2주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 R markdown으로 HTML, PDF 모두 가능합니다. **.R이나 .ipynb 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 5시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

Warnings가 뜨는 경우 R Markdown에서 warning=FALSE로 설정해서 뜨지 않게 해주세요.

Chapter 1 모델링을 위한 데이터 전처리

이번주는 본격적으로 모델링을 해보겠습니다. 이번 패키지를 통해 기본적인 모델링의 프로세스를 익히는 것이 목표입니다. 지도학습에는 예측하는 자료의 형태에 따라 분류/회귀로 나뉘는데 이번 데이터는 Y값이 연속형값인 회귀모델링을 해보겠습니다. 저번주에 배웠던 전처리 방법을 복습하며 EDA를 통해 간단히 데이터를 살펴보고, 모델링을 본격적으로 하기 전 데이터를 전처리해봅시다.

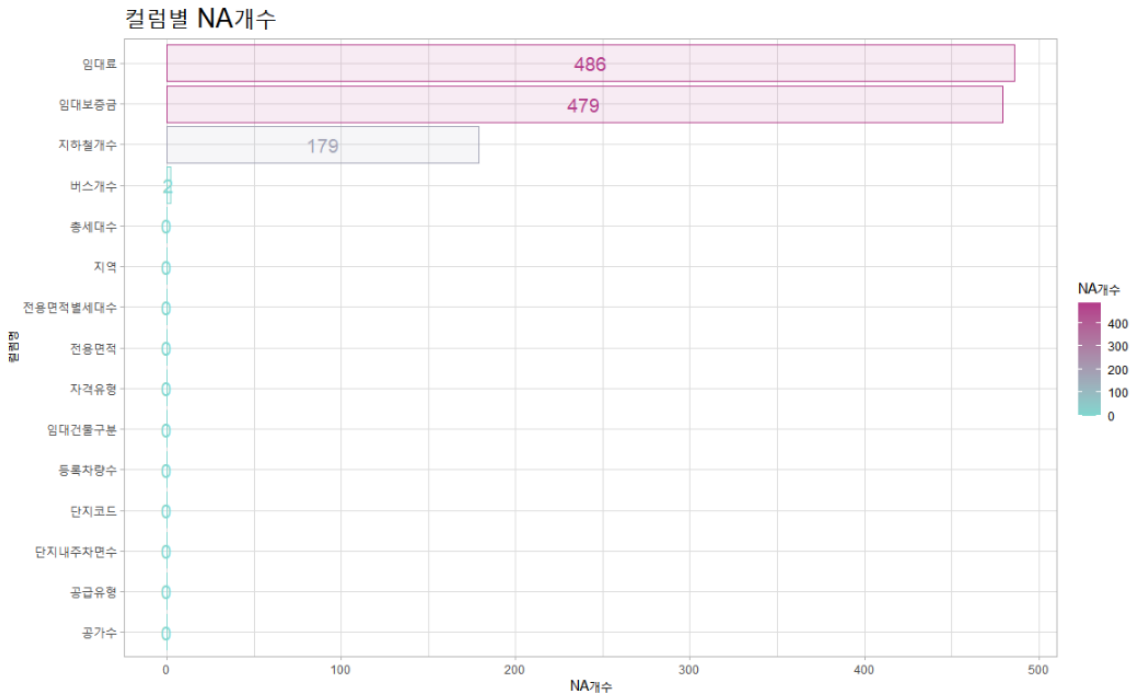
문제0. (기본 세팅) 0번 txt파일을 실행하세요. (패키지 불러오기, 디렉토리 설정)

문제1. Train데이터를 불러온 뒤 기본 구조를 파악하고 데이터 개수, 변수 개수, 데이터 형식을 확인해보세요.

문제2. 각 데이터의 컬럼명을 확인해보세요. '도보 10분거리 내 지하철역 수(환승노선 수 반영)' 컬럼이름과 '도보 10분거리 내 버스정류장 수' 컬럼이름이 너무 길기에 각각 '지하철개수', '버스개수'로 컬럼명을 바꾸어주세요.

