강우 예측

머신 러닝 모델링

AI_17_오형석

CONTENTS

1. 프로젝트 개요 (데이터 선정)

- 시나리오(문제 정의)

- 전세계적으로 코로나19 감소, 여행에 대한 수요가 증가할 것으로 예상

- 겨울에도 따뜻한 나라에서 액티비티 여행을 원하는 고객들을 위해
  계절이 반대인 호주에 새로운 여행 상품을 기획

- 액티비티 활동이 중요한 상품이라서 강우 여부에 따라
  상품 철회가 발생할 수 있어 사측의 수익이 줄어들고 고객과의 신뢰성이 하락할 수 있다.

- 상품을 기획할 호주의 강우 여부를 예측하는 모델을 통해 상품 스케줄을 기획한다면
  액티비티 상품의 취소,환불이 줄어들 것으로 예상

1. 프로젝트 개요 (목표 및 가설 설정)

프로젝트 목표

- 수집한 데이터를 통해 강우 여부를 예측하는 머신러닝 모델을 완성

가설 설정

1. 습도가 높으면 비가 내릴 확률이 높을 것이다
2. 여름 시즌이 비시즌(다른 계절)보다 비가 올 확률이 높을 것이다
3. 기온에 따라 비가 올 확률이 다를 것이다

2. 데이터 전처리 (EDA)

- 데이터셋 정보

- 출처 :캐글(https://www.kaggle.com/)
- 원본 출처 : 호주기상청 http://www.bom.gov.au/climate/data/

- 주요 Feature Engineering

- 결측치 비율이 30% 이상인 컬럼 drop
- 계절 구분을 위해 'Date' 컬럼을 분리,연도/월은 따로 컬럼 생성
- 'Summer' 컬럼 생성, 여름 시즌 (1), 비시즌(0) 저장 (호주의 여름은 12~2월)
- float64 타입의 결측치를 mean으로 대체. object 타입의 결측치는 최빈값으로 대체
- 분류 문제로 해결하기 위해 Boolean 값을 int타입의 0과 1로 변경
- 평균 기온, 습도, 바람속도, 대기압 계산하여 새로운 컬럼 생성 후 계산에 사용된 컬럼은 삭제

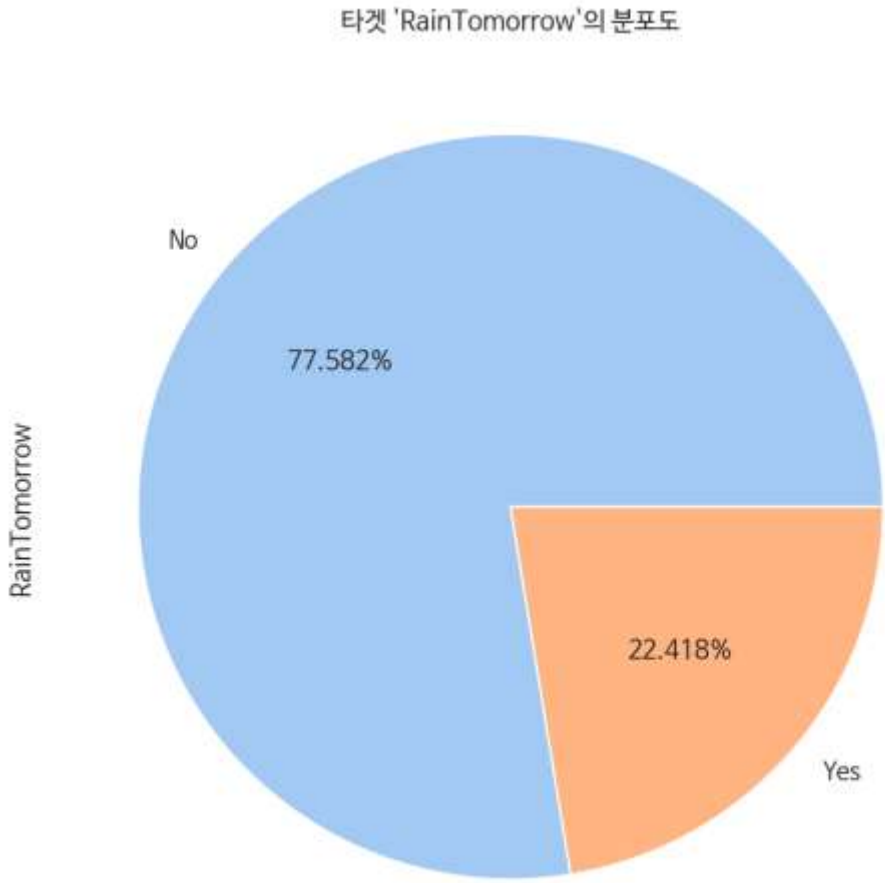
2. 데이터 전처리 (EDA)

