DATA MINING

데이터 마이닝 프로젝트

NBA선수 연봉 분류예측

단국대학교 정보통계학과 최현석, 황성진

목차

DATA MINING

01 주제 및 데이터 소개

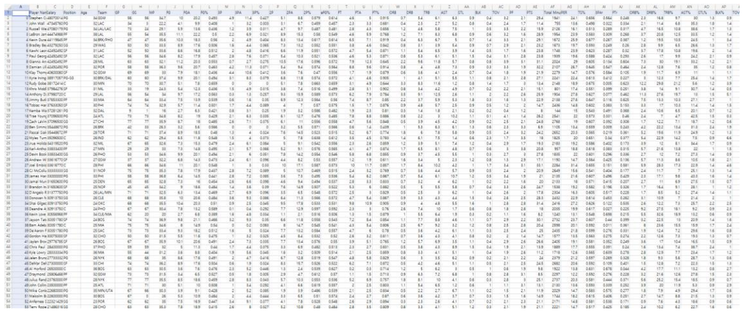
1-1. 주제 및 데이터 소개

주제 배경

NBA

NBA는 샐러리 캡(Salary Cap) 제도로 팀별 연봉 총액을 제한하고 있으며, 이에 따라 선수 연봉은 경기 성과와 구단 전략을 반영하는 지표로 작용한다. 그렇기에 본 프로젝트에서는 NBA 경기 데이터를 활용해 연봉을 예측하고자 한다.

NBA_Player_Salaries(22-23).csv



데이터 소개

데이터이름 : NBA_Player_Salaries(2022-23)
출처 : Welsh, J. Kaggle.
https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season
수집기관 : Basketball-reference(경기기록/통계), HoopsHype(연봉)
샘플크기 : 471개 관측치, 52개 변수
수집기간 : 2022-2023 Season

수집방법

본 프로젝트에서는 Kaggle에서의 제임스 윌슨 사용자가 NBA 경기 기록, 통계, 연봉 등 제공하는 사이트에서 웹 크롤링을 통해 수집한 NBA 선수 연봉 및 통계 데이터를 활용하였다.

1-2. 분석목표

NBA 선수의 연봉 분류 예측 모델 구축

Percent Cap	Cluster
< 2% (Min)	1
2-4% (Vet Min)	2
4-7% (MLE, Bi-annual)	3
7-10%	4
10-13%	5
13-16%	6
16-19%	7
> 19% (Maxes)	8

Table 2 Clustering bounds used to group data, with various exceptions listed

[분류기준]

Percent Cap	구간	분류	선수 의미
<2%	1	C	최소 보장 계약, 2-way, 벤치 선수 등
2-4%	2	B	로테이션~스타팅급 선수
4-7%	3		
7-10%	4		
10-13%	5		
13-16%	6	A	팀의 에이스, 스타, 올스타급 선수
16-19%	7		
>19%	8		

C급 선수 (Low): 13% 이상 = 연봉 $ 16,075,150 달러 이상
B급 선수 (Mid): 13% 미만 = 연봉 $ 16,075,150 달러 미만
A급 선수 (High): 2% 미만 = 연봉 $ 2,473,100 달러 미만

22-23시즌 기준 샐러리 캡(Salary Cap) = $ 123,655,000

PART 1

참고문헌: Wu, William, et al. (2018). *Classification of through Player Statistics*. Sports Analytics Group at Berkeley,

1-3. 변수 소개

[변수] - 선수/성과 지표: 총 51개의 변수

기본 정보		
index	변수명	설명
1	Player.Name	선수 이름
2	Salary	연봉 (달러)
3	Position	포지션
4	Age	나이
5	Team	소속팀

득점 능력		
index	변수명	설명
9	FG	필드골 성공 수
10	FGA	필드골 시도 수
11	FG.	필드골 성공률
12	X3P	3점슛 성공 수
13	X3PA	3점슛 시도 수
14	X3P.	3점슛 성공률
15	X2P	2점슛 성공 수
16	X2PA	2점슛 시도 수
17	X2P.	2점슛 성공률
18	eFG.	유효 필드골 성공률 (eFG%)
19	FT	자유투 성공 수
20	FTA	자유투 시도 수
21	FT.	자유투 성공률
30	PPG	득점

비율/효율 지표		
index	변수명	설명
33	TS.	진정한 슈팅 비율 (TS%)
34	X3PAr	3점슛 시도 비율
35	FTr	자유투 시도 비율
36	ORB.	공격 리바운드 비율
37	DRB.	수비 리바운드 비율
38	TRB.	총 리바운드 비율
39	AST.	어시스트 비율
40	STL.	스틸 비율
41	BLK.	블록 비율
42	TOV.	턴오버 비율
43	USG.	사용률 (USG%)

경기 참여도		
index	변수명	설명
6	GP	경기 수
7	GS	선발 경기 수
8	MP	경기당 평균 출전 시간
31	Total.Minutes	총 출전 시간

리바운드 및 수비		
index	변수명	설명
22	ORB	공격 리바운드
23	DRB	수비 리바운드
24	TRB	총 리바운드
26	STL	스틸
27	BLK	블록
29	PF	파울

고급 지표		
index	변수명	설명
32	PER	선수 효율성 지수 (PER)
44	OWS	공격 Win Shares
45	DWS	수비 Win Shares
46	WS	총 Win Shares
47	WS.48	48분당 WS
48	OBPM	공격 BPM
49	DBPM	수비 BPM
50	BPM	총 BPM
51	VORP	VORP (대체 선수 대비 가치)

플레이메이킹 및 실수		
index	변수명	설명
25	AST	어시스트
28	TOV	턴오버

02 데이터 전처리 및 시각화

2-1. 데이터 전처리

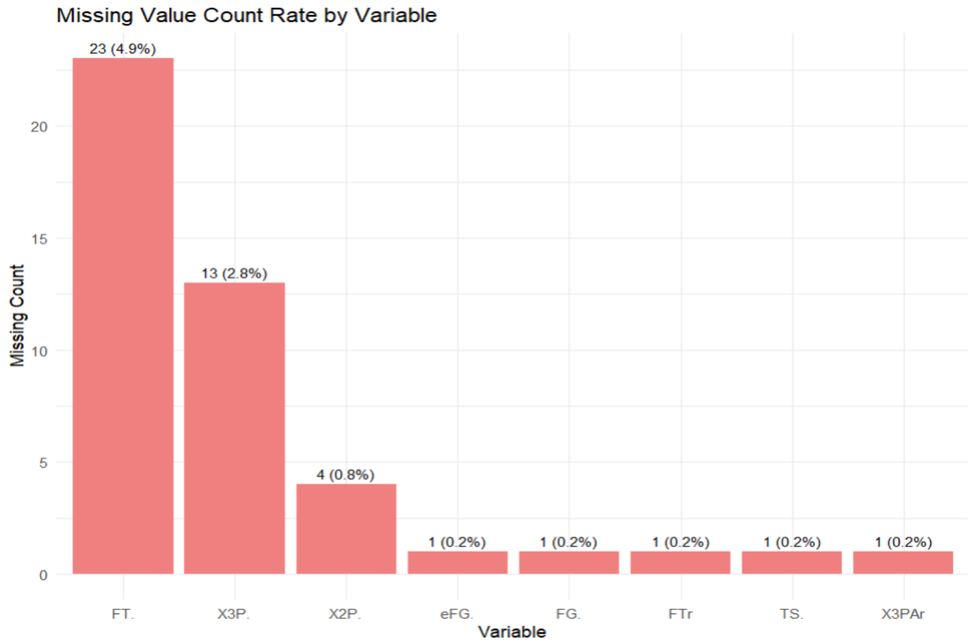
[1. ID 제거]

```
nba <- nba[, -1]
```

[2. 결측치 제거]

```
nba <- na.omit(nba)
```

- 총 34명 선수 제거



[3. 혼합포지션을 한 포지션으로]

```
nba$Position <- sub("-.*", "", nba$Position)
```

2-1. 데이터 전처리

[4. 변수 생성]

1 연봉 구간화 - 샐러리캡 비율 (pct_cap), 연봉 분류 (sal_tier)

샐러리캡 비율 = (연봉/샐러리캡)*100

*샐러리 캡(Salary Cap) = $ 123,655,000

High	샐러리캡 2% 미만	$ 2,473,100 달러 미만
Mid	샐러리캡 13% 미만	$ 16,075,150 달러 미만
Low	샐러리캡13% 이상	$ 16,075,150 달러 이상

```
salary_cap = 123655000
nba$pct_cap <- (nba$Salary / salary_cap) * 100


nba$sal_tier <- cut(nba$pct_cap,
                    breaks = c(0, 2, 13, 100),
                    labels = c("Low", "Mid", "High"),
                    right = FALSE)
```