문제3. 각 데이터에 '임대료', '임대보증금'이 문자형식으로 되어있습니다. 이유를 찾아 수치형으로 바꾸어 주세요. (HINT : NA 형식 확인)

문제4. 열별로 NA 개수를 확인해보세요. 확인 한 후 다음과 같이 시각화 해보세요.



- 변수별 결측치 개수에 데이터프레임을 만들면 편리합니다.
- 크기 순으로 정렬해주세요.
- Bar graph의 색상은 결측치 개수가 큰 쪽이 #B43C8A, 작은 쪽은 #81D8D0이고, 투명도는 0.1입니다.
- theme는 light입니다.

문제5. 데이터에서 범주형변수를 Factor변수로, 정수형변수를 수치형(Numeric)으로 바꾸어주세요.

문제6. NA값이 있는 행을 확인한 후 NA값을 열 별 평균으로 대체해주세요.

문제7. 공급유형이 '장기전세'인 경우 임대료가 0입니다. 데이터가 잘못되어 있는 경우 확인하고 고쳐주세요.

문제8. 면적당 임대료 면적당 임대보증금을 계산하여 파생변수를 만들어주세요.

문제9. 임대료, 임대보증금, 단지코드는 모델링에 사용하지 않을 예정입니다. 삭제해주세요.

Chapter 2 랜덤포레스트 및 교차검증

대부분의 예측 모델링 데이터의 경우 Test셋의 타겟 값이 주어지지 않기에, Train셋만 사용하는 경우 Train셋에 과적합되는 문제가 발생할 수도 있습니다. 과적합을 방지하기 위해서 Hold-out Validation 또는 K-fold Cross Validation (CV) 등이 사용되는데, 이 2가지 방법을 통하여 예측 모델링을 하고 비교해보려합니다.

이번에 사용할 모델은 랜덤포레스트입니다. 랜덤 포레스트는 앙상블 학습 방법의 일종으로, 다수의 결정트리를 사용하여 동작하는 모델입니다. 랜덤포레스트와 같이 모델의 하이퍼파라미터가 많은 경우 최적의 하이퍼파라미터 조합을 찾는 것이 필요한데, 이 과정을 '튜닝'이라고 합니다. 튜닝에는 여러가지 방법이 있지만 이번에는 그리드 서치를 통해 모델링을 해보겠습니다.

[Hold-out]

문제1. 데이터를 층화추출을 사용하여 Train셋와 Validation셋을 7:3 비율로 나누어주세요. (Seed : 2728 / p = 0.7)

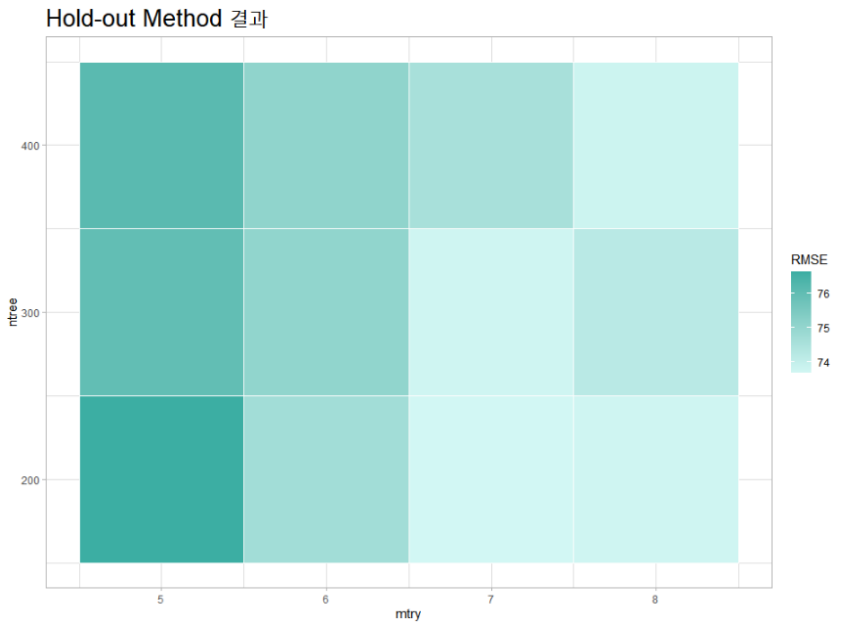
문제2. 랜덤포레스트의 하이퍼파라미터에 대해서 간단히 적어주세요.

