프로젝트 소개

서울자전거 따릉이 소개

서울자전거 SEOUL BIKE 따릉이

- 회원 수 100만 명 돌파
- 누적 대여량 1536만 여 건
- 하루 평균 이용자 수 약 2만 4천명
- 2019년 3만대로 늘릴 예정

서울자전거 따릉이 소개

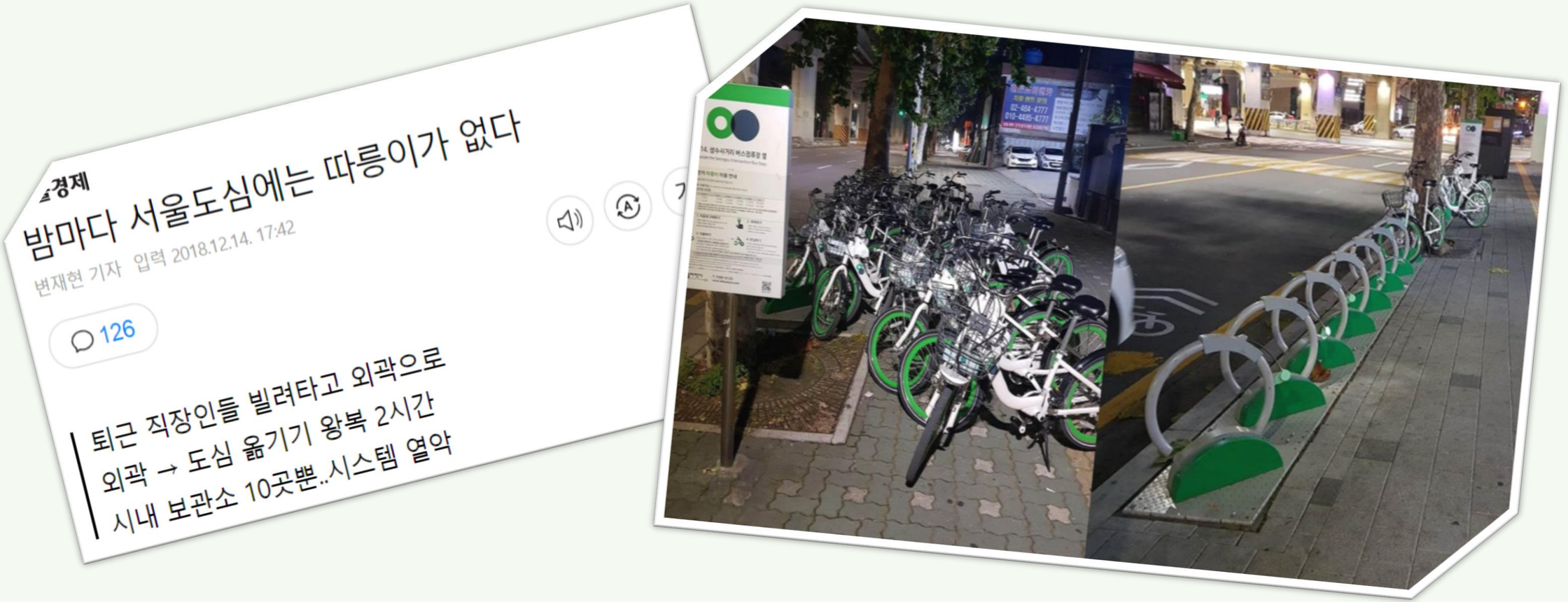
따릉이의 특성

- 모든 대여소에 반납 가능
- 출근 수요 1000 ~ 1500 대
- 퇴근 수요 4000 ~ 5000 대

따릉이 관리팀

- 강북·강남관리소 총 40개 팀 3교대 2조
- 기본적으로 70%의 거치율을 유지
- 매일 효율적인 업무를 위한 동선 선정

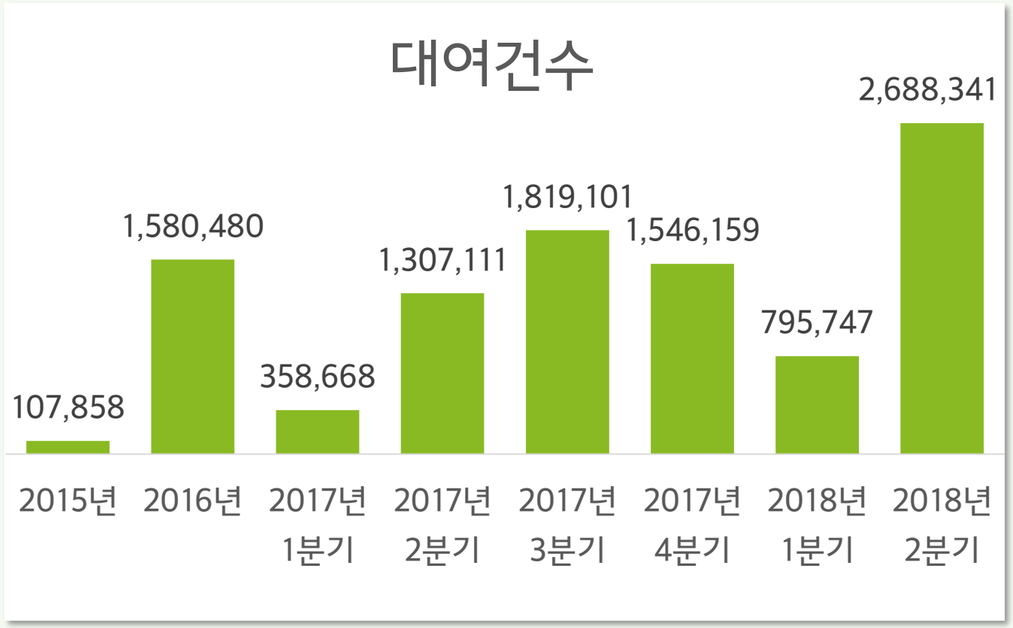
프로젝트 소개

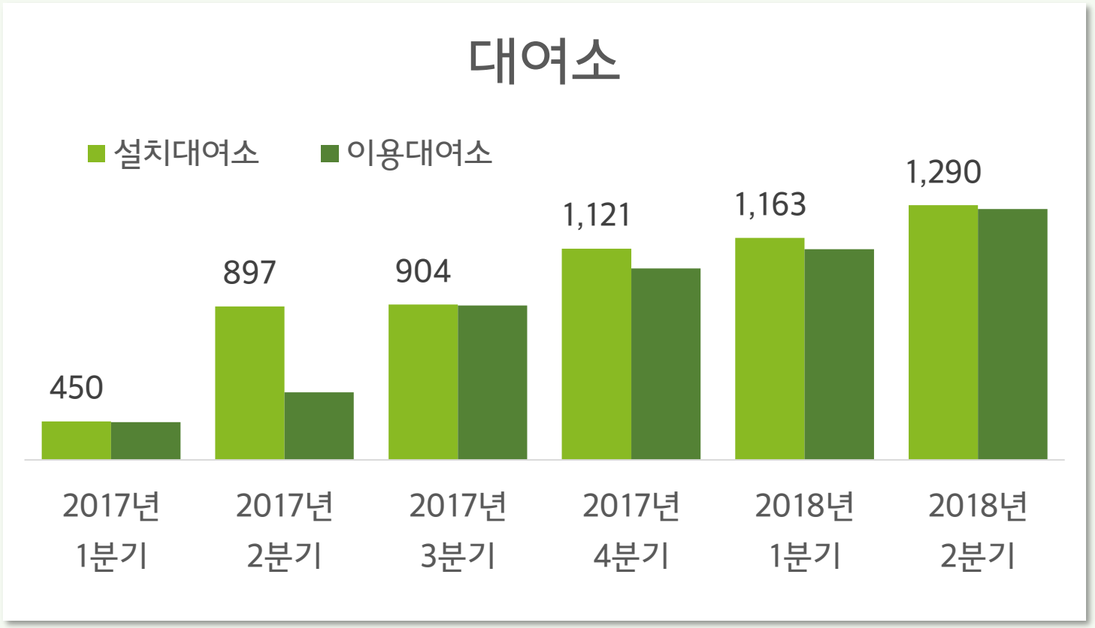


시간대 및 대여소별 수요예측 → 따릉이 재분배 최적화 기여

서울 자전거 대여이력 분석

서울 자전거 대여이력 개요



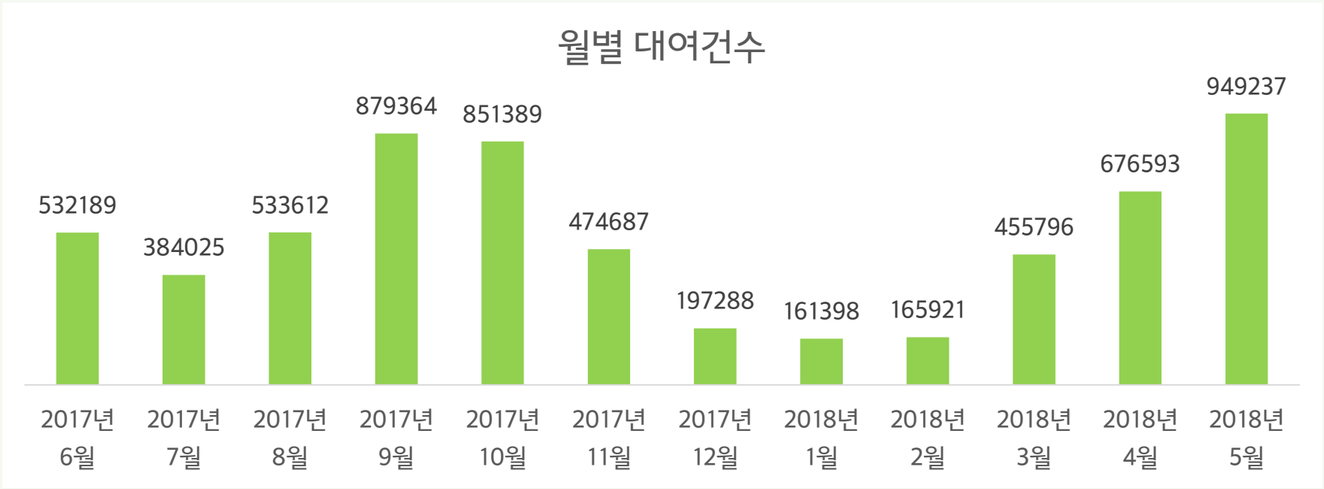


