Team Project 2nd - 으쌰으쌰 팀😎

팀장 가채원, 부팀장 김범중, 김진연, 윤진훈

한국가스공사 가스 사용량 예측하기

가스공급량 수요예측 모델개발

한국가스공사의 시간단위 공급량 내부데이터와 기상정보 및 가스외 발전량 등 외부데이터를 포함한 데이터셋을 구축하여 90일 한도 일간 공급량을 예측하는 인공지능 모델을 개발한다.

가설 세우기

> 💡 **기온이 낮을수록 도시가스 공급량이 늘어난다.**

- 가정용 도시가스는 난방에 사용된다.
- 기온이 낮으면 사람들은 난방을 더 많이 사용한다.

∴ 기온이 낮으면 도시가스 공급량이 늘어날 것이다.

가설 확인하기

1. 1, 4분기가 2, 3분기에 비해 기온이 낮다.
2. 1, 4분기에 2, 3분기보다 공급량이 많다.
3. 가스 공급량과 기온 데이터 중 이상치가 없는지 확인한다.

기본 데이터 분석

https://public.tableau.com/views/_16347799122130/sheet0?:showVizHome=no&:embed=true#2

분석 내용

분기별 공급량 합계

[상대적] 1분기와 4분기의 공급량 합계 > 2분기와 3분기의 공급량 합계

⇒ '가설 확인하기-2' 확인 가능

공급사별 공급량 합계

공급사별 공급량 차이 존재

⇒ 공급사 : 공급량 예측의 중요 변수

공급사별 월별 공급량 합계

월별 분석 결과, 가스 공급량 U자 그래프 확인 가능

⇒ 기온 그래프가 A자 그래프를 그린다면, 기온이 낮아질 시 가스 공급량 증가의 가설 신빙성 증가

대한민국 평균 기온 분석

출처: 기상청 기상자료개방포털 링크

https://public.tableau.com/views/20132018/sheet1?:showVizHome=no&:embed=true

분석 내용

월별 평균기온

⇒ 기온 그래프: A자 그래프 확인 가능

⇒ 2013 - 2018년 중 다른 년도 대비 이상치를 보이는 연도 X

분기별 평균기온

⇒ 분기별 가스 공급량 합계와 반대 모습(U↔A) 확인 가능

채택하지 않은 가설

> ## 천연가스 가격 상승 시, 가스 공급량 감소 예상

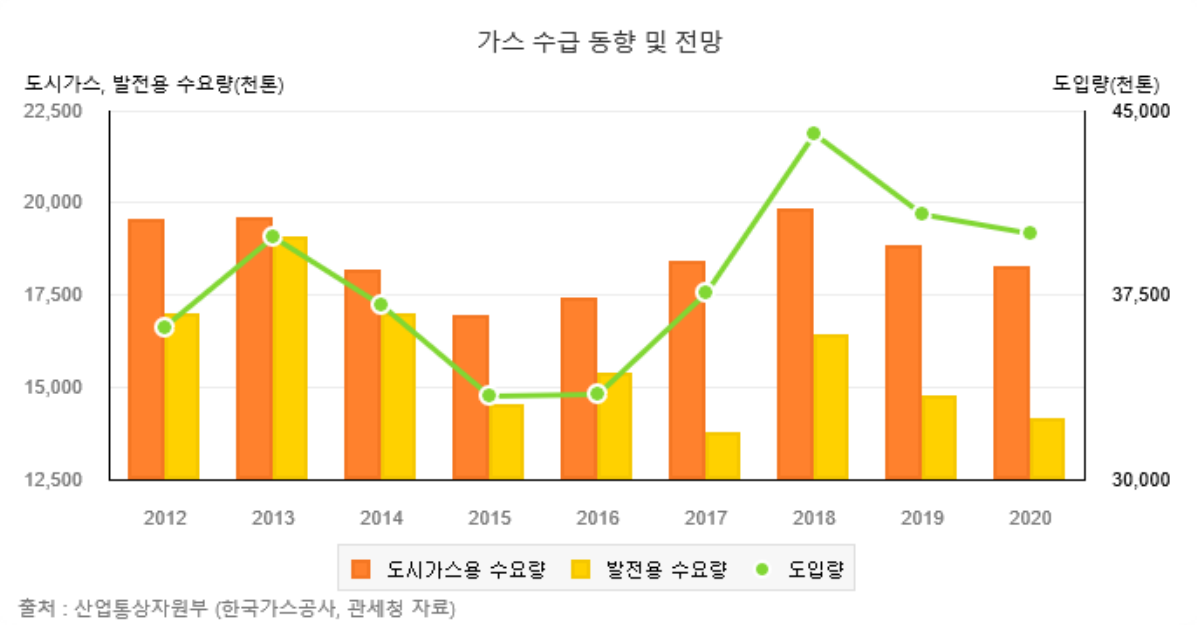
국제 천연가스 가격 상승 ⇒ 가스 공급에 차질 발생 가능 ⇒ 가스 공급량 감소