- 타겟(target) = ‘RainTomorrow’

- 모델의 성능을 비교하기 위한 기준 모델 Baseline 설정

- 최빈값을 이용해 생성

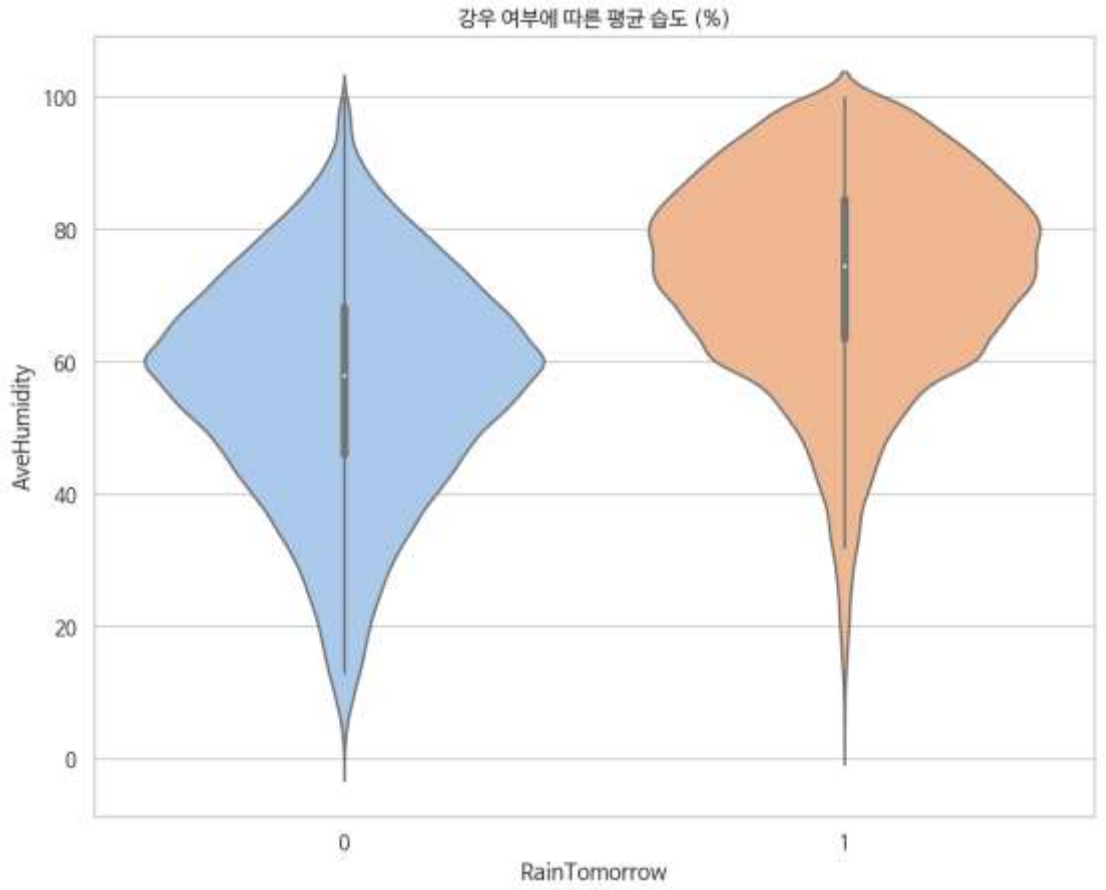
- 내일 비가 오지 않을 확률 약 77.6%



2. 데이터 전처리 (가설 검증)

