장기 천연가스 수요예측 모델 개발
- 다중회귀와 D-Linear을 기반으로

TEAM 소대장은 실망했다

Index

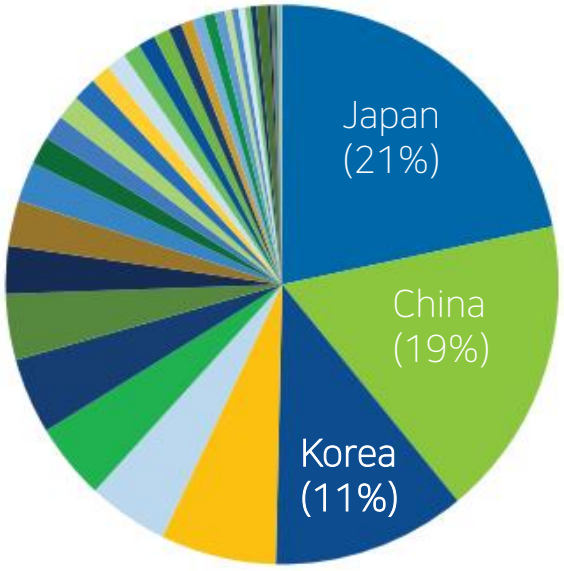
Part 1,

Introduction

과제 이해

천연가스(Natural gas)

LNG, CNG, PNG 등으로 분류되며 도시가스, 난방, 천연가스버스 등에 이용된다.



<2021년 천연가스 수입량>

<천연가스 가격 시세>

미래의 천연가스 수요에 대한 관심 필요

제14차 장기 천연가스 수급계획(2021-2034) 공고

총 천연가스 수요(기준수요)는 연평균 **1.09%** 상승 예상

도시가스용 수요는 **가정·일반용 수요 증가세가 둔화, 산업용 수요**와 LNG 벙커링, 수소차 등 **신규 수요 증가**로 연평균 1.73% 상승 기대

