7조 보고서

담당교수 : 이경은 교수님
과 목 : 데이터 마이닝 및 실험
팀 원 : 통계학과 김성은
통계학과 정예원
글솝전공 노태현
글솝전공 정명원

KNU KYUNGPOOK NATIONAL UNIVERSITY

목차

1. 연구 목적 및 데이터 전처리

1) 연구 목적

축구를 보면서 자신이 좋아하는 팀을 응원하는 경우가 있을 것이다. 하지만 어느 팀이 경기에서 이길 지 예측해보는 것도 관심사 중 하나일 것이다. 하지만 개인이 주관적인 기준으로 경기의 승자를 예측하기에는 어려움이 있어 보인다. 우리는 이번 프로젝트를 통해 각 팀의 선수들의 데이터들을 통해 개인 선수 기록이 승률에 어떤 영향을 미칠지, 이어 축구 경기의 승률을 예측할 수 있을지 알아보려고 한다.

2) 데이터 전처리

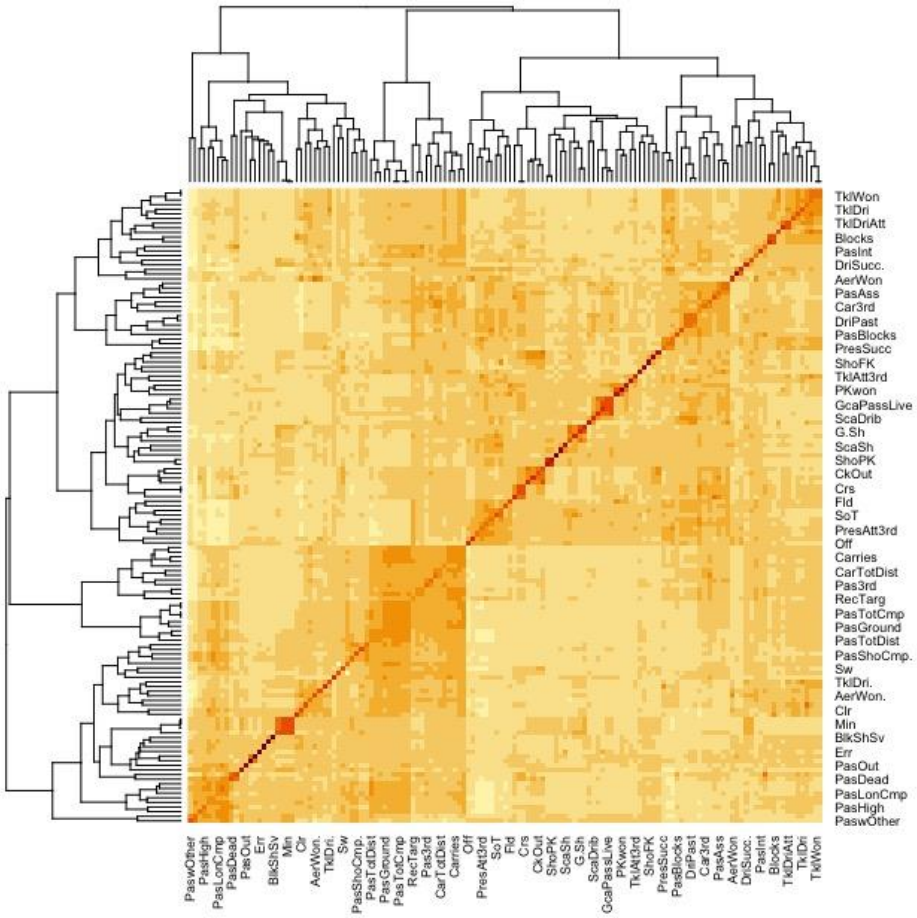
Rk	Player	Nation	Pos	Squad	Comp	Age	Born	MP	Starts	Min	90s	Goals	Shots	SoT
1	Max Aarons	ENG	DF	Norwich City	Premier Leagı	22	2000	30	28	2521	28	0	0.43	0.07
2	Yunis Abdelhamid	MAR	DF	Reims	Ligue 1	34	1987	31	31	2713	30.1	0.07	0.6	0.2
3	Salis Abdul Samed	GHA	MF	Clermont Foc	Ligue 1	22	2000	30	28	2398	26.6	0.04	0.68	0.19
4	Laurent Abergel	FRA	MF	Lorient	Ligue 1	29	1993	31	31	2708	30.1	0	0.86	0.23
5	Charles Abi	FRA	FW	Saint-Étienne	Ligue 1	22	2000	1	1	45	0.5	0	0	0
6	Dickson Abiama	NGA	FW	Greuther Fürt	Bundesliga	23	1998	23	5	723	8	0	2.25	0.5
7	Matthis Abline	FRA	FW	Rennes	Ligue 1	19	2003	7	1	103	1.1	0	1.82	0
8	Tammy Abraham	ENG	FW	Roma	Serie A	24	1997	34	33	2815	31.3	0.48	2.78	0.96
9	Luis Abram	PER	DF	Granada	La Liga	26	1996	8	6	560	6.2	0	0.32	0
10	Francesco Acerbi	ITA	DF	Lazio	Serie A	34	1988	27	26	2266	25.2	0.16	0.56	0.2
11	Ragnar Ache	GER	MF-FW	Eint Frankfurt	Bundesliga	23	1998	13	1	259	2.9	0	3.79	0.69
12	Marcos Acuña	ARG	DF	Sevilla	La Liga	30	1991	27	22	1900	21.1	0.05	0.57	0.19
13	Che Adams	SCO	FW	Southamptor	Premier Leagı	25	1996	28	23	2002	22.2	0.32	2.16	1.08
