

삼삼사자

데이콘 가스 수요량 예측

박승규, 이성준, 노현곤





목차

01

프로젝트 개요

02

프로젝트 팀 구성 및 역할

03

프로젝트 수행 절차 및 방법

04

프로젝트 수행 결과

05

자체 평가 의견



대회명 : 데이콘 가스공급량 수요예측 모델개발 대회

<https://dacon.io/competitions/official/235830/overview/description>

- ◆ 주제 : 한국가스공사의 시간단위 가스 공급량 데이터와 기상 데이터 및 유가 데이터를 종합한 데이터셋을 구축하여 90일 한도 일간 공급량을 예측하는 인공지능 모델을 개발
- ◆ 주최 및 주관 : 한국가스공사
- ◆ 대회 개요: 한국 가스 공사가 보유한 다년간 시간 단위 공급량 데이터를 기반으로 미래 공급량을 예측하는 모델을 만든다.



프로젝트 팀 구성 및 역할

성 명	분담 내용	역 할
박승규	<ul style="list-style-type: none">- 데이터 (날씨, 유가, 가스 공급량) 병합 및 결측치 처리- feature Engineering을 통한 변수 생성(통계량을 이용한 특성생성, 모델을 이용한 특성 생성), 모델 성능 개선- 모델 평가(교차 검증, 그리드 리서치 함수 등을 활용)- plotly를 이용한 데이터 시각화, 발표 자료 준비	팀장
이성준	<ul style="list-style-type: none">- MSE, RMSE를 활용하고 NMAE평가 지표를 만든 후 모델 최종 평가- 교차 검증 및 파라미터 튜닝(max_depth, n_estimators, max_features), test_size를 크기를 달리해서 비교- matplotlib, seaborn 등을 활용한 데이터 탐색 및 시각화- Pycaret을 이용한 ML 성능 향상- ppt 자료 정리	부팀장
노현곤	<p>데이터 수집</p> <ul style="list-style-type: none">- 데이터 탐색 및 수집	팀원

프로젝트 수행 절차 및 방법

가정·상업용 도시가스의 경우는 소비의 많은 부분이 기온변화로 설명될 수 있다. 수요가수가 포화시점에 근접하였으며 대부분 난방용으로 사용되는 가정·상업용 도시가스 소비 증감율은 난방도일¹⁾의 증감
출처 : 산업용 도시가스 수요변화 요인분석(에너지경제연구원)

행정구역별	2018		
	보급률 (%)	공급권역내 총 가구수 (가구)	도시가스 수요가구수 (가구)
합계	85.0	21,674,404	18,429,378
경기도	88.0	5,306,214	4,669,015
서울특별시	98.2	4,263,868	4,186,336
부산광역시	92.3	1,480,468	1,367,105
인천광역시	92.9	1,115,997	1,201,455
경상남도	75.3	1,360,084	1,023,557
대구광역시	96.4	1,021,266	984,148
경상북도	65.9	1,179,225	776,786
충청남도	70.8	916,667	649,389
광주광역시	99.9	603,107	602,499
대전광역시	94.8	624,965	592,467
전라북도	71.6	806,235	576,948
충청북도	66.2	705,471	466,926
울산광역시	93.9	461,756	433,523



기상청

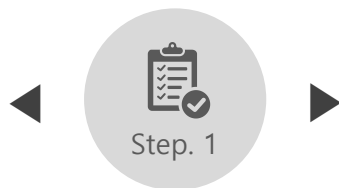
Korea Meteorological
Administration

도시가스 수요가구수(가구)가 가장 큰 서울특별시 데이터를 데이터셋으로 pycaret을 활용하여 가장 좋은 성능의 ML기법 확인 및 활용

프로젝트 수행 결과

데이터 확인

날씨 데이터에서 결측치
다수 발견



전처리

결측치에서 가장 가까운 값들의 평균으로 처리
ex) 2015-01-01 00:00 (NaN)
(2014-12-31 23:00 기온 + 2015-01-01 01:00 기온) / 2

```
1 df[df['기온(* c)'].isnull()]
```

	지점	지점명	연월일	시간	기온(* c)
17519	108	서울	2014-12-31	24	NaN
41512	108	서울	2017-09-26	17	NaN
41894	108	서울	2017-10-12	15	NaN
41895	108	서울	2017-10-12	16	NaN
41896	108	서울	2017-10-12	17	NaN
51810	108	서울	2018-11-29	19	NaN

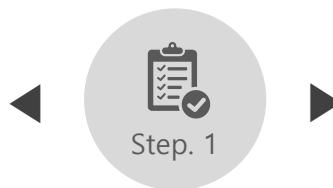
```
1 # 17519번 뒤 아래 값 평균으로 대체
2 df.loc[17519] = (108, '서울', '2014-12-31', 24, (-6.2-7.4)/2)
3 df.loc[[17518, 17519, 17520], :]
```

	지점	지점명	연월일	시간	기온(* c)
17518	108	서울	2014-12-31	23	-6.2
17519	108	서울	2014-12-31	24	-6.8
17520	108	서울	2015-01-01	1	-7.4

프로젝트 수행 결과

데이터 확인

공급량 데이터의 날짜를
year, month, day, weekday로 나눠줌



전처리

공급량 데이터의 날짜를
year, month, day, weekday로 나눠줌

ex) 연월일 2013-01-01

Year	month	day	weekday
2013	01	01	1

```
1 df["year"] = df["연월일"].dt.year
2 df["month"] = df["연월일"].dt.month
3 df["day"] = df["연월일"].dt.day
4 df["hour"] = df["시간"]
5 df["weekday"] = df["연월일"].dt.weekday
```

1 df.head()

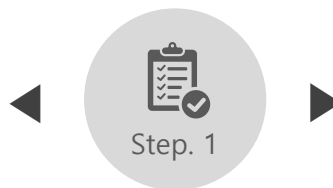
	지점	지점명	연월일	시간	기온(* C)	year	month	day	hour	weekday
0	108	서울	2013-01-01	1	-8.5	2013	1	1	1	1
1	108	서울	2013-01-01	2	-8.4	2013	1	1	2	1
2	108	서울	2013-01-01	3	-8.1	2013	1	1	3	1
3	108	서울	2013-01-01	4	-8.2	2013	1	1	4	1
4	108	서울	2013-01-01	5	-8.2	2013	1	1	5	1

