태양광 패널 데이터 분석을 통한 모델별 성능 파악 및 선정

2024.08.14.

So What 조(4조)

강현주 / 권아연 / 구정모

김동엽 / 박동융 / 조시윤

역할 분담

강현주

- 선형회귀분석
- 군집분석
- 이동평균 파악
- PPT 작성

권아연

- 발표
- 주성분분석
- 지역별 선형회귀분석
- 의사결정나무 + k-fold

구정모

- 기상청 데이터 수집
- 랜덤포레스트 분류분석
- 앙상블 _ soft voting
- 주성분 분석

김동엽

- 로지스틱 회귀분석
- XG Boost
- 주성분 분석
- 시계열 분석

박동융

- 데이터 전처리
- 로지스틱 회귀분석
- 랜덤 포레스트 회귀분석
- 시계열 분석

조시윤

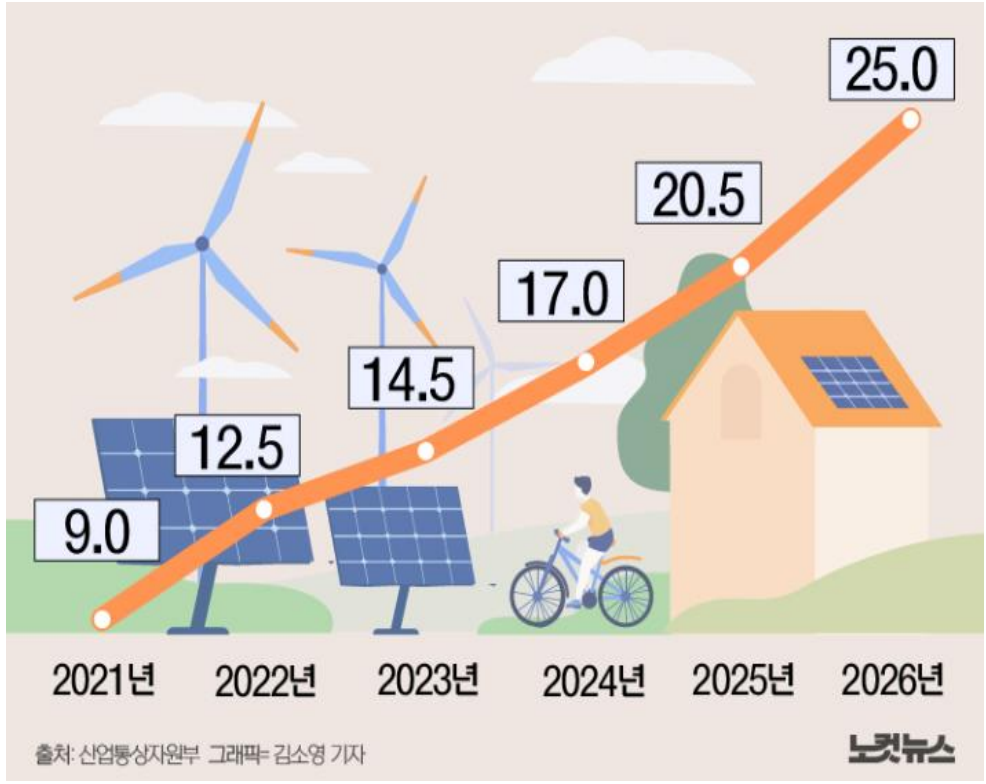
- 랜덤포레스트 회귀분석
- 의사결정나무 분석
- XG Boost
- GBM Boost
- Ada Boost

방향성 _ 추진 배경

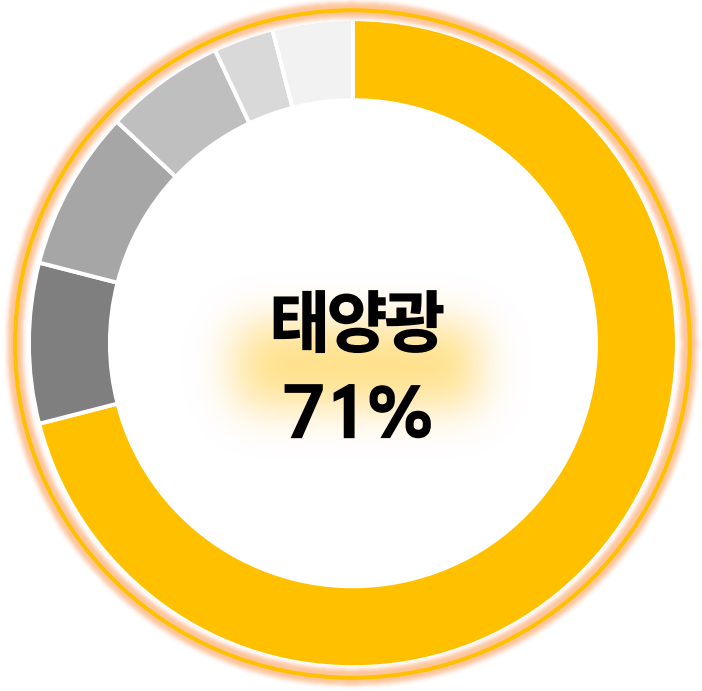
신재생에너지공급 의무화 비율이 꾸준한 상승세를 보이며, **태양광 에너지 발전 설비 비중이 71%로 압도적임**에 따라

