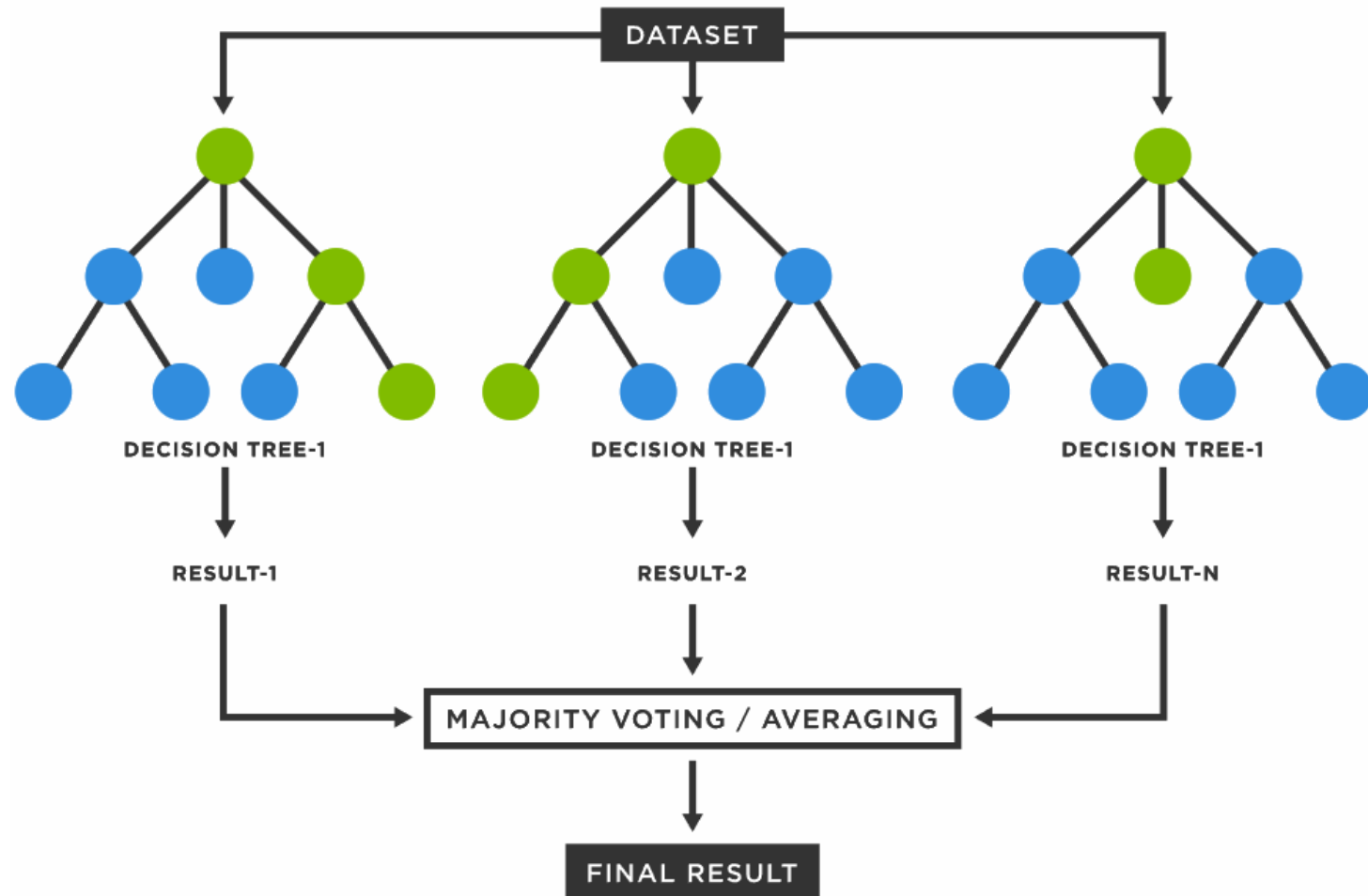
03 모델 학습

(1) 모델 종류

1. 로지스틱 회귀
2. 랜덤 포레스트
3. 엑스트라 트리
4. KNN(K-Nearest Neighbor)
5. LGBM(Light Gradient Boosting Machine)
6. LDA(Linear Discriminant Analysis)

