

제4회 빅스타(빅데이터 · 스타트업) 경진대회

장기 천연가스 수요예측 모델 개발



KOGAS
KOREA GAS CORPORATION

Team DDoGas

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

3. 아이디어 제안

4. Appendix

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

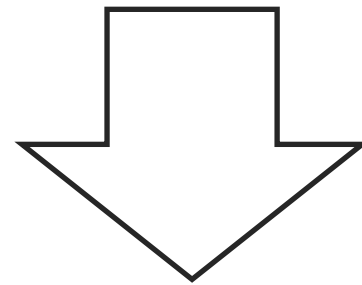
(5) 결과 요약

3. 아이디어 제안

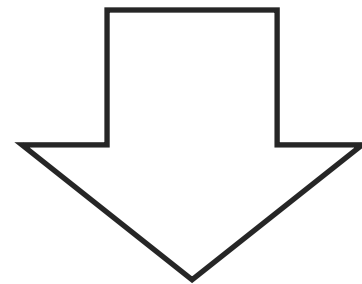
4. Appendix

1. 개발 배경

- ✓ 기후 변화와 **환경 보호**에 대한 지속적인 관심
- ✓ **시계열 데이터** 예측과 연구 진행



- ✓ ESG 탄소국경세 도입 -> **천연가스의 중요성 증대**
- ✓ 대형 AI 모델과 같은 **첨단 연구**에 대해서도 **환경 문제 대두**



- ✓ 복합 영향 인자의 고려 -> **높은 정확도의 수급관리수요 예측**
- ✓ 온실가스 배출 감소, 저탄소·친환경 **연구 기여**

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

3. 아이디어 제안

4. Appendix

2. 방법론 설명 1) 데이터 소싱

(1) 외부 데이터 추가

기후 관련 데이터

온도

전국, 서울, 부산의 최고 / 최저 / 평균 온도 및 일교차의 데이터로
직관적으로 큰 영향을 줄 것이라 예상

습도

전국, 서울, 부산의 최고 / 최저 / 평균 습도 데이터로
직관적으로 큰 영향을 줄 것이라 예상

유라시아 눈 덮임 정도

기온 저하 및 한파와 관련된 자연 현상 데이터로
한파 가능성에 대한 수요 변화 확인을 위해 추가

해빙 크기

지구 온난화의 영향으로 줄어들고 있는 해빙크기를 관측한 데이터로,
여름철 고온 및 미래 기후 트렌드 파악을 위해 추가

2. 방법론 설명 1) 데이터 소싱

(1) 외부 데이터 추가

경제 관련 데이터

GDP

OECD 예측 국가별 GDP 데이터로
기존 KOGAS 모델* 생성 시 피처로 활용하기 위해 추가

월별 수출입

한국 월별 수출입 데이터가 산업용 천연가스 수요에
영향을 미칠 것이라 예상하고 분석을 위해 추가

소비 매출

한국 소비 매출 데이터가 산업용 천연가스 수요에
영향을 미칠 것이라 예상하고 분석을 위해 추가

글로벌 천연가스 생산량

글로벌 천연가스 생산량 데이터가 산업용 천연가스 수요에
영향을 미칠 것이라 예상하고 분석을 위해 추가

2. 방법론 설명 1) 데이터 소싱

(1) 외부 데이터 추가

기타 추가 데이터

한국 인구

한국 측 국가별 GDP 데이터로
기존 KOGAS 모델* 생성 시 피쳐로 활용하기 위해 추가

음의 북극진동

북극 지역에 존재하는 찬 공기의 소용돌이를 관측한 데이터로
강한 음의 북극 진동은 국내 한파 가능성을 높이기 때문에 추가

통계청 내 크롤링 데이터

접근 가능한 외부 데이터와
천연가스 수요 증감의 관련성 분석을 위해 추가

글로벌 수심별 수온 편차 / 연평균 해양 열용량

해양과 관련된 데이터에 대한 상관성 분석을 위해 추가

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

3. 아이디어 제안

4. Appendix

2. 방법론 설명 2) 데이터 엔지니어링

(1) 월별 데이터로 변환 및 Imputation



연도별, 분기별로 되어 있는 데이터를 월별로 변환 시 각 월 값에 대한 Imputation이 필요

Stock & Flow 데이터 유형에 따라 Imputation 방식을 별도로 적용

Stock 데이터

특정 시점에 측정된 값으로
해당 시점에 남아있는 양을 의미



선형 보간
ex. 한국 인구

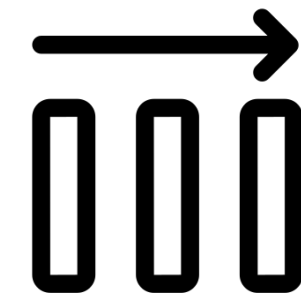


Flow 데이터

두 특정 시점 사이에 측정된 값으로
단위 기간에 생성된 양을 의미



산술 평균
ex. GDP



(2) Feature Engineering

1) QVA 2015

기존 QVA의 경우 통화가치 정보가 포함되어 있어 실질적으로 생산된 부가가치를 제대로 반영하지 못함

실제 부가가치와 천연가스 수요량과의 관계를 파악하기 위해 2015년을 기준으로 Normalizing한 변환 피처 생성



$\ast (\text{분기별 QVA 합} / \text{Commodity Price Index 평균}) \ast (\text{분기별 2015년 Commodity Price Index})$

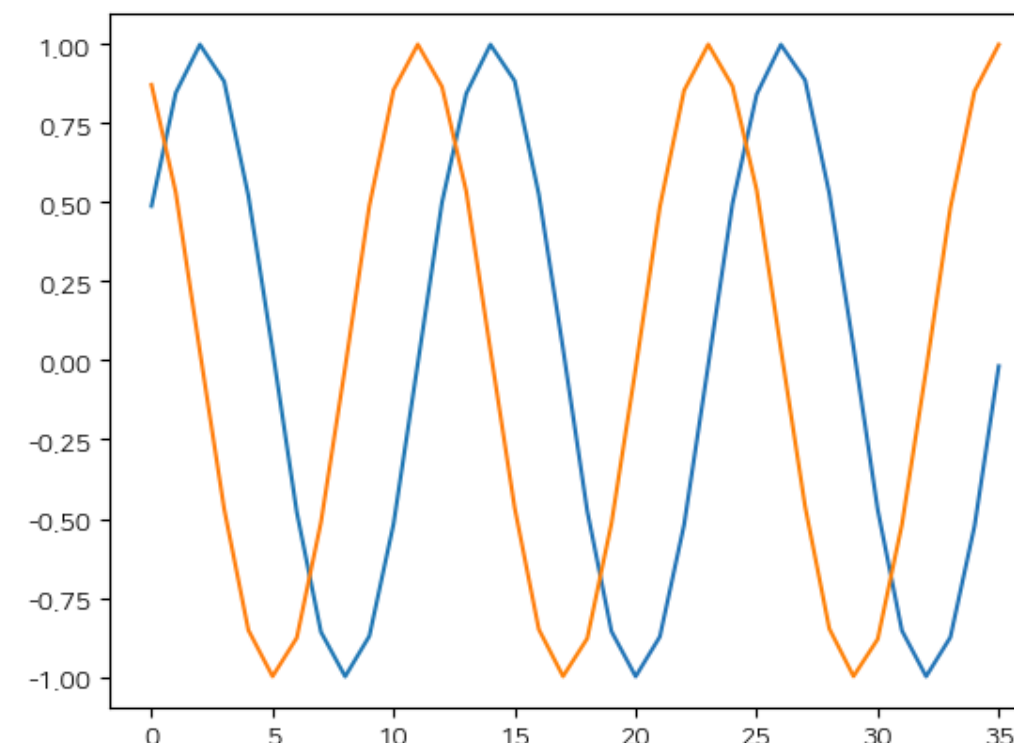
2) 계절

월을 계절 단위로 변환한 데이터 추가 (겨울 기준은 12월, 1월, 2월)