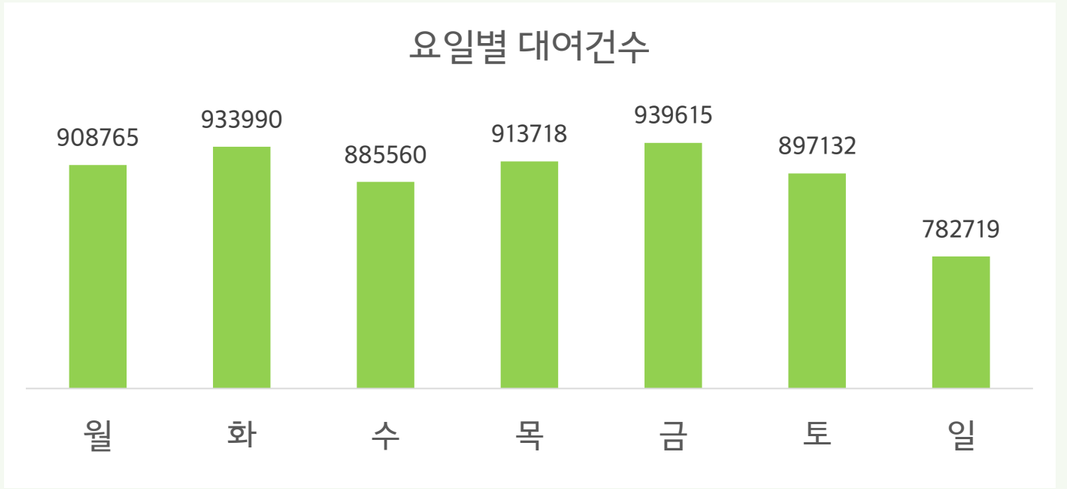
2017년 5월부터 모든 자치구에서 공공자전거 운영

↓

2017년 6월 ~ 2018년 6월 데이터 이용하기로 결정

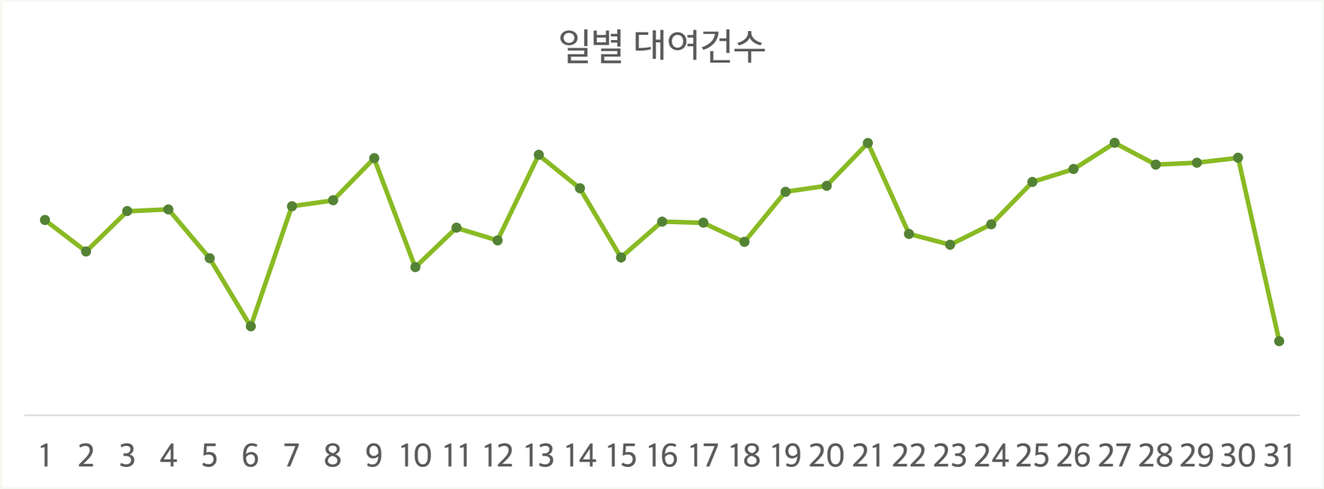
서울 자전거 대여이력 개요

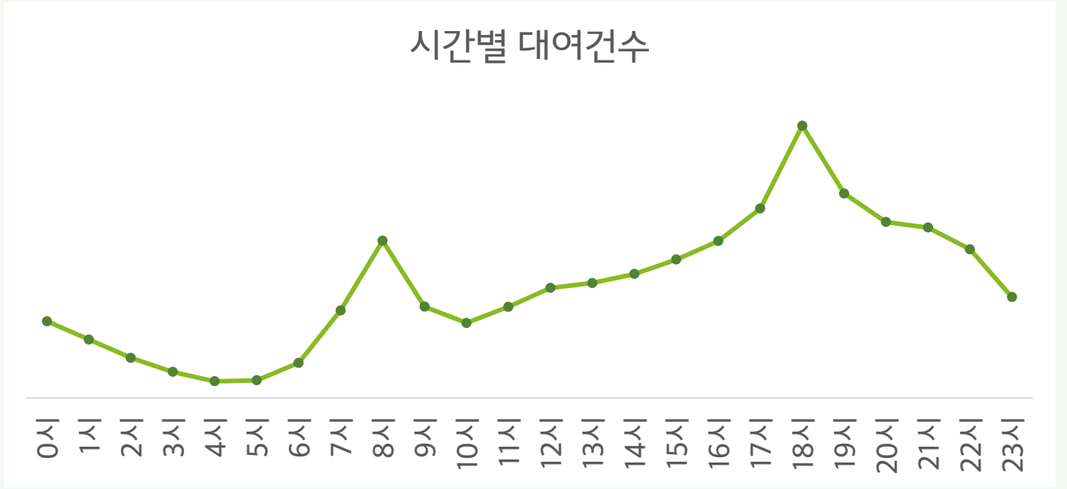




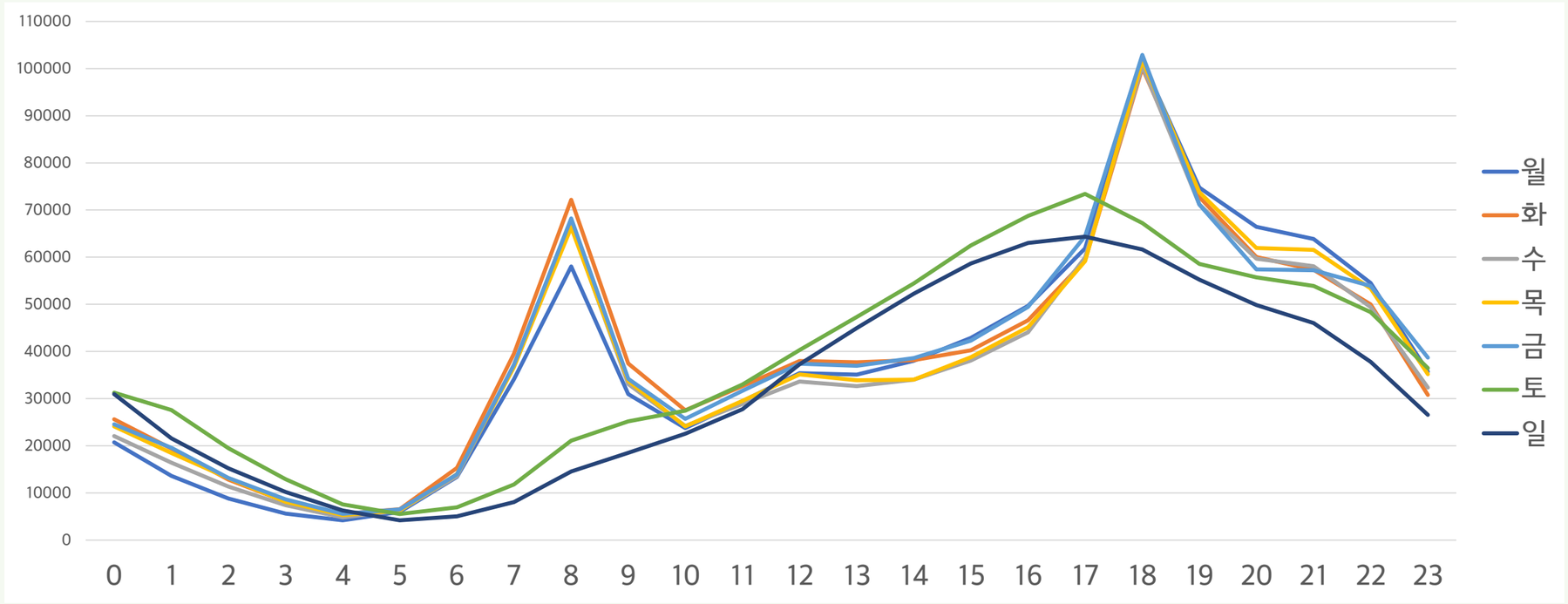
- 월별 대여건수로 계절성 파악

- 요일별 대여건수로 평일/주말 이용량 차이 파악





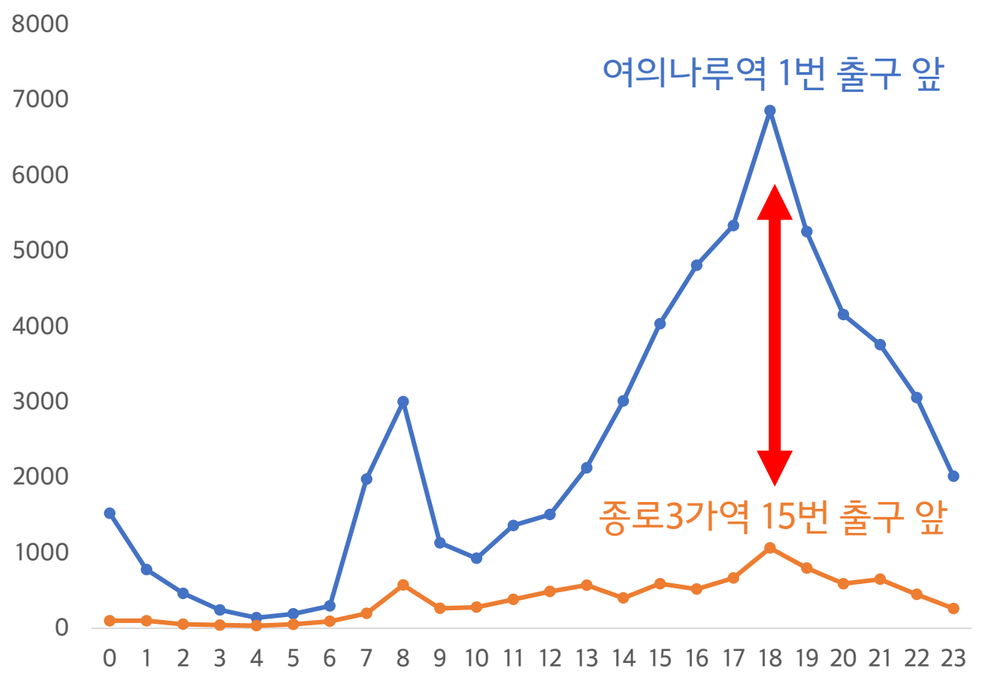
시간대별 및 요일별 대여이력 분석



- 평일 출근 시간대(6~9시) 및 퇴근 시간대(18~19시) 이용빈도 상승
- 주말 오후시간대(13~17시) 이용빈도 상승

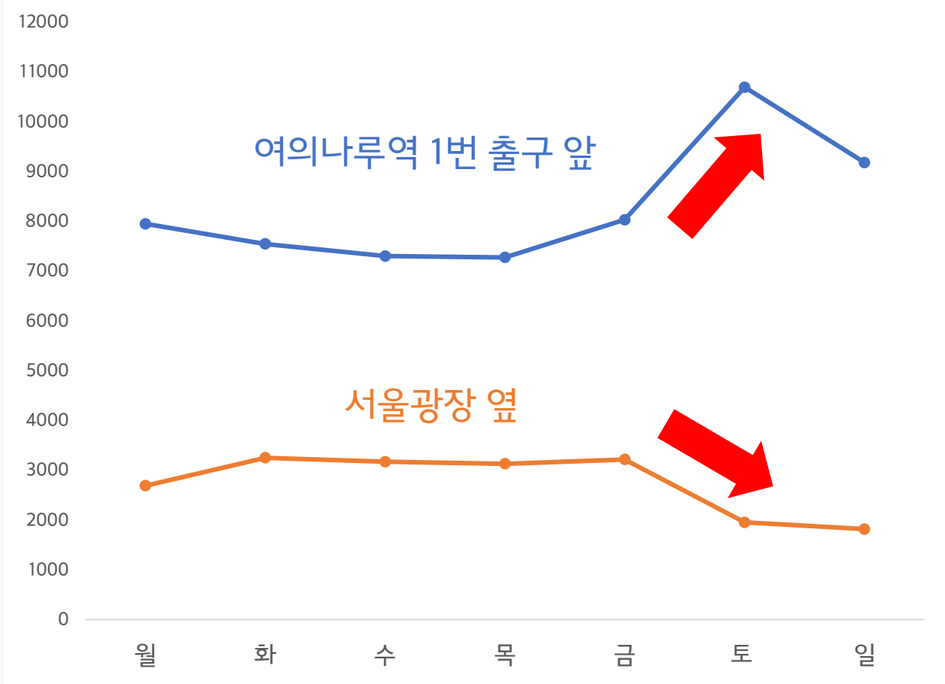
위치별 대여이력 분석

특정 대여소 간 대여건수 비교



- 대여소의 특성에 따라 대여건수에서 큰 차이가 남