2-1. 데이터 전처리

[4. 변수 생성]

2 경력 (Years.of.Service)
각 선수의 NBA 데뷔 연도와 분석 시점을 기준으로 계산
경력 = 2023 – NBA Debut

Giannis Antetokounmpo

Pronunciation: \YAHN-iss ah-dedo-KOON-bo\
Giannis Antetokounmpo ▪ Instagram: Giannis_An34
(last name previously spelled Adetokunbo)
(The Greek Freak, The Alphabet)
Position: Power Forward, Small Forward, Point Guard, and Shooting Guard ▪ **Shoots:** Right
6-11, 242lb (211cm, 109kg)
Team: Milwaukee Bucks
Born: December 6, 1994 (Age: 30-189d) in Athens, Greece
Relatives: Brothers Thanasis Antetokounmpo, Kostas Antetokounmpo, Alex Antetokounmpo
Draft: Milwaukee Bucks, 1st round (15th pick, 15th overall), 2013 NBA Draft
NBA Debut: October 30, 2013
Experience: 11 years

*출처: Basketball-reference

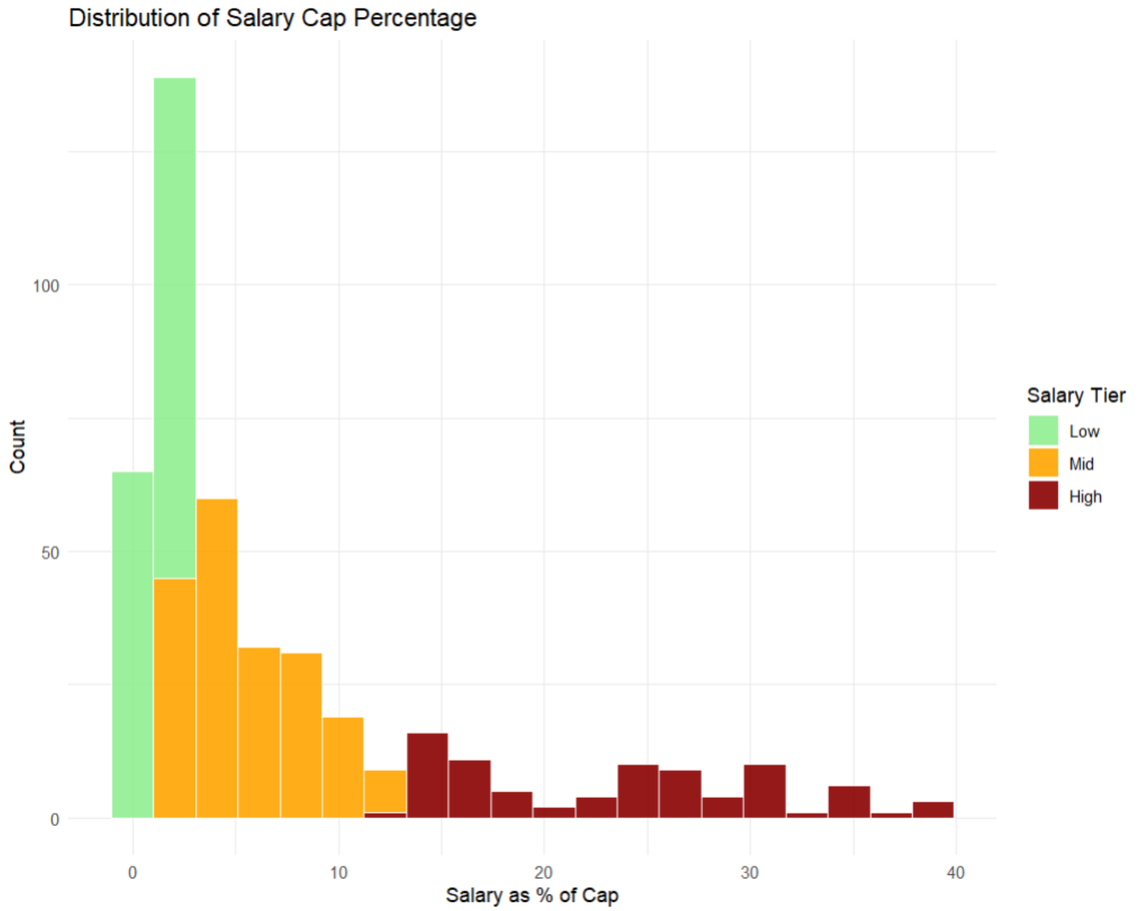
PART 2

2-2. 데이터 시각화

[시각화1]
- 샐러리캡 연봉 분류 히스토그램

```
> table(nba$sal_tier)

 Low  Mid High
 159  195   83
```

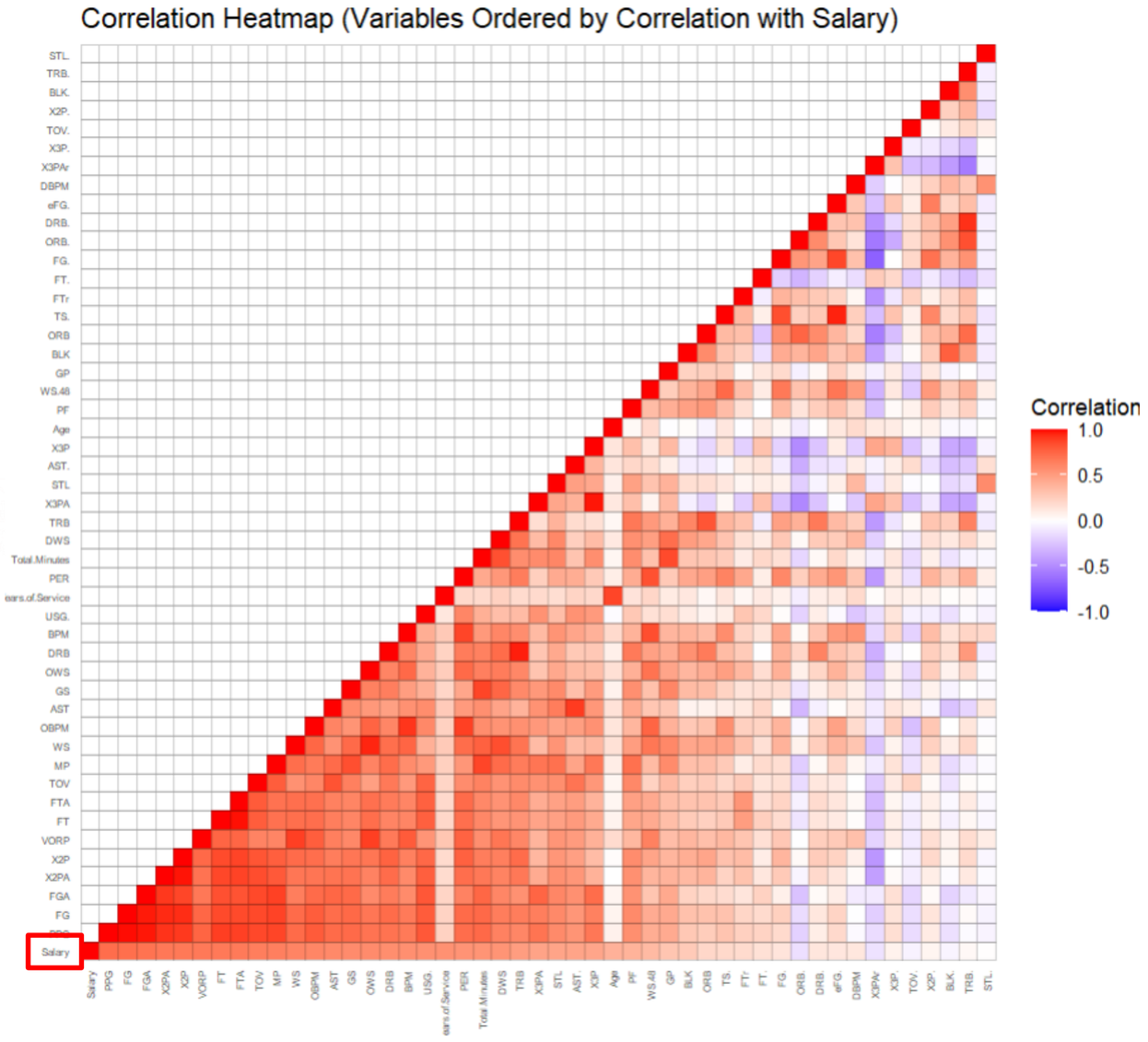


PART 2

2-2. 데이터 시각화

[시각화2]
- 변수별 상관계수 히트맵
(연봉 상관계수 크기 정렬)