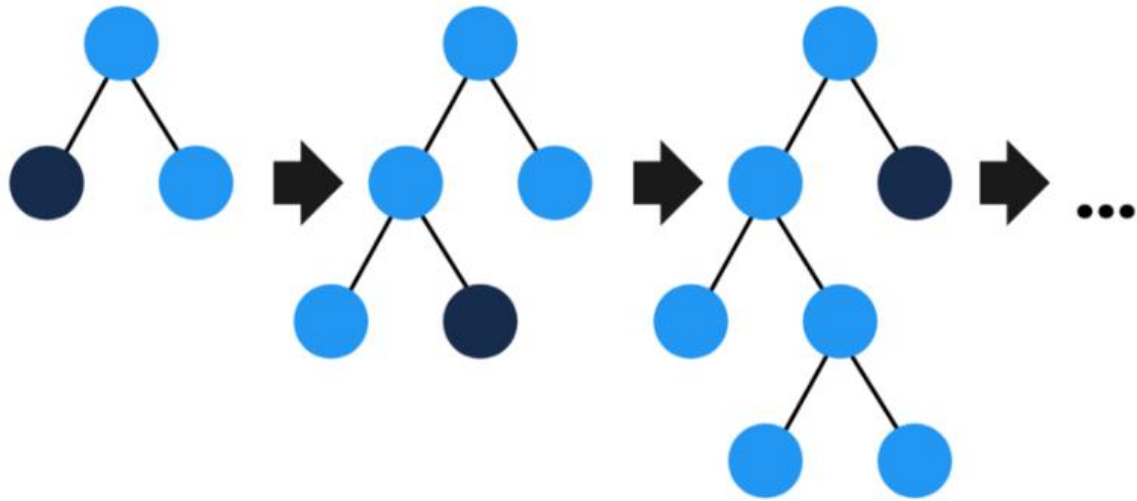


(1) 모델 종류



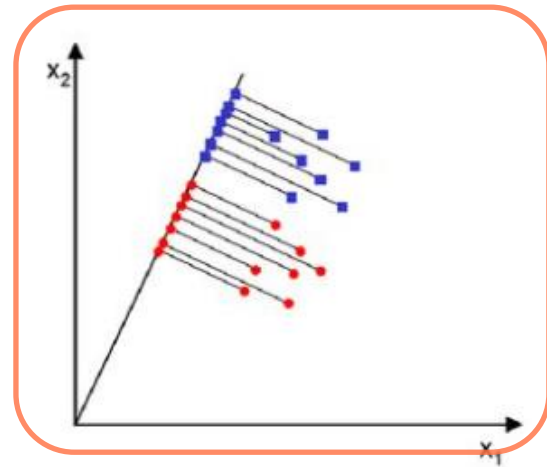
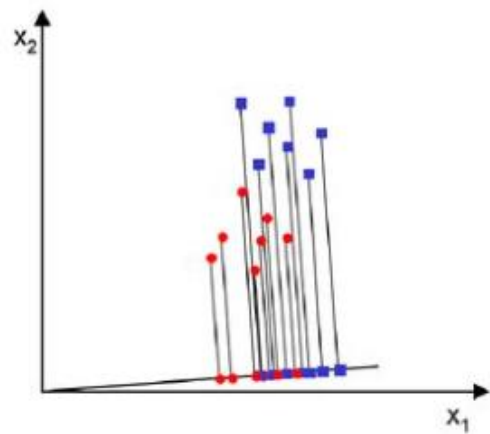
랜덤 포레스트 : 복원추출
엑스트라 트리 : 비복원추출

(1) 모델 종류



이미지 출처 : <https://morioh.com/p/5744808a7324>

LGBM : 틀린 것에 가중치를 뒤서 학습



이미지 출처 : <https://slidesplayer.org/slide/16218884/>

LDA : 클래스 간 분산 최대화 & 클래스 내부 분산 최소화
→ 차원 축소

(2) 평가 지표

	실제 클래스	예측 클래스
TP (True Positive)	Positive	Positive
TN (True Negative)	Negative	Negative
FP (False Positive)	Negative	Positive
FN (False Negative)	Positive	Negative