  - 여의나루역 1번 출구 앞 : 141~6859건
  - 종로3가역 15번 출구 앞 : 35~1064건

특정 대여소 간 요일별 대여패턴 비교



- 대여소의 특성에 따라 교통수단 또는 여가수단으로 활용됨
  - 여의나루역 1번 출구 앞 : 주말 대여건수 증가 → 여가수단
  - 서울광장 옆 : 주말 대여건수 감소 → 교통수단

이용 데이터 내역

이용 데이터 내역

자전거 관련 데이터

시간대 및 대여소별
자전거 대여이력

대여소별 지리적 정보
(한강/공원과의 거리, 고도의 변화,
근접 공원 넓이, 지하철역 접근성)

외부 데이터

날짜 정보(시간, 요일 및 공휴일)

대여소별 날씨 정보
(풍속, 기온, 강수유무, 습도, PM10 등)

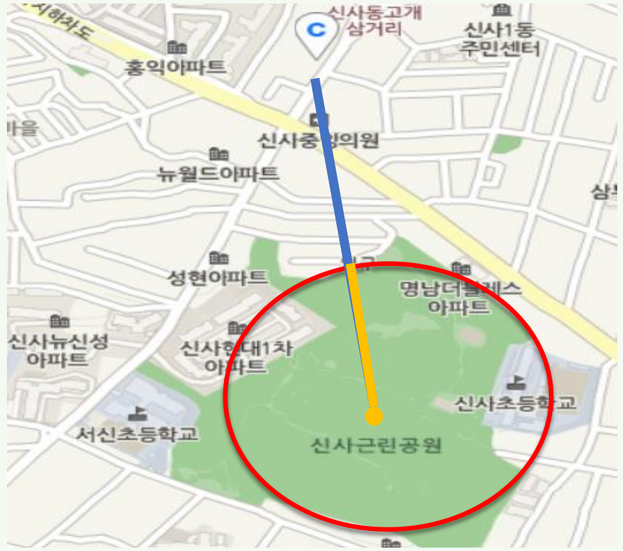
시간대 및 지역별 유동인구 수
(성별 및 나이대별 집계)

이용 데이터 내역 - 자전거 관련 데이터

대여소별 지리적 정보
(한강/공원과의 거리, 고도의 변화,
근접 공원 넓이, 지하철역 접근성)