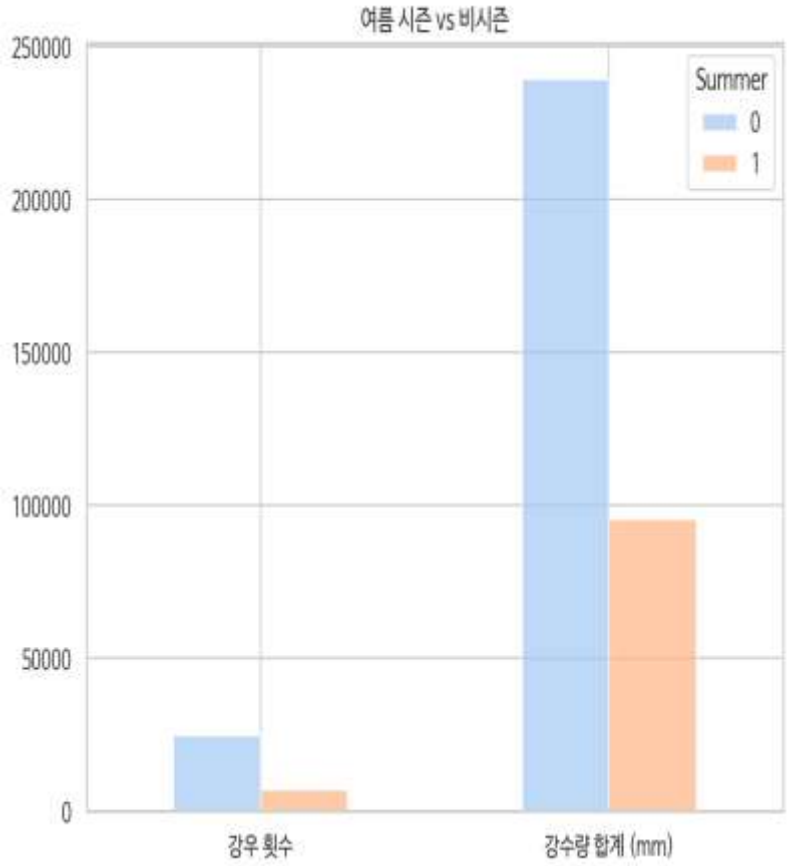
1. 습도가 높으면 비가 내릴 확률이 높을 것이다

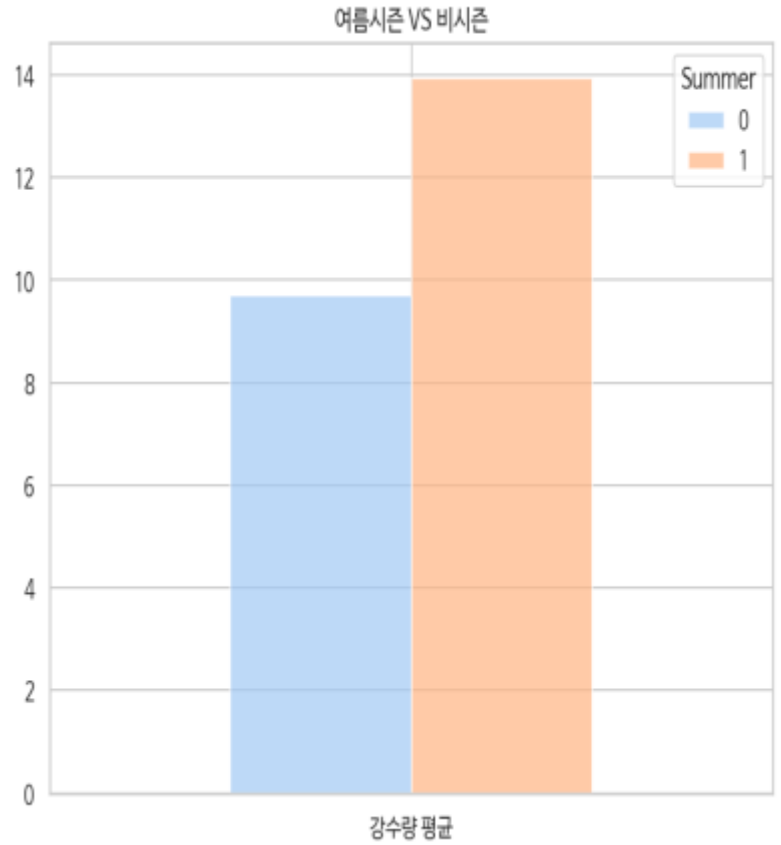
- 내일 비가 올 확률이 높을 경우 평균 습도가 높게 형성되어 있다
- 다만 습도가 높다고 해서 무조건 비가 오는 것은 아니다.