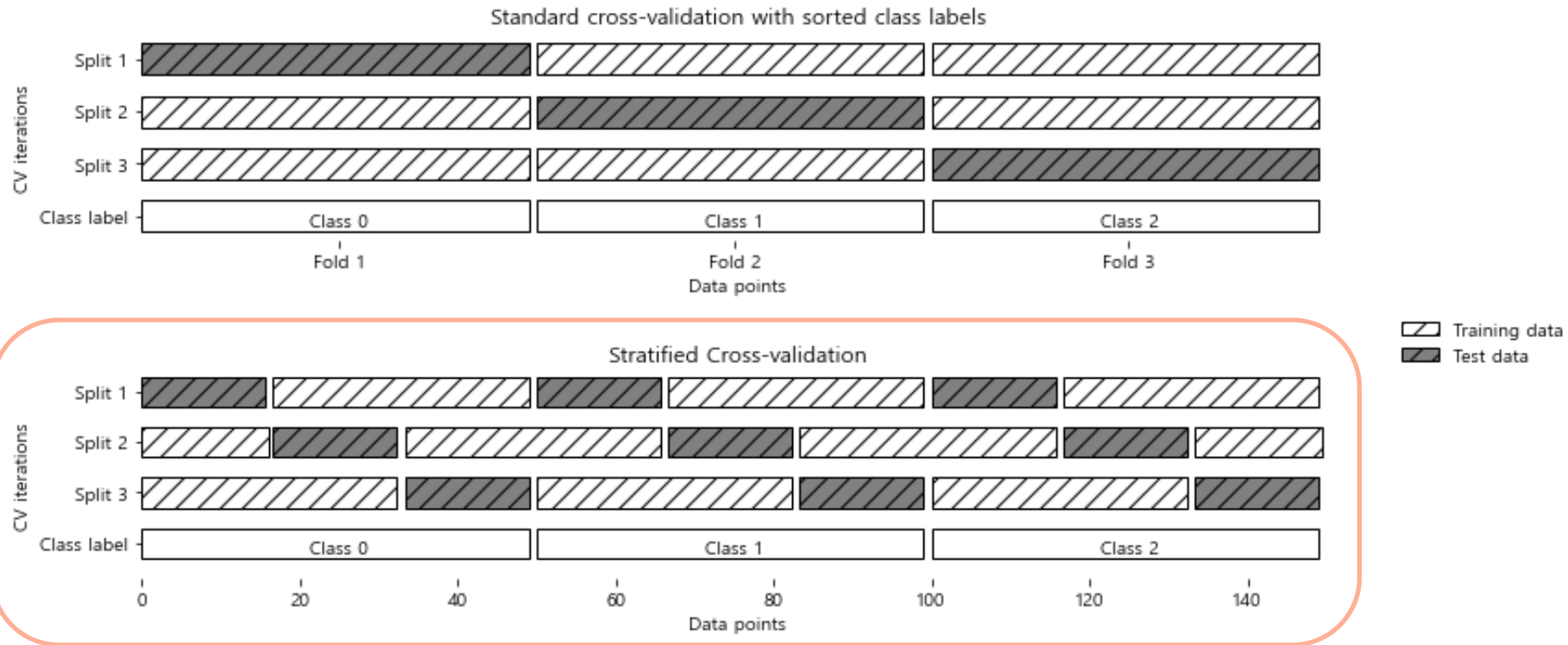
정확도(accuracy) : 실제 데이터 중 맞게 예측한 데이터의 비율 $\frac{TP + TN}{TP + TN + FP + FN}$

정밀도(precision) : Positive로 예측한 데이터 중 실제로 Positive인 데이터의 비율 $\frac{TP}{TP + FP}$

재현율(recall) : 실제 Positive인 데이터 중 Positive로 예측한 데이터의 비율 $\frac{TP}{TP + FN}$

F1 점수 : 정밀도와 재현율의 조화평균 $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

(3) 교차 검증

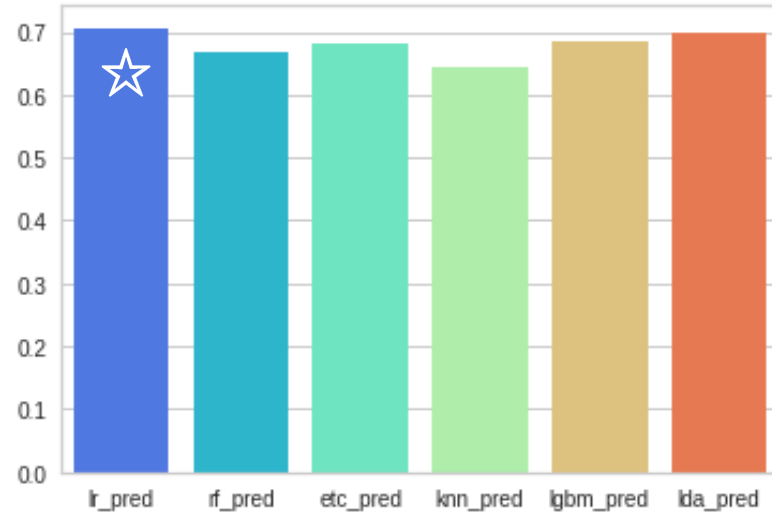


계층별 k-겹 교차 검증(Stratified k-fold cross validation)