대분류	중분류	상세 설명
대여소 지리적 정보	근접 한강 거리	임의로 지정한 포인트 기준과의 거리
	근접 공원 거리	공원을 원으로 가정한 중심과의 거리
	근접 공원 넓이	공원정보 데이터 기준
	고도 변화(1km)	반경 1km 이내 100m 간격 표준편차
	근접 지하철 수(1km)	반경 1km 이내 지하철역 수

대여소 근접 한강 거리 계산 방식



대여소 근접 공원 거리 계산 방식

대여소별 특징을 변수로 반영

→ 변수로 대여소를 구별하고자 함

이용 데이터 내역 - 외부 데이터

날짜 정보(시간, 요일 및 공휴일)

- 교통수단과 여가수단으로 활용성이 높은 특성을 고려
- 특정 기간 및 주기마다 반복되는 패턴 분석하여 날짜 정보의 중요성을 파악

날짜	공휴일 전날	공휴일	공휴일 다음날
2017-02-29	1	0	0
2017-03-01	0	1	0
2017-03-02	0	0	1
2017-03-03	0	0	0

대분류	중분류	상세 설명
날짜	0시	one-hot encoding으로 처리
	1시	
	2시	
	…	
	22시	
	23시	
	월	
	화	
	…	
	토	
	일	
	공휴일 전날	
	공휴일	
	공휴일 다음날	

→ 시간적인 맥락을 반영하기 위한 변수

이용 데이터 내역 - 외부 데이터

대여소별 날씨 정보
(풍속, 기온, 강수유무, 습도, PM10 등)

- 대여소 위·경도 좌표와 근접한 관측소를 연결
- 초미세먼지는 비어 있는 데이터가 약 40%이므로 제외함

분류	기호	기상현상	기호	기상현상	기호	기상현상
물	01) ●	비	02) ●	이슬비	03) ●	착빙성의 비
	04) ●	소낙비	05) ✕	눈	06) ●	진눈깨비
	07) ●	착빙성의 이슬비	08) ✱	소낙눈	09) ●	소낙성 진눈깨비
	10) ✕	싸락눈	11) ∴	가루눈	12) ▲	어는비
	13) △	싸락우박	14) ▲	우 박	15) ↔	얼 음 침
현	16) ≡≡	안개, 낮은안개	17) ≈	땅 안 개	18) ≋	얼음안개
	19) =	박 무	20) ⫩	땅 날린눈	21) ⫩	높은 날린눈
	22) ⫩	눈 보 라	23) ○	이 슬	24) ◎	언 이슬
상	25) ㄴ	서 리	26) ㄱ	서 릿 발	27) ㄱ	무 빙
	28) ㄴ	수 상	29) ∨	수 빙	30) ∨	조 빙
	31) ∿	우 빙	32) ㄴ	결 빙	33) ㅍ	용 오 름
	34) ㅁ	해 빙	35) ㅁ	유 빙	36) ㅈ	해 명

대분류	중분류	소분류	상세 설명
날씨	기상 관측	풍속	km/h
		기온	℃
		기압	hPa
		누적강수	mm
		습도	%
	공기질	PM10	㎍/㎥
		아황산가스	ppm
		이산화질소	ppm
		오존농도	ppm
		일산화탄소	ppm
		황산가스	ppm
	현천 코드	눈	0 or 1
		비	0 or 1
		…	0 or 1
		소나기	0 or 1
		이슬	0 or 1
		안개	0 or 1

기상관측 출처 : https://ir.freemeteo.com
공기질 출처 : 에어코리아
현천코드 출처 : 기상청 ASOS

→ 환경적인 요소를 반영하기 위한 변수

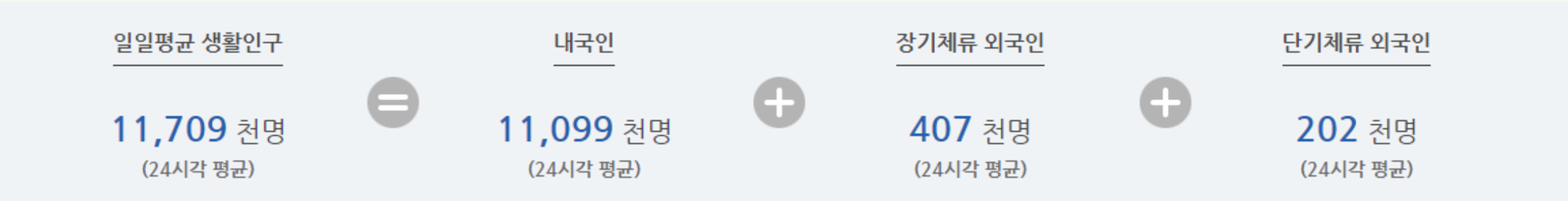
이용 데이터 내역 - 외부 데이터

시간대 및 지역별 유동인구 수
(성별 및 나이대별 집계)

대분류	중분류	상세 설명
유동인구	남자10대	남10대~70대 여10대~70대 각 행정동의 시간대별 유동인구 수
	…	
	남자70대	
	여자10대	
	…	
	여자70대	

- 서울 생활 인구

  : 서울시와 KT가 공공 빅데이터와 통신 데이터를 이용하여 추계한 인구



→ 시 • 공간적 의미를 추가하기 위한 변수

이용 데이터 내역 - 외부 데이터

시간대 및 지역별 유동인구 수
(성별 및 나이대별 집계)

대분류	중분류	상세 설명
유동인구	남자10대	남10대~70대 여10대~70대 각 행정동의 시간대별 유동인구 수
	…	
	남자70대	
	여자10대	
	…	
	여자70대	

- 전처리 과정에서 발생한 문제점

법정동

행정동

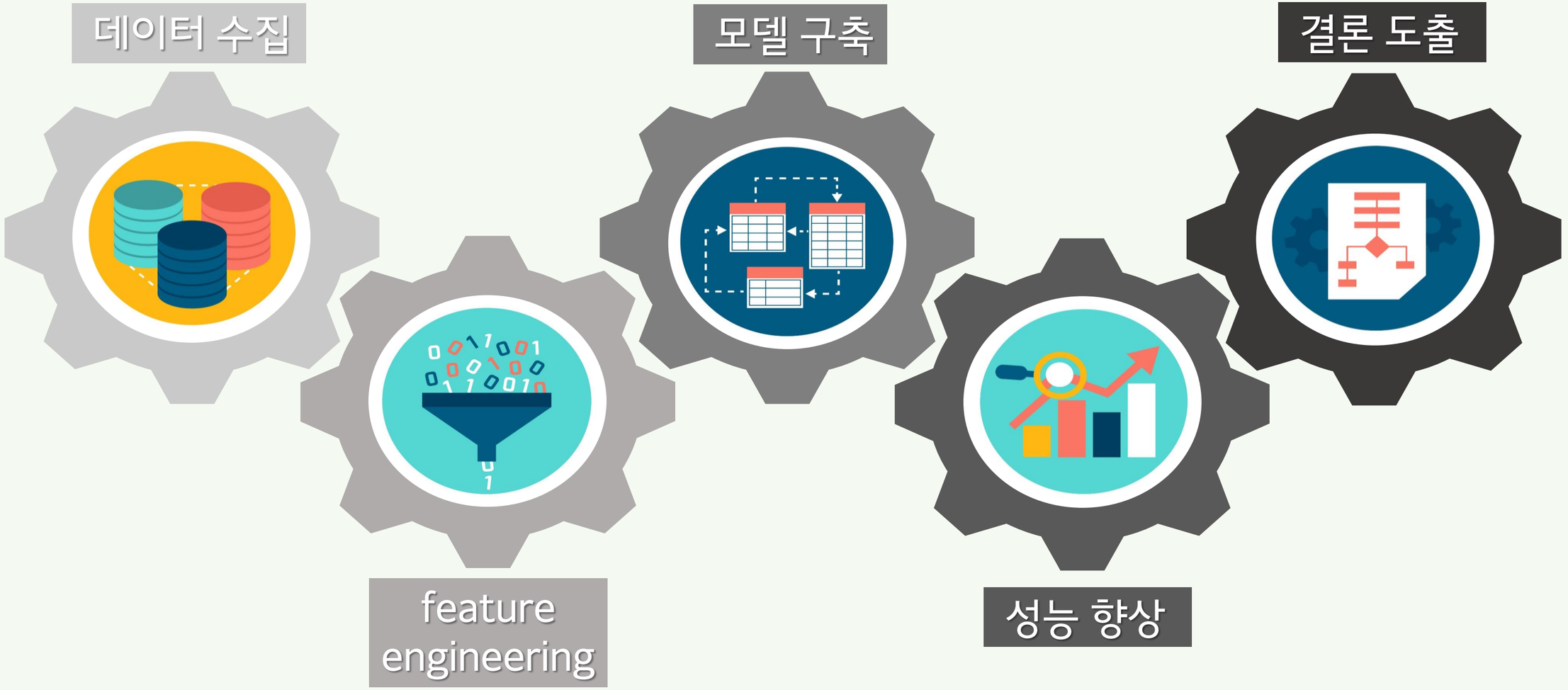
데이터 기준 지역이 일치하지 않는 문제 발생 (유동인구 : 행정동 기준 / 대여소 : 법정동 기준)