3) 시간 데이터

$\sin(x)$, $\cos(x)$, 선형 함수를 활용해 연 단위의 주기성 데이터 생성

 $= \sin(\frac{2\pi m}{12})$  $= \cos(\frac{2\pi m}{12})$



목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

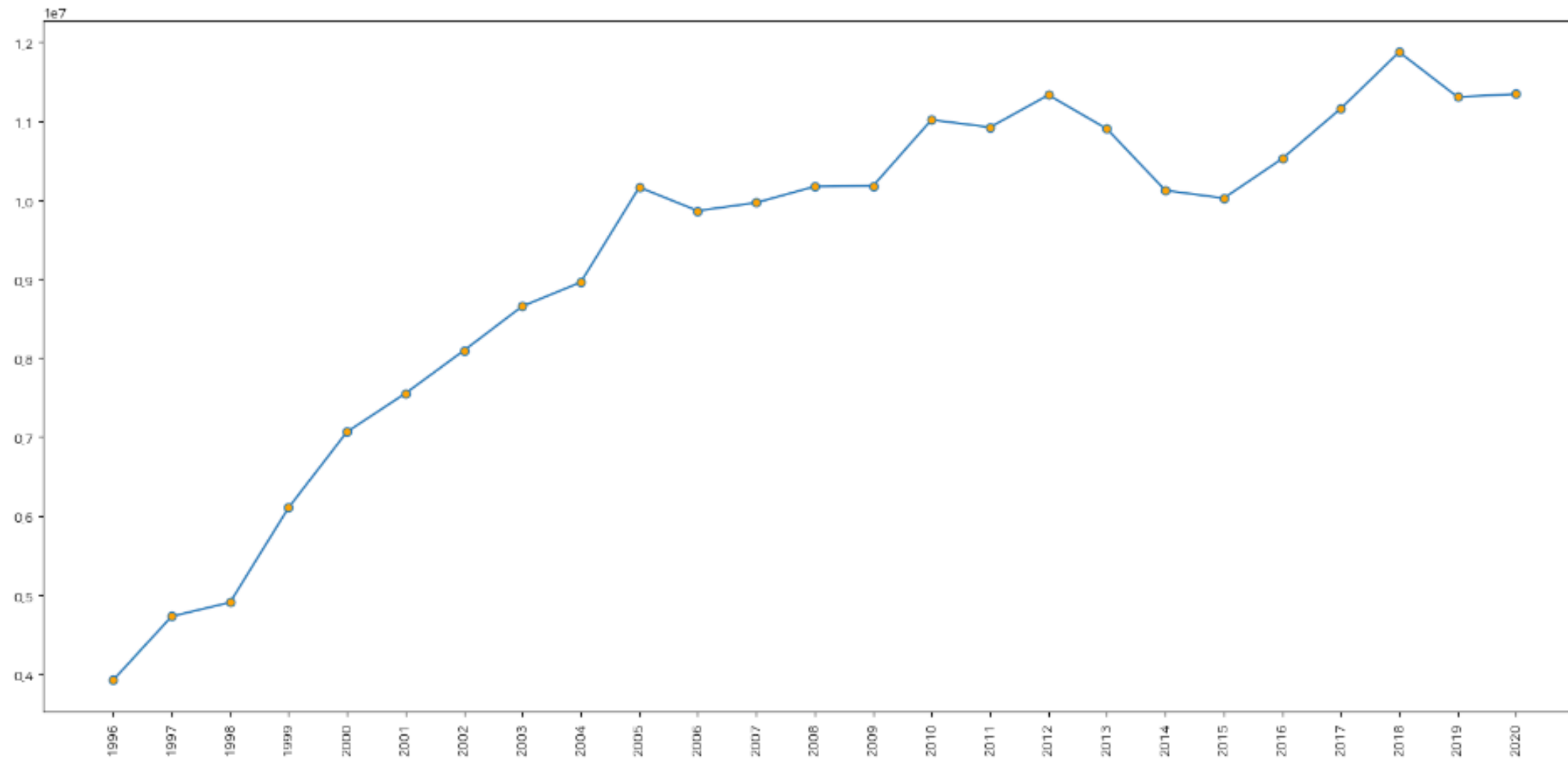
3. 아이디어 제안

4. Appendix

2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

1996년 ~ 2020년 천연가스 수요 추이



✓ 2005년까지 가파른 상승세를 보임

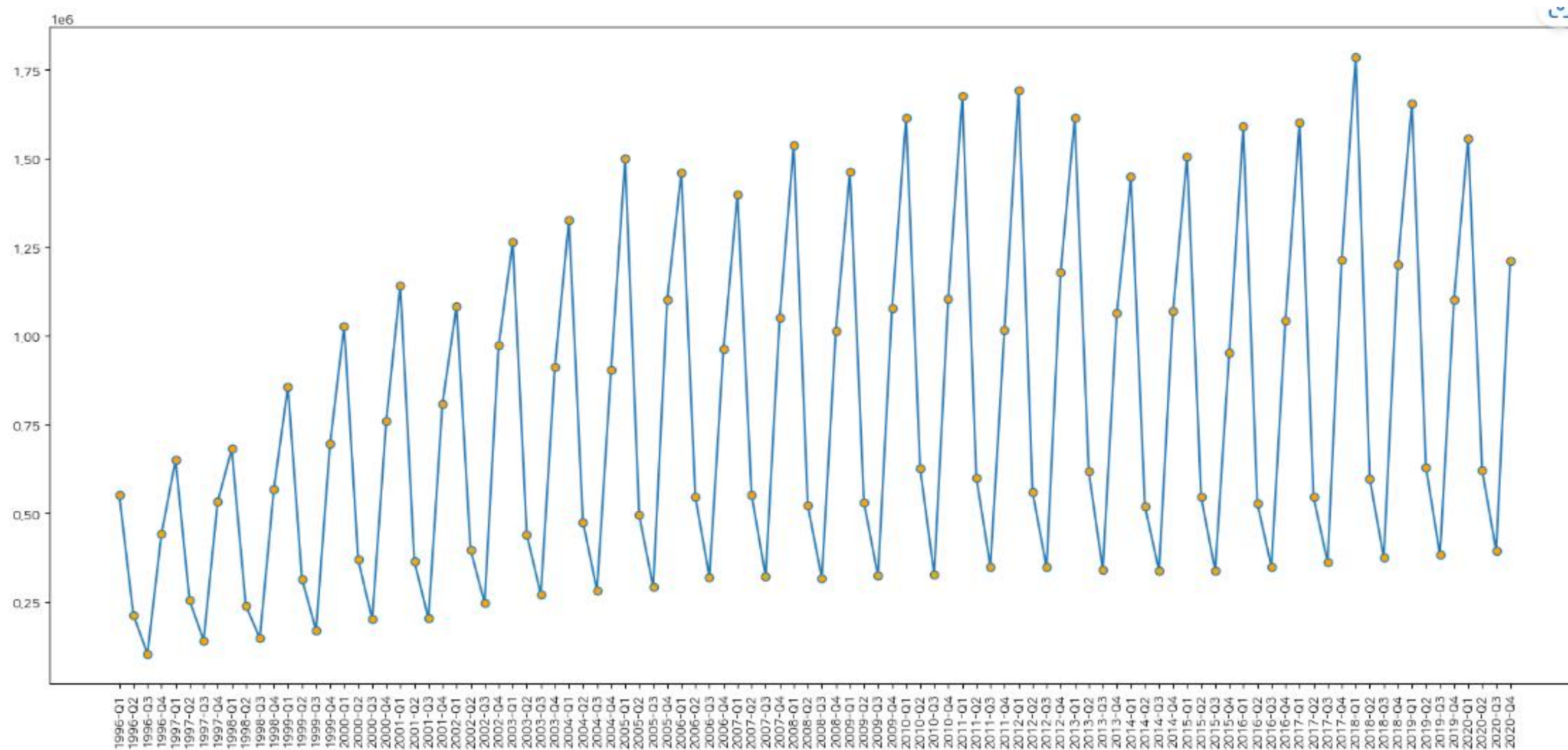
✓ 2005년 이후에는 약간 증가세에 있으나, 횡보국면으로 접어들음

✓ 1996을 기점으로 최근까지 수요가 약 200% 상승

2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

1996년 ~ 2020년 분기별 천연가스 수요 추이



- ✓ 분기별로 주기가 존재한다는 것을 확인
- ✓ 시계열 데이터에 특화된 알고리즘이 예측을 잘 할 것이라고 예상 (ex. ARIMA)
- ✓ 2005년 이후 변화폭이 적기 때문에 Rule-Based 알고리즘을 활용한 방법도 좋게 작용할 것이라 예상

2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

초단기(T+1) 민수용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
연중 월 cosine 변환	0.8658
민수용 천연가스 수요량	0.8380
서울 최고기온	-0.7848
서울 평균기온	-0.7768
전국평균 최고기온	-0.7744

유사 Feature 제거*



Feature	상관계수
연중 월 cosine 변환	0.8658
민수용 천연가스 수요량	0.8380
서울 최고기온	-0.7848
유라시아 눈 덮임 면적	0.7421
부산 평균습도	-0.7343

* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

단기(T+24) 민수용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
민수용 천연가스 수요량	0.9733
전국평균 최고기온	-0.9143
서울 최고기온	-0.9142
서울 평균기온	-0.9092
부산 최고기온	-0.9077

유사 Feature 제거*



Feature	상관계수
민수용 천연가스 수요량	0.9733
전국평균 최고기온	-0.9143
유라시아 눈 덮임 면적	0.9018
부산 평균습도	-0.8171
연중 월 cosine 변환	-0.7333

* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

(1) 민수용 천연가스 수요 데이터

중기(T+120) 민수용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
민수용 천연가스 수요량	0.7170
산업용 천연가스 수요량	0.7047
국내 해수면 높이	0.6311
한국 인구	0.5810
OECD 천연가스 생산량	0.5704

유사 Feature 제거*



Feature	상관계수
전국평균 최고기온	-0.9593
유라시아 눈 덮임 면적	0.9476
민수용 천연가스 수요량	0.8938
부산 평균습도	-0.8549
연중 월 cosine 변환	0.7594


* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

(1) 민수용 천연가스 수요 데이터

장기(T+168) 민수용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
전국평균 최고기온	-0.9563
서울 최고기온	-0.9549
부산 최고기온	-0.9498
서울 평균기온	-0.9490
전국평균 평균기온	-0.9487

유사 Feature 제거*



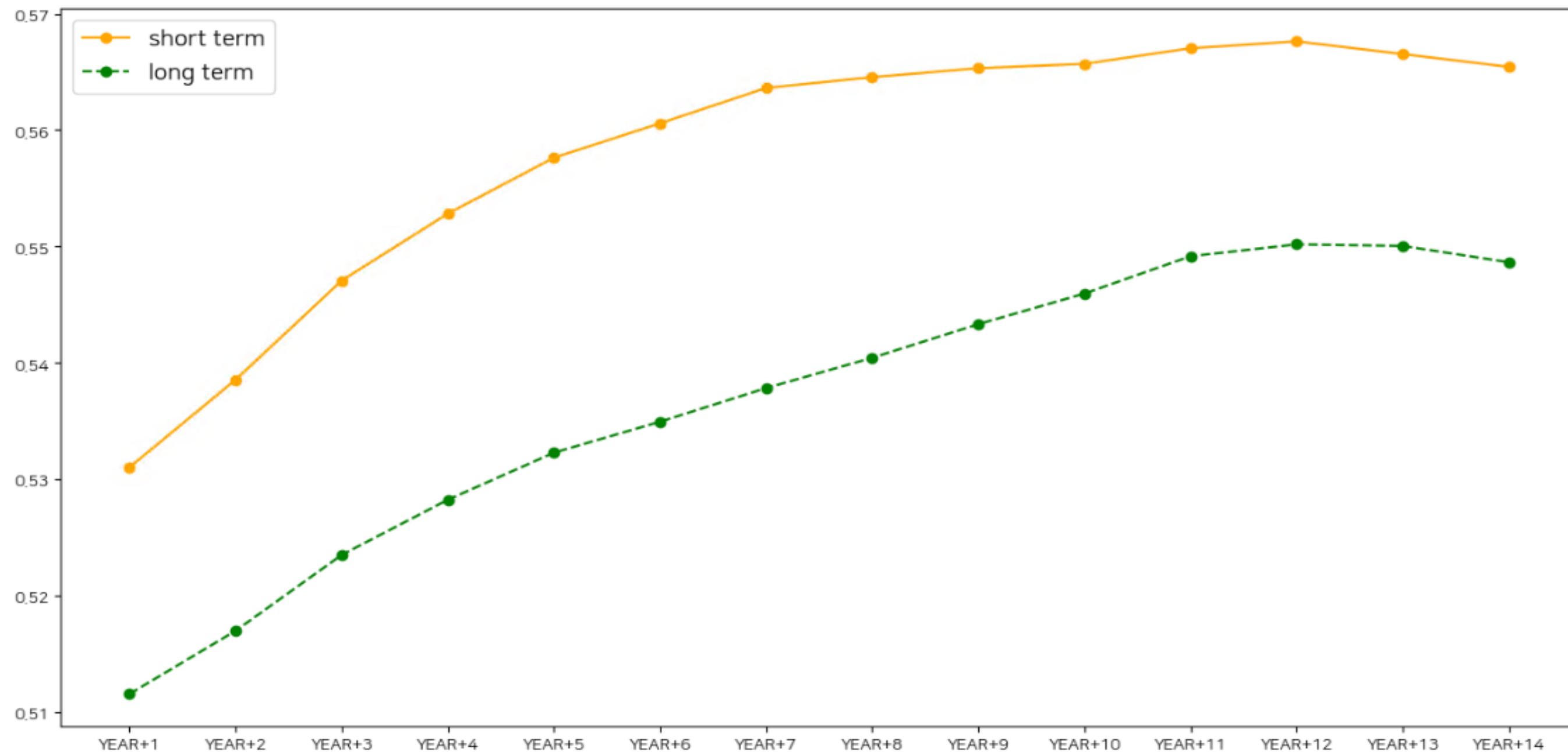
Feature	상관계수
전국평균 최고기온	-0.9563
유라시아 눈 덮임 면적	0.9423
부산 평균습도	-0.8830
민수용 천연가스 수요량	0.8659
산업용 대비 민수용 천연가스 수요량 비중	0.7757

* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

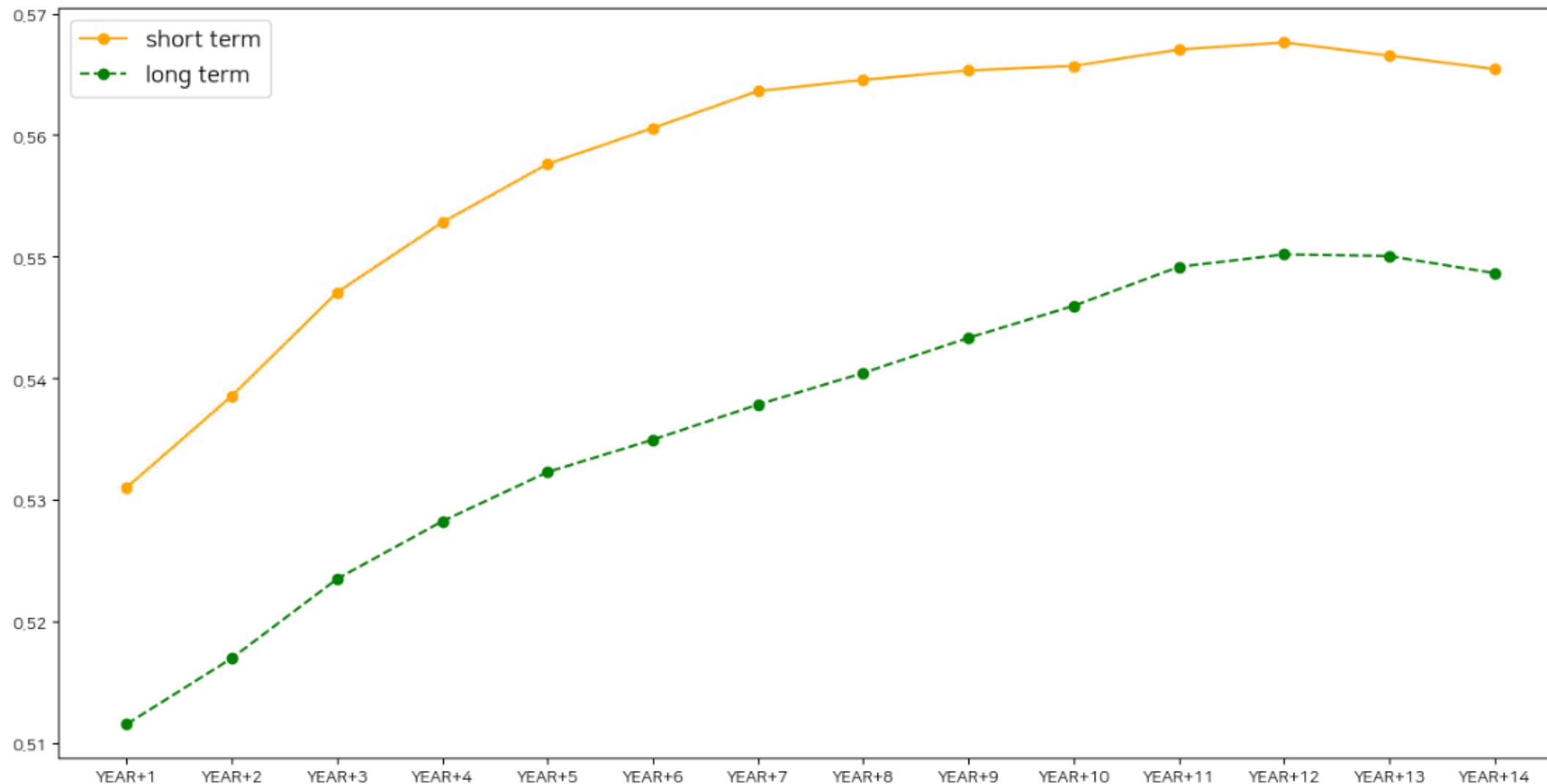
연도별 단기(T+24) & 장기(T+168) Top5 Feature 평균 상관관계



