2021. 11. 11

MACHINE LEARNING PROJECT

Weather Forecast In Australia

It will be raining tomorrow?

한다운
dawun.han@gmail.com
DAWUNHAN @Github
CODESTATES
AI BOOTCAMP 7기

CONTENTS

Predict Raining tomorrow

날씨 데이터,
머신러닝
강우 가능성

프로젝트 개요

01. 데이터셋

선정 이유

Outdoor Activities가 많은 호주 특성 상 강우 여부는 매우 중요.

데이터 소개

날짜, 일조 시간, 바람 속도, 바람 방향, 습도, 대기압, 구름량, 온도, 오늘의 강우 여부

02. 프로젝트 목표

목표

수집한 날씨 데이터를 통해 **강우 여부를 예측**하는 머신러닝 모델을 완성한다.

03. 데이터 가설

가설 1

여름 시즌이 다른 계절에 비해 비가 올 확률이 높다

가설 2

대기 습도가 높을수록 비가 올 확률이 높다.

Baseline Model 및 평가지표

01. 베이스라인 모델이란?

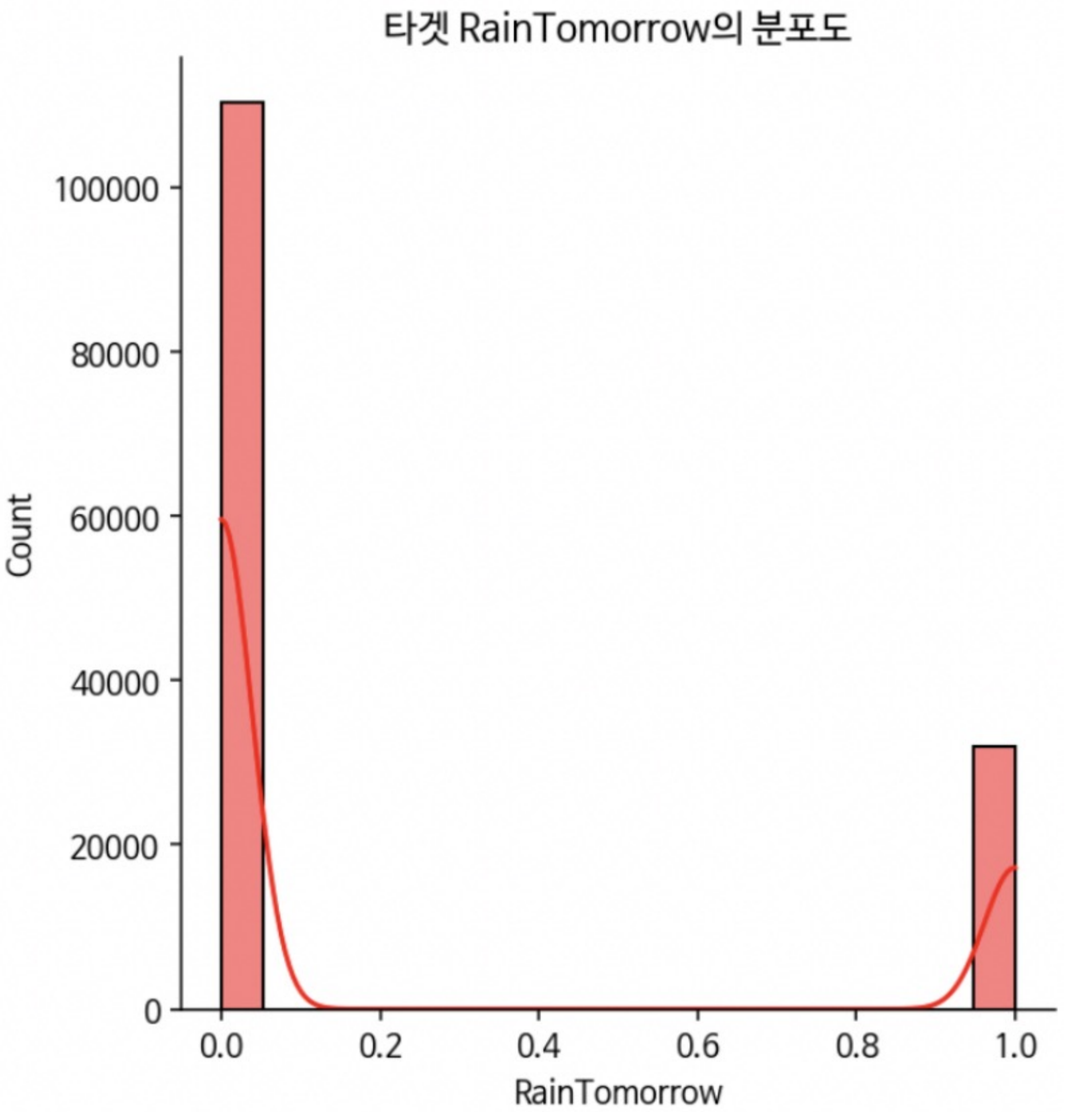
모델의 성능을 비교하기 위한
초기 모델.

비 오는 경우	0.22
비 안오는 경우	0.77

최빈값을 통해 비가 오지 않는 경우의
확률이 77% 임을 확인.

02. 평가 지표

내일 비가 오는지 안오는지 분류 문제.
AUC score를 이용해서 평가.



가설 검증 01

01. 여름 시즌이 다른 계절에 비해 비가 올 확률이 높다.