프로젝트 수행 결과

데이터 확인

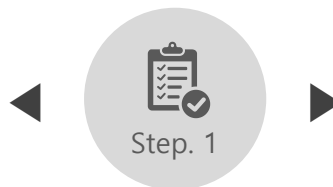
전처리

구분값이 문자열임



구분값을 숫자열로 바꿔줌
ex) M → 2

날씨 데이터와 공급량 데이터
의 시간 표기 차이 발견

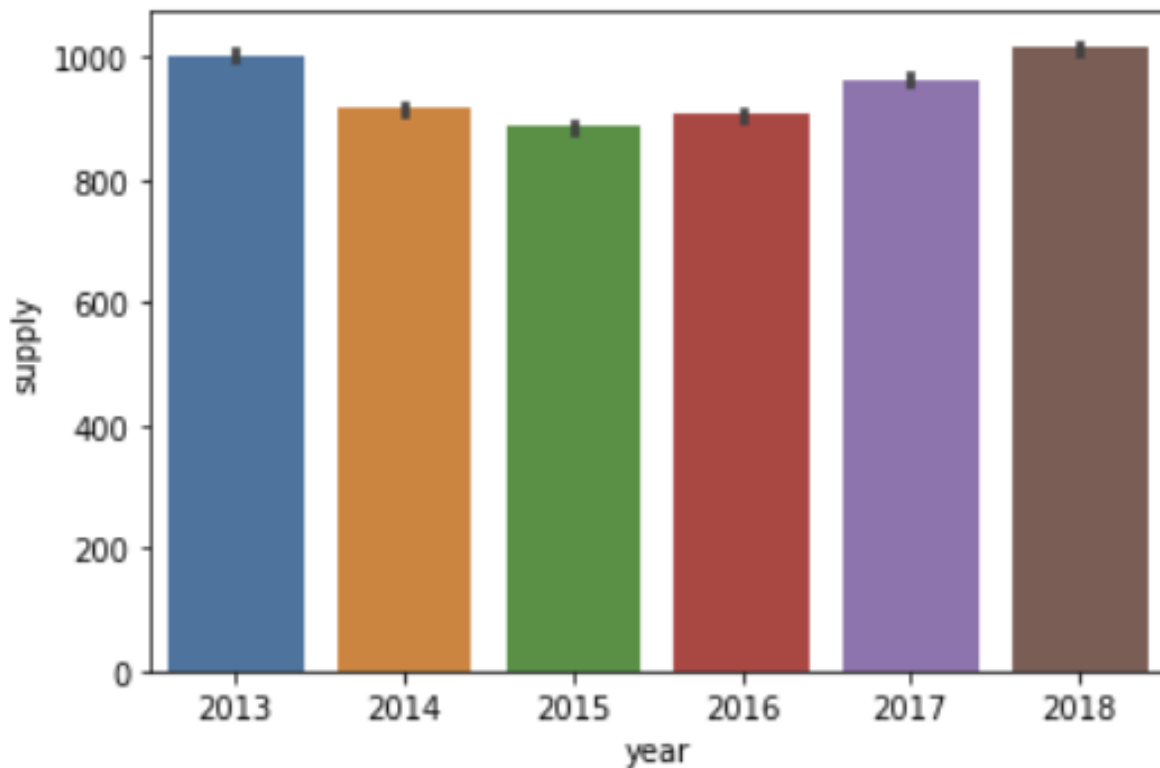


공급량 데이터 기준으로 변경
ex) 2014-01-01 0:00
→ 2013-12-31 | 24

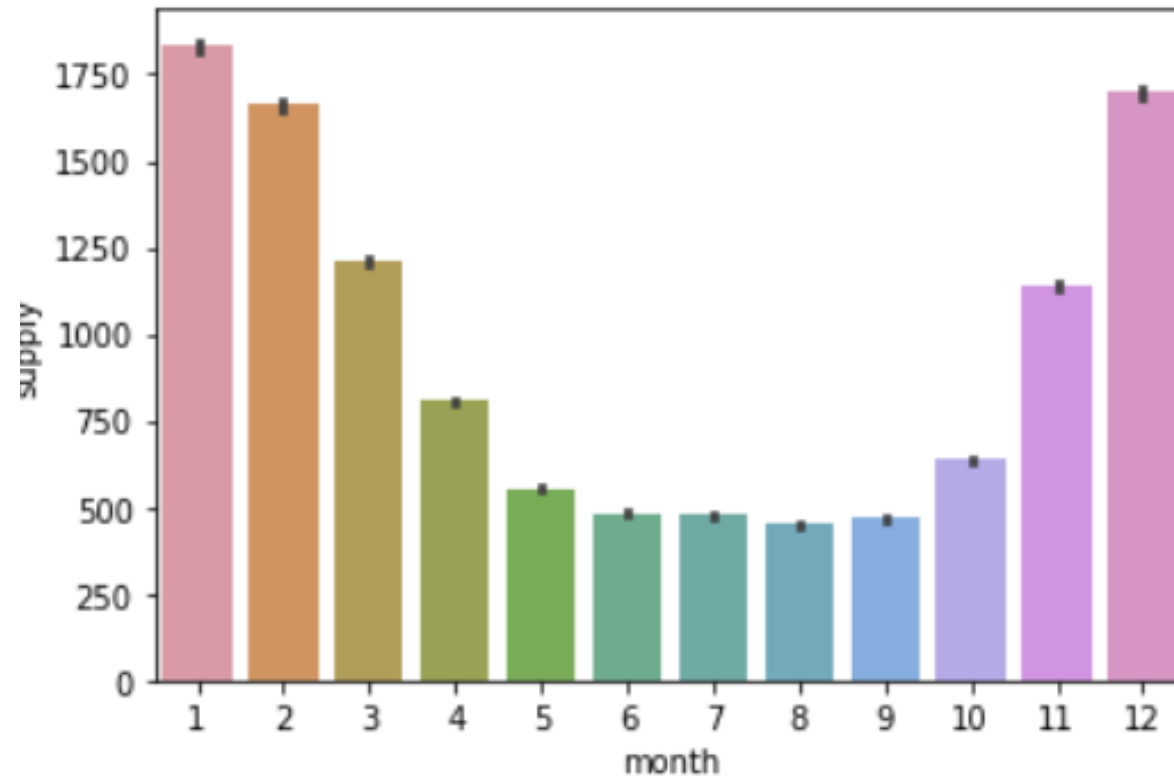
	연월일	시간	일시	year	month	day	hour	weekday	구분	구분_E	공급량	기온(°C)
368078	2018-12-31	15	2018-12-31 14:00:00	2018	12	31	14	0	H	6	525.488	1.800000
368079	2018-12-31	16	2018-12-31 15:00:00	2018	12	31	15	0	H	6	518.009	2.066667
368080	2018-12-31	17	2018-12-31 16:00:00	2018	12	31	16	0	H	6	542.360	0.933333