채택하지 않은 이유

e-나라지표 산업통상자원부 산하 한국가스공사 LNG 수급동향 정보

e-나라지표 지표조회상세

https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1165



천연가스(LNG) 수급 동향 지표의 의의 및 활용도

- LNG 수요는 청정 고급 에너지의 선호에 따라 매년 큰 폭 증가

  ∴ LNG 중장기 도입계약 등을 통한 공급 안정성 확보 필수

- 연도별 물량의 변화

  ⇒ 당해년도 LNG 수요는 당해년도 LNG 공급가능물량(기초재고 + 도입물량)으로 충당

  [유의사항] 당해년도 LNG 수요량과 도입량 간 차이 발생

LNG 수급은 중장기 도입계약으로 이루어지며, 공급량은 당해년도 공급가능물량에 기반하고 있다.

LNG 수요량과 도입량에는 차이 발생 가능

> 💡 **국제시장에서 변동되는 천연가스 가격과 가스 공급량은 큰 관련성 X**

기온 데이터 수집하기

- 분기별 가스 공급량의 변화와 기온의 변화가 음의 상관관계를 가지며, 모델의 성능을 높이고자 기온 데이터 추가.

시간별 기온 데이터 수집하기

기상청 기상자료개방포털에서 제공하는 자료는 일별 기온 자료만 제공한다.

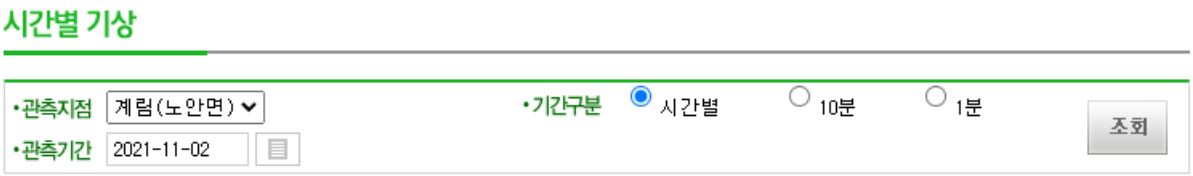
시간별 기온 자료는 따로 수집하여 전처리 할 필요가 있었다.

시간별 기온 데이터 출처: 나주시 농업기상정보시스템

:::나주시 농업기상정보시스템:::

나주시농업기상정보시스템

https://weather.naju.go.kr/agri_meteo/agri_time.html



시간별 기상자료

관측시간 (년-월-일 시:분)	기온 (℃)	습도 (%)	풍향	풍속(km/h) 평균	풍속(km/h) 최대순간	기압 (hPa)	강수량 (mm)	이슬 유무	일사량 (W/㎡)	지중 온도 (%)	초상 온도 (℃)	토양 수분 (%)
2021-11-02 22:00	5.7	83.2	서남서	0	2.5	-	0.0		-	7.9	6.0	28.4
2021-11-02 21:00	7.1	82.3	남서	0	2.2	-	0.0		-	8.7	7.0	28.7
2021-11-02 20:00	8.1	79.2	서남서	0	2.9	-	0.0		-	9.7	8.0	28.9
2021-11-02 19:00	10.7	70.9	서남서	1.8	5.4	-	0.0		-	11.3	9.8	29.2
2021-11-02 18:00	12.5	63.7	서남서	2.9	9.7	-	0.0		-	13.3	12.2	29.3
2021-11-02 17:00	14.1	56.3	서남서	4.7	15.1	-	0.0		-	15.2	14.4	29.4
2021-11-02 16:00	15.8	50.1	서남서	5.4	14.8	-	0.0		-	17.0	16.5	29.3
2021-11-02 15:00	16.7	47.4	서남서	6.8	17.6	-	0.0		-	18.3	17.9	29.3

데이터 수집 및 전처리 과정 중 겪은 문제점

▼ **웹 크롤링** 코드

▼ 주요 문제

- 사이트 날짜, 기온 데이터 결측치 존재
  - 해결1) 해당 날짜의 시간별 리스트 만들어 대조해서 확인.
  - 해결2) 날짜가 없으면 해당 날짜 채우고, 기온이 없을 경우 NaN값으로 처리.
- bs에 담은 자료에서 날짜, 기온의 데이터 인덱스 순서가 달라지는 문제
  - 해결1) 인덱스 규칙에 따라 함수 작성

