



기상청 빅데이터 콘테스트 결로예측모델 개발

513호 세입자

우나영 / 조주영 / 한효선

CONTENTS

01 목표

02 변수설명

- 기존데이터
- 외부데이터
- 파생변수

03 전처리 및 EDA

04 모델링

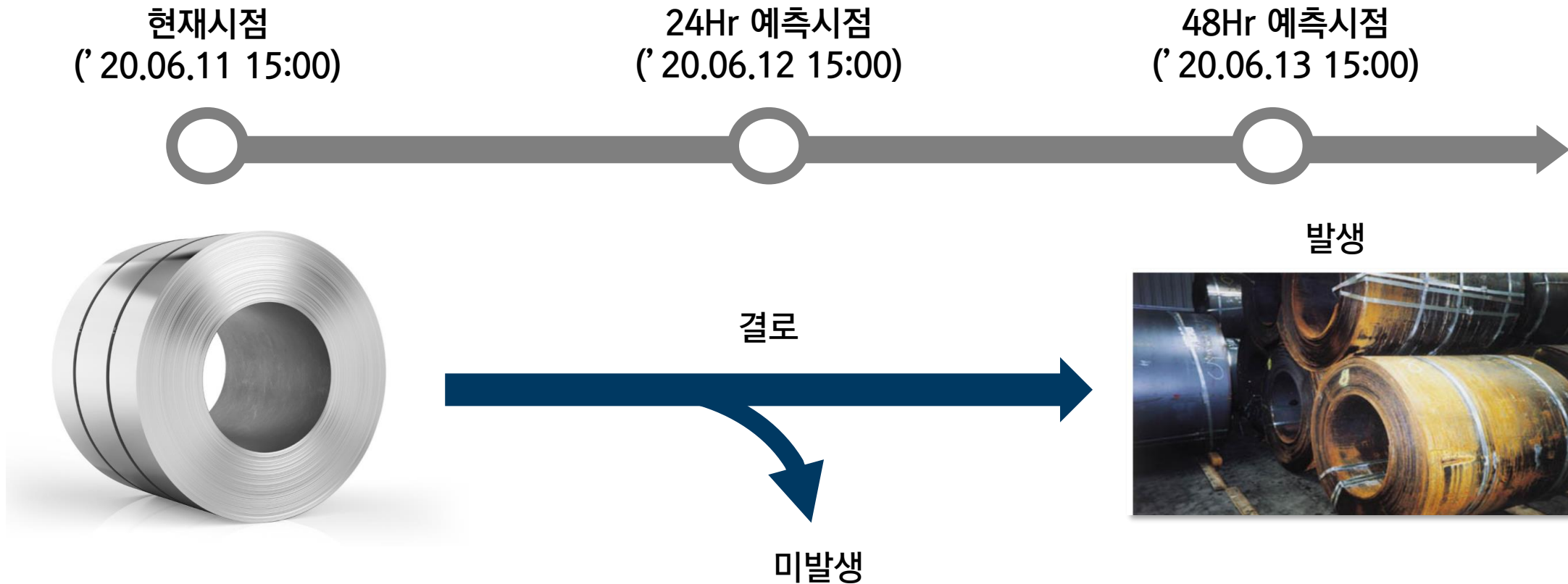
05 활용방안

A blue-tinted photograph of an industrial facility, likely a power plant or refinery, featuring several tall smokestacks and complex piping structures. The image is framed by a white border.

01 목표

Problem : 결로

- 공장 내 제품 결로 발생 시 경제적 손실이 크다.
- 결로 방지 시스템 구동을 위해서는 최소 24시간 이전에 결로를 예측해야 한다.



공장 내 보관 중인 코일의 결로 발생 예측 모형 개발이 필요하다.

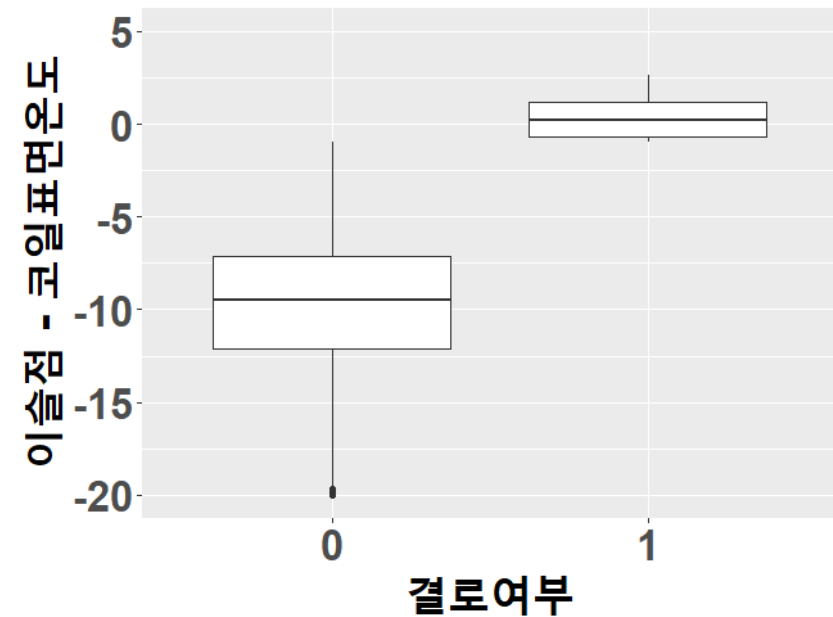
Problem : 결로

결로

- 표면 온도 < 이슬점
- 물체의 표면에 물방울이 맺히는 현상
- 이슬점 공식

$$D_p = \frac{243.12 \cdot (\ln\left(\frac{Rh}{100}\right) + \frac{17.62 \cdot T}{243.12 + T})}{17.62 - (\ln\left(\frac{Rh}{100}\right) + \frac{17.62 \cdot T}{243.12 + T})}$$

Where T : 온도, Rh : 습도



표면온도가 이슬점보다 낮을 때 결로가 발생한다.

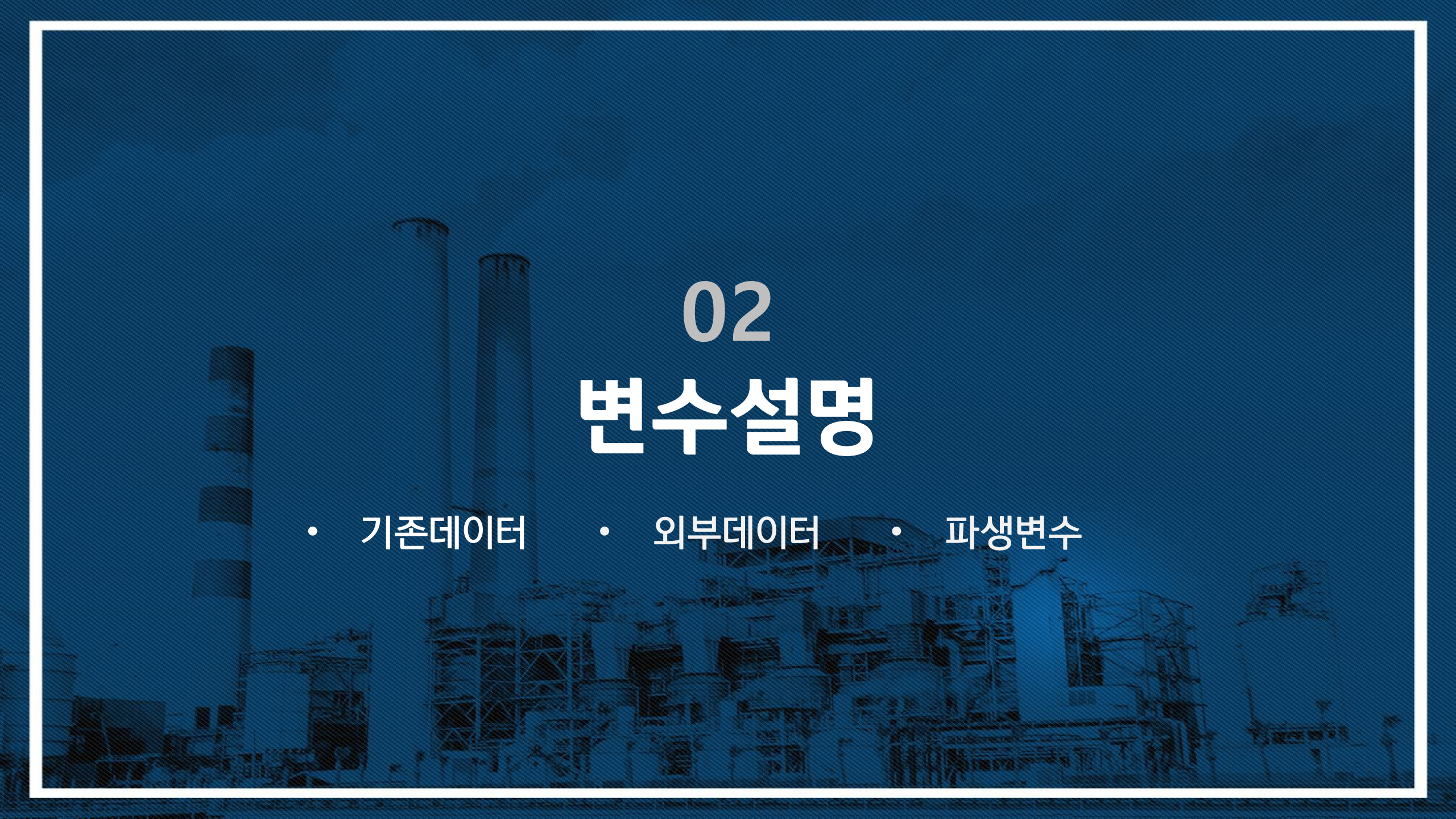
결로

현재 시점에 주어진 자료만 사용하여 결로 발생을 예측하는 모형 개발

Where T : 온도, Rh : 습도

결로여부

표면온도가 이슬점보다 낮을 때 결로가 발생한다.



02

변수설명

- 기존데이터
- 외부데이터
- 파생변수

변수설명 : 기존 데이터

[Plant1, Plant2 train data]

Variable	STR	
MEA_DDHR	Character	데이터 측정 일자 및 시간
tem_in_loc	Num	공장 내부 대기온도 측정값
hum_in_loc	Num	공장 내부 대기 상대습도 측정값
tem_out_loc	Num	공장 외부 대기온도 측정값
hum_out_loc	Num	공장 외부 대기 상대습도 측정값
tem_coil_loc	Num	공장 내부 철강제품(코일) 표면 온도 측정값
cond_loc	Int	공장 내부 결로발생 여부

변수설명 : 기존 데이터

[Plant1, Plant2 train data]

Variable	STR	
MEA_DDHR	Character	데이터 측정 일자 및 시간
tem_in_loc	Num	공장 내부 대기온도 측정값
hum_in_loc	Num	공장 내부 대기 상대습도 측정값
tem_out_loc	Num	공장 외부 대기온도 측정값
hum_out_loc	Num	공장 외부 대기 상대습도 측정값
tem_coil_loc	Num	공장 내부 철강제품(코일) 표면 온도 측정값
cond_loc	Int	공장 내부 결로발생 여부

변수설명 : 기존 데이터

[Plant test data]

Variable	STR	
MEA_DDHR	Character	데이터 측정 일자 및 시간
plant	Num	공장
loc	Num	공장 내부 위치
tem_in	Num	(해당 공장 및 위치) 대기온도 측정값
hum_in	Num	(해당 공장 및 위치) 대기 상대습도 측정값
tem_coil	Num	(해당 공장 및 위치) 철강제품(코일) 표면 온도 측정값
tem_out_Loc1	Num	(해당공장) 외부 대기온도 측정값
hum_out_Loc1	Num	(해당 공장) 외부 대기 상대습도 측정값
X24H/X48H_TMA	Character	(기준) 24, 48시간 후 일자 및 시간
X24H/X48H_cond_Loc	Int	(기준) 24, 48시간 후 결로발생여부 예측 값
cond_Loc_prob	Int	24, 48시간 후 결로발생여부 예측 확률

기존데이터 관측 시간 단위

[Plant1, Plant2 train data]

Variable
일자
공장내부습도
공장외부기온
코일표면온도
결로여부

Train data

Plant1 - 기간 : 2016-04-01 00:00 ~ 2019-03-31 23:50

Plant2 - 기간 : 2016-07-01 18:00 ~ 2019-03-31 23:50

Test data

기간 : 2019-04-01 0:00 ~ 2020-03-31 23:50

관측기간별 관측 시간 단위

기간 : 2016-04-01 00:00 ~ 2016-12-26 21:00 → 3시간

기간 : 2016-12-26 22:00 ~ 2018-03-22 00:00 → 1시간

기간 : 2018-03-22 00:30 ~ 2018-06-12 00:00 → 30분

기간 : 2018-06-12 00:10 ~ 2020-03-31 23:50 → 10분

변수설명 : 기존 데이터

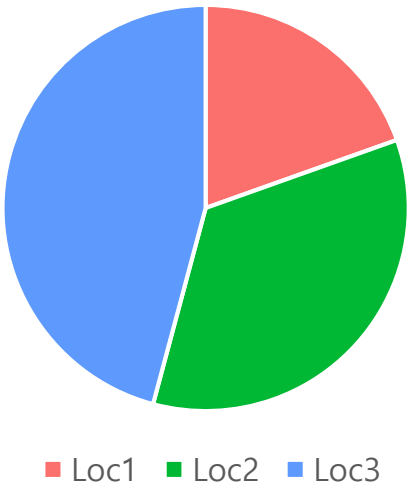
[Plant1, Plant2 train data]