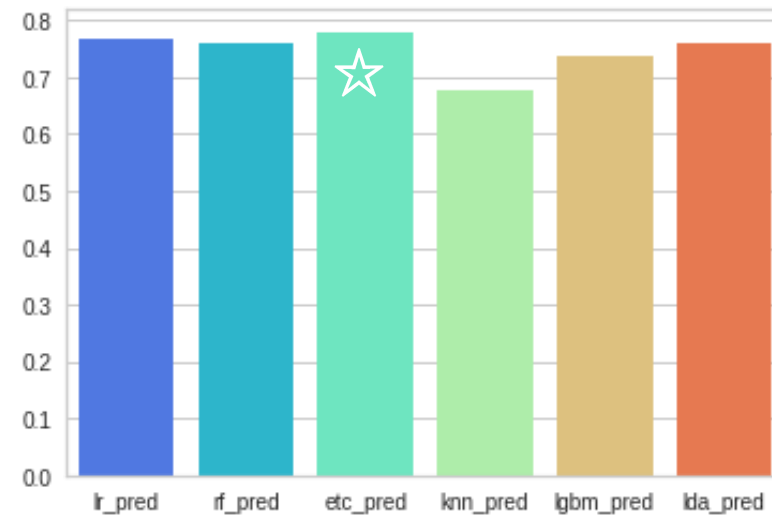
: 클래스 비율이 전체 데이터셋의 클래스 비율과 같도록 데이터를 나눔

(4) 모델 평가

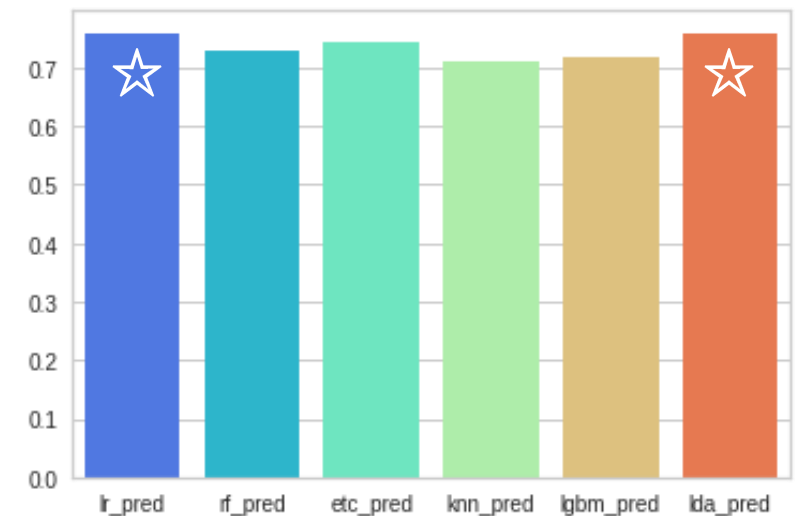
<Accuracy>



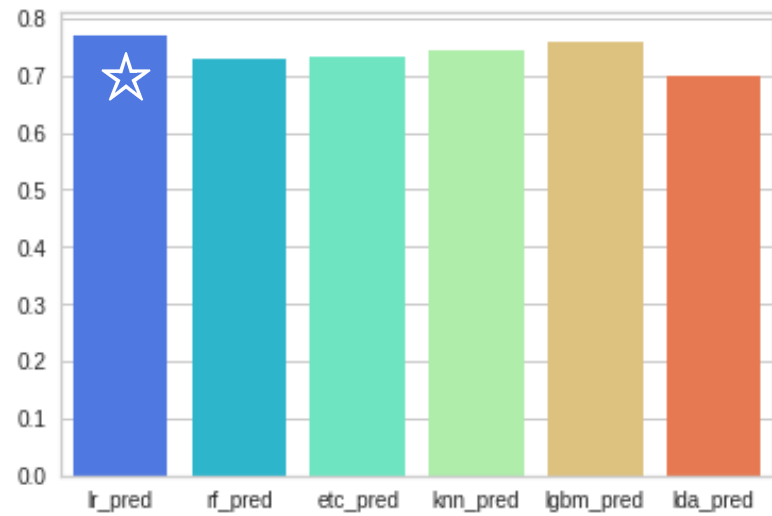