프로젝트 수행 결과

연도별 가스 공급량

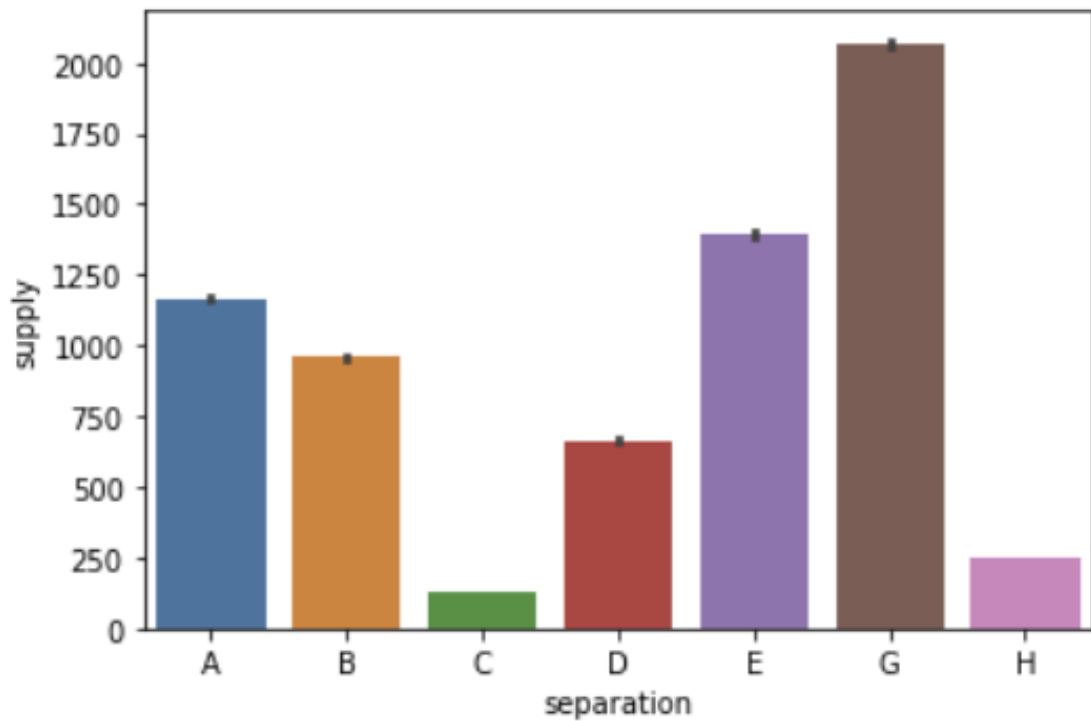


월별 가스 공급량

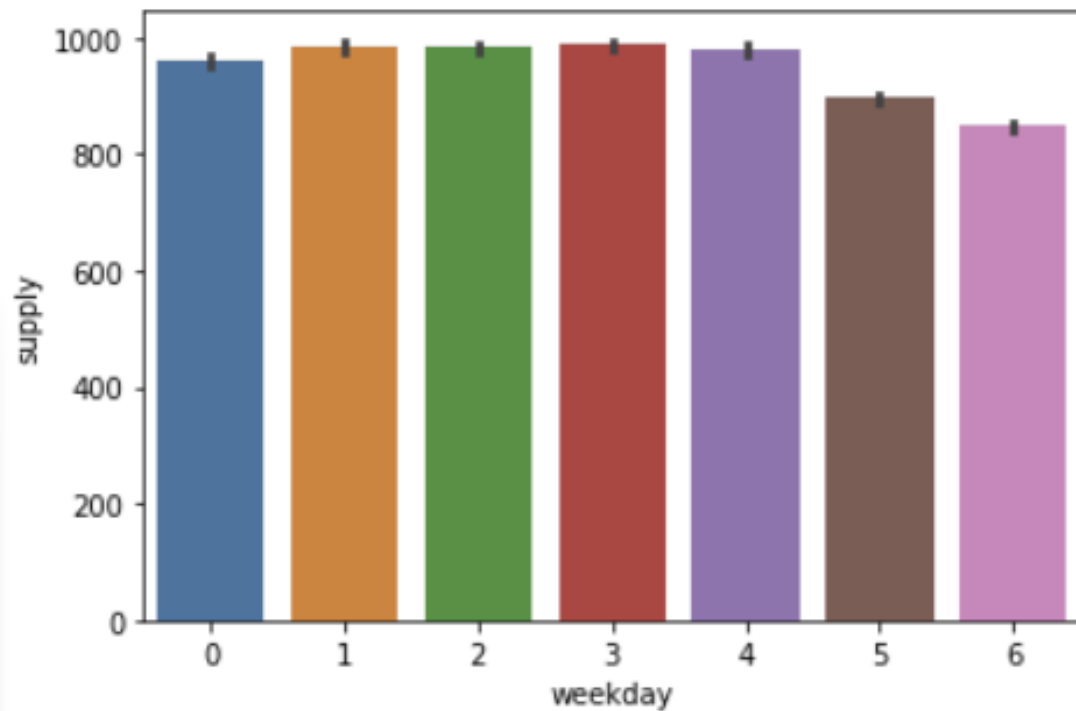


프로젝트 수행 결과

구분별 가스 공급량

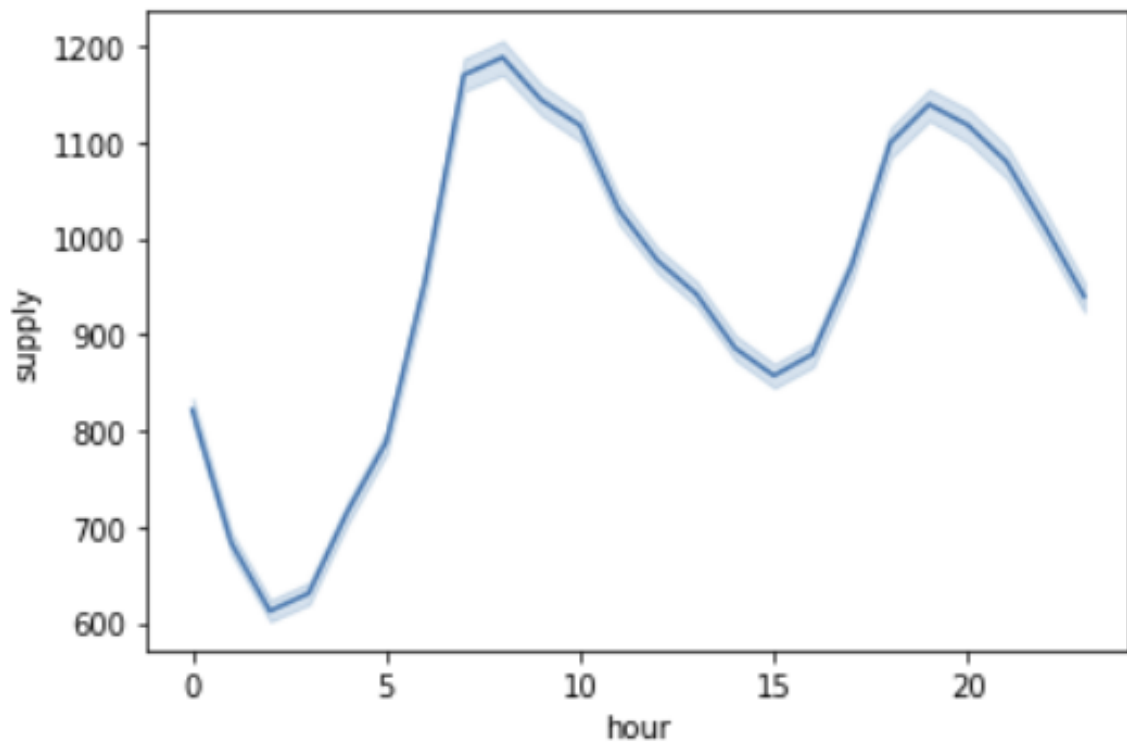


요일별 가스 공급량