문제3. 그리드서치를 위해 다음과 같이 데이터 프레임을 만들어 주세요. (For문을 통해 모델링을 진행할 과정입니다.) (HINT : expand.grid 함수)

문제4. For문을 활용하여 등록차량수를 예측하는 랜덤포레스트 모델링을 진행한 후 Validation셋의 RMSE를 계산하여 앞서 만들었던 표에 RMSE값을 넣으세요.

문제5. (기존필수/신입선택) 결과를 다음과 같이 시각화하고, 간단히 해석해보세요.

	mtry	ntree	RMSE
1	5	200	NA
2	6	200	NA
3	7	200	NA
4	8	200	NA
5	5	300	NA
6	6	300	NA
7	7	300	NA
8	8	300	NA
9	5	400	NA
10	6	400	NA
11	7	400	NA
12	8	400	NA



- 색상은 큰 쪽이 #3CAEA3, 작은 쪽은 #D2F7F4입니다.

- theme는 light입니다.

문제6. RMSE가 가장 낮은 하이퍼파라미터 조합을 출력하세요.

[5-fold CV]

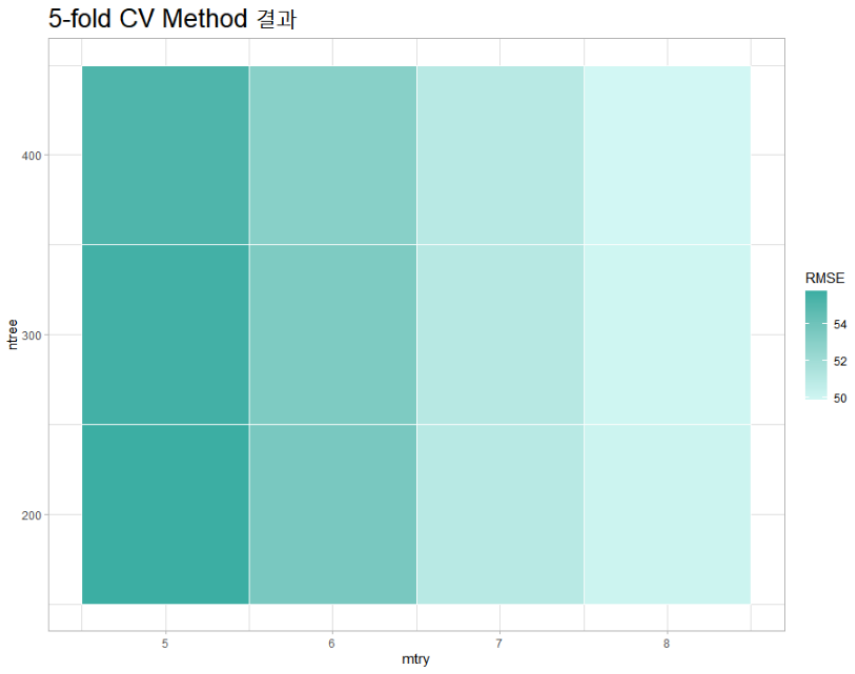
문제7. 5-fold 교차검증을 위해 층화추출을 사용하여 CV인덱스를 만들어주세요. (Seed : 2728)

문제8. 그리드서치를 위해 다음과 같이 데이터 프레임을 만들어 주세요. (For문을 통해 모델링을 진행할 과정입니다.) (HINT : expand.grid 함수)

문제9. 이중 For문을 활용하여 등록차량수를 예측하는 랜덤포레스트 모델링을 진행한 후 Validation셋의 RMSE를 계산하여 앞서 만들었던 표에 RMSE를 넣으세요. (HINT : 첫번째 For문 – 하이퍼파라미터 변경 / 두번째 For문 – Val셋 변경)

	mtry	ntree	RMSE
1	5	200	NA
2	6	200	NA
3	7	200	NA
4	8	200	NA
5	5	300	NA
6	6	300	NA
7	7	300	NA
8	8	300	NA
9	5	400	NA
10	6	400	NA
11	7	400	NA
12	8	400	NA

문제10. (기존필수/신입선택) 결과를 다음과 같이 시각화하고, 간단히 해석해보세요.

