FIFA 2018 Statistics

김명준 이은주 하도연

*[변수설명]* (변수중에서 numeric한 것만 고른 것)

Goal Scored 득점 수 ( Winner와 Loser 을 가르는 데만 쓰이고 삭제할 변수)

Ball Possession % 볼 점유율

Attempts 득점 시도 수

On-Target 유효 슈팅 수

Off-Target 비유효 슈팅 수

Blocked 상대 팀의 득점 시도를 막은 횟수

Corners 코너 킥 횟수

Offsides 오프사이드 수

Free Kicks 프리 킥 횟수

Saves 골키퍼가 골을 막은 횟수

Pass Accuracy % 패스 정확도

Passes 패스 횟수

Distance Covered (Kms) 선수들이 뛴 거리

Fouls Committed 파울 수

Yellow Card

Yellow & Red

Red

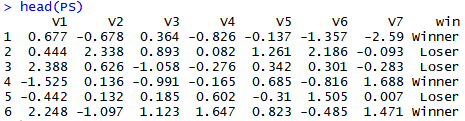
1st Goal 첫 번째 골이 나온 시간

*[Win 변수 만들어주기]*

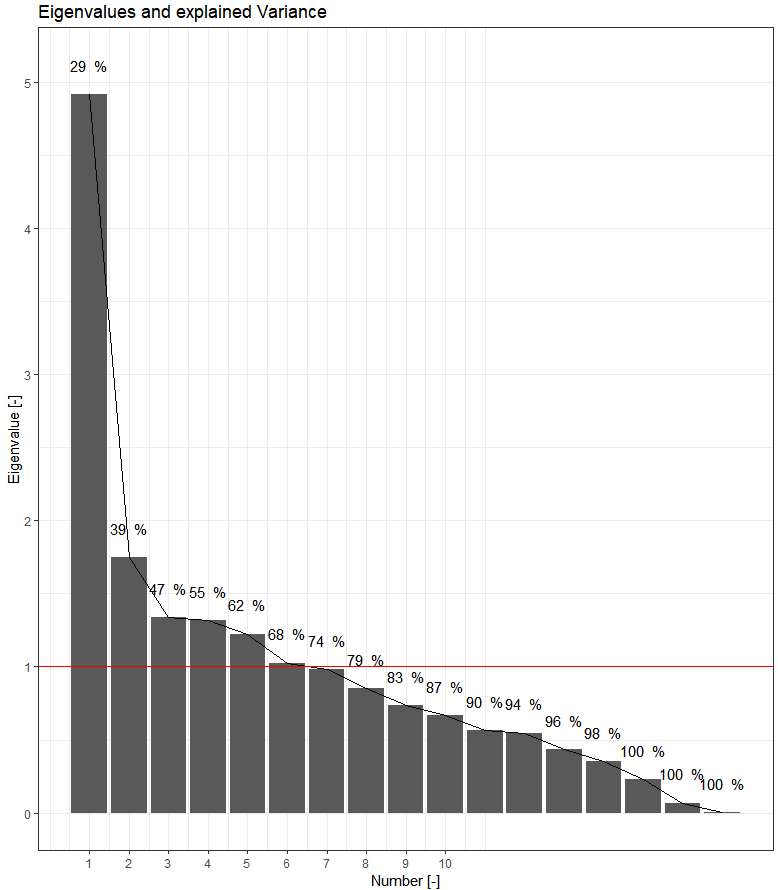
데이터를 보면 2열당 1경기를 나타내고 있다. 따라서 득점 수를 비교해서 승패를 가린다.

동점인 경우에서는 승부차기를 하는 경우 승부차기 득점 수 변수를 비교해서 승패를 가리고

동점인데도 승부차기를 안 하는 경우는 제거한다. ( 랜덤포레스트 모형에서 AUC와 ROC 곡선을 구하기 위해서 )

