제3회 한국가스공사 (KOGAS)

빅데이터·인공지능 스타트업 경진대회

가스공급량 수요예측 모델 개발

가스 공급량 수요예측 최적화와 활용방안

데린이 팀

한국가스공사

1. 배경 및 개요

도시가스는 공공재 → 국민의 편익 위해 정확한 수요예측이 필요

배경

- 수요 예측은 원료 수급, 제품 생산, 수요처 공급, 재고관리 등 산업의 전 Value-Chain에 큰 영향을 미침
- 급변하는 가스 수요 환경에 대응하지 못할 경우 정부, 기업, 국민에게 사회적 비용 손실이 발생함

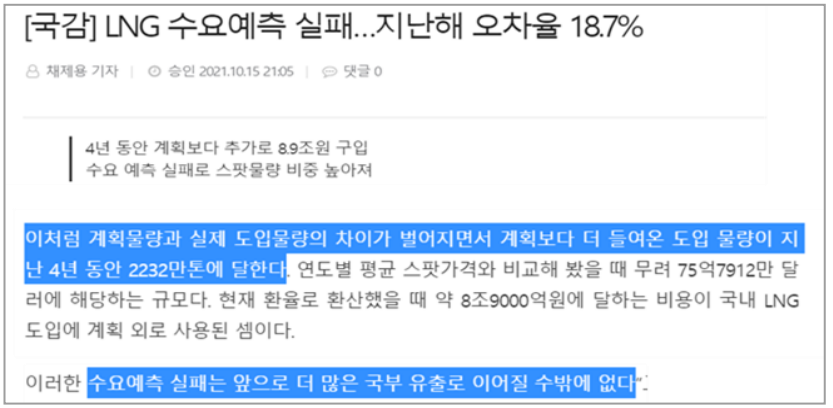
급감하는 산업용 도시가스, 내년 전망도 '먹구름'

A 송승은 기자 | 입력 2020.12.10 18:04 | 수정 2020.12.10 18:10 | 댓글 0

A사 30%·B사 21% 감소, 우회직수입 확대에 우려 증폭
LPG 연료전환 지속, 원료비연동제 효과도 아직은 미미

정부는 지난 8월부터 주택용과 일반용을 제외한 도시가스 전 용도(산업용, 열병합용)의 원료비를 기존 홈수월 조정에서 매월 자동조정키로 결정한 바 있다. 이에 따라 산업용 도시가스 수요 회복에 긍정적 효과를 줄 수 있을 것으로 전망되기도 했으나 기대 만큼의 효과는 나타나지 않고 있는 것이 현실이다.

가스공사 노조 관계자는 "직수입으로 인한 산업용 물량의 이탈 현상은 동절기 위주의 도입계약 체결이 불가피 하게 만들면서 결국 연료선택권이 없는 국민들이 요금상승의 피해를 입게된다"고 주장했다.



[국감] LNG 수요예측 실패...지난해 오차율 18.7%

채제용 기자 | 승인 2021.10.15 21:05 | 댓글 0

4년 동안 계획보다 추가로 8.9조원 구입
수요 예측 실패로 스팟물량 비중 높아져

이처럼 계획물량과 실제 도입물량의 차이가 벌어지면서 계획보다 더 들여온 도입 물량이 지난 4년 동안 2232만톤에 달한다. 연도별 평균 스팟가격와 비교해 봤을 때 무려 75억7912만 달러에 해당하는 규모다. 현재 환율로 환산했을 때 약 8조9000억원에 달하는 비용이 국내 LNG 도입에 계획 외로 사용된 셈이다.

이러한 수요예측 실패는 앞으로 더 많은 국부 유출로 이어질 수밖에 없다"

과제 개요

- 6년간 공급사/시간별 공급량 데이터를 통해 향후 90일간의 공급량 수요를 예측

	데이터	예측
7개 공급사	2013.1.1 ~ 2018.12.31	2019.1.1 ~ 3. 31

2. 사용 데이터

제공 데이터 외 다양한 외부데이터의 활용 가능성 검토

제공 데이터

	연월일	시간	구분	공급량
0	2013-01-01	1	A	2497.129
1	2013-01-01	2	A	2363.265
2	2013-01-01	3	A	2258.505
3	2013-01-01	4	A	2243.969
4	2013-01-01	5	A	2344.105
...	...	...	...	...
368083	2018-12-31	20	H	681.033
368084	2018-12-31	21	H	669.961
368085	2018-12-31	22	H	657.941
368086	2018-12-31	23	H	610.953
368087	2018-12-31	24	H	560.896