Step

1. 여름 확인 컬럼 생성

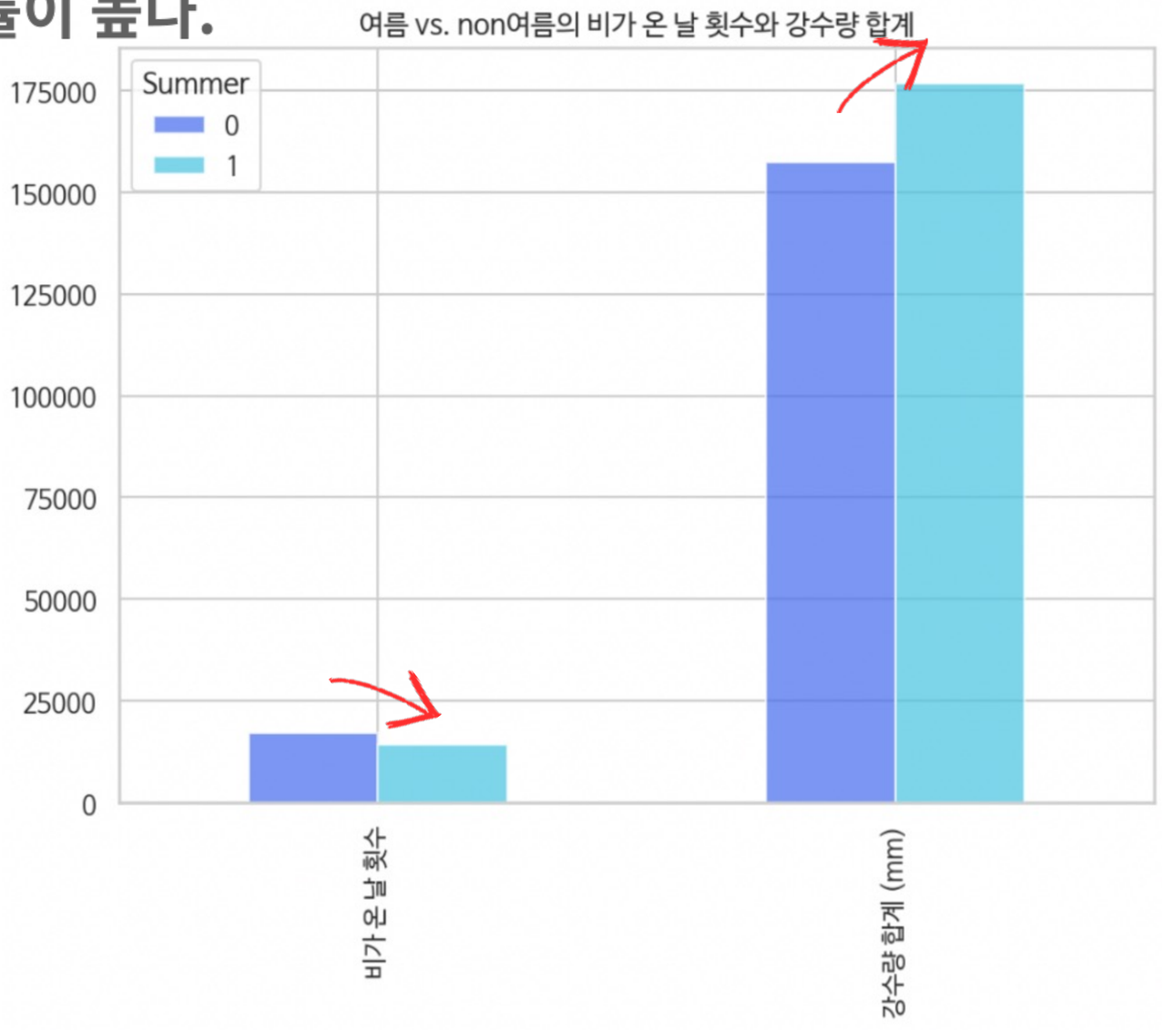
Summer time이 시작되는 10월 부터 summer time이 끝나는 4월까지를 여름으로 지정.

2. 여름 vs. non-여름

여름 : 1, 2, 3, 10, 11, 12월
non 여름 : 4, 5, 6, 7, 8, 9월

3. 가설 검정

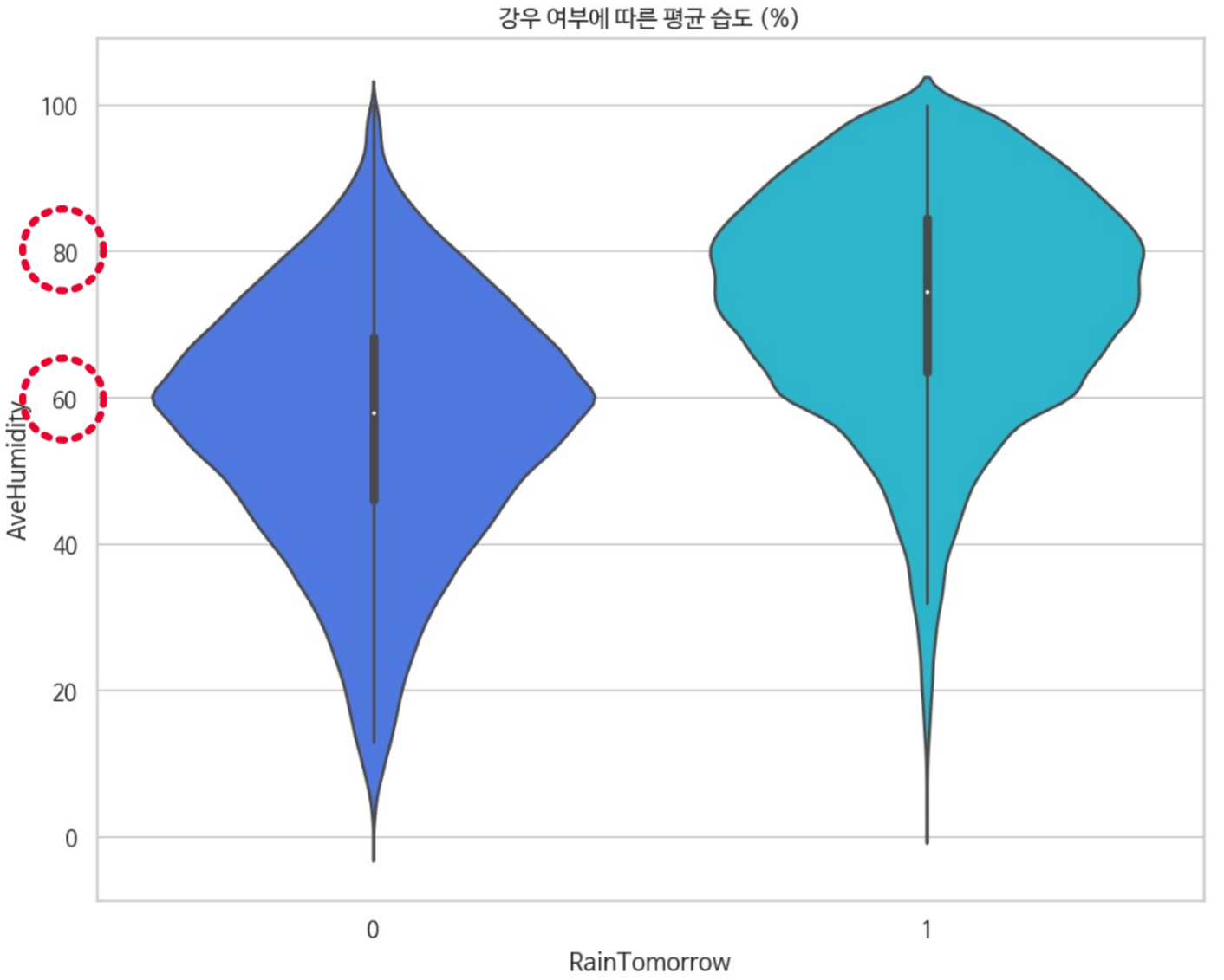
강수량 총량은 더 높지만 여름에 비 오는 횟수는 더 적다.
그러므로 가설은 틀린 것으로 판단되며, 모델링 이후에 다시 확인.