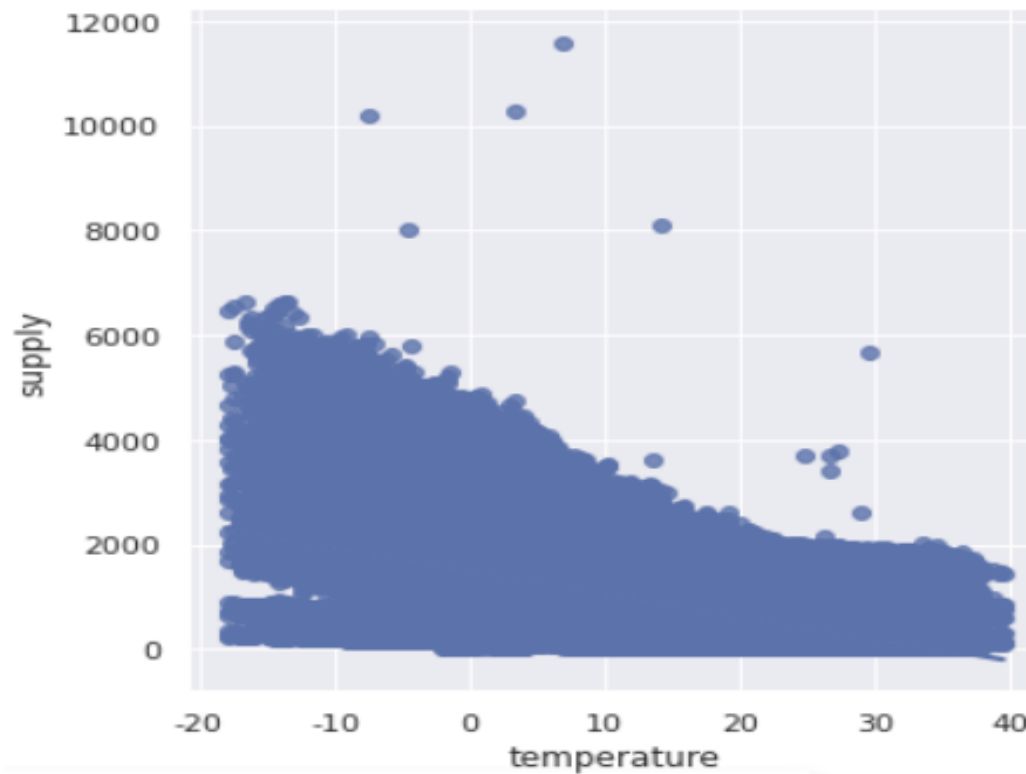


프로젝트 수행 결과

시간별 가스 공급량

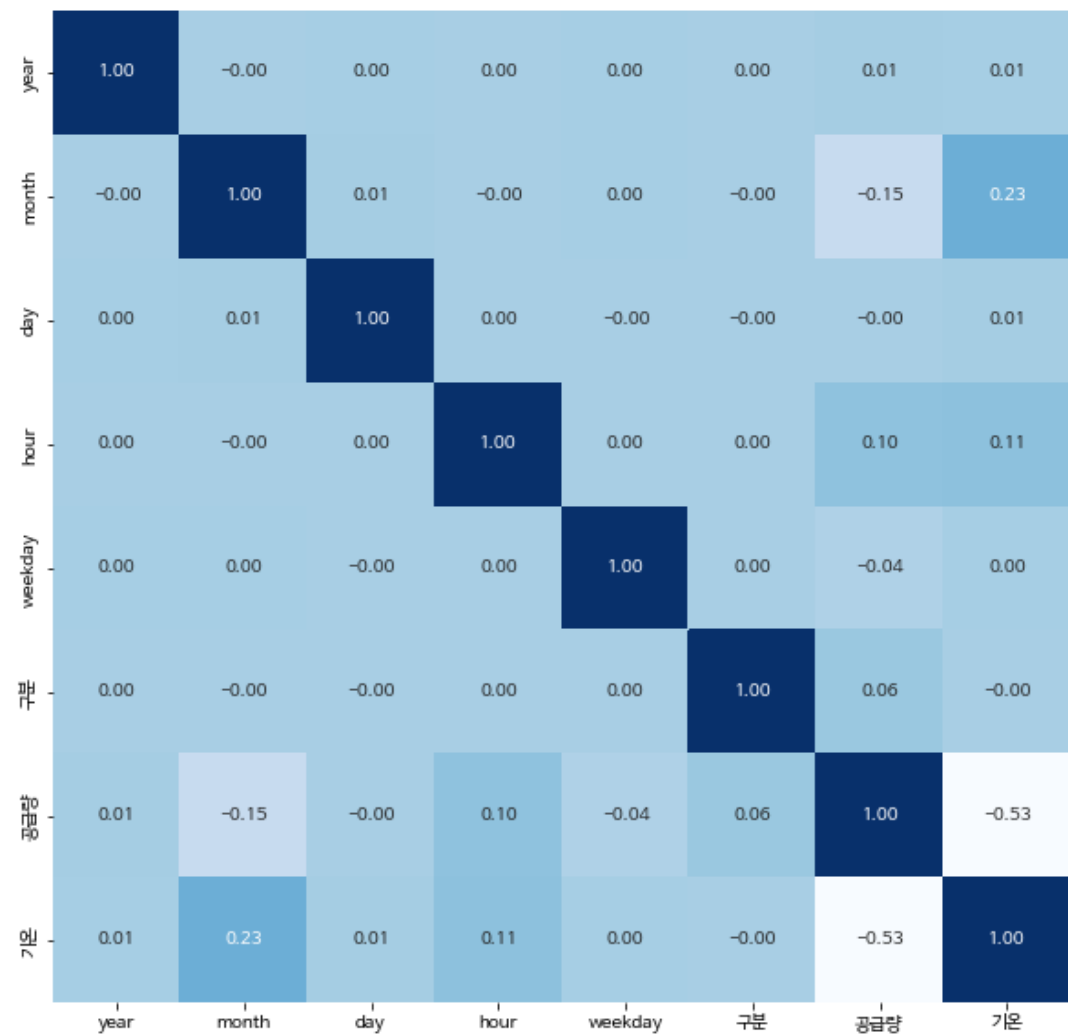


온도 가스 공급량 관계



프로젝트 수행 결과

데이터 상관관계



가스공급량과의 상관관계

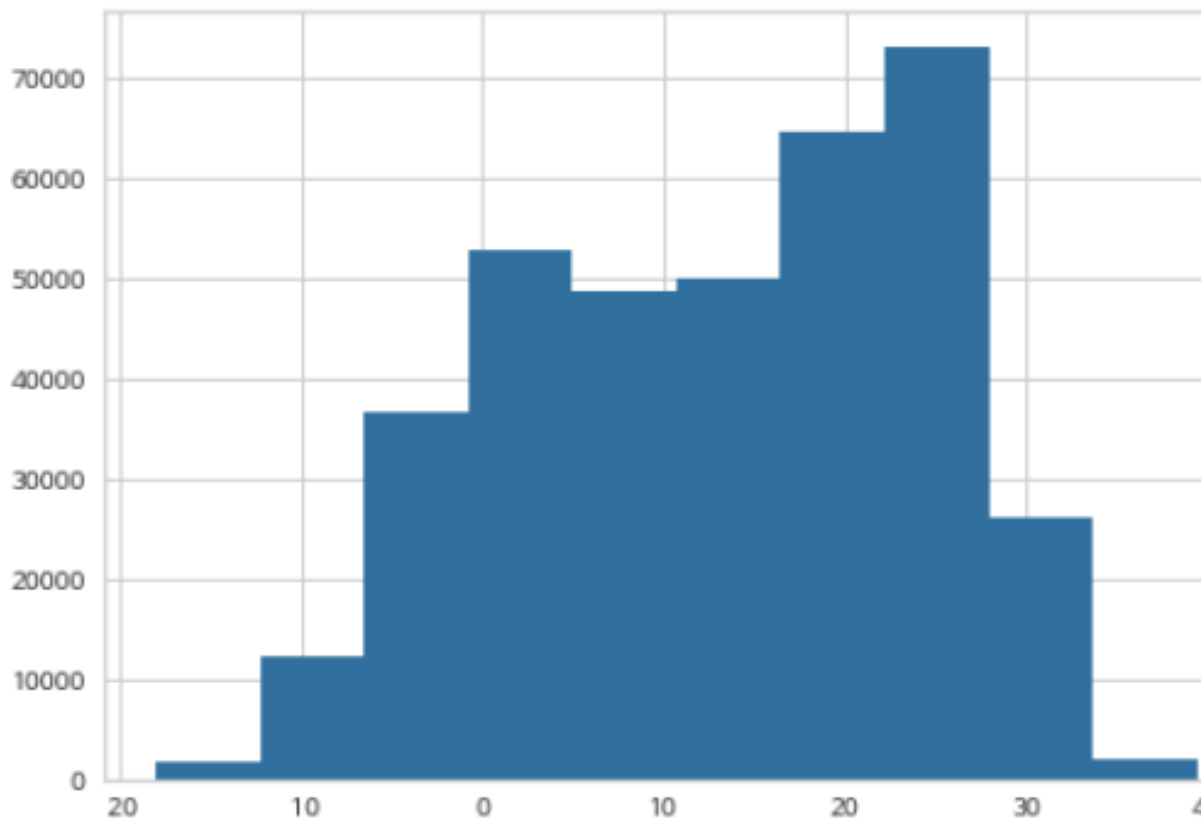