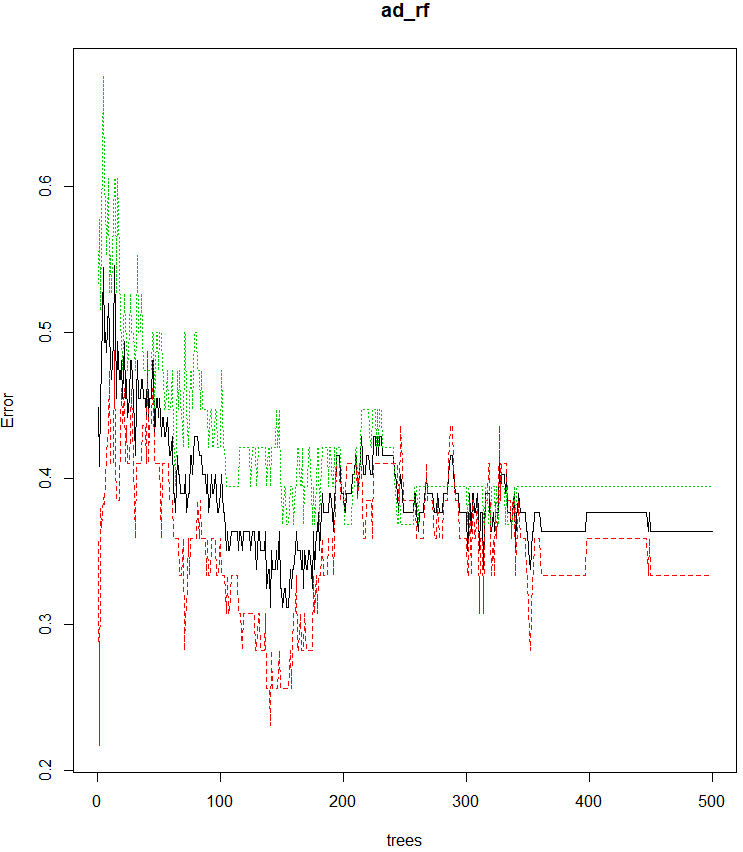


*[주성분분석]*

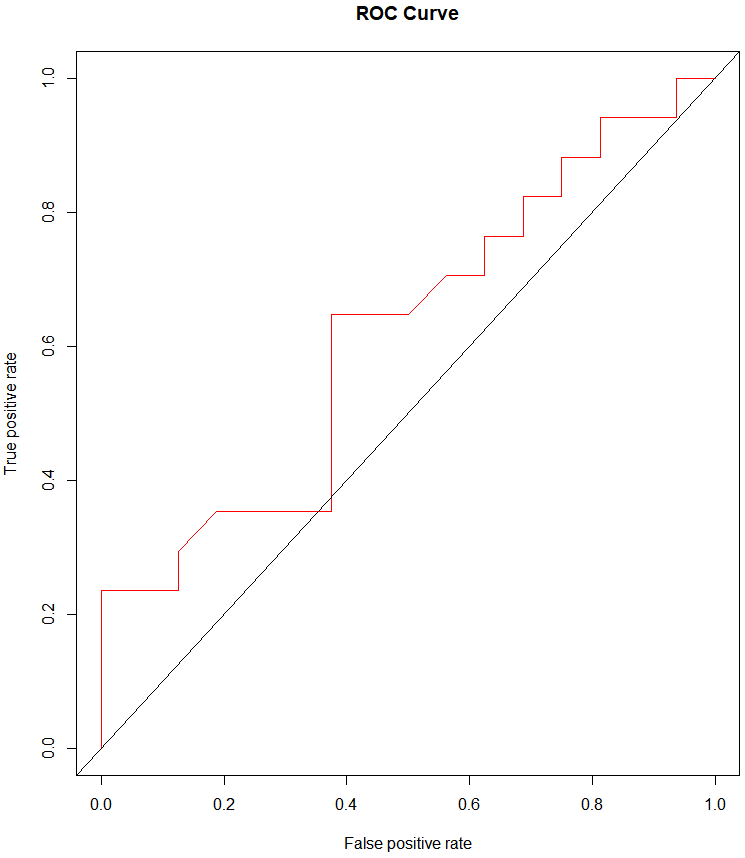


* 70%이상의 설명력을 가지기 위하여 주성분 7개 선택
* 73.8%의 설명력을 가진다.
* 변수차원 18에서 7으로 축소 시킨 후 랜덤포레스트 실시