가설 검증 02

02. 대기 습도가 높을수록 비가 올 확률이 높다.

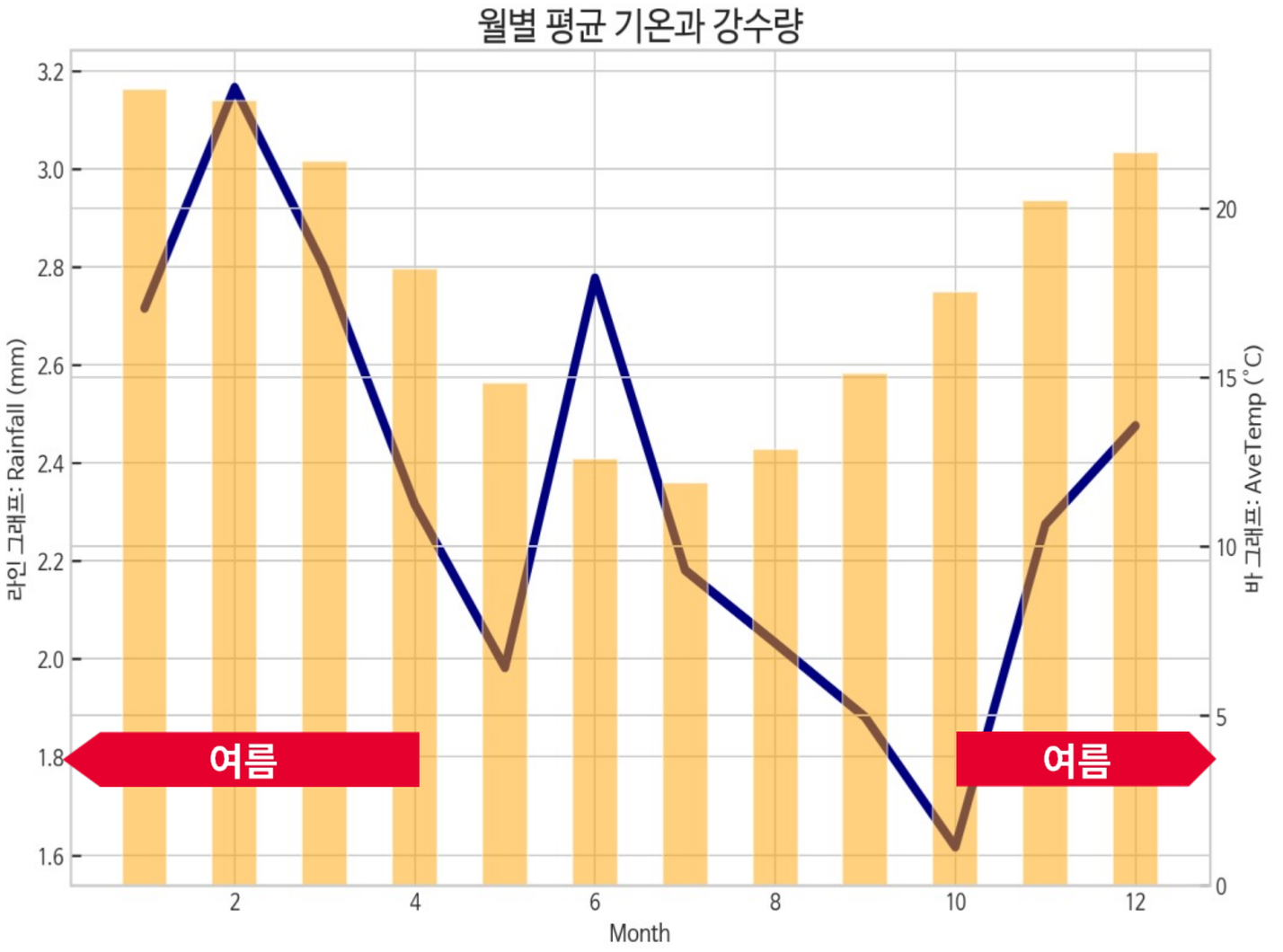
- 내일 비가 오는 지에 따라 평균 습도에 차이가 있다.
- 내일 비가 올 확률일 경우 평균 습도가 높다.



