공장 내 철강 제품의 결로(結露) 발생 예측 모형 개발

203675 신민용, 김지윤, 이다정

목차

1. 공모배경

1.1 결로 발생 조건

상대적으로 높은 습도

차가운 표면

이슬점보다 표면온도가
낮으면 결로 발생

" 이슬점 +3° 일때 안전 "

1. 공모배경

1.2 결로 발생시 문제점

코일 표면에 부식 발생 → 결로를 제거하기 위한 내부 프로세스 추가 → 제품 생산 비용 증가

결로 발생 예측을 하지 못한다면?

표면 부식 결함을 인식하지 못하고 최종 고객사까지 출하 → 고객사의 불만 제기에 따른 금전적인 손해비용 보상 신용도 하락 → 제조사 이미지 실추

1. 공모배경

1.2 결로 발생시 문제점

[출처] 금융감독원 전자공시시스템

3. 재고자산의 보유 및 실사 내역 등

가. 최근 3사업연도의 재고자산의 사업부문별 보유현황

(단위 : 백만원, %)

사업부문	계정과목	제56기 1분기	제55기	제54기	비고
철강업	제 품	1,365,998	1,332,155	1,001,359	-
	부산물	4,630	3,760	5,951	
	상 품	162,547	129,602	117,196	
	반제품	705,058	830,319	801,564	
	재공품	338,113	377,637	341,789	
	원재료	1,192,135	1,224,410	1,113,033	
	저장품	744,698	742,933	735,216	
	미착자재	768,746	774,736	803,362	
	합 계	5,281,925	5,415,552	4,919,470	
총자산대비 재고자산 구성비율(%) [재고자산합계÷기말자산총계×100]		15.5	15.8	14.8	
재고자산회전율(회수) [연환산 매출원가÷{(기초재고+기말재고)÷2}]		3.32	3.70	4.15	

2020년 3월 말 기준

총 재고자산: 5조 2819억원
- **상품: 1625억원,**
- **제품: 1조 3659억원**

" 결로예측을 통한 재고 가치 상실 방지 필요 "

1. 공모배경

1.3 예측의 중요성(장점)

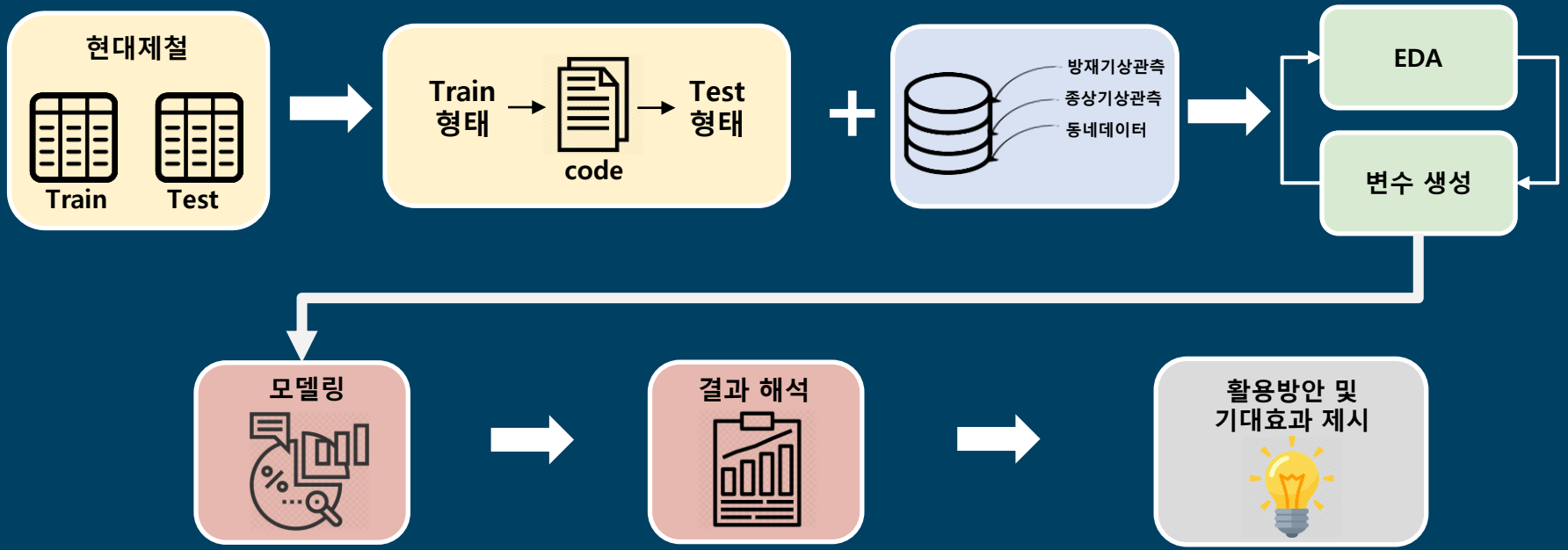
잠재적 품질 손실위험 감소

품질 관리 및 생산비용 절감

고객 만족도 향상

2. 활용 데이터 정의

2.1 분석 프로세스



2. 활용 데이터 정의

2.2 현대제철 내부데이터

Train (116663,15)

시간	대기온도_1번위치	상대습도_1번위치	대기온도_2번위치	상대습도_2번위치	대기온도_3번위치	상대습도_3번위치	외부 대기 온도	외부 상대 습도	코일표면 온도_1번 위치	코일표면 온도_2번 위치	코일표면 온도_3번 위치	결로발생 _1번위치	결로발생 _2번위치	결로발생 _3번위치
2016-04-01 0:00	16	24	11	14	23	11	13	32	10	9	42	0	0	0
2016-04-01 3:00	14	28	10	12	32	9	11	42	7	7	59	0	0	0
2016-04-01 6:00	13	33	10	11	37	9	10	44	7	6	56	0	0	0
2016-04-01 9:00	13	33	10	11	35	9	10	41	8	18	30	0	0	0
2016-04-01 12:00	16	28	10	15	27	11	14	30	9	18	20	0	0	0
2019-03-31														

Test (3539,12)

시간	공장	공장 내부위치	대기온도	대기 상대습도	코일 표면 온도	외부 대기 온도	외부 대기 상대습도	24시간 후 일자 및 시간	24시간 후 결로발생여부 예측 값	48시간 후 일자 및 시간	48시간 후 결로발생여부 예측 값
2019-04-01 0:00	2	3	8.17	40.42	10.1	4.2	54.82	2019-04-02 0:00	NA	2019-04-03 0:00	NA
2019-04-01 4:30	1	2	10.03	48.81	10.79	6.09	59.34	2019-04-02 4:30	NA	2019-04-03 4:30	NA
2019-04-01 10:30	2	3	9.45	40.93	9.07	12.26	32.14	2019-04-02 10:30	NA	2019-04-03 10:30	NA
2019-04-01								2019-04-02		2019-04-03	

2. 활용 데이터 정의

2.3 외부데이터

본 분석에서는 대기온도와 습도 외에 다양한 외부 날씨의 변화가 공장 내부의 결로 현상에 영향을 미칠 것이라고 판단함

방재기상관측(AWS), 종관기상관측(ASOS)

✓ 공장 주위의 환경을 결정할 수 있는 날씨 요소인 강수 유무, 풍속 등을 포함하고 있기 때문에 AWS, ASOS 데이터를 수집함

동네예보

✓ 동네예보 데이터는 뇌전, 하늘 상태 등 다양한 날씨 정보를 포함하고 있음

✓ 당진 1공장 및 2공장과 거리가 매우 가까우므로 공장과 매우 유사한 정보를 담을 것이라고 판단하여 동네예보 데이터를 수집함

2. 활용 데이터 정의

2.3 외부데이터 - 방재기상관측(AWS)

방재기상관측(AWS)

지진, 태풍, 홍수, 가뭄 등 기상현상에 따른 자연재해를 막기위해 실시하는 지상관측을 말함

일시	기온(°C)	1분 강수량(mm)	강수유무(유무)	풍향(deg)	풍속(m/s)	현지기압	해면기압	습도	일사	일조
2016-03-01 0:01	-3.5	0	0	304.2	2.5	1026.6	1029.5		0	0
2016-03-01 0:02	-3.6	0	0	296.3	1.3	1026.6	1029.5		0	0
2016-03-01 0:03	-3.6	0	0	304.3	1.7	1026.6	1029.5		0	0
2016-03-01 0:04	-3.7	0	0	308.1	1.4	1026.6	1029.5		0	0
2020-04-30 23:57	17.1	0	0	208.1	4.6	1011.9	1014.2	88.8		
2020-04-30 23:58	17.1	0	0	208.5	3.8	1011.9	1014.2	88.8		
2020-04-30 23:59	17.1	0	0	211.8	5	1011.9	1014.2	88.9		
2020-05-01 0:00	17.2	0	0	207.2	3.9	1011.9	1014.2	88.8		

당진(2221466,12)

신평(2716943,12)

2. 활용 데이터 정의

2.3 외부데이터 - 종관기상관측(ASOS)

종관기상관측(ASOS)

종관규모의 날씨를 파악하기 위하여 정해진 시각에 모든 관측소에서 같은 시각에 실시하는 지상관측

*종관규모는 일기도에 표현되어 있는 보통의 고기압이나 저기압의 공간적 크기 및 수명을 말하며, 주로 매일의 날씨 현상을 뜻함

일시	온도	누적 강수량	풍향	풍속	현지기압	해면기압	습도	일사량	일조량
2016-04-01 0:00	8.8	0	49.9	1.5	1012.1	1015.6	40.9	13.61	31980
2016-04-01 3:00	7.5	0	345.8	0.4	1011.4	1014.9	40.3	0	0
2016-04-01 6:00	3.7	0	76	0.4	1011.9	1015.5	66.2	0	0
2016-04-01 9:00	12.9	0	294.2	0.7	1013.2	1016.7	41.2	1.39	6960
2019-03-30 23:10	4.1	2.3	319.6	4.9	1012.8	1016	69.9	10.71	12780
2019-03-30 23:20	4.2	2.3	310.3	5.1	1012.9	1016.1	69.9	10.71	12780
2019-03-30 23:30	4.2	2.3	337	3.6	1013.1	1016.3	69	10.71	12780
2019-03-30 23:40	4.2	2.3	315.8	4.6	1013.2	1016.4	69.9	10.71	12780