14	Tyler Adams	USA	MF	RB Leipzig	Bundesliga	23	1999	22	12	1313	14.6	0	0.14	0
15	Sargis Adamyan	ARM	FW-MF	Hoffenheim	Bundesliga	28	1993	13	2	331	3.7	0.27	0.81	0.54
16	Martin Adeline	FRA	MF-FW	Reims	Ligue 1	18	2003	8	2	352	3.9	0	1.03	0.26
17	Amine Adli	FRA	FW-MF	Leverkusen	Bundesliga	21	2000	25	13	1256	14	0.21	2.71	1.36
18	Yacine Adli	FRA	MF-FW	Bordeaux	Ligue 1	21	2000	35	25	2234	24.8	0.04	1.29	0.6
19	Michel Aebischer	SUI	MF	Bologna	Serie A	25	1997	10	2	287	3.2	0	0.63	0.63
20	Felix Afena-Gyan	GHA	FW-MF	Roma	Serie A	19	2003	17	6	668	7.4	0.27	2.43	0.95
21	Martin Agirregabiria	ESP	DF	Alavés	La Liga	25	1996	21	16	1501	16.7	0	0.24	0
22	Julen Agirrezabala	ESP	GK	Athletic Club	La Liga	21	2000	4	4	360	4	0	0	0
23	Lucien Agoume	FRA	MF	Brest	Ligue 1	20	2002	25	20	1763	19.6	0	0.82	0.05
24	Kevin Agudelo	COL	FW-MF	Spezia	Serie A	23	1998	20	9	1033	11.5	0.26	1.48	0.7
25	Nayef Aguerd	MAR	DF	Rennes	Ligue 1	26	1996	29	29	2555	28.4	0.07	0.88	0.25
26	Sergio Agüero	ARG	FW	Barcelona	La Liga	33	1988	4	2	151	1.7	0.59	3.53	0.59
27	Ruben Aguilar	FRA	DF	Monaco	Ligue 1	29	1993	25	19	1870	20.8	0	0.38	0.05
28	Álvaro Aguirre	ESP	FW	Rayo Vallecar	La Liga	22	2000	1	0	9	0.1	0	0	0
29	Naouirou Ahamada	FRA	MF	Stuttgart	Bundesliga	20	2002	3	0	61	0.7	0	0	0
30	Anel Ahmedhodzic	BIH	DF	Bordeaux	Ligue 1	23	1999	12	12	1080	12	0	0.5	0.33
31	Jean-Eudes Aholou	CIV	MF	Strasbourg	Ligue 1	28	1994	22	10	886	9.8	0.1	1.43	0.2
32	Joseph Aidoo	GHA	DF	Celta Vigo	La Liga	26	1995	28	26	2277	25.3	0	0.24	0.04