EDA

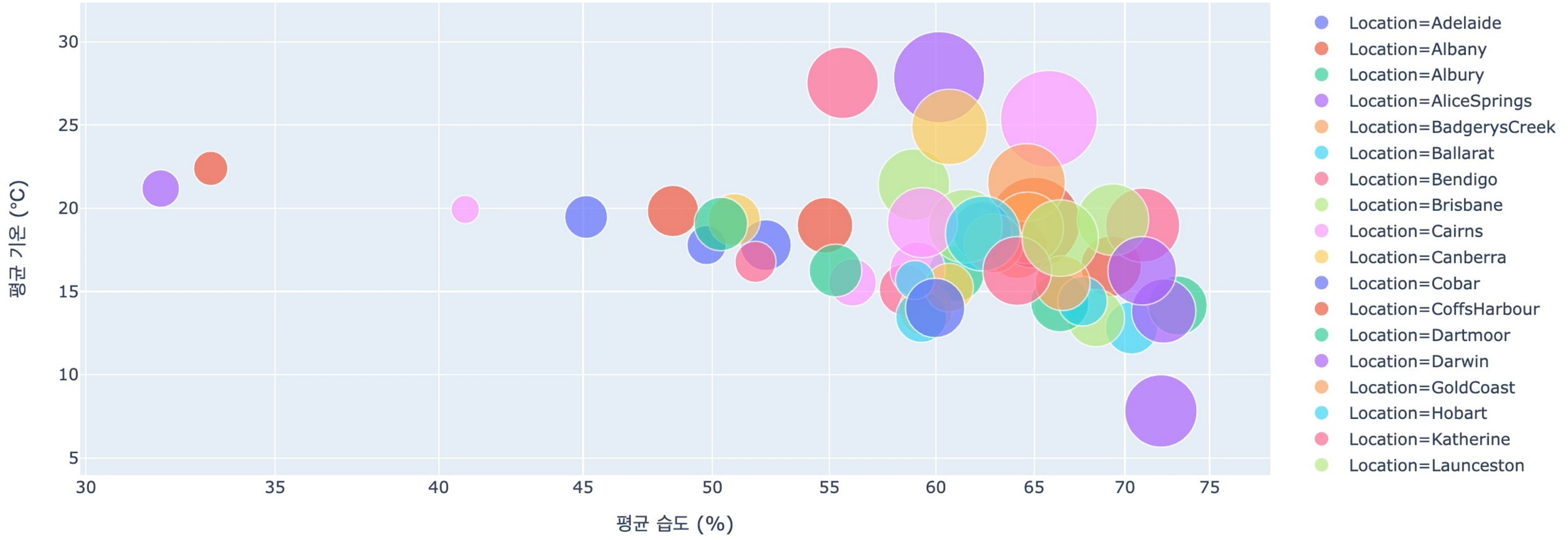
01. 월별 평균 기온과 강수량 경향

- 평균 기온이 낮아질수록 강수량이 낮아진다.
- 여름이 시작되는 시기인 10월부터 강수량이 급증한다.



EDA

02. 각 지역의 습도와 기온에 따른 강수량 변화

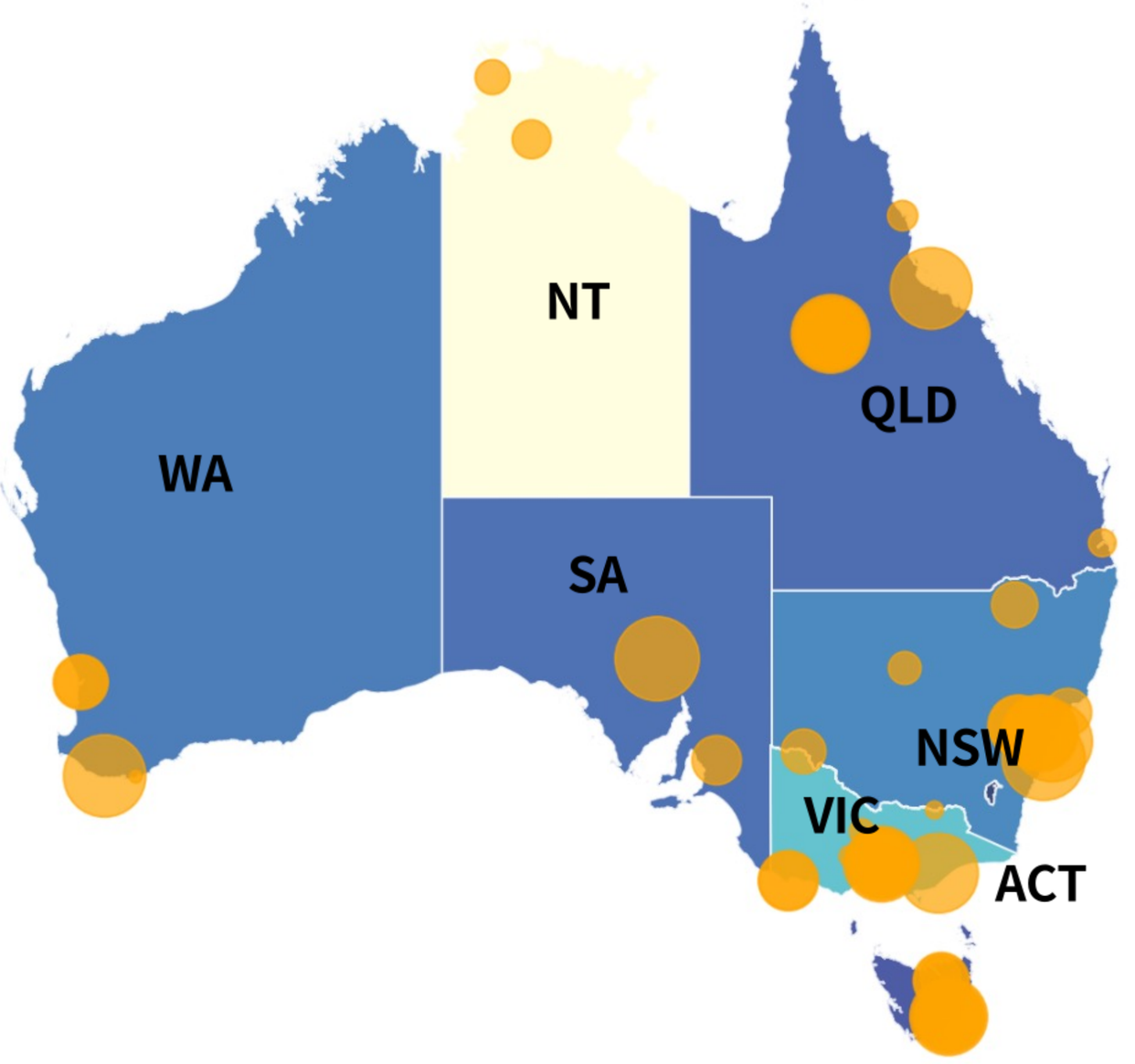


- 습도가 높아짐에 따라 강수량 크기가 커진다.
- 평균 기온은 강수량과 큰 연관성이 나타나지 않는다.