Variable
일자
공장내부습도
공장외부기온
코일표면온도
결로여부

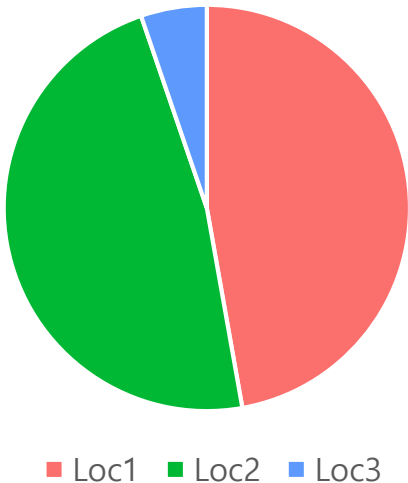
[전체 결로 발생 비율]



[Plant1 location별 결로 비율]



[Plant2 location별 결로 비율]



전체 데이터에서 **결로 발생 비율은 1% 미만으로 매우 낮은** 것을 확인할 수 있다. 즉 결로 데이터는 불균형 데이터이다.
이를 plant, location별로 나누어 더 자세히 살펴보면, plant1에서는 Loc3, Loc2, Loc1순으로 결로가 많이 발생하는 것을 알 수 있다.
이는 앞서 확인한 plant1 공장 내 습도 그래프에서 Loc3, Loc2, Loc1순으로 습도가 높아 나타난 결과라고 생각된다. 또한 코일 표면온도가 가장 높았던 plant1 Loc1, plant2 Loc3가 각 plant에서 가장 결로 발생 비율이 적은 점을 볼 때 코일 표면온도가 결로 여부에 영향을 미친다고 추측할 수 있다.

전체 결로 발생 비율은 매우 적다.

변수설명 : 외부 데이터

[외부데이터 from ASOS, AWS]

Variable	STR	
temp(129, 616, 637)	Num	(해당 관측소) 대기온도 측정값
wind_dir(616, 637)	Num	(해당 관측소) 풍향 측정값
wind_speed(129, 616, 637)	Num	(해당 관측소) 풍속 측정값
humid(129, 616)	Num	(해당 관측소) 상대습도 측정값
spot_press(637)	Num	(해당 관측소) 현지기압 측정값
sea_press(637)	Num	(해당 관측소) 해면기압 측정값
rain_cum(129)	Num	(해당 관측소) 누적 강수량
rain_status(616, 637)	factor	(해당 관측소) 강수유무 측정값
rain(637)	factor	(해당 관측소) 1분 강수량 측정값
solar_rad(129)	Num	(해당 관측소) 일사 측정값
solar_amount(129)	Num	(해당 관측소) 일조 측정값

변수설명 : 외부 데이터

[외부데이터 from 동네예보]

Variable	STR	
temp_3hour(24, 48)	Num	기온(3시간) 예보
humid(24, 48)	Num	습도 예보
wind_dir(24, 48)	Num	풍향 예보
wind_speed(24, 48)	Num	풍속 예보
rain_prob(24, 48)	Num	강수 확률 예보
sky_status(24, 48)	factor	하늘 상태 예보

24 : 24시간 뒤 예측에 사용되는 변수

48 : 48시간 뒤 예측에 사용되는 변수

변수설명 : 파생변수

Variable	STR	
month	Num	데이터 측정 월
day	Num	데이터 측정 일
season	Num	계절 (1 : 3-5월 / 2 : 6-8월 / 3 : 9-11월 / 4 : 12-2월)
plant	Num	공장
location	Num	공장 내부 위치
time	Num	예측 시점 (24시간 후, 48시간 후)
dp_in_loc	Num	공장 내부 1/2/3번 위치 이슬점 값
dp129	Num	서산 이슬점 값
fore_dp	Num	이슬점 예보 값



03

전처리 및 EDA

NA 처리

Train data

전체 데이터 중 NA는 0.6%로 데이터분석에 영향이 적을것으로 판단되어 해당 row를 제거했다.

Test data

모델링에 사용된 외부변수를 test 데이터에 join 해보았을 때 전체데이터의 9% 정도가 NA로 나타난다.

적합된 모델을 사용하여 결로 예측을 하기 위해서는 반드시 해당 NA를 채워넣어야 한다.

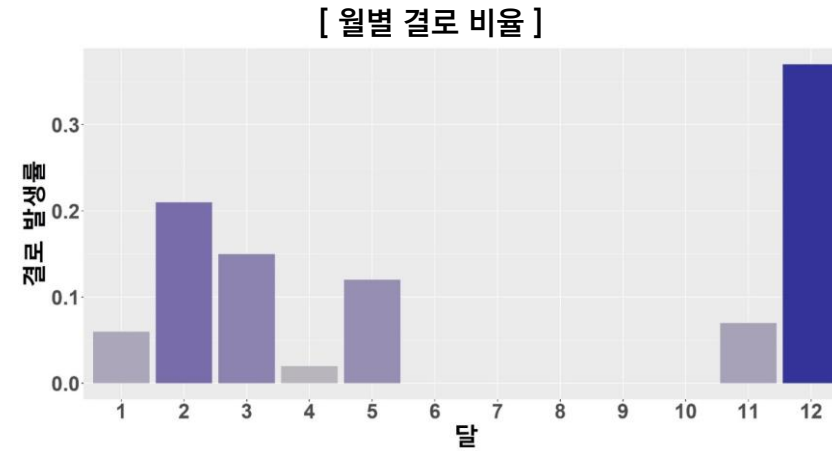
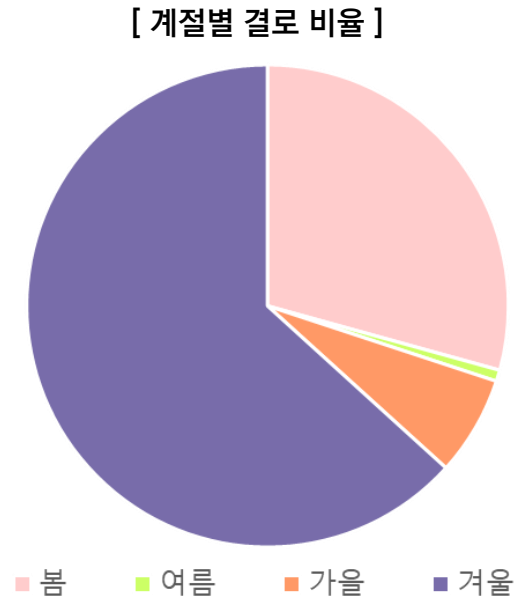
NA는 다음과 같은 규칙을 사용하여 처리하였다.

1. 기온, 풍향, 풍속, 습도, 현지기압, 해면기압, 1분 강수량, 강수 유무

- 결측치가 존재하는 경우 **가까운 지역의 관측값으로 대체**하였다.
- 예) 서산 결측치는 당진 AWS에서 측정된 값으로 대체

2. 누적강수량, 일사, 일조

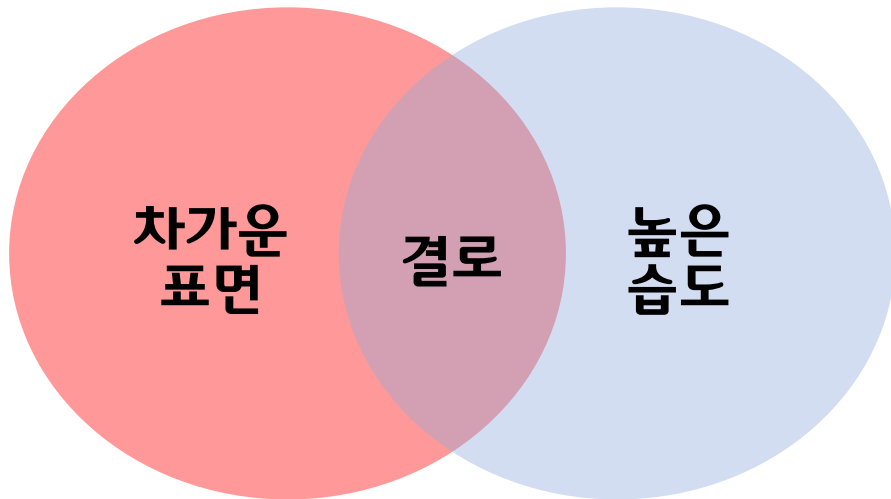
- 서산지역에서만 측정 가능한 기상변수는 다른 지역의 값으로 대체할 수 없기 때문에 **KNN을 사용**하여 처리하였다.



계절별 결로 비율을 살펴보면 **겨울에 결로가 가장 많이 나타나**는 것을 알 수 있다.
계절을 더 세부적으로 나누어보면 12월과 2월에 결로 발생률이 가장 높음을 확인할 수 있다..

결로는 겨울철에 많이 발생한다.

앞에서 살펴봤듯이
결로는 다음과 같은 상황에서 발생한다.

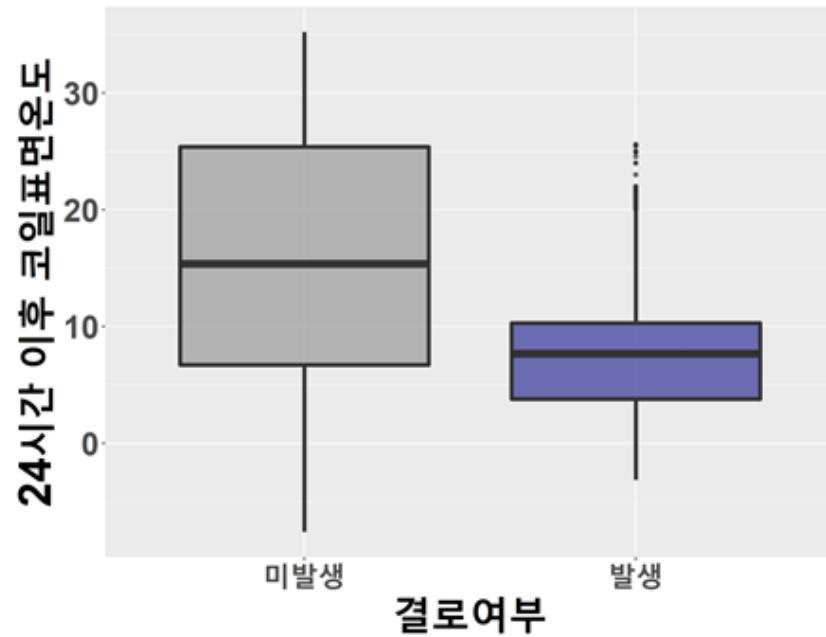


따라서 다음과 같은 기상변수들이 결로 예측에
많은 영향을 미칠 것이라고 예상하고 EDA를 진행하였다.