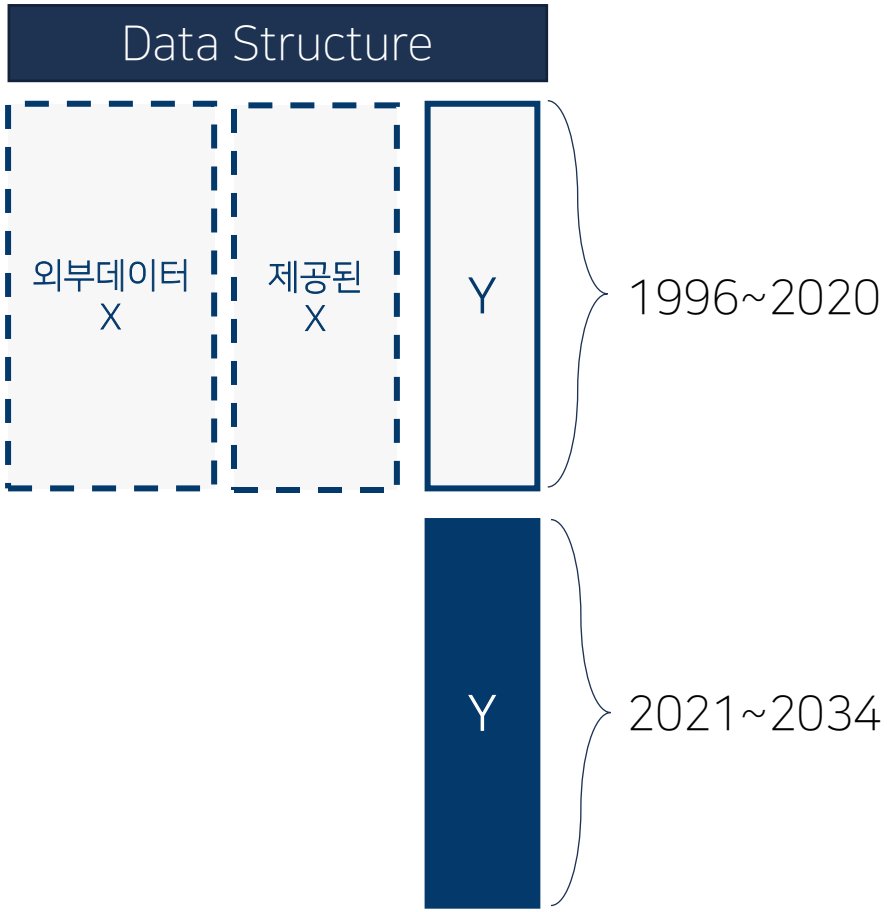
과제 이해

과제1. 장기 천연가스 수요예측 모델 개발

과거부터 기록된 민수용/산업용 천연가스 소비량을 참고하여 다양한 입력 데이터를 활용해 미래의 천연가스 소비량을 예측

X

target



다양한 외부 데이터

기존에는 미래의 수요량을 예측하기 위해서 다양한 외부 데이터를 활용

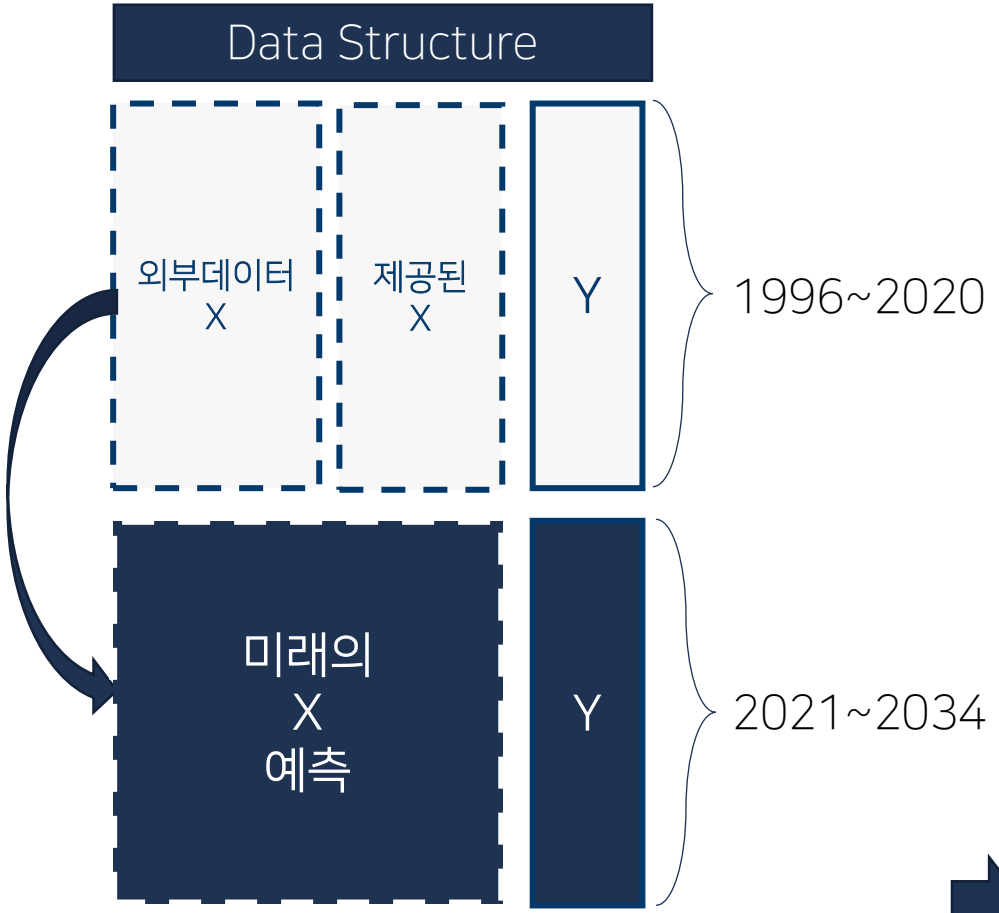
시계열 데이터

주어진 데이터들은 모두 가스사용량과 관련된 시계열 데이터

다양한 외부지표를 이용한 타겟 시계열 데이터 예측이 중요!