→ 법정동/행정동 코드를 이용하여 행정동으로 데이터 기준 통일

- 행정동 : 행정편의 및 관리를 위하여 재편성한 단위

수요예측 과정 및 결과

진행 과정



실험 설계

STEP 1

수요 데이터만 이용한 모델 VS 수요 + 대여소 지리적 정보 데이터를 이용한 모델

대여소 지리적 정보 데이터가 의미 있는 데이터인지 판단
& 가장 성능이 좋은 모델 판단

STEP 2

수요 + 대여소 정보 + 날씨 데이터를 이용한 모델

수요 + 대여소 정보 + 날짜 데이터를 이용한 모델

수요 + 대여소 정보 + 유동인구 데이터를 이용한 모델

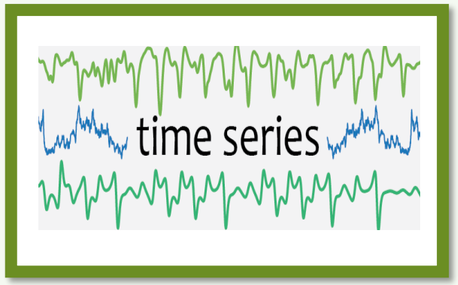
날씨, 날짜, 유동인구 데이터가 의미 있는 데이터인지 판단

STEP 3

수요 + 대여소 정보 + 날씨 + 날짜 + 유동인구 데이터를 이용한 모델

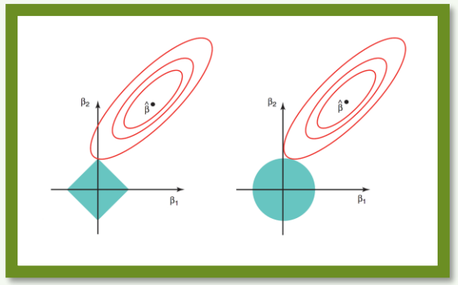
이용 모델

ARIMA
(AutoRegressive Integrated Moving Average)



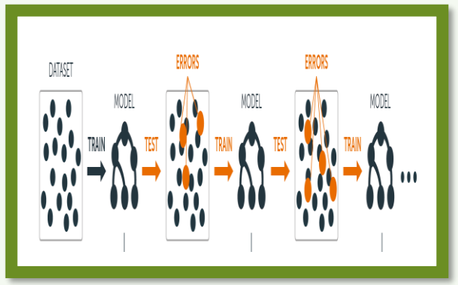
- 과거의 관측값과 오차를 사용해서 현재 시계열 값을 설명하는 기법

LASSO & Ridge



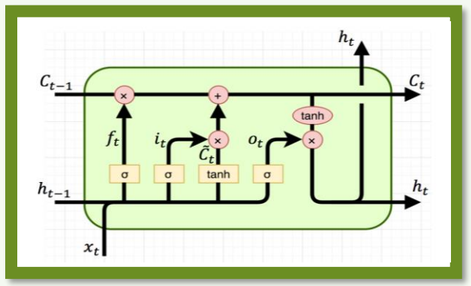
- 회귀 모형에 패널티 항을 더하여 과적합을 방지하는 기법

Boosting



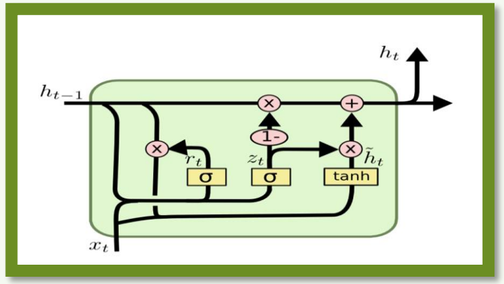
- 오답에 가중치를 부여하여 학습하는 앙상블 기법

LSTM
(Long Short Term Memory networks)



- RNN의 hidden state에 cell-state를 추가한 구조의 기법

GRU
(Gated Recurrent Unit)



- LSTM을 보다 간단하게 구조화한 기법

Shallow Learning ← → **Deep Learning**

수요예측 작업 환경

문제점 : Data 크기 >> 6GB → Memory Error 발생

해결책 : Google Cloud Platform 이용

- 구글의 클라우드 컴퓨팅 서비스
- GCP의 VM 인스턴스 / ML 엔진을 이용

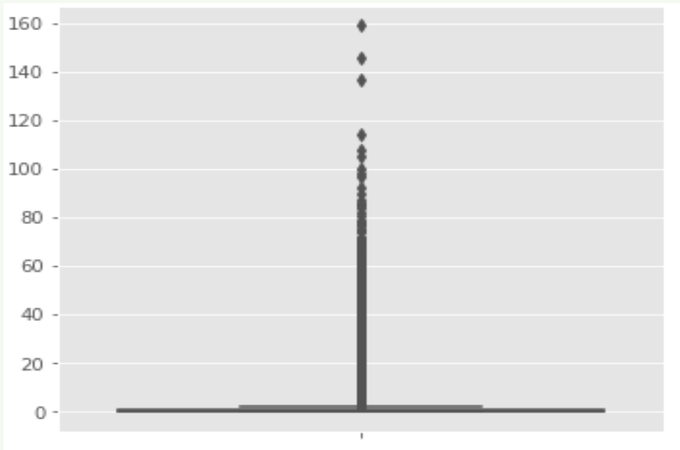
클라우드	GCP	ML engine
OS	Ubuntu 16.04	-
CPU	vCPU 12	vCPU 8
GPU	Tesla K80 X1	-
RAM	45G	52G
언어	Python	Python
IDE	Jupyter	-

수요예측 작업 환경

- performance measure

	MAPE	RMSE
의미	실제 값에 대한 오차의 비율	$\sqrt{\text{오차 제곱의 평균}}$
특징	다소 작은 값의 예측에 민감	절대적으로 큰 값의 예측에 민감
문제점	실제 값이 0인 경우 계산하기 어려움	예측 대상의 크기에 영향을 받음
수식	$\frac{1}{n}\sum_{j=1}^{n}\left\|\frac{y_j - \hat{y_j}}{y_j}\right\| \cdot 100$	$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2}$

수요량 boxplot



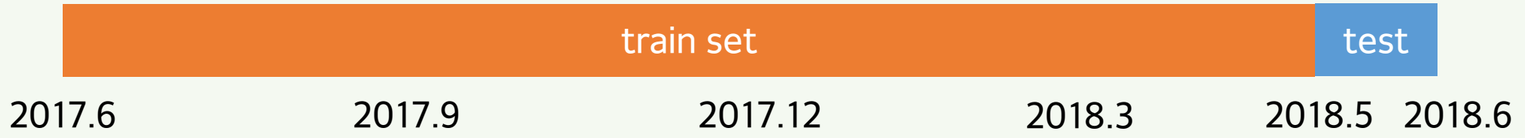
- 수요량이 0인 경우가 많음
- 절대적으로 큰 값도 많이 분포되어 있음

  → RMSE 사용하기로 결정

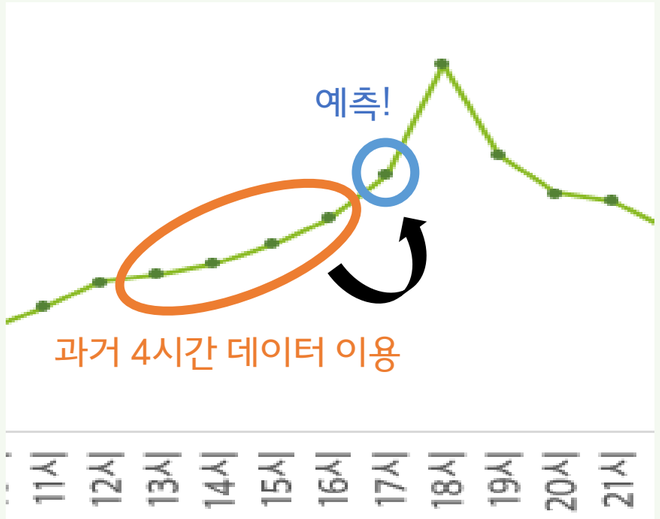
- 모델별 hyperparameter 설정
  - ARIMA : p = 0, d = 1, q = 1
  - LASSO : alpha = 0.001237
  - Ridge : alpha = 0.0001376
  - XGBoost : max_depth = 5
  - LSTM, GRU : learning rate = 0.001
    n_step = 4 / n_hidden = 32
    n_multicell = 1~2
    epoch = 5 / batch_size = 256