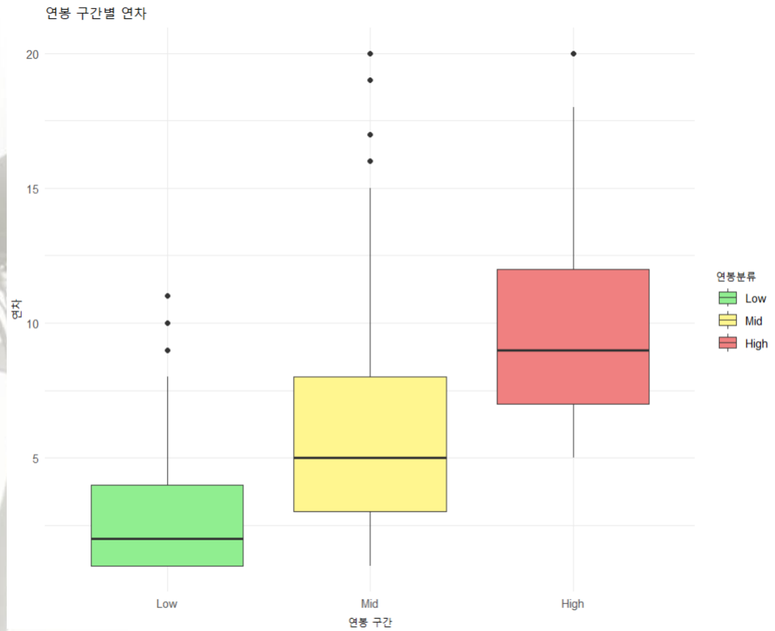
- 연봉과 비율 지표간 작은 상관관계
- 공격관련 변수끼리 큰 상관관계



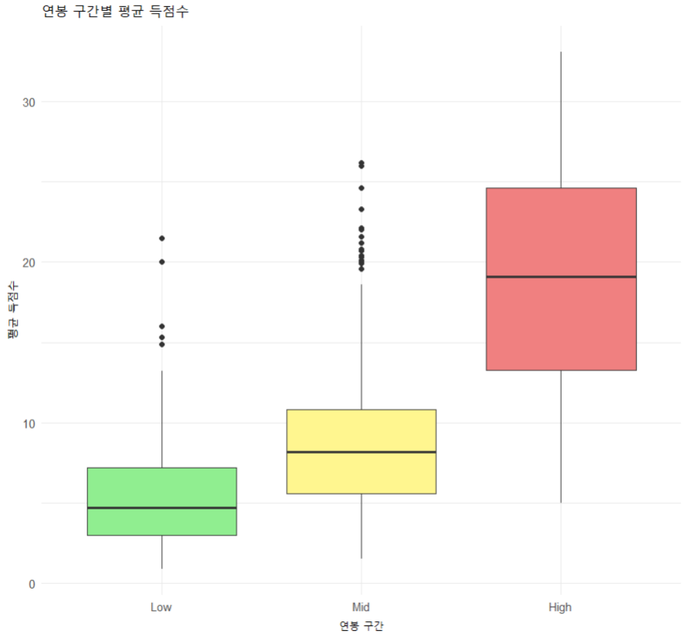
PART 2

2-2. 데이터 시각화

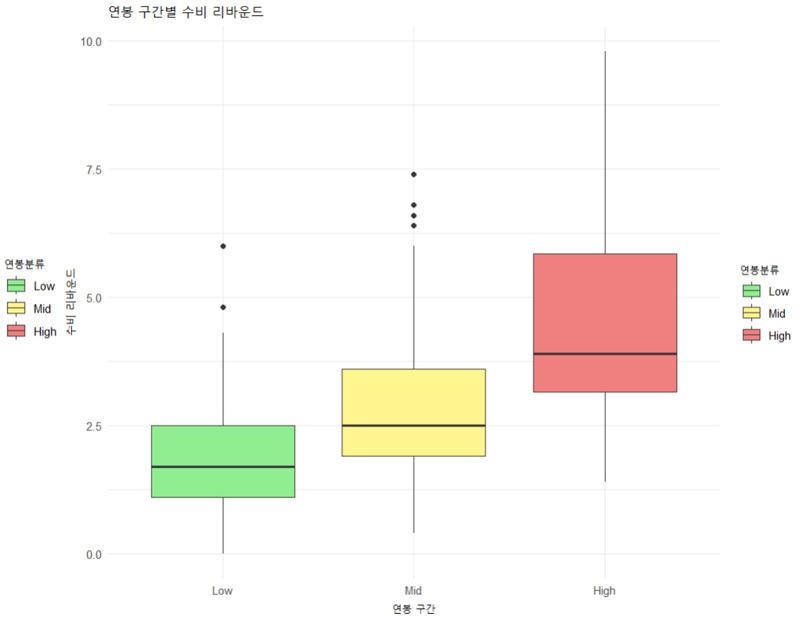
[시각화3]
- 연봉 구간별 주요변수 박스플롯



선수정보 – 연차(Years.of.Service)



공격지표 – 평균 득점수(PPG)



수비지표 – 수비 리바운드(DRB)

PART 2

2-3. 도메인 지식 기반

[변수 선택]

- 비율 관련 지표 제외 (단, TS., USG. 예외)
- 그 외에도
  소속팀(Team), WS.48(48분당 WS) 제외

득점 능력		
index	변수명	설명
9	FG	필드골 성공 수
10	FGA	필드골 시도 수
11	FG.	필드골 성공률
12	X3P	3점슛 성공 수
13	X3PA	3점슛 시도 수
14	X3P.	3점슛 성공률
15	X2P	2점슛 성공 수
16	X2PA	2점슛 시도 수
17	X2P.	2점슛 성공률
18	eFG.	유효 필드골 성공률 (eFG%)
19	FT	자유투 성공 수
20	FTA	자유투 시도 수
21	FT.	자유투 성공률
30	PPG	득점

비율/효율 지표		
index	변수명	설명
33	TS.	진정한 슈팅 비율 (TS%)
34	X3PAr	3점슛 시도 비율
35	FTr	자유투 시도 비율
36	ORB.	공격 리바운드 비율
37	DRB.	수비 리바운드 비율
38	TRB.	총 리바운드 비율
39	AST.	어시스트 비율
40	STL.	스틸 비율
41	BLK.	블록 비율
42	TOV.	턴오버 비율
43	USG.	사용률 (USG%)