(393622,10)

2. 활용 데이터 정의

2.3 외부데이터 - 동네예보

동네예보(충청남도 당진시 송악읍)

전국의 읍,면,동 단위별 상세한 날씨를 매시각 제공하는 실황 관측자료

일시	강수량	형태	기온	뇌전	습도	풍속	풍향	하늘 상태
2016-04-01 0:00	0	0	13.8	0	38	0.4	236	1
2016-04-01 3:00	0	0	20.2	0	22	1	29	2
2016-04-01 6:00	0	0	19.8	0	24	4.4	307	4
2016-04-01 9:00	0	0	18.2	0	25	1.8	311	4
2016-04-01 12:00	0	0	19.1	0	49	1.3	270	1
2019-03-30 23:10	0	0	17.9	0	81	1.5	266	1
2019-03-30 23:20	0	0	16.8	0	90	1.3	207	3
2019-03-30 23:30	0	0	16.8	0	91	1.4	236	4
2019-03-30 23:40	0	0	17	0	91	0.7	270	4

(35064,10)

3. 데이터 처리 방안 및 분석기법

3.1 문제 정의

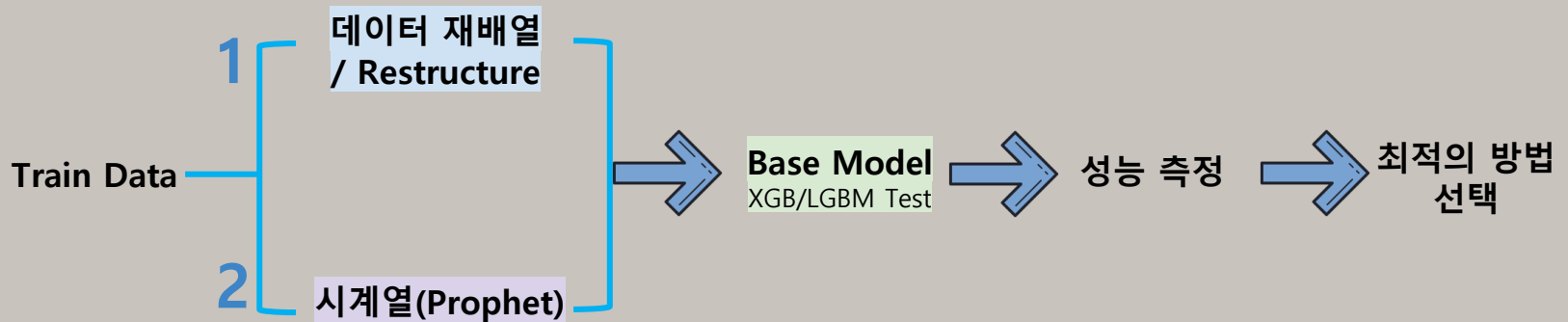
예측 대상

- 24시간 후 결로현상
- 48시간 후 결로현상

방법론

1. 현재 시간의 설명변수로 24시간 후 / 48시간 후 결로 발생 여부 예측
2. 예측해야하는 24시간 후 48시간 후의 설명변수들로 그 시간의 결로현상을 예측

방법론 선택 과정



3. 데이터 처리 방안 및 분석기법

3.1-1) 방법론 1(재배열 / Restructure) 데이터셋 구축

시간	대기 온도 _1번 위치	상대 습도 _1번 위치	대기 온도 _2번 위치	상대 습도 _2번 위치	대기 온도 _3번 위치	상대 습도 _3번 위치	외부대기 온도	외부상대 습도	코일표면 온도 _1번위치	코일표면 온도_2번 위치	코일표면 온도_3번 위치	결로발생 _1번위치	결로발생 _2번위치	결로발생 _3번위치
2016-04-01 0:00	16	24	11	14	23	11	13	32	10	9	42	1	0	1
2016-04-01 3:00	14	28	10	12	32	9	11	42	7	7	59	0	1	0
2016-04-01 6:00	13	33	10	11	37	9	10	44	7	6	56	0	0	1
2016-04-02 0:00	13	33	10	11	35	9	10	41	8	18	30	0	1	0
2016-04-02 3:00	16	28	10	15	27	11	14	30	9	18	20	1	1	0
2016-04-02 6:00	10.52	36.39	10.52	10.55	35.2	11.17	9.85	37.88	9.79	6.4	43.86	1	1	0
2016-04-03 0:00	10.52	37.09	10.44	10.55	35.53	11.09	9.88	38.4	9.72	6.34	44.02	0	1	1
2016-04-03 3:00	10.43	37.18	10.56	10.49	35.65	11.21	9.85	38.22	9.91					
2019-04-03														

Train 데이터의 각 Row별 24시간 / 48시간 후의 결로현상 데이터 추출

시간	공장	공장 내부 위치	대기온도	대기 상대 습도	코일 표면 온도	외부 대기 온도	외부 대기 상대습도	24시간 후 일자 및 시간	24시간 후 결로발생여부 예측 값	48시간 후 일자 및 시간	48시간 후 결로발생 여부 예측 값
2016-04-01 0:00	1	1	16	24	10	13	32	2016-04-02 0:00	0	2016-04-03 0:00	0
2016-04-01 0:00	1	2	11	14	9	13	32	2016-04-02 0:00	1	2016-04-03 0:00	1
2016-04-01 0:00	1	3	23	11	42	13	32	2016-04-02 0:00	0	2016-04-03 0:00	1
2016-04-01	1	1	14	28							

"Train 데이터를 24시간 / 48시간 후의 결로 현상을 예측할 수 있는 Test 데이터셋 형식으로 재배열

3. 데이터 처리 방안 및 분석기법

3.1-1) 방법론 1(재배열 / Restructure) 데이터셋 구축 - 데이터 전처리

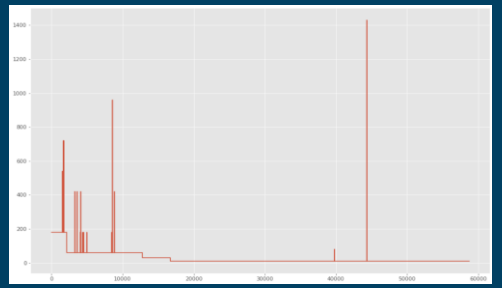
<Train24 이상치 확인>

```
0 days 00:10:00.000000000    42040
0 days 01:00:00.000000000    10575
0 days 00:30:00.000000000     3936
0 days 03:00:00.000000000     2182
0 days 02:00:00.000000000        5
0 days 07:00:00.000000000        4
0 days 01:20:00.000000000        1
3 days 12:00:00.000000000        1
0 days 23:50:00.000000000        1
NaT                              1
0 days 12:00:00.000000000        1
0 days 09:00:00.000000000        1
0 days 16:00:00.000000000        1
Name: time_interval, dtype: int64
```

<Train48 이상치 확인>

```
0 days 00:10:00.000000000    42040
0 days 01:00:00.000000000    10640
0 days 00:30:00.000000000     3936
0 days 03:00:00.000000000     2161
0 days 07:00:00.000000000        4
0 days 02:00:00.000000000        4
0 days 01:20:00.000000000        1
3 days 12:00:00.000000000        1
0 days 23:50:00.000000000        1
NaT                              1
0 days 12:00:00.000000000        1
0 days 09:00:00.000000000        1
0 days 16:00:00.000000000        1
Name: time_interval, dtype: int64
```

<좌측의 시간간격 패턴>



데이터의 시간 간격 분포를 확인해봤을 때
간격이 일정하지 않은 데이터 존재

모수가 적기 때문에 해당 케이스는
무시하고 데이터셋 구축하기로 결정

<분석에 사용한 데이터 형식>

시간	공장	공장 내부위치	대기온도	대기 상대습도	코일 표면 온도	외부 대기온도	외부 대기 상대습도	24시간 후 일자 및 시간	24시간 후 결로발생여부 예측 값	48시간 후 일자 및 시간	48시간 후 결로발생여부 예측 값
2016-04-01 0:00	1	1	16	24	10	13	32	2016-04-02 0:00	0	2016-04-03 0:00	0
2016-04-01 0:00	1	2	11	14	9	13	32	2016-04-02 0:00	1	2016-04-03 0:00	1
2016-04-01 0:00	1	3	23	11	42	13	32	2016-04-02 0:00	0	2016-04-03 0:00	1
01-01 3:00	1	1	14	28	7	11					

3. 데이터 처리 방안 및 분석기법

3.1-1) 방법론 2(시계열 - Prophet) 데이터셋 구축

시간	공장	공장 내부위치	대기온도	대기 상대습도	코일 표면 온도	외부 대기온도	외부 대기 상대습도	24시간 후 일자 및 시간	24시간 후 결로발생여부 예측 값	48시간 후 일자 및 시간	48시간 후 결로발생여부 예측 값
2019-04-01 0:00	1	1	16	24	10	13	32	2019-04-02 0:00	NA	2019-04-03 0:00	NA
2019-04-01 0:00	1	2	11	14	9	13	32	2019-04-02		2019-04-03	NA
2019-04-01 0:00	1	3	23	11	42	13					

+24 Hour

+48 Hour

시간	공장	공장 내부위치	대기온도	대기 상대습도	코일 표면 온도	외부 대기온도	외부 대기 상대습도
2019-04-02 0:00	1	1	16	24	10	13	32
2019-04-02 0:00	1	2	10	14	9	12	32
2019-04-02 0:00	1	3	21	19	41	13	31
2019-04-02 3:00	1	1	14	28	7	11	42
2019-04-02	1	2					