수요예측 작업 환경

- data 기간 설정
  - train set에 계절성이 반영되도록 1년치 데이터를 이용 (2017.6~2018.5)
  - test set은 이후 1달 데이터를 이용 (2018.6)

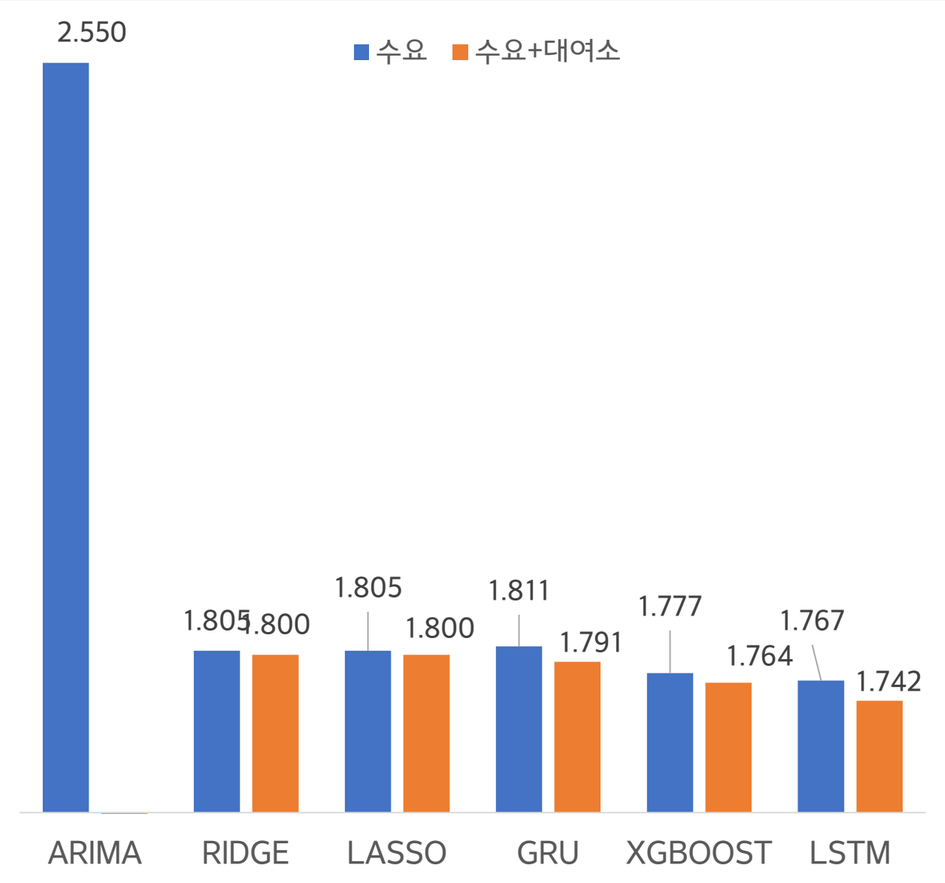


- 과거 수요 데이터를 변수로 입력
  - 시계열 데이터의 자기 회귀적인 특성 반영
    - ✓ 자기회귀(autoregressive) : p 시점 전의 자료가 현재 자료에 영향을 줌
  - 교통량 예측 연구 시 통상적으로 사용하는 과거 4시간 데이터를 이용