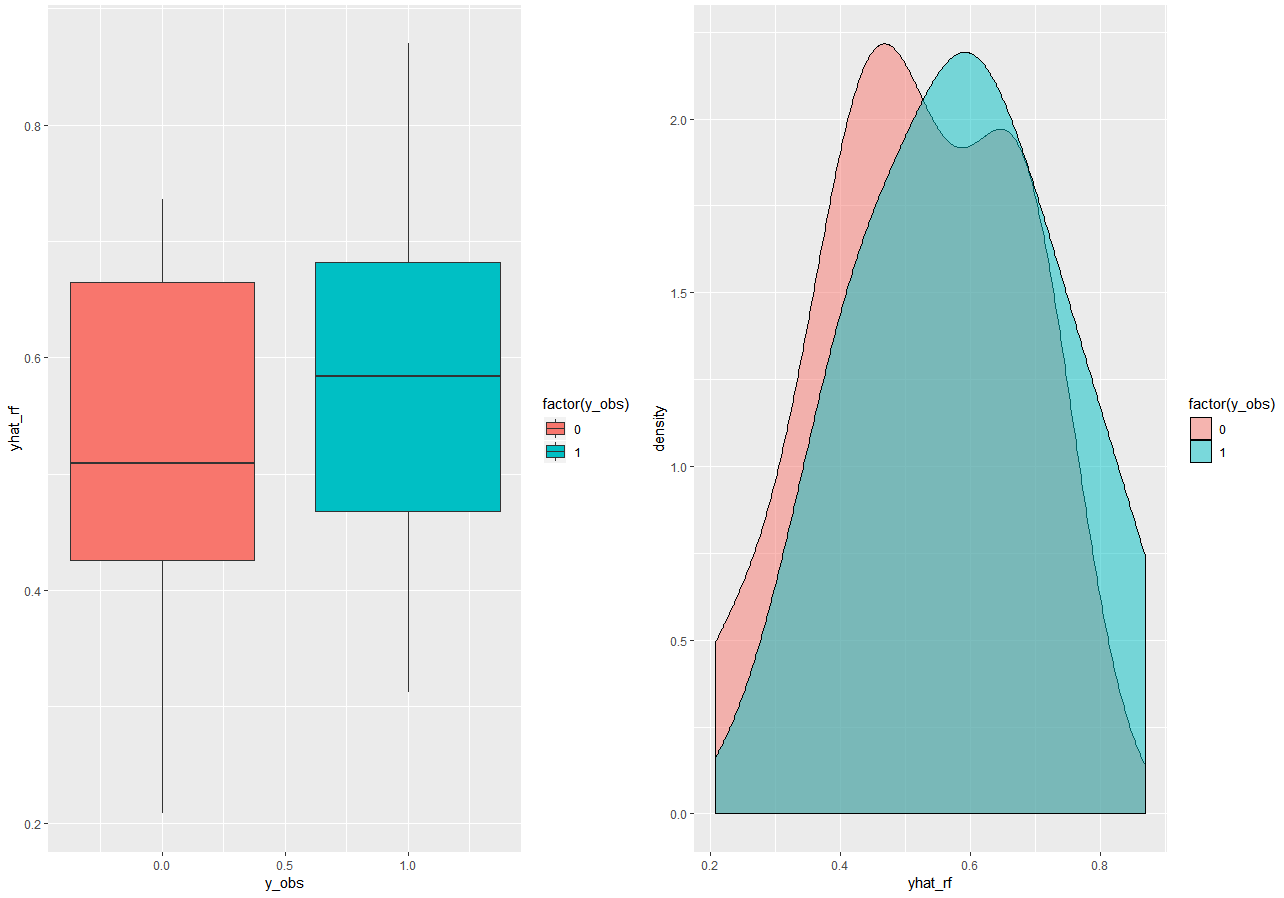
*[랜덤포레스트]*



* 설명변수가 모두 수치형 자료라서 그런지 오차 변동이 너무 컸다…
* 트리수 500개로 정하고 랜덤포레스트 실시

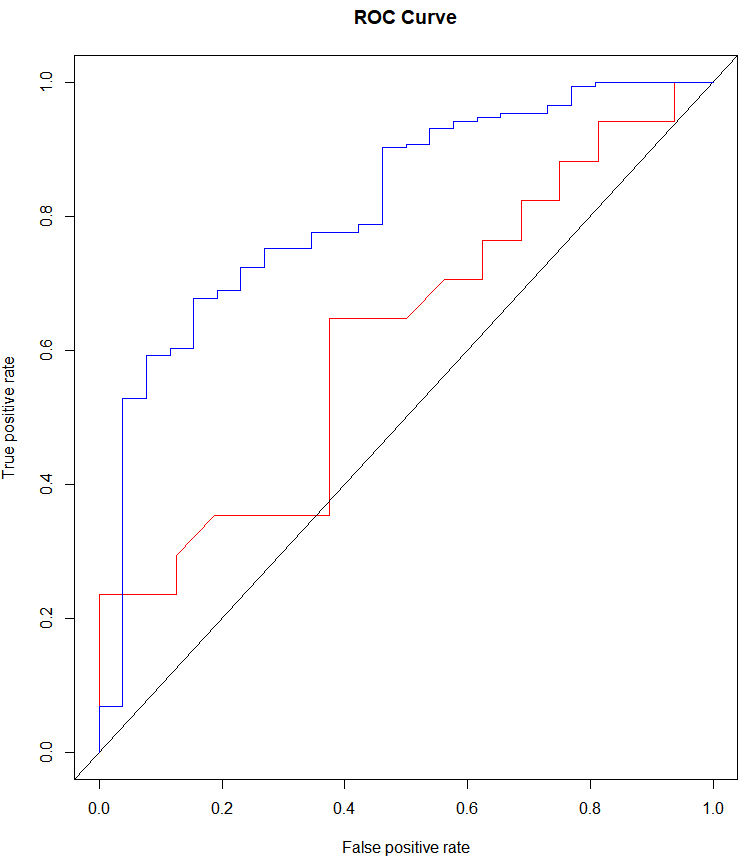