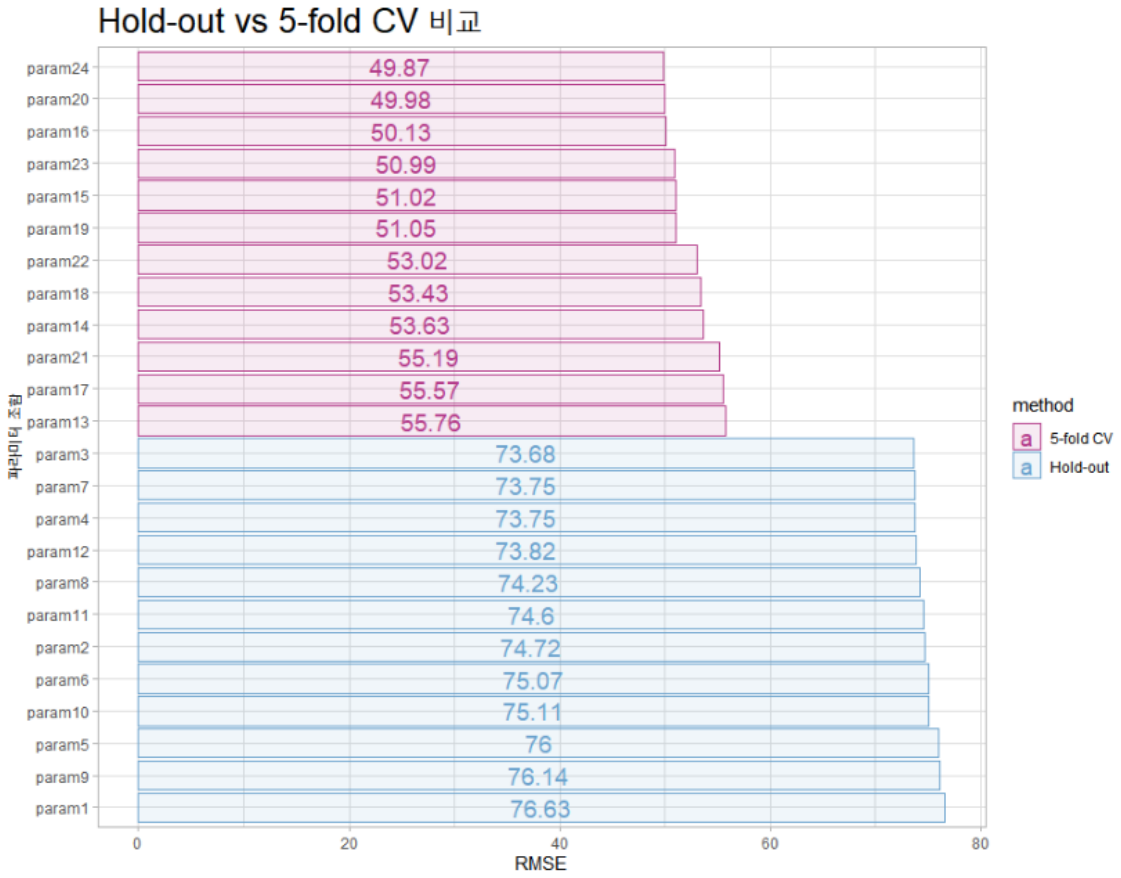


- 색상은 큰 쪽이 #3CAEA3, 작은 쪽은 #D2F7F4입니다.
- theme는 light입니다.

문제11. RMSE가 가장 낮은 하이퍼파라미터 조합을 출력하세요.