- **연월일**
  공급량 측정일
- **시간**
  공급량 측정시간
- **구분**
  1개 권역내 7개 공급사
- **공급량**
  비식별화 된 공급량 측정값

외부 데이터

- **도시가스 월별 상대가격 지수 데이터** (출처 : 한국가스공사)
  - 내용: 민수용, 산업용 가스 상대가격
  - 적용: 변수 추가, 예측기간의 데이터는 18년 12월 데이터로 대체
  - 결과: 테스트 결과 적용 효과가 약함
- **기상 정보 데이터** (출처 : 기상자료개방포털)
  - 내용: 온도, 강수, 풍속 등 기상정보
  - 적용: 월평균 데이터 생성 후 3개월 Lagging하여 적용
  - 결과: 예측기간의 실제 데이터와 차이 발생 → 공급량 예측력 저하
- **지역별 특수일 효과** (출처 : 한국가스공사)
  - 내용: 지역별 평일 대비 가스 수요 감소 비율
  - 적용: 특수일에 해당 변수 적용
  - 결과: 공급량 예측 대상 권역을 알 수 없어 적용 어려움
- **공휴일 데이터** (출처 : 한국천문연구원 특일정보 API)
  - 내용: 명절, 국경일 등 공휴일 정보
  - 적용: 공휴일 변수 추가
  - 결과: 공휴일의 공급량 예측값 오차 증가

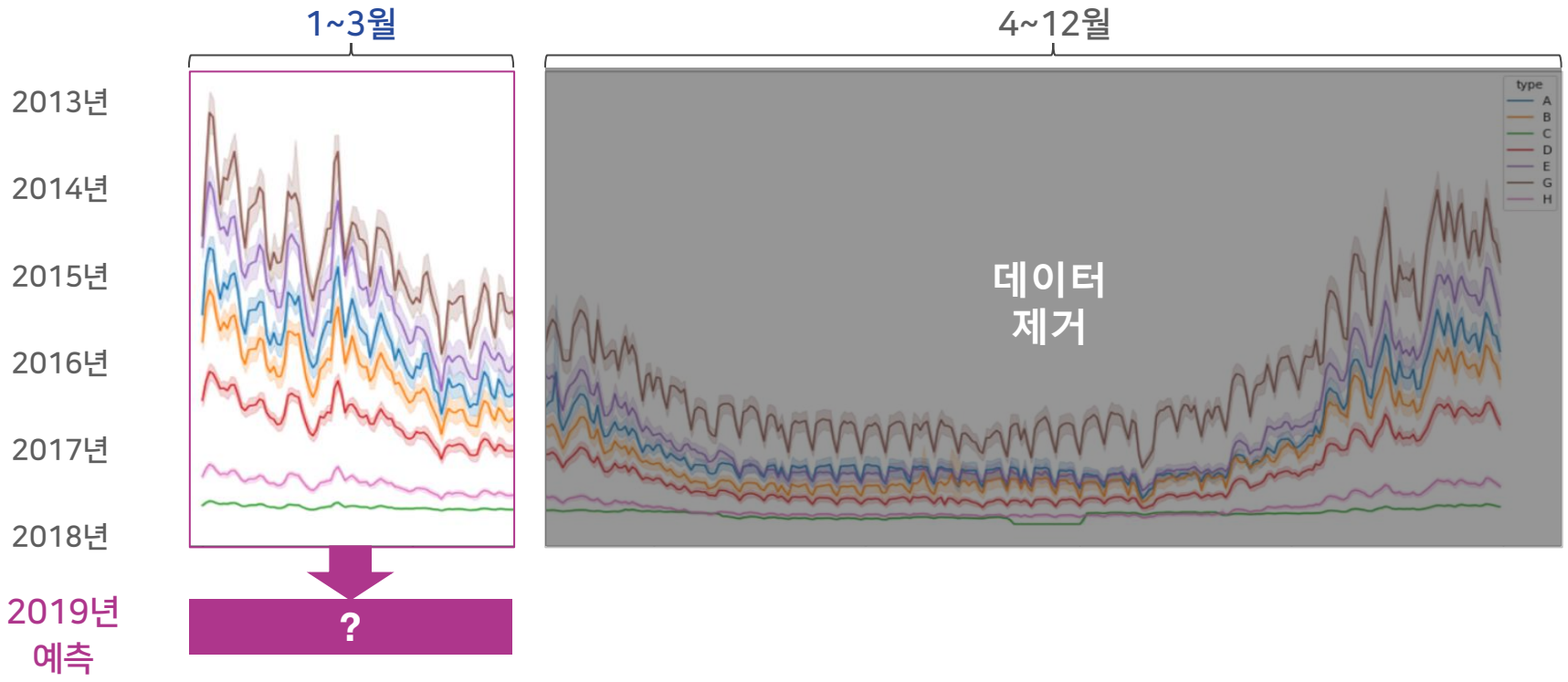
데이터 사용 기간 제한(3개월 전), **낮은 효과**로 외부 데이터 활용 X