<Recall>



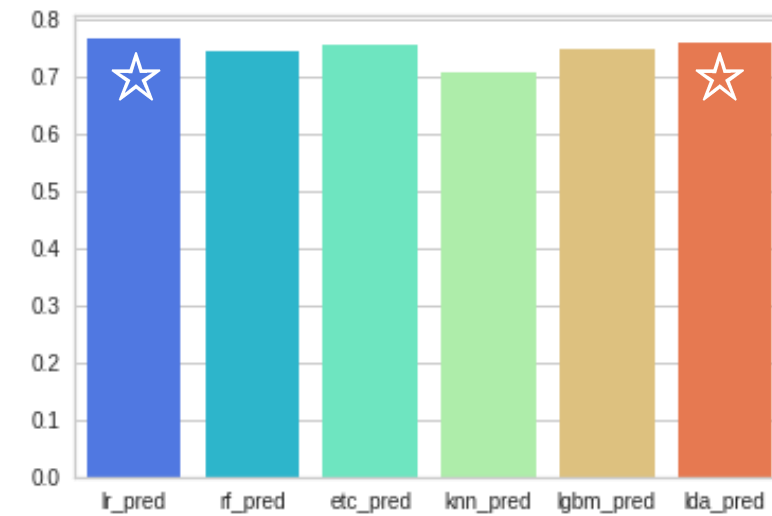
<AUC>



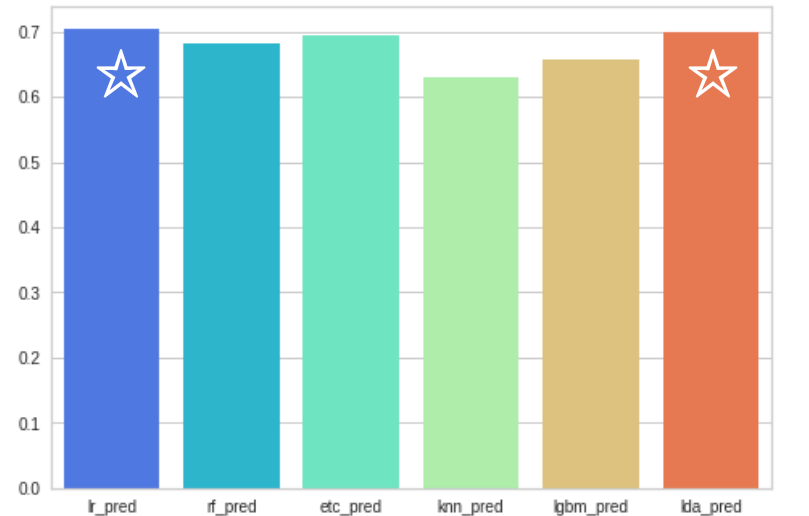
<Precision>



<F1>

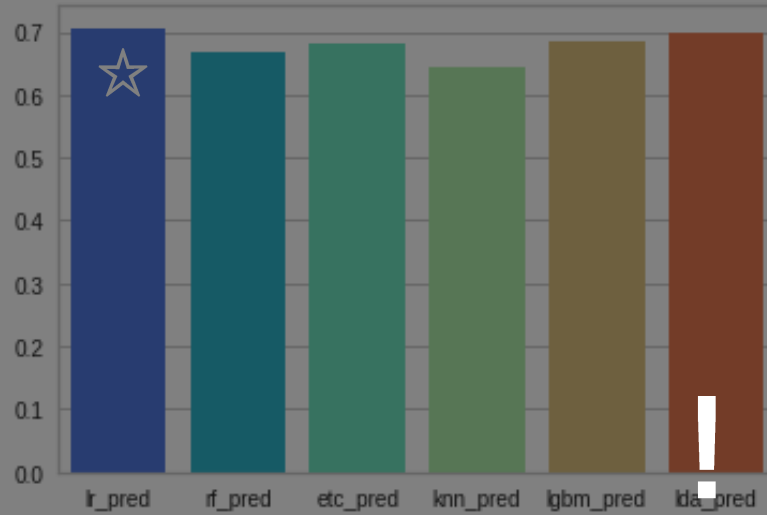


<Cross Validation Score>