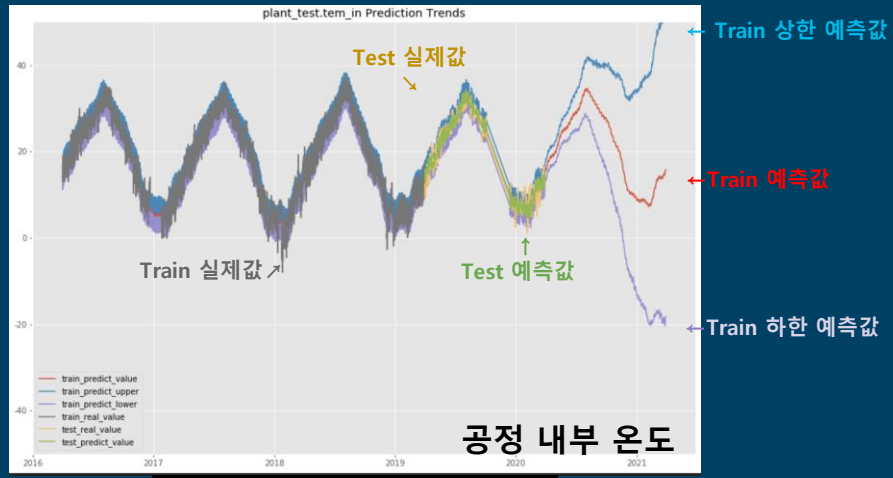
시간	공장	공장 내부위치	대기온도	대기 상대습도	코일 표면 온도	외부 대기온도	외부 대기 상대습도
2019-04-03 0:00	1	1	15	24	19	13	32
2019-04-03 0:00	1	2	11				

" 예측해야하는 24시간 / 48시간 후의 설명 변수값을 시계열 모델(Prophet)로 예측 "

3. 데이터 처리 방안 및 분석기법

3.1-2) Prophet 예측 성능 검증

>> 시계열 예측 성능을 높이기 위해서 Test Data 설명 변수들도 같이 학습한 결과

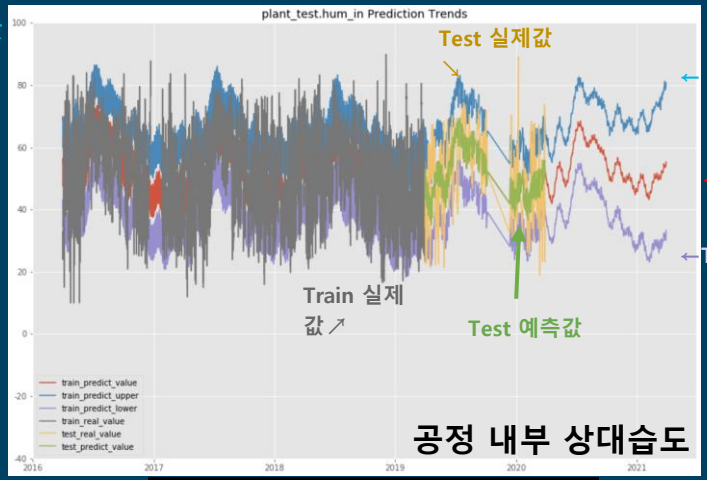


← Train 상한 예측값

←Train 예측값

←Train 하한 예측값

Predict RMSE : 1.8552



← Train 상한 예측값

←Train 예측값

←Train 하한 예측값

Predict RMSE : 11.6850

- 예측 패턴 : 온도의 예측 패턴은 실제값과 크게 차이나지 않지만, 습도의 예측 패턴은 실제값과 크게 차이남
- 예측 오차 : 온도는 비교적 낮은 RMSE, 습도는 비교적 높은 RMSE

➡ *"시계열로 예측했을 때 설명변수간 오차율의 차이가 큼. 즉, 각 변수마다 RMSE가 1.5에서 16까지 다양한 값을 보임"*

3. 데이터 처리 방안 및 분석기법

3.1-3) 방법론별 성능 검증

방법론1 데이터셋

방법론2 데이터셋

Base Model (XGB)

실 예측 성능

```
In [ ]: print(pd.Series(clf_xgb24.predict(test)).value_counts())
0.0    3528
1.0      11
dtype: int64

In [48]: print(pd.Series(clf_xgb48.predict(test)).value_counts())
0.0    3538
1.0       1
dtype: int64
```

>>방법론1 실제 예측 분류 [결과1]

```
In [28]: pd.Series(pred24_value).value_counts()
0.0    3539
dtype: int64

In [29]: pd.Series(pred48_value).value_counts()
0.0    3539
dtype: int64
```

>>방법론2 실제 예측 분류 [결과2]

" 실예측(Test) 결과값(방법론1)과 시계열 예측(방법론2)에 대한 오차를 종합적으로 비교하여 판단 "

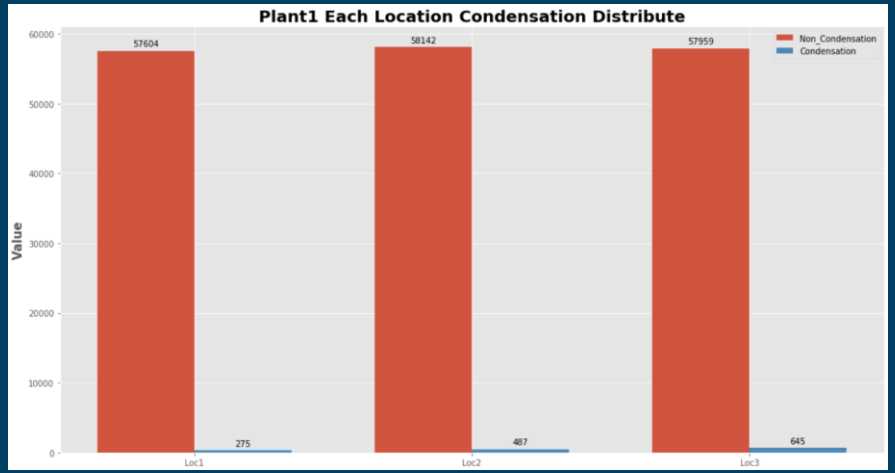
1. 결과2에서 실제 Test 데이터에 대한 예측을 아예 하지 못하는 것을 확인
2. Prophet의 예측 성능 검증 결과를 통해 결로예측오차율의 범위가 더 커질 것으로 예상

방법론1(재배열/Restructure) 채택

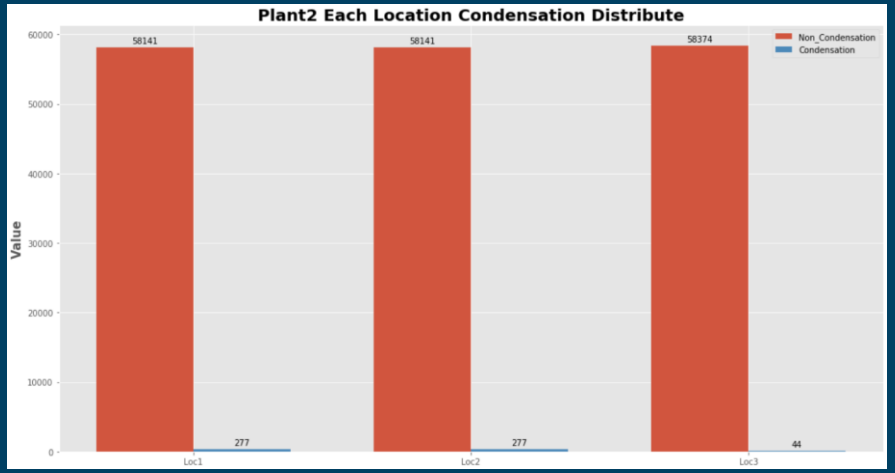
3. 데이터 처리 방안 및 분석기법

3.2 데이터 시각화

<Plant1 - Location별 결로발생현황>



<Plant2 - Location별 결로발생현황>



전체 데이터 중 결로 현상이 나타난 데이터가 약 0.4%로 매우 적음

특히 Plant2에서 데이터 불균형성이 더 심한 것을 확인

→ 과적합 발생 위험 증가

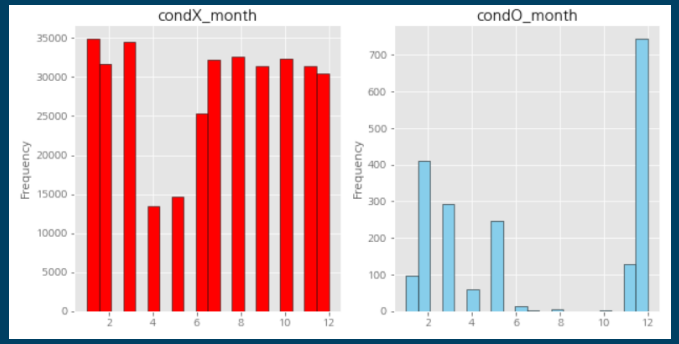
3. 데이터 처리 방안 및 분석기법

3.2 데이터 시각화

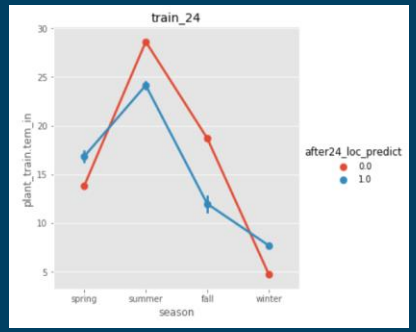
*3 ~ 5월 : spring / 6 ~ 8월 : summer

9 ~ 11월 : fall / 12 ~ 2월 : winter

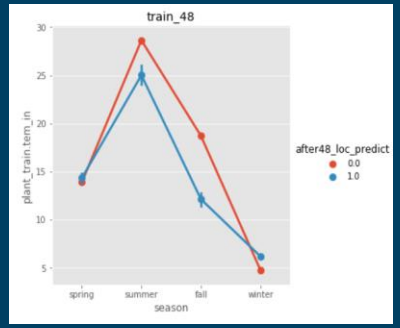
>> 결로 발생 여부 별 월 패턴



월별 패턴 확인



<Train_24>



<Train_48>

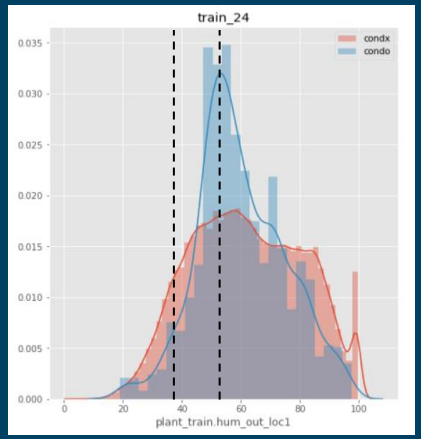
" 결로는 12월에 가장 많이 발생하며 9월에 가장 적게 발생 "