3. 사용 모델기술 – 1) 데이터 전처리

불필요한 학습 데이터를 제거하여 모델 학습의 효율을 개선

데이터 사용범위 설정

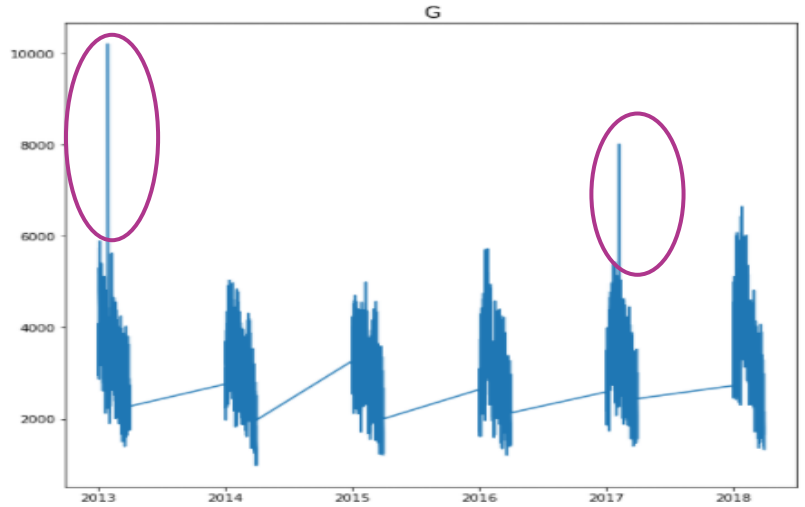
- Noise제거 및 모델 학습의 효율 및 예측력을 향상시키기 위해 평가 데이터와 같은 기간(1~3월)의 데이터 사용