EDA

03. 각 주 (State)와 도시의 비 온 횟수

- geometry 데이터 출처
https://data.gov.au/
- NT 주 외에는 비가 온 횟수의 차이가 크지 않다.
- 내륙 지역과 해안 지역의 강우 횟수의 차이가 나타나지 않는다.

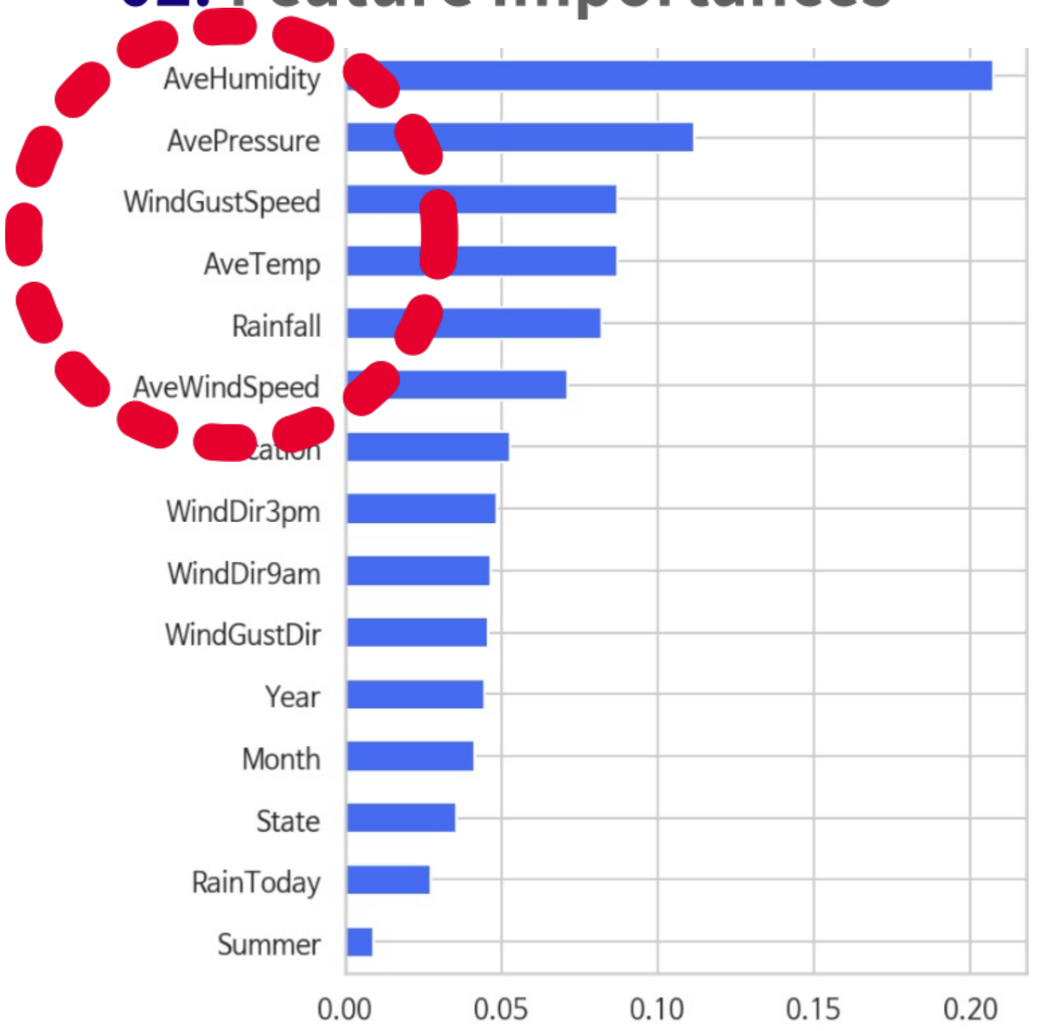


Random Forest

01. 성능 평가

정확도	0.8448
정밀도	0.7401
재현율	0.4740
F1 Score	0.5779
AUC	0.8650

02. Feature Importances



XGBoost Classifier

01. 성능 평가

최적의 하이퍼 파라미터를 찾아주는 RandomizeSearchCV를 이용.

XGBoost	하이퍼 파라미터 튜닝 전	하이퍼 파라미터 튜닝 후
정확도	0.8414	0.8524
정밀도	0.7249	0.7396
재현율	0.4715	0.5273
F1 Score	0.5713	0.6156
AUC	0.8606	0.8775

02. Permutation Importance

어떤 특성이 가장 중요한가를 보여주는 지표

Weight	Feature
0.1068 ± 0.0039	AveHumidity
0.0329 ± 0.0006	WindGustSpeed
0.0221 ± 0.0018	AvePressure
0.0131 ± 0.0022	AveWindSpeed
0.0113 ± 0.0023	State
0.0112 ± 0.0014	Location
0.0098 ± 0.0025	AveTemp
0.0082 ± 0.0007	Rainfall
0.0057 ± 0.0013	RainToday
0.0041 ± 0.0007	Month
0.0036 ± 0.0019	WindDir3pm
0.0023 ± 0.0012	WindGustDir
0.0022 ± 0.0003	Year
0.0016 ± 0.0006	WindDir9am
0.0006 ± 0.0008	Summer