태양광 에너지 발전 사업에 대한 관심 필요

| 신재생에너지공급 의무화 비율 |



| 신재생E 발전 설비 비중 |



■ 태양광 ■ 풍력 ■ 일반수력 ■ 바이오 ■ 연료전지 ■ 기타

출처 : 전력거래소 '2020년도 발전설비현황' 통계자료

방향성 _ 현상 파악

에너지 저장 장치(ESS)의 과부화로 화재 건수가 증가하는 추세 + 출력량이 많아 잦은 출력 제어 발생

출력량을 예측하여 에너지 저장 장치 과부하 방지 필요

| 다시 늘어나는 ESS 화재 |

| 잦은 출력 제어 시스템으로 인한 문제 |

올해 3월까지 총 60회 출력제어 발생

양이원영 더불어민주당 의원실에 따르면 지난해 풍력발전과 태양광 발전의 출력제어는 총 132회 발생했다. 올해 지난 3월까지 총 60회의 출력제어가 발생했고 3월 한 달 동안 태양광은 총 14회의 출력제어가 일어났다.

ㄱ씨가 운영하는 제주도 서귀포시 406키로와트(kW) 규모의 태양광 발전소는 올해 1월부터 이달 17일까지 29차례 '출력 제어'가 이뤄졌다. 지난달에는 8일부터 나흘 연속 약 3~4시간 씩 설비를 멈춰세웠다.

출력제어는 전력당국이 해당 발전소의 전력망 접속을 차단해 전력 생산을 중단하는 조처다. 사업자들은 '영업정지'로 받아들인다. 출력 제어가 되면 전기를 팔 수 없기 때문이다.

데이터 수집(1)

● 결측치 및 이상치 처리 , 파생변수 생성

항목	의미	유형	이상치	결측치	확인결과	정제방안
위치	A,B,C,D,E 지역 분류	범주형	B지역 값	-	B지역 데이터 이상치 발생	B지역 제거, A,C,D,E 지역 값 대치 (A,C,D,E ▶당진, 울산, 영광, 동해)
일자	관측일자	범주형	-	-	-	Datetime 날짜형 변환
계절	관측일자로 계절 파생변수 생성	범주형	-	-	-	파생변수
현재발전출력	측정된 전력의 출력 값	연속형	-	-	-	목표변수 Y 설정
평균기온(°C)	관측 기준 평균 기온	연속형	-	O	NaN값 0 대치	기상청 2016~2021년도 데이터 수집 및 병합
강수 계속시간(Hr)	관측 기준, 강수 지속 시간	연속형	-			
1시간 최다강수량(mm)	시간당 최다강수량	연속형	-			
일강수량(mm)	일 강수량(total)	연속형	-			
가조시간(hr)	발전 설비 운영 X 시간 or 발전 중단된 시간	연속형	-			

데이터 수집(2)