        - 해결2) 조건문으로 적용

▼ 그 외 문제

- 달마다 날짜가 다른 문제
    - 윤년과 30, 31일 존재
        - 해결1) 특정 연, 월, 일 크롤링 함수 작성
        - 해결2) 조건문으로 적용
- 중복 날짜 데이터가 수집되고, 날짜가 내림차순 정렬되있는 문제
    - 해결) 날짜 데이터를 set으로 변환해 중복 제거, list로 변환해 오름차순 정렬.
- 기온데이터에서 해당 날짜의 24시가 다음날 0시로 표현되는 문제.
    - 해결) 해당 날짜의 마지막 0시를 하루 당겨 24시로 변경.

▼ **데이터 전처리** 코드

▼ 주요 문제

- 특정 날짜, 오후, 오전 시간 때 기온 데이터 없는 문제.
    - 해결) 같은 달, 일, 시간의 기온 평균으로 결측치 처리.
- 매년 12월31일 24시 기온 데이터 없는 문제.
    - 해결) 연도별 바로 전시간 23시 기온으로 결측치 처리.

이슈 정리 표

머신러닝 모델 비교

사용한 머신러닝 모델

- LinearRegression, Lasso, Ridge, KNeighborRegression, DecesionTreeRegressor
- 앙상블 모델(RandomForest, XGBRegressor, LGBMRegressor)

머신러닝 모델 비교

Aa 머신러닝 모델	# 정확도	# 교차검증 정확도 평균

Aa 머신러닝 모델	# 정확도	# 교차검증 정확도 평균
LinearRegression	0.038	0.037
Lasso(alpha=0.01)	0.034	0.034
Ridge(alpha=0.0001)	0.034	0.034
KNeighborRegression	0.773	0.664
DecesionTreeRegressor	0.983	0.958
RandomForest	0.99	0.976
XGBRegressor	0.983	0.982
LGBMRegressor	0.971	0.97

기온 데이터 추가하기

2019년 기온 데이터를 머신러닝 모델로 추가,이를 기반으로 다시 공급량을 추정하는 모델 작성 코드

1. `XGBRegressor` 를 이용하여 2019년 기온 데이터 예측
    a. 기온 데이터 정확도: `0.976`
2. `XGBRegressor` 를 이용하여 공급량 추정하는 모델 작성 후 검증

머신러닝 모델 비교(기온 데이터 추가)

Aa 머신러닝 모델	# 정확도	# 교차검증 정확도 평균
LinearRegression	0.268	0.268
Lasso(alpha=0.01)	0.263	0.263
Ridge(alpha=0.0001)	0.263	0.263
KNeighborRegression	0.613	0.468
DecesionTreeRegressor	0.967	0.951
RandomForest	0.984	0.976
XGBRegressor	0.986	0.984
LGBMRegressor	0.979	0.978

하이퍼파라미터 튜닝하기

```
param_grid = {
    'n_estimators': [100, 150, 200],
    'max_depth': [3, 6, 9, 12],
    'min_samples_split': [0.01, 0.05, 0.1]
}

kf = KFold(random_state=30,
           n_splits=10,
           shuffle=True,
          )

grid_search = GridSearchCV(estimator=model_random,
                           param_grid=param_grid,
                           cv=kf,
                           n_jobs=-1,
                           verbose=2
                          )

grid_search.fit(X_train, y_train)

print(grid_search.best_params_)
### {'max_depth': 12, 'min_samples_split': 0.01, 'n_estimators': 200}
```

하이퍼파라미터 튜닝 후

Aa 머신러닝 모델	기온	튜닝	# 정확도	# 교차검증 정확도 평균
RandomForest	미포함	O	0.923	0.927
RandomForest	포함	O	0.923	0.927
RandomForest	미포함	X	0.99	0.976
RandomForest	포함	X	0.984	0.976