03 모델구축

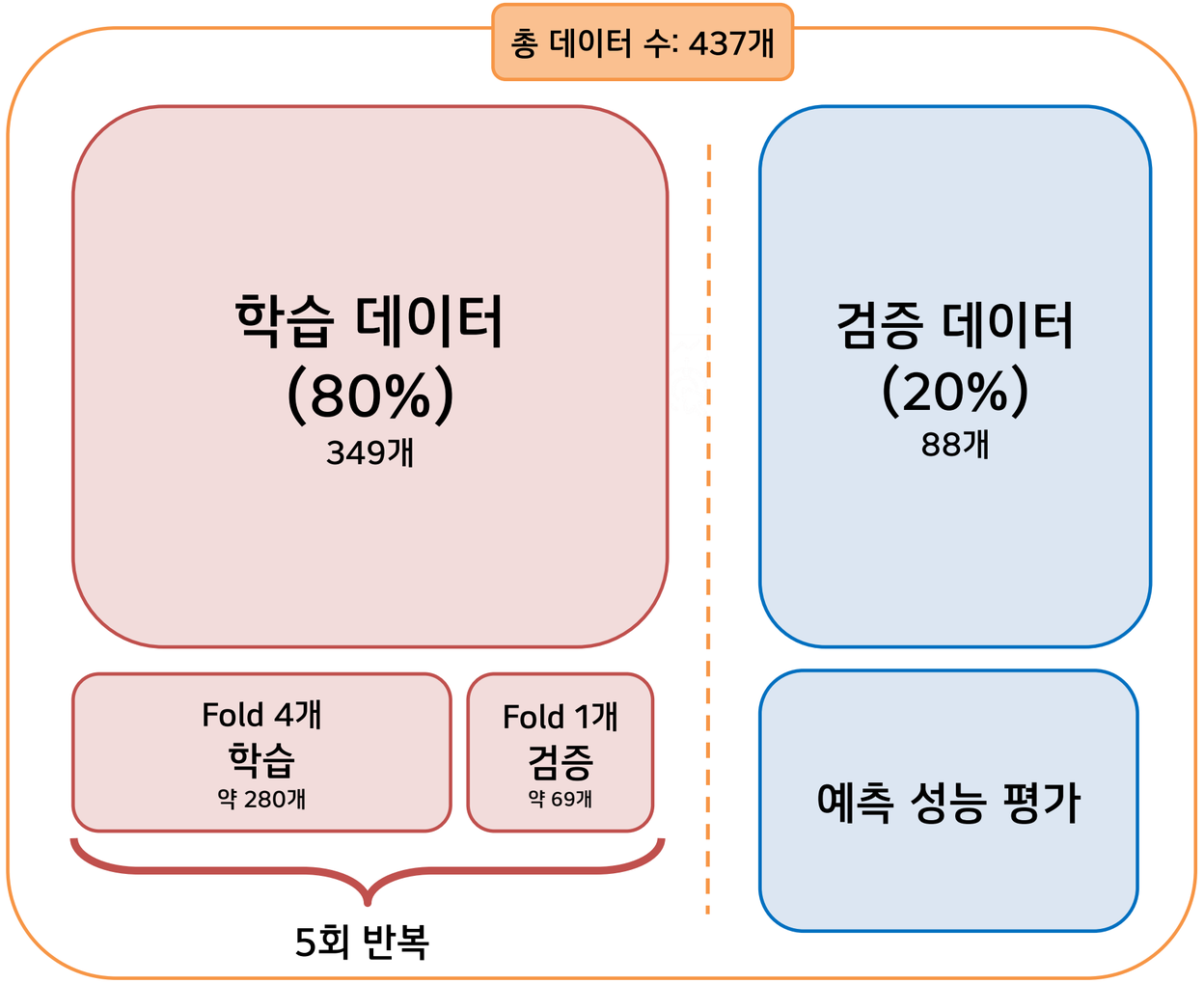
3-1. 모델 구축

[데이터 분할]

Seed number: 1

교차검증: 5-fold
반복: 5 repeats

총 25개 fold

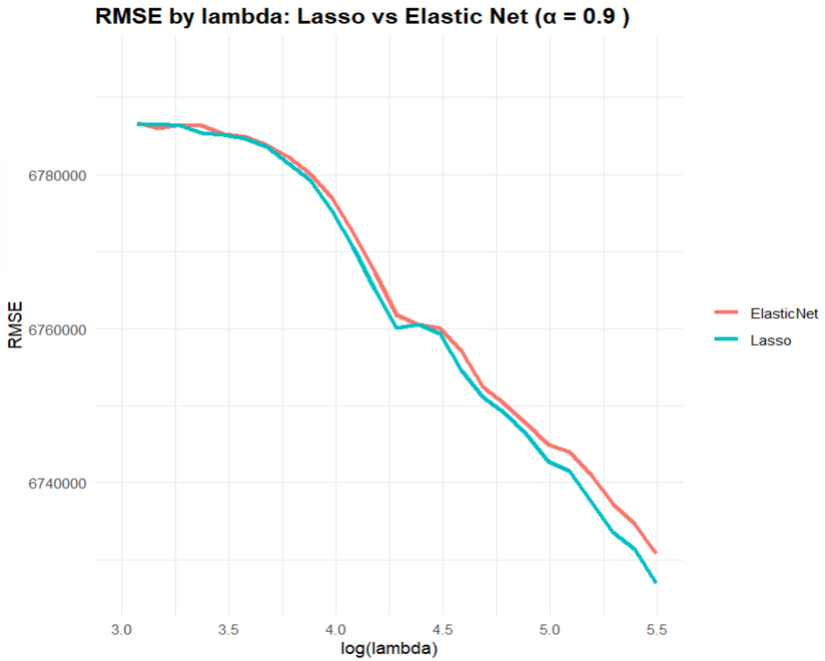


3-2. 변수 선택

[규제 기반의 최종 변수 선택]

*Elastic Net: L1(라쏘) 규제와 L2(릿지) 규제를 동시에 적용 (0<alpha<1) – (논문에서 선택한 방법)

- Lasso(alpha=1)와 Elastic Net(alpha=0.9) 비교



대략 Log(lambda) 3.5이상일 시,
RMSE 성능 Lasso > ElasticNet

L1(Lasso)

교차검증 기반의
lambda.1se 기준으로
변수 선택

최종 변수 9개
Years.of.Service
GS
FGA
X2PA
FTA
PPG
AST
TOV
VORP

PART 3

참고문헌: Zhang, Y. (2022). *NBA Player Salary Analysis based on Multivariate R.*

3-3. KNN 모델 1

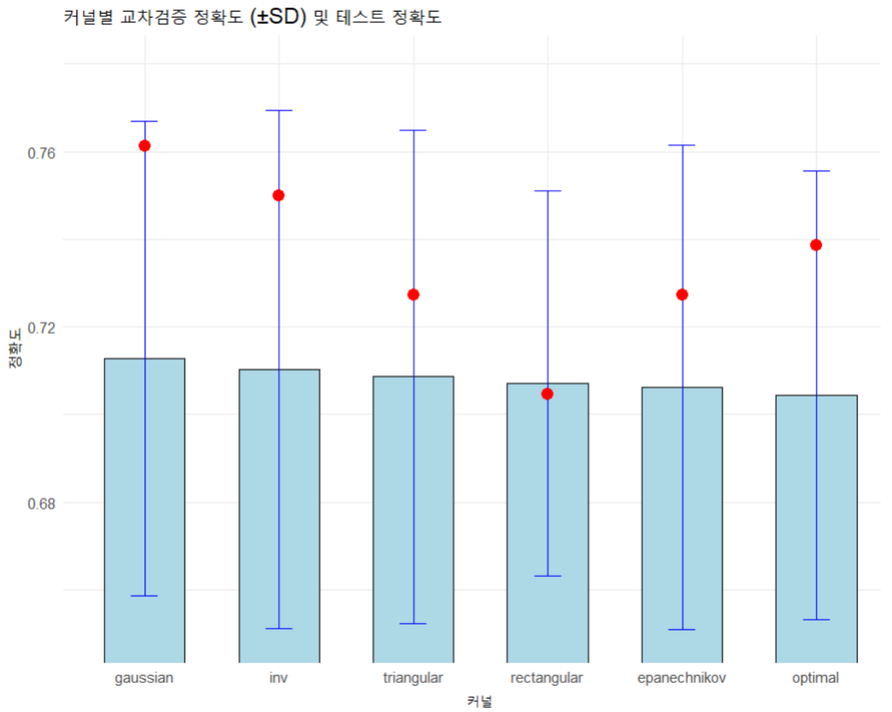
- kknn() 모델 (커널 선택 – 고정 거리: Euclidean 2)

사용한 커널: 6가지
(rectangular, triangular, epanechnikov, gaussian, inv, optimal)

커널	CV 평균 정확도	CV 표준편차	테스트 정확도	해석 요약
gaussian	0.713	0.054	0.761	가장 우수한 성능 및 예측 안정성 확보
inv	0.710	0.059	0.761	테스트 정확도가 우수하지만 불안정
triangular	0.709	0.056	0.727	무난한 성능
rectangular	0.707	0.044	0.705	안정적이지만, 테스트 정확도 낮음
epanechnikov	0.706	0.055	0.727	무난한 성능
optimal	0.704	0.051	0.739	가장 낮은 CV 평균 정확도

⇨ gaussian로 최적 커널 선택

[커널별 교차검증/테스트 정확도 시각화]



막대: CV(교차검증) 평균 정확도
에러바: CV_SD(표준편차)
점: 테스트 정확도

*knn 분석 시 사용되는 변수는 모두 수치형이므로, z-score 스케일(정규화)을 적용함.