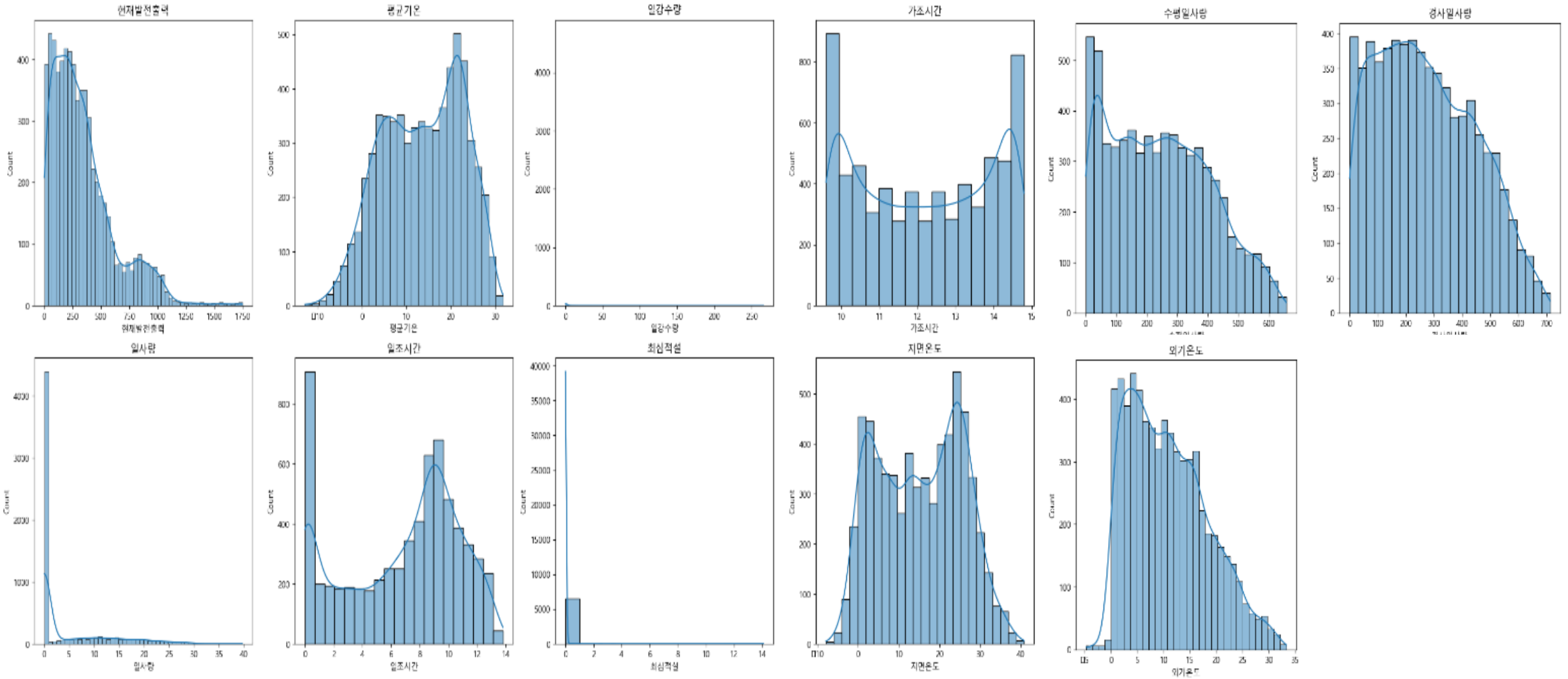
● 결측치 및 이상치 처리 , 파생변수 생성

결측치 있는 항목	의미	유형	이상치	NaN 값	확인결과	정제방안
합계 일조시간(hr)	태양빛이 지면에 도달하여 관측된 시간 합계	연속형	-	0	NaN값 0 대치	기상청 2016~2021년도 데이터 수집 및 병합
평균 전운량(1/10)	하늘의 구름 덮임 정도 (구름의 양/10)	연속형	-			
평균 중하층운량(1/10)	중간 높이, 낮은 높이에 위치한 구름 덮임 정도(구름의 양/10)	연속형	-			
평균지면온도(°C)	지표면에서 측정된 온도의 평균값	연속형	-			
수평 일사량	• 패널의 수평 그리드 일사량 • 수평 일사량 1, 2 의 평균치	연속형	-	-	-	파생변수
경사 일사량	• 패널의 경사 그리드 일사량 • 경사 일사량 1, 2 의 평균치	연속형	-	-	-	파생변수
외기온도	• 대기 중의 공기 온도 • 외기 온도1, 2의 평균치	연속형	-	-	-	파생변수

데이터 분석 계획 수립 내용

목적	분석방법
설명력이 우수한 모델 선정	랜덤포레스트
	선형회귀분석
	GBM Boost
	XG Boost
	Ada Boost
	의사결정나무
	랜덤포레스트
	XG Boost
	로지스틱 분석