* AUC는 0.62정도로 나온다.
* 좋은 모델 아니다.

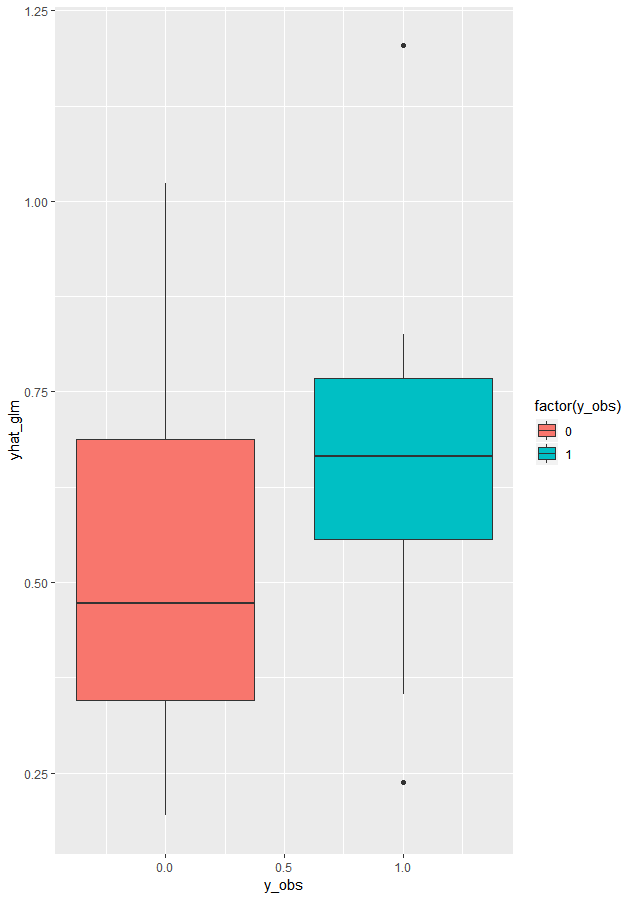


* 첫번째 그래프에서 실제값이 0일 때와 1일 때의 예측값의 상자그림이 차이가 적다.
* 두번째 그래프에서 실제값이 0일 때와 1일 때의 밀도함수 차이가 확실하게 나타나지 않는다.