데이터 이해

데이터활용방안



1. 외부 데이터 수집

- Target에 대한 설명력이 높은 외부변수들을 활용하여 보다 정확한 예측 가능
- Test 시점의 외부 데이터들이 있어야 다양한 모델링 기법에 적용 및 회귀식 적합가능

2. 외부 데이터의 시계열 변환

- Target이 시계열 데이터이고, 미래 시점의 결과를 예측해야 하기 때문에 본 과제는 시계열 과제임
- 외부 데이터들로 시계열을 예측한 값들을 활용하여 보다 정확하고 효과적인 미래 수요량 예측 가능

3. 외부 데이터와 타겟의 관계를 분석

- 과거의 X를 이용하여 미래의 y를 한번에 예측하기에는 적시성과 예측력이 부족함
- 미래의 Y와 과거의 Y값의 데이터 비율이 비슷하기 때문에 과거의 X 값만을 이용한 학습은 한계가 있음

외부 데이터의 미래 값을 예측한 후, 회귀 모형을 이용하여 타깃 예측

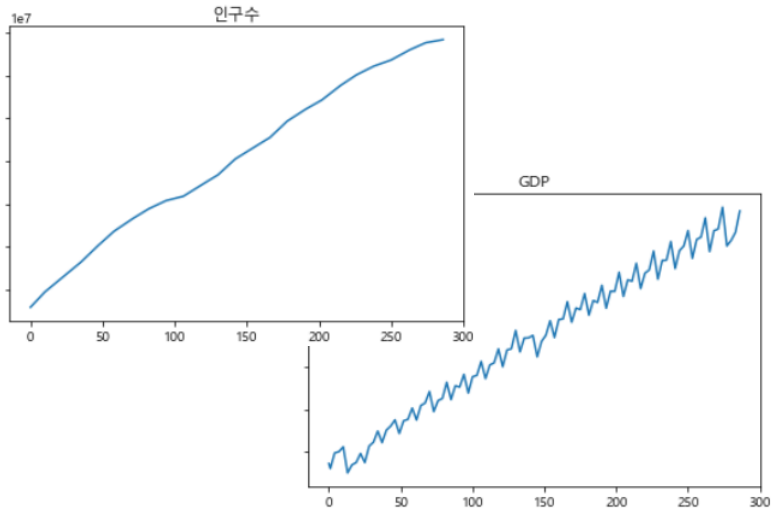
EDA, 데이터 선별

천연가스의 수요량은 높은 **계절성**을 보임

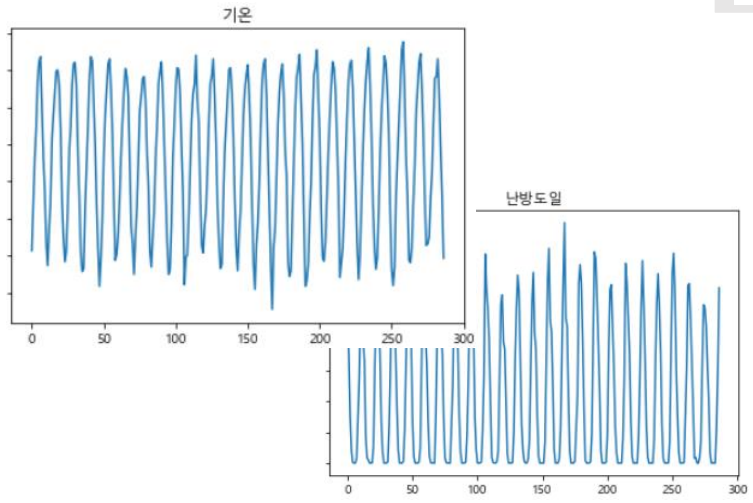
따라서 수요량의 Trend와 Seasonal을 잘 파악할 수 있는 외부변수들과 천연가스 예측에 용이하다고 판단되는 한국가스공사의 천연가스 관련 데이터와 대체효과를 보여줄 수 데이터 등을 수집