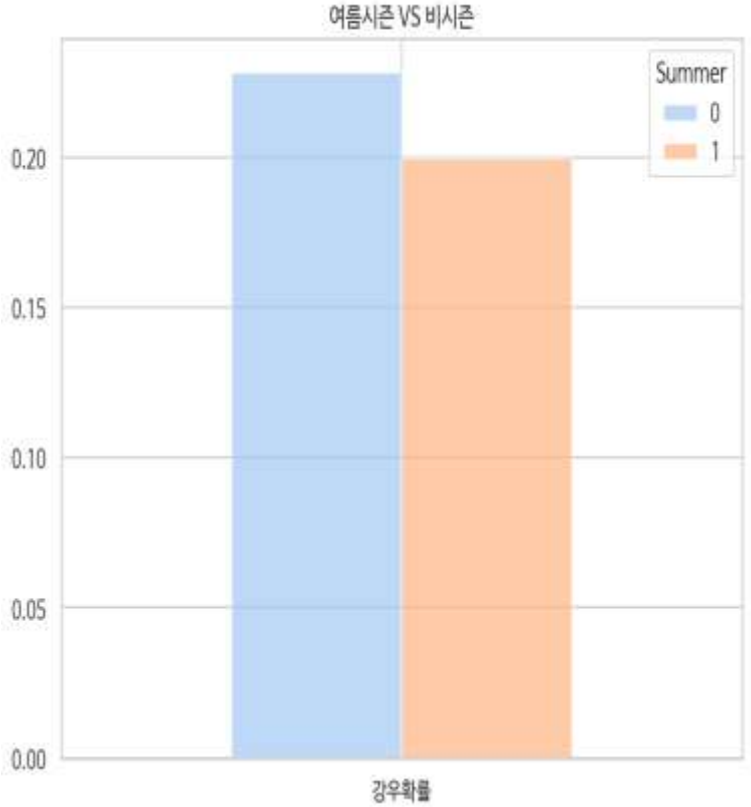


2. 데이터 전처리 (가설 검증)