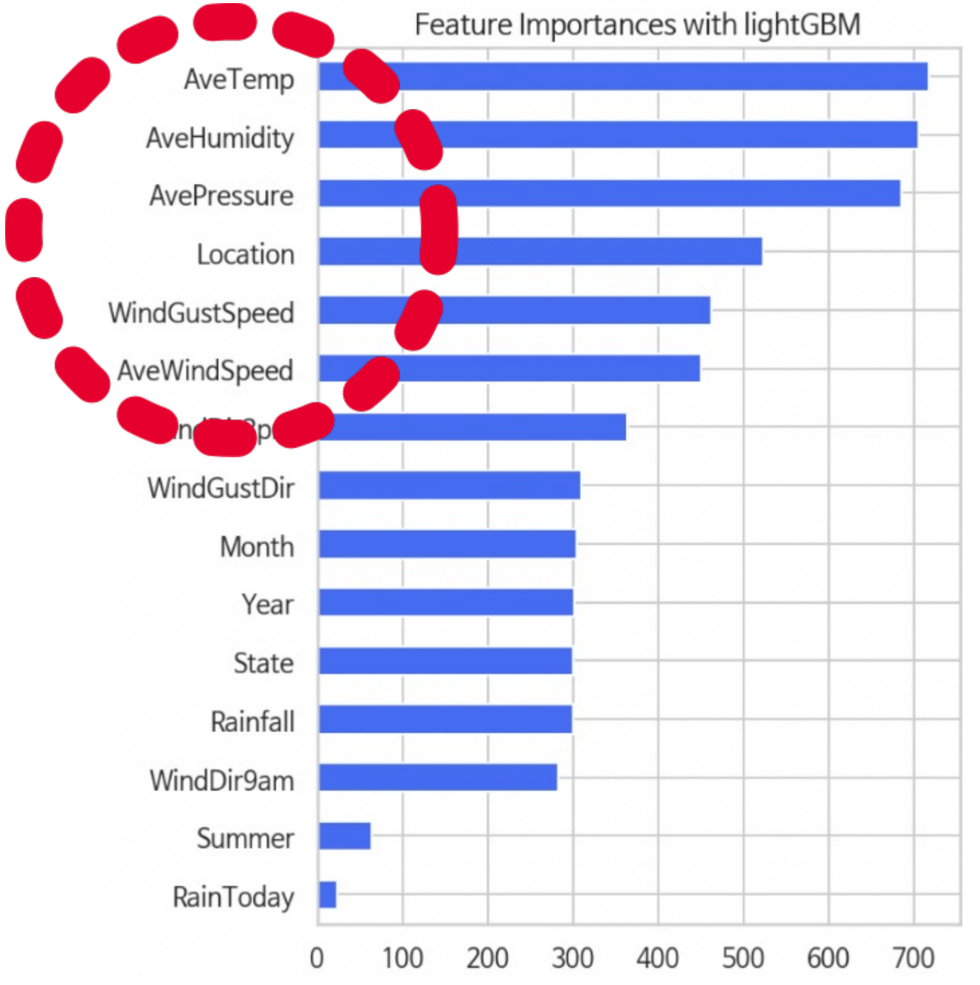
LightGBM Classifier

01. 성능 평가

최적의 하이퍼 파라미터를 찾아주는 RandomizeSearchCV를 이용.

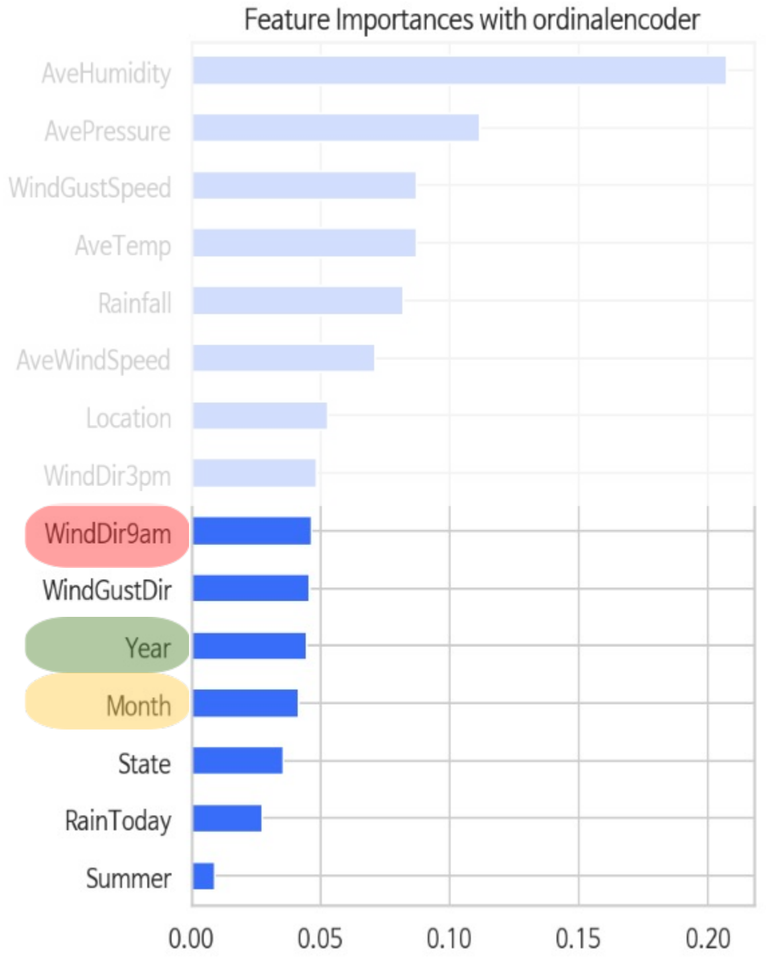
LGBM	하이퍼 파라미터 튜닝 전	하이퍼 파라미터 튜닝 후
정확도	0.8486	0.8521
정밀도	0.7348	0.7375
재현율	0.5083	0.5282
F1 Score	0.6009	0.6156
AUC	0.8741	0.8759