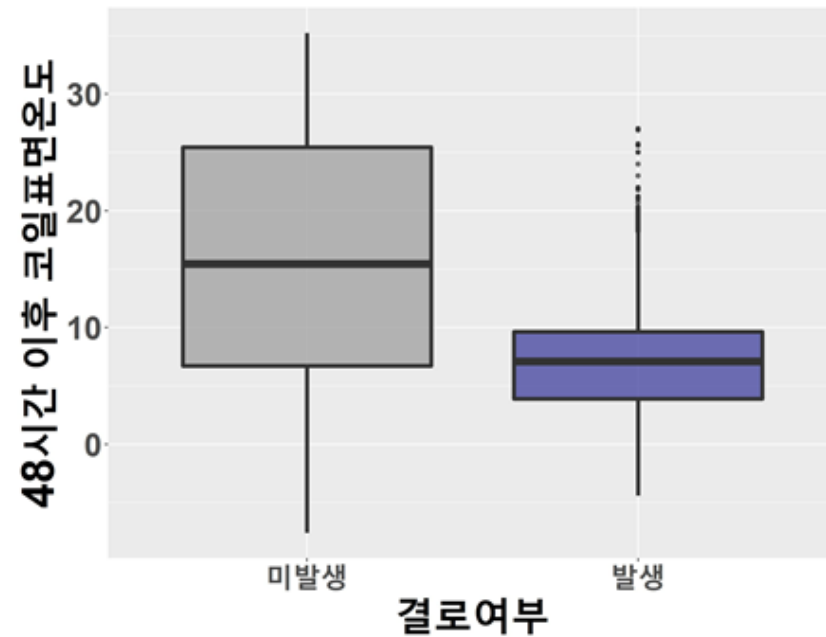
1. 코일 표면 온도
2. 이슬점 온도
3. 습도

EDA : 코일 표면 온도

[결로 여부 별 코일 표면 온도 분포 (24시간)]



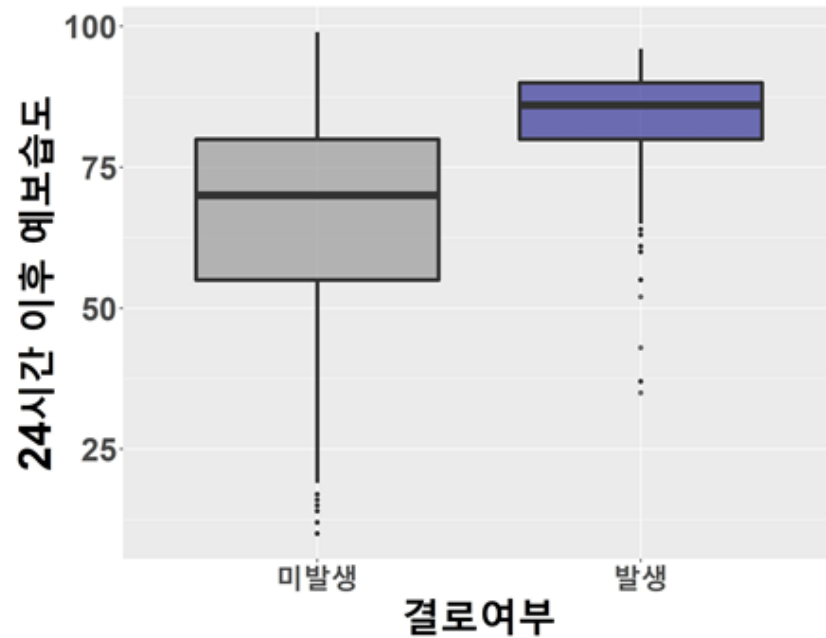
[결로 여부 별 코일 표면 온도 분포 (48시간)]



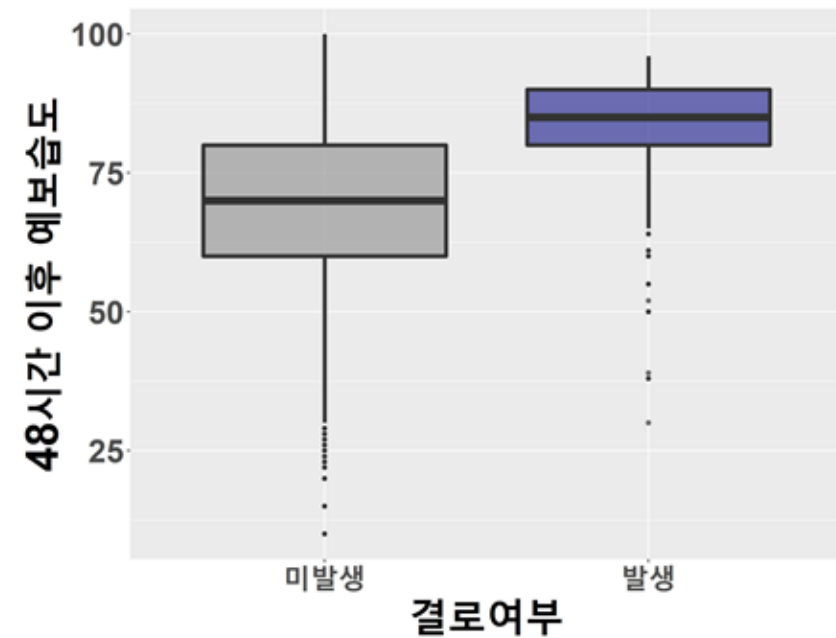
코일 표면 온도가 낮을수록 결로 발생률이 높다.

EDA : 예보 습도

[결로 여부 별 예보 습도 분포 (24시간)]



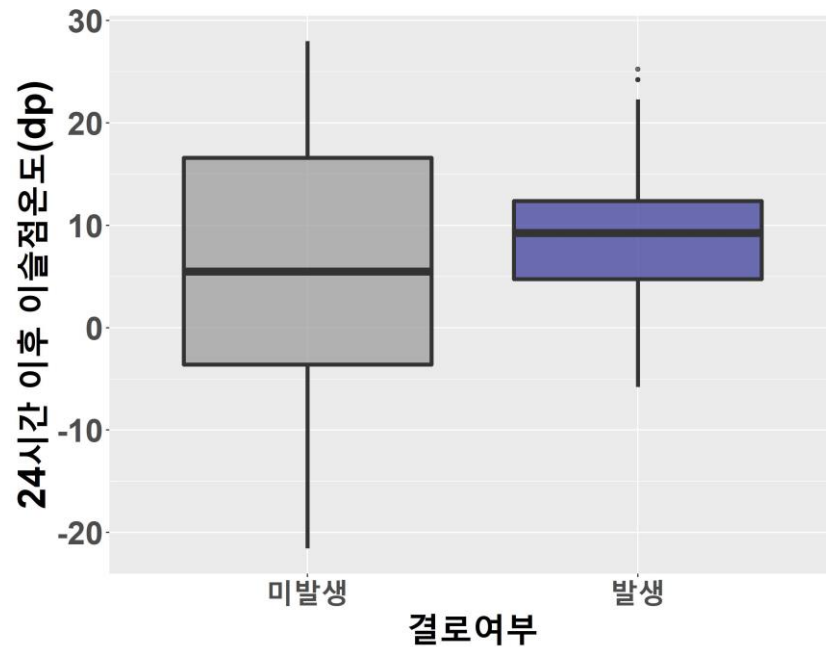
[결로 여부 별 예보 습도 분포 (48시간)]



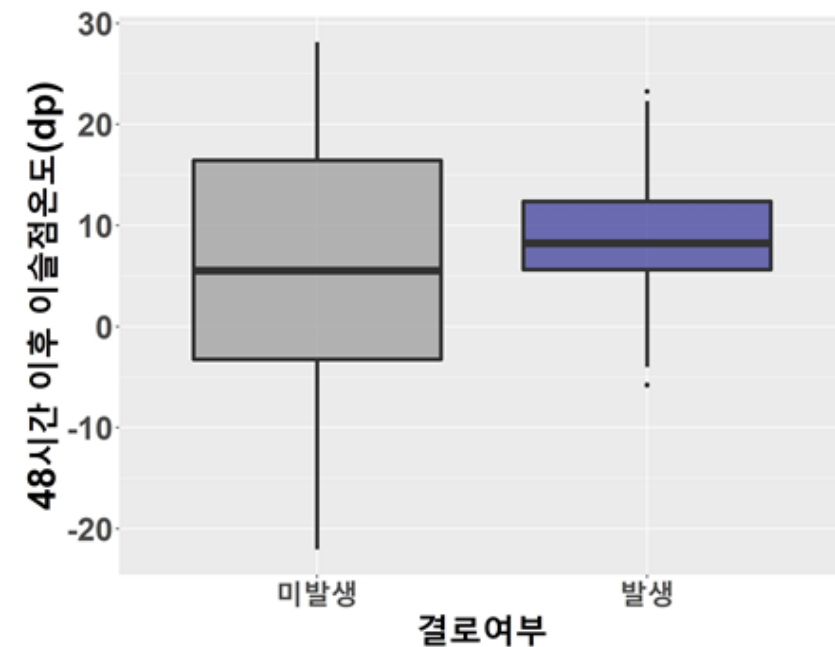
예보 습도가 높을수록 결로 발생률이 높다.

EDA : 예보 이슬점온도

[결로 여부 별 예보 이슬점온도 분포 (24시간)]



[결로 여부 별 예보 이슬점온도 분포 (48시간)]

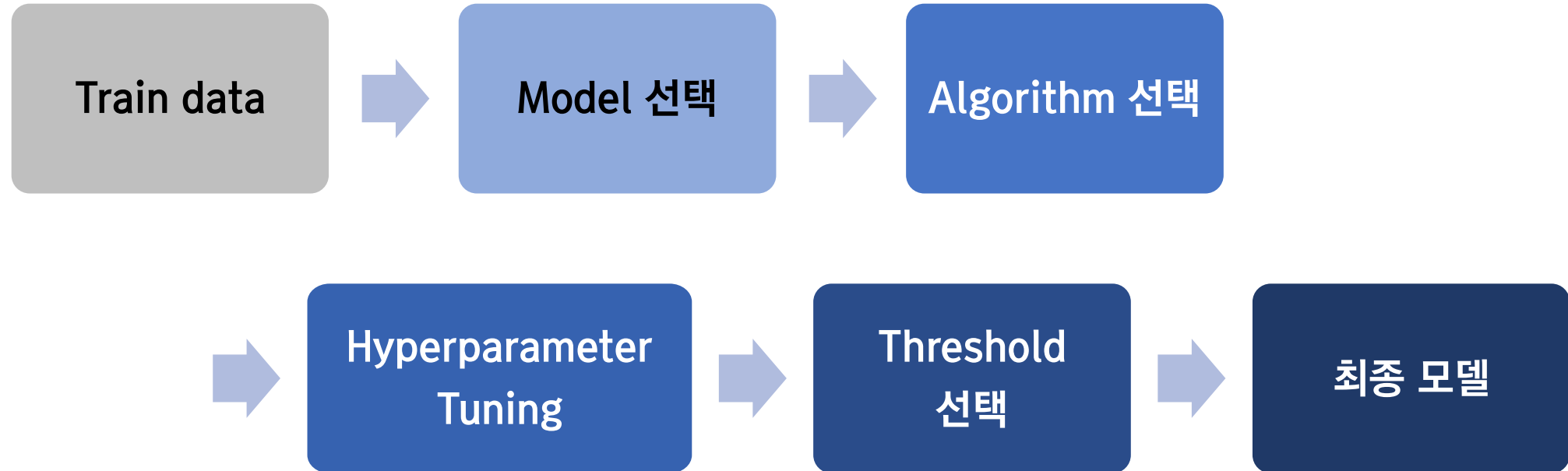


예보 이슬점온도가 높을수록 결로 발생률이 높다.

A blue-tinted photograph of an industrial facility, likely a power plant or refinery, featuring several tall smokestacks and complex piping structures. The image is framed by a white border.

04 모델링

Modeling 순서



Confusion Matrix

		True Condition	
		True	False
Predicted Condition	True	True positive (TP)	False positive (FP)
	False	False negative (FN)	True negative (TN)

$$CSI = \frac{TP}{TP+FP+FN} \times 100$$

- 예측과 관측에서 사건발생과 관련된 경우 총합에서 옳은 예측의 비율을 나타낸다.
- 즉, 사건발생을 잘못 예측한 경우(FP), 사건 발생을 예측하지 못한 경우(FN) 모두 패널티를 부과한다.
- 100에 가까울수록 완벽한 예측 분류를 의미한다.
- 사건의 수가 작거나 범주의 기상학적 빈도수가 거의 같은 경우에도 실질적인 정보 제공이 가능하다.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

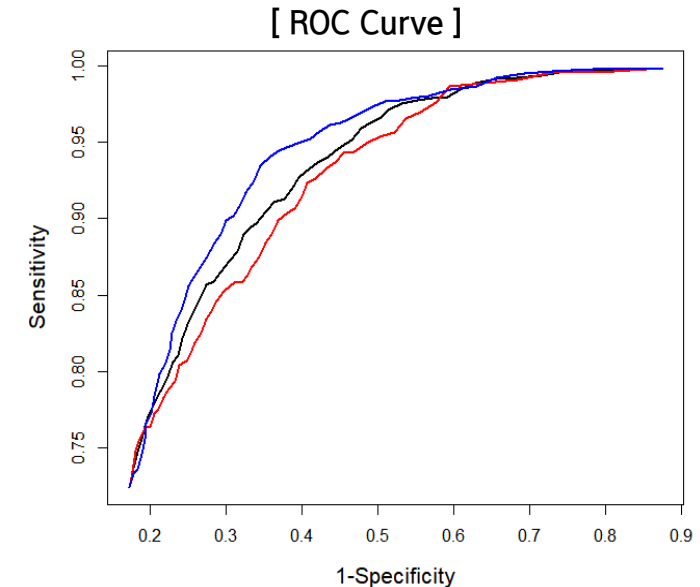
- 실제값이 1인 것 중 1이라고 예측한 정도를 나타낸다.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

- 실제값이 0인 것 중 0이라고 예측한 정도를 나타낸다.

AUC

- ROC curve의 곡선 아래 영역 넓이를 나타내며 클수록 좋은 성능을 의미한다.