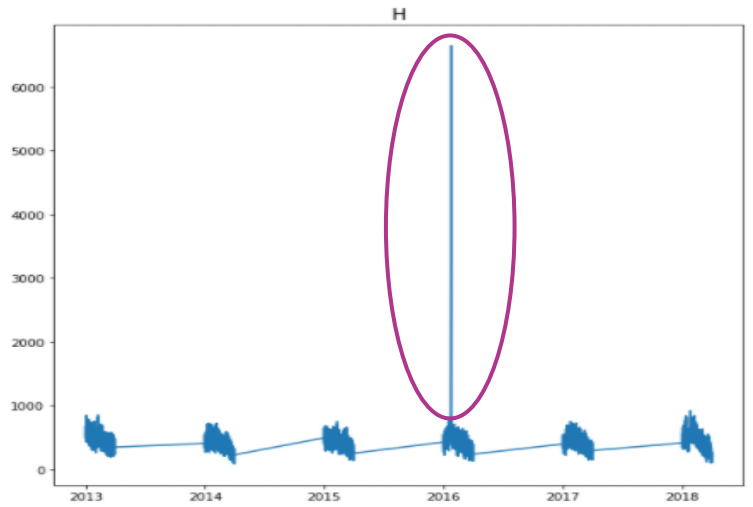


이상치 처리의 최소화로 Domain 정보의 손실을 예방

EDA를 통한 이상치 처리

- IQR 등 일반적인 이상치 처리 방법 적용시 Domain 정보가 과도하게 손실될 우려가 있음
- EDA 결과를 바탕으로 학습에 방해가 되는 구간에 한하여 최소한의 이상치 처리 실시






```
# 2016년 H 공급사 이상치는 +1, -1 일 동시간의 공급량 평균으로 대체
for row in train[(train['type'] == 6) & (train['year'] == 2016) & (train['month'] == 1) & (train['day'] == 24) & (train['hour'] < 4)&(train['amount'] > 800)].index:
  train.loc[row, 'amount'] = (train.loc[(row-24), 'amount'] + train.loc[(row+24), 'amount'])/2

# 공급량 6644이상인 것은 이상치로 판단하고 +1, -1 시간의 공급량 평균으로 대체
for row in train[train['amount'] > 6644].index:
  train.loc[row, 'amount'] = (train.loc[(row-1), 'amount'] + train.loc[(row+1), 'amount'])/2
```

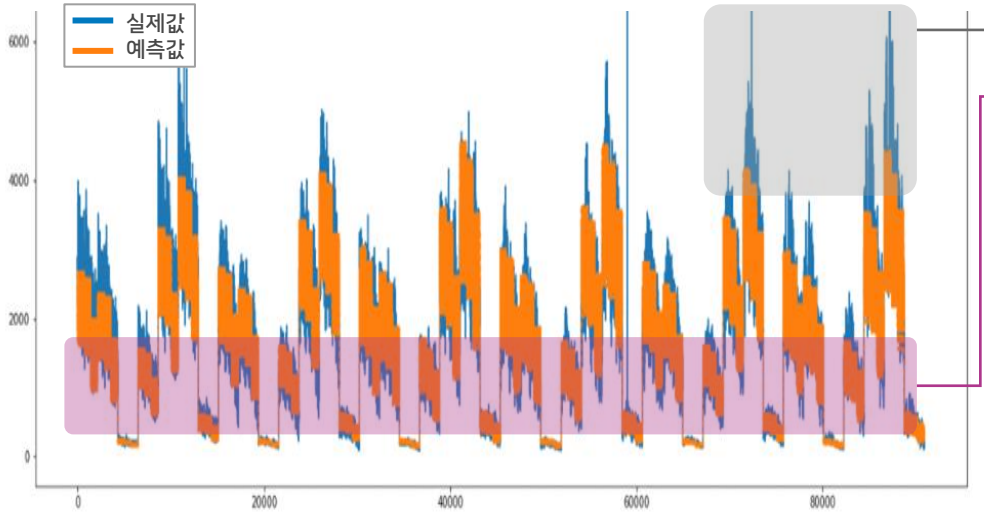
※ 전후 1시간 또는 전후 1일 같은 시간의 공급량 평균으로 대체

3. 사용 모델기술 – 1) 데이터 전처리

예측이 부정확한 구간을 변수화하여 모델 예측력을 강화

변수 추가

- fold별 예측값과 실제 공급량의 차이를 비교 ☞ 특정구간에서 예측력이 떨어짐을 확인
- 평가지표(NMAE)특성상 공급량이 적은 구간에서의 예측력이 중요할 것임
- 해당 구간을 변수로 추가하여 부정확한 예측을 보완



case1) 실제값 6000, 예측값 4000, NMAE = 0.3

case2) 실제값 500, 예측값 1500, NMAE = 2.0

	year	month	day	hour	type	amount	val_pred	nmae
23738	2014	3	31	3	3	220.036	613.538700	1.788356
30218	2014	3	31	3	6	85.804	236.073044	1.751306
30219	2014	3	31	4				1.608272
19418	2014	3	31	3				1.511834
23737	2014	3	31	2	3	267.956	645.606722	1.409376
17258	2014	3	31	3	0	420.265	1008.993792	1.400851
19417	2014	3	31	2	1	387.493	903.970612	1.332870
25898	2014	3	31	3	4	549.389	1279.768349	1.329439