	기온	month	hour	구분	weekday	year	day
Corr	-0.53	-0.15	0.10	0.06	-0.04	0.01	0.00
Corr 절대값	0.53	0.15	0.10	0.06	0.04	0.01	0.00

Top 3

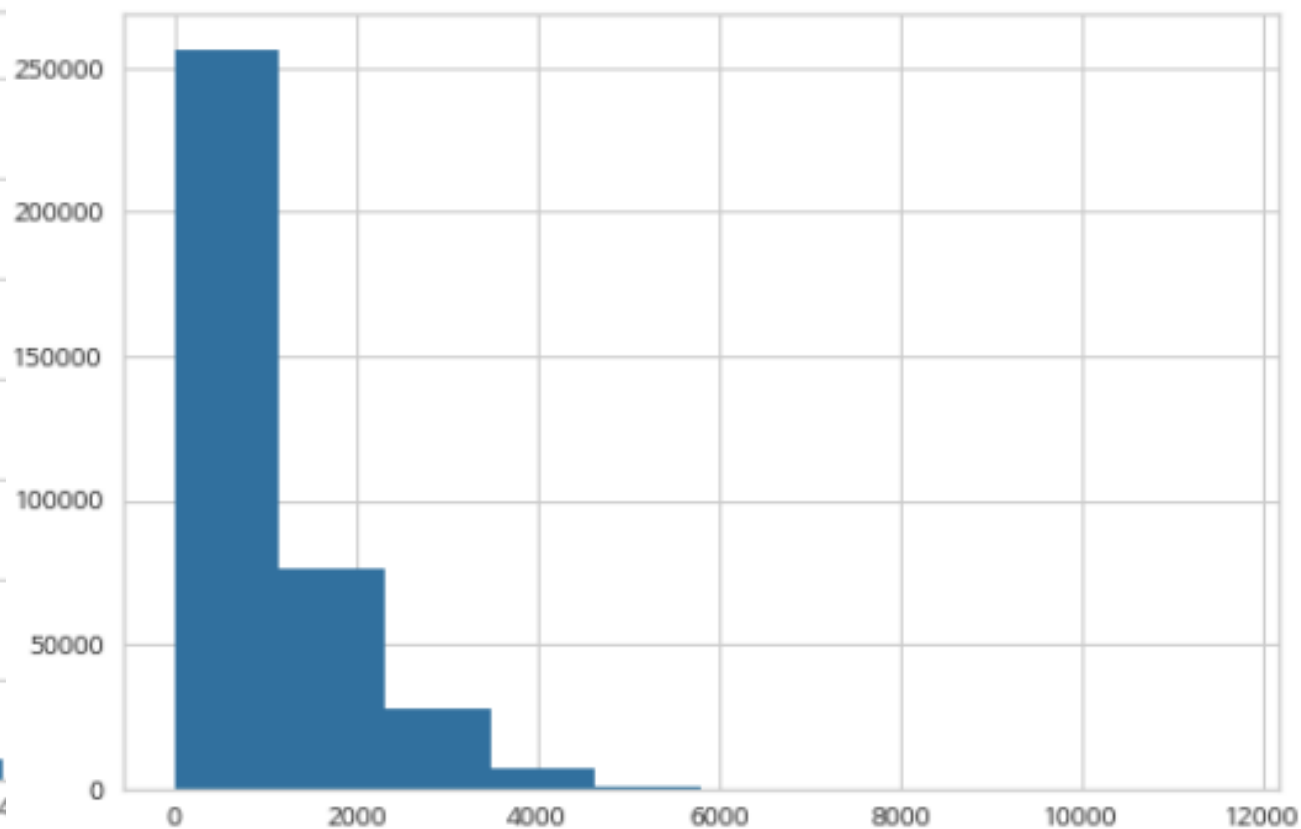
1. Month : 0.15
2. Hour : 0.10
3. 구분 : 0.06