Model 선택

Model	설명		AUC	CSI (Threshold : Train 결로 평균)
1 model	<ul style="list-style-type: none"> Plant, location, time 변수 추가 		0.9842	63.96%
2 model(plant)	<ul style="list-style-type: none"> Plant1 모델 Plant2 모델 Location, time 변수 추가 		0.9656	59.06%
2 model(time)	<ul style="list-style-type: none"> Time24 모델 Time48 모델 Plant, location 변수 추가 		0.9735	62.54%
4 model	<ul style="list-style-type: none"> Plant1, Time24 모델 Plant1, Time48 모델 Plant2, Time24 모델 Plant2, Time48 모델 Location 변수 추가 		0.96	54.86%
12 model	<ul style="list-style-type: none"> Plant1, Time24, location1 모델 Plant1, Time24, location2 모델 Plant1, Time24, location3 모델 Plant1, Time48, location1 모델 Plant1, Time48, location2 모델 Plant1, Time48, location3 모델 	<ul style="list-style-type: none"> Plant2, Time24, location1 모델 Plant2, Time24, location2 모델 Plant2, Time24, location3 모델 Plant2, Time48, location1 모델 Plant2, Time48, location2 모델 Plant2, Time48, location3 모델 	0.9413	48.57%

총 모델의 수를 정하기 위해 각 모델을 평가해보았다. 이 때 모델링은 **LightGBM을 이용**하여 수행하였고 결과는 **리더보드의 값을 이용**하였다.

1 model의 성능이 가장 좋기 때문에 변수가 각 plant, location에 미치는 영향이 비슷하다고 추측할 수 있다.

성능이 가장 좋은 1 model을 선택하였다.

1 model structure

- Plant1 train

Mea_ddhr	tem_in_loc1	hum_in_loc1	tem_coil_loc1	...	tem_in_loc3	hum_in_loc3	tem_coil_loc3	tem_out_loc1	hum_out_loc1	Cond_loc1	Cond_loc2	Cond_loc3
2016-04-01 0:00	16.0	24	11		13	32	10	9	42	0	0	0
2016-04-01 3:00	14.0	28	10		11	42	7	7	59	0	0	0
2016-04-01 6:00	13.0	33	10		10	44	7	6	56	0	0	0



Mea_ddhr	tem_in_loc	hum_in_loc	tem_coil_loc	...	fore_dp	month	day	season	Plant	Location	time	Cond_loc
2016-04-01 0:00	16	24	11		5.647393	4	1	1	1	1	24	0
2016-04-01 0:00	14	23	11		5.647393	4	1	1	1	2	24	0
2016-04-01 0:00	13	32	10		5.647393	4	1	1	1	3	24	0
2016-04-01 0:00	16	24	11		6.2280047	4	1	1	1	1	48	0
2016-04-01 0:00	14	23	11		6.2280047	4	1	1	1	2	48	0
2016-04-01 0:00	13	32	10		6.2280047	4	1	1	1	3	48	0

1 model structure

- Plant1 train

Mea_ddhr	tem_in_loc1	hum_in_loc1	tem_coil_loc1	...	tem_in_loc3	hum_in_loc3	tem_coil_loc3	tem_out_loc1	hum_out_loc1	Cond_loc1	Cond_loc2	Cond_loc3
2016-04-01 0:00	16.0	24	11		13	32	10	9	42	0	0	0
2016-04-01 3:00	14.0	28	10		11	42	7	7	59	0	0	0
2016-04-01 6:00	13.0	33	10		10	44	7	6	56	0	0	0



Mea_ddhr	tem_in_loc	hum_in_loc	tem_coil_loc	...	fore_dp	month	day	season	Plant	Location	time	Cond_loc
2016-04-01 0:00	16	24	11		5.647393	4	1	1	1	1	24	0
2016-04-01 0:00	14	23	11		5.647393	4	1	1	1	2	24	0
2016-04-01 0:00	13	32	10		5.647393	4	1	1	1	3	24	0
2016-04-01 0:00	16	24	11		6.2280047	4	1	1	1	1	48	0
2016-04-01 0:00	14	23	11		6.2280047	4	1	1	1	2	48	0
2016-04-01 0:00	13	32	10		6.2280047	4	1	1	1	3	48	0

1 model structure

- Plant1 train

Mea_ddhr	tem_in_loc1	hum_in_loc1	tem_coil_loc1	...	tem_in_loc3	hum_in_loc3	tem_coil_loc3	tem_out_loc1	hum_out_loc1	Cond_loc1	Cond_loc2	Cond_loc3
2016-04-01 0:00	16.0	24	11		13	32	10	9	42	0	0	0
2016-04-01 3:00	14.0	28	10		11	42	7	7	59	0	0	0
2016-04-01 6:00	13.0	33	10		10	44	7	6	56	0	0	0



Mea_ddhr	tem_in_loc	hum_in_loc	tem_coil_loc	...	fore_dp	month	day	season	Plant	Location	time	Cond_loc
2016-04-01 0:00	16	24	11		5.647393	4	1	1	1	1	24	0
2016-04-01 0:00	14	23	11		5.647393	4	1	1	1	2	24	0
2016-04-01 0:00	13	32	10		5.647393	4	1	1	1	3	24	0
2016-04-01 0:00	16	24	11		6.2280047	4	1	1	1	1	48	0
2016-04-01 0:00	14	23	11		6.2280047	4	1	1	1	2	48	0
2016-04-01 0:00	13	32	10		6.2280047	4	1	1	1	3	48	0

Algorithm 선택

Model	Oversampling O		Oversampling X	
	AUC	CSI (Threshold : Train 결로 평균)	AUC	CSI (Threshold : Train 결로 평균)
Ridge Regression	0.9744	7.01%	0.9836	8.49%
Lasso	0.9846	7.33%	0.9946	12.47%
Random Forest	0.9848	19.26%	0.9881	73.92%
SVM	0.7562	24.83%	0.7382	27.96%
XGBoost	0.9604	69.23%	0.979	72.95%
LightGBM	0.9984	69.67%	0.9991	80.61%
Neural Network	0.9856	24.37%	0.9873	25.4%

여러가지 알고리즘을 이용하여 1 model을 모델링하고 평가하였다. 결로 데이터는 불균형 데이터이기 때문에 **oversampling을 시행**해 보았다. 각 알고리즘은 oversampling의 이용 여부에 따라 각각 두번씩 구하였다. 이 때 **cross validation(K = 5)**을 이용하였고, 각 알고리즘은 **파라미터를 조정해가며 가장 높은 성능을 기준**으로 하였다. 결과적으로 Oversampling을 하지 않는 LightGBM의 성능이 가장 좋다.

Oversampling을 하지 않은 LightGBM 알고리즘을 선택하였다.

Hyperparameter Tuning

Parameters

- Learning Rate : 훈련량
- Num Iterations : 부스팅 수행 횟수. 너무 크면 overfitting을 일으킴
- Early Stopping Round : Validation set의 발전이 없으면 그만두게 설정
- Max Depth : 나무 깊이
- Bagging Fraction : Row sampling, 다양성을 높임
- Feature Fraction : Column Sampling, 다양성을 높임, 보통 정확도가 높아짐
- Scale Pos Weight : 불균형 데이터에서 weight를 줌

[최종 Parameter]

Parameter	Value
Learning Rate	0.01
Num Iterations	2000
Early Stopping Rounds	500
Max Depth	5
Bagging Fraction	0.8
Feature Fraction	0.7
Scale Pos Weight	1

RandomizedSearch, GridSearch를 이용하여 parameter 튜닝을 수행하였다.

이 때 cross-validation(K = 5)을 이용하였고, AUC, CSI(Threshold = Train 결로 비율)을 이용하여 최종 parameters를 선정하였다.

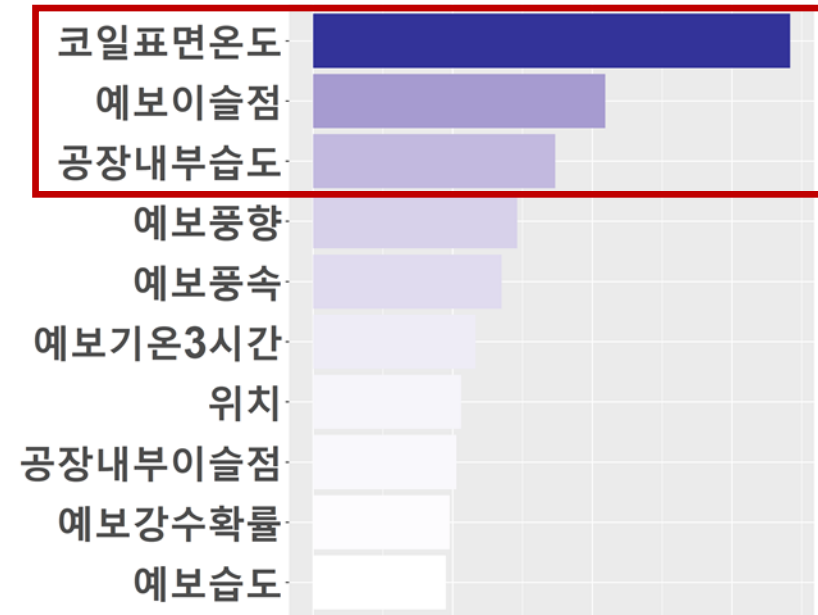
Weight를 주는 Scale Pos Weight Parameter의 값이 1인 점은 앞의 알고리즘 선택에서 Oversampling의 성능이 더 좋지 않았던 결과와 일치한다.

Threshold & Modeling 결과

최종 성능 : [cut-off] 0.0075(24시간), 0.008(48시간)

AUC	CSI
0.9869	66.36%

[변수 중요도 그래프]



최종적인 성능은 AUC 0.9869, CSI 66.36%로 나타났다.

앞서 사전 조사를 통해 결로는 이론적으로 표면온도가 이슬점보다 낮으면 발생한다는 것을 알 수 있다. 이때, 이슬점은 주변 온도와 습도로 계산 가능하다. 최종적으로 선택된 모델을 통해 얻은 **중요 변수** TOP3은 **코일표면온도, 예보 이슬점, 공장 내부 습도**이다. 이는 앞서 살펴본 결로의 이론적 배경과 일치하는 것을 확인할 수 있으며, 높은 AUC를 갖기 때문에 우리의 **최종 모델**이 결로 현상을 잘 설명하며 **공장 내 결로 예측에 적합한 모델**이라 생각된다.

최종 예측 모델이 결로현상을 잘 설명하고 있다.

A blue-tinted photograph of an industrial facility, likely a power plant or refinery, featuring several tall smokestacks and complex piping structures. The image is framed by a white border.

05 활용방안

결로 예방을 위한 예측 경보 시스템

- 실시간 기상 데이터로 예측된 결로 위험 수준에 대한 모니터링 및 위험 수준별 경보 시스템



예측 정보 시스템 운영방안

결로 위험 수준별 정보 시스템

- 정보시스템의 위험 수준은 다음과 같이 6개의 단계로 구성되어 있다.



- 24시간 후의 위험 수준에 맞추어 관리자에게 다음과 같은 정보 알람 메시지를 전달한다.

[경보 해제]



1단계에서 결로 미발생으로 넘어갈 때
관리자에게 알람 메시지를 전달한다.

[경보 1단계]



결로 1단계가 예측되면
관리자에게 알람 메시지를 전달한다.

[경보 2단계 이상]



결로 2단계 이상이 예측될 때 결로 방지
시스템이 미가동일 경우 자동 가동하고
관리자에게 알람 메시지를 전달한다.

예측 정보 시스템 운영방안

모니터링 시스템

- 모니터링 시스템은 10분 단위로 업데이트되며 다음과 같이 구성되어 있다.

[메인 화면]

2020-06-05 10:00

공장 1 공장 2

시간 별 예측 화면 이동 버튼

2020-06-06 10:00 (24시간)

발생 Location 1

미발생 Location 2

미발생 Location 3

2020-06-07 10:00 (48시간)

발생 Location 1

발생 Location 2

미발생 Location 3

경보 3단계 경보 2단계 경보 1단계 미발생 미발생 미발생

메인 화면 이동 버튼

[시간 별 경로 예측 화면]