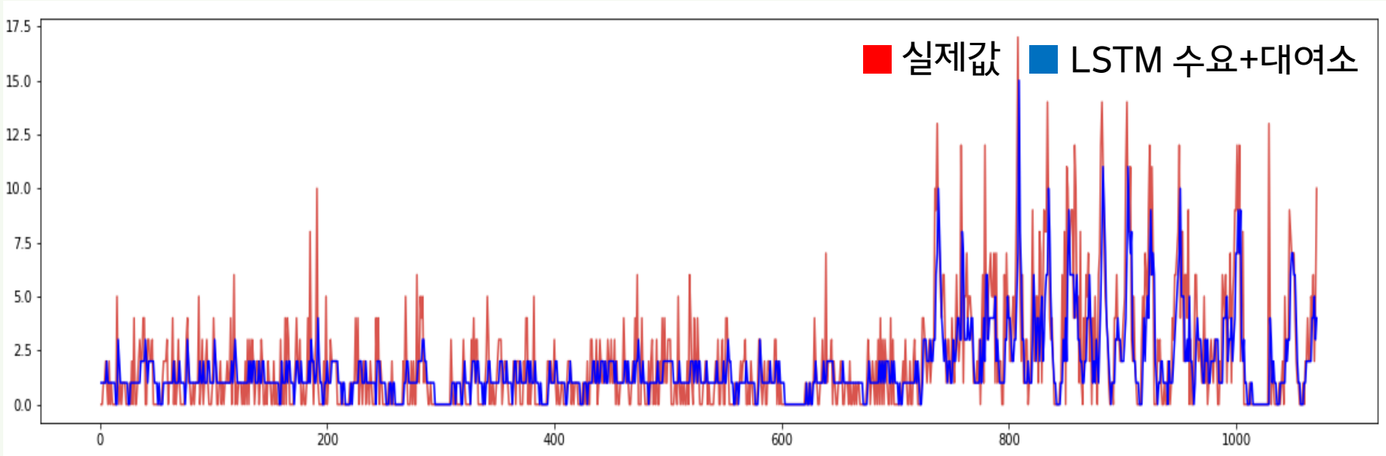
수요예측 결과 - STEP 1 : 수요 및 대여소 정보 반영 결과 비교

모델별 수요 및 수요+대여소 정보 RMSE 값 비교

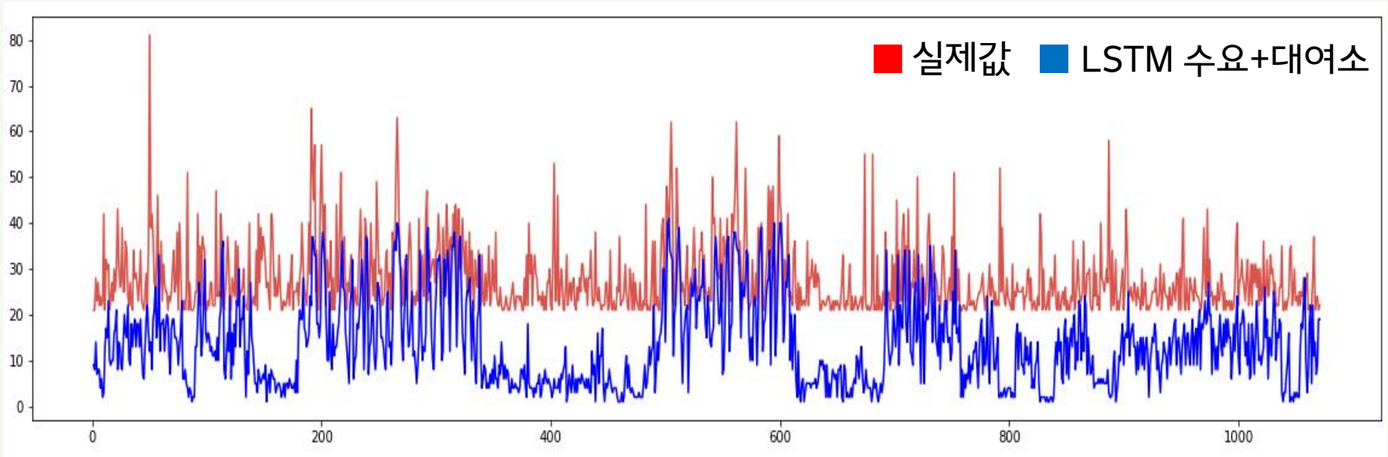


- 대여소 정보를 넣어도 성능에 큰 차이가 나타나지 않음
- LSTM 모델이 가장 좋은 성능을 나타냄

시작점부터 1000번째 까지 결과 비교

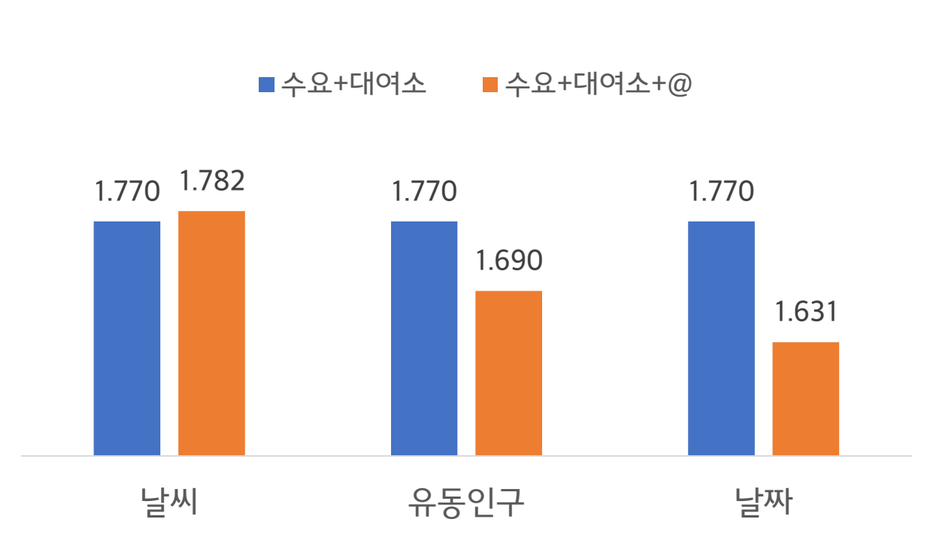


실제 대여건수가 20 이상인 경우 비교



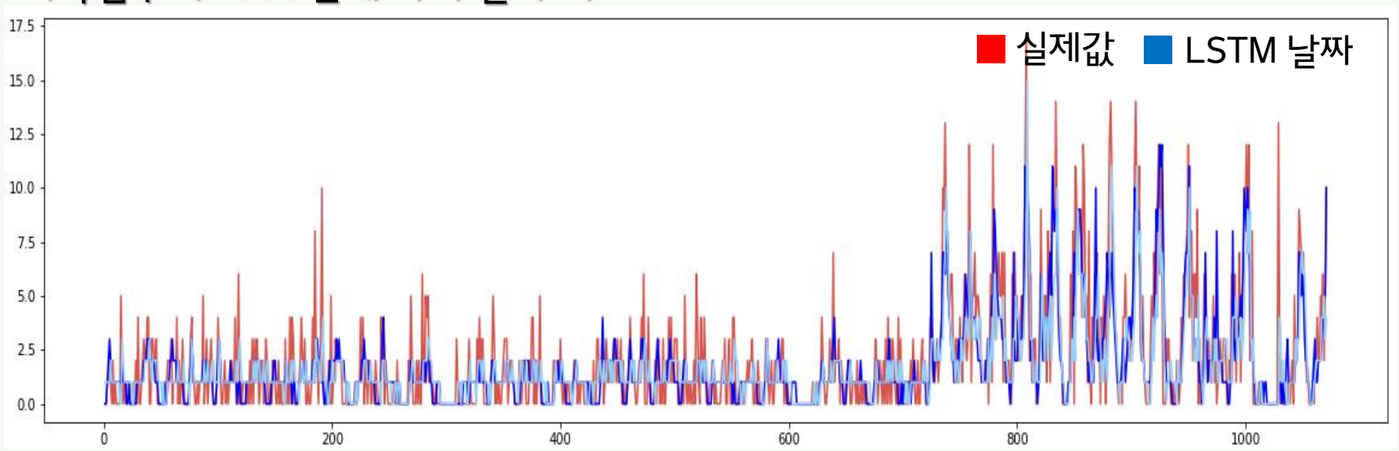
수요예측 결과 - STEP 2 : LSTM 모델에 새로운 feature 추가

새로운 feature 추가 시 RMSE 값 비교



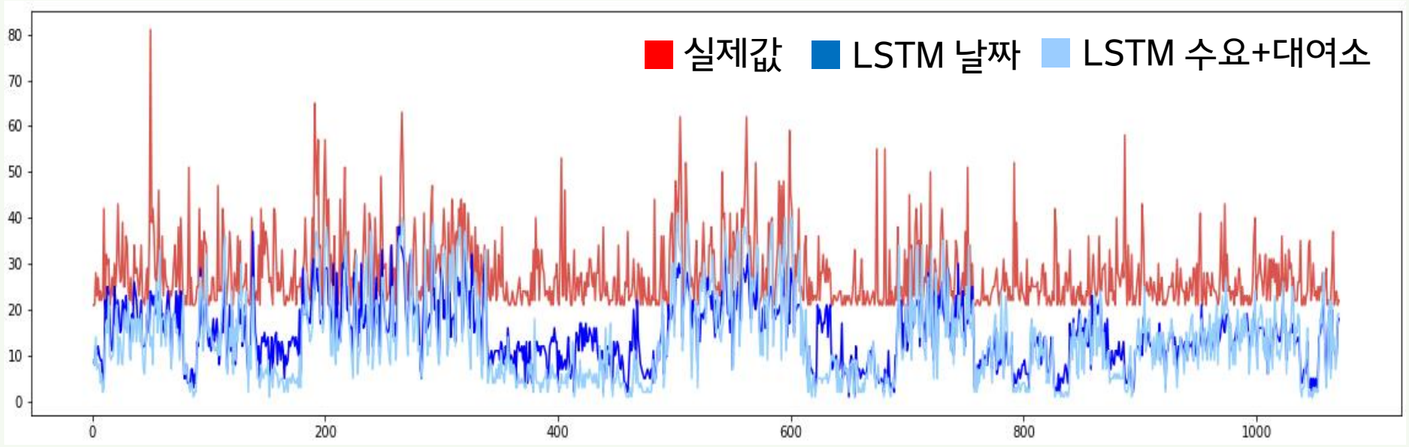
Feature	Parameters				
	Input	step	L rate	epoch	batch
대여소	11	4	0.001	5	256
날씨	33	4	0.001	5	256
유동인구	25	4	0.001	5	256
날짜	45	4	0.001	5	256