[결과비교]

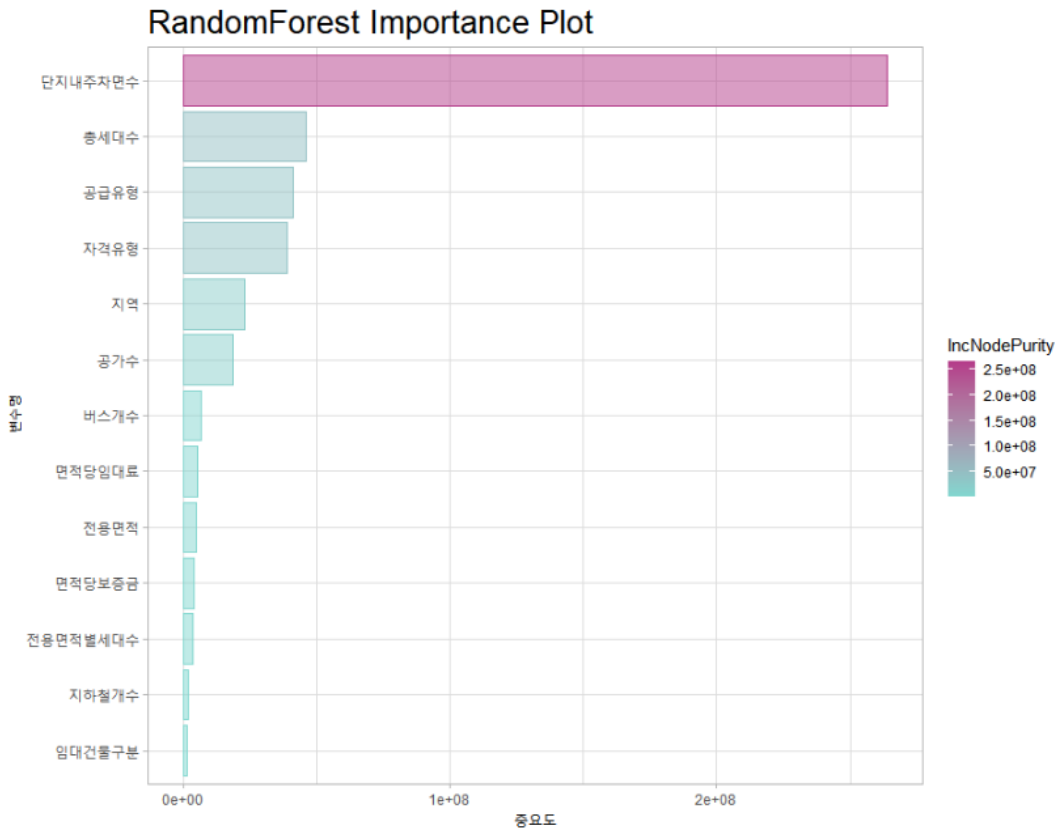
문제12. 앞에서 Hold-out을 사용하여 튜닝한 결과와 5-fold CV를 사용하여 튜닝한 결과를 다음과 같이 시각화 하고 해석해주세요.



- Hold-out색상은 #6AA2CD, 5-fold CV색상은 #B43C8A, alpha는 0.1입니다.
- theme는 light입니다.
- X축은 행번호에 'param'을 붙인 것 입니다. (param1 ~ param24까지 있음)
- 순서는 RMSE를 기준으로 정렬했습니다.

문제 13. 랜덤포레스트에서 Importance 계산이 어떻게 되는지 간단하게 적어주세요.

문제 14. 가장 좋게 나온 하이퍼파라미터 조합에 대하여 전체 Train 에 대하여 학습 후 Importance Plot 을 그린 후 해석해주세요.



- IncNodePurity사용
- 색상은 큰 쪽이 #B43C8A, 작은 쪽은 #81D8D0, alpha는 0.5입니다.
- theme는 light입니다.

Chapter3 Xgboost

이번에 사용할 모델은 XGBoost입니다. Xgboost는 대표적인 부스팅 모델 중 하나입니다. 앞에서는 그리드서치를 사용하여 튜닝을 했었지만, 이번에는 랜덤튜닝을 사용하여 튜닝을 해보겠습니다.