2020-06-05 10:00

공장 1 공장 2

2020-06-05 10:00

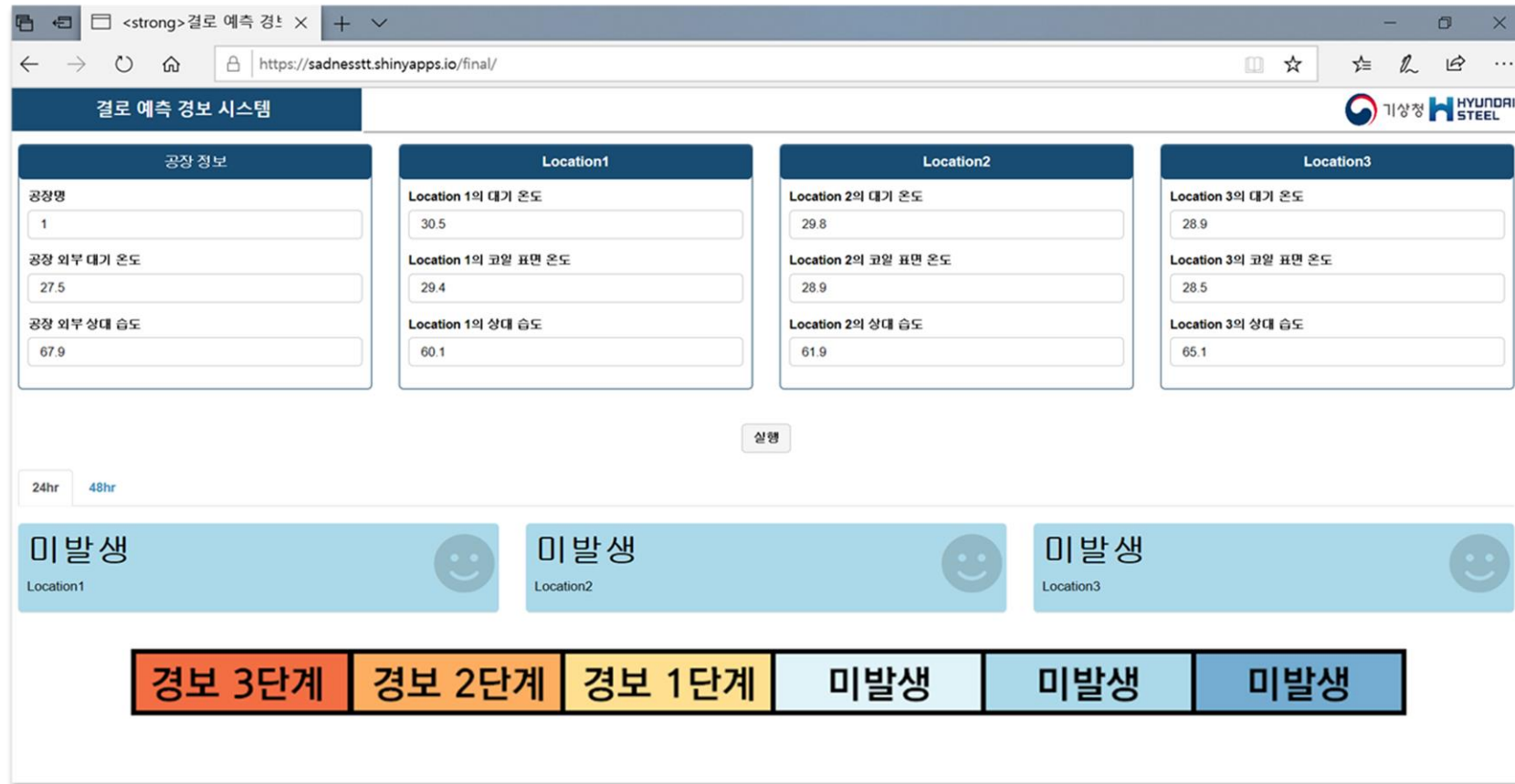
날짜	Location 1	Location 2	Location 3
2020-06-05 10:00	미발생	미발생	미발생
2020-06-05 10:10	미발생	미발생	미발생
2020-06-05 10:20	미발생	미발생	미발생
2020-06-05 10:30	경보 1단계	미발생	미발생
⋮	⋮	⋮	⋮
2020-06-07 10:00	경보 2단계	경보 1단계	미발생

24시간 후, 48시간 후 Location 별 경로 예측 상황을 확인할 수 있다.

현재 시간부터 48시간까지 10분 단위의 경로 예측 상황을 확인할 수 있다.

예측 경보 시스템 실제 예시

예측 경보 시스템 앱 주소



결로 예측 경보 시스템

기상청 HYUNDAI STEEL

공장 정보	Location1	Location2	Location3
공장명 1	Location 1의 대기 온도 30.5	Location 2의 대기 온도 29.8	Location 3의 대기 온도 28.9
공장 외부 대기 온도 27.5	Location 1의 코일 표면 온도 29.4	Location 2의 코일 표면 온도 28.9	Location 3의 코일 표면 온도 28.5
공장 외부 상대 습도 67.9	Location 1의 상대 습도 60.1	Location 2의 상대 습도 61.9	Location 3의 상대 습도 65.1

실행

24hr 48hr

미발생
Location1

미발생
Location2

미발생
Location3

경보 3단계 | 경보 2단계 | 경보 1단계 | 미발생 | 미발생 | 미발생

URL : <https://sadnesstt.shinyapps.io/final/>



데이터 관리의 용이성과 데이터를 활용한 공정 환경 구축이 가능



자동화로 인한 결로 관리의 효율성 극대화 및 관리 비용을 절감



실시간 결로 경보 알림으로 빠르게 대응하여 상품의 손상을 최소화



THANK YOU



팀원 참여도

구분	우나영	조주영	한효선
문제 이해 및 자료 조사	33.4%	33.3%	33.3%
데이터 전처리	33.3%	33.4%	33.3%
데이터 모델링	33.3%	33.3%	33.4%
분석결과 정리 및 보고서 작성	33.3%	33.4%	33.3%
활용방안 아이디어 제시	33.3%	33.3%	33.4%

A photograph of an industrial facility, possibly a refinery or chemical plant, featuring several tall smokestacks and complex piping structures. The image is overlaid with a dark blue gradient and a white border.

APPENDIX

변수설명 : 기존 데이터

[Plant1, Plant2 train data]

Variable

일자

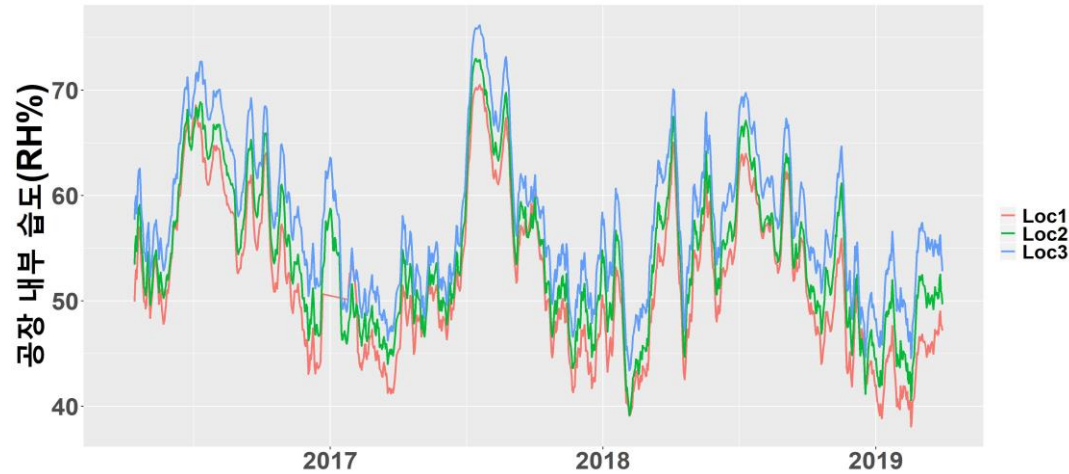
공장내부습도

공장외부기온

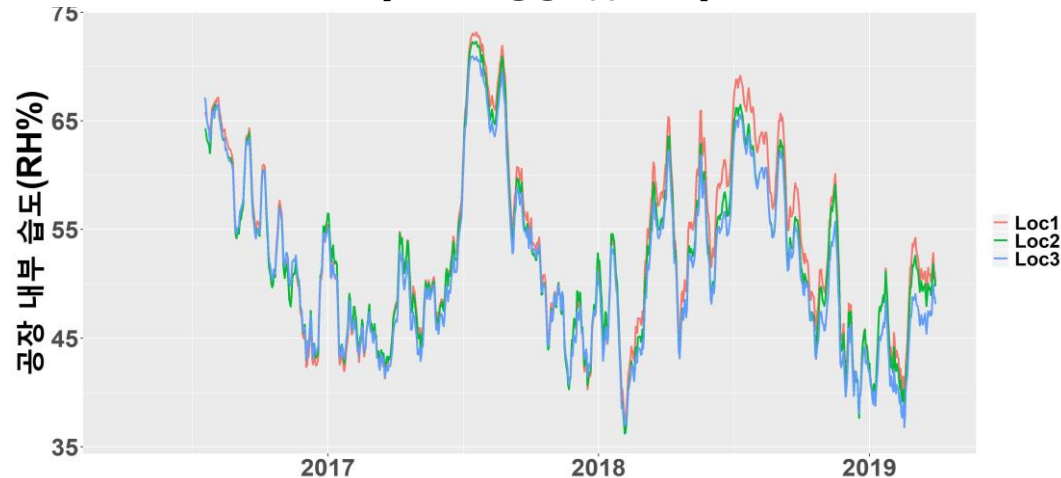
코일표면온도

결로여부

[Plant1 공장 내부 습도]



[Plant2 공장 내부 습도]



기상 변수의 경향성을 뚜렷하게 파악하기

위해 **Moving average**(MA, 14일 이동평균)를 사용하여 그래프를 그렸다.

Plant1에서 측정된 공장 내부 습도는

각 location별 **일정한 차이**를 가진다.

반면에, Plant2에서 측정된 공장 내부 습도는

각 location별 서로 비슷하게 나타난다.

변수설명 : 기존 데이터

[Plant1, Plant2 train data]

Variable

일자

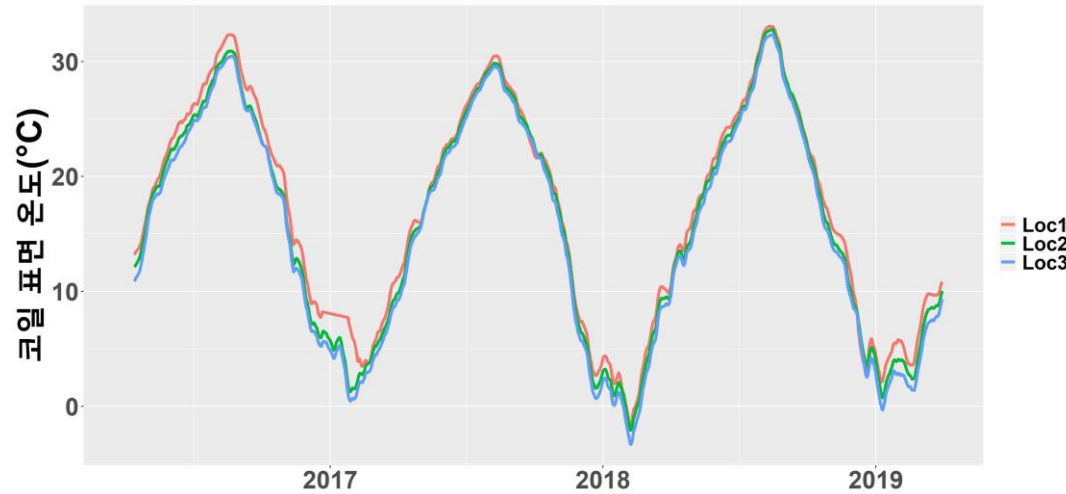
공장내부습도

공장외부기온

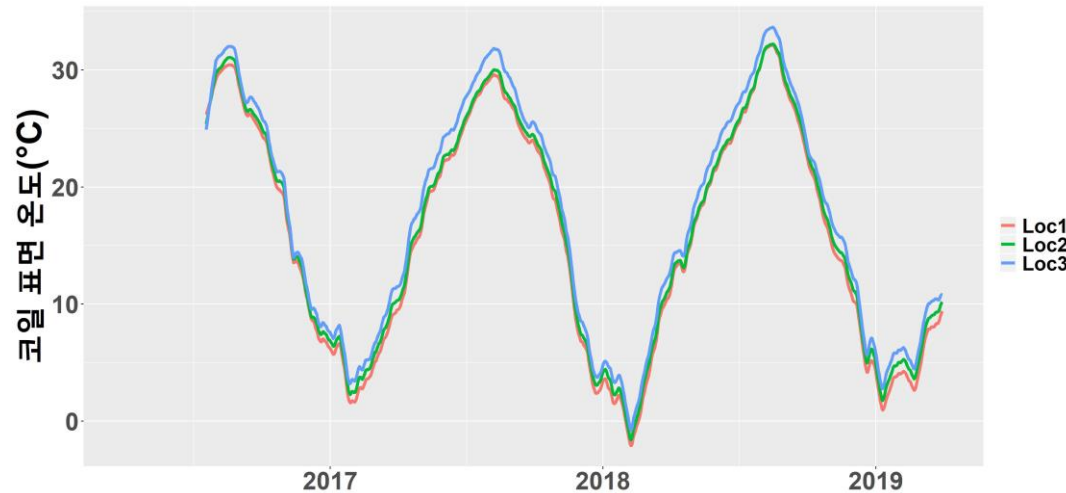
코일표면온도

결로여부

[Plant1 코일 표면 온도]



[Plant2 코일 표면 온도]



Plant, location별 측정된 코일 표면 온도는 서로 비슷한 양상을 보이며

여름철에는 약 30°C, 겨울철에는 약 0°C 까지 변화하는 것을 확인할 수 있다.

Plant1의 경우 location1의 코일 표면온도가 대체로 가장 높고, plant2의 경우는 location3의 코일 표면 온도가 가장 높다.