시작점부터 1000번째 까지 결과 비교



- 트렌드를 더 잘 맞춰가는 경향을 보이고 있음

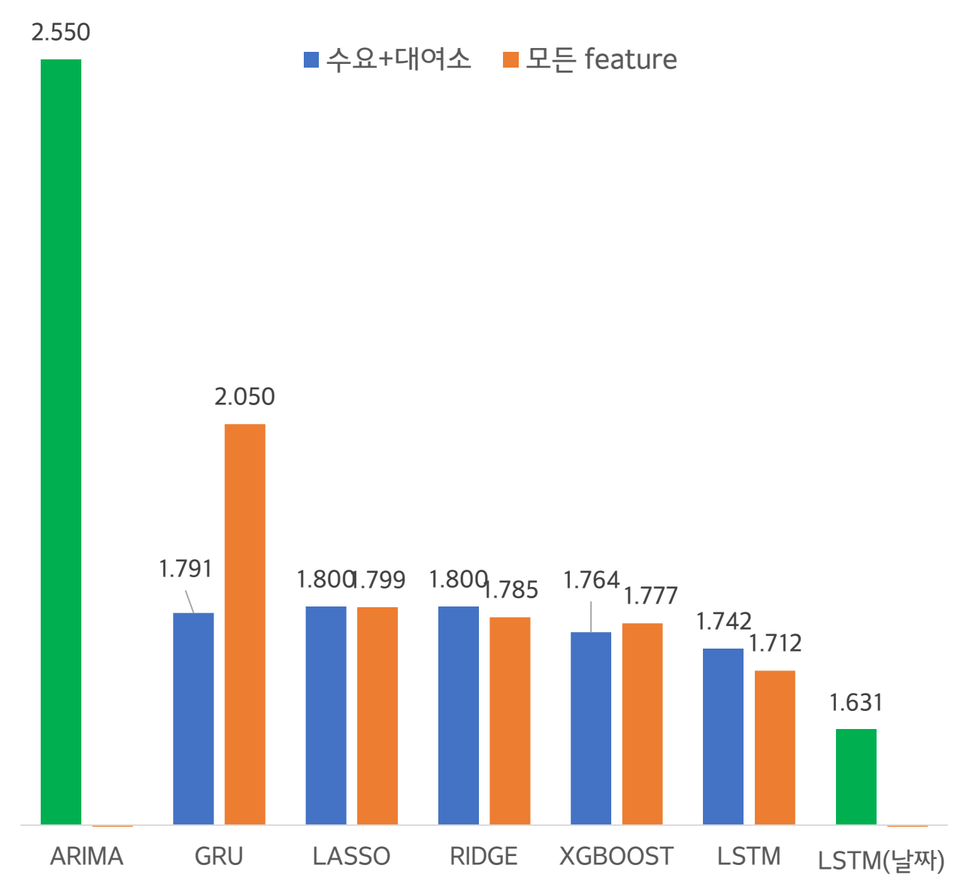
실제 대여건수가 20 이상인 경우 비교



- 큰 값의 예측력은 낮아지지만 상대적으로 낮은 값을 조금 높이는 경향을 보이고 있음

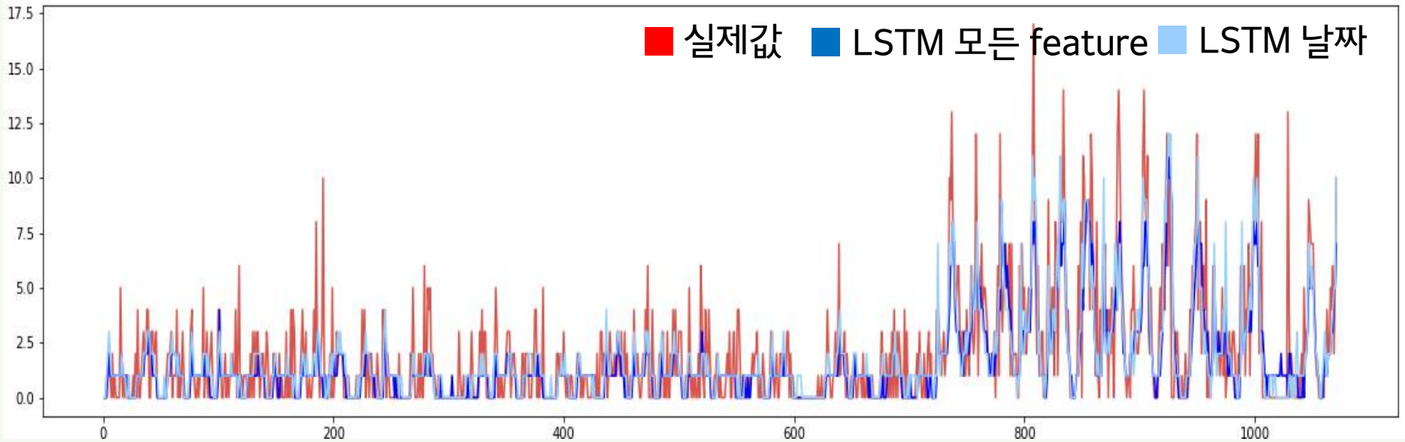
수요예측 결과 - STEP 3 : 모든 feature를 반영한 최종 모델

최종 모델 RMSE 값 비교



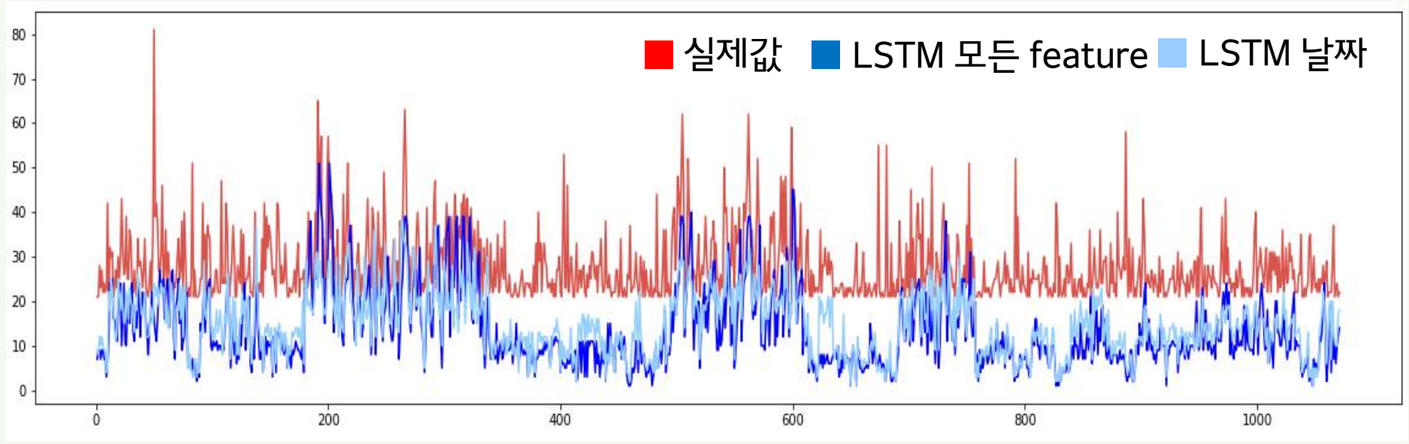
- 모든 변수를 넣은 모델들보다
  LSTM 날짜만을 넣은 모델의 성능이 가장 좋음

시작점부터 1000번째 까지 결과 비교



- 작은 값에 대해서는 예측력이 떨어짐

실제 대여건수가 20 이상인 경우 비교



- 큰 값에 대해서는 예측력이 나아짐

결론

한계점

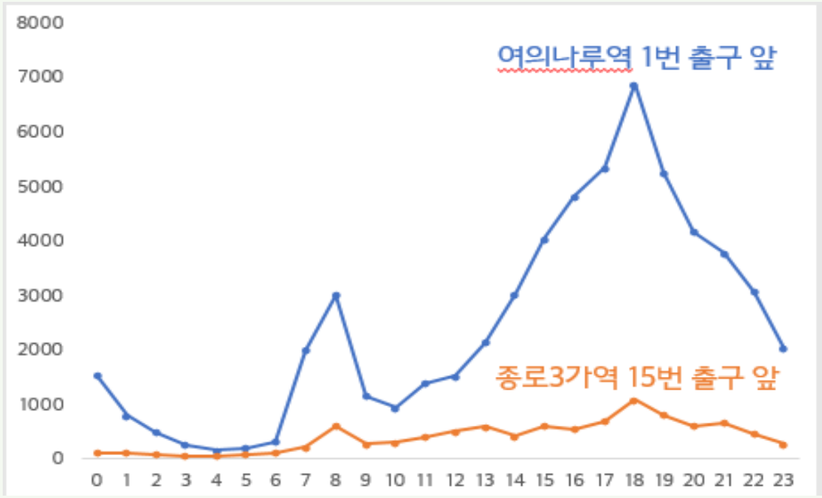
1. 일반화 모델 설계의 오류

- 대여소 특징을 구분하는 feature가 유의미하지 않았음

  → 일반화 모델의 성능 저하

- 자전거 이용 목적을 고려

  → 주변 건물과 같은 추가적인 feature 고려 필요



1. 유의미한 Feature 선정 및 engineering

- 카테고리가 아닌 세분화된 feature별로 유의미한지 판단 필요
- 날씨 현천코드, 주 이용연령 등에 대한 세부사항 고려 필요

대분류	소분류
날씨	풍속
	기온
	기압
	누적강수
	습도
	…

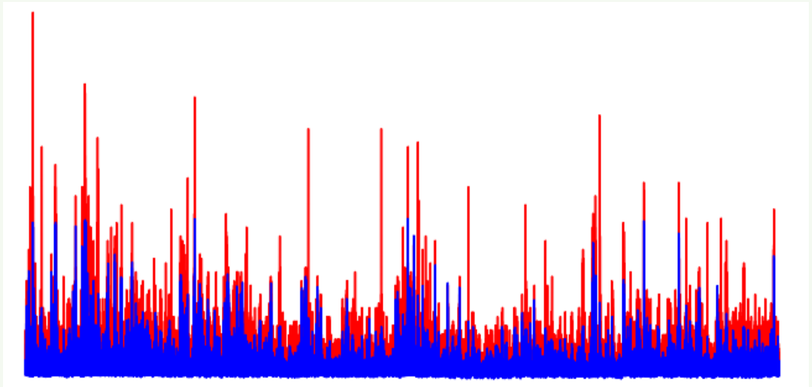
앞으로 유의미한지 판단해야 할 단위

현재 유의미한지 판단해본 단위

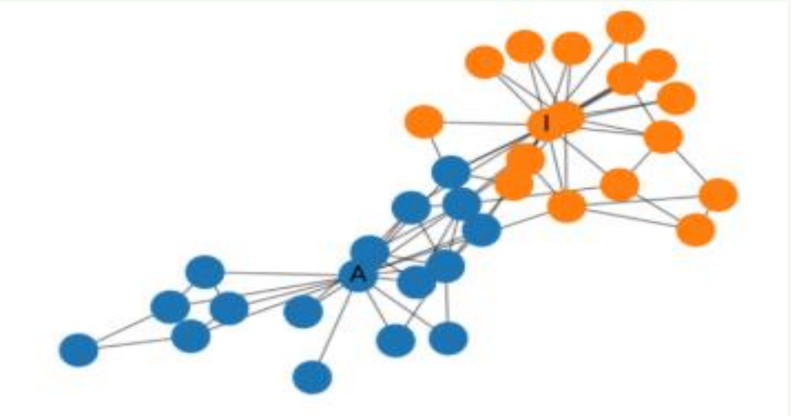
문제점 및 개선방향

1. 모델 평가지표 선정 – 프로젝트 목적에 맞는 방향성
   - 대여건수가 대부분 0 ~ 10 사이의 범위에 속함
   - 모든 값의 정확한 예측보다

     대여건수가 급증하는 경우에 대한 예측이 중요



2. 대여소별 특징을 고려하는 모델 선정
   - 군집분석 (K-means, DTW 등)을 통해 특성에 맞는 모델을 각각 설계
   - GCN과 같은 대여소 간 상관관계를 고려한 모델 설계



Q & A

THANK YOU