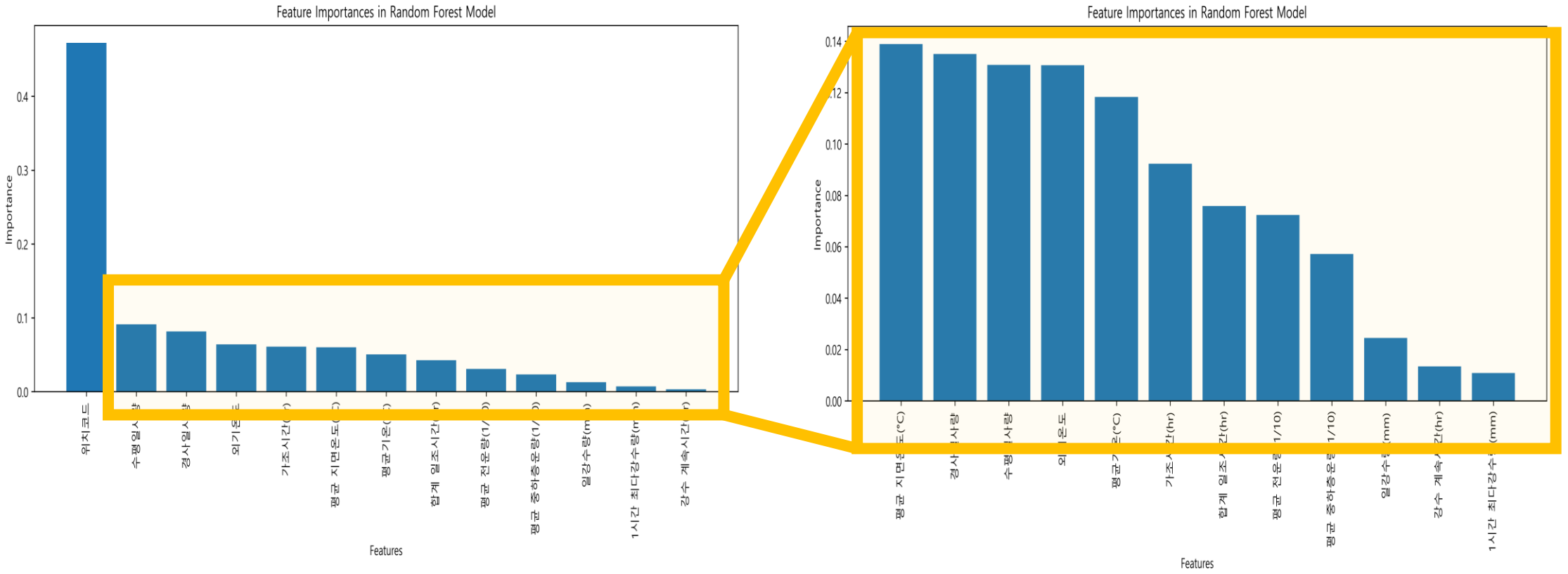
데이터 분포 확인


현재발전출력
평균기온
일강수량
가조시간
수평일사량
경사일사량
일사량
일조시간
최심적설
지면온도
외기온도
Count

데이터 분석 결과(1)_회귀분석 데이터 선정

데이터의 특성 파악을 위해 특정 회귀 분석(Random Forest)진행 결과, 위치데이터의 영향력이 큼을 확인.

위치데이터 제거 후에 회귀 분석 재진행



데이터 분석 결과(1)_회귀분석(Y=출력량, X= 위치데이터 제외한 나머지 데이터)

6가지 분석에 대한 결정계수 확인 결과, Radom Forest의 설명력이 가장 우수했기에

Random Forest 분석을 선택

Random Forest
$r^2\ score$
0.214

Xg boost
$r^2\ score$
0.2088

Gbm boost
$r^2\ score$
0.2061

선형회귀분석
$r^2\ score$
0.064

의사결정트리
$r^2\ score$
0.0652

Ada boost
$r^2\ score$
0.0053

데이터 분석 결과(1)_회귀분석(Y=출력량, X= 위치데이터 제외한 나머지 데이터)

5가지 분석에 대한 결정계수 확인 결과, Radom Forest의 설명력이 가장 우수했기에

Random Forest 분석을 선택

Random Forest

$r^2\ score$

0.214

선형회귀분석

$r^2\ score$

0.064

[성능 지표]