3-3. KNN 모델 1

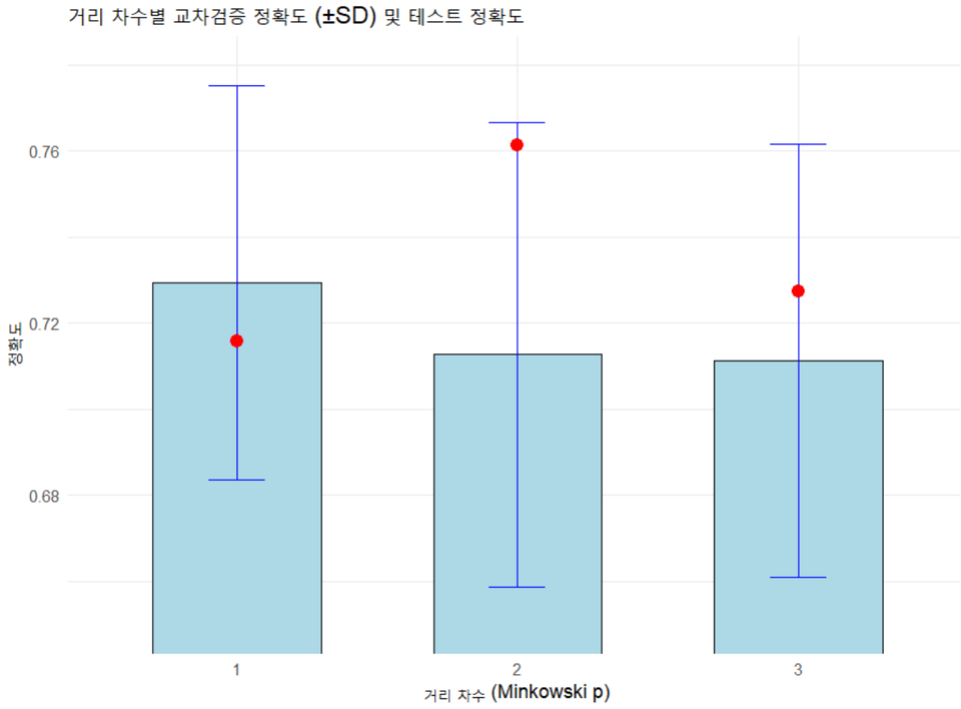
- kknn() 모델 (거리 선택 – 고정 커널: guassian)

사용한 거리: 3가지
(1: Manhattan, 2: Euclidean, 3: Minkowski)

거리	CV 평균 정확도	CV 표준편차	테스트 정확도	해석 요약
Manhattan	0.729	0.046	0.716	CV 성능이 가장 뛰어나고 안정적, 테스트 성능은 다소 낮음
Euclidean	0.713	0.054	0.761	테스트 정확도가 우수하지만 다소 불안정
Minkowski	0.711	0.050	0.727	무난한 성능

➡ 모델 안정성과 신뢰도를 고려하여
Manhattan(p=1)로 최적 거리 선택

[거리별 교차검증/테스트 정확도 시각화]



막대: CV(교차검증) 평균 정확도
에러바: CV_SD(표준편차)
점: 테스트 정확도

PART 3

*knn 분석 시 사용되는 변수는 모두 수치형이므로, z-score 스케일(정규화)을 적용함.

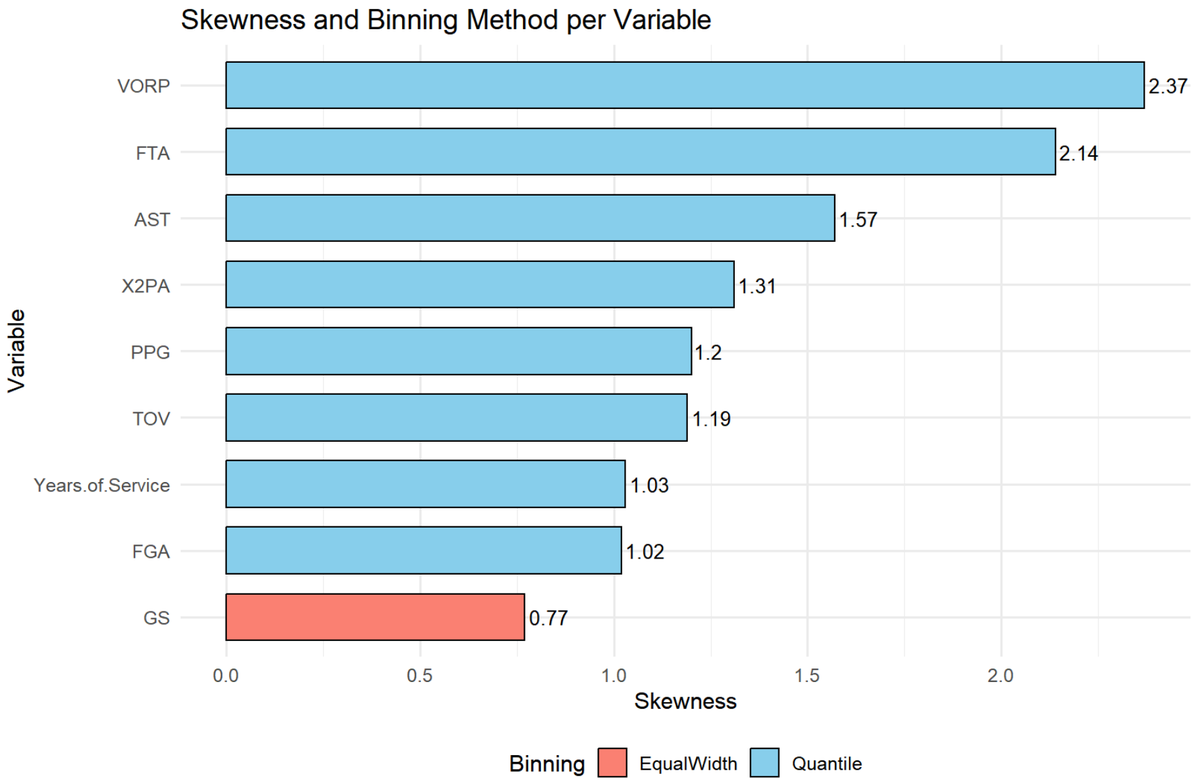
3-4. 나이브베이즈 모델 2

- 수치형 예측 변수의 범주화 기준

기준	적용방식
Skewness ≥ 1	Quantile (사분위수) 구간
Skewness < 1	Equal Width (등간격) 구간

변수의 분포 특성에 따라 왜도(Skewness)를 기준으로 유연한 구간화를 적용해 이상치 영향을 최소화함.

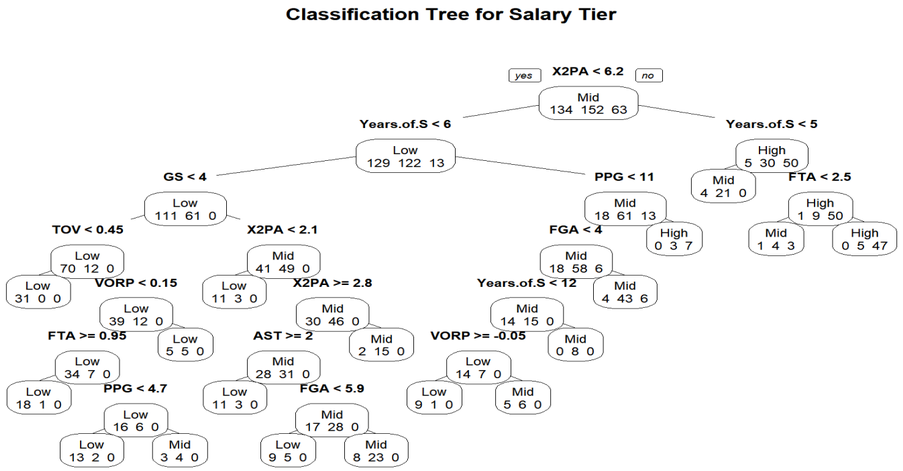
[왜도 기준 범주화 적용방식 시각화]