사용 외부데이터 목록

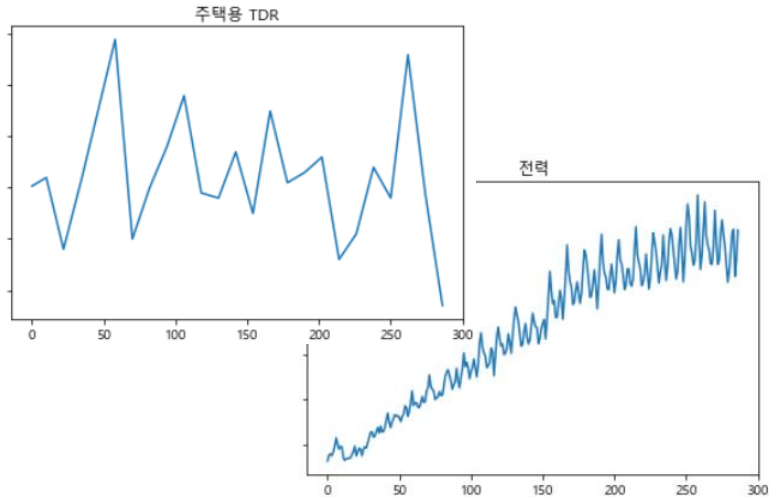
- 한국가스공사_천연가스_용도별_TDR (출처 – 한국가스공사)
- 한국가스공사_지역별_용도별_도시가스_수요_기온반응함수 (출처 : 한국가스공사)
- 수도권_도시가스수요의 기온효과 (출처 : 한국가스공사)
- 도시가스 수요 월별 기온 민감도 (출처 : 한국가스공사)
- 최종에너지 원별 소비(출처 kosis)- data
- 주요 인구지표 성비 인구성장률 연구구조 부양비 등 (출처 : kosis)
- 소비자물가지수(출처 : kosis)
- 성 및 연령별 추계인구 (출처 : kosis)
- 냉난방도일 (출처 : 기상청)
- 경제활동별 GDP 및 GNI 원계열 실질 분기 및 연간 (출처 : kosis)
- 서울지역 월별기온분포 (출처 : 기상청)



Trend를 잘 파악할 수 있는 변수
(예 : 인구수, GDP, GNI 등)



Seasonal을 잘 파악할 수 있는 변수
(예 : 기온, 냉난방도일 등)



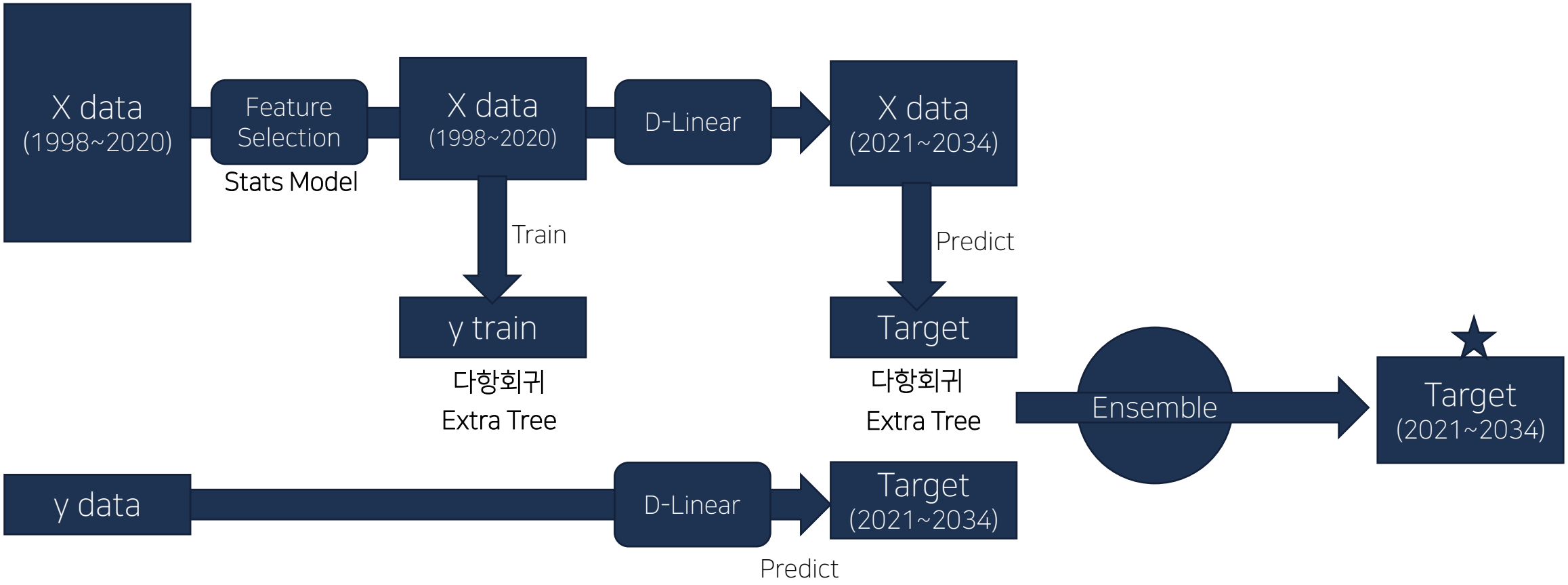
천연가스 및 에너지 관련 데이터
(예 : 주택용 TDR, 전력 등)