3월말, 2~7시에 NMAE가 최대

```
# 예측력이 떨어지는 기간 피쳐화 (3월말, 2~4시)
for df in [train, test]:
    df['little_gas'] = 0
    df.loc[(df['month'] == 3) & (df['day'] >=29) & (df['hour'] >=2) & (df['hour'] <=4), 'little_gas'] = 1
    df['26~31'] = df['day'].apply(lambda x : 1 if x >= 26 else 0)
    df['2~7'] = df['hour'].apply(lambda x : 1 if x >= 2 and x <= 7 else 0)
```

특정 구간관련 변수 추가

3. 사용 모델기술 – 2) 모델 선정

유사 과제에 사용된 모델을 분석/평가하여 최종 모델 선정

모델 후보 검토

· 관련 논문, 유사 경진대회에 사용된 모델 7종에 대해 분석 및 성능 평가 실시

구분	ARIMA	FBProphet	LSTM
종류	시계열	시계열	딥러닝
장점	추세와 계절성을 반영한 시계열 예측의 대표 모델	추세 변화 인지, 이상치 처리 탁월	대규모 시계열자료에서 성능이 우수
단점	대규모, 다차원 데이터셋에서 좋은 성능을 발휘하지 못함		예측값이 직전값에 의존
평가결과	Bad	Not Bad	Bad

3. 사용 모델기술 – 2) 모델 선정

유사 과제에 사용된 모델을 분석/평가하여 최종 모델 선정

모델 후보 검토

· 관련 논문, 유사 경진대회에 사용된 모델 7종에 대해 분석 및 성능 평가 실시

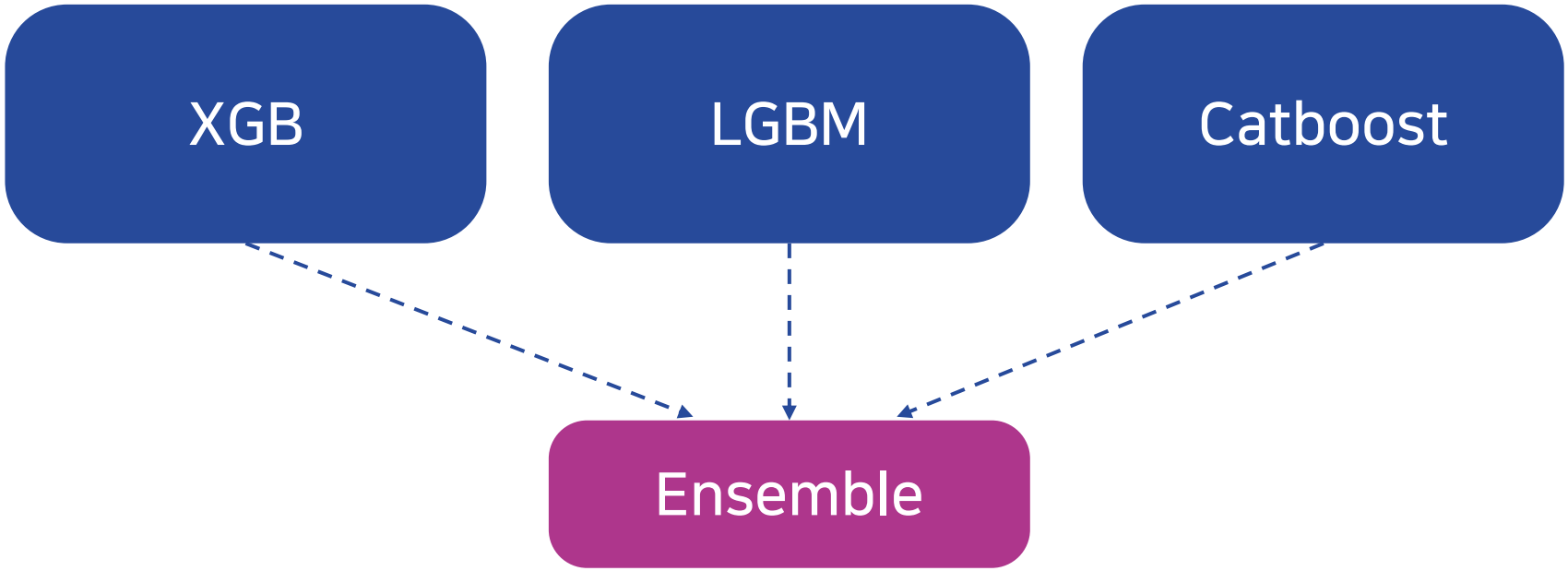
구분	RandomForest	XGBOOST	LGBM	CATBOOST
종류	트리-배깅	트리-부스팅		
장점	다수 DT 사용으로Variance 가 낮아짐	트리계열 중 시계열 예측이 우수	빠른 연산속도	범주형 변수 예측 성능이 우수
단점	느린 연산속도	트리모델 특성상 추세를 반영하기 어려움		
평가결과	Not Bad	Good	Good	Good