변수설명 : 파생변수

[파생변수]

Variable

월

일

계절

공장

공장내부위치

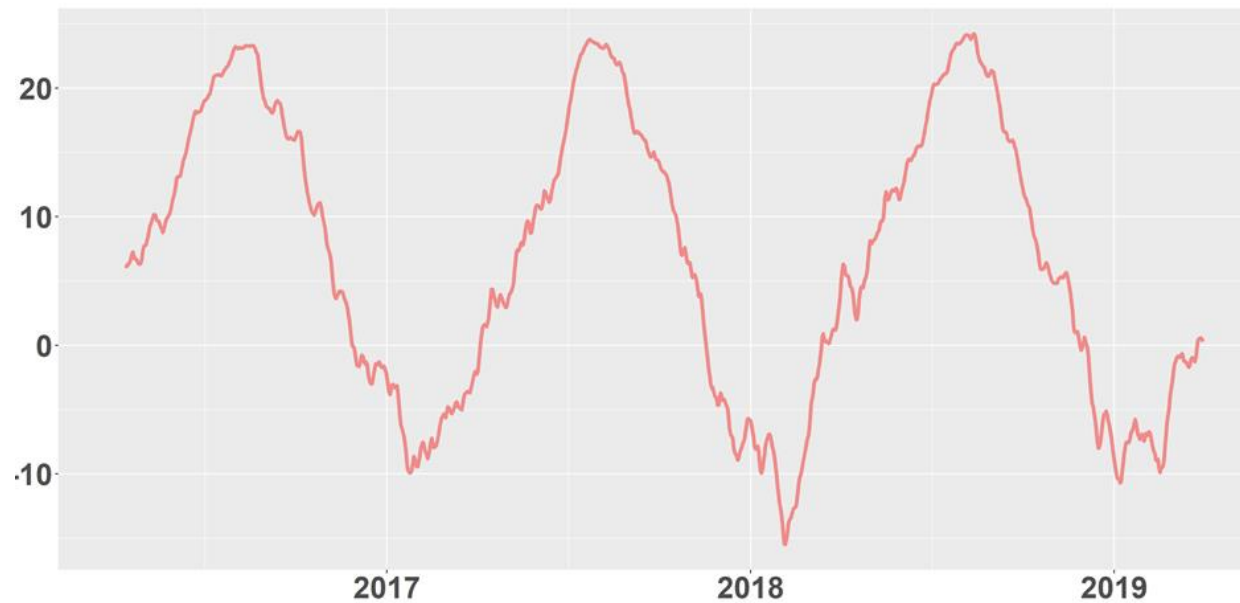
예측시점

공장내부 이슬점

서산 이슬점

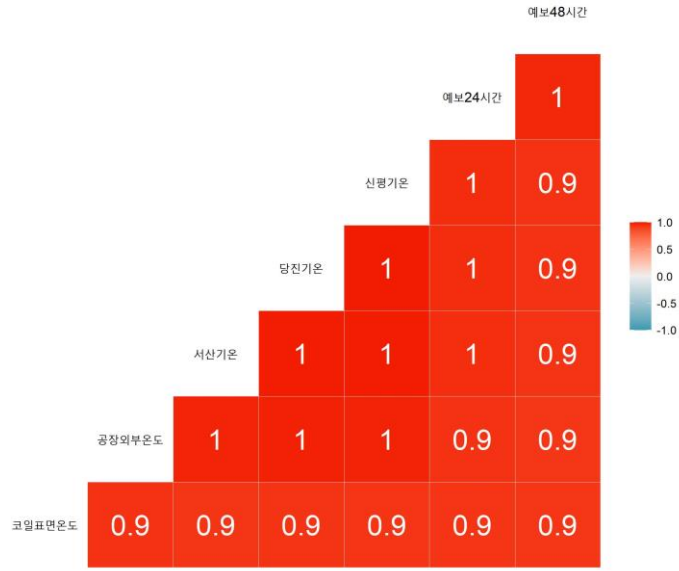
예보 이슬점

[공장내부, 서산, 예보 이슬점 추이]



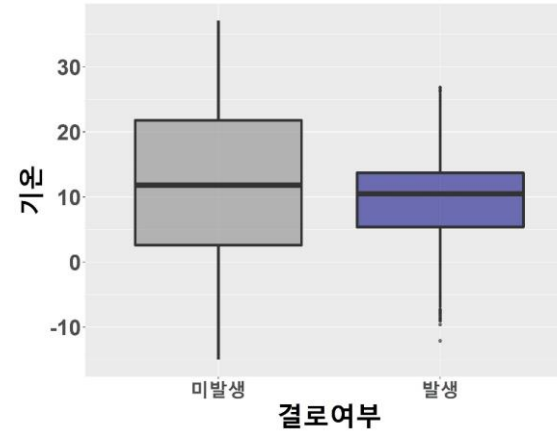
앞서 설명한 이슬점 공식을 이용하여 이슬점 파생변수를 생성하였다. 공장내부, 서산, 예보 이슬점은 서로 비슷한 양상을 보이며 여름철에는 높고 겨울철에는 낮은 것을 확인할 수 있다.

EDA : Temperature

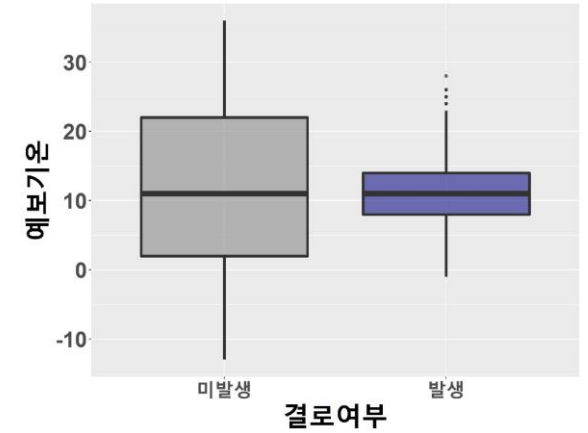


공장 외부, 코일표면, 서산, 당진, 신평, 예보 온도 간의 상관관계는 모두 0.95 이상이다. 서산 기온과 24, 48시간 결로 여부와의 관계를 나타내는 box-plot은 다음과 같다.
현재 관측기온이 낮을 수록 48시간 후의 결로 발생이 높은 것을 확인할 수 있다.

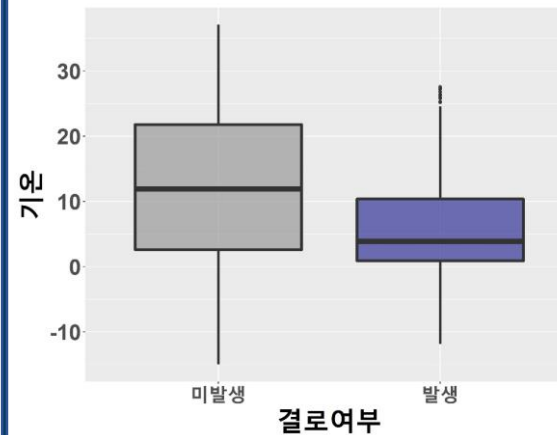
[결로 여부 별 서산 기온 분포 (24시간)]



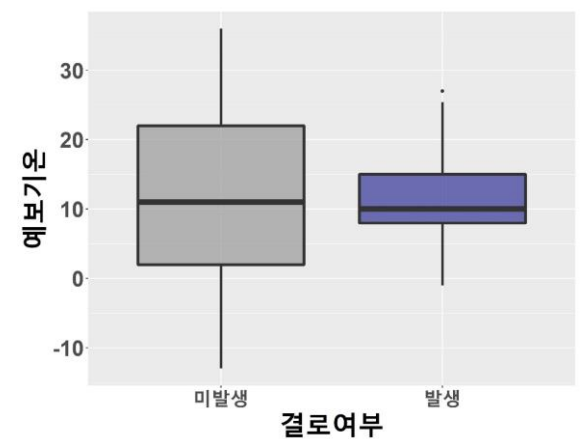
[결로 여부 별 예보 기온 분포 (24시간)]



[결로 여부 별 서산 기온 분포 (48시간)]



[결로 여부 별 예보 기온 분포 (48시간)]

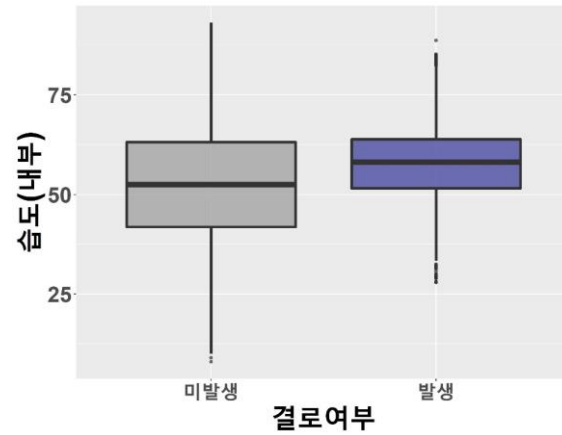


현재 관측 기온이 낮을수록 48시간 후 결로 발생 높다

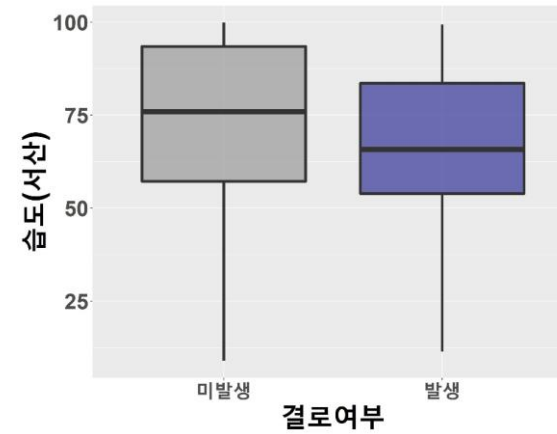
EDA : Humidity

- 공장 내부, 서산, 당진, 예보 습도와 24, 48시간 결로 여부와의 관계를 나타내는 box-plot은 다음과 같다.

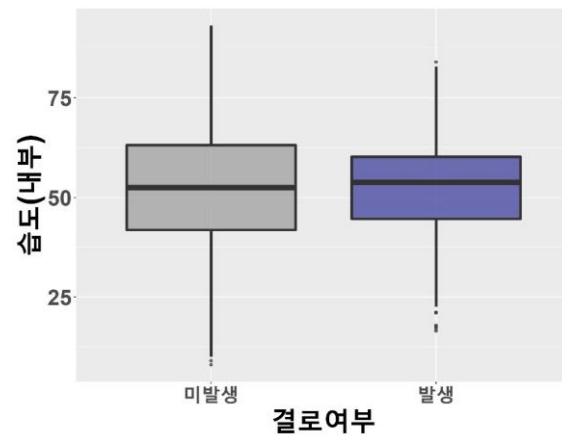
[결로 여부 별 내부 습도 분포 (24시간)]



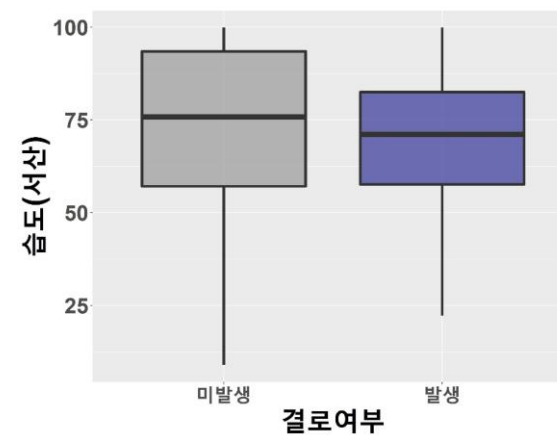
[결로 여부 별 서산 습도 분포 (24시간)]



[결로 여부 별 내부 습도 분포 (48시간)]

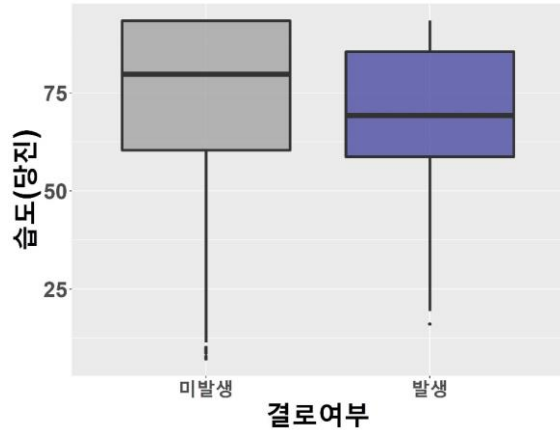


[결로 여부 별 서산 습도 분포 (48시간)]