2. 방법론 설명 3) EDA

(1) 민수용 천연가스 수요 데이터

연도별 단기(T+24) & 장기(T+168) Top5 Feature 평균 상관관계

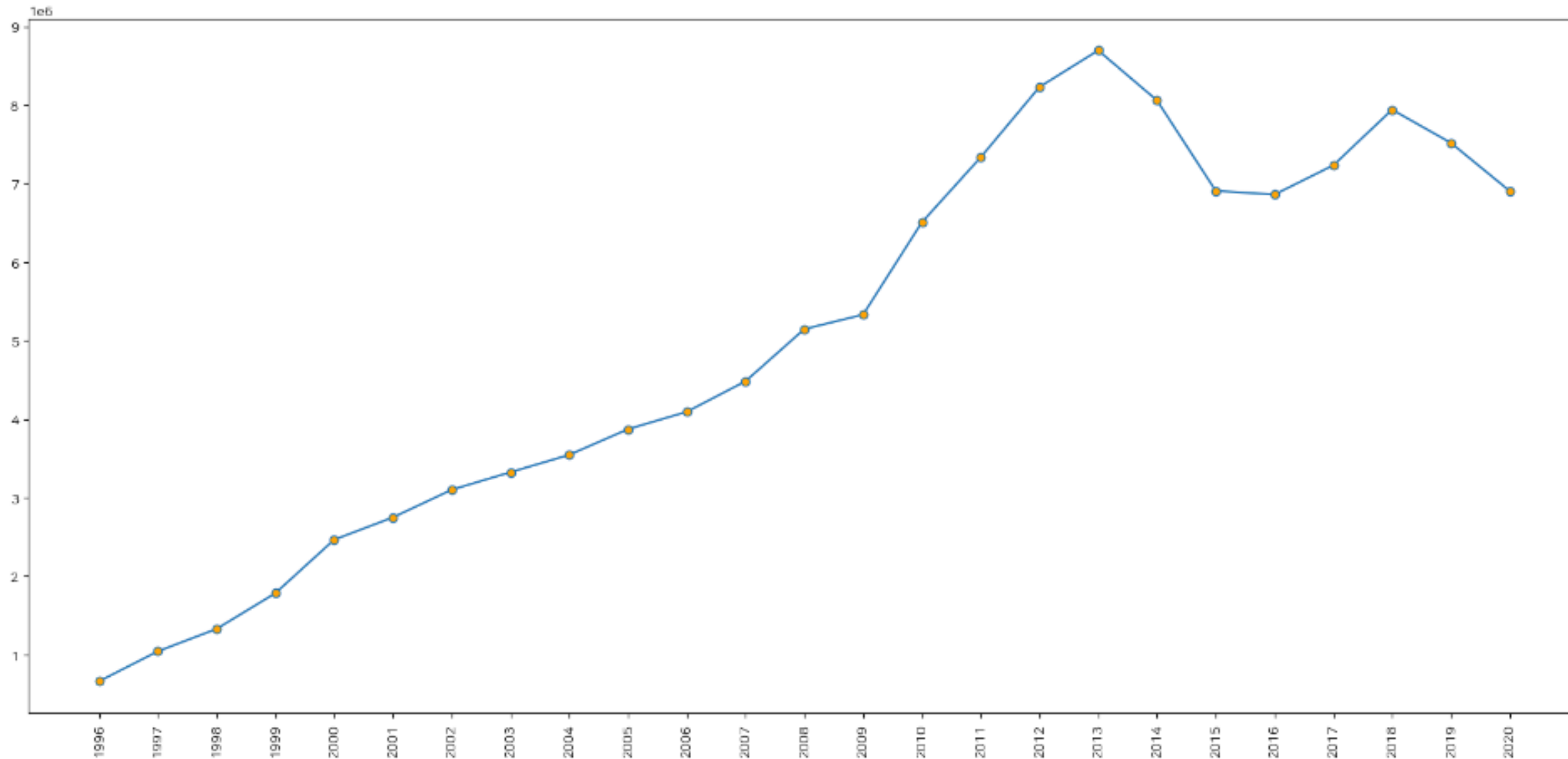


- ✓ 단기적 상관성이 높은 feature group과 장기적 상관성이 높은 feature group과의 상관관계의 차이가 그리 크지 않음
-> 장단기 예측 기간에 상관없이 상관성이 고름
- ✓ 기후 관련 feature의 상관성이 다른 feature에 비해 상대적으로 높았음
- ✓ 온도의 경우 최고기온이, 습도의 경우 평균습도가 미세하게 상관성이 높았음

2. 방법론 설명 3) EDA

(2) 산업용 천연가스 수요 데이터

1996년 ~ 2020년 천연가스 수요 추이

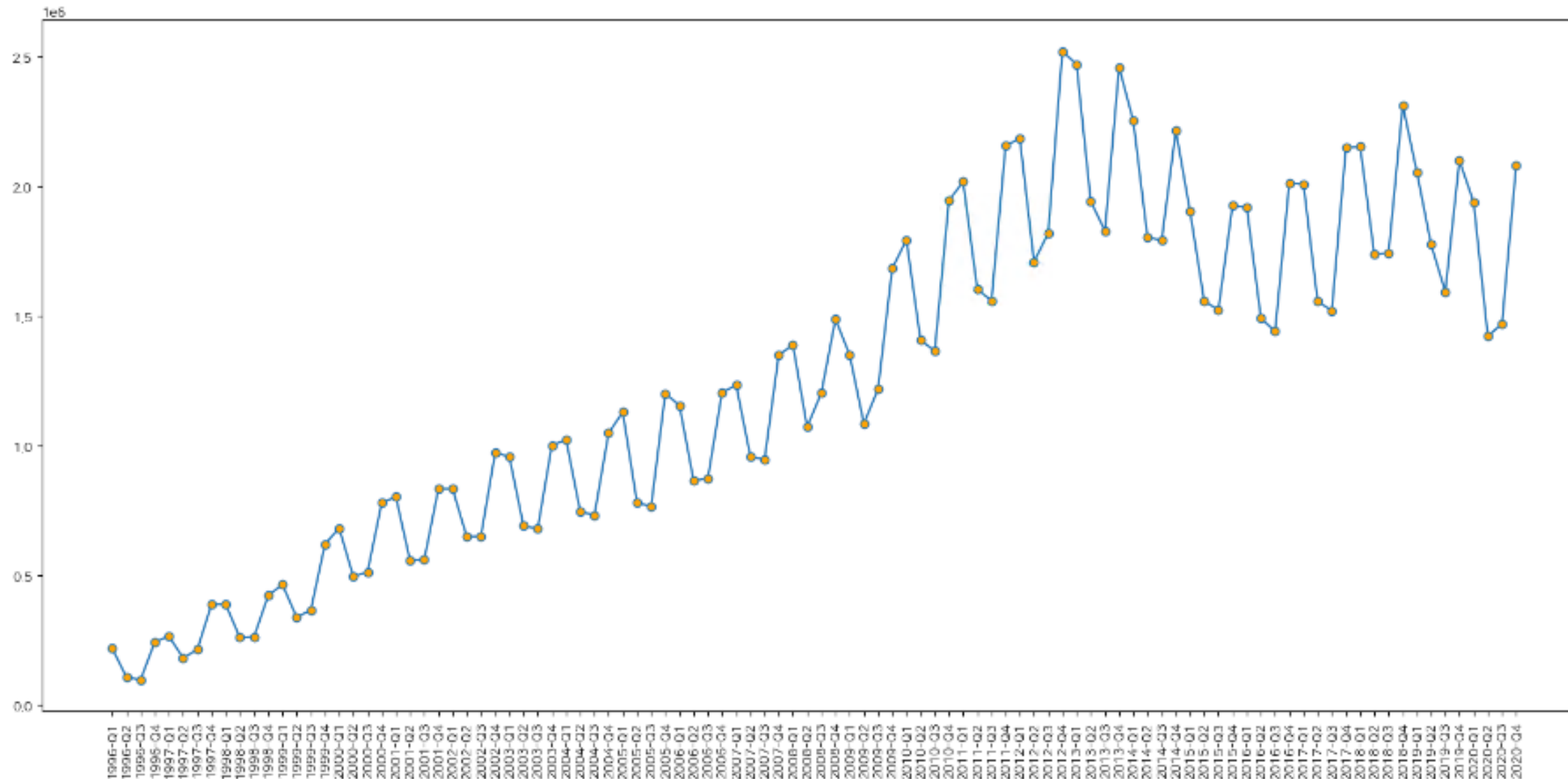


- ✓ 2009년까지 점진적 상승세를 보인다, 2010년부터 2013년까지 가파른 상승세를 보임
- ✓ 2015년, 2016년 저점을 보인 후 회복하는 모습을 보여줌
- ✓ 1996을 기점으로 최근까지 수요가 약 700% 상승
-> 민수용 수요에 비해(200%) 큰 상승폭을 보임

2. 방법론 설명 3) EDA

(2) 산업용 천연가스 수요 데이터

1996년 ~ 2020년 분기별 천연가스 수요 추이




- ✓ 일정한 주기가 존재하는 것으로 보이나
트렌드는 존재한다고 하기 어려움
→ 특정 기간에 대해서는 상승 트렌드를 보임
- ✓ 1분기, 4분기에 수요가 높아지는 것으로 보아
기후나 계절에 영향을 일부 받는 것으로 보임
- ✓ 민수용 천연가스 수요보다 **경제적요인, 지정학적
요인 등이 상대적으로 크게 작용**하기에 높은
복잡도를 가진 모델이 적합할 것으로 판단
→ ML/DL 모델을 통한 접근 시도

(2) 산업용 천연가스 수요 데이터

초단기(T+1) 산업용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
산업용 천연가스 수요량	0.9621
수출액	0.9044
QVA	0.8975
Non-OECD 아시아 천연가스 생산량	0.8957
Non-OECD 미국 생산량	0.8908

유사 Feature 제거*



Feature	상관계수
산업용 천연가스 수요량	0.9621
수출액	0.9044
QVA	0.8975
Non-OECD 아시아 천연가스 생산량	0.8957
한국 인구	0.8848


* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

(2) 산업용 천연가스 수요 데이터

단기(T+24) 산업용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
산업용 천연가스 수요량	0.9275
Non-OECD 아시아 천연가스 생산량	0.8757
Non-OECD 미국 천연가스 생산량	0.8566
QVA	0.8528
Non-OECD 중동 천연가스 생산량	0.8442

유사 Feature 제거*



Feature	상관계수
산업용 천연가스 수요량	0.9275
Non-OECD 아시아 천연가스 생산량	0.8757
QVA	0.8528
한국 인구	0.8439
수출	0.8368


* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

(2) 산업용 천연가스 수요 데이터

중기(T+120) 산업용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
민수용 천연가스 수요량	0.7170
산업용 천연가스 수요량	0.7047
국내 해수면 높이	0.6311
한국 인구	0.5810
OECD 천연가스 생산량	0.5704

유사 Feature 제거*



Feature	상관계수
민수용 천연가스 수요량	0.7170
산업용 천연가스 수요량	0.7047
한국 인구	0.5810
OECD 천연가스 생산량	0.5704
지표면 5M 지점 온도	0,5650

* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

(2) 산업용 천연가스 수요 데이터

장기(T+168) 산업용 천연가스 수요에 대한 상관관계 Top5

Feature	상관계수
연중 월 cosine 변환	0.7497
서울 최고기온	-0.7165
전국평균 최고기온	-0.7122
서울 평균기온	-0.7121
서울 최저기온	-0.7094