2. 여름 시즌이 비시즌(다른 계절)보다 비가 올 확률이 높을 것이다

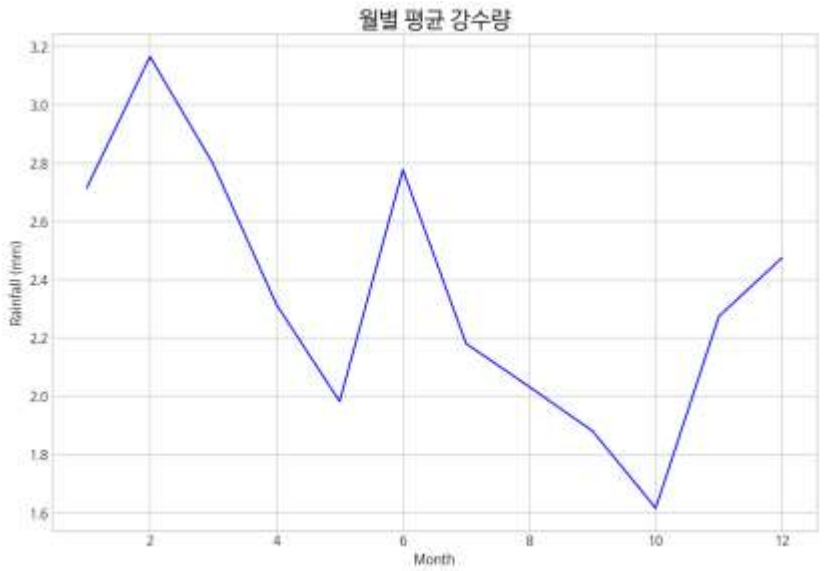


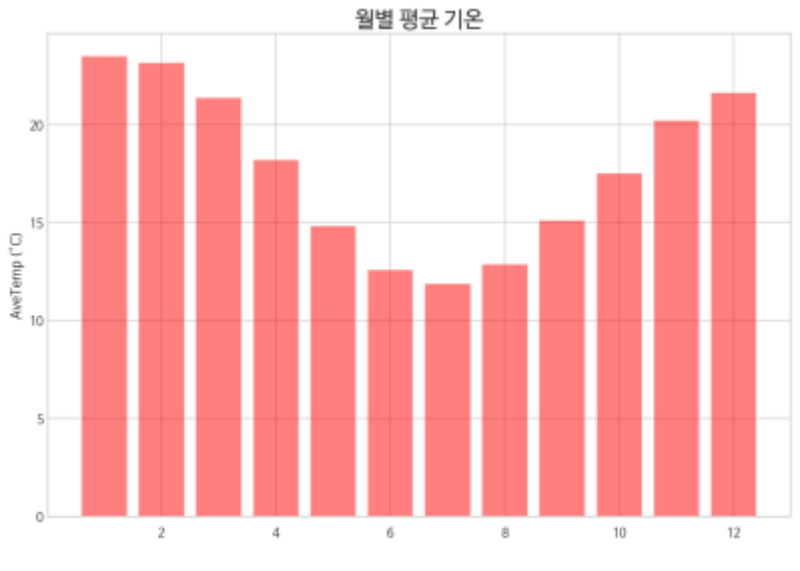


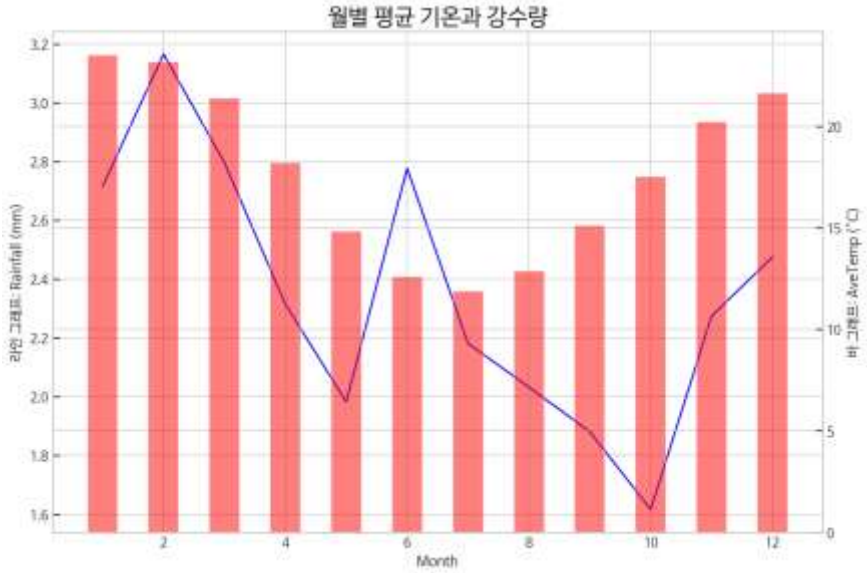


2. 데이터 전처리 (가설 검증)

3. 기온에 따라 비가 올 확률이 다를 것이다







- 계절 변화에 따라 평균 기온이 달라짐
- 계절 변화에 따라 강수량이 달라짐
- 평균기온과 강수량으로 비가 올 확률은 알 수 없었음

3. 모델링 (RandomForest)

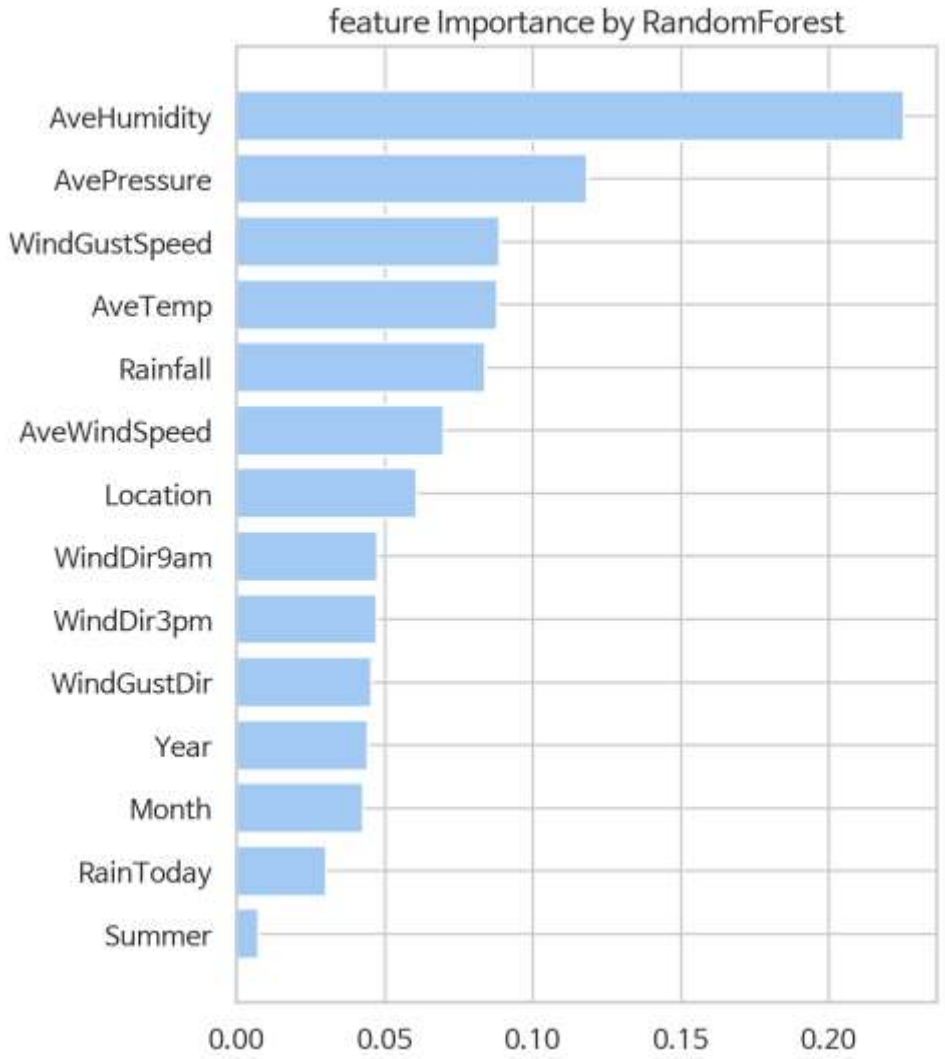
- 성능 평가

```
AUC score :  0.8603117914776812
              precision    recall  f1-score   support

           0       0.85      0.96      0.90     17651
           1       0.74      0.43      0.54      5100

    accuracy                           0.84     22751
   macro avg       0.80      0.69      0.72     22751
weighted avg       0.83      0.84      0.82     22751
```