Part 2,

Modeling

Modeling

방법론

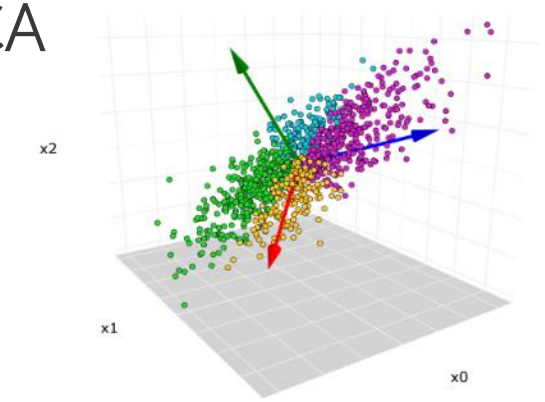


Feature Engineering

사용기법

목표 → Target과 새로운 관계를 유추할 수 있는 Feature 찾기!

PCA



- 이질적인 값들을 추출해 target과의 새로운 관계를 파악하기 위해 사용
- 기존의 원변수들끼리 높았던 상관관계를 낮출 수 있음
- N_components = 10으로 설정

Polynomial

$$a_n x^n + a_{n-1} x^{n-1} + \cdots\cdots + a_1 x + a_0 = 0$$

- 우리가 예측해야 하는 target은 시계열 형태이므로, 독립변수와 비선형적으로 관계가 있음
- 2차 Polynomial Feature를 추가

Feature Selection

- 전체 X_DataFrame을 모두 적합시켜 결과를 확인
- 변수가 70개 이상으로 상당히 많아 p-value<0.3을 기준으로 feature를 선정
- 인구통계에서 총합과 0~5세, 5~10세와 같이 공산성이 큰 feature들은 대표성을 띌 수 있는 총합과 같은 feature로 선정해서 selection해 줌
- 다시 model에 적합 후 대회 측에서 제공해준 QVA, GAS_PRICE변수를 추가

```
minsu_use_cols = ['QVA(제조업부가가치/단위:십억원)','GAS_PRICE(산업용도시가스)'
                  ,'평균기온(℃)','주택용', '일반용','계','수도권 기온반응도','cos_month', '전력', '열']

sanup_use_cols = minsu_use_cols + ['난방도일 (도일)','전기업','수도권 기온효과']
```

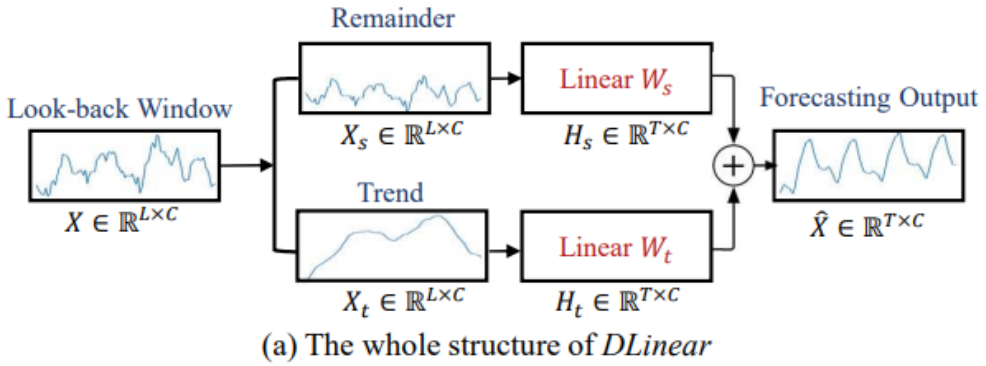
[최종 선택된 columns]