유사 Feature 제거*



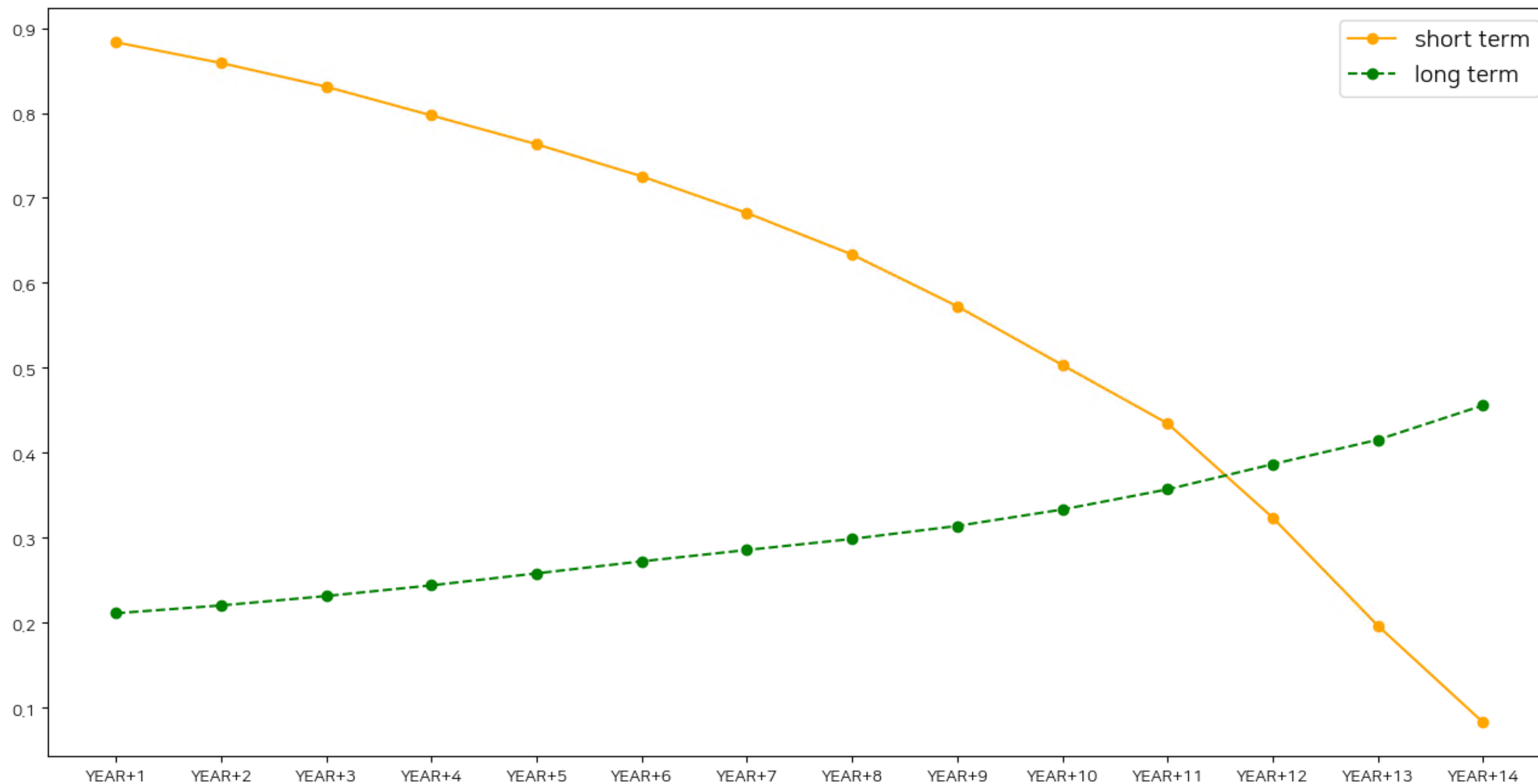
Feature	상관계수
연중 월 cosine 변환	0.7497
서울 최고기온	-0.7165
유라시아 눈 덮임 면적	0.6977
부산 평균습도	-0.6738
민수용 천연가스 수요량	0.6401

* 같은 카테고리 내 있는 feature 중 랭크가 낮은 것을 제거

2. 방법론 설명 3) EDA

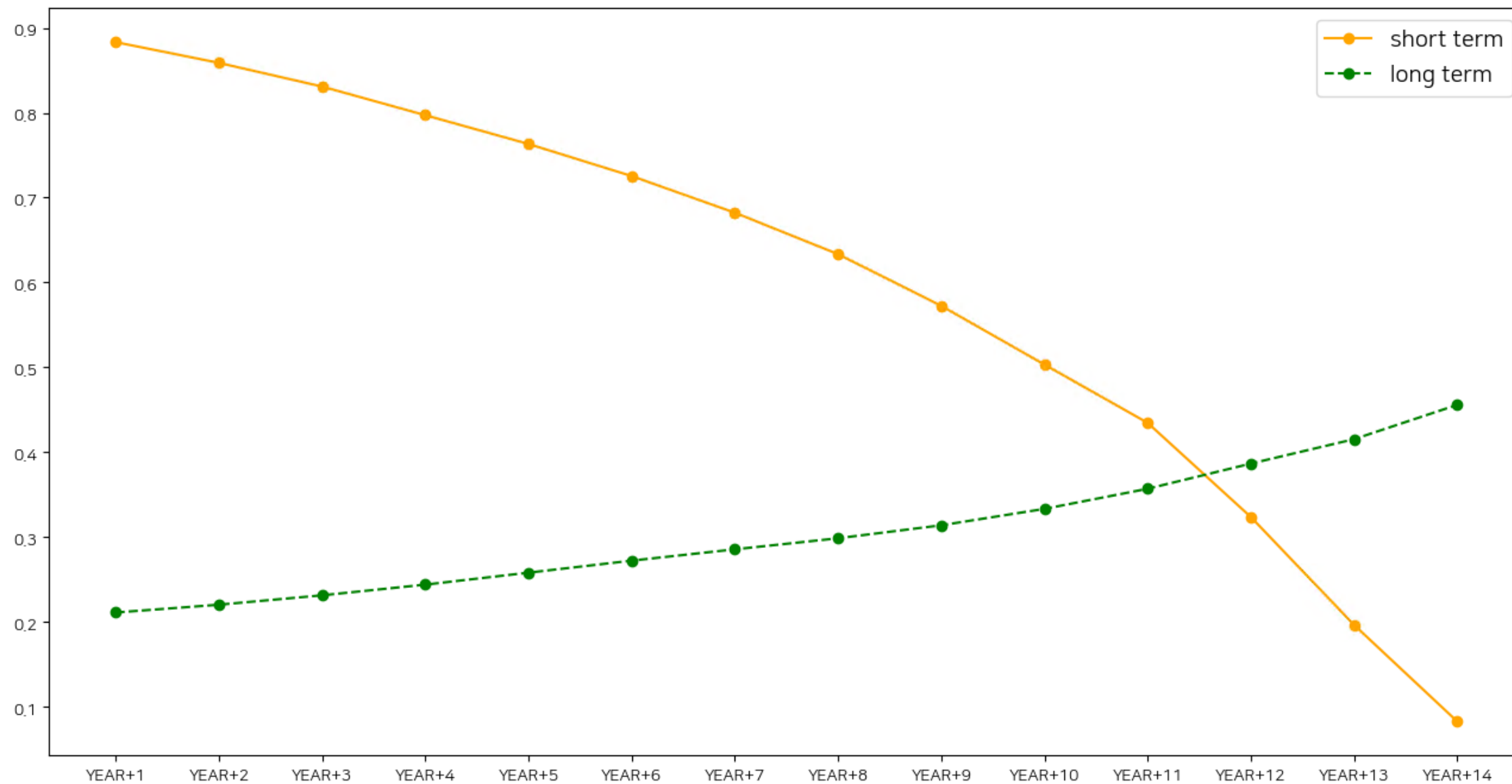
(2) 산업용 천연가스 수요 데이터

연도별 단기(T+24) & 장기(T+168) Top5 Feature 평균 상관관계



(2) 산업용 천연가스 수요 데이터

연도별 단기(T+24) & 장기(T+168) Top5 Feature 평균 상관관계



- ✓ 단기적 상관성이 높은 feature group과 장기적 상관성이 높은 feature group과의 상관관계의 크게 나타남
-> YEAR+12 부터 장단기 상관관계 수치가 역전
- ✓ 단기적일수록 경제적 요인과 상관성을 많이 띄고, 장기적일수록 기후적 요인과 상관성을 많이 띄움
-> 기후적 요인은 기간에 상관없이 비교적 상관성이 고름
- ✓ 외부데이터로 추가한 **한국 월별 수출액**과 **Non-OECD 아시아 천연가스 생산량(중국 포함)** 데이터가 단기적으로 큰 상관성을 보임