정확도	0.84
정밀도	0.74
재현율	0.43
F1 score	0.54
AUC	0.8603



3. 모델링 (XGBoost)

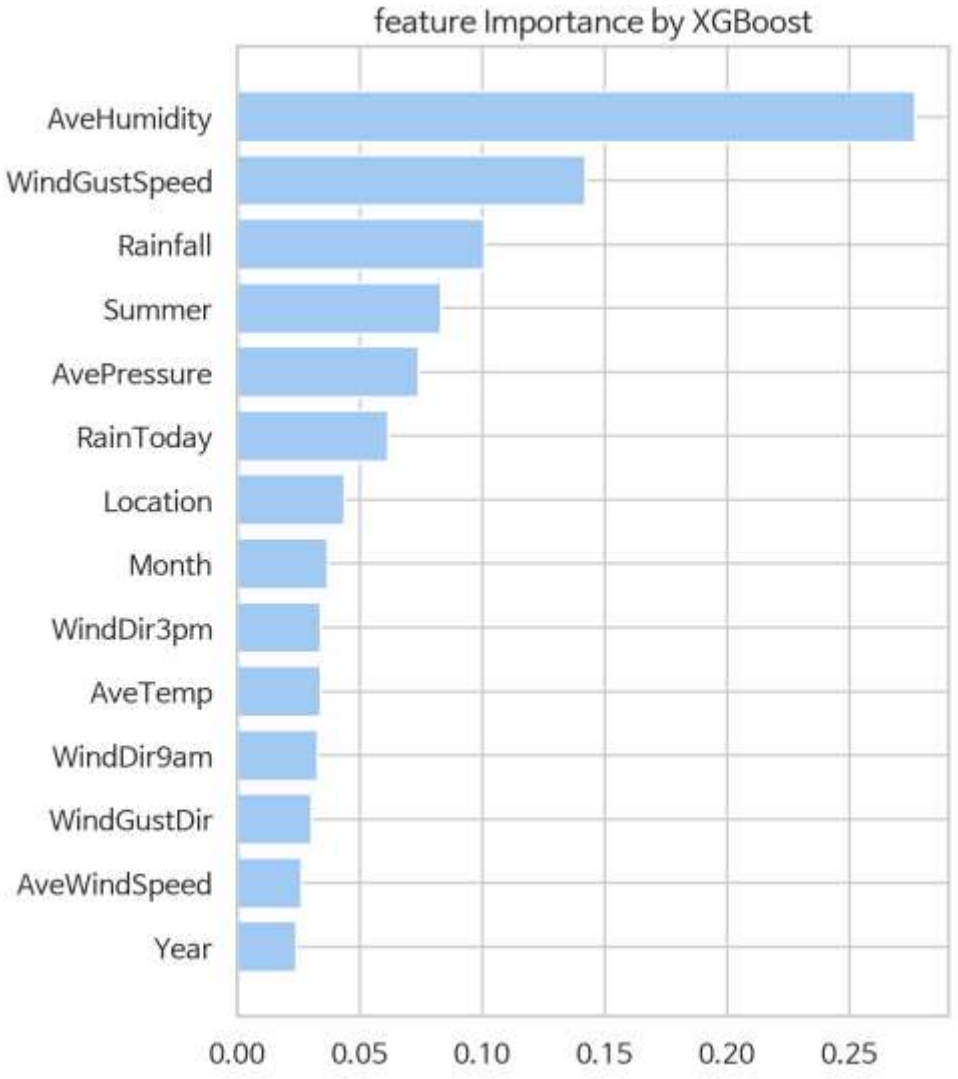
- 성능 평가

```
AUC score :  0.8685311002764939
              precision    recall  f1-score   support

           0       0.87      0.95      0.91     17651
           1       0.73      0.50      0.59      5100

    accuracy                           0.85     22751
   macro avg       0.80      0.72      0.75     22751
weighted avg       0.84      0.85      0.84     22751
```

정확도	0.85
정밀도	0.73
재현율	0.50
F1 score	0.59
AUC	0.8685



3. 모델링 (XGBoost – RandomizedSearchCV)