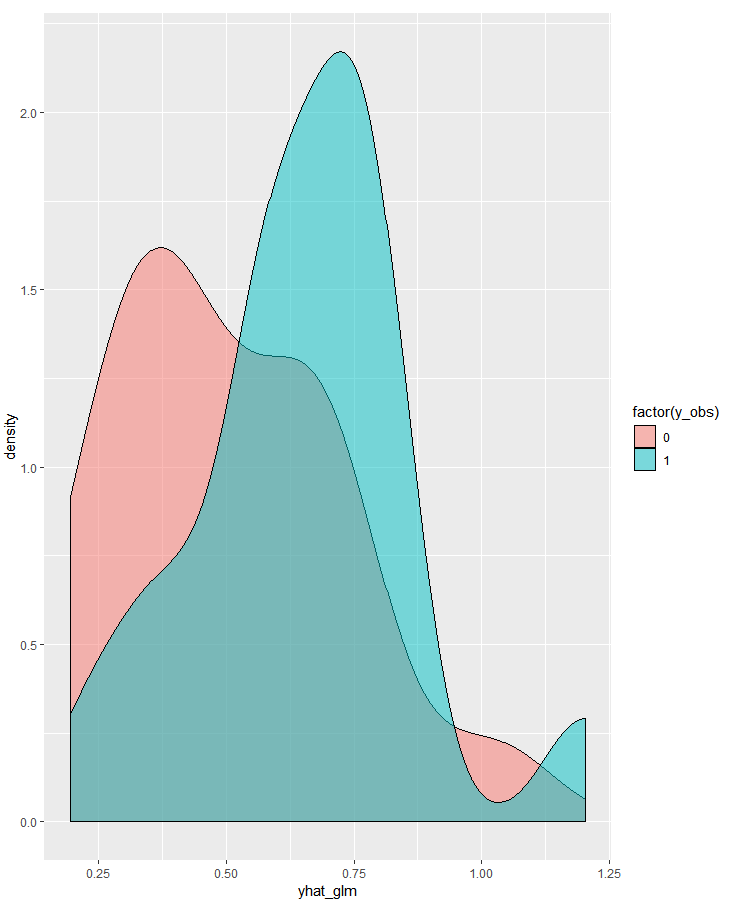
*[GLM]*



* AUC 값이 약 0.82가 나온다.
* 랜덤포레스트보다 좋은 모형

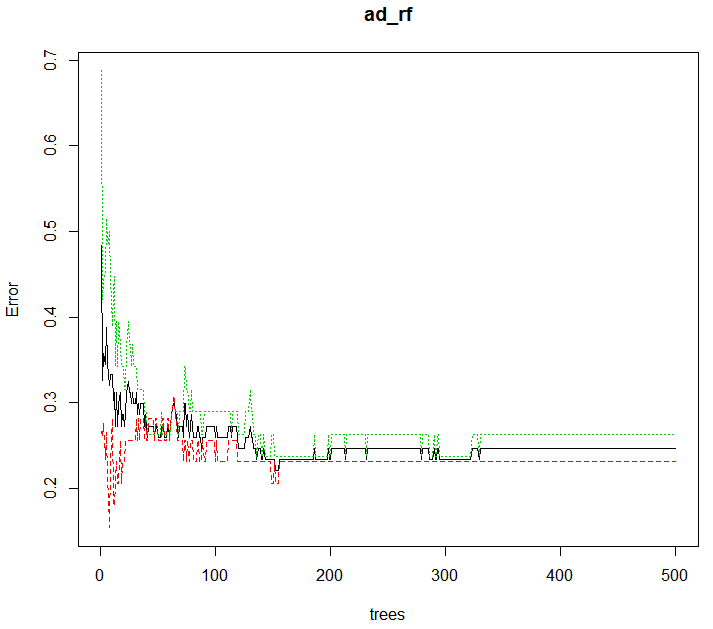
*[실제값과 예측값 간의 관계 살펴보기]*

* 분포가 확연하게는 아니지만 랜덤포레스트 모형과 비교하면 다르다.