문제 1. Xgboost는 numeric 변수만 받으므로 범주형 변수를 encoding을 해야합니다. 범주형 변수들을 One-hot 인코딩해주세요.

문제 2. Xgboost의 하이퍼파라미터에 간단히 적어주세요.

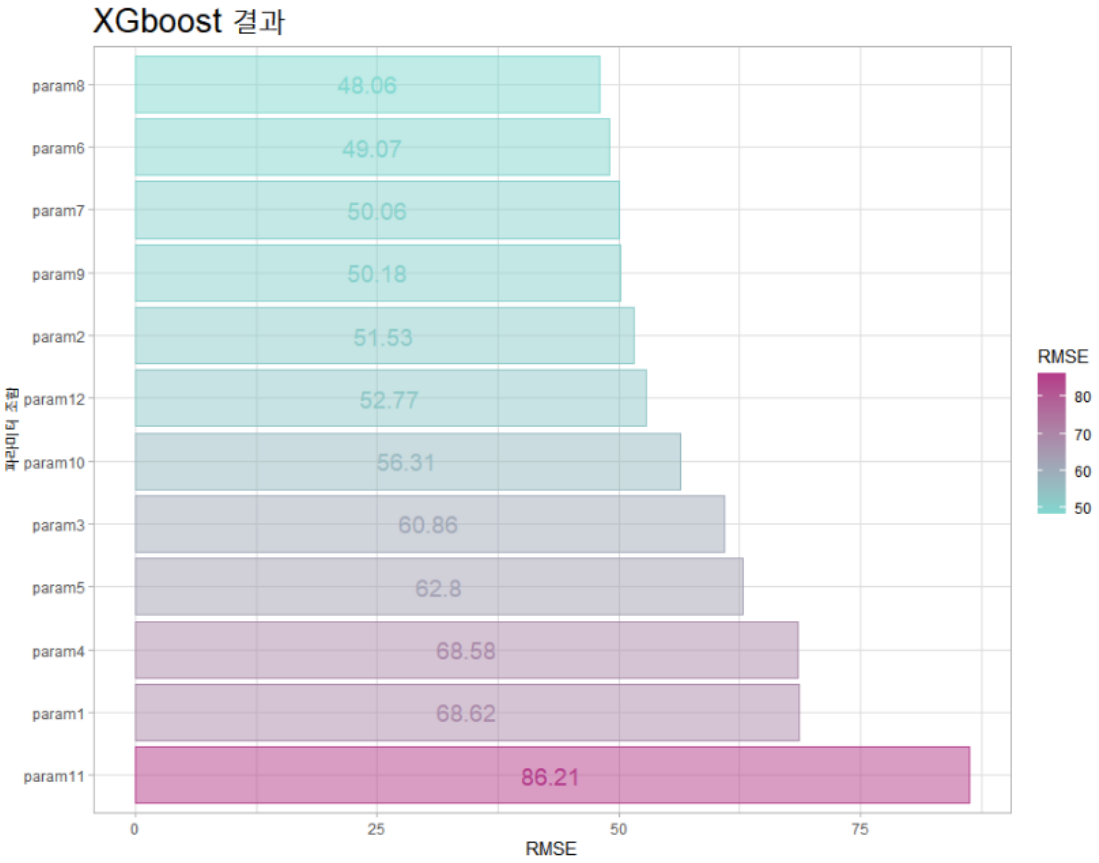
문제3. 랜덤튜닝을 활용하여 튜닝을 진행할 예정입니다. 아래의 튜닝데이터 범위를 참고하여 서치할 하이퍼파라미터를 골라주세요. (Seed : 2728)

- 랜덤튜닝수는 12번입니다.
- max_depth : 4~10
- min_child_weight : 4~10
- subsample : 0.5 ~ 1
- colsample_bytree : 0.5 ~ 1

문제4. 이중 For문을 활용하여 등록차량수를 예측하는 Xgboost 회귀모델링을 진행한 후 Validation의 RMSE를 계산해주세요. (HINT : 첫번째 For문 - 파라미터 변경 / 두번째 For문 - Val셋 변경)