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

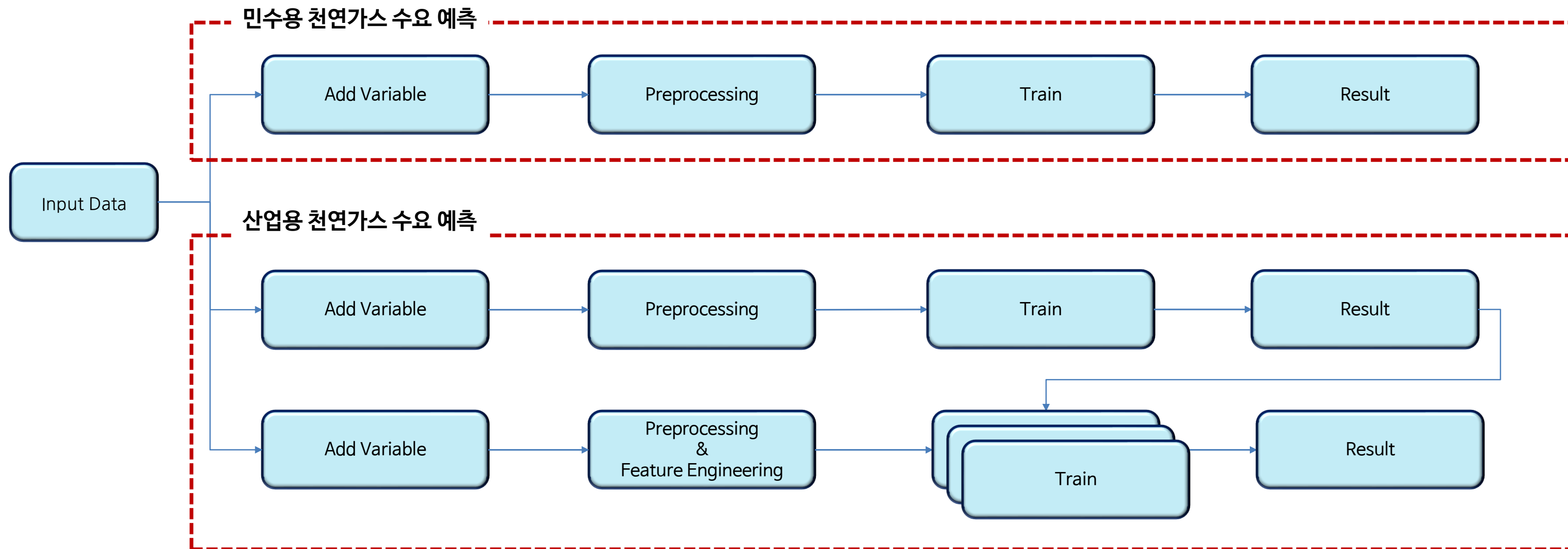
3. 아이디어 제안

4. Appendix

2. 방법론 설명 4) 모델링

(1) 아키텍처 개요

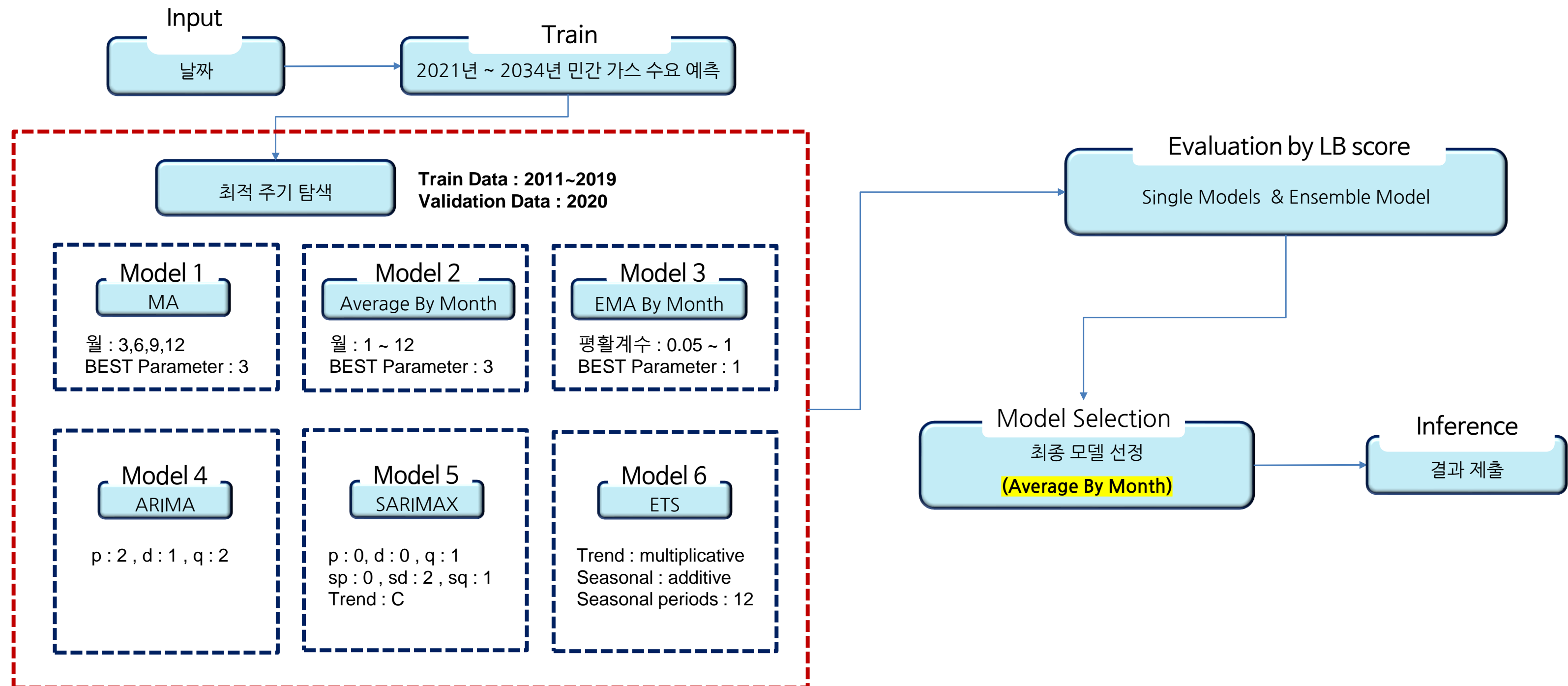
전체 모델 구조



2. 방법론 설명 4) 모델링

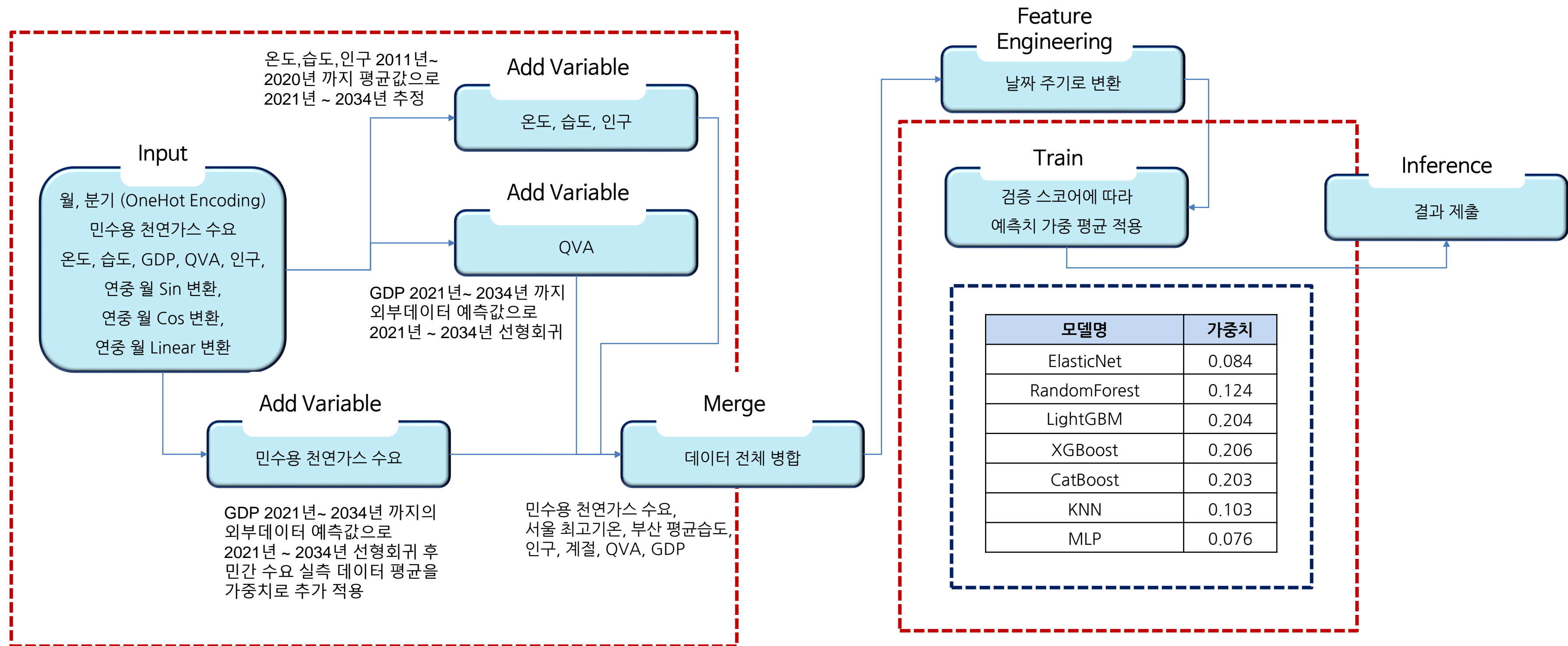
(1) 아키텍처 개요

민수용 천연가스 수요 예측 모델



(1) 아키텍처 개요

산업용 천연가스 수요 예측 모델



목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

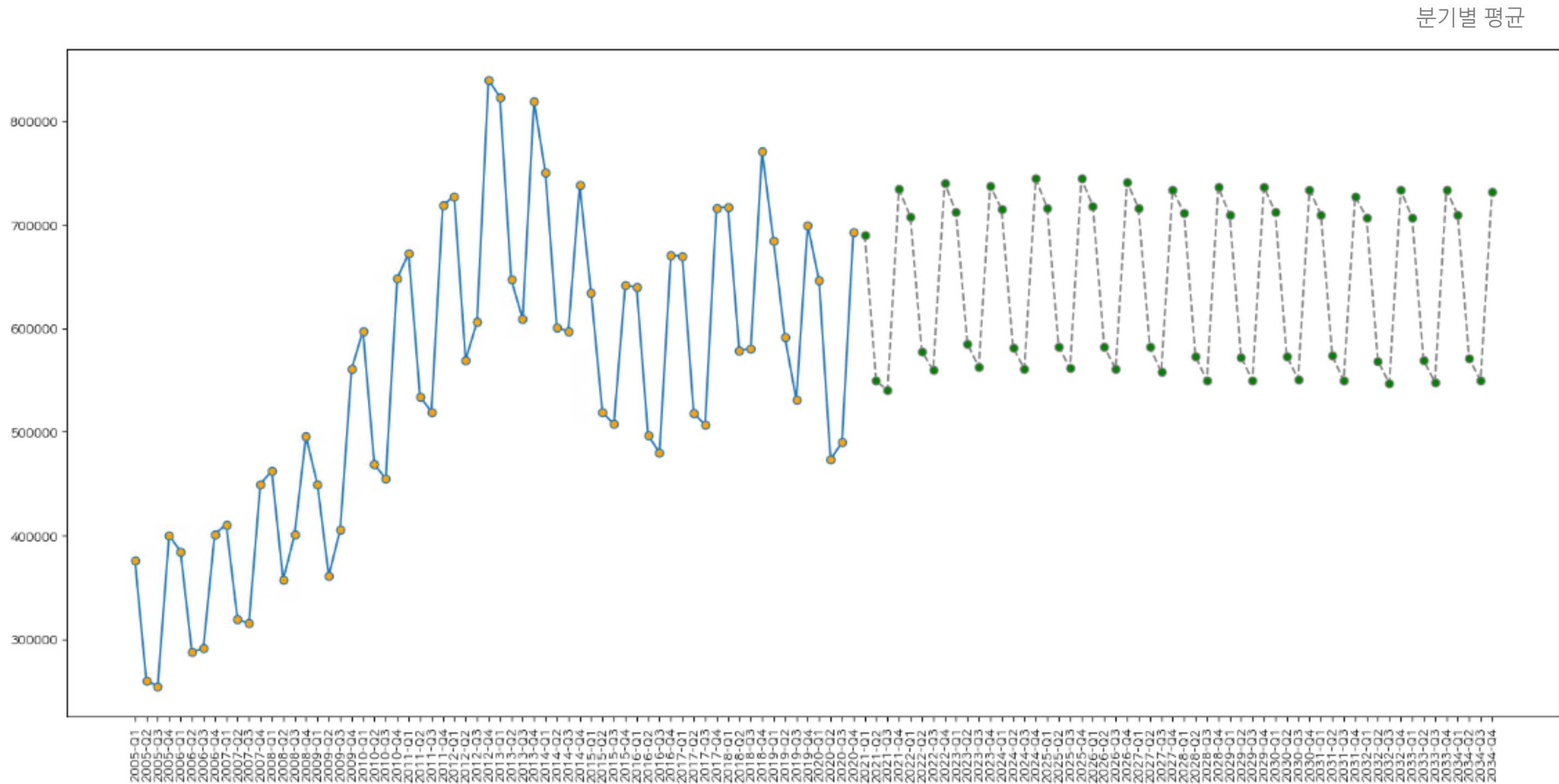
(5) 결과 요약

3. 아이디어 제안

4. Appendix

2. 방법론 설명 5) 결과 요약

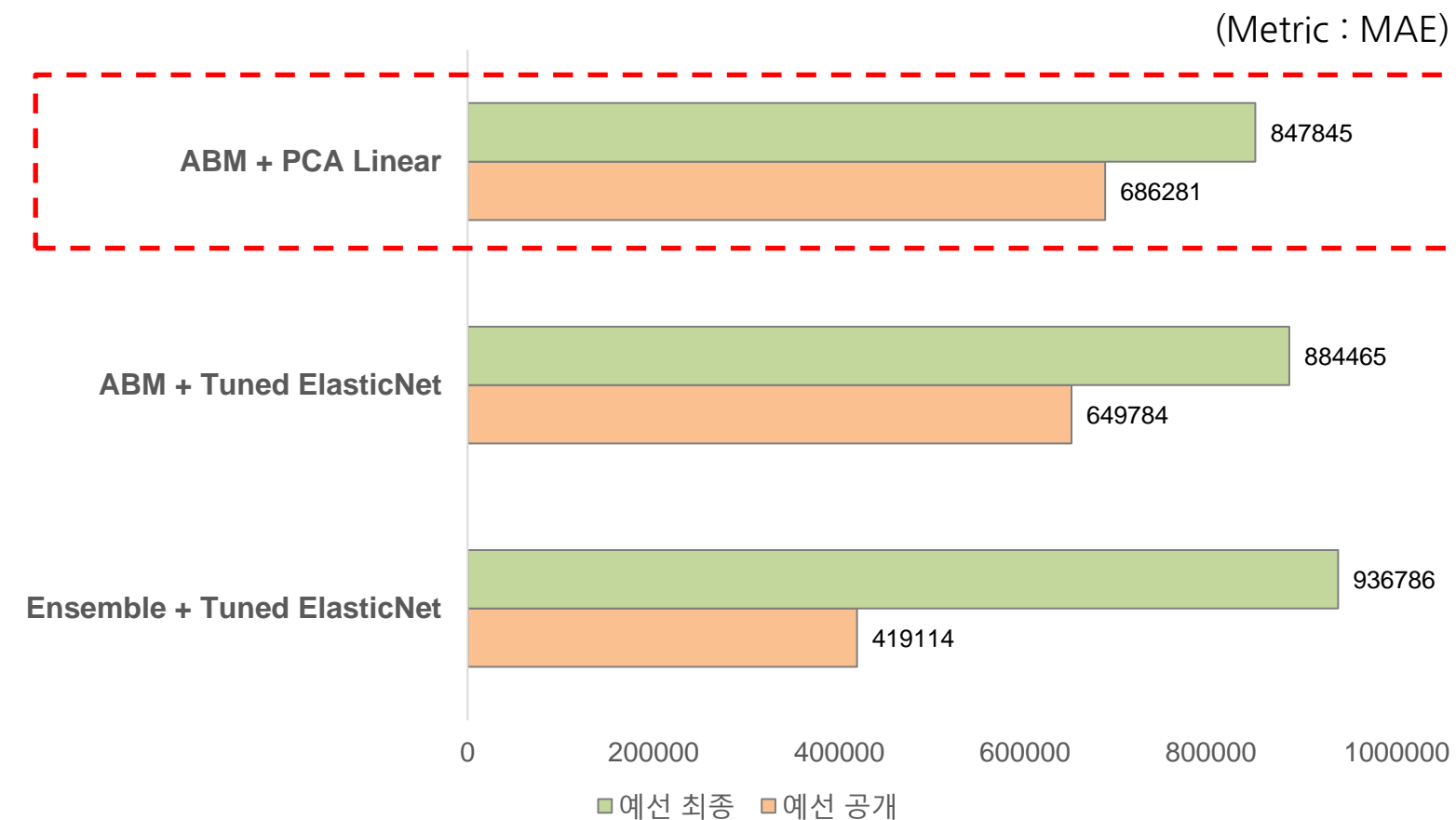
(1) 예측 결과 시각화



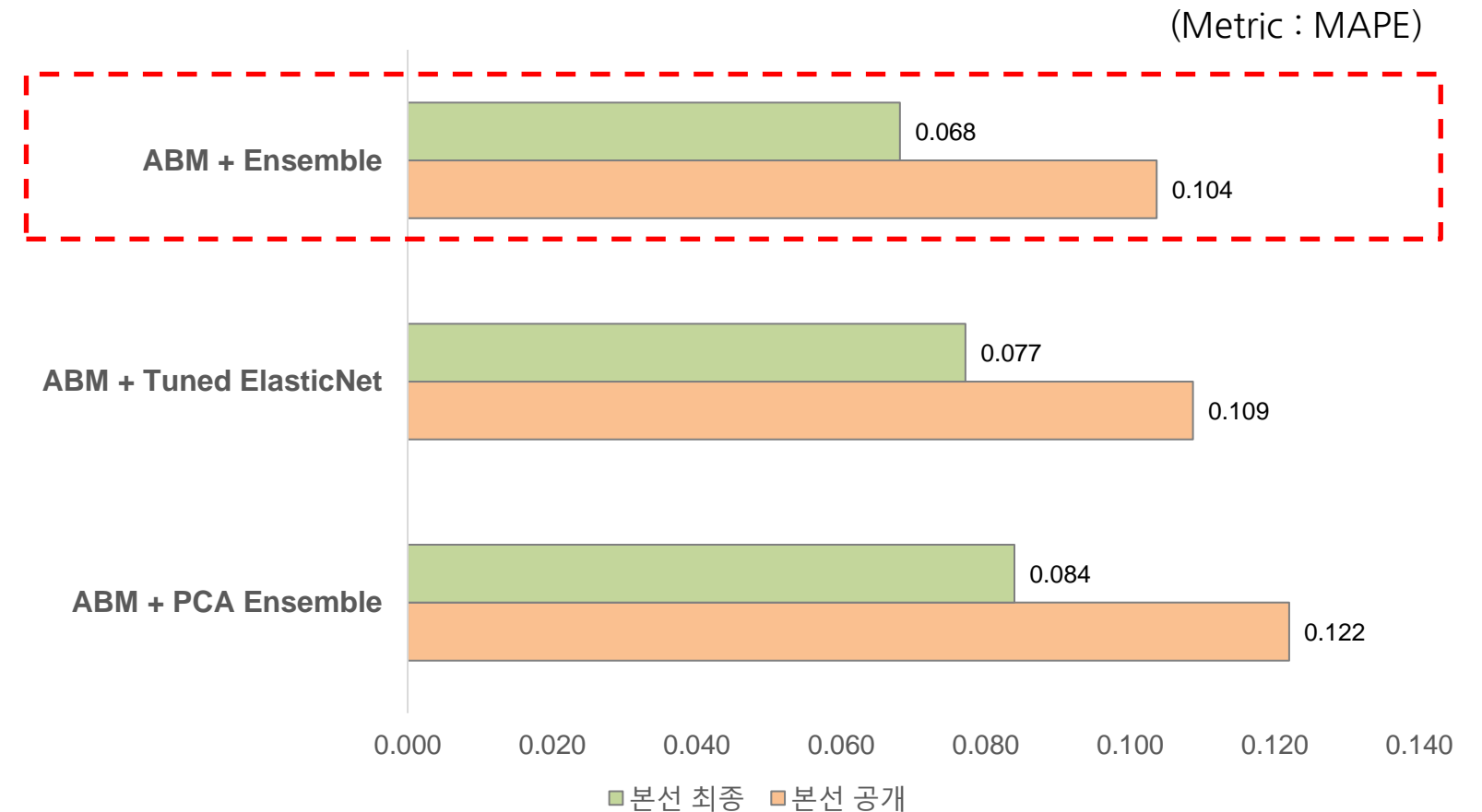
2. 방법론 설명 5) 결과 요약

(2) 스코어 결과

예선 결과



본선 결과



스테이지	점수	순위
본선 리더보드 결과	MAPE : 0.068	7위
예선 리더보드 결과	MAE : 847,845	13위
개선 수준	-	6위 상승

- ✓ Feature의 미래값을 예측 후 최종 모델에 활용하는 Forecast Feature & Model Ensemble 아키텍처가 가장 우수
- ✓ 내부 모델 검증 시스템이 본선 스테이지에서 Public-Private Score Correlation을 보임

목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