(4) 모델 평가

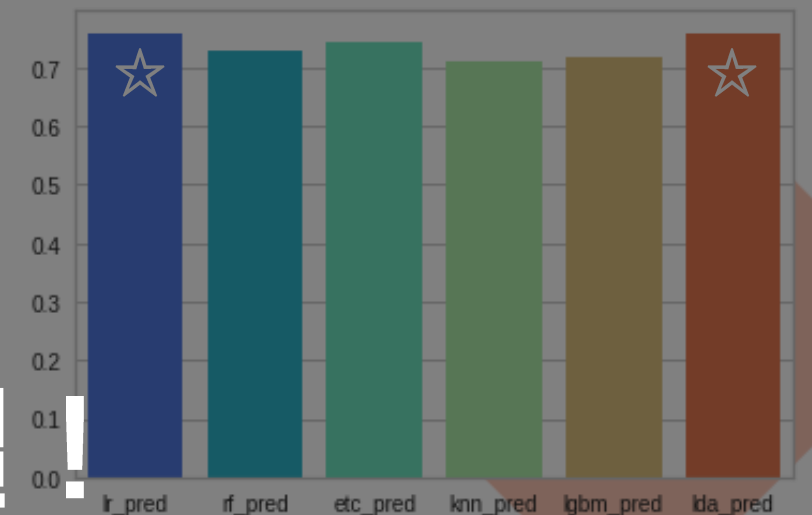
<Accuracy>



<Recall>

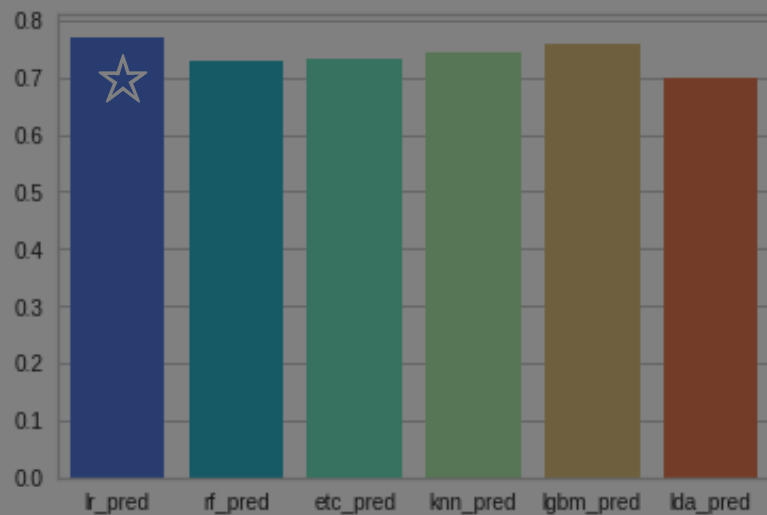


<AUC>

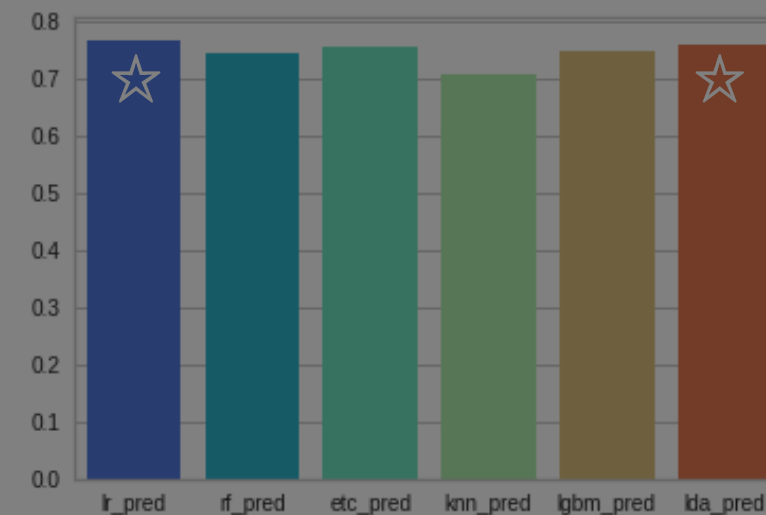


! 로지스틱 회귀 모델 !