프로젝트 수행 결과

기온 히스토그램

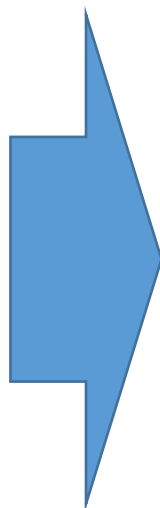
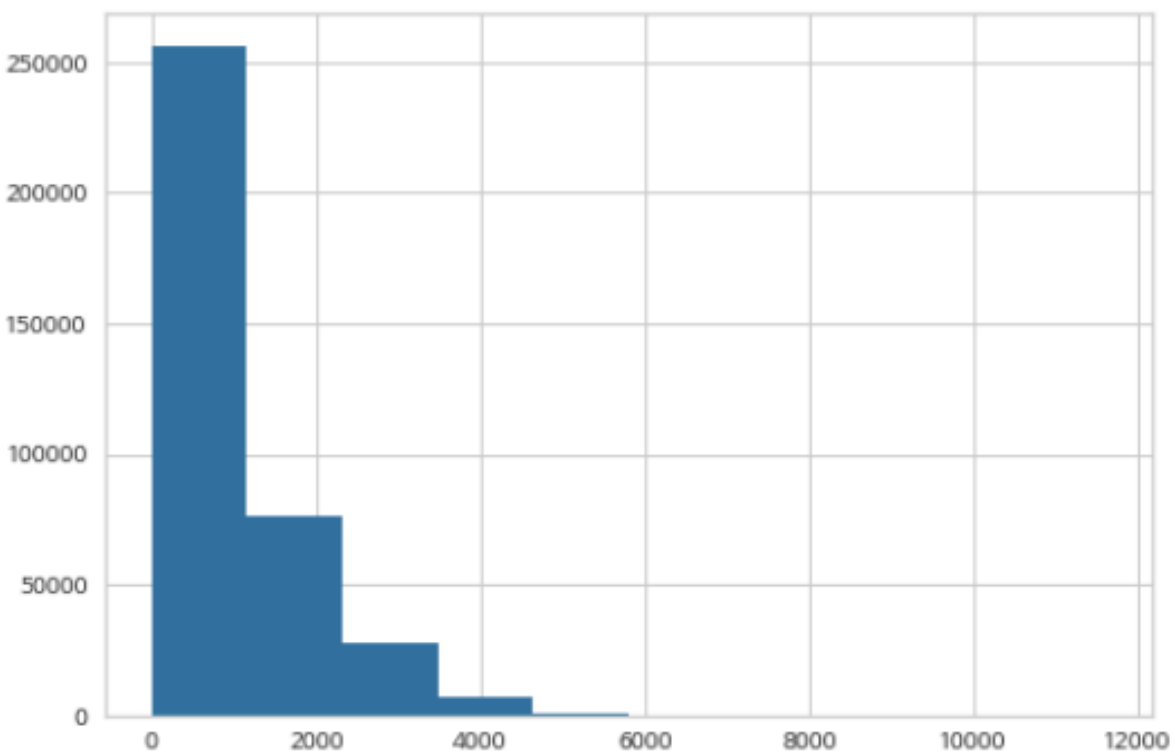


공급량 히스토그램

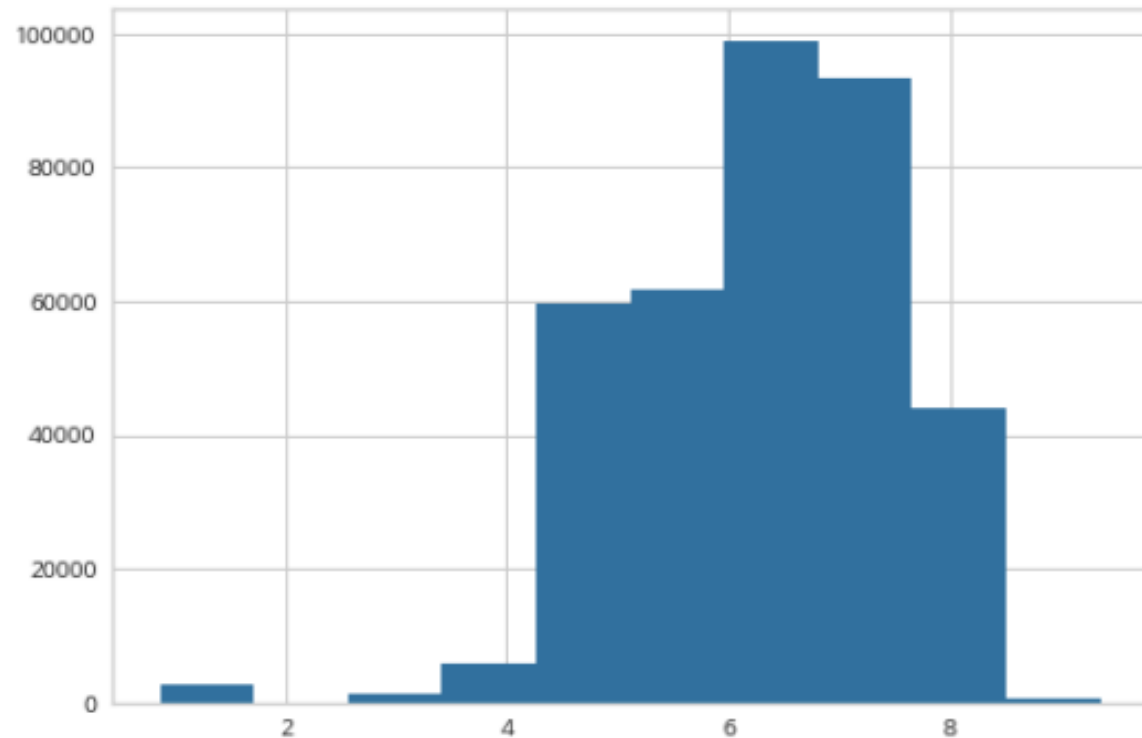


로그화 필요

공급량 히스토그램



log(공급량 히스토그램)