3. 사용 모델기술 – 2) 모델 선정

유사 과제에 사용된 모델을 분석/평가하여 최종 모델 선정

최종 모델 선정

- 모델의 적용 가능성 분석 및 성능 평가를 통해 최종 모델 3개 선정
- 3개 모델을 Ensemble 하여 예측값을 도출



사용 변수 최소화 및 Custom Tuning으로 일반화된 예측모델 구현

변수 제한

- 사용하는 변수를 최소화하여 과적합 가능성이 낮은 일반화된(Generalized) 예측 모델을 구현

```
vars = ['hour', 'type', 'month',  'weekday', 'little_gas', '26~31','2~7']

X = train[vars]
y = train['amount']
log_y = np.log1p(y)
```

Custom Metric

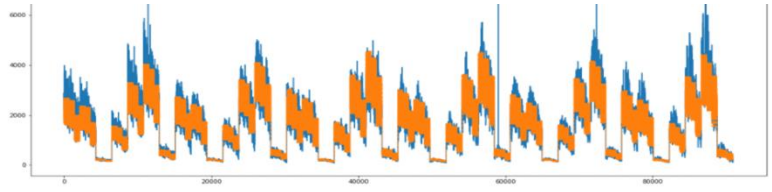
- 모델 학습시 평가지표(NMAE)와 동일한 Metric 적용

```
def nmae(y_pred, train_data):
    y_true = train_data.get_label()
    y_pred = np.expm1(y_pred)
    y_true = np.expm1(y_true)
    score = np.mean((np.abs(y_true-y_pred))/y_true)
    return 'nmae', score, False

model = lgb.train(params, train_set=d_training, feval = nmae)
```