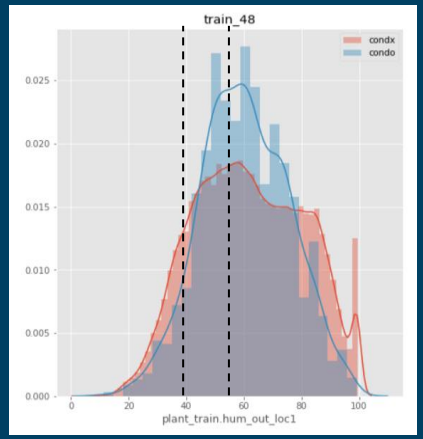
'*Season*' 컬럼 추가

" 여름에 결로가 발생할 가능성이 가장 높고 겨울에 결로가 발생할 가능성이 가장 낮다. "

3. 데이터 처리 방안 및 분석기법

3.2 데이터 시각화

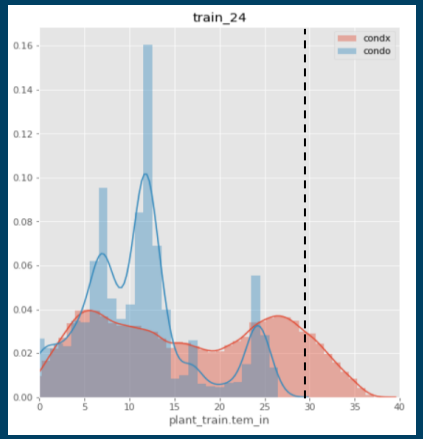


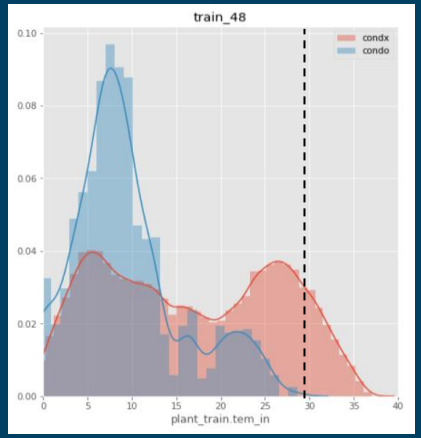


두 데이터 모두 외부 습도가 43~61일때
결로 발생 확률이 높다.

외부 습도가 43 ~ 61일때 1, 그외는 0

파생변수 생성





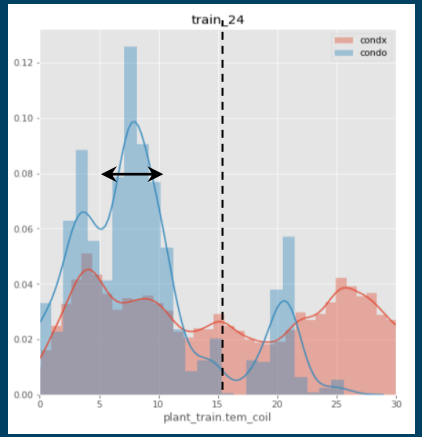
공장 내부온도가 26 이상이면
결로발생확률이 낮다.

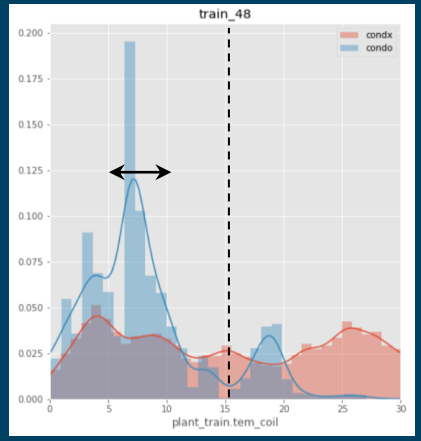
내부 온도가 26 이상이면 0, 아니면 1인

파생변수 생성

3. 데이터 처리 방안 및 분석기법

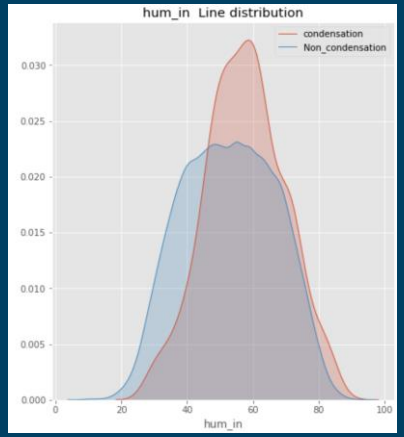
3.2 데이터 시각화

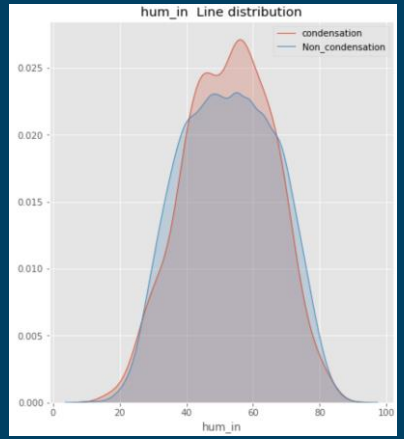




코일 온도가 5~10도일 때 결로 발생 확률이 가장 높으며 25이 상이면 결로 발생 확률이 적다.

코일 온도가 25 이상일 때 0, 아니면 1인 파생변수 생성
코일 온도가 5~10도이면 1, 아니면 0인 파생변수 생성

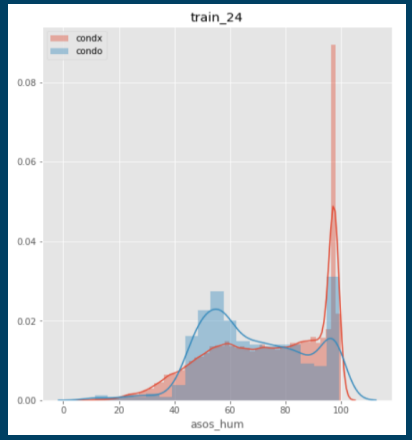


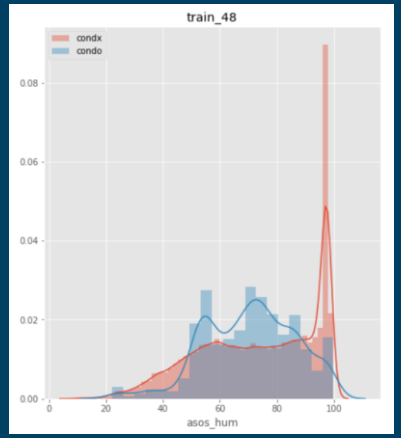


- 24시간 후의 결로현상일수록 내부 습도가 높은 특징을 가진다.
- 48시간 후의 결로 현상여부에 따른 내부 습도는 특정 패턴을 가지고 있지 않다

3. 데이터 처리 방안 및 분석기법

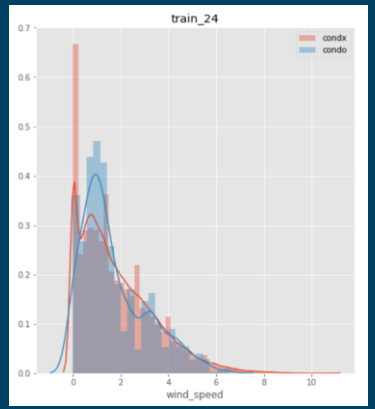
3.2 데이터 시각화



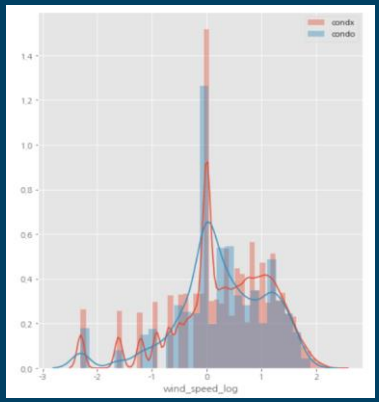


ASOS 서산에서 수집한 습도는 95이상이면
결로발생확률이 적다.

**asos_hum이 95이상이면 0, 아니면 1인
파생변수 생성**



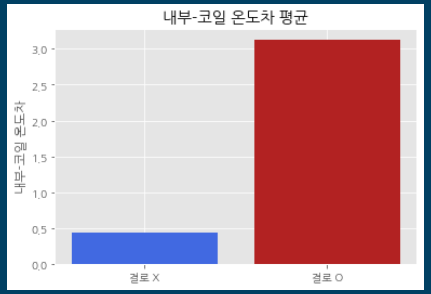
log

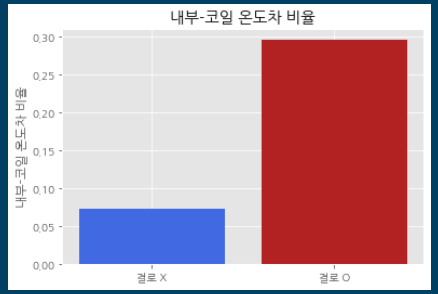


**ASOS 서산에서 수집한 풍속의 분포가 한쪽으로
치우쳐져 있으므로 로그변환한 파생변수 생성**

3. 데이터 처리 방안 및 분석기법

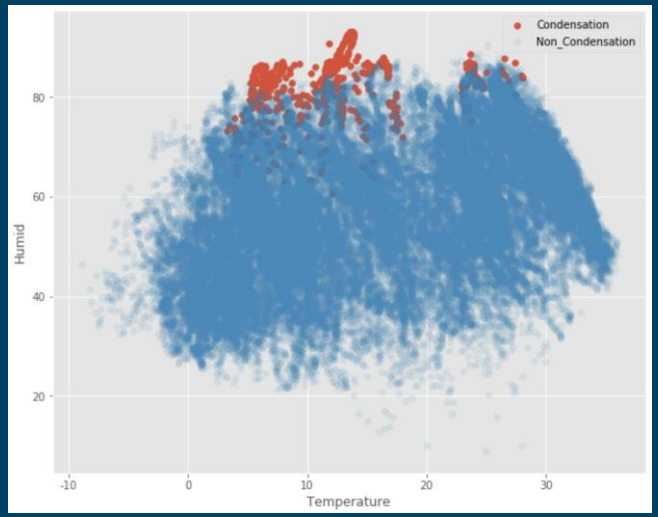
3.2 데이터 시각화





결로현상이 일어나는 경우에 결로현상이 일어나지 않을 때보다
내부 온도와 코일의 온도차의 평균이 큰 것을 알 수 있다.
내부 공기와 접촉하는 코일의 온도의 차이가 약 30% 정도면
결로현상이 일어날 가능성이 더욱 커진다

'내부-외부 온도차', '내부-외부 온도차 비율'
파생변수 생성



결로현상일수록 온도 분포는 크게
연관이 없지만 습도의 분포는 뚜렷하게
차이나는 것을 볼 수 있다.