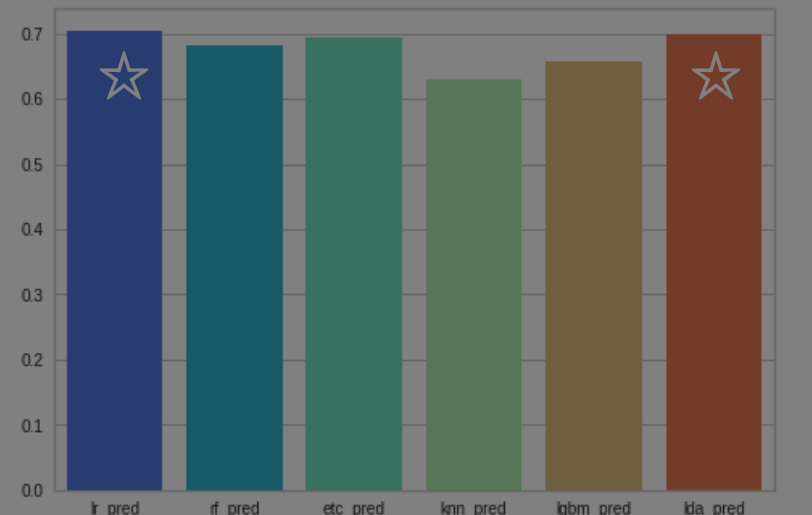
<Precision>



<F1>



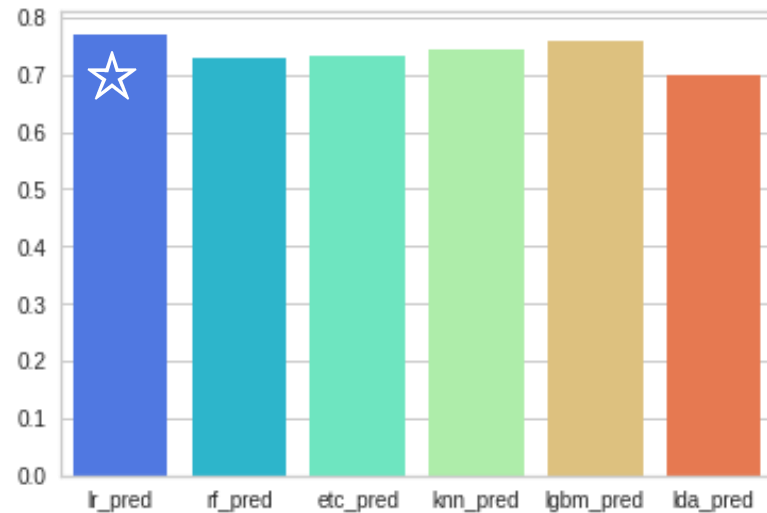
<Cross Validation Score>



(4) 모델 평가

> 자본이 적은 small club <
오래 활동할 것이라고 예측한 선수들이
실제로 오래 활동하는 것이 중요

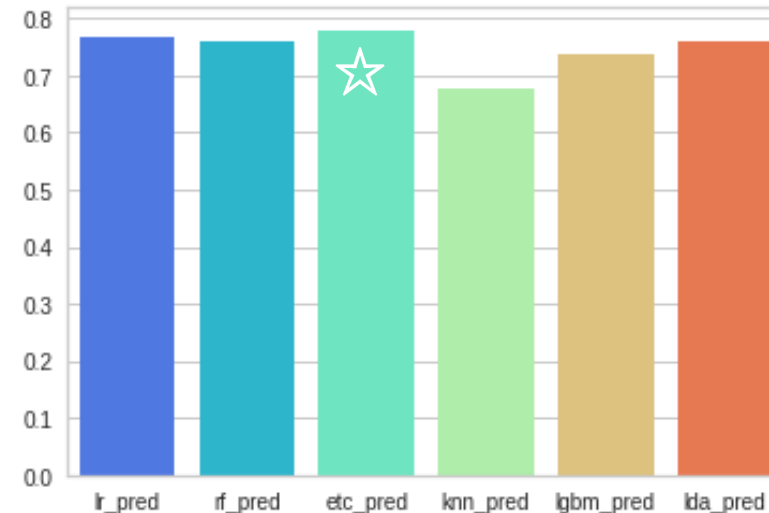
<Precision>



! 로지스틱 회귀 모델 !

> 자본이 많은 big club <
실제로 오래 활동할 선수들을
최대한 많이 예측하는 것이 중요

<Recall>



! 엑스트라 트리 모델 !