3. 아이디어 제안

4. Appendix

3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(1) 개선점 발굴

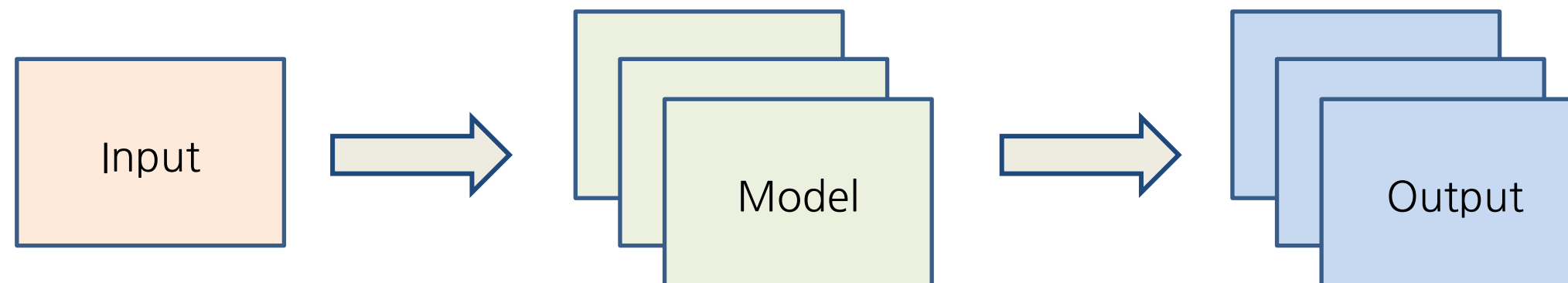


기존 KOGAS 아키텍처는 미래 시점의 수요를 예측하기 위해,
학습에 필요한 해당 시점의 변수를 예측하는 방식을 활용하고 있다고 추정*

이러한 구조는 변수 내 편향이 최종 '수요 예측 모델'에 반영되어 더 큰 오차 발생 가능성 존재



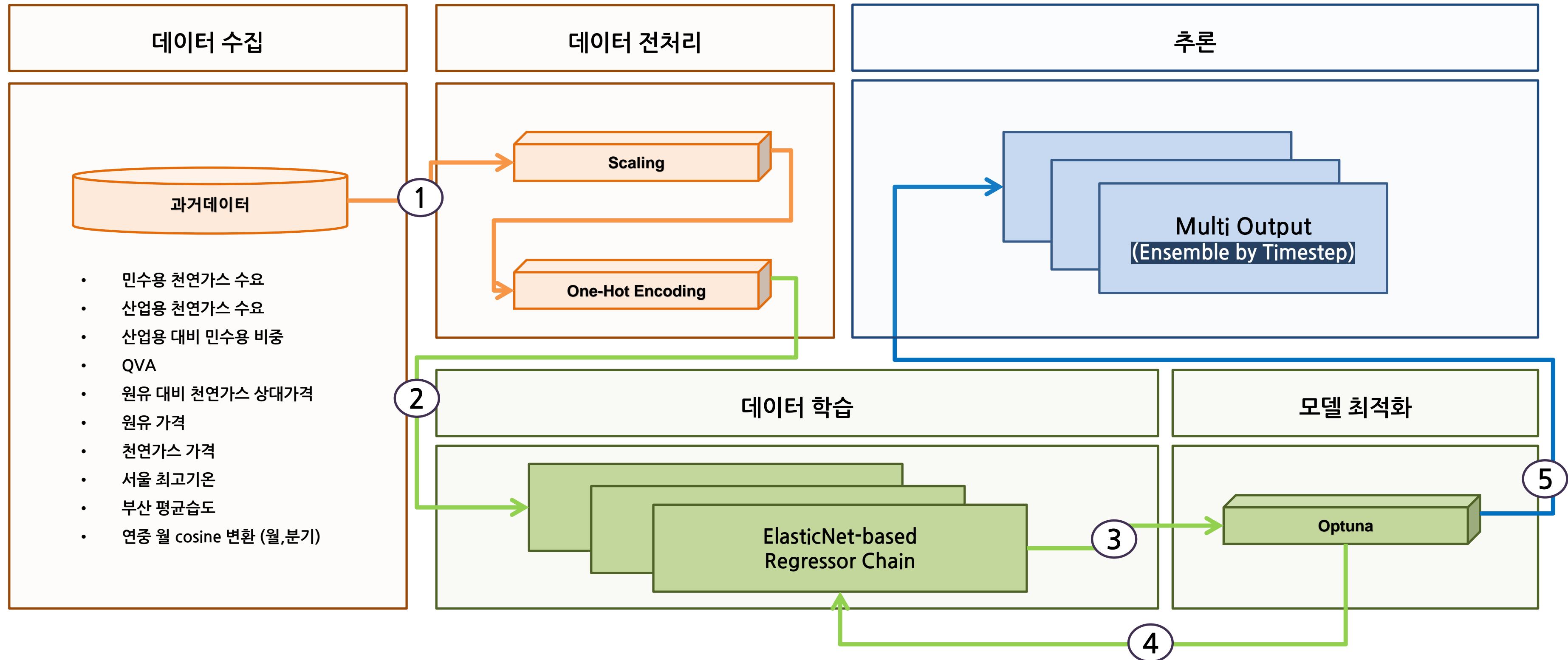
‘과거 데이터만을 활용해 예측을 하는 One-To-Many 아키텍처 개발을 시도’



* 에너지 포커스 문서 내 '에너지 수요 전망을 위한 (중략) 기온은 2020년 11월 30일까지의 일평균 실적치를 사용하였고, 이후 전망 기간에 대해서는 과거 10년의 일평균 기온 평균값을 이용하였다' 에서 해당 방식을 추정

3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

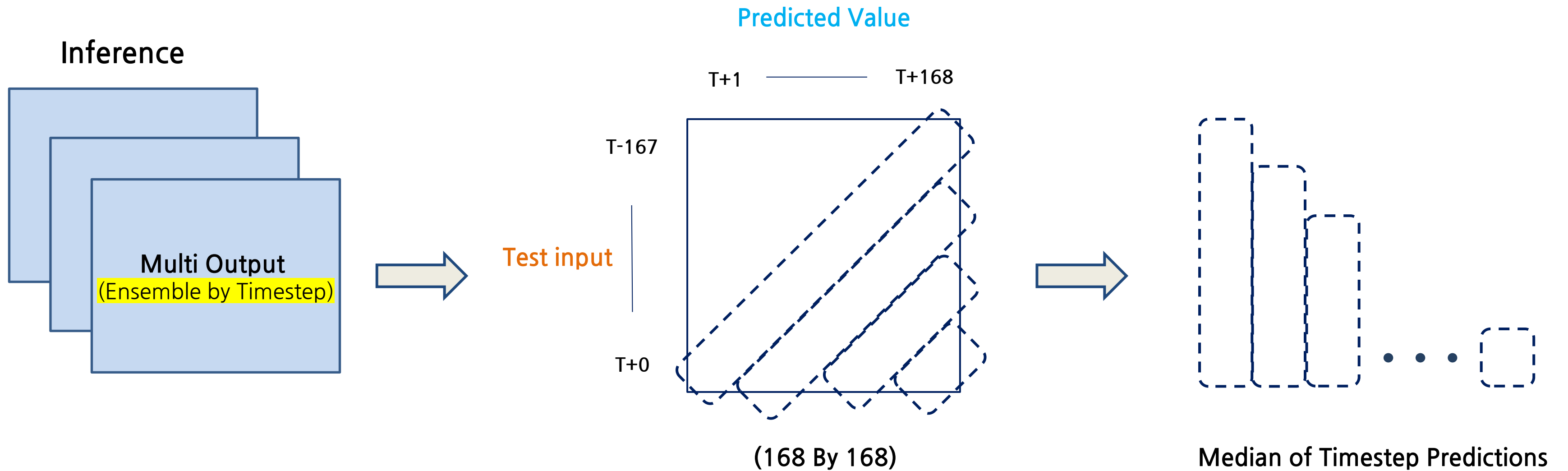
(2) 아키텍처 개요



3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(3) 차별점

Post-Inference Operation : Ensemble by Timestep



3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(4) 검증 스코어 요약

Fold	MAE	MAPE
1	54771	9.4%
2	50228	8.5%
3	56037	9.6%
4	48379	8.5%
Average	52354	9.0%