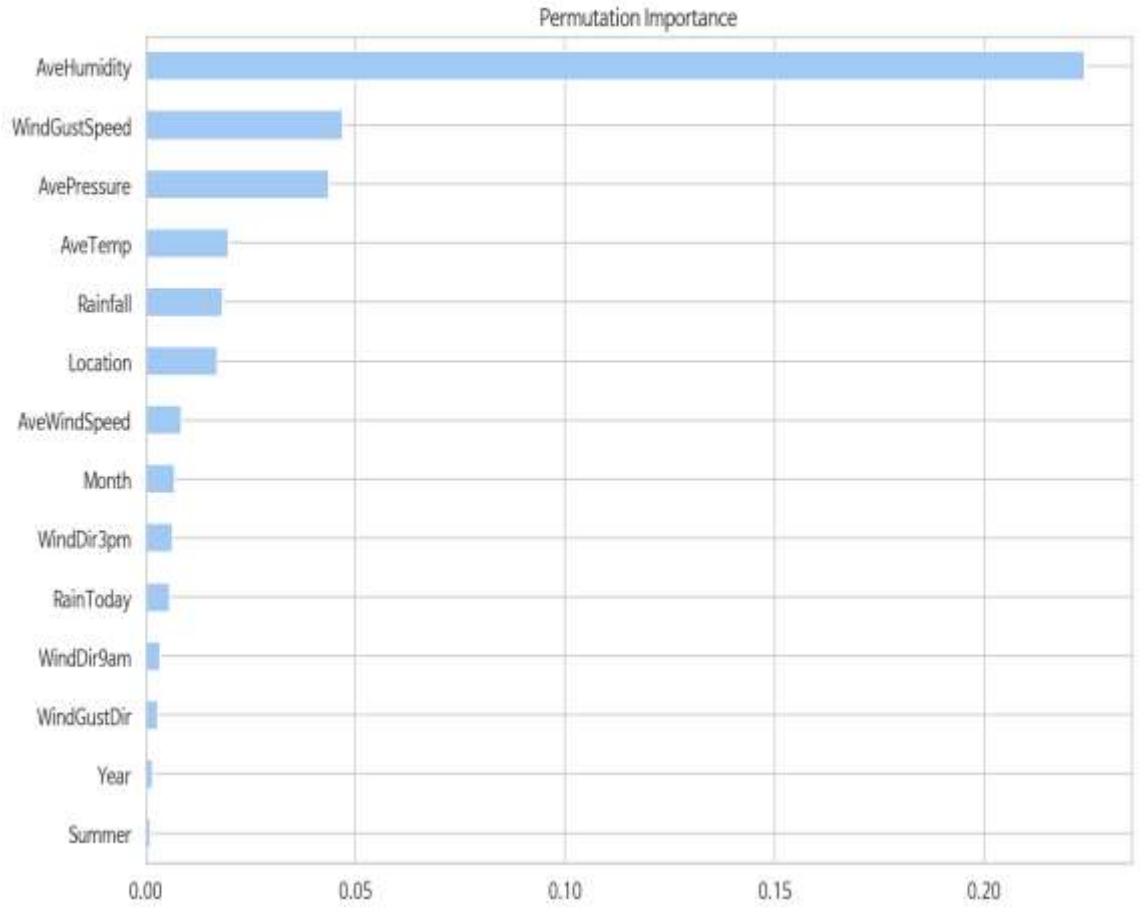
- 성능 평가
- RamdomizedSearchCV로 최적의 하이퍼 파라미터 탐색 후 다시 모델 학습

```
AUC score :  0.869698378473252
              precision    recall  f1-score   support

           0       0.87      0.95      0.91     17651
           1       0.73      0.50      0.59      5100

    accuracy                           0.85     22751
   macro avg       0.80      0.72      0.75     22751
weighted avg       0.84      0.85      0.84     22751
```

정확도	0.85
정밀도	0.73
재현율	0.50
F1 score	0.59
AUC	0.8696



4. 결론 (최종 모델)