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.763e+06	2.36e+07	-0.117	0.907	-4.94e+07	4.38e+07
a0	-1.658e+05	1.04e+05	-1.602	0.111	-3.7e+05	3.85e+04
a1	2629.6859	1370.465	1.919	0.057	-75.750	5335.122
a2	-876.3813	750.285	-1.168	0.244	-2357.519	604.756
a3	-4.1651	14.274	-0.292	0.771	-32.344	24.013
a4	13.6603	22.537	0.606	0.545	-30.830	58.150
a5	-49.0690	134.075	-0.366	0.715	-313.747	215.609
a6	-63.1651	138.833	-0.455	0.650	-337.235	210.905
a7	6.53e+04	7.56e+04	0.864	0.389	-8.39e+04	2.14e+05
a8	6.525e+04	7.55e+04	0.864	0.389	-8.38e+04	2.14e+05
a9	-6.524e+04	7.55e+04	-0.864	0.389	-2.14e+05	8.39e+04
a10	9.6077	171.776	0.056	0.955	-329.495	348.710
a11	-7.1099	174.825	-0.041	0.968	-352.231	338.012
a12	-5.9454	3.567	-1.667	0.097	-12.987	1.096
a13	-2.526e+04	1.05e+04	-2.406	0.017	-4.6e+04	-4535.877
a14	8883.5846	6.75e+04	0.132	0.895	-1.24e+05	1.42e+05
a15	-6.115e+06	1.95e+07	-0.314	0.754	-4.46e+07	3.24e+07
a16	-744.2773	1355.403	-0.549	0.584	-3419.980	1931.425
a17	757.4475	1488.252	0.509	0.611	-2180.512	3695.407
a18	718.1567	296.408	2.423	0.016	133.018	1303.296

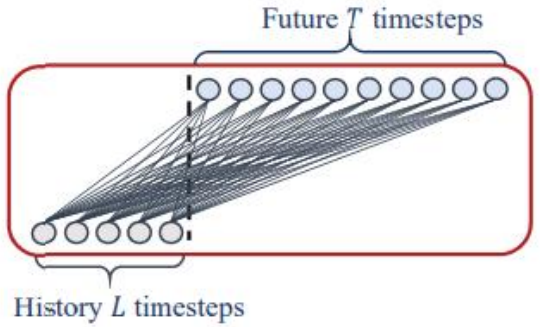
딥러닝(D-Linear)

D-Linear

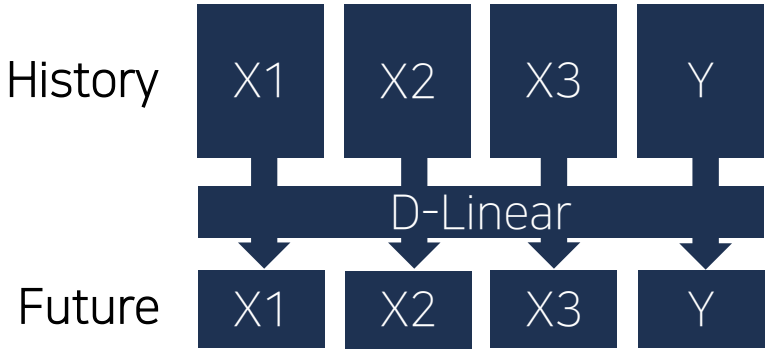
시계열 분해를 이용하여 LTSF(Long Time Series Forecasting)과제에 SOTA를 달성한 모델



(a) The whole structure of *DLinear*



(b) One Linear Layer



모델 활용

- 시계열의 단위를 hour 에서 **Month**로 바꿈
- History timesteps 와 Future timestep을 지속적으로 변환하며 **각 모델마다 최적의 t**를 찾아냄
- Future timestep보다 더 긴 time series를 예측하기 위해 **결과값을 다시 input으로 받아들이는 구조**로 변경
- 각 변수마다 새로운 모델 적용