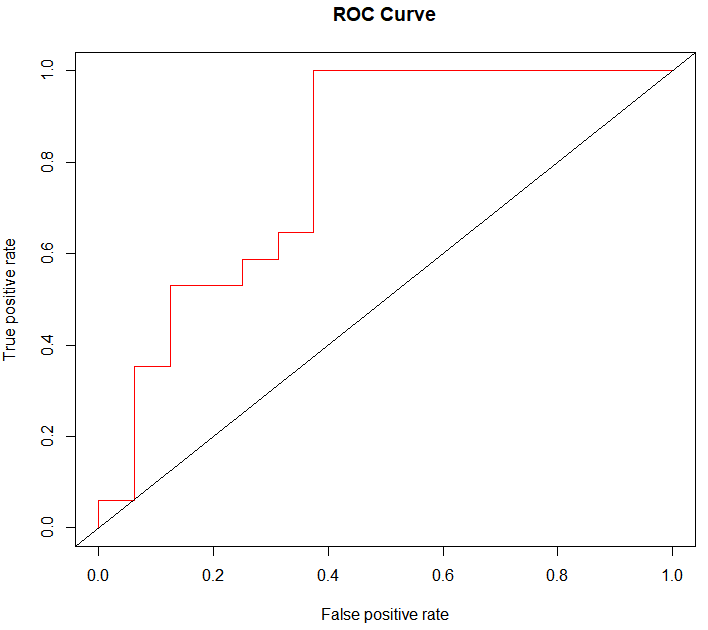


*[변수 축소 전과 비교]*

*랜덤포레스트*

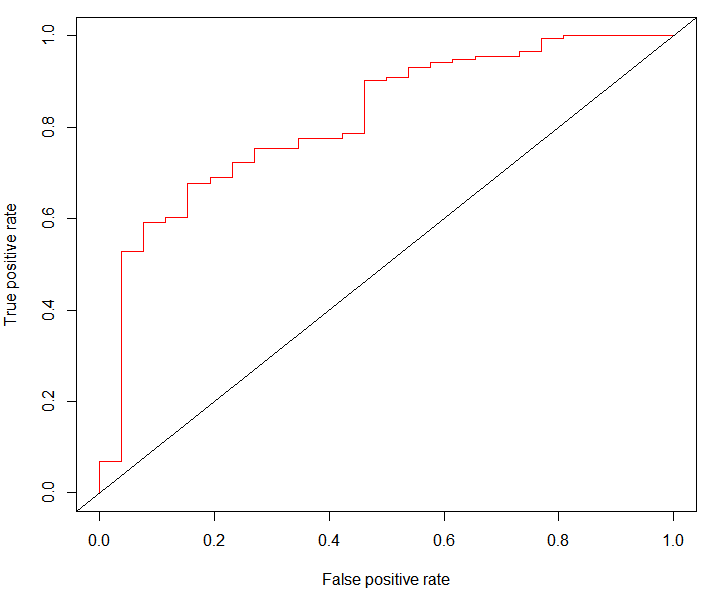


* 오차 변동이 변수 축소한 후보다 더 안정적이다.



* AUC가 0.79

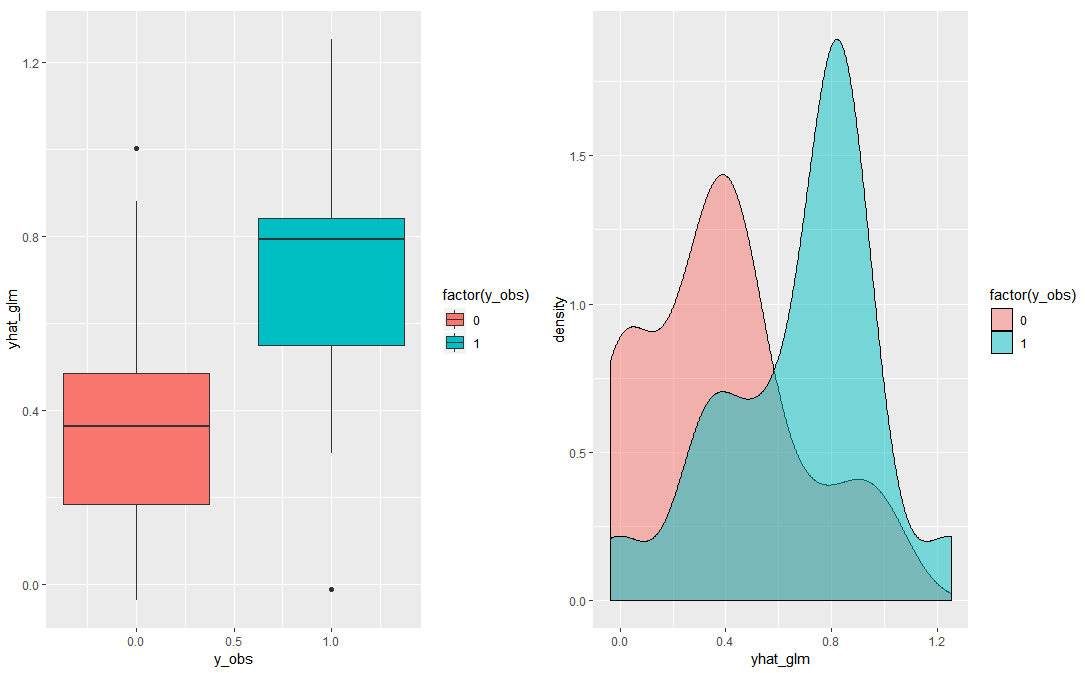
랜덤포레스트 모형에서는 변수 축소의 영향이 컸다.

*glm*

-> AUC 변수 축소 전과 후 같다.

0.82

-> ROC곡선도 완전히 똑같이 나와서 이상해서 그래프 더 그려봄



아까 변수 축소 후 그래프와 다르게 나옴

AUC 와 ROC 곡선은 완전히 일치하지만 변수 축소 전이 0과 1사이의 분포 차이가 더 크다.