3. 데이터 처리 방안 및 분석기법

3.3 이슬점 관련 변수 생성

이슬점 변수 생성 이유

이슬점은 결로현상이 의심되거나
우려될 때 기준점으로 사용됨

따라서, 결로현상을 예측하는데
중요한 요인으로 판단하여
이슬점 관련 파생변수 생성

내부 이슬점

$$\gamma(T, \mathrm{RH}) = \ln\left(\frac{\mathrm{RH}}{100}\right) + \frac{bT}{c+T};$$

$$T_{\mathrm{dp}} = \frac{c\gamma(T, \mathrm{RH})}{b - \gamma(T, \mathrm{RH})};$$

b: 상수
c: 상수
T: 내부온도
RH: 습도
r: gamma
Tdp: 이슬점

b와 c는 상수이며, 보편적으로 사용되는 값은 온도 범위 -45 < T < 60에서 오차율 +-0.35%를 보이는 값을 사용
→ b: 17.42 c: 243.12

최소 안전 이슬점

이슬점(Tdp) + 3

내부 이슬점 및 안전 이슬점과 내부 온도 차이

내부온도 - 이슬점(Tdp)
내부온도 - 안전 이슬점(Tdp+3)

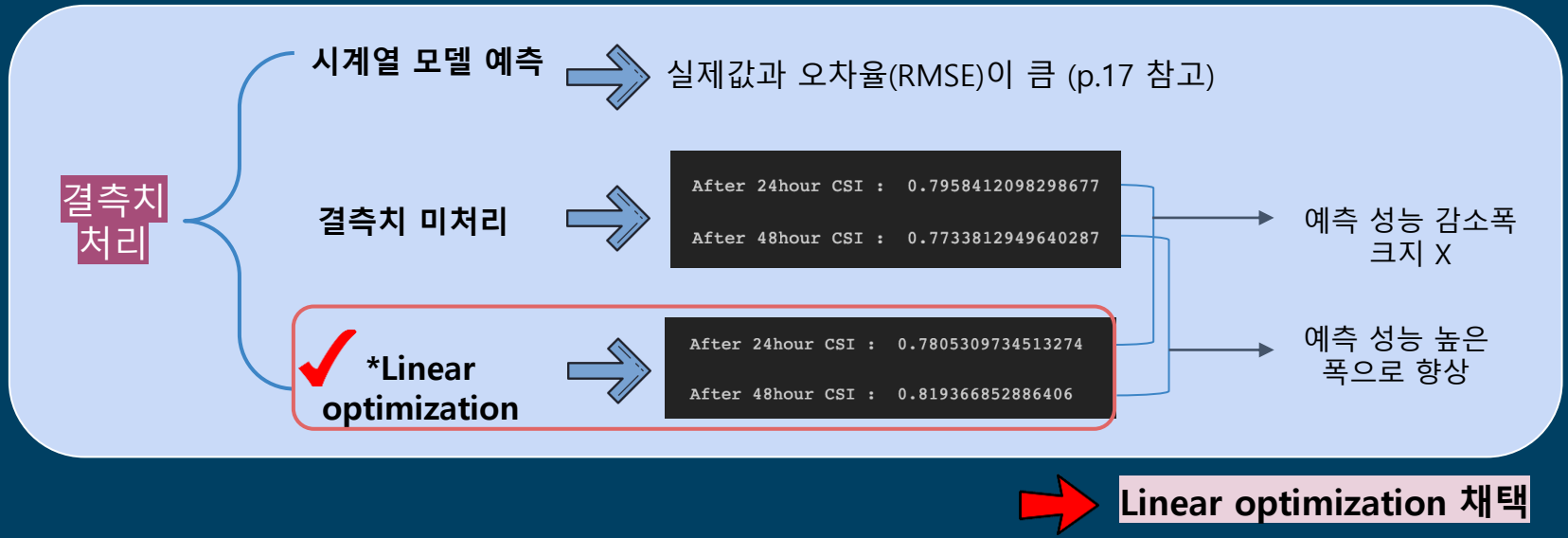
내부 이슬점 및 안전 이슬점과 코일 온도 차이

코일온도 - 이슬점(Tdp)
코일온도 - 안전 이슬점(Tdp+3)

3. 데이터 처리 방안 및 분석기법

3.4 결측치 처리

예측에 사용되는 변수들의 특성 파악
⇒ 온도, 습도, 일사량 등 대부분의 변수가 시계열적인 특성

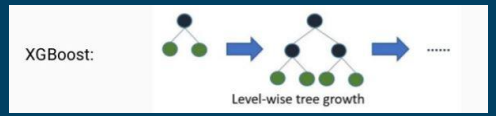


*Linear optimization: 선형적으로 증감하는 날씨 변수의 특성에 맞게 실제 값들 사이의 결측치를 값 사이의 간격 폭으로 증감 값을 구해서 채우는 방식

3. 데이터 처리 방안 및 분석기법

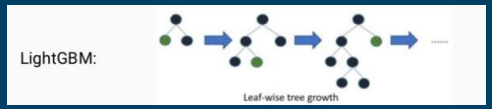
3.5 모델링

XGBoost



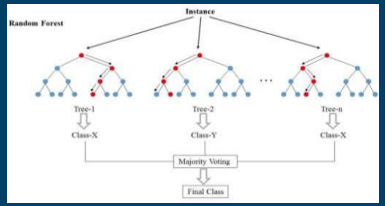
- 약한 분류기를 여러개 만들어서 강한 하나의 분류기를 만든다
- 각 모델에서 생긴 오차를 개선하는 방향으로 학습이 진행된다

LightGBM



- XGB와 다르게 나무를 수직으로 확장 시키며 loss(손실)를 줄이는 방향으로 학습진행
- Goss, EFM방식을 통해서 속도적인 측면을 기존 Boosting모델보다 비약적으로 개선

Random Forest



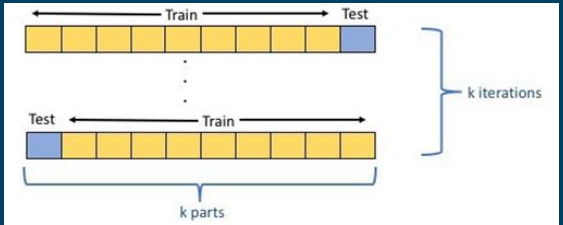
- 의사결정나무를 앙상블한 기법으로 기존 단일 나무에서 생기는 과적합을 방지함
- 하나의 예측에 여러가지 알고리즘 (의사결정 나무)의 결과를 Voting하는 방식으로 최종결과를 반환

*RF는 학습속도가 느리고, 부스팅 기반 모델과 앙상블시 성능 저하를 확인 따라서 RF는 최종 모델에 반영 X

3. 데이터 처리 방안 및 분석기법

3.5-1) 모델링하는 과정에서 공통된 정책

모든 모델에 대해 교차검증 수행(K = 5)



모델 학습과정에서 생길 수 있는 과적합을 방지

평가 Metric

		Predict	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

AUC: Sensitivity / (1 - Specificity)

CSI: TP / (TP + FN + FP)

Parameter

Grid RandomSearch

Ensemble

산술평균 (arithmetic mean)	$\mu = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$

4. 분석결과

4.1 모델링 종합 정확도 – 기본변수

XGB

● **24시간 후**

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	275	224
	Non Condensation	28	86461

CSI : 52.1% / AUC : 0.775

● **48시간 후**

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	185	317
	Non Condensation	17	86190

CSI : 35.6% / AUC : 0.684

LGBM

● **24시간 후**

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	377	122
	Non Condensation	141	86348

CSI : 58.9% / AUC : 0.877

● **48시간 후**

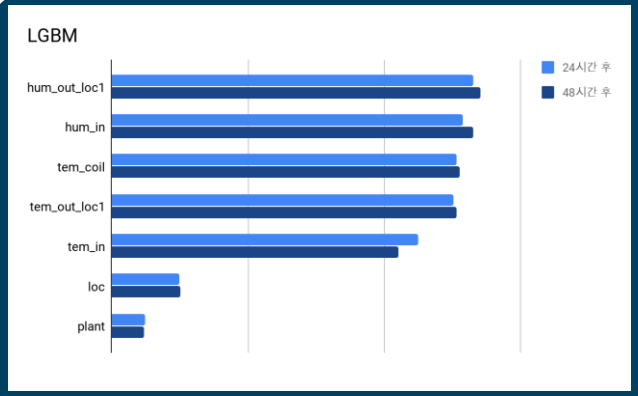
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	319	183
	Non Condensation	199	86008

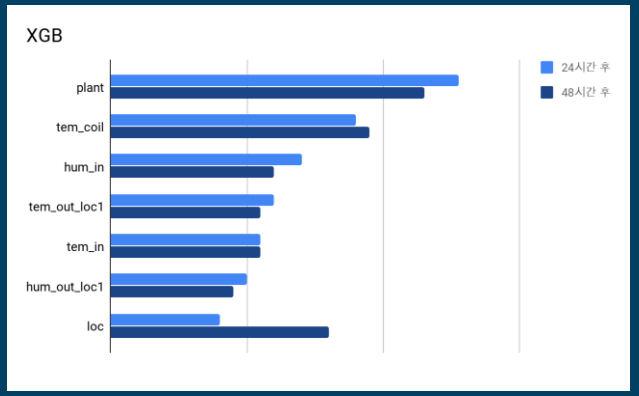
CSI : 45.5% / AUC : 0.816

▶▶ 기본변수의 경우, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 높은 것을 확인

4. 분석결과

4.1-1) 기본변수 변수중요도 및 앙상블





" 전체적으로 모델 성능에 기여하는 중요도가 높은 상위 변수에 습도, 온도와 관련된 변수가 분포해 있는 것을 알 수 있다. "

	confusion matrix				CSI	AUC
Ensemble (24시간 후)	LGBM + XGB		Predict		74.5%	0.916
			Condensation	Non Condensation		
	Actual	Condensation	435	67		
		Non Condensation	134	86073		
Ensemble (48시간 후)	LGBM + XGB		Predict		73.9%	0.920
			Condensation	Non Condensation		
	Actual	Condensation	425	77		
		Non Condensation	124	86083		

