AIR Regression HW

201811526 이은주

1. Cross-validation(교차검증)이 무엇인지, 그리고 k-fold교차검증은 어떤 방식으로 진행되는지 간단하게 서술

#Cross-validation(교차검증)

:교차검증이란, 테스트 데이터를 어떻게 골라내느냐에 따라 모형의 성능이 달라지게 되는데 한 개의 테스트 데이터만 사용하는 것이 아니라 각기 다른 방법으로 서로 다른 테스트 데이터를 여러번 골라내서 복수의 테스트를 실시하는 것이다. 이 교차 검증을 통한 모형 성능은 보통 오차 평균 또는 오차 분산으로 나타난다.

#k-fold 교차검증

:k개의 fold를 만들어서 진행하는 교차검증이다. training을 k개의 fold로 나누어 한 개의 fold에 있는 데이터를 다시 k개로 쪼갠 다음, k-1개는 training data, 마지막 한 개는 validation data set으로 지정한다. 이후 모델을 생성하고 예측을 진행하여 이에 대한 에러값을 추출한다. 다음 fold에서는 validation셋을 바꿔서 지정하고 이전 fold에서 validation 역할을 했던 set은 다시 training set으로 활용한다. 이를 k번 반복하여 각각의 fold의 시도에서 기록된 에러를 바탕으로 최적의 모델을 찾는다. 해당 모델을 바탕으로 전체 training set의 학습을 진행하고 해당 모델을 처음에 분할했던 test set을 활용하여 평가한다. 이러한 과정을 통해 데이터 셋에 대하여 정확도를 향상시킬 수 있다.

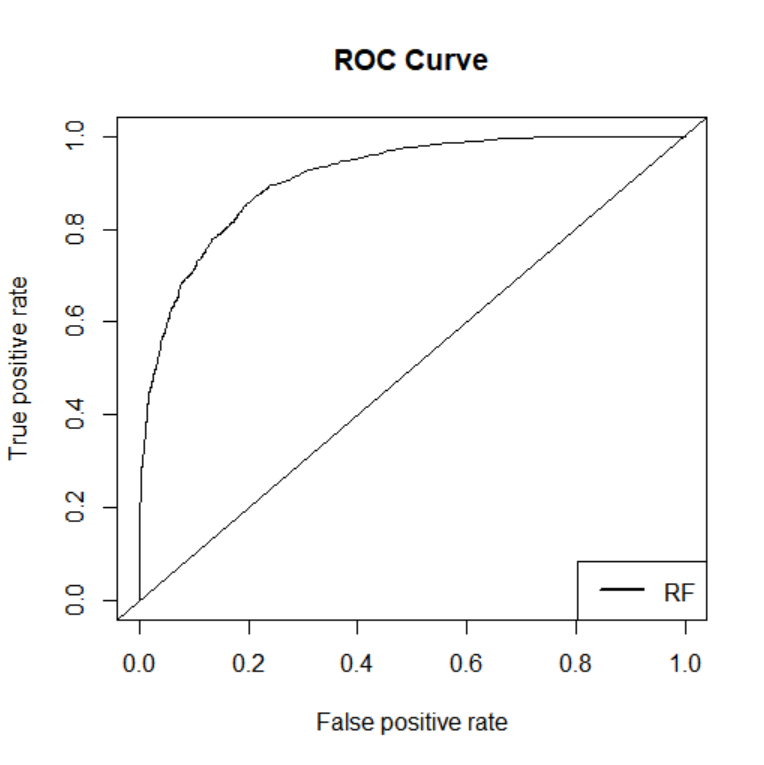
2. 의사결정나무, 랜덤포레스트를  validation에 적합한 결과를 아래 표에 캡쳐해서 첨부할 것.

|  |  |  |
| --- | --- | --- |
|  | Validation의 AUC 값 | Validation의 ROC곡선 |
| 의사결정나무 | 0.7378385 |  |
| 랜덤포레스트 | 0.914292 |  |

3.validation의 AUC값이 큰 모델을 선택해서 test에 적합 시키고, 그때의 AUC와 ROC곡선을 첨부할 것.

: 랜덤포레스트

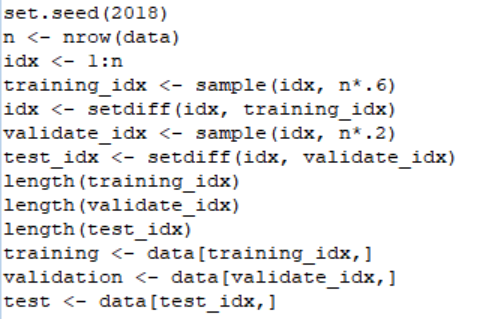




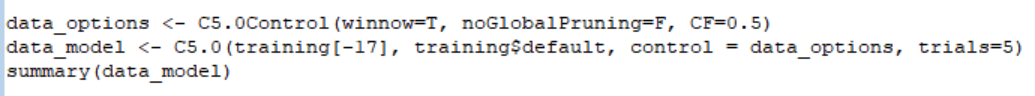
4. 모델링을 하는 코드를 캡처해서 올릴 것.

#의사결정나무

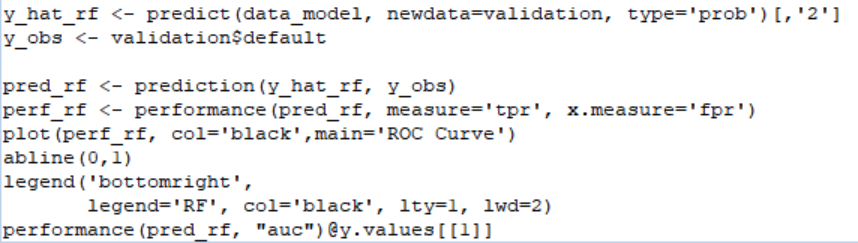
1)Sampling



2)모델링



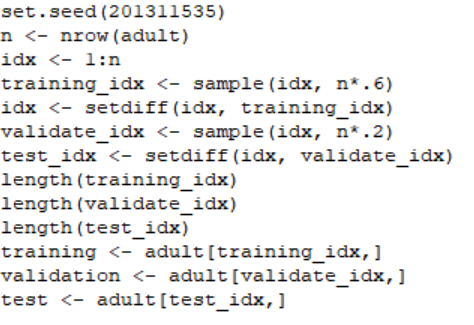
3)AUC와 ROC곡선



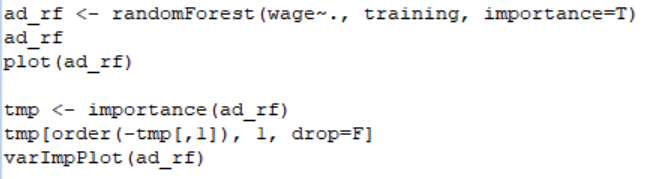
-> 결과는 위의 표에 작성

#랜덤포레스트

1)Sampling



2)모델링



-> 결과는 위의 표에 작성