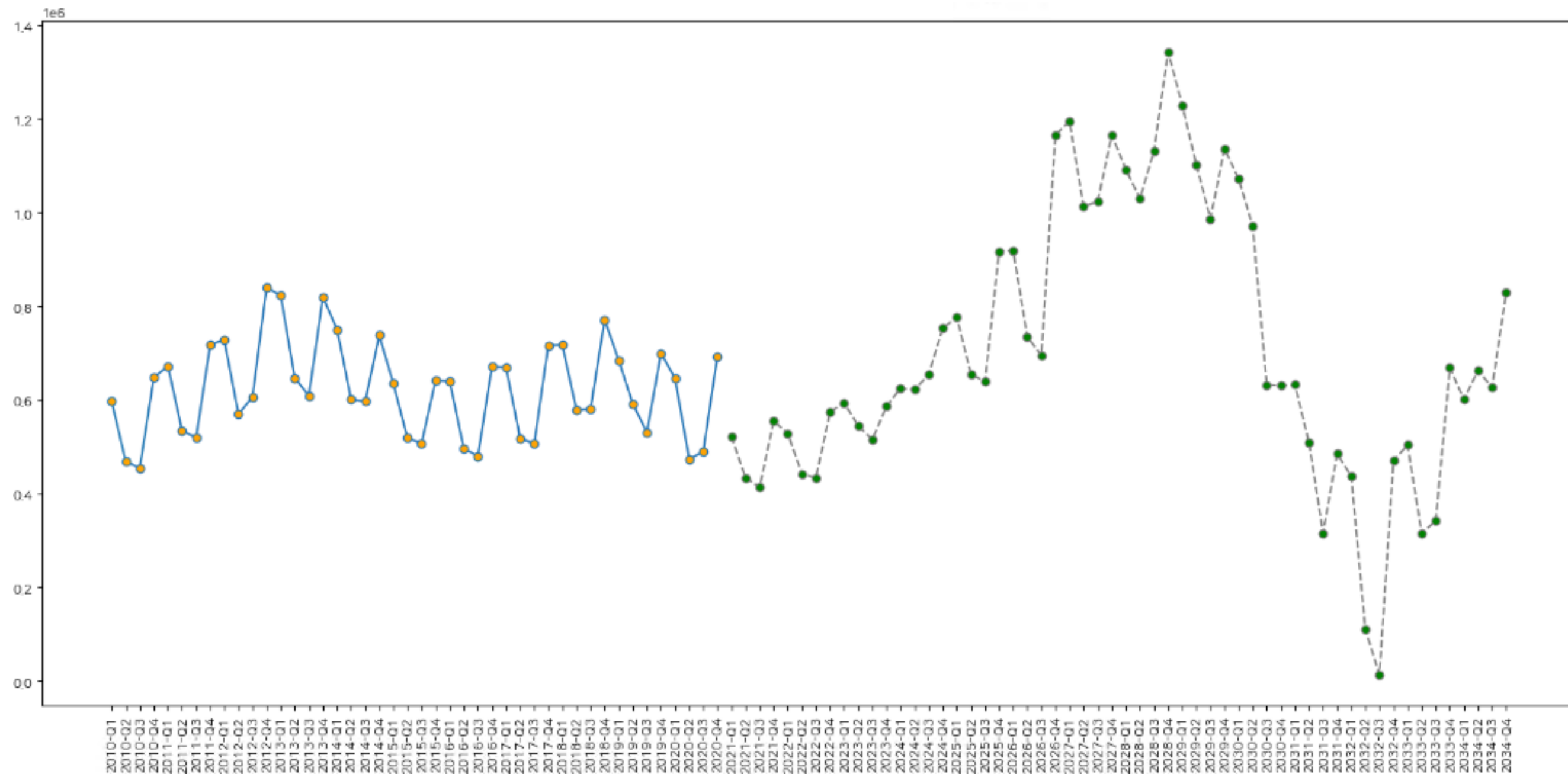
- ✓ 2006년의 각 분기를 Fold의 검증데이터, 그 시점 직전으로부터 60개(5년)의 데이터를 훈련데이터로 사용
- ✓ 검증데이터에 대해 MAPE가 최소가 되도록 하이퍼 파라미터 튜닝 (Optuna 모듈 활용)
- ✓ 추론 시에는 Fold로 학습시킨 모델의 결과를 산술 평균하여 최종 예측치 산출

3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(5) 예측 시각화

마지막 Timestep 만을 이용한 예측

분기별 평균

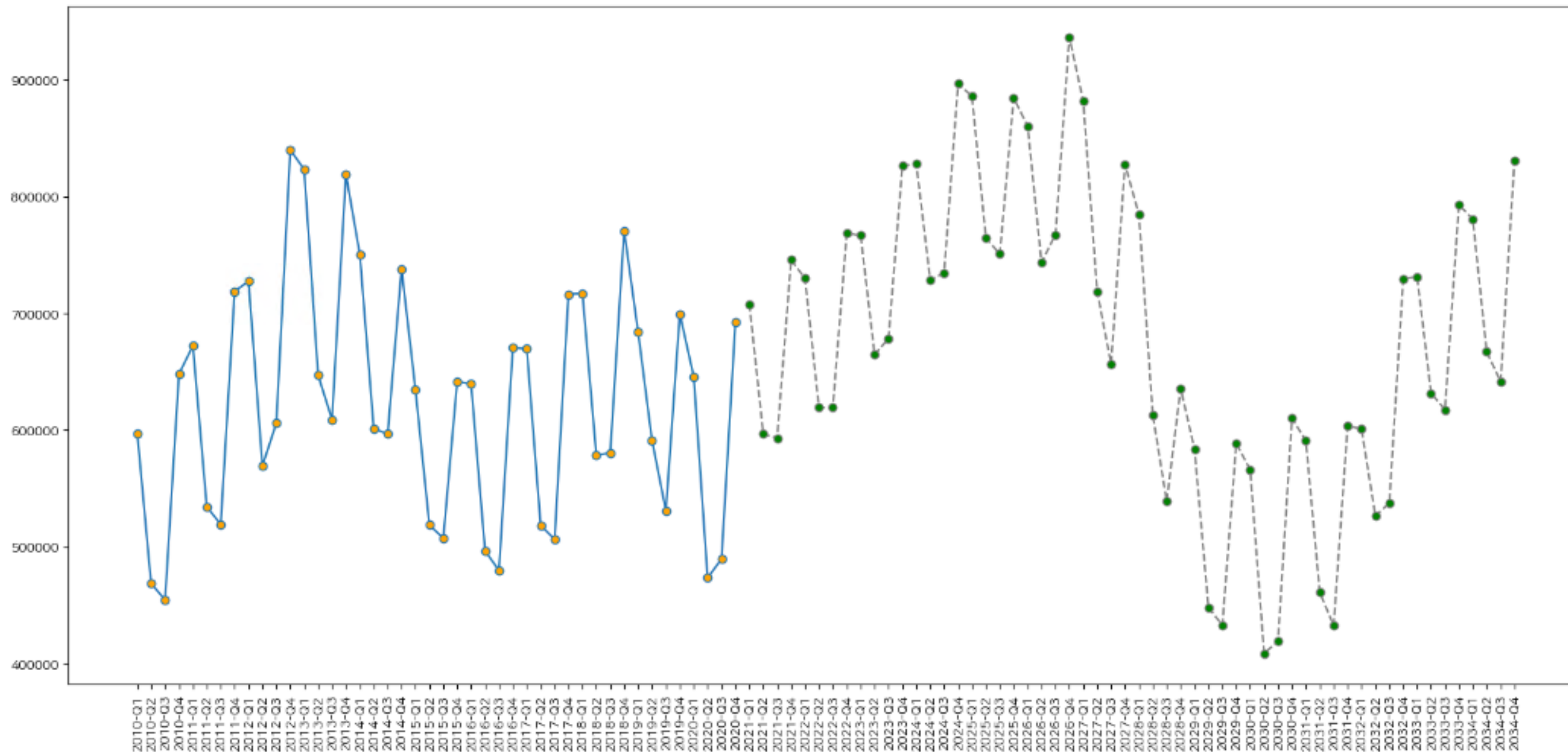


3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(5) 예측 시각화

168개 Timestep을 Ensemble하여 이용한 예측

분기별 평균

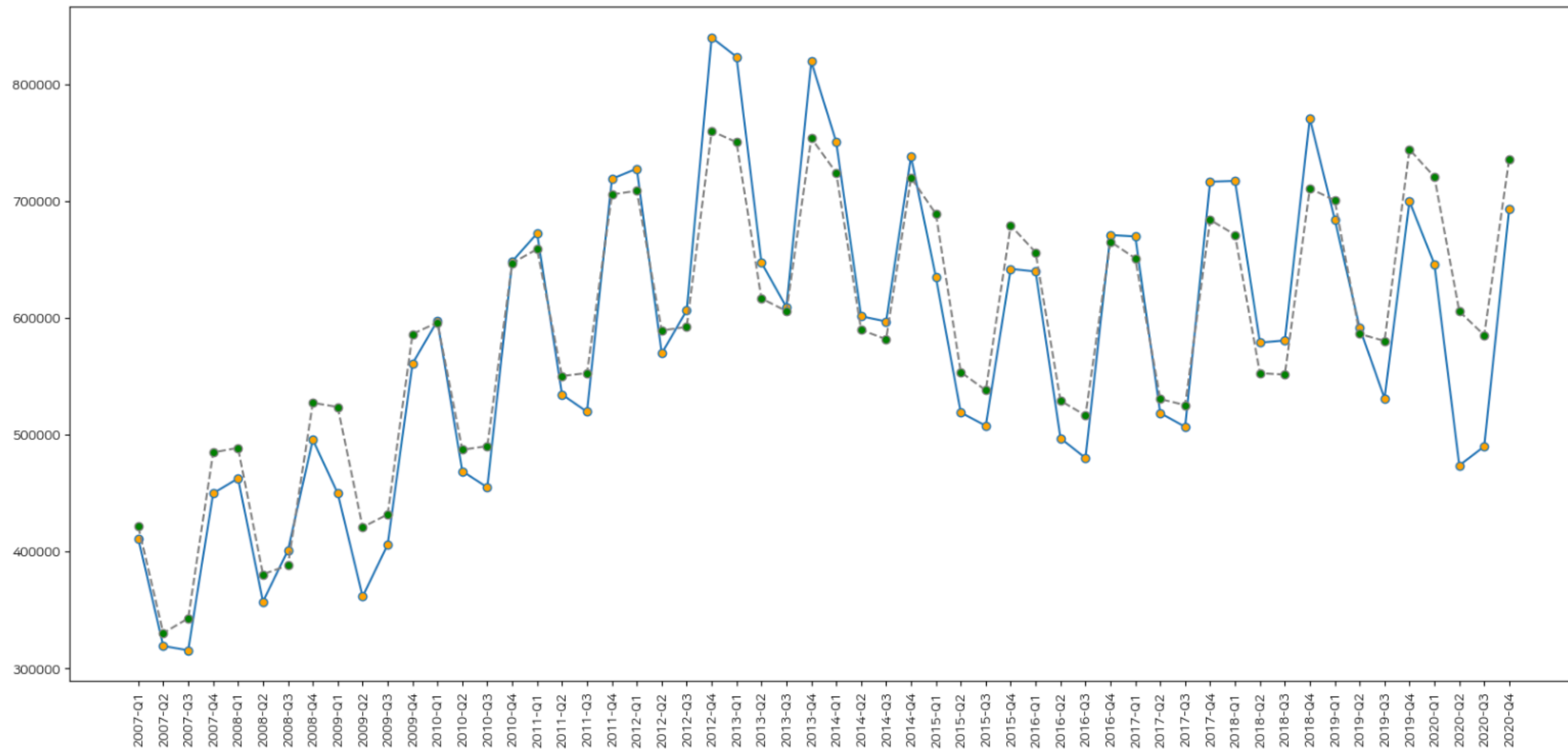


3. 아이디어 제안 산업용 천연가스 수요 ML 기반 One-To-Many 아키텍처

(5) 예측 시각화

훈련데이터 Ground Truth & Prediction 비교 (2007년 ~ 2020년)