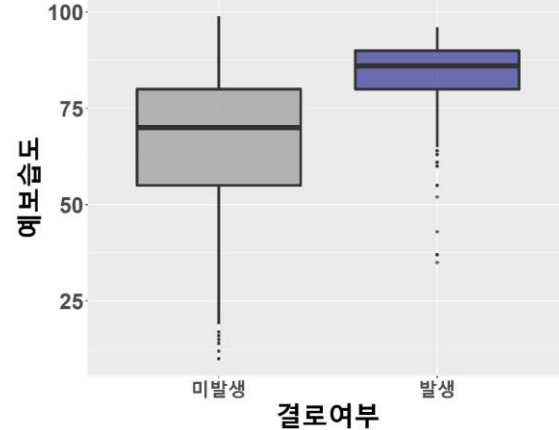


EDA : Humidity

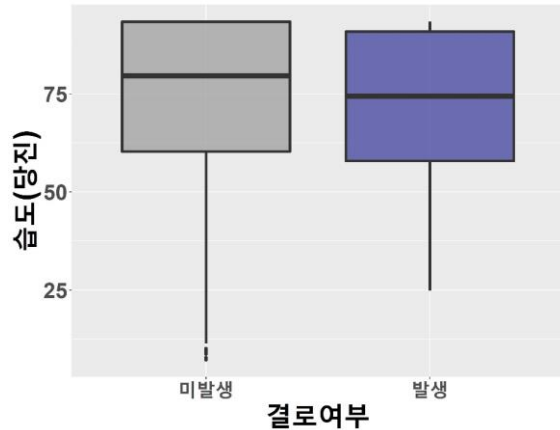
[결로 여부 별 당진 습도 분포 (24시간)]



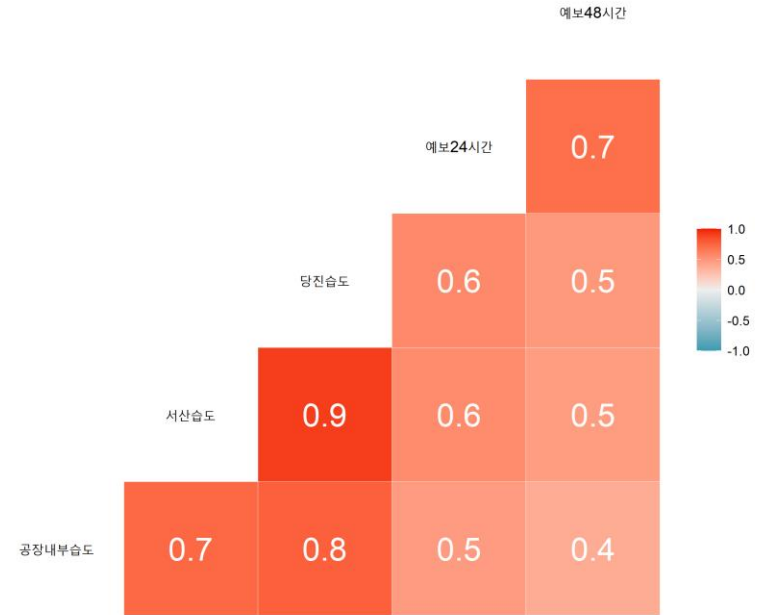
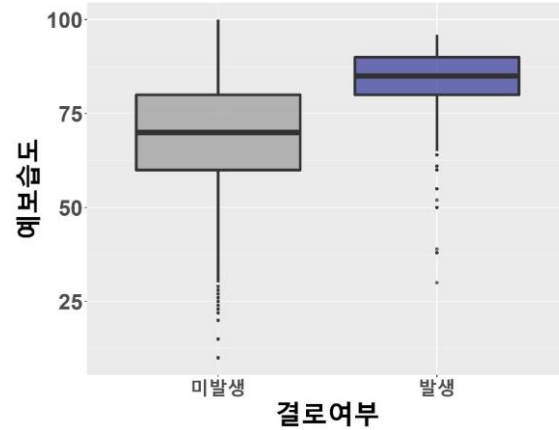
[결로 여부 별 예보 습도 분포 (24시간)]



[결로 여부 별 당진 습도 분포 (48시간)]



[결로 여부 별 예보 습도 분포 (48시간)]



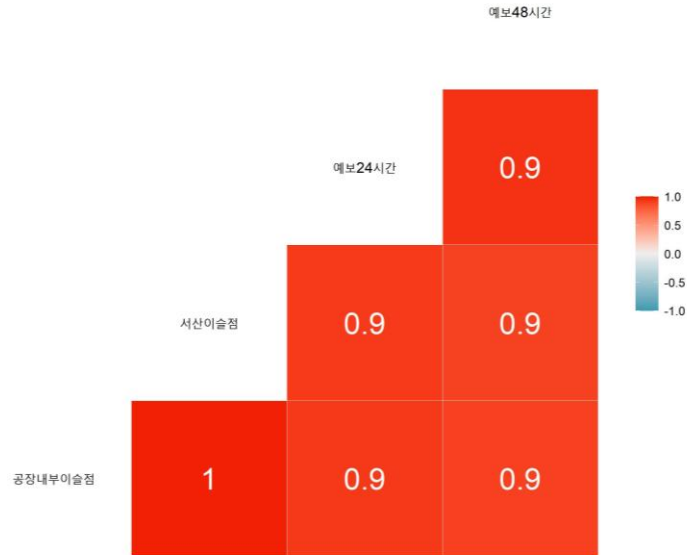
공장 내부, 서산, 당진, 예보 습도 간의 상관관계는 높지 않다.

습도 그래프를 살펴보면 현재 습도가 높을수록 미래의 결로 여부는 낮은 반면, 예보습도가 높을수록 결로 여부 높아진다.

이는 현재 습도가 높을 시 결로 대비로 인해 다음날의 실제 결로 발생 비율이 낮아지기 때문이라 추측된다.

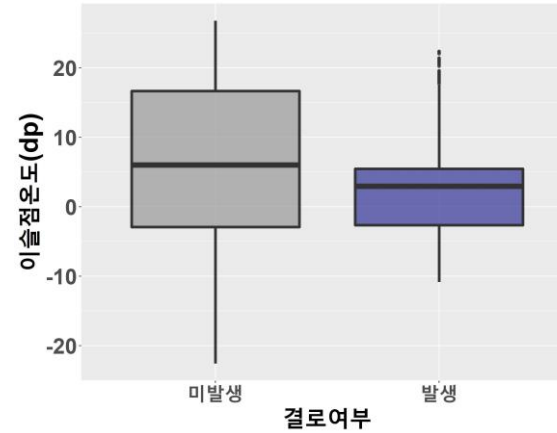
예보 습도가 높을수록 결로 발생률이 높아진다.

EDA : Dp

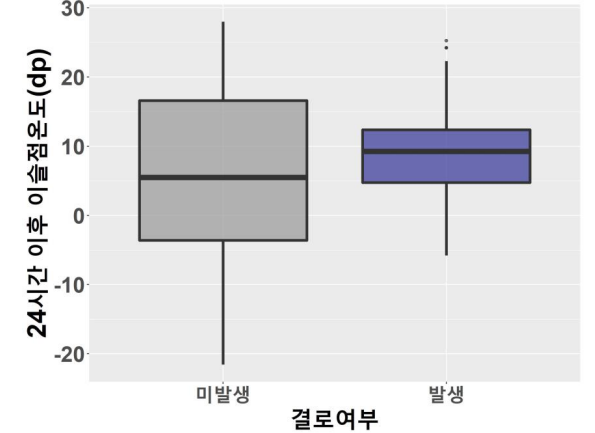


공장 내부, 서산, 예보 이슬점 간의 상관관계는 모두 0.95 이상이다. 서산 이슬점과 24, 48시간 결로 여부와의 관계를 나타내는 box-plot은 다음과 같다.
예보 이슬점 온도가 높을수록 결로 발생률이 높아진다.

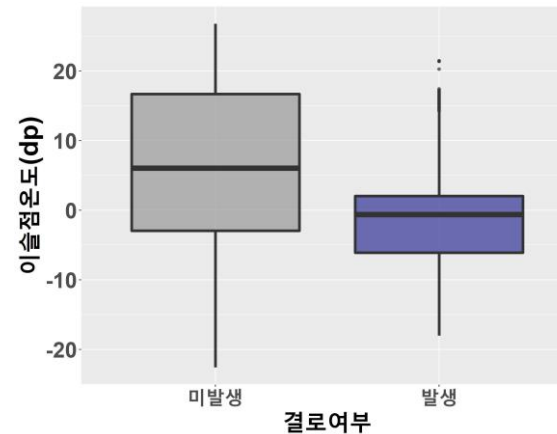
[결로 여부 별 서산 이슬점온도 분포 (24시간)]



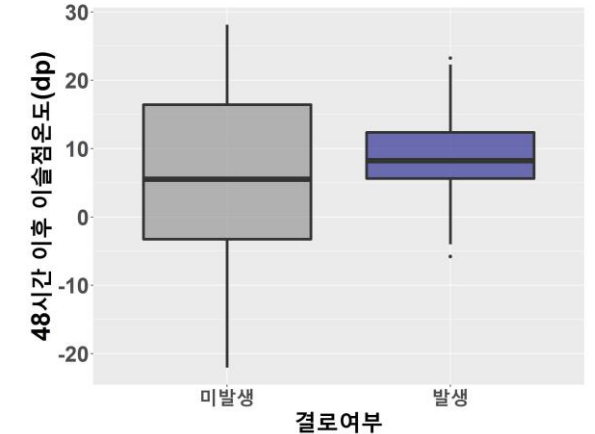
[결로 여부 별 예보 이슬점온도 분포 (24시간)]



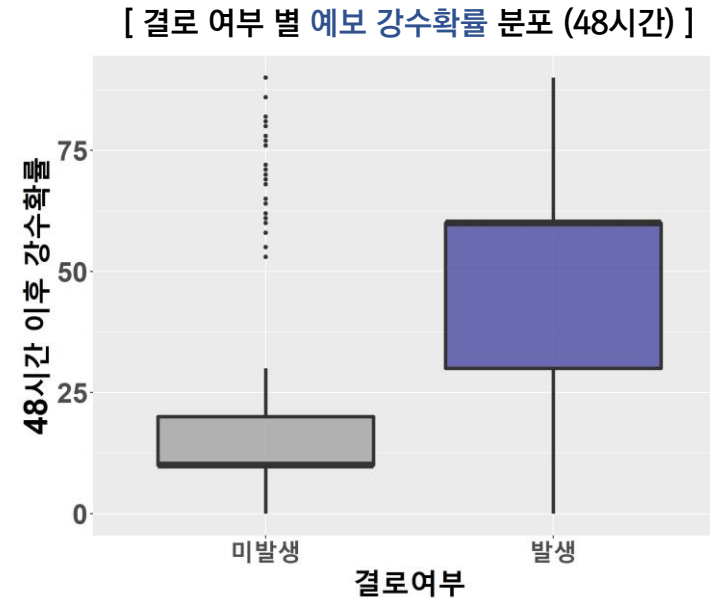
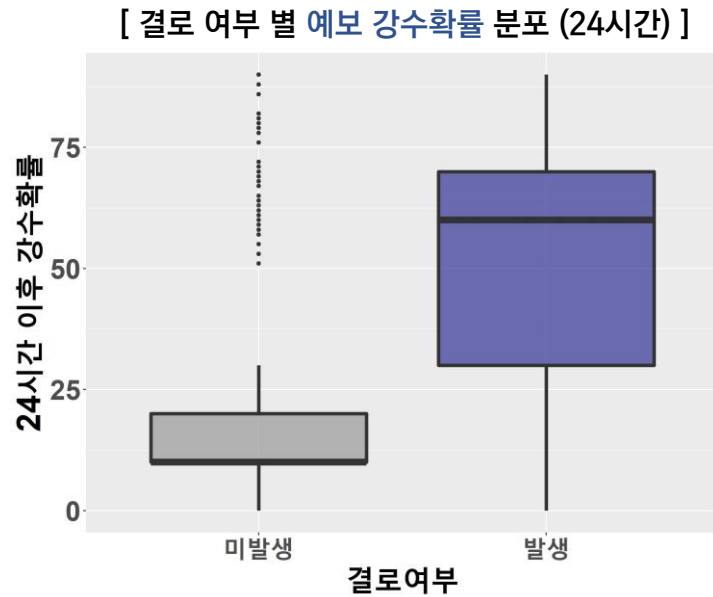
[결로 여부 별 서산 이슬점온도 분포 (48시간)]



[결로 여부 별 예보 이슬점온도 분포 (48시간)]



예보 이슬점 온도는 높을수록 결로 발생률이 높아진다.



강수확률이 높을수록 결로가 많이 발생하는 것을 확인할 수 있다.
이는 즉, 결로가 습도의 영향을 많이 받는다는 것을 알 수 있다.

예보 강수확률이 높을수록 결로가 많이 발생한다.