1.MAE : 179.62

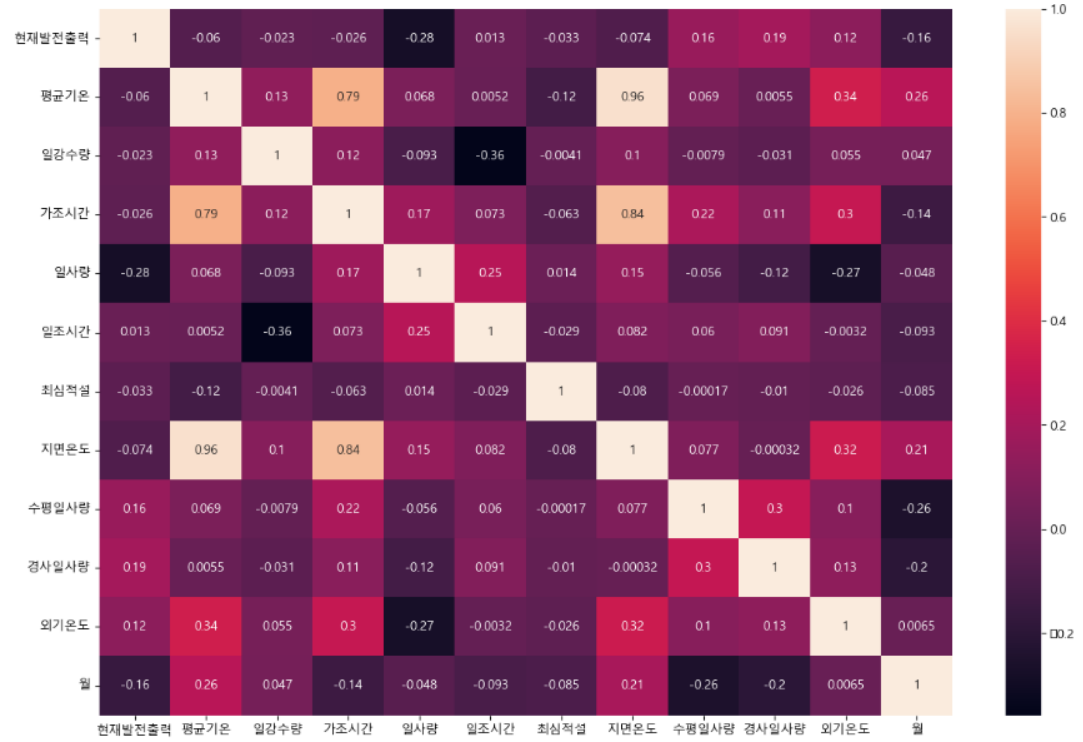
2.MSE : 58621.74

3.$r^2\ score$: 0.2140

4.explained_variance : 0.2161

데이터 분석 결과(2) _ 전체 지역 로지스틱 회귀분석

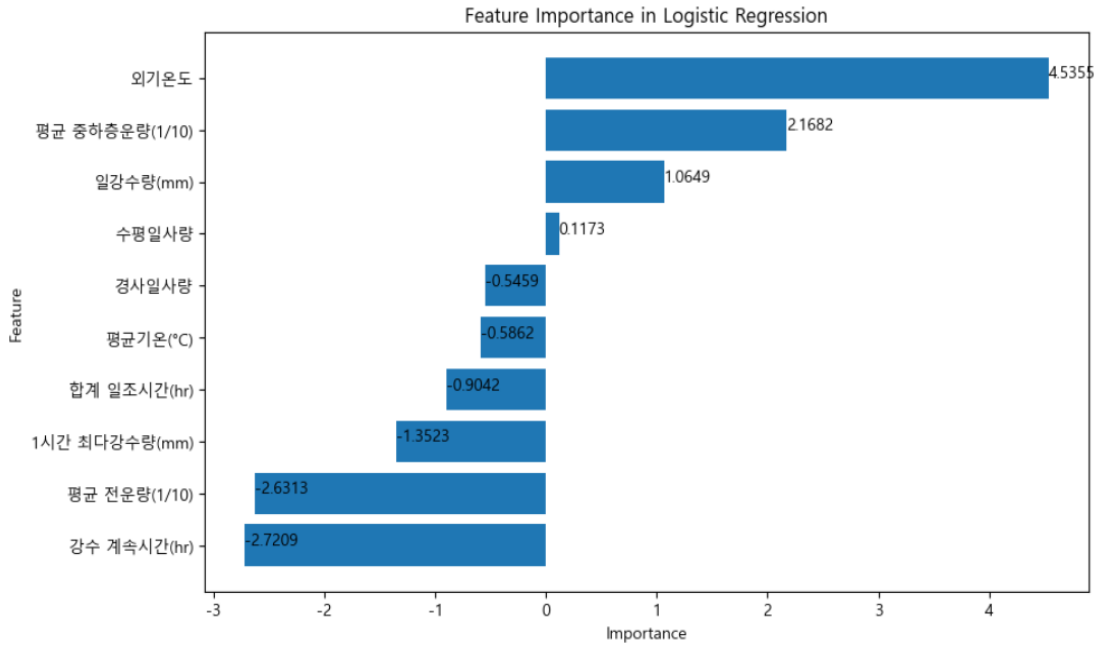
변수 간 상관관계 히트맵 시각화



▶ 변수 간 상관관계를 시각화한 결과, **선형 관계를 파악**

→ 로지스틱 회귀분석을 위한 독립변수 선정(상관관계 높은 변수 제거)