<2021-2022 시즌 유럽축구 5대 리그 선수 기록 + 팀 기록>[1]

위의 이미지는 2021-2022 시즌 유럽축구 5대 리그 선수 기록과 팀기록을 합친 자료의 일부이다. 총 설명변수는 160개, 데이터의 개수는 2856개로 이루어져 있는 데이터이다. 연구에 앞서 이 데이터에 대해 전처리를 진행하였다.

팀 승률을 계산 후 변수로 추가하였다. 순위 변수는 제거하였고, 출생연도는 생일과 중복되기에 제거하였다. 팀 승률을 제외한 팀 기록 관련 변수를 제거하고 수치형 변수를 제외한 범주형 변수를 제거하였다. 결측치의 경우 한 선수의 국적이 표기 되어있지 않던 것을 수정하였다. 마지막으로 모형을 적합시켜야 하기 때문에 총 데이터를 Training Data와 Test Data를 7:3으로 분리하였다.

[1] "2021-2022 Football Player Stats", Kaggle, 2022년5월29일 수정, 2022년5월22일 접속
https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats



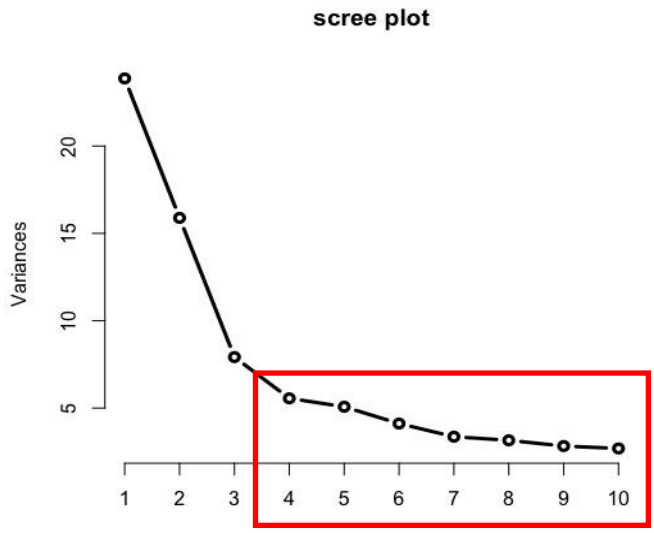
<전처리 후 상관관계 확인>

변수 간 상관관계를 확인하기 위해 r에서 heatmap 메서드를 실행해본 결과이다. 데이터를 관측한 결과 다중공산성과 VIF는 확인 불가능 하였다.

2. 분석 적용

1) PCA

현재 데이터셋은 설명 변수가 너무 많기 때문에 먼저 주성분 분석을 통해 차원 축소를 진행해 보았다.



procmp 함수를 통해 적절한 주성분의 개수를 확인해보았고 주성분은 4 개를 사용하기로 하였다.

```
Call:
lm(formula = win_rate ~ PC1 + PC2 + PC3 + PC4, data = pca.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38314 -0.11420 -0.01659  0.10516  0.52811

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.156e-01  1.053e-02  20.470  < 2e-16 ***
PC1         -6.881e-04  4.779e-05 -14.398  < 2e-16 ***
PC2         -8.723e-04  1.229e-04  -7.100 1.73e-12 ***
PC3         -8.277e-04  1.879e-04  -4.406 1.11e-05 ***
PC4         -3.040e-04  1.318e-04  -2.307   0.0211 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1537 on 1994 degrees of freedom
Multiple R-squared:  0.1343,    Adjusted R-squared:  0.1326
F-statistic: 77.35 on 4 and 1994 DF,  p-value: < 2.2e-16
```

주성분 회귀분석 결과 Adjusted R2 값으로 13.2%가 나오는 것을 확인하였다.