PART 3

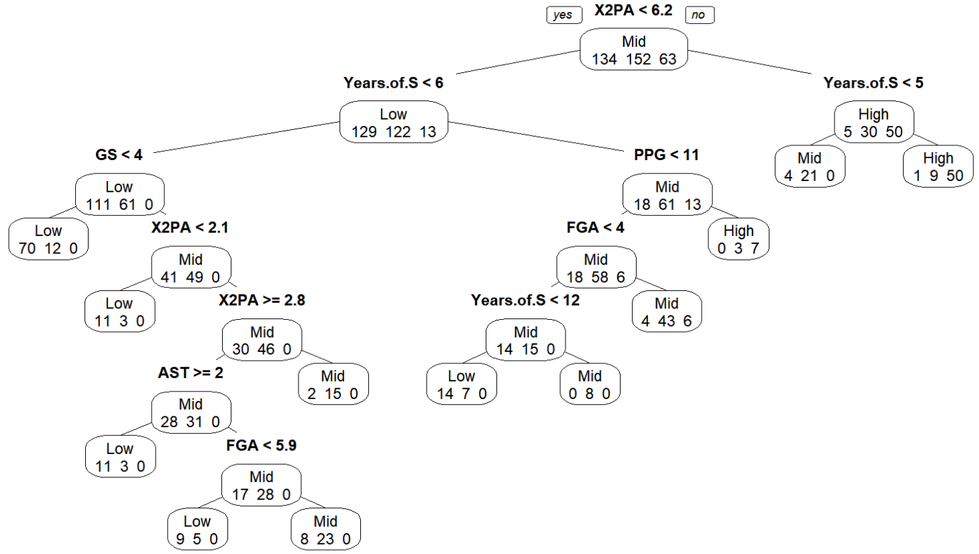
3-5. 트리 기반 모델 [결정트리] 3

1) 완전성장트리



High 분류는 아주 간단
Mid과 Low 분류는 복잡

2) 가지치기



규칙 예시	해석
X2PA<6.2	경기당 평균 2점 슛 시도가 6.2번 미만이면 Low, 아니면 High
Years.of.S<6	선수 경력이 6년 차 미만이면 Low, 아니면 Mid

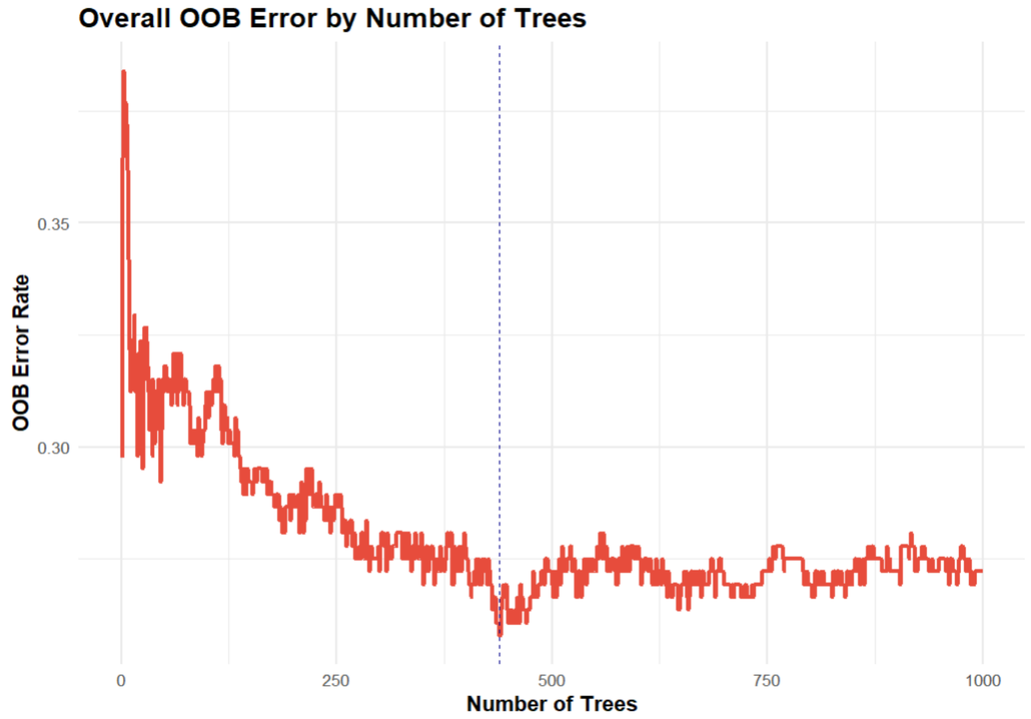
PART 3

3-5. 트리 기반 모델 [랜덤포레스트] 4

- 트리의 수(B)에 따른 OOB Error

*Out-Of-Bag Error?

랜덤포레스트가 학습 도중 자체적으로 추정하는 검증 오류율
즉, 별도의 검증 셋 없이도 모델 성능 평가하는 내장된 교차검증 방식



트리 수가 439개가 최적
하지만, 비교를 위해 500개 트리 수
(B=500)로 고정

04 분석 결과 및 결론

4. 분석 결과 및 결론 [4-1 모델 별 전체 결과 비교]

모델	교차검증(CV)			검증 데이터	
	정확도	표준편차	kappa	정확도	kappa
KNN(gaussian)	0.7340	0.0345	0.5753	0.7614	0.4886
Naive Bayes	0.6439	0.0498	0.4455	0.6512	0.4540
결정나무	0.6549	0.0491	0.4510	0.750	0.6101
랜덤포레스트	0.7260	0.0436	0.5604	0.7386	0.5859

[해석]

KNN(gaussian)	검증 데이터에서 가장 높은 정확도(76.1%)와 가장 높은 교차검증 결과들을 기록했지만, 검증 데이터에서의 Kappa(0.49)로 클래스 불균형에는 상대적으로 덜 강함.
Naive Bayes	전체적으로 모든 정확도(64.4%, 65.1%)와 Kappa(0.45)가 낮아 성능이 부족하여 클래스 간 구분력이 약함.
결정나무	검증 데이터 정확도(75%)와 Kappa 점수(0.61)가 가장 높아 분류의 일관성이 우수하지만, 교차검증 결과가 낮아 일관된 성능을 내지 못함. -> 과적합 가능성
랜덤포레스트	교차검증과 검증 데이터 모두 비교적 높은 정확도(72.6%/73.9%)와 Kappa(0.56/0.59)를 달성해 안정적이고 균형 잡힌 모델로 평가됨.

PART 4

*kappa: 우연히 맞춘 분류를 제외하고 실제 일치 정도를 측정하는 지표

4. 분석 결과 및 결론 [4-2 모델 별 클래스 결과 비교]

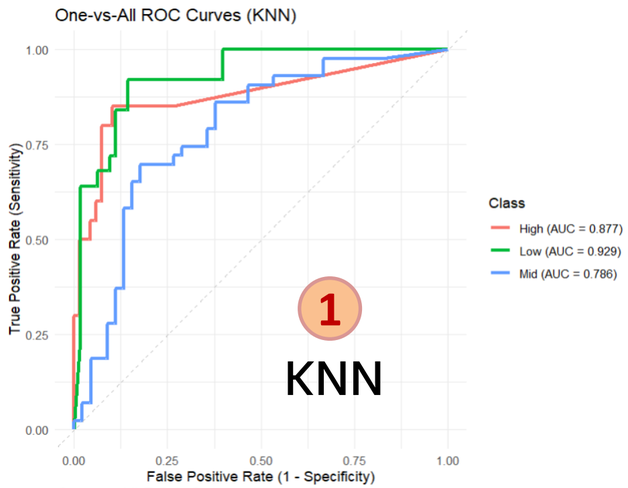
모델	Class	민감도	특이도	정밀도	균형 정확도
KNN(gaussian)	Low	0.8400	0.8730	0.7241	0.8565
	Mid	0.6977	0.8222	0.7400	0.7599
	High	0.8000	0.9265	0.7619	0.8632
Naive Bayes	Low	0.6129	0.8909	0.7600	0.7519
	Mid	0.6410	0.6809	0.6250	0.6609
	High	0.7500	0.8714	0.5714	0.8107
결정나무	Low	0.8000	0.8571	0.6897	0.8286
	Mid	0.6977	0.8222	0.7895	0.7599
	High	0.8000	0.9265	0.7619	0.8632
랜덤포레스트	Low	0.7200	0.8730	0.6923	0.7965
	Mid	0.7209	0.7556	0.7381	0.7382
	High	0.8000	0.9412	0.8000	0.8706

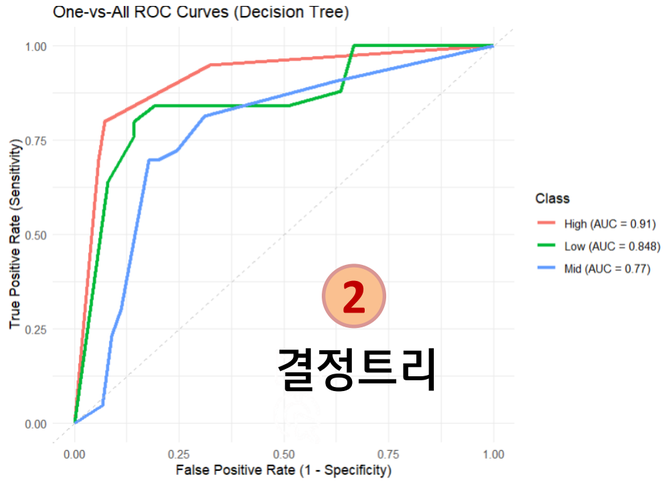
[해석]