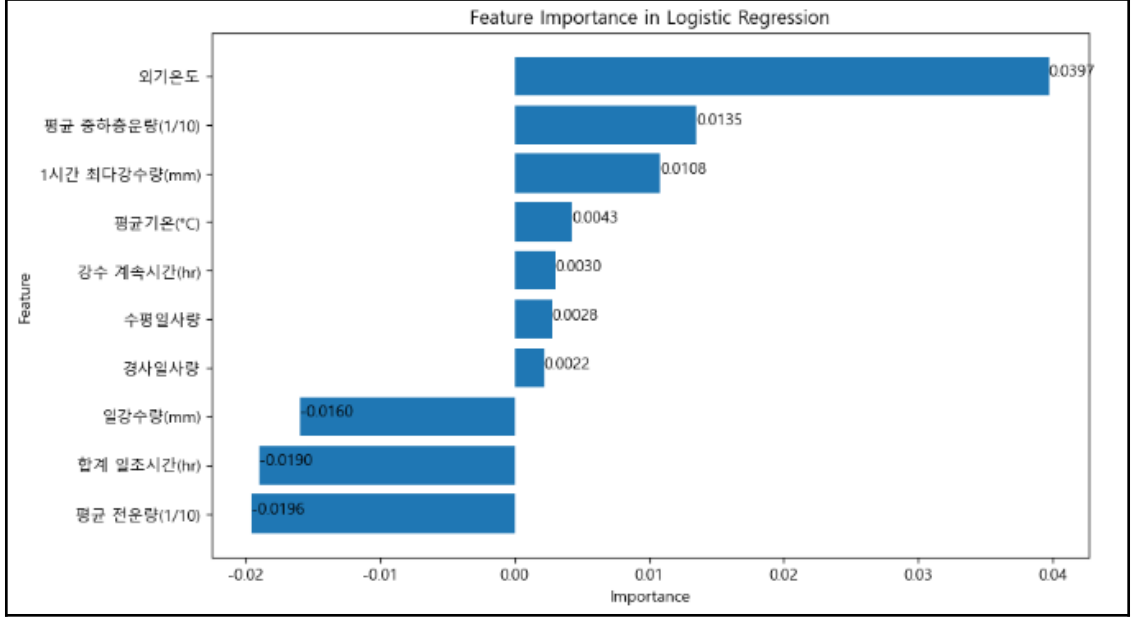
'전체' 지역 로지스틱 회귀분석



	정확도	정밀도	재현율	F1-score
비위험군	0.8156	0.86	0.93	0.89
과부하 위험군 (상위 20%)		0.36	0.20	0.26

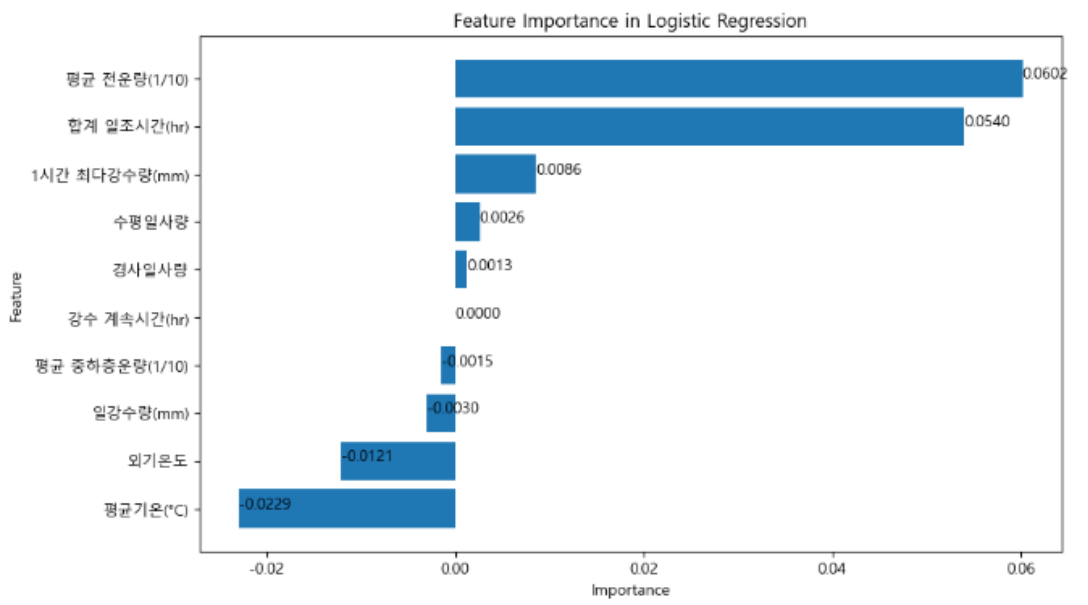
데이터 분석 결과(2) _ 지역별 로지스틱 회귀분석

‘당진’ 지역 로지스틱 회귀분석



	정확도	정밀도	재현율	F1-score
비위험군	0.7739	0.85	0.87	0.86
과부하 위험군 (상위 20%)		0.40	0.36	0.38

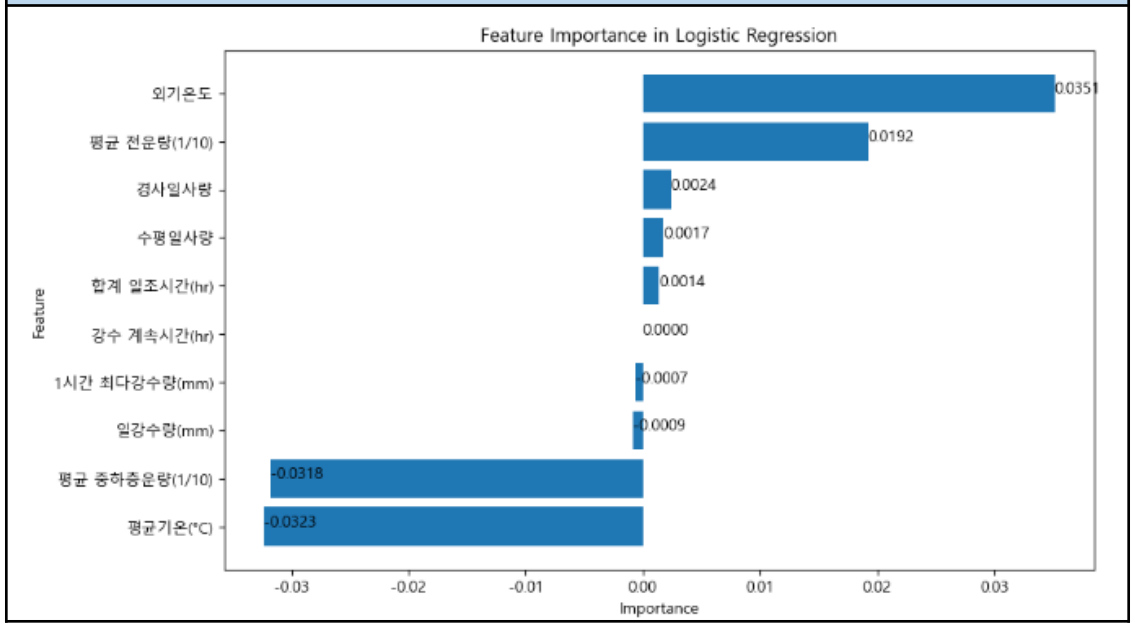
‘울산’ 지역 로지스틱 회귀분석



	정확도	정밀도	재현율	F1-score
비위험군	0.7922	0.86	0.89	0.87
과부하 위험군 (상위 20%)		0.47	0.39	0.43

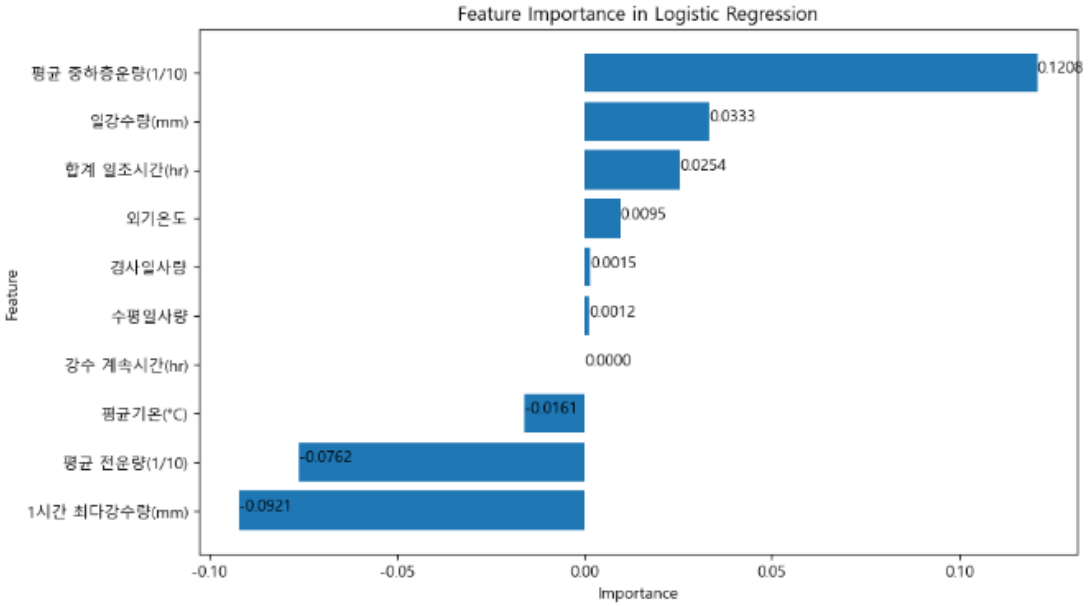
데이터 분석 결과(2) _ 지역별 로지스틱 회귀분석

'영광' 지역 로지스틱 회귀분석



	정확도	정밀도	재현율	F1-score
비위험군	0.8059	0.84	0.95	0.89
과부하 위험군 (상위 20%)		0.49	0.21	0.30

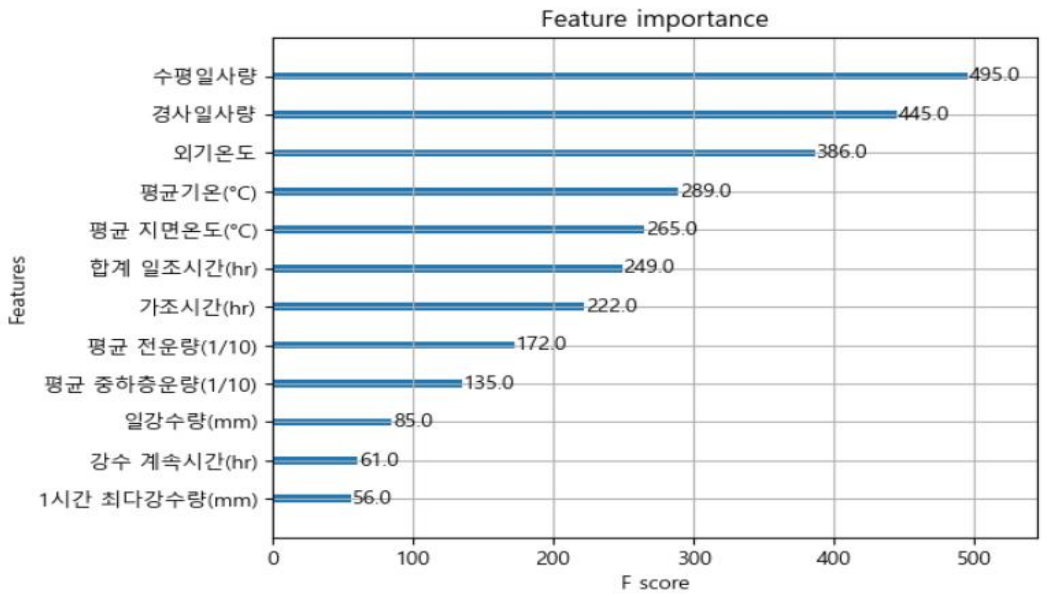