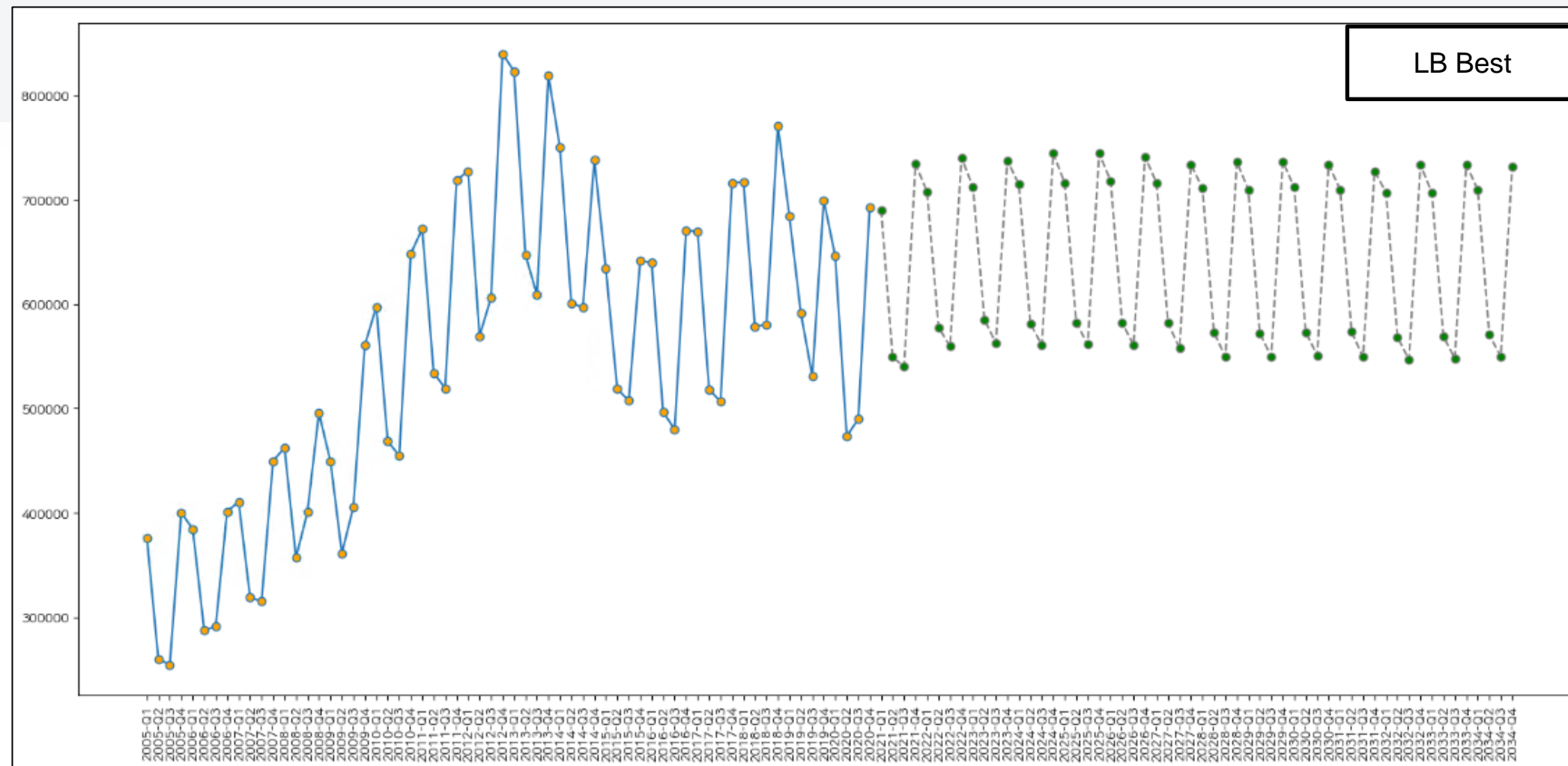
분기별 평균



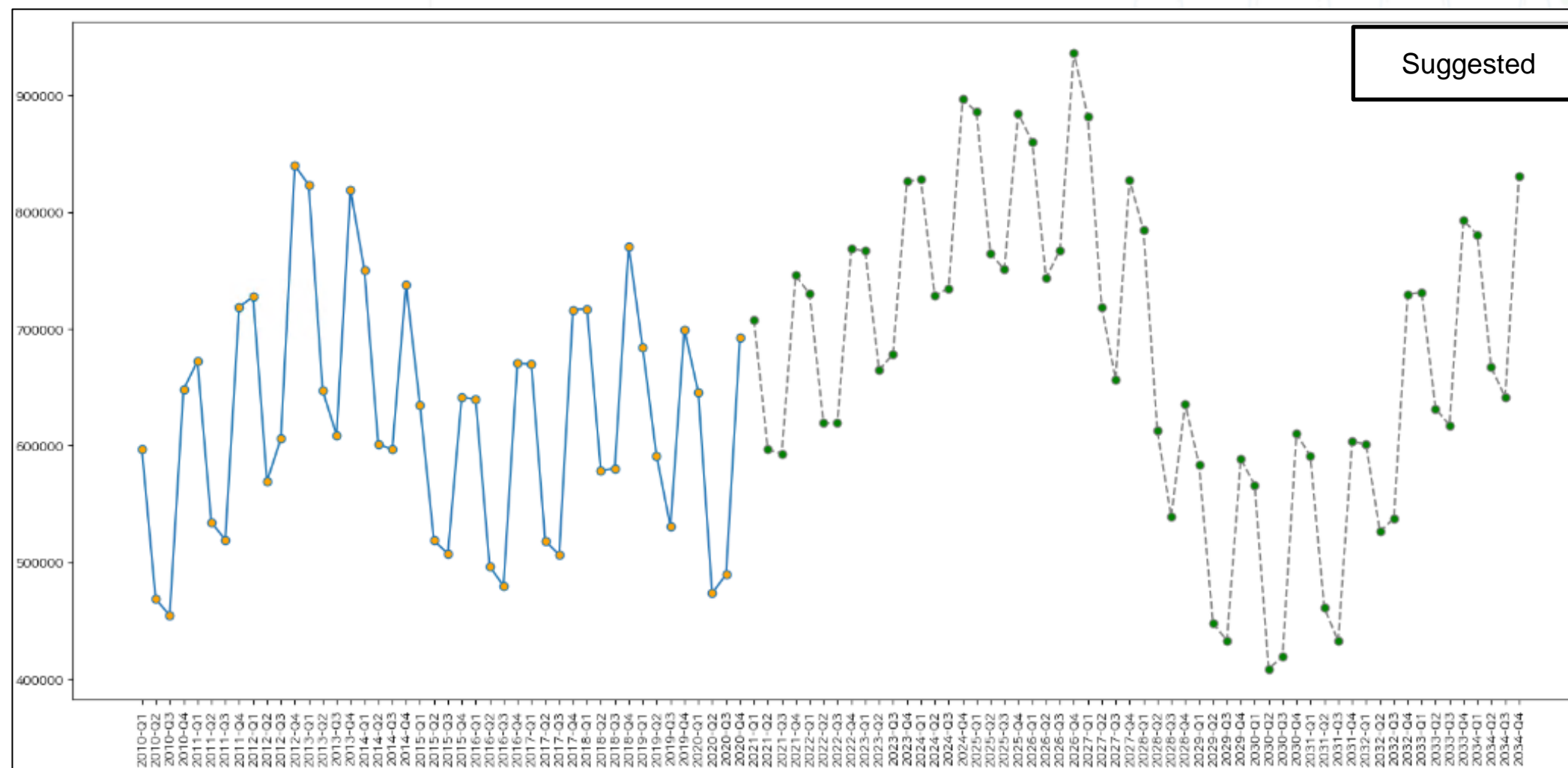
Too-Many 아키텍처

2007년 ~ 2020년)

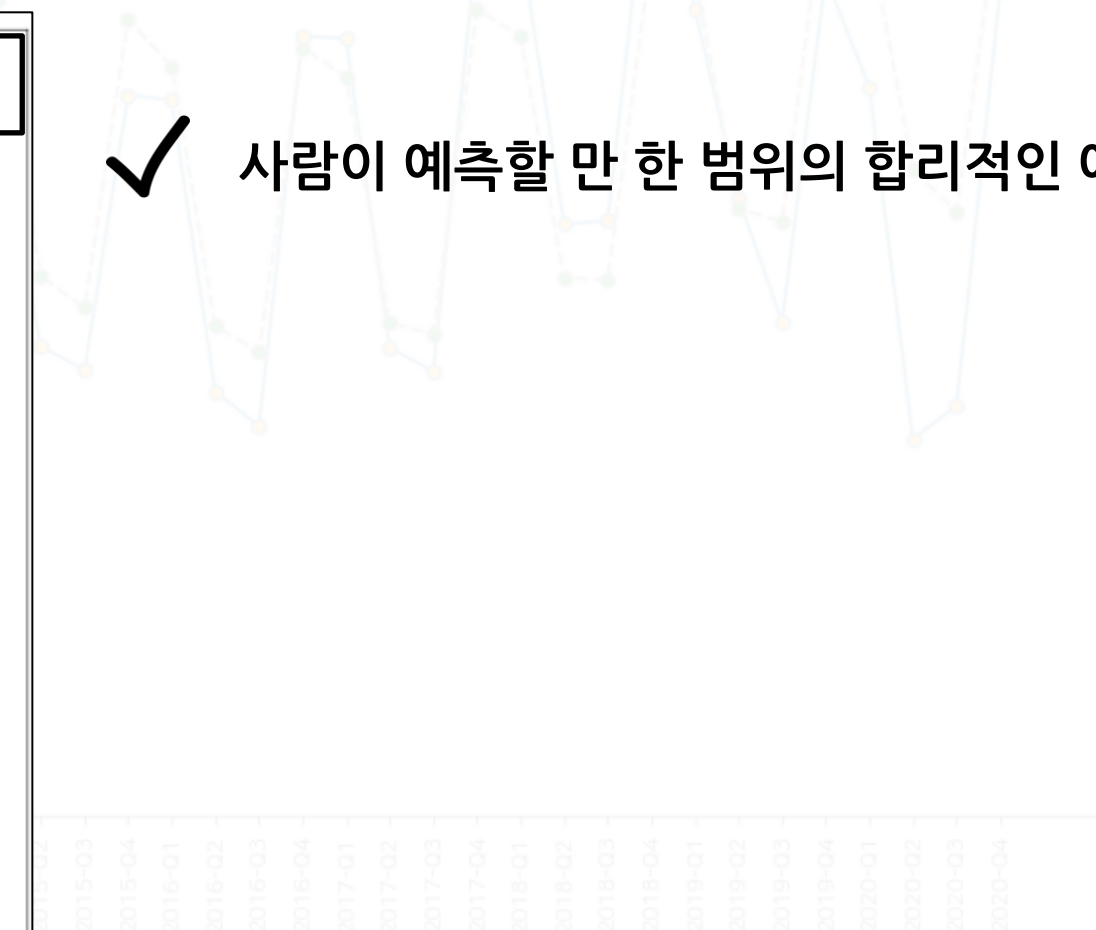
분기별 평균



✓ 최근 홍보하는 트렌드를 잡지 못하는 LB Best 모델과는 다르게, Suggested 모델은 잡아내는 모습을 보임



✓ 사람이 예측할 만 한 범위의 합리적인 예측 결과를 보임



목차

1. 개발 배경

2. 방법론 설명

(1) 데이터 소싱

(2) 데이터 엔지니어링

(3) EDA

(4) 모델링

(5) 결과 요약

3. 아이디어 제안

4. Appendix

4. Appendix 출처

외부 데이터

1. 기상청 <조건별 통계 - 온도 / 습도>

<https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do?pgmNo=179>

2. 통계청 <장래 인구 추계 2020~2070년>

<http://kostat.go.kr/assist/synap/preview/skin/miri.html?fn=d5ee78458568914110605&rs=/assist/synap/preview>

3. Enerdata <Energy Statistical Yearbook 2022>

<https://www.enerdata.co.kr/publications/world-energy-statistics-supply-and-demand.html>

4. OECD <Real GDP long-term forecast>

<https://data.oecd.org/gdp/real-gdp-long-term-forecast.htm#indicator-chart>

5. 종합 기후변화감시정보

http://www.climate.go.kr/home/09_monitoring/main

레퍼런스

1. 에너지경제연구원 <2022~2021년 에너지 수요 전망>

http://www.keei.re.kr/keei/download/focus/ef2103/ef2103_70.pdf

2. 산업통상자원부 <제14차 장기 천연가스 수급계획(2021-2034)>

<https://eiec.kdi.re.kr/policy/materialView.do?num=213255&topic=>

3. 전력 거래소 < 전력 수요 전망 주 모형 전망 프로세스 >

<https://new.kpx.or.kr/menu.es?mid=a10403030200>

Thank You



Q&A