04 모델 수정

(1) AutoML

```
best_3 = compare_models(sort = 'Accuracy', n_select = 5, fold = 5, round = 3)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.717	0.000	0.824	0.744	0.781	0.383	0.390	0.018
lda	Linear Discriminant Analysis	0.713	0.775	0.815	0.743	0.777	0.376	0.381	0.024
lr	Logistic Regression	0.709	0.770	0.800	0.746	0.771	0.372	0.376	0.386
ada	Ada Boost Classifier	0.695	0.738	0.771	0.743	0.755	0.349	0.352	0.134
gbc	Gradient Boosting Classifier	0.681	0.735	0.764	0.731	0.745	0.319	0.322	0.228
lightgbm	Light Gradient Boosting Machine	0.681	0.722	0.757	0.734	0.744	0.321	0.323	0.086
et	Extra Trees Classifier	0.674	0.734	0.768	0.721	0.742	0.300	0.305	0.542
rf	Random Forest Classifier	0.669	0.726	0.764	0.718	0.738	0.287	0.291	0.612
qda	Quadratic Discriminant Analysis	0.647	0.734	0.579	0.792	0.668	0.310	0.328	0.022
knn	K Neighbors Classifier	0.643	0.690	0.728	0.704	0.714	0.237	0.239	0.142
nb	Naive Bayes	0.643	0.753	0.517	0.841	0.640	0.324	0.363	0.020
svm	SVM - Linear Kernel	0.640	0.000	0.791	0.703	0.721	0.186	0.199	0.020
dummy	Dummy Classifier	0.614	0.500	1.000	0.614	0.761	0.000	0.000	0.016
dt	Decision Tree Classifier	0.610	0.594	0.666	0.690	0.677	0.185	0.186	0.024

(1) AutoML

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6599	0.0	0.7778	0.7000	0.7368	0.2598	0.2627
1	0.7466	0.0	0.8989	0.7407	0.8122	0.4336	0.4532
2	0.7603	0.0	0.8333	0.7895	0.8108	0.4844	0.4857
3	0.6849	0.0	0.7444	0.7444	0.7444	0.3337	0.3337
4	0.7329	0.0	0.8667	0.7429	0.8000	0.4050	0.4161
Mean	0.7169	0.0	0.8242	0.7435	0.7809	0.3833	0.3903
Std	0.0382	0.0	0.0565	0.0283	0.0332	0.0786	0.0815

Ridge Classifier : 0.7169

0	0.6667	0.7561	0.7778	0.7071	0.7407	0.2770	0.2795
1	0.7123	0.7400	0.8090	0.7423	0.7742	0.3800	0.3827
2	0.7534	0.8502	0.8111	0.7935	0.8022	0.4750	0.4752
3	0.6986	0.7298	0.7333	0.7674	0.7500	0.3712	0.3718
4	0.7397	0.7716	0.8556	0.7549	0.8021	0.4264	0.4336
Mean	0.7142	0.7695	0.7974	0.7530	0.7738	0.3859	0.3885
Std	0.0306	0.0428	0.0405	0.0285	0.0256	0.0659	0.0659

Logistic Regression : 0.7142

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6939	0.7601	0.7778	0.7368	0.7568	0.3447	0.3456
1	0.6986	0.7413	0.8539	0.7103	0.7755	0.3287	0.3419
2	0.7877	0.8605	0.8222	0.8315	0.8268	0.5525	0.5526
3	0.6370	0.6782	0.6556	0.7284	0.6901	0.2550	0.2570
4	0.7260	0.7852	0.8444	0.7451	0.7917	0.3962	0.4029
Mean	0.7086	0.7651	0.7908	0.7504	0.7682	0.3754	0.3800
Std	0.0490	0.0594	0.0725	0.0421	0.0453	0.0994	0.0980

LGBM : 0.7086

0	0.7211	0.7649	0.8111	0.7526	0.7807	0.3990	0.4012
1	0.6849	0.7282	0.8202	0.7087	0.7604	0.3075	0.3146
2	0.7329	0.8169	0.7889	0.7802	0.7845	0.4332	0.4333
3	0.6986	0.7544	0.7000	0.7875	0.7412	0.3835	0.3873
4	0.7123	0.7423	0.8333	0.7353	0.7812	0.3660	0.3722
Mean	0.7100	0.7613	0.7907	0.7529	0.7696	0.3778	0.3817
Std	0.0168	0.0304	0.0476	0.0290	0.0166	0.0416	0.0392

GBC : 0.7100

(2) 앙상블 메서드 : 로지스틱 회귀 모델

1. Bagging

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6803	0.7519	0.8000	0.7129	0.7539	0.3019	0.3060
1	0.7055	0.7353	0.8090	0.7347	0.7701	0.3632	0.3664
2	0.7466	0.8446	0.7889	0.7978	0.7933	0.4659	0.4659
3	0.6918	0.7250	0.7333	0.7586	0.7458	0.3547	0.3551
4	0.7397	0.7738	0.8444	0.7600	0.8000	0.4304	0.4353
Mean	0.7128	0.7661	0.7951	0.7528	0.7726	0.3832	0.3858
Std	0.0262	0.0426	0.0361	0.0284	0.0212	0.0581	0.0575

추출 10회 : 0.7128

2. Boosting

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6667	0.7558	0.7778	0.7071	0.7407	0.2770	0.2795
1	0.6986	0.7394	0.7865	0.7368	0.7609	0.3546	0.3560
2	0.7534	0.8437	0.8000	0.8000	0.8000	0.4786	0.4786
3	0.6918	0.7317	0.7111	0.7711	0.7399	0.3632	0.3650
4	0.7397	0.7710	0.8556	0.7549	0.8021	0.4264	0.4336
Mean	0.7100	0.7683	0.7862	0.7540	0.7687	0.3800	0.3825
Std	0.0320	0.0400	0.0463	0.0313	0.0274	0.0684	0.0685

추출 50회 : 0.7100

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6531	0.7509	0.7778	0.6931	0.7330	0.2425	0.2458
1	0.7397	0.7329	0.8539	0.7525	0.8000	0.4317	0.4388
2	0.7808	0.8667	0.8333	0.8152	0.8242	0.5334	0.5336
3	0.6575	0.7321	0.6778	0.7439	0.7093	0.2948	0.2967
4	0.7260	0.7542	0.8667	0.7358	0.7959	0.3876	0.3998
Mean	0.7114	0.7674	0.8019	0.7481	0.7725	0.3780	0.3829
Std	0.0493	0.0505	0.0691	0.0393	0.0437	0.1024	0.1023

추출 50회 : 0.7114




05 한계 및 의의



한계

target을 예측할 변수가 제한적
데이터 양의 부족
데이터 자체의 결함



의의

평가지표에 따라 모델 선정
기준이 달라질 수 있음을 깨달음
제한된 데이터 속에서 파생변수를
형성함으로써 기존의 데이터로
부족했던 측면을 채울 수 있었음
정확도 향상을 위해 다양한
처리를 해보는 경험을 쌓음



KUBIG 16기 유우혁 이수찬 하예은

들어주셔서
감사합니다