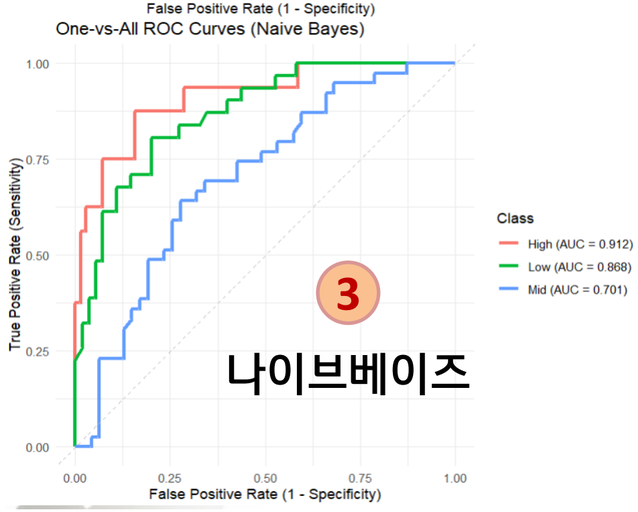
- 클래스별 성능은 High > Low > Mid 순으로 뚜렷하게 구분.
- KNN은 Low 클래스 예측에 강점.
- 랜덤 포레스트는 High 클래스에서 전반적으로 우수한 성능을 보임.
- Mid 클래스는 모든 모델에서 비교적 예측이 어려운 구간으로 나타남.

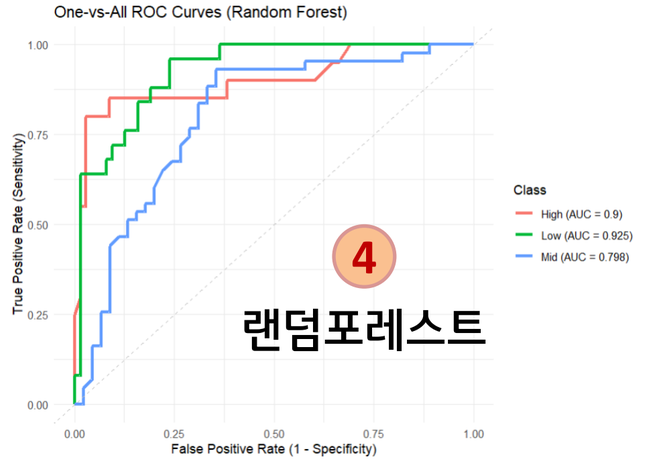
민감도(재현율)이 높으면?
실제 연봉자를 잘 찾아냄 → 이득 기회 손실 줄임
정밀도가 높으면?
잘못된 예측 줄임 → 불필요한 보상/선발 줄임

4. 분석 결과 및 결론 [4-3 모델 별 클래스 ROC 곡선 및 AUC 비교]









모델	해석
랜덤포레스트	전반적으로 가장 높은 AUC, Mid 구분력 최고
KNN	Low, High 구분 양호, Mid 구분 약함
결정트리	해석 용이하고 성능 무난, Mid 성능 아쉬움
나이브베이즈	단순한 모델이라도 High와 Low 구분 잘함, Mid 분류력 떨어짐

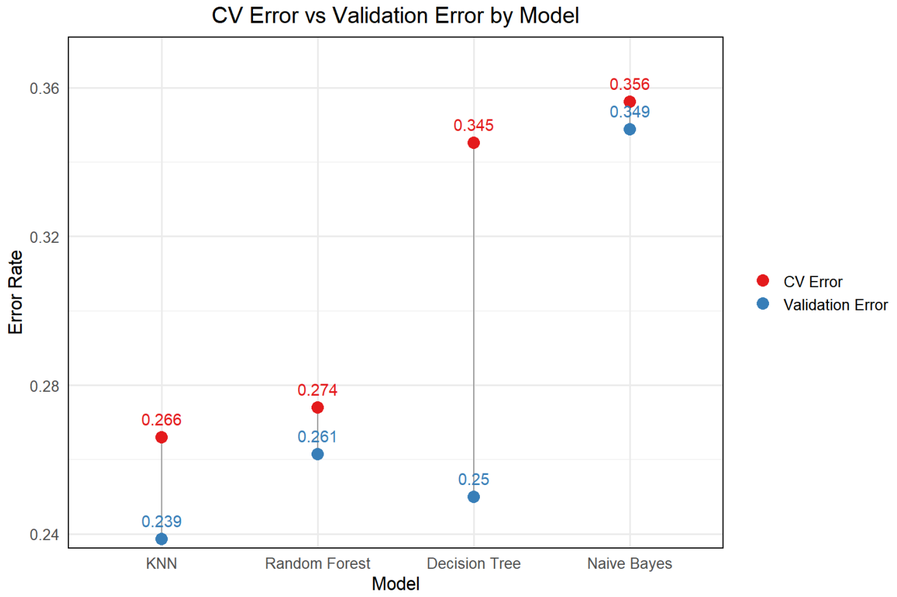
- 전체적으로 Mid 분류력이 낮음.

*Mid -> Mid

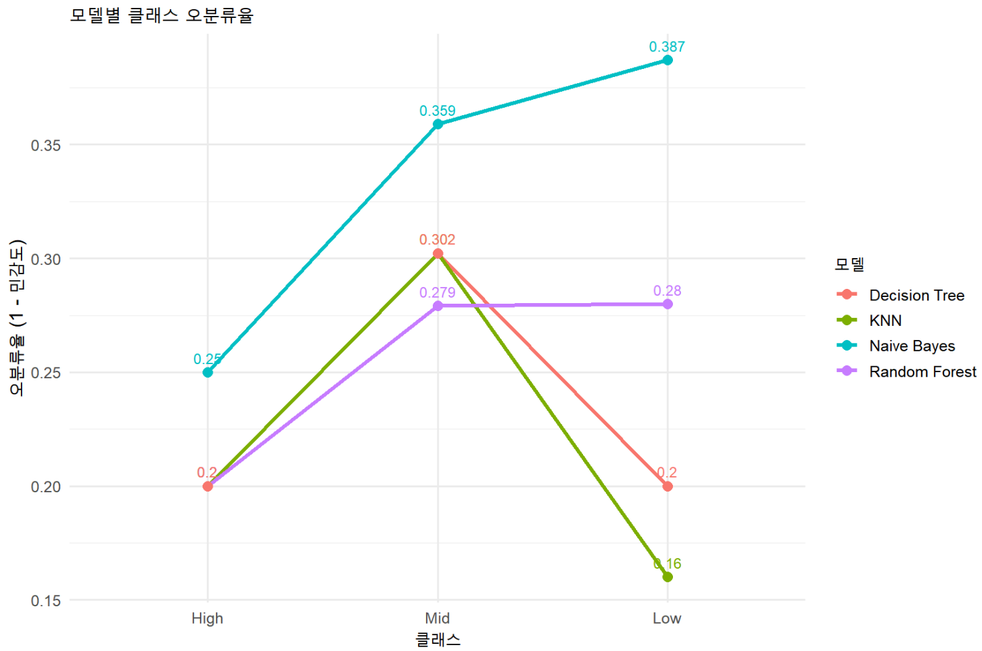
PART 3

4. 분석 결과 및 결론 [4-4 모델 별 오분류율 비교]

모델 별 CV와 검증 오분류율



모델 별 클래스에 따른 검증 오분류율



성능: [KNN > 랜덤포레스트 > 결정트리 > 나이브베이즈]