하지만 각 주성분이 무엇을 의미하는 지 이해하기 힘들었기 때문에 요인분석을 진행하였다.

2) 요인분석

주성분 분석만으로는 각 주성분이 무엇을 의미하는지 이해하기 힘들었기 때문에 요인분석을 진행하였다.

principal 함수를 사용하였고, Varimax, oblimin 회전을 한 결과이다.

요인1

PasTotCmp	0.9310
PasCmp	0.9310
PasLive	0.9304
PasGround	0.9239
PasTotAtt	0.8996
PasAtt	0.8996

요인2

Tkl	0.8273
Tkl.%	0.8059
TklDef3rd	0.7213
TklWon	0.6960
TklW	0.6960
PresDef3rd	0.6873

요인3

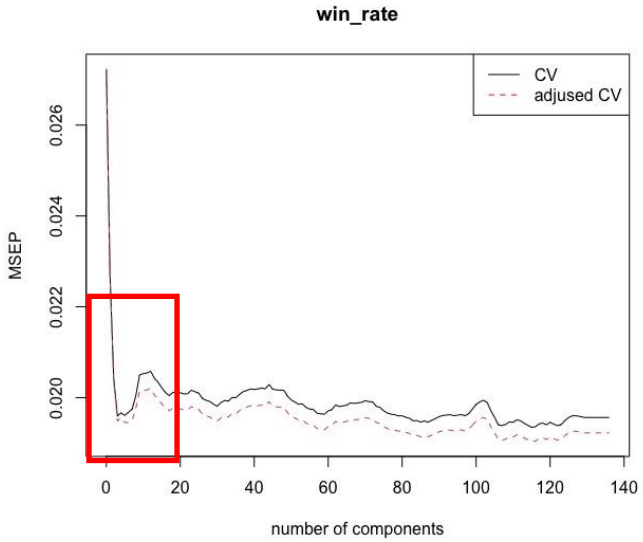
TouAtt3rd	0.7230
PasAss	0.7146
Crs	0.7114
PasCrs	0.7114
SCA	0.6646
CK	0.6131

요인4

PasHigh	-0.6853
PasLonAtt	-0.6808
TouDef3rd	-0.6684
PasTotPrgDist	-0.6673
PasDead	-0.6668
RecProg	-0.6562

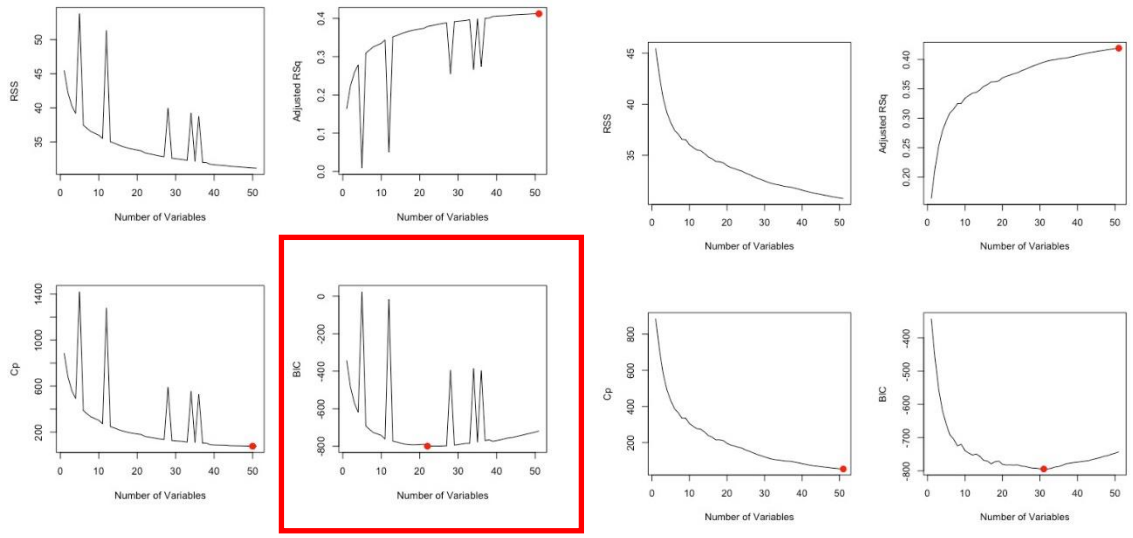
결과는 주성분 분석과 비슷하였고, 각 요인들을 1 부터 4 까지 "패스", "태클,압박", "롱패스, 크로스", "수비 진영에서 공격진영으로 롱패스"로 구분하였다.