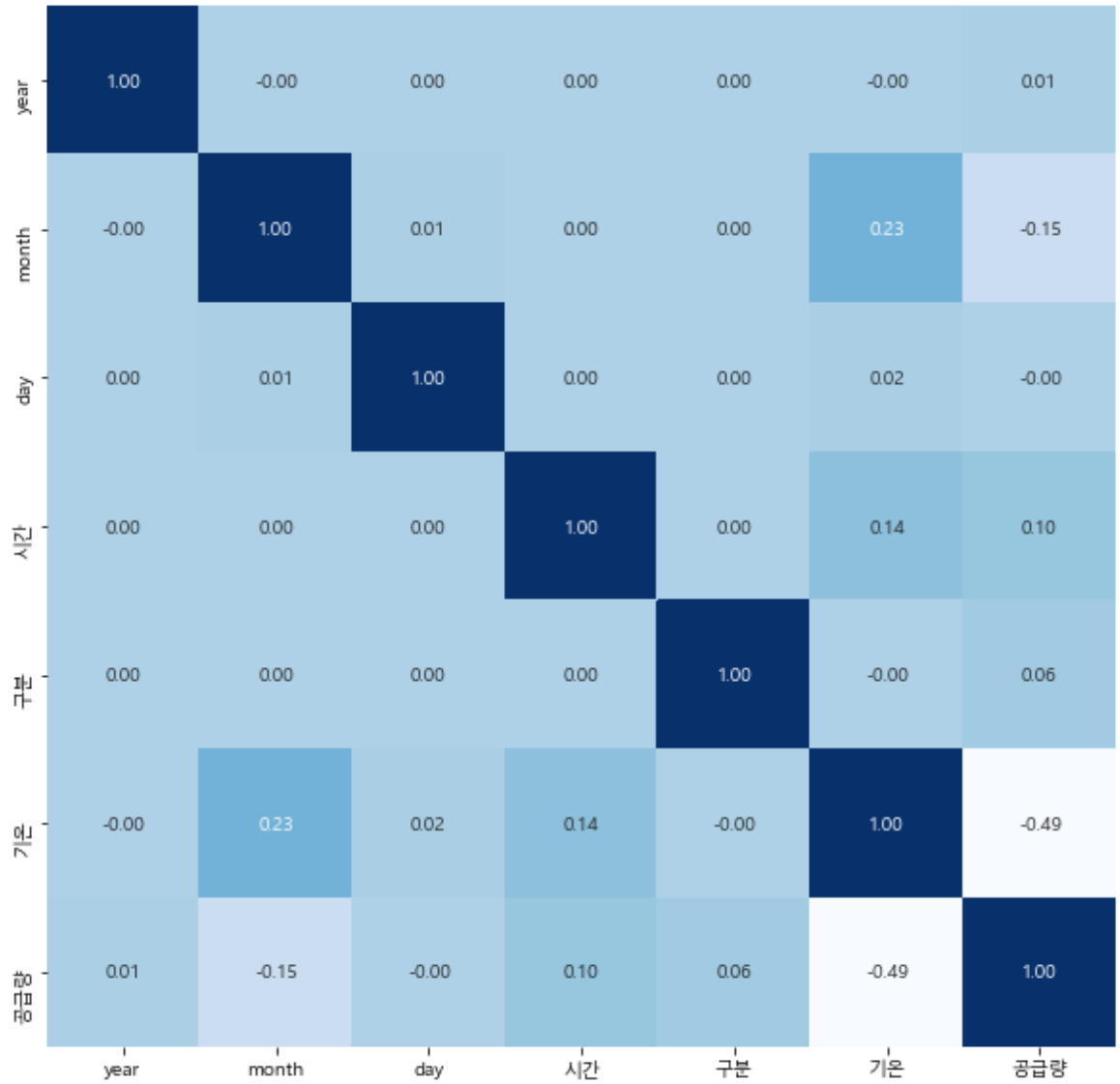
하이퍼파라미터 조정 후 정확도가 낮아지는 결과, 하지만

> **실제로 제출해보니 점수는 0.1653877251로 시도해 본 것 중 가장 높았다.**

왜 정확도가 낮은 모델이 더 높은 점수를 얻었을까?

💡 **기존에 시도하던 머신러닝 모델들이 과적합되었다.**

특성 선택 변경



- 공급량과 상관관계가 큰 특성만 선택
  - 'month', '시간', '구분', '기온'
    - 모델 : RandomForestRegressor
    - 점수 : 0.1181514058
- 모든 특성 사용 시보다 자체 결정 계수는 떨어졌지만 제출 점수는 상승.
  - 적절한 특성 선택이 과대적합을 해소한 것으로 판단
  - 자체 평가 점수
    - 전 : 0.95 ➡️ 후 : 0.87
  - 제출 점수
    - 전 : 0.16 ➡️ 후 : 0.11

기말 발표까지 목표

- [ ] 기존 머신러닝 모델들의 과적합 문제 해결
- [ ] RandomForest 외 다른 모델들 하이퍼파라미터 튜닝
- [ ] 딥러닝 모델로 학습
- [ ] 기온 데이터 외에 적용 가능한 외부 데이터 적용
- [ ] polynomialfeature와 특성선택을 이용