최종 모델
Model : XGBoost Classifier

Hyper parameter tuning

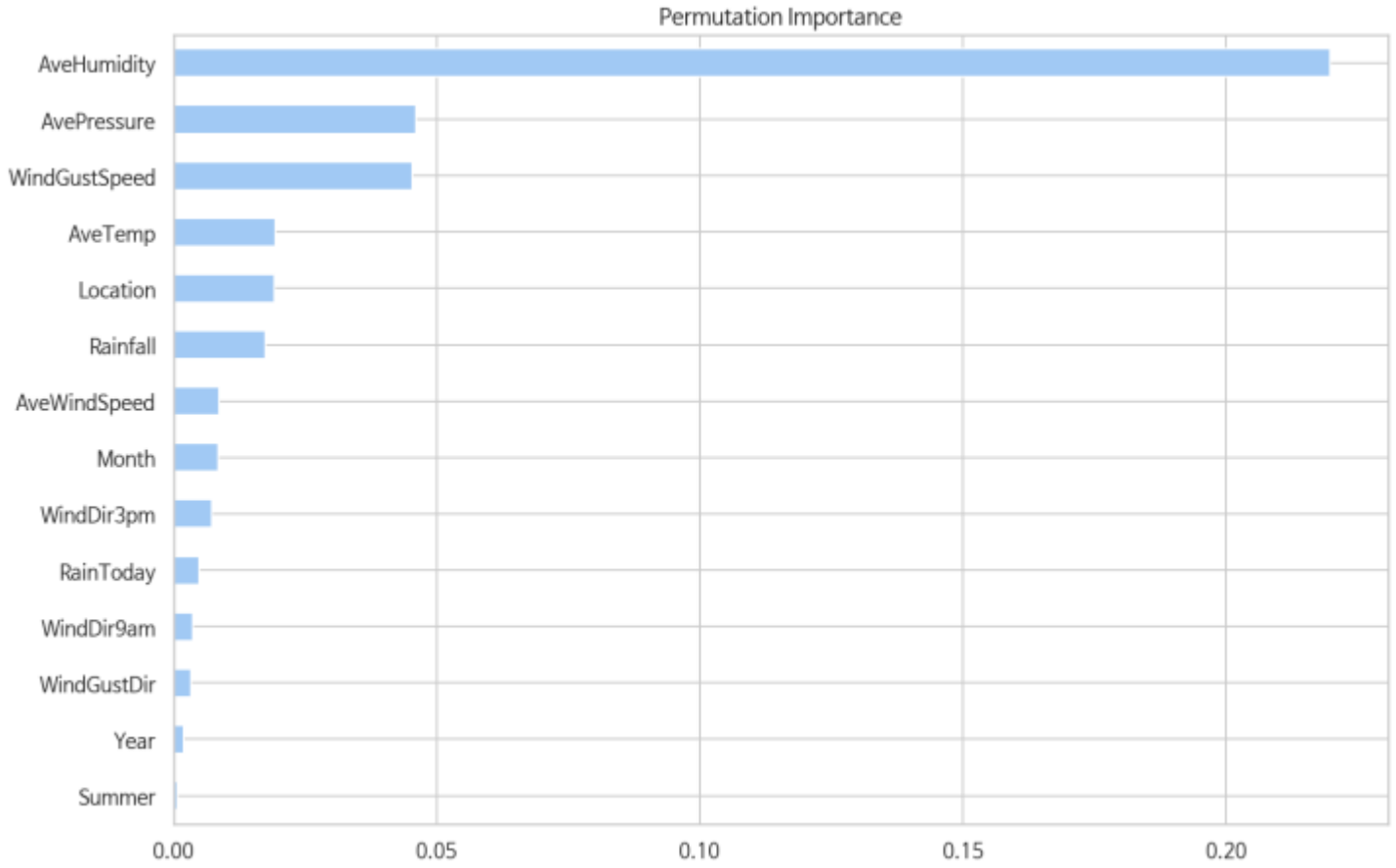
- n_estimators=2000
- min_child_weight=2
- max_depth=4
- learning_rate=0.2

AUC score : 0.8745186374042002

	precision	recall	f1-score	support
0	0.87	0.95	0.91	22064
1	0.73	0.52	0.61	6375
accuracy			0.85	28439
macro avg	0.80	0.73	0.76	28439
weighted avg	0.84	0.85	0.84	28439

정확도	0.85
정밀도	0.73
재현율	0.52
F1 score	0.61
AUC	0.8745

4. 결론 (모델 해석)



1. 평균 습도가 다른 특성에 비해 타겟값에 많은 영향을 준다.

2. 평균 습도, 평균대기압, 바람속도, 평균 기온 순으로 다음날 강우 여부에 영향을 주는 특성임을 확인할 수 있다.

3. 해당 모델 성능 평가와 시각화 자료만으로는 타겟과 특성간 음양 관계를 알 수 없다.
(e.g. 평균 대기압이 올라갈수록 강우 확률이 올라간다? -> 알 수 없음)

4. 결론 (한계)

- 추가적인 검증, 분석이 없어서 순열중요도로 알아낸 주요 특성들의 영향력을 정확히 파악x

- 호주는 굉장히 넓은 나라, 지역이 상위권의 중요 특성으로 나왔지만 영향력 알 수 없음

- 모델을 통해 다음 날 강우 여부를 예측할 수 있지만 설득력이 부족
(음양 관계를 알 수 없어 새로운 데이터가 들어왔을 때 예측을 잘할거라는 설득력 부족)

THANK YOU

AI_17_오형석