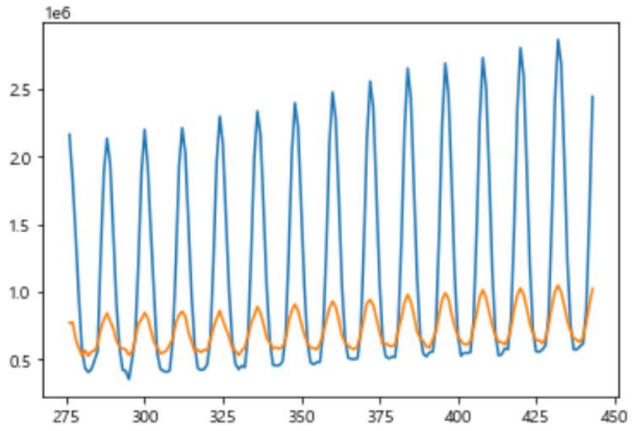
다중회귀, ML 모델, DL 모델

사용모델

statsmodels

Sample의 수는 적고, Column의 수가 많아
단순한 회귀식이 더 효과적임

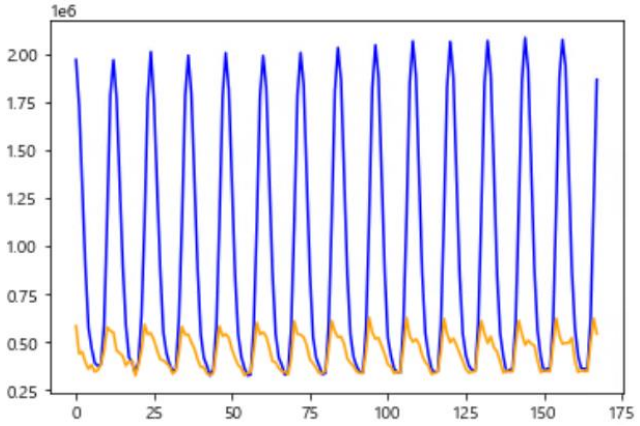
여러 개의 변수들과 target 값과의 관계를 파악

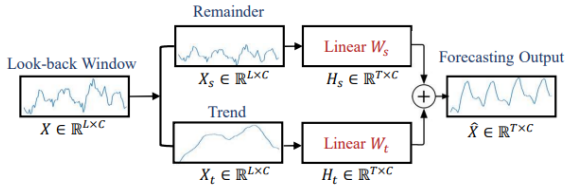


scikit learn
Extra Trees

Random Forest 보다 무작위성 ↑

많은 Tree들을 앙상블하기 때문에
일반화 성능 향상



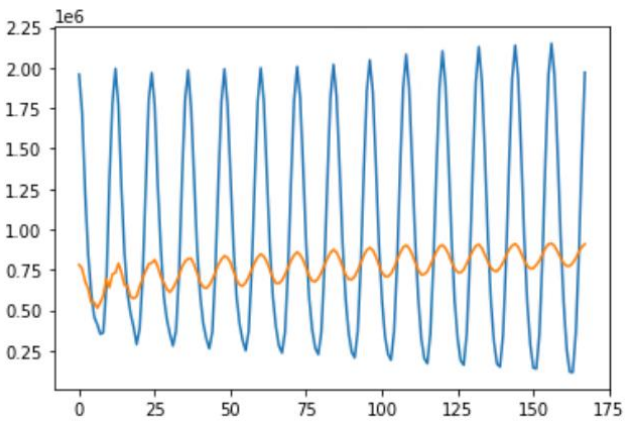


(a) The whole structure of *DLinear*

D-Linear

주어진 Train 시점의 수요량(Y_train)을 활용해
미래 시점의 수요량(Y_test) Forecasting

데이터의 주기성과 계절성을 분리하여 예측에
활용하므로 보다 정확한 값 예측 가능



Ensemble

기하평균&가중평균

Polynomial ✖ 0.7

PCA ✖ 0.3

Extra Trees