3) PLS

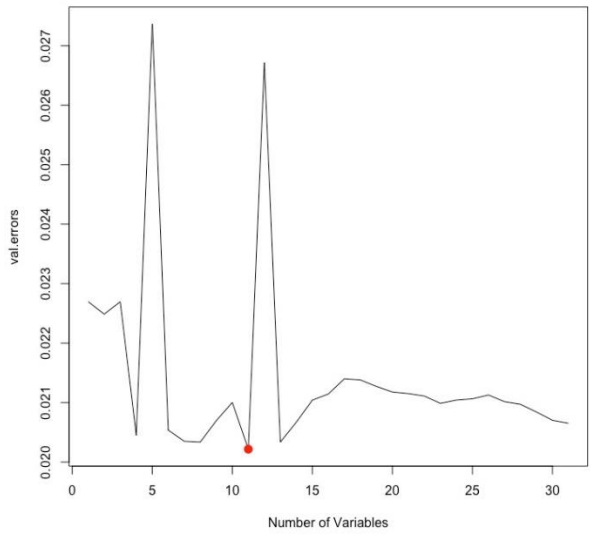


PLS 에서는 적절한 주성분의 개수가 5 개 인 것을 확인할 수 있었다. 주성분 개수 5 개로 예측을 해본 결과 Adjusted R2 가 27.2%로 이전의 PCA 방법과 비교하여 14%정도 높은 정확도를 보여주었다.