Pycaret

데이터 전처리, 모델링, 하이퍼파라미터 튜닝 등 여러 단계의 머신러닝 프로세스를 자동화

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	3.0148	15.1009	3.8860	0.8812	0.5163	0.8050	1.207
catboost	CatBoost Regressor	3.0163	15.1161	3.8879	0.8810	0.5169	0.8060	8.592
gbr	Gradient Boosting Regressor	3.0224	15.1601	3.8936	0.8807	0.5097	0.8082	7.964
et	Extra Trees Regressor	3.0314	15.2658	3.9071	0.8799	0.5190	0.8104	10.469
dt	Decision Tree Regressor	3.0314	15.2658	3.9071	0.8799	0.5190	0.8104	0.216
rf	Random Forest Regressor	3.0315	15.2678	3.9074	0.8798	0.5191	0.8106	13.742
knn	K Neighbors Regressor	3.3030	18.1851	4.2643	0.8569	0.5515	0.9103	1.514
ada	AdaBoost Regressor	3.7391	21.6513	4.6530	0.8296	0.5974	0.8716	6.128
huber	Huber Regressor	9.2390	120.5888	10.9812	0.0510	0.8585	3.1370	0.937
lar	Least Angle Regression	9.3362	118.9846	10.9079	0.0636	0.8542	3.0252	0.039
br	Bayesian Ridge	9.3362	118.9846	10.9079	0.0636	0.8542	3.0252	0.060
ridge	Ridge Regression	9.3362	118.9846	10.9079	0.0636	0.8542	3.0252	0.034
lr	Linear Regression	9.3362	118.9846	10.9079	0.0636	0.8542	3.0252	0.045
en	Elastic Net	9.3581	119.0508	10.9110	0.0631	0.8567	3.0404	0.048
lasso	Lasso Regression	9.3631	119.0897	10.9128	0.0628	0.8571	3.0438	0.046
omp	Orthogonal Matching Pursuit	9.3788	120.4342	10.9742	0.0522	0.8561	3.0387	0.036
llar	Lasso Least Angle Regression	9.7363	127.0848	11.2731	-0.0001	0.8874	3.2027	0.036
par	Passive Aggressive Regressor	13.2506	292.7407	16.7125	-1.3071	1.1549	4.4420	0.205

Top 5 (MAE 기준)

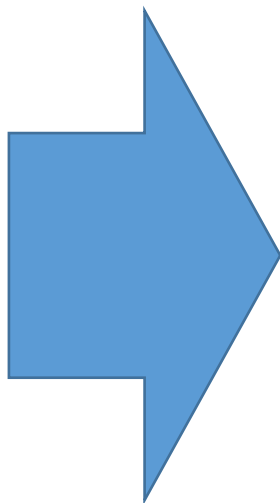
1. LightGBM
2. CatBoost Regressor
3. Gradient Boosting Regressor
4. ExtraTrees Regressor
5. Decision Tree Regressor

프로젝트 수행 결과

19년도 기온예측

하이퍼파라미터 튜닝 전 평가지표

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	3.0053	15.0805	3.8834	0.8815	0.5099	0.7934
1	3.0001	14.9380	3.8650	0.8812	0.5068	0.7756
2	3.0226	15.1874	3.8971	0.8802	0.5112	0.8180
3	3.0120	15.1324	3.8900	0.8797	0.5132	0.8138
4	3.0066	15.1261	3.8892	0.8811	0.5135	0.7906
5	3.0359	15.1797	3.8961	0.8801	0.5153	0.8141
6	3.0092	15.0429	3.8785	0.8812	0.5128	0.7929
7	3.0239	15.1762	3.8957	0.8809	0.5140	0.8082
8	3.0199	15.2055	3.8994	0.8800	0.5170	0.8335
9	3.0168	15.0725	3.8823	0.8810	0.5192	0.8062
Mean	3.0152	15.1141	3.8877	0.8807	0.5133	0.8046
SD	0.0101	0.0783	0.0101	0.0006	0.0034	0.0158

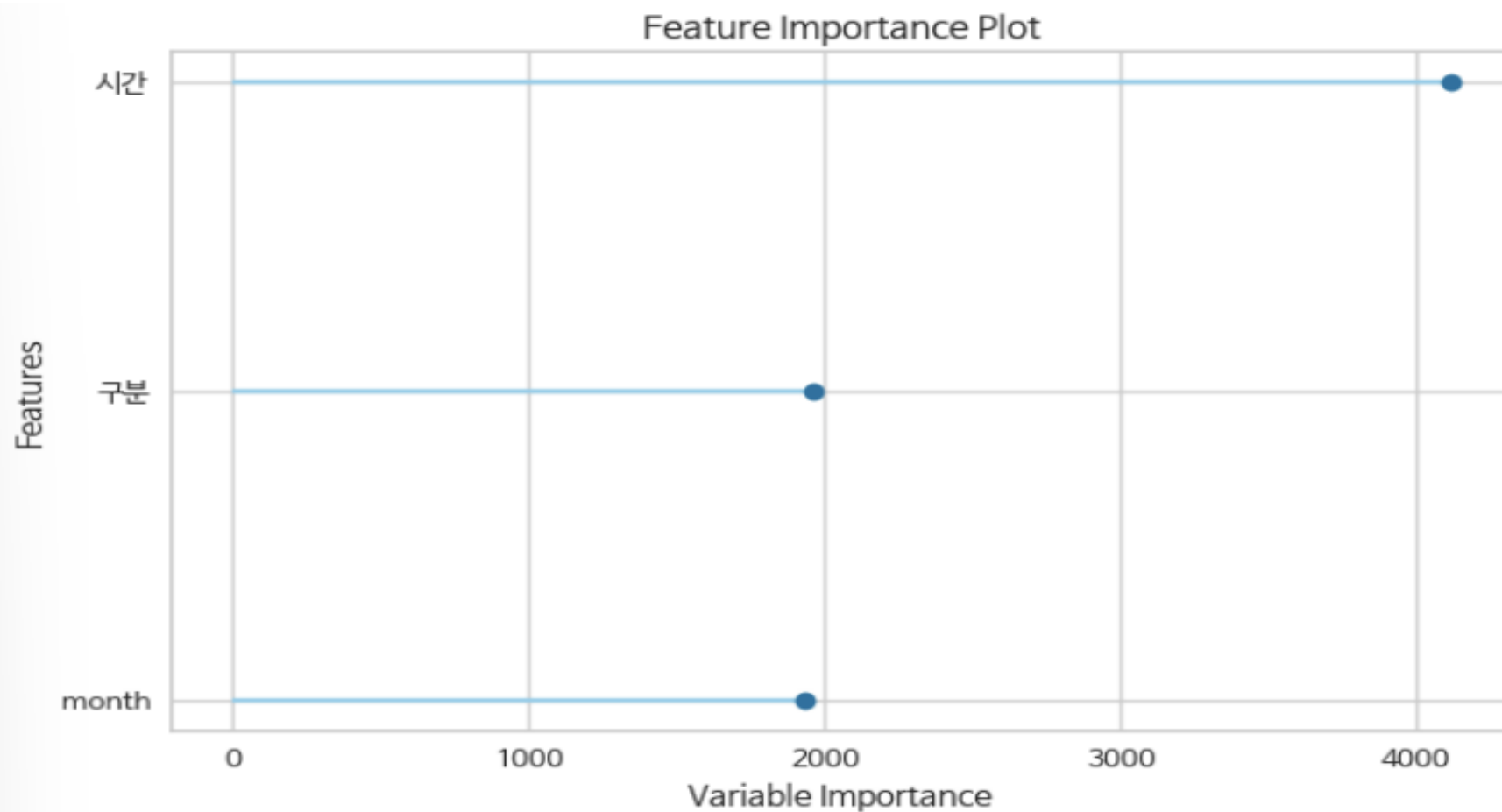


하이퍼파라미터 튜닝 후 평가지표

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	3.0166	15.1986	3.8985	0.8805	0.5121	0.7973
1	3.0120	15.0607	3.8808	0.8803	0.5104	0.7790
2	3.0331	15.2997	3.9115	0.8793	0.5143	0.8228
3	3.0221	15.2504	3.9052	0.8788	0.5169	0.8179
4	3.0201	15.2632	3.9068	0.8801	0.5170	0.7985
5	3.0488	15.3169	3.9137	0.8790	0.5177	0.8208
6	3.0208	15.1661	3.8944	0.8803	0.5154	0.7997
7	3.0356	15.2980	3.9113	0.8799	0.5172	0.8118
8	3.0327	15.3418	3.9169	0.8790	0.5215	0.8398
9	3.0291	15.1988	3.8986	0.8800	0.5216	0.8112
Mean	3.0271	15.2394	3.9038	0.8797	0.5164	0.8099
SD	0.0103	0.0805	0.0103	0.0006	0.0034	0.0161