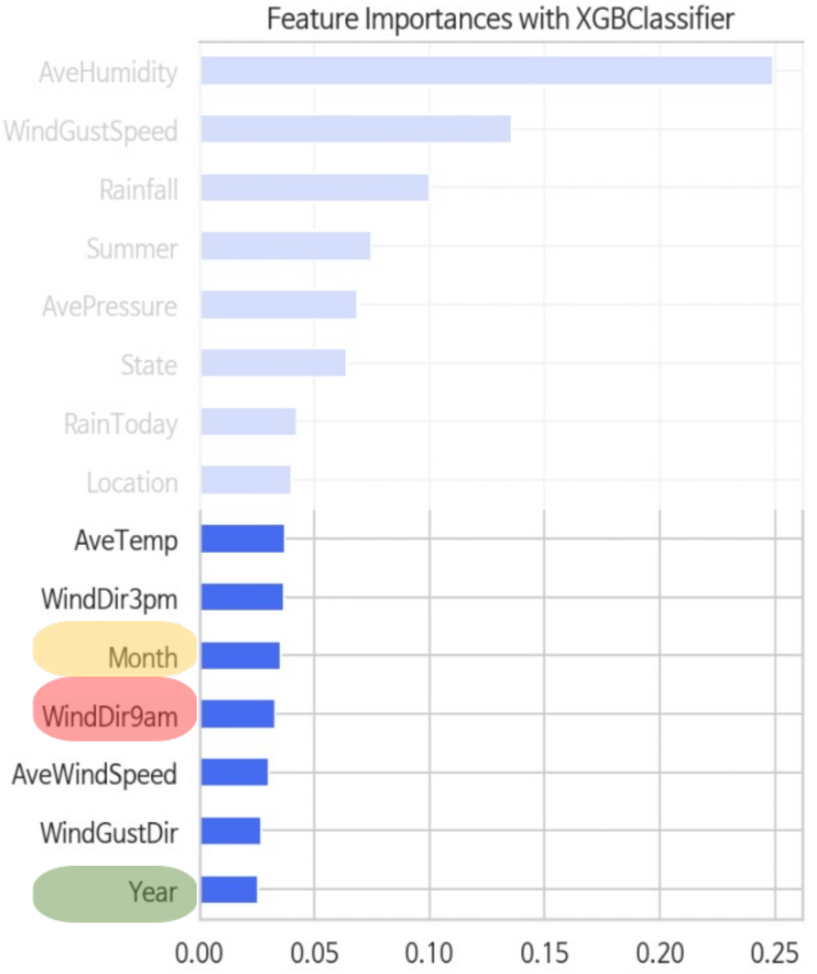
02. Feature Importance

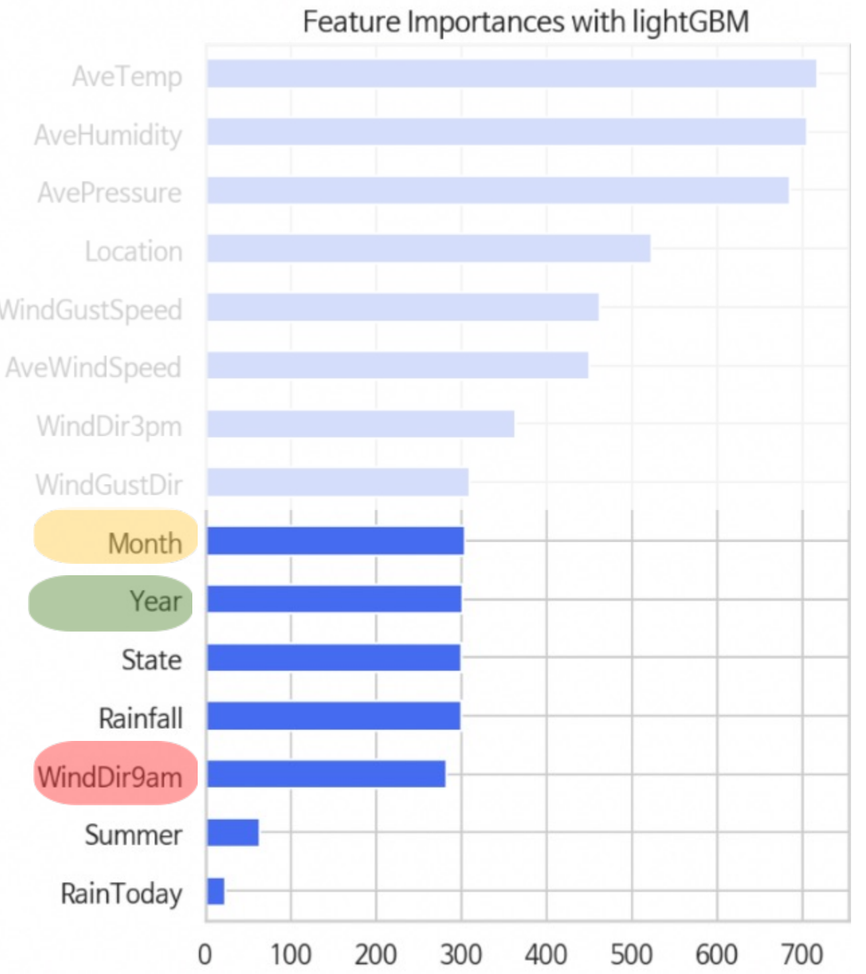


모델들 Feature Importance 비교

세 개의 모델의 특성 중요도 중 공통적으로 하위 7위에 속하는 특성을 찾는다.
WindDir9am, Year, Month 특성은 모든 모델에서 중요하지 않으므로 제거한다.