4) Best Subset Selection

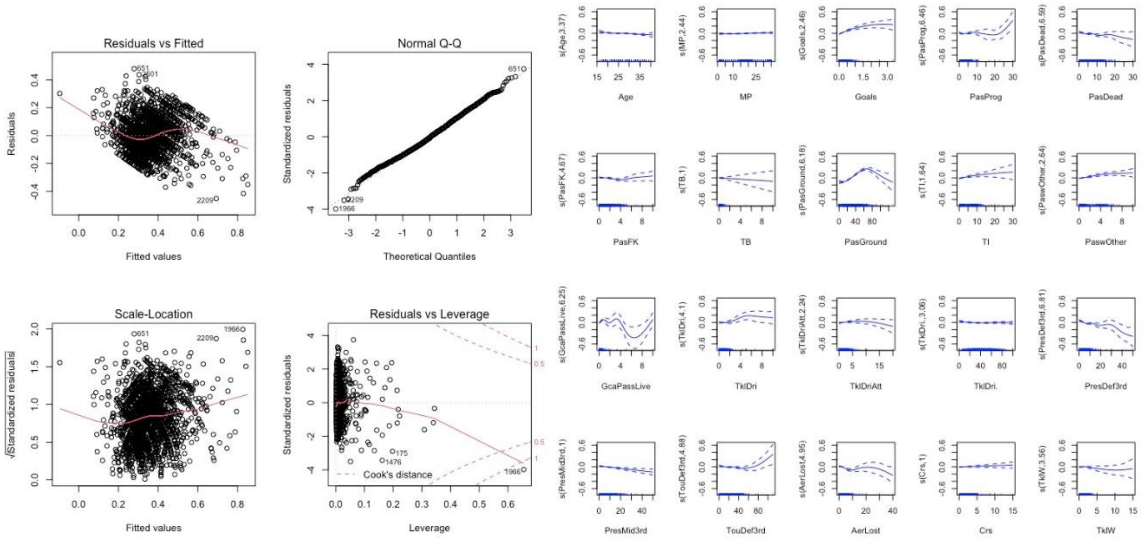


<좌: 단계선택법, 전진선택법, 우: 후진제거법>



<Test MSE 기준>

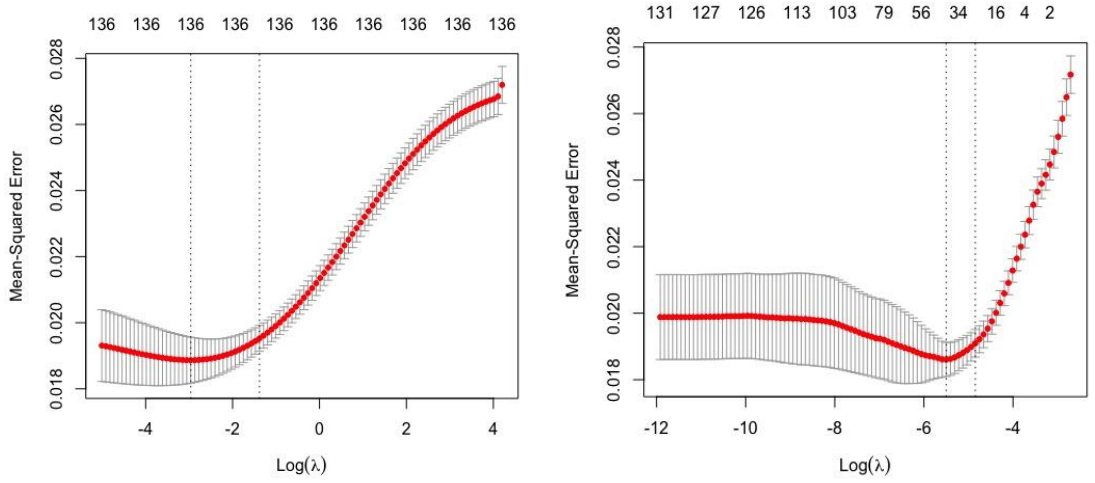
단계선택법과 전진선택법 기준으로는 최적의 변수가 22개, 후진 제거법의 경우 31개, Test MSE의 경우 11개가 각 경우에 대해서 최적의 결과가 나왔다. 최종 적합 변수를 정하기 위해 vif함수를 통해 다중공산성을 확인하였고 변수 정리를 해본결과 변수개수가 각각 11개, 22개, 31개일 때 Adjusted R2가 28.6% 35.8%, 34.1%로 나왔다. 따라서 우리는 최종 적합 변수 개수를 22개로 하였다.



<좌: 다중회귀분석, 우: GAM>

다중 회귀 분석에서 주요변수는, Goals, PaswOther, GcaPassLive 변수가 선택되었고, GAM에서는 PassGround, PaswOther, GcaPassLive, PresDef3rd, AerLost 변수가 선택되었다.