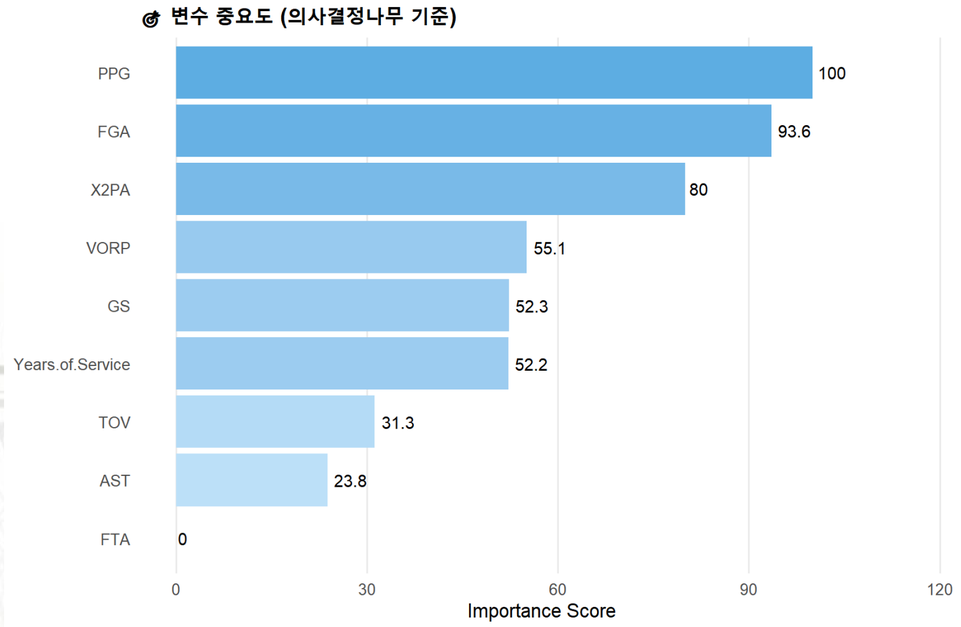
전체 클래스에서 안정적으로 좋은 모델: 랜덤포레스트
성능지표가 좋은 모델: KNN

예측의 안정성, 실제 운영 전략, 변수 해석 뿐만 아니라
지금까지의 전체적인 결과를 종합적으로 보았을 때
랜덤포레스트 모델이 적합

PART 4

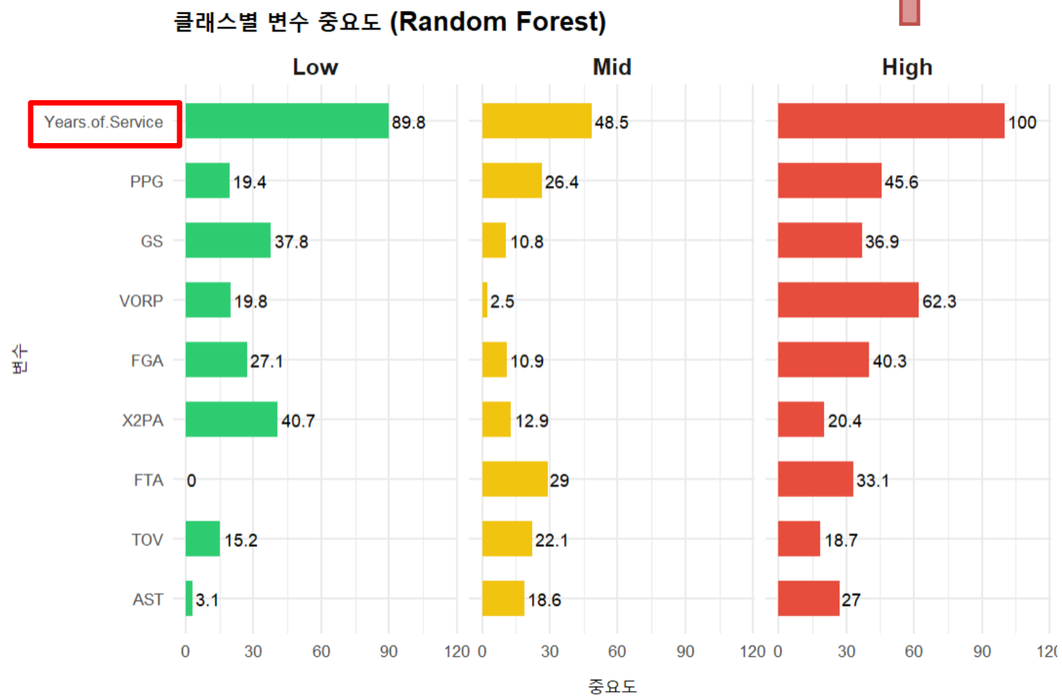
4. 분석 결과 및 결론 [4-5 변수 중요도]

[결정트리]



[랜덤포레스트]

High에서 모든 변수 중요도가 비교적 높음



득점력, 슛 시도, 기여도, 선발경기, 경력 순으로
결정트리 모델의 연봉 분류에서 핵심 지표

클래스	중요 변수
High	Years.of.Serive(경력), VORP(선수 기여도), PPG(득점 수) 등
Mid	Years.of.Service(경력), FTA(자유투 시도), PPG(득점 수), TOV(실책) 등
Low	Years.of.Service(경력), X2PA(2점 슛 시도), GS(선발경기 수) 등

PART 4

4. 분석 결과 및 결론 [4-5 변수 중요도]

연봉 분류에 영향을 주는 **핵심 지표**

1. Year.of.Service (경력) – 리그 내 누적 경험 및 경력

2. PPG (경기당 평균 득점 수) – 득점력 및 팀 내 영향력

3. VORP (대체선수 대비 기여도) – 팀 기여도 전체를 종합적으로 평가

4. GS (선발 출전한 경기 수) – 팀 내 핵심 선수로의 기용 여부

특이 변수 - TOV (경기당 평균 실책 수) – 선수 책임 비중 및 기여도

4. 분석 결과 및 결론 [4-6 인사이트] – 랜덤포레스트 기반

[과대 예측: Mid → High]

선수 이름: Darius Garland

-PPG 21.6 / GS 69 / FGA 16.4 / 경력 5년
-퍼포먼스만 보면 고연봉자에 준하는 수준.
하지만 루키 계약으로 인해 실제 연봉은 Mid에 머무름

시사점:
오분류의 측면: 루키 계약으로 인한 일시적 Mid 연봉을 반영하지 못함
인재 발굴의 측면: 모델이 Garland를 고성과-저연봉의 가성비 인재로 판단

[과소 예측: High → Mid]

선수 이름: Al Horford

- PPG 9.8 / VORP 2.5 / 경력 16년
- 성과는 평범하지만, 베테랑 프리미엄과 리더십 가치를 반영해 실제 연봉은 높음
- 모델은 수치 중심이라 경력·시장 평가는 누락됨

시사점:
'현재 시즌 성과'만으로는 베테랑 가치판단이 어려움

PART 4

05 한계 및 소감

5-1. 한계

1. 농구 통계 지표 간에는 구조적으로 높은 상관관계가 존재
라쏘 회귀로 변수 선택을 하여도 **다중공선성**이 완전히 해소되지 않음

2. 단일 시즌 데이터를 기반으로 분석했기 때문에 해당 시즌 내에서는 의미가 있지만,
시간에 따른 일반화에 한계가 존재
-> 여러 시즌을 고려한 분석이 향후 필요

3. 경기 성과 중심의 통계 지표만을 활용했지만,
실제 연봉에는 계약 구조, 시장 상황, 선수 개인적 요소(부상 이력, 인기 등)에 의해 **복합적**으로 결정
-> FA 여부, 계약 연도, 부상 기록, SNS 영향력 등 **비정량적 요소**를 포함한
확장된 데이터셋을 활용해 정교한 예측 모델 구축 필요

5-2. 소감

최현석

처음으로 데이터마이닝 프로젝트를 진행하다 보니 변수 선택과 모델 설계 과정에서 많은 고민이 있었지만, 수업 시간에 배운 내용과 농구 연봉 예측 관련 논문들을 참고하면서 조금씩 방향을 잡을 수 있었다. 농구 연봉 분류라는 이번 프로젝트를 통해 데이터마이닝이 단순한 분석을 넘어서 전략적인 판단에도 활용될 수 있다는 점이 흥미로웠고, 앞으로도 다양한 주제를 바탕으로 더 정교한 예측 모델을 설계해보고 싶다는 생각이 들었다.

농구 선수 연봉 분류 모델을 만들면서, 성과 지표만으로는 연봉을 충분히 설명할 수 없고, 계약 구조나 인기와 같은 비정량적 요인들도 큰 영향을 미친다는 것을 실감했다. 그래도 제한된 데이터 안에서 다양한 모델을 비교하고 중요한 변수를 찾아보는 과정은 흥미로웠다. 평소 관심이 있던 스포츠 데이터를 주제로 데이터 마이닝 기법을 직접 적용해볼 수 있어 의미 있는 경험이었다.

황성진

THANK YOU

단국대학교 정보통계학과 최현석, 황성진