4. 분석결과

4.2 모델링 종합 정확도 – 기본변수 + ASOS

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	403	96
	Non Condensation	24	86465

CSI : 77.1% / AUC : 0.904

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	385	117
	Non Condensation	21	86186

CSI : 73.6% / AUC : 0.883

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	474	25
	Non Condensation	306	86183

CSI : 58.9% / AUC : 0.973

● 48시간 후

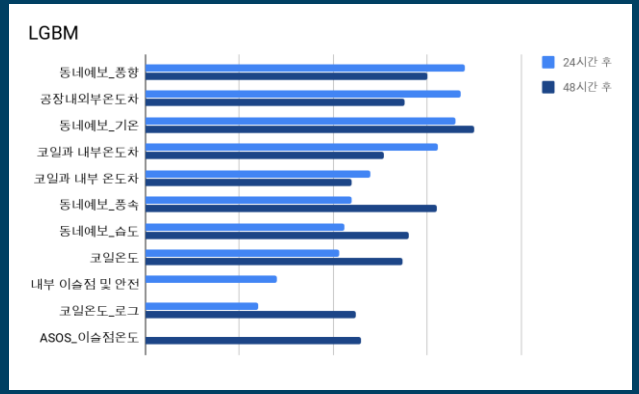
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	458	44
	Non Condensation	110	86097

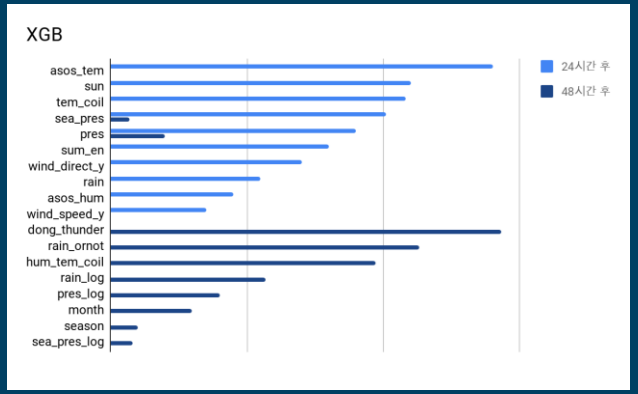
CSI : 74.8% / AUC : 0.956

▶▶ XGB의 경우에는, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 더 높고,
LGBM의 경우에는, 48시간 이후의 예측이 24시간 이후의 예측보다 성능이 더 높은 것을 확인

4. 분석결과

4.2-1) 기본변수+ ASOS 변수중요도 및 앙상블





" 전체적으로 모델 성능에 기여하는 중요도가 높은 상위 변수에 동네예보 관련, 코일 온도 관련된 변수가 분포해 있는 것을 알 수 있다. "

	confusion matrix	CSI	AUC
Ensemble (24시간 후)	LGBM+XGB / Predict: Condensation, Non Condensation Actual: Condensation — 457, 42 Actual: Non Condensation — 121, 86368	73.7%	0.957
Ensemble (48시간 후)	LGBM+XGB / Predict: Condensation, Non Condensation Actual: Condensation — 458, 44 Actual: Non Condensation — 110, 86097	74.8%	0.956

4. 분석결과

4.3 모델링 종합 정확도 – 기본변수 + AWS

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	389	110
	Non Condensation	22	86467

CSI : 74.1% / AUC : 0.889

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	383	119
	Non Condensation	17	86190

CSI : 73.7% / AUC : 0.881

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	444	55
	Non Condensation	104	86385

CSI : 73.6% / AUC : 0.944

● 48시간 후

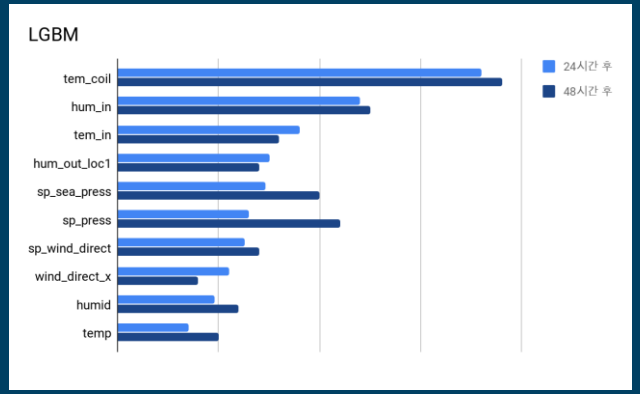
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	447	55
	Non Condensation	101	86106

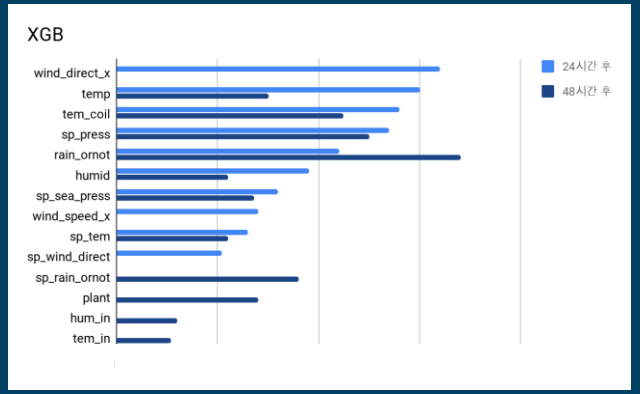
CSI : 74.1% / AUC : 0.944

▶▶ XGB의 경우에는, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 더 높고,
LGBM의 경우에는, 48시간 이후의 예측이 24시간 이후의 예측보다 성능이 더 높은 것을 확인

4. 분석결과

4.3-1) 기본변수 + AWS 변수중요도 및 앙상블





	confusion matrix				CSI	AUC
Ensemble (24시간 후)	LGBM + XGB		Predict		76.2%	0.918
			Condensation	Non Condensation		
	Actual	Condensation	418	81		
		Non Condensation	49	86440		
Ensemble (48시간 후)	LGBM + XGB		Predict		77.2%	0.918
			Condensation	Non Condensation		
	Actual	Condensation	420	82		
		Non Condensation	42	86165		

" "

전체적으로 기본 데이터와
aws데이터도 마찬가지로
온도 / 습도 변수가 모델에서 중요한
팩터로 작용하는 것을 알 수 있다.

4. 분석결과

4.4 모델링 종합 정확도 – 기본변수 + 동네예보

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	407	92
	Non Condensation	36	86463

CSI : 77.5% / AUC : 0.907

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	369	133
	Non Condensation	14	86193

CSI : 71.5% / AUC : 0.867

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	474	25
	Non Condensation	581	85908

CSI : 43.8% / AUC : 0.971

● 48시간 후

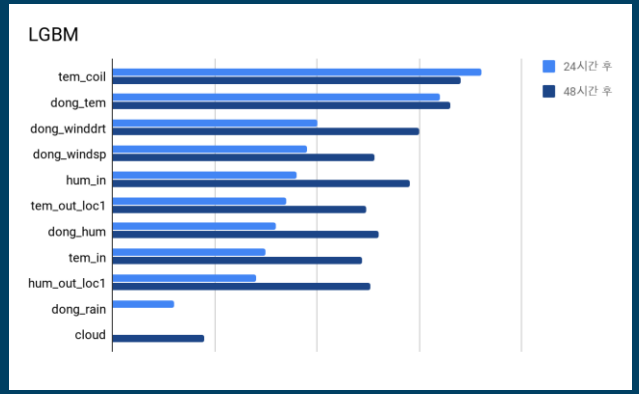
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	453	49
	Non Condensation	130	86077

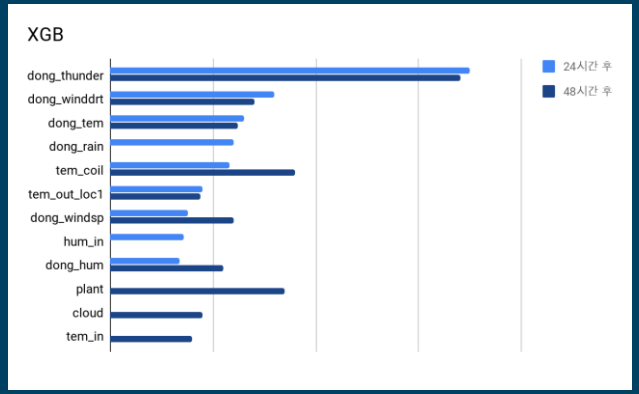
CSI : 71.6% / AUC : 0.950

▶▶ XGB의 경우에는, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 더 높고,
LGBM의 경우에는, 24시간 이후의 CSI성능이 48시간 이후의 CSI성능보다 더 낮지만,
AUC는 더 높은 것을 확인

4. 분석결과

4.4-1) 기본변수 + 동네예보 변수중요도 및 앙상블





	confusion matrix	CSI	AUC
Ensemble (24시간 후)	LGBM + XGB: Predict (Condensation / Non Condensation); Actual (Condensation: 446, 53; Non Condensation: 101, 86388)	74.3%	0.946
Ensemble (48시간 후)	LGBM + XGB: Predict (Condensation / Non Condensation); Actual (Condensation: 412, 90; Non Condensation: 39, 86168)	76.1%	0.910

" 동네예보 관련 변수들도 결로현상에 영향을 미치는 온도 / 습도 관련 변수가 큰 팩터로 모델에 작용하고 있는 것을 알 수 있다. "

4. 분석결과

4.5 모델링 종합 정확도 – 기본변수 + AWS + ASOS + 동네예보

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	418	81
	Non Condensation	32	86457

CSI : 78.7% / AUC : 0.918

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	385	117
	Non Condensation	21	86186

CSI : 73.6% / AUC : 0.883

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	457	42
	Non Condensation	21	86368