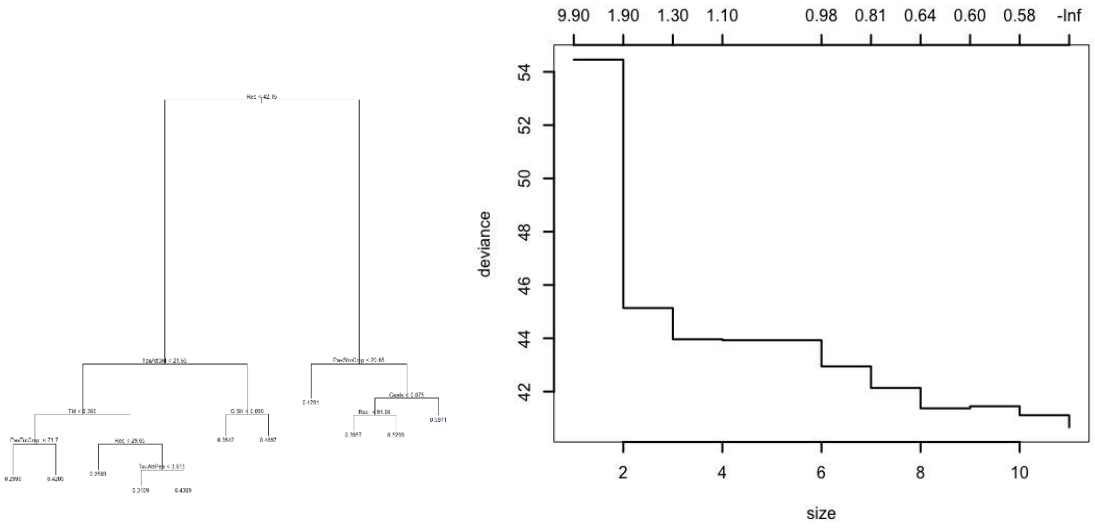
5) Ridge, Lasso Regression



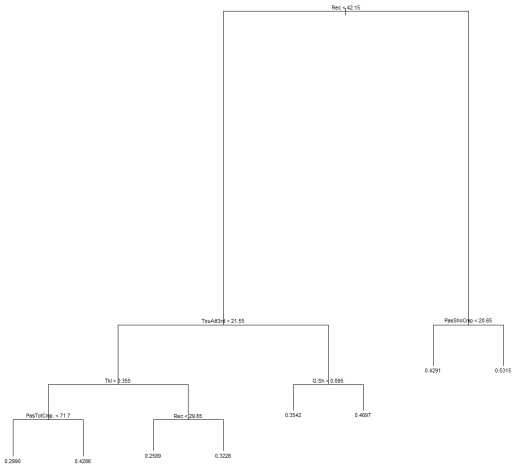
<좌: Ridge, 우: Lasso>

Ridge regression의 경우 모든 변수를 이용하여 Adjusted R2가 28.3%가 나왔고 Lasso의 경우 설명변수 44개를 선별하여 27.0%가 나왔다.

6) Decision Tree



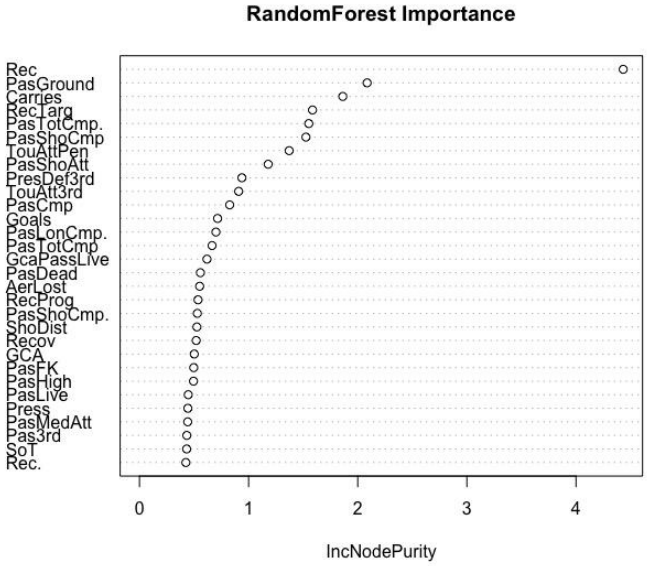
tree 함수를 사용하였을 때는 terminal node가 11개로 나왔지만 cv.tree 함수를 사용하여 적절한 node의 사이즈는 8개로 정하였다.



과적합 방지를 위해 가지치기를 진행한 결과 주요 변수는 Rec, Touatt3rd, Tkl, PasTocCmp, G/Sh, PassShoCmp가 선택되었다.

7) Random Forest

randomForest 함수를 사용하였고 인자는 ntree=500, ntry=42를 사용하였다.



변수 중요도의 순서로는 Rec, PasGround, Carries의 순서대로 선택되었지만, Rec의 경우 PasGround에 비해 2배 이상 높은 것을 확인할 수 있었다.

3. 분석 비교

분석 방법	정확도	RMSE	R square
PCA	0.1061844	0.2694188	0.05658971
PLS	0.2625437	0.1424434	0.27232271
다중회귀분석 (p = 22)	0.2683781	0.1443591	0.25520460
GAM (p = 22)	0.2695449	0.1344502	0.33628401
Ridge Regression	0.2695449	0.1394890	0.28356583
Lasso Regression	0.2590432	0.1412812	0.27041692
Decision Tree	0.2520420	0.1420488	0.26046351
Random Forest	0.3045508	0.1244505	0.42779978

앞서 다양한 방법으로 모델을 적합시켜 보았지만, 가장 정확도가 높은 Random Forest 방식도 정확도가 30%밖에 나오지 않았다. 따라서 우리는 다음과 같은 결과를 도출해내었다.