'동해' 지역 로지스틱 회귀분석



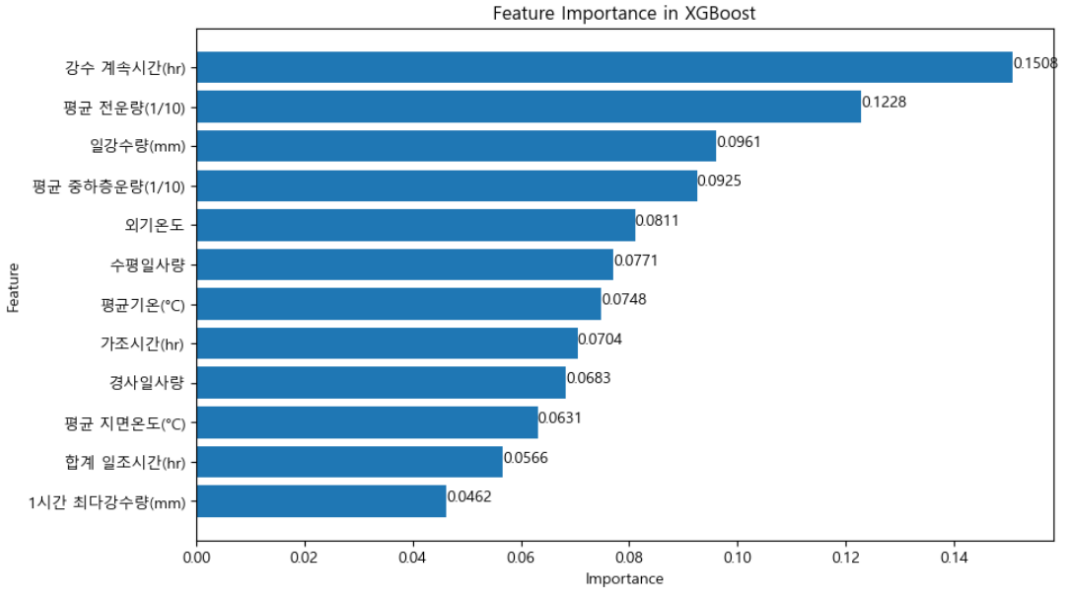
	정확도	정밀도	재현율	F1-score
비위험군	0.7716	0.80	0.95	0.87
과부하 위험군 (상위 20%)		0.38	0.12	0.18

데이터 분석 결과(3) _ 특정 독립변수 XGBoost

Gain 값 기반 XGBoost



Weight 값 기반 XGBoost



	정확도	정밀도	재현율	F1-score
비위험군	0.8013	0.83	0.94	0.88
과부하 위험군(상위 20%)		0.60	0.31	0.41

효율성 값이 하위 30% 이하면 0 (낮은 효율성), 그렇지 않으면 1 (높은 효율성)

데이터 분석 결과(4) _ 분류 분석(Random Forest)

위험군 설정

- 현재발전출력 상위 20% 기준 : 과부하 위험군

플래그 생성

- 하위 80% : 0
- 상위 20% : 1

→ 플래그 생성하여 이진 클래스로 분류

정규화

- X, Y 선정

(상관관계 높은 변수 제외)
(출력량 극단적인 당진, 영광 제외)

- Min-Max scaling

→ 정규화

모델 훈련 및 예측

- 데이터 분할

(Test data 20%)

- 랜덤포레스트 분류 모델 훈련 및 예측

성능평가

- 과부하 위험군