최종 모델링

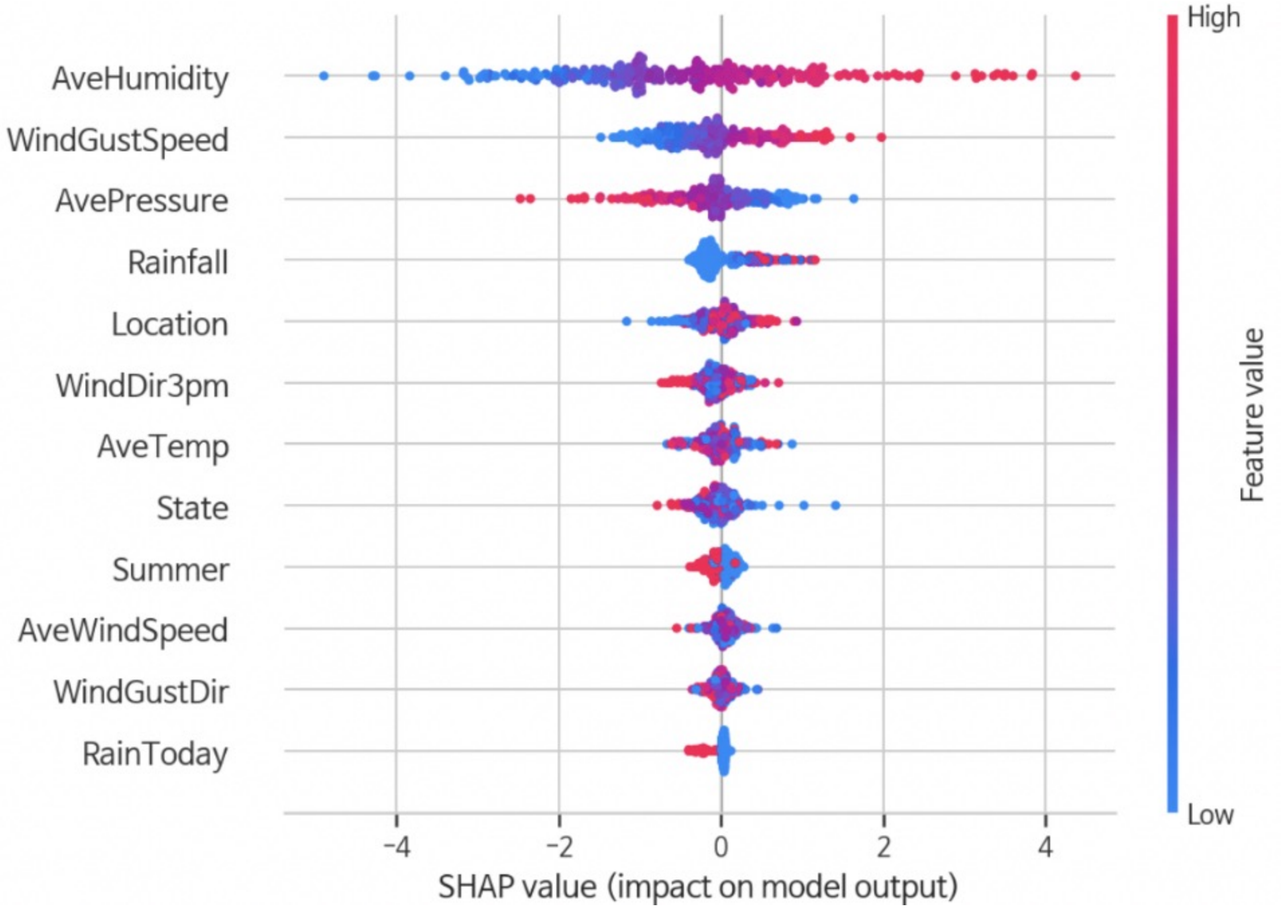
01. 모델 선택 및 테스트셋 최종 평가

AUC 값이 좀 더 높은 XGBoost Classifier 이용.

	XGBoost (최종 전 모델)	LGBM (최종 전 모델)	최종 모델
정확도	0.8524	0.8521	0.8468
정밀도	0.7396	0.7375	0.7259
재현율	0.5273	0.5282	0.5089
F1 Score	0.6156	0.6156	0.5983
AUC	0.8775	0.8759	0.8747

최종 모델 Insight

02. 머신러닝 모델 해석



	비가 올 확률 높아진다	비가 올 확률 낮아진다
평균 습도 AveHumidity	높을수록	낮을수록
가장 강한 풍속 WindGustSpeed	높을수록	낮을수록
평균 기압 AvePressure	낮을수록	높을수록
강수량 Rainfall	높을수록	낮을수록

03. 가설 확인

가설 1. 여름 시즌이 다른 계절에 비해 비가 올 확률이 높다.

FALSE

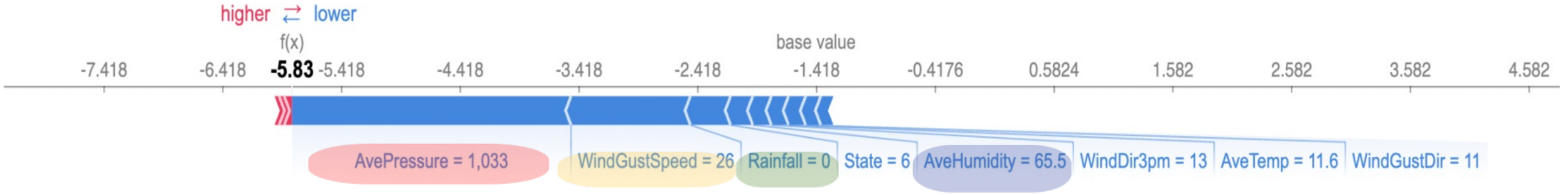
가설 2. 대기 습도가 높을수록 비가 올 확률이 높다.

TRUE

최종 모델 Insight

03. 특정 데이터를 이용하여 모델 해석

Location	Rainfall	WindGustDir	WindGustSpeed	WindDir3pm	RainToday	AveTemp	AveHumidity	AveWindSpeed	AvePressure	Summer	State
Melbourne	0.0	N	26.0	NNE	0	11.6	65.5	9.5	1033.0	0	Victoria



	비가 올 확률 높아진다	비가 올 확률 낮아진다
평균 습도 AveHumidity	높을수록	낮을수록
가장 강한 풍속도 WindGustSpeed	높을수록	낮을수록
평균 기압 AvePressure	낮을수록	높을수록
강수량 Rainfall	높을수록	낮을수록

내일 비가 오지 않으리라 예측!

최종 모델 Insight

04. 모델의 한계

지역별 차이

호주 지역별 기상 차이

호주는 거대한 면적을 갖고 있으므로
각 지역별로 기상 차이가 큽니다.
이 모델은 각 지역별 기상 차이를
전혀 고려하지 못합니다.

하이퍼파라미터

튜닝 할수록 성능 하락

최적화된 하이퍼 파라미터를 구한뒤
모델을 학습시켰지만
오히려 성능이 더 하락하는
현상이 발생하였습니다.

이상 기온 현상

이상 기상 현상 반영 못함

시간이 흐름에 따라
지구 온난화가 악화되면서
발생하는 이상 기상 현상을
반영하지 못합니다.
시간의 흐름을 반영하여
그에 따른 예외적인 현상도
예측 가능할 필요가 있습니다.

Weather Forecast in Australia

THANK YOU.