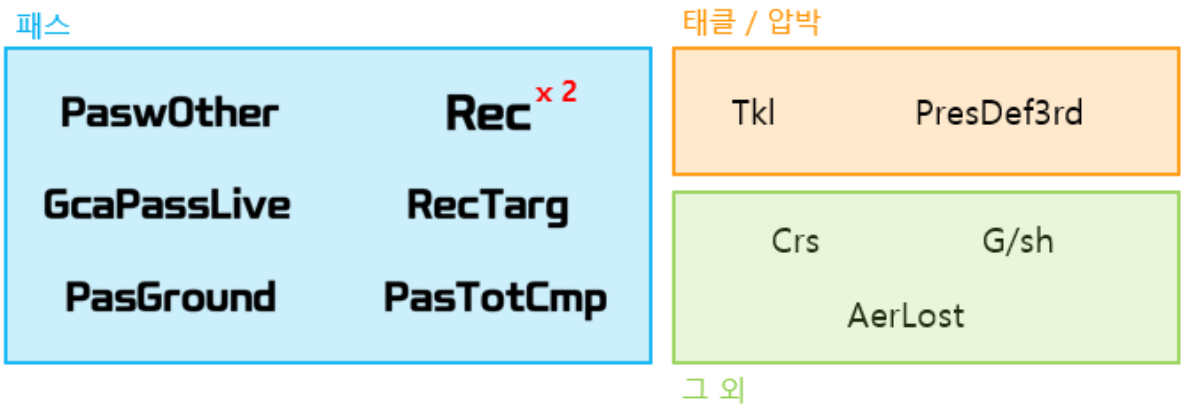
(1) 선수 개인의 역량으로 승리를 예단하기에는 한계가 있다. 즉, 축구는 팀플레이다.

(2) 패스를 얼마나 잘 하는지가 중요하다

(3) 패스를 얼마나 잘 받는지도 패스를 하는 만큼 중요하다.

4. 승률 예측

우리가 적합한 모델을 통해 2022 FIFA 월드컵 카타르에서 한국이 대한민국 조별예선에서 이길 확률을 예측해보려고 한다.



이를 위한 변수로 위의 사진과 같은 변수들을 선택하였고, Rec의 경우 중요도가 높았기 때문에 가중치를 주게 되었다. 팀 승률을 종속변수로 할 때, 회귀계수를 사용하였다.

변수	의미	손흥민	에이유	수아레즈	호날두
PaswOther	발/머리를 제외한 부위로 패스 시도 수	0.17	0.33	0.67	0.76
GcaPassLive	골로 연결되는 패스 수	0.27	0.19	0.21	0.15
PasGround	땅볼 패스 수	26.7	23.9	14.4	25.2
PasTotCmp	성공한 전체 패스 수	28.7	25.4	17.4	26.4
Rec	패스를 받은 횟수	33.1	21.0	26.1	35.2
RecTarg	패스의 표적이 된 횟수	47.4	48.3	44.9	55.1
Tkl	성공한 태클 수	0.54	2.43	0.41	0.38
PresDef3rd	수비 진영에서 상대 선수 압박 횟수	2.13	5.42	0.72	0.84
AerLost	공중볼 경합 패배 횟수	1.59	2.66	1.64	2.09
G/sh	슛 대비 골 전환 비율	0.26	0.06	0.19	0.15
Crs	크로스 수	2.06	2.24	0.46	0.46

<각 변수의 의미 및 수치>

예측 결과 손흥민은 0.59793, 에이유는 0.50462, 수아레즈는 0.48, 호날도는 0.57495로 카타르 월드컵 조별예선에서 우리나라는 3승 0무 0패를 할 것으로 예상할 수 있었다.