매개변수 조정

- Optuna, Grid Search 사용시 과적합 발생
  ☞ 예측값의 시각화 결과를 확인하며 수동으로 조정



KFold

- KFold는 3~24 까지 적용 후 모델별 최적 Fold수 선택

XGB	LGBM	Catboost
24 Fold	6 Fold	12 Fold

과대추정 방지를 위해 Minimum Ensemble 적용

Ensemble

· 수요예측 비즈니스의 특성과 평가지표(NMAE)를 고려하여 과대추정 보다는 과소추정이 합리적이라고 판단

case1) 실제값 6000, 예측값 4000, NMAE = 0.3
case2) 실제값 500, 예측값 1500, NMAE = 2.0

→ 실제값이 작을때 작게 예측하는 것이 중요

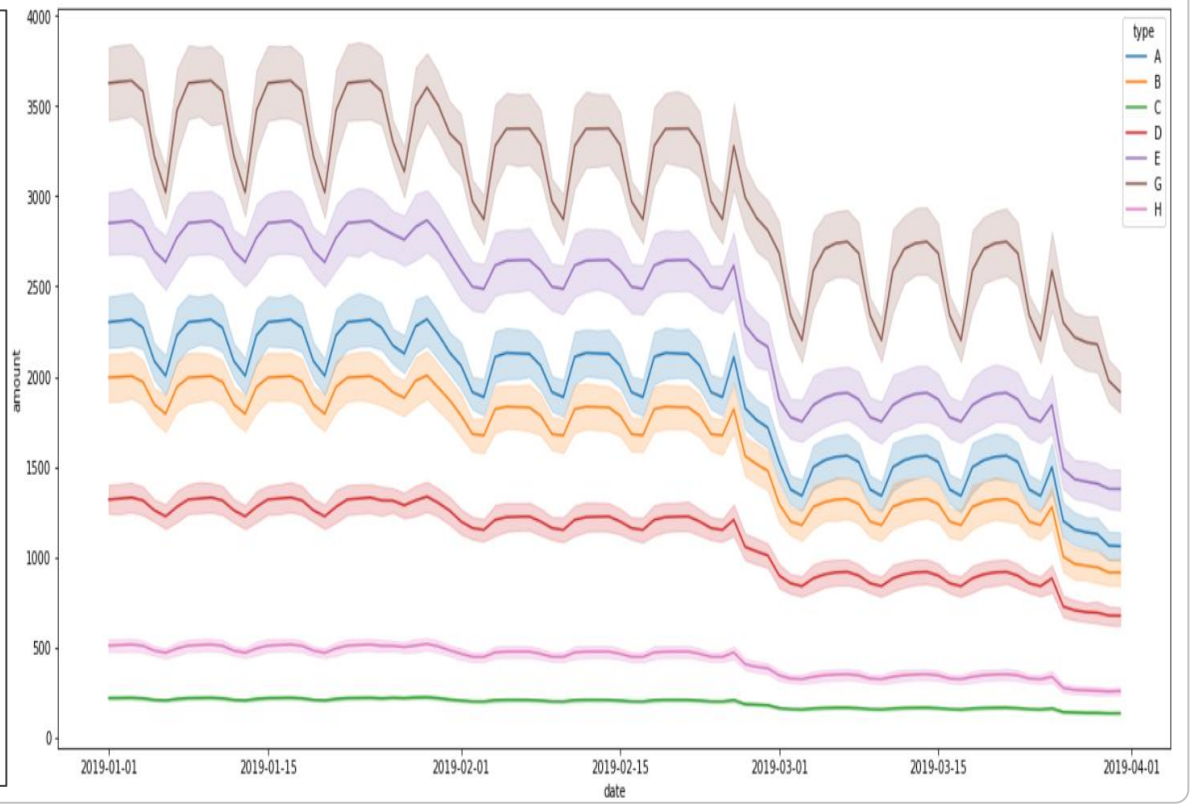
· 3개 모델의 예측값 중 최소값을 최종 예측값으로 선택하는 'Minimum Ensemble' 적용

	일자|시간|구분	공급량	lgb_pred	xgb_pred	cat_pred
0	2019-01-01 01 A	2061.970535	2101.111564	2061.970535	2161.203070
1	2019-01-01 02 A	1792.151836	1809.501060	1794.805046	1792.151836
2	2019-01-01 03 A	1689.663193	1692.503973	1689.663193	1697.251391
3	2019-01-01 04 A	1727.195526	1727.287962	1727.195526	1737.477559
4	2019-01-01 05 A	1918.173037	1922.780823	1928.554031	1918.173037

4. 모델링 결과

예측값

	일자	시간	구분	공급량
0	2019-01-01	01	A	2061.970535
1	2019-01-01	02	A	1794.805046
2	2019-01-01	03	A	1689.663193
3	2019-01-01	04	A	1727.195526
4	2019-01-01	05	A	1922.059821
...	...	...	...	...
15115	2019-03-31	20	H	333.837695
15116	2019-03-31	21	H	328.356260
15117	2019-03-31	22	H	320.012548
15118	2019-03-31	23	H	283.641309
15119	2019-03-31	24	H	280.750799



최종 성적

10th	데린이		0.09486	14

9. 참고 자료

논문

- 도시가스 수요량 예측을 위한 시계열 모형 개발(2009) 최보승, 강현철, 이경윤, 한상태
- 국내 도시가스의 시간대별 수요 예측(2016) 한정희, 이근철
- 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측(2018) 배상완, 유정석
- 외재적 변수를 이용한 딥러닝 예측 기반의 도시가스 인수량 예측(2019) 김지현, 김지은, 박상준, 박운학
- 함수 주성분 분석을 이용한 일별 도시가스 수요 예측(2020) 최용옥, 박혜성
- 도시가스 일 최대수요 예측에 관한 연구(2020) 박철웅, 박철호

데이터

- 도시가스 월별 상대가격 지수 데이터 (출처 : 한국가스공사)
- 기상 정보 데이터 (출처 : 기상자료개방포털)
- 지역별 특수일 효과 (출처 : 한국가스공사)
- 공휴일 데이터 (출처 : 한국천문연구원 특일정보 API)

경진대회

- 전력 사용량 예측 AI 경진대회
- 태양광 발전량 예측 AI 경진대회