CSI : 73.7% / AUC : 0.957

● 48시간 후

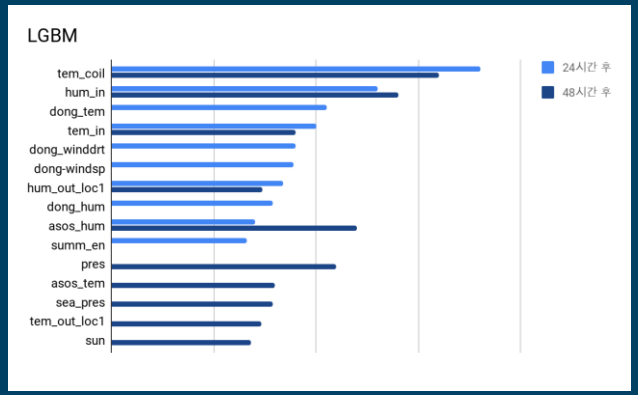
XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	458	44
	Non Condensation	110	86097

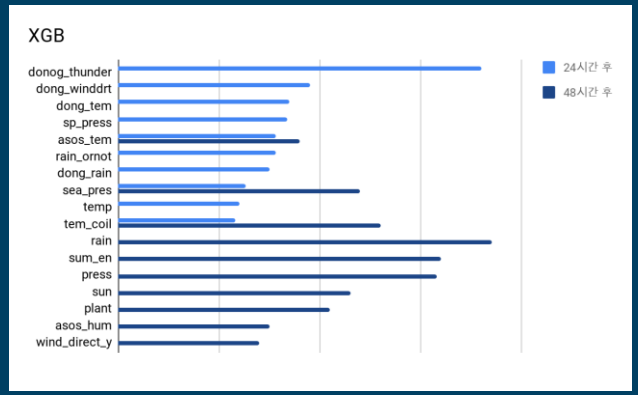
CSI : 74.8% / AUC : 0.956

▶▶ XGB의 경우에는, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 더 높고,
LGBM의 경우에는, 48시간 이후의 CSI성능이 24시간 이후의 CSI성능보다 높지만,
AUC는 조금 더 낮은 것을 확인

4. 분석결과

4.5-1) 기본변수 + AWS + ASOS + 동네예보 변수 중요도 및 앙상블





	confusion matrix	CSI	AUC
Ensemble (24시간 후)	LGBM+XGB / Predict: Condensation, Non Condensation; Actual: Condensation (449, 52), Non Condensation (102, 86385)	79.7%	0.957
Ensemble (48시간 후)	LGBM+XGB / Predict: Condensation, Non Condensation; Actual: Condensation (428, 84), Non Condensation (36, 86161)	74.8%	0.956

" 전체적으로 모델 성능에 기여하는 중요도가 높은 상위 변수에 동네예보 관련, 코일 온도 관련된 변수가 분포해 있는 것을 알 수 있다. "

4. 분석결과

4.6 모델링 종합 정확도 – 기본변수 + AWS + ASOS + 동네예보 + 파생변수

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	275	224
	Non Condensation	28	86461

CSI : 77.5% / AUC : 0.911

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	185	317
	Non Condensation	17	86190

CSI : 78% / AUC : 0.902

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	458	41
	Non Condensation	132	86357

CSI : 72.5% / AUC : 0.958

● 48시간 후

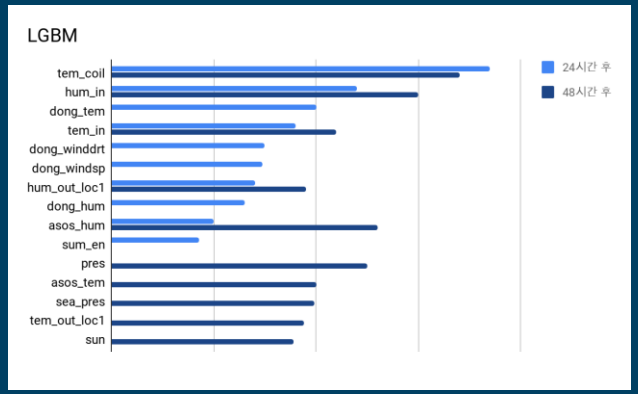
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	465	37
	Non Condensation	104	86103

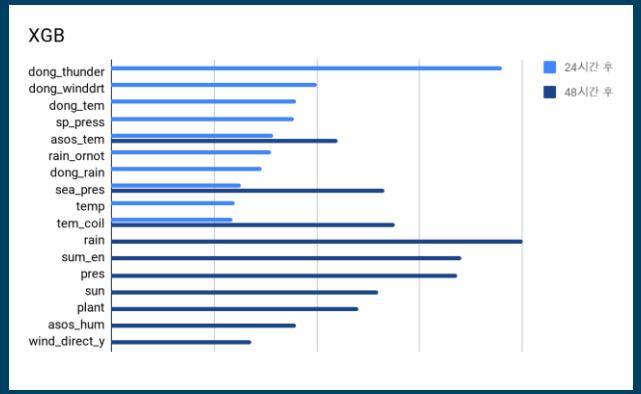
CSI : 76.7% / AUC : 0.962

▶▶ XGB의 경우에는, 24시간 이후의 CSI성능이 48시간 이후의 CSI성능보다 낮지만 AUC는 더 높고, LGBM의 경우에는, 48시간 이후의 예측이 24시간 이후의 예측보다 성능이 더 높은 것을 확인

4. 분석결과

4.6-1) 기본변수 + AWS + ASOS + 동네예보 + 파생변수 변수중요도 및 앙상블





" 전체적으로 모델 성능에 기여하는 중요도가 높은 상위 변수에 동네예보 관련, 코일 온도, aws와 관련된 변수가 분포 해 있는 것을 알 수 있다. "

	confusion matrix	CSI	AUC
Ensemble (24시간 후)	LGBM + XGB: Actual (Condensation: 445, 57 / Non Condensation: 124, 86083) vs Predict (Condensation, Non Condensation)	76.5%	0.927
Ensemble (48시간 후)	LGBM + XGB: Actual (Condensation: 440, 62 / Non Condensation: 129, 86078) vs Predict (Condensation, Non Condensation)	79.8%	0.932

4. 분석결과

4.7 모델링 종합 정확도 – 기본변수 + AWS + ASOS + 동네예보 + 파생변수 + 이슬점

XGB

● 24시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	412	87
	Non Condensation	29	86460

CSI : 78% / AUC : 0.912

● 48시간 후

XGB		Predict	
		Condensation	Non Condensation
Actual	Condensation	403	99
	Non Condensation	18	86189

CSI : 77.5% / AUC : 0.901

LGBM

● 24시간 후

LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	451	48
	Non Condensation	105	86384

CSI : 74.6% / AUC : 0.951

● 48시간 후

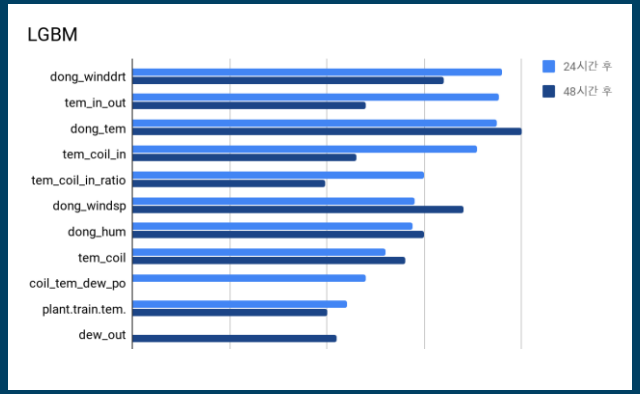
LGBM		Predict	
		Condensation	Non Condensation
Actual	Condensation	469	33
	Non Condensation	135	86072

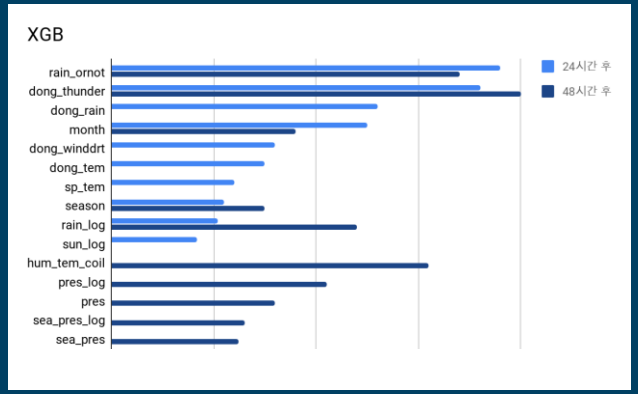
CSI : 73.6% / AUC : 0.966

▶▶ XGB의 경우에는, 24시간 이후의 예측이 48시간 이후의 예측보다 성능이 더 높고,
LGBM의 경우에는, 24시간 이후의 CSI성능이 48시간 이후의 CSI성능보다 높지만,
AUC는 더 낮은 것을 확인

4. 분석결과

4.7-1) 기본변수 + AWS + ASOS + 동네예보 + 파생변수 + 이슬점 변수중요도 및 앙상블





" 전체적으로 모델 성능에 기여하는 중요도가 높은 상위 변수에 동네예보 관련, 코일 온도 관련된 변수가 분포해 있는 것을 알 수 있다. "

	confusion matrix	CSI	AUC
Ensemble (24시간 후)	LGBM + XGB: Actual (Condensation: 436, 63 / Non Condensation: 56, 86433) vs Predict (Condensation, Non Condensation)	78.5%	0.936
Ensemble (48시간 후)	LGBM + XGB: Actual (Condensation: 440, 62 / Non Condensation: 35, 86172) vs Predict (Condensation, Non Condensation)	81.9%	0.938

4. 분석결과

4.8 하이퍼파라미터 최적화 및 변수소거

Grid
Randomsearch

Min_child_weight
Subsample
Gamma
Colsample_bytree
Max_depth
Learning_rate
N_estimators

: 파라미터를 증가시키면서 탐색
- 파라미터가 증가할수록 오버피팅을 감소 시킴

: 파라미터를 감소시키면서 탐색
- 파라미터가 감소할수록 오버피팅을 감소 시킴

Baseline Optimization

LGBM

gamma : 0.4
colsample_bytree : 0.8
min_child_weight : 0.5
subsample : 0.8
max_depth : 5
learning_rate : 0.6
n_estimators : 1000
Early_Stopping : 300

XGB

gamma : 0.5
colsample_bytree : 0.7
min_child_weight : 0.4
subsample : 0.9
max_depth : 5
learning_rate : 0.7
n_estimators : 1000
Early_Stopping : 300