- 5-fold CV 사용 (앞에서 뽑은 CV index 사용)
- 앞에서와 같이 서치할 파라미터를 데이터 프레임으로 만들면 편합니다.
- 랜덤튜닝수는 12번입니다.
- Seed : 2728
- eta : 0.01
- nrounds : 1000
- early_stopping_rounds : 0.05*nrounds

문제5. 결과를 다음과 같이 시각화하고 가장 좋은 하이퍼파라미터 조합을 보여주세요. 그리드서치와 랜덤서치의 개념, 그리고 랜덤서치가 가지는 장점과 단점을 설명해주세요.



- 색상은 큰 쪽이 #B43C8A, 작은 쪽은 #81D8D0, alpha는 0.5입니다.
- theme는 light입니다.
- X축은 행번호에 'param'을 붙인 것 입니다. (param1 ~ param12까지 있음)

Chapter4 비교

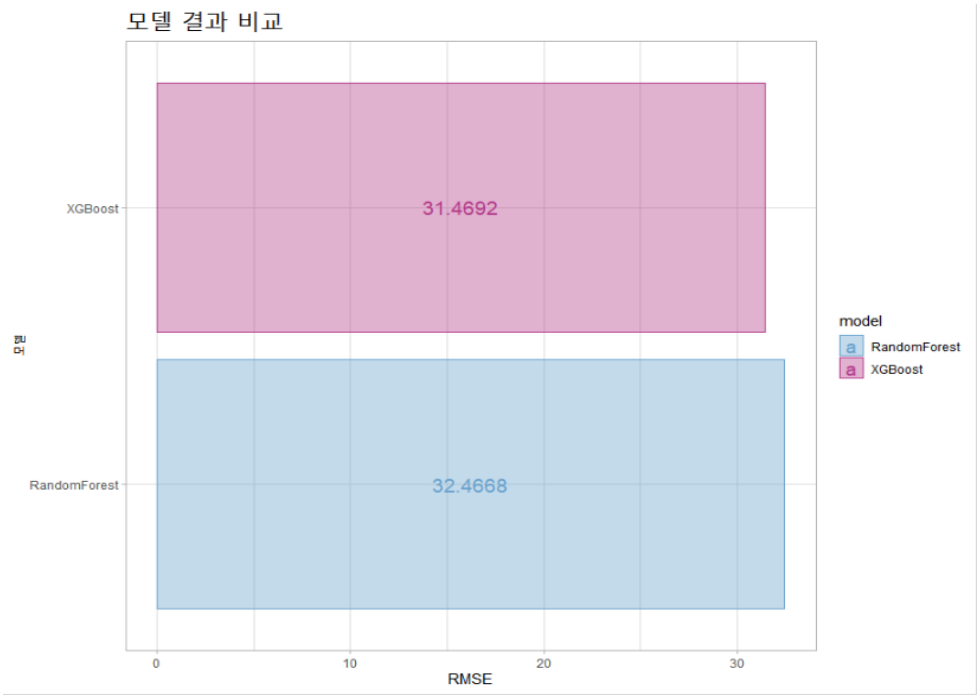
Test셋에 대하여 예측을 한 후 RMSE를 사용하여 성능을 평가한 후 간단히 모델들을 비교해봅시다. 앞에서 튜닝한 파라미터를 사용해주세요.

문제 1. Test셋을 불러와 Train셋과 똑같이 전처리 해주세요.

문제 2. RandomForest에서 가장 잘 나온 하이퍼파라미터 조합을 사용하여 전체 Train셋을 학습시킨 후 Test셋에 대한 RMSE를 계산하세요.

문제 3. Xgboost에서 가장 잘 나온 하이퍼파라미터 조합을 사용하여 전체 Train셋을 학습시킨 후 Test셋에 대한 RMSE를 계산하세요.

문제 4. 2개의 모델링 결과를 다음과 같이 시각화한 후 해석해보세요.



- RandomForest색상은 #6AA2CD, Xgboost색상은 #B43C8A, alpha는 0.4입니다.
- Theme는 light입니다.