  - ✓ 정확도 : 0.87
  - ✓ 정밀도 : 0.74
  - ✓ 재현율: 0.56
  - ✓ F1-score : 0.64
- 비위험군
  - ✓ 정확도 : 0.87
  - ✓ 정밀도 : 0.90
  - ✓ 재현율: 0.95
  - ✓ F1-score : 0.92

인사이트 도출

● 분석을 통한 인사이트 도출안

	당진	울산	영광	동해
과부하 위험군 F1-score	0.38	0.43	0.30	0.18

지역별 로지스틱 회귀 분석 후 F1-Score 확인 결과,

울산 지역의 F1-Score 수치가 가장 우수함을 확인할 수 있었다.

▶ 울산 데이터의 성능이 좋음을 파악했고, 울산 데이터로 추가적인 분석 필요

● 최종 선택 모델

: 랜덤 포레스트 분류 모델

20%	과부하 위험군 (출력 제어 명령 발생 위험군)
80%	비위험군

● Review . .

1) 깨달은 점

실제 데이터를 기반으로 진행하니, 생각보다 R squared 수치가 낮게 나왔다.

이에 이론적으로 배운 값과는 수치적으로 차이가 있어 의미가 없는 줄 알았으나,

실무 데이터에서는 0.2XX 값이면 높은 수치임을 알게 되었다.

2) 좋았던 점

이론적으로 배웠던 머신러닝 분석 기법을 프로젝트를 통해 익힐 수 있어 좋았다.

감사합니다.

iM DiGital Banker Academy