프로젝트 수행 결과

19년도 기온예측



	month	구분	시간	Label
0	01	0	01	-3.395948
1	01	0	02	-3.818450
2	01	0	03	-3.877256
3	01	0	04	-4.202990
4	01	0	05	-4.699295
...
15115	03	6	20	7.071337
15116	03	6	21	6.520227
15117	03	6	22	5.881322
15118	03	6	23	5.566545
15119	03	6	24	5.043025

프로젝트 수행 결과

19년도 공급량예측

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.1452	0.1027	0.3204	0.9281	0.0860	0.0450	1.420
rf	Random Forest Regressor	0.1584	0.1322	0.3635	0.9075	0.0932	0.0467	50.246
knn	K Neighbors Regressor	0.1650	0.1299	0.3603	0.9091	0.0911	0.0485	1.012
gbr	Gradient Boosting Regressor	0.1675	0.1196	0.3458	0.9163	0.0903	0.0504	11.989
et	Extra Trees Regressor	0.1682	0.1557	0.3945	0.8911	0.1026	0.0485	40.906
dt	Decision Tree Regressor	0.1780	0.1782	0.4220	0.8753	0.1106	0.0503	0.653
ada	AdaBoost Regressor	0.5890	0.6250	0.7883	0.5625	0.1511	0.1155	8.688
huber	Huber Regressor	0.8161	1.1138	1.0553	0.2204	0.1762	0.1734	1.200
lr	Linear Regression	0.8413	1.0981	1.0479	0.2314	0.1745	0.1748	0.049
lar	Least Angle Regression	0.8413	1.0981	1.0479	0.2314	0.1745	0.1748	0.038
br	Bayesian Ridge	0.8413	1.0981	1.0479	0.2314	0.1745	0.1748	0.069
ridge	Ridge Regression	0.8413	1.0981	1.0479	0.2314	0.1745	0.1748	0.032
en	Elastic Net	0.8596	1.1094	1.0533	0.2235	0.1755	0.1782	0.050
lasso	Lasso Regression	0.8729	1.1328	1.0643	0.2071	0.1771	0.1807	0.048
omp	Orthogonal Matching Pursuit	0.8741	1.1429	1.0690	0.2001	0.1775	0.1806	0.039
llar	Lasso Least Angle Regression	0.9646	1.4288	1.1953	-0.0000	0.1932	0.1978	0.036
par	Passive Aggressive Regressor	1.5940	4.0956	1.9505	-1.8700	0.3238	0.2956	0.267

Top 5 (MAE 기준)

1. LightGBM
2. Random Forest Regressor
3. K.Neighbors Regressor
4. Gradient Boosting Regressor
5. Extra Trees Regressor

프로젝트 수행 결과






하이퍼파라미터 튜닝 전 평가지표

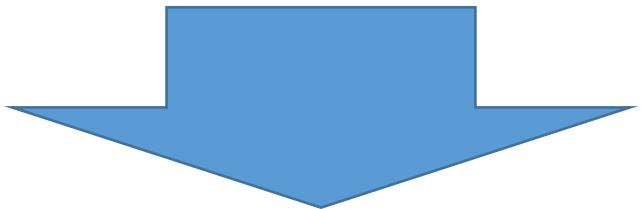
	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.1469	0.1089	0.3300	0.9246	0.0890	0.0468
1	0.1471	0.1094	0.3307	0.9240	0.0896	0.0473
2	0.1442	0.1009	0.3177	0.9293	0.0848	0.0442
3	0.1467	0.1041	0.3227	0.9270	0.0862	0.0452
4	0.1456	0.1032	0.3213	0.9270	0.0865	0.0454
5	0.1431	0.0941	0.3067	0.9329	0.0814	0.0423
6	0.1449	0.1033	0.3214	0.9278	0.0867	0.0454
7	0.1455	0.1039	0.3224	0.9281	0.0868	0.0455
8	0.1438	0.0990	0.3147	0.9308	0.0844	0.0440
9	0.1441	0.1004	0.3168	0.9296	0.0850	0.0444
Mean	0.1452	0.1027	0.3204	0.9281	0.0860	0.0450
SD	0.0013	0.0043	0.0067	0.0026	0.0022	0.0013

하이퍼파라미터 튜닝평가지표

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.1602	0.1153	0.3396	0.9201	0.0904	0.0495
1	0.1590	0.1154	0.3397	0.9198	0.0908	0.0497
2	0.1556	0.1065	0.3264	0.9254	0.0859	0.0465
3	0.1592	0.1107	0.3328	0.9223	0.0876	0.0478
4	0.1559	0.1083	0.3291	0.9235	0.0876	0.0475
5	0.1540	0.1002	0.3166	0.9286	0.0828	0.0447
6	0.1566	0.1084	0.3293	0.9242	0.0877	0.0477
7	0.1597	0.1111	0.3334	0.9232	0.0883	0.0485
8	0.1555	0.1043	0.3229	0.9271	0.0854	0.0464
9	0.1544	0.1046	0.3234	0.9266	0.0857	0.0464
Mean	0.1570	0.1085	0.3293	0.9241	0.0872	0.0475
SD	0.0022	0.0046	0.0070	0.0027	0.0023	0.0015

프로젝트 수행 결과

162	벙류2		0.22919	2	17일 전
163	BluBerry	  	0.2466	5	1분 전
164	TonyStank		0.33589	4	12일 전



47	BluBerry	  	0.11191	11	한 시간 전
----	----------	---	---------	----	--------



자체평가 의견

- 현재까지는 단순히 피쳐를 이용하였음
추후에는 단순한 피쳐만을 이용하지 않고 구간분할을 통해 새로운 학습예정
- 시간문제상 다양한 모델을 비교하지 못하여 추후에 다양한 모델을 제출하고 비교해 볼 예정
- 처음에는 모든 feature를 사용하였는데 상관관계를 보고 feature를 선택하는 것이 더 좋은 성능을 가져온 것을 확인하였다.
- 파라미터 튜닝시 임의로 설정하여 하는 것보다 AutoML을 사용하여 튜닝하는 것이 더 좋은 결과를 가져왔다.
- 데이터의 특성을 잘 살피고 정규화를 시키는 것이 중요하다.