[결로 여부 별 **신평지역 1분강수량** 빈도표]

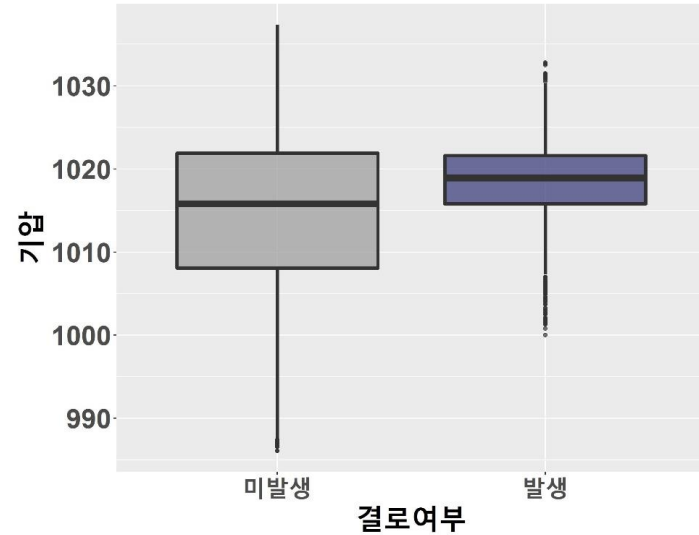
		1분 강수량 (신평)				
		0	0.5	1	1.5	2
24시간 뒤 결로여부	0	340224	1221	72	12	6
	1	1979	0	0	0	0

[결로 여부 별 **신평지역 1분강수량** 빈도표]

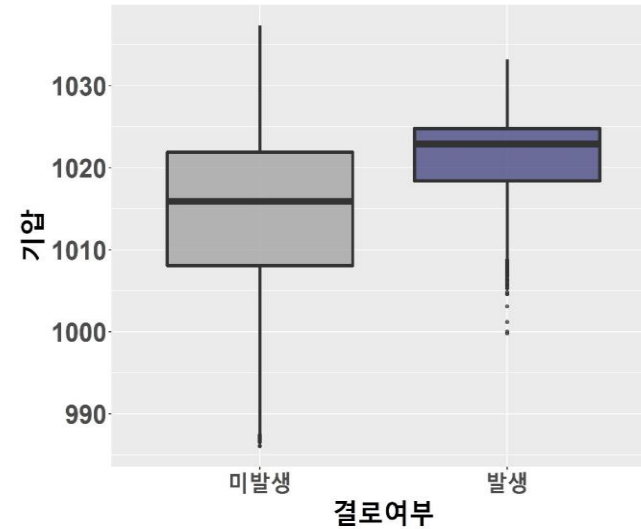
		1분 강수량 (신평)				
		0	0.5	1	1.5	2
48시간 뒤 결로여부	0	339122	1212	72	12	6
	1	1967	2	0	0	0

신평 지역 현재 시점에서 비가 많이 오는 경우 24시간, 48시간 뒤의 결로 현상은 나타나지 않았다.
이는 즉, 비가 오게 되는 경우 습도와 관련되어 결로현상이 나타날 수 있기 때문에 **미리 결로 방지 action을 취하는** 것이라 추측된다.

[결로 여부 별 신평 현지기압 분포 (24시간)]



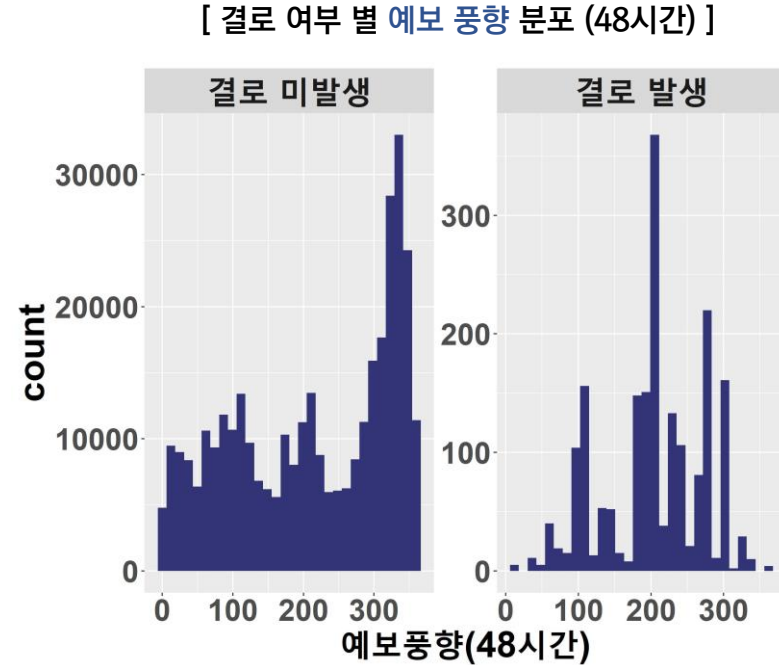
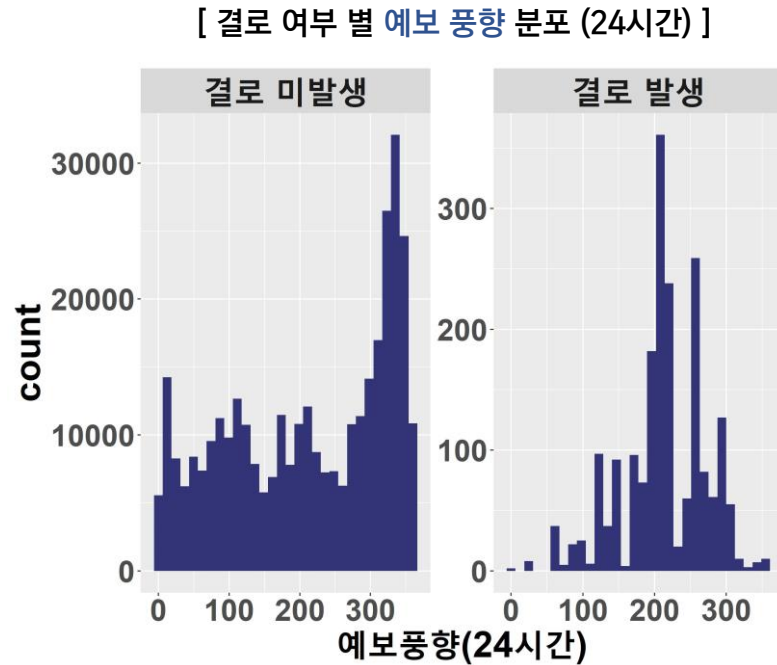
[결로 여부 별 신평 현지기압 분포 (48시간)]



24, 48시간 결로에 따른 boxplot을 그려본 결과, 현지기압이 높을수록 결로가 발생하는 것으로 보인다.
대체로 현지기압은 겨울철에 높는데 결로 또한 겨울철에 많이 발생하는 것을 확인할 수 있다.

현지기압이 높을수록 결로가 발생한다.

EDA : Wind Direction

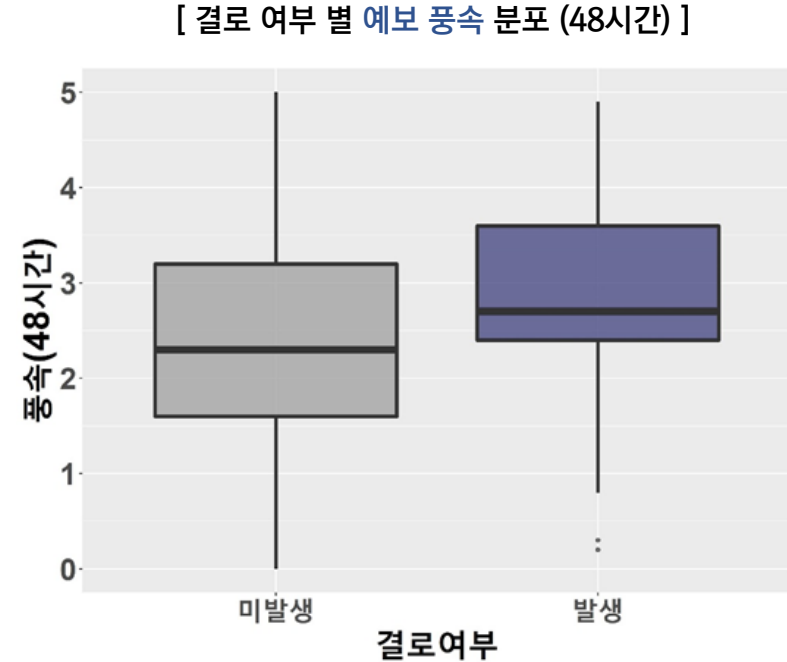
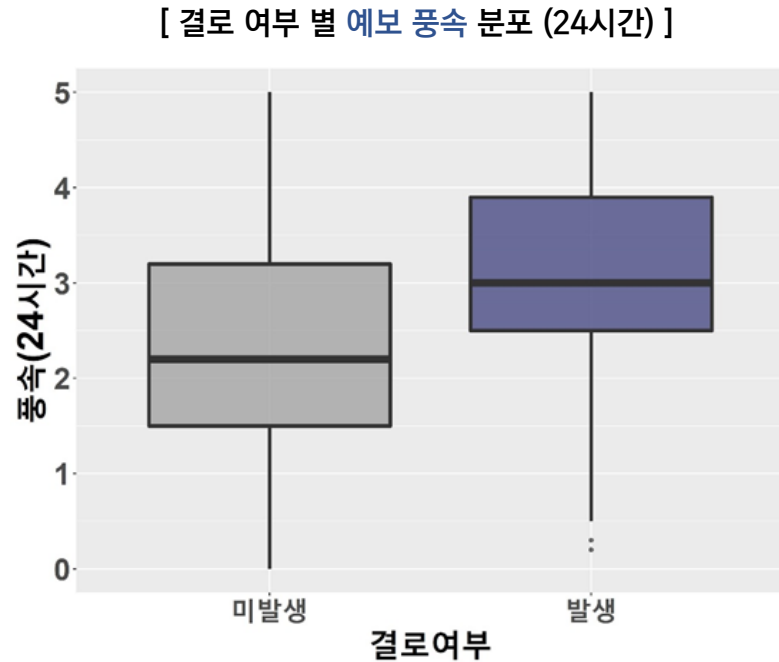


결로 여부 별 예보 풍향 분포를 그래프로 나타내었다.

결로가 발생한 경우와 미발생한 경우의 예보 풍향 분포가 다르기 때문에
결로여부와 예보 풍향은 상관성이 있다고 할 수 있다.

예보 풍향은 결로에 영향이 있다.

EDA : Wind Speed



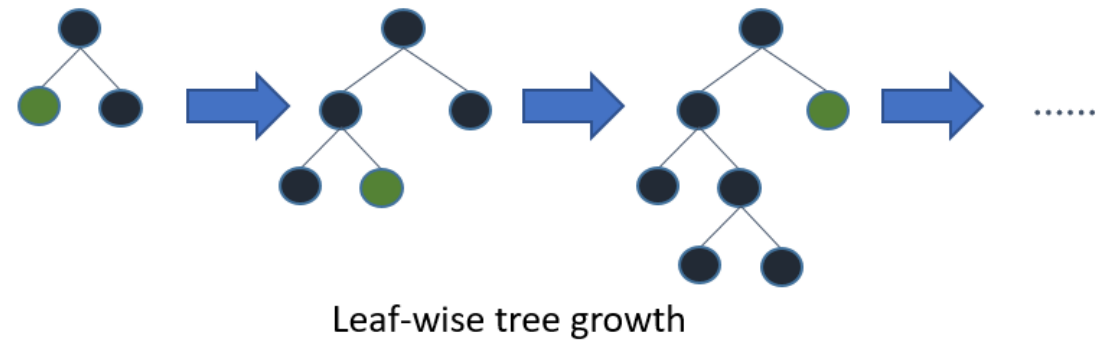
결로 여부 별 예보 풍속 분포를 그래프로 나타내었다. 예보 풍속이 높을수록 결로가 발생한 경우가 많기 때문에 결로 여부와 예보 풍속은 상관성이 있다고 할 수 있다. 우리나라의 경우 주로 겨울철에 풍속이 높고, 대부분의 결로 현상은 겨울철에 발생하기 때문에 위와 같은 결과가 나타났을 것이라고 추측된다.

예보 풍속은 결로에 영향이 있다.

LightGBM Algorithm

LightGBM

- 히스토그램 기반 알고리즘을 사용한다.
- Discrete bins로 연속 값을 대체하여 메모리 사용량을 줄인다.
- Leaf-Wise 방식을 사용하여 훨씬 복잡한 트리를 생성한다.
- XGBOOST에 비해 training 시간을 현저히 단축시킨다.
- 병렬 학습을 지원한다.



예측 경보 시스템 실제 예시

시스템 작동 프로세스

<https://youtu.be/MJic-VRxWaY>

공장 정보	Location1	Location2	Location3
공장명 1	Location 1의 대기 온도 30.5	Location 2의 대기 온도 29.8	Location 3의 대기 온도 28.9
공장 외부 대기 온도 27.5	Location 1의 코일 표면 온도 29.4	Location 2의 코일 표면 온도 28.9	Location 3의 코일 표면 온도 28.5
공장 외부 상대 습도 67.9	Location 1의 상대 습도 60.1	Location 2의 상대 습도 61.9	Location 3의 상대 습도 65.1

실행

공장에서의 관측 데이터를 입력하면 해당 시간 기상청의 기상 관측 데이터 및 동네 예보 데이터 크롤링 하여 예측 모형을 실행한다.

그 결과를 바탕으로 결로 예측 경보 시스템을 작동한다.

Reference

- I. 정일영. 돔형 장스판 공장지붕 시스템의 결로 방지에 관한 연구, 2012.
- II. 김태명, 지석근, 김영완. 이슬 결로점 기반 수배전반 결로 방지 장치 제작, 2018.
- III. 기상청, 2018년 기상기후 빅데이터 융합서비스