Prediction

Selected Feature

statsmodels

Prediction

Geometric Mean

민수용 : 0.2
산업용 : 0.1

민수용 : 0.8
산업용 : 0.9

D-Linear Y_pred

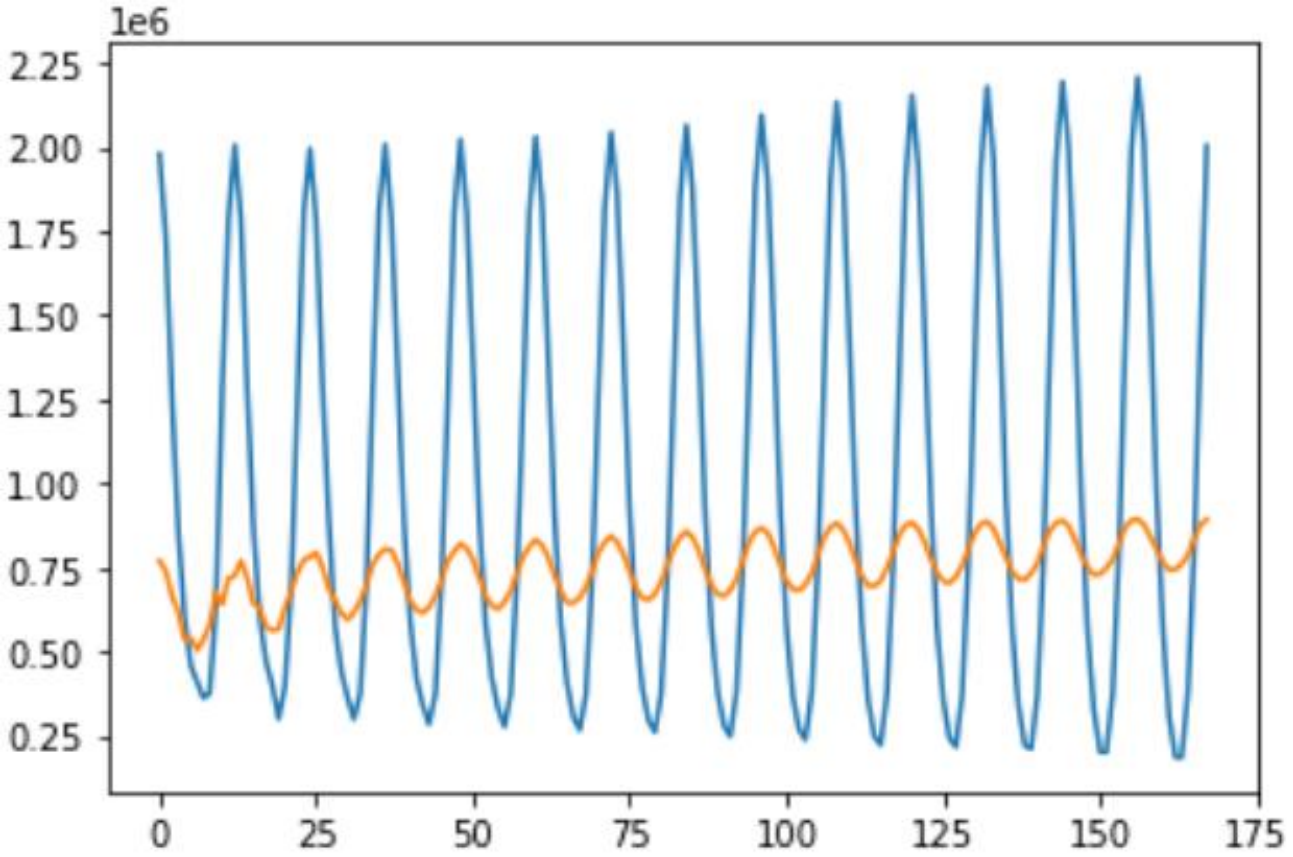
Final Prediction!

Part 3,

Conclusion

Conclusion

결과



최종 예측 결과 (파란색: 민수용, 주황색: 산업용)

민수용의 사용량은 기존과 비슷하지만, 산업용의 사용량은 꾸준히 증가하는 추세를 따른다

Conclusion

결론 및 제언

- 기존의 단순히 target만을 이용한 시계열 모델에서 벗어나 유의미한 외부변수를 활용하면 더욱 효과적인 예측이 가능함
- 필요한 외부변수를 선정할 때 다항회귀를 통해 통계적으로 유의미한 변수들만을 선택함으로써 효율성을 극대화함
- 변수의 미래를 예측하는 시계열 모델에서 변수별로 미세조정을 해준 것이 효과적이었음
- 단 하나의 방법론으로 예측하기 보다 여러가지 모델들을 동시에 활용하고 앙상블 하여 과적합을 방지할 수 있었음

Thank You

TEAM 소대장은 실망했다