예측 데이터셋

XGB

24시간 후 결로 예측

48시간 후 결로 예측

LGBM

24시간 후 결로 예측

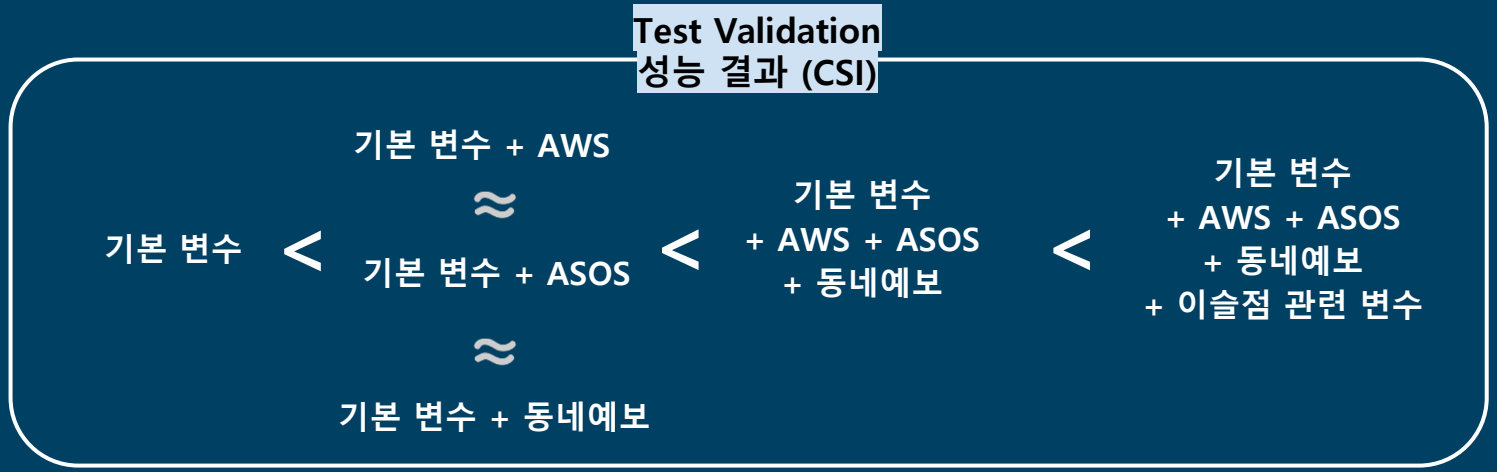
48시간 후 결로 예측

변수중요도가 0인 변수 소거

최종 결과에 반영

4. 분석결과

4.9 최종 제출 모델 및 데이터



변수 중요도 및 Validation 종합 성능을 통해서 24시간 뒤, 48시간 뒤의 결로 예측 모델에 사용되는 변수를 선택하면서
동네예보와 관련된 변수, 이슬점 변수가 중요한 역할을 하는 것을 확인

따라서 이슬점은 결로 현상에 영향을 주는 중요한 요인으로 작용하며,
각 공정에 가장 가까운 동네예보 데이터가 결로 현상을 예측하는데 큰 영향을 미치는 것으로 판단

4. 분석결과

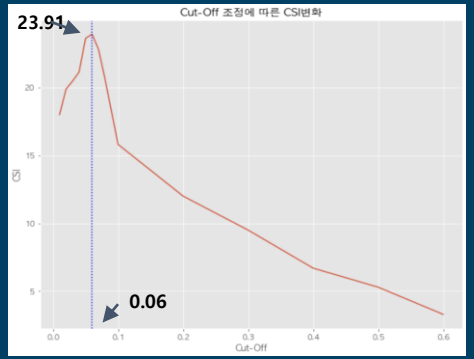
4.9-1) 최종 제출 모델 및 데이터

데이터 불균형

결과 신뢰성

과적합

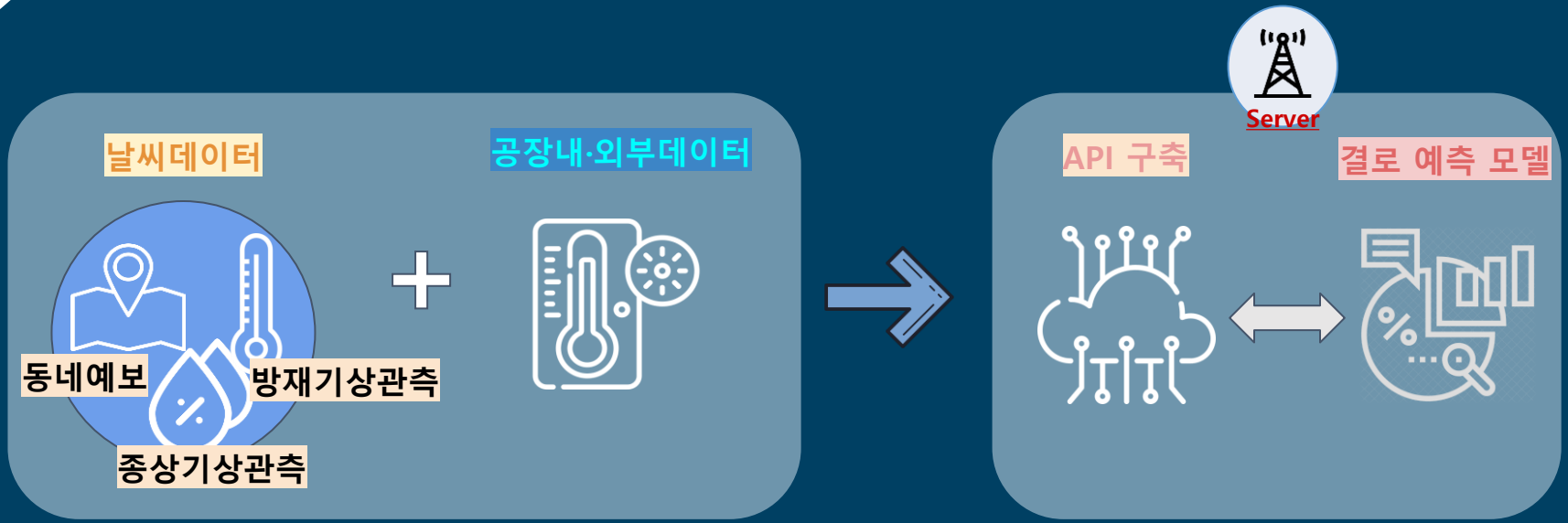
성능 최적화를 위한 Cut-Off 조정 ▶ Test Validation의 한계점 파악 ▶ Test Validation으로 추출한 CSI와의 차이가 존재 ▶ 과적합을 막기 위한 제반 장치 (CV, hyper parameter)를 두어도 해당 문제를 완전히 해결치 못함 ▶ Public Score로 최종 검증하기로 결정



▶ 0.06을 임계점으로 성능이 반등하는 것을 볼 수 있음 ▶ 최종 Cut-off 0.06으로 설정

5. 활용방안 및 기대효과

5.1 예측 정보 시스템



>> 결로현상예측에 필요한 input data를 연동하는 api구축

5. 활용방안 및 기대효과

5.1 예측 정보 시스템

Server

결로 예측 모델

A공장 B위치에서
24시간 후 결로 발생 예측

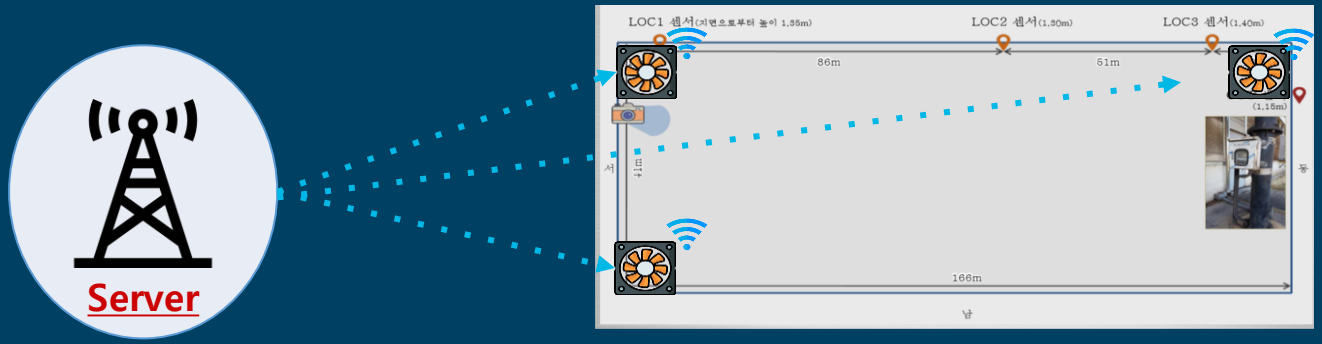
현대제철 근무자용 앱
클라이언트로 값 전달

개별 근로자에게
알림경보

사용자의 판단으로
유연한 대처 가능

5. 활용방안 및 기대효과

5.1 예측 정보 시스템



온도, 습도가 적정 임계치 초과시 송풍팬 자동 가동 → 효율적인 품질관리

매각 후 재고 창고로 사용 할 전기로 열연 공장에 도입

6. 활용 데이터 및 분석 도구

● 활용 데이터 목록

데이터	출처	기준년도
공장 내·외부 기상, 코일온도, 결로발생여부 관측데이터	현대제철	2016년 4월 ~ 2020년 3월
서산 종관기상관측(ASOS)데이터	기상청(기상자료개방포털)	2016년 4월~2020년 3월
당진 방재기상관측(AWS)데이터	기상청(기상자료개방포털)	2016년 4월~2020년 3월
신평 방재기상관측(AWS)데이터	기상청(기상자료개방포털)	2016년 4월~2020년 3월
충청남도 당진시 송악읍 동네예보/초단기실황분석자료	기상청(기상자료개방포털)	2016년 4월~2020년 3월

● 분석도구

전처리, 모델링

전처리

● 참고문헌

금융감독원 전자공시시스템
- 현대제철 증권신고서(채무증권) 예비투자설명서

7. 콘테스트 공모 문제해결 과정별 팀원 참여도

구분	신민용	김지윤	이다정
문제이해 및 자료조사	30	30	40
데이터 전처리	40	30	30
데이터 모델링	40	30	30
분석결과 정리 및 보고서 작성	35	30	35
활용 방안 아이디어 제시	30	35	35

감사합니다.