

TEAM ML 7

Average
is
All you
need



Q Outline

1 팀 소개

2 대회 소개

3 EDA & Feature
Engineering

4 모델링

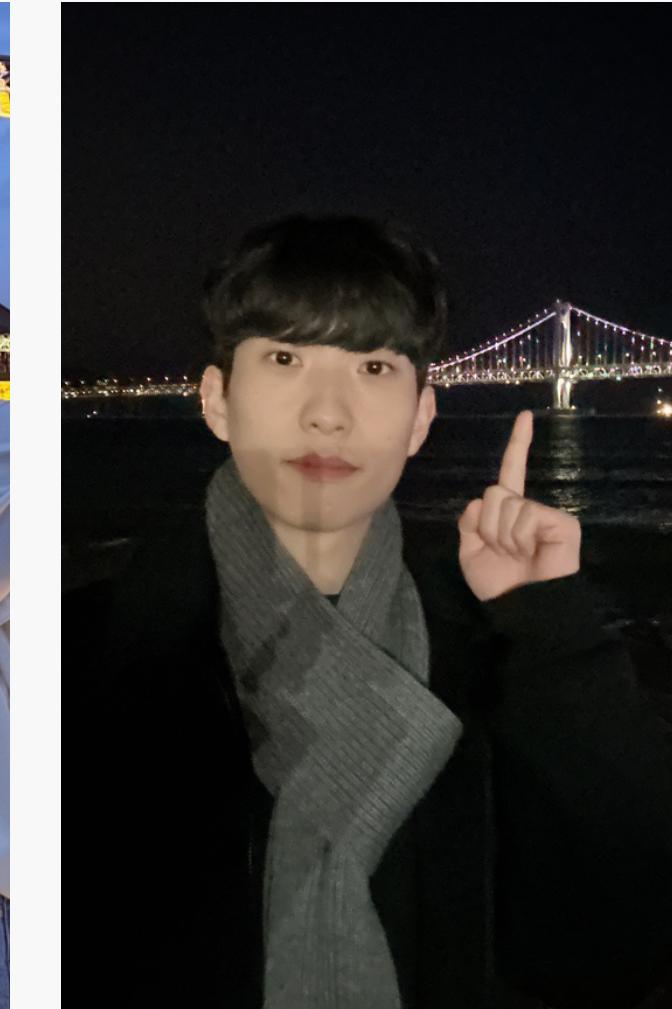
5 결론

6 그룹 스터디 진행 소감

1



팀 소개



이준형

- MBTI : INTJ
- 관심분야 : Multimodal
- 역할 : Feature Engineering & Modeling

이명진

- MBTI : INTP
- 관심분야 : Investing Model
- 역할 : Feature Engineering & Modeling

서재현

- MBTI : ENFJ
- 관심분야 : 컴퓨터비전
- 역할 : Feature Engineering & Modeling

이영훈

- MBTI : ESFJ
- 관심분야 : 추천시스템
- 역할 : Feature Engineering & Modeling

신주용

- MBTI : ISFP
- 관심분야 : LLM & 추천시스템
- 역할 : Feature Engineering & Modeling

2

대회 소개

진행한 대회는 업스테이지 ML Competition인 AI Stages에서 진행한 대회로, 서울 부동산 실거래가 예측 대회입니다.

개요

✓ 데이터셋

- 주제 : 서울 부동산 실거래가 예측 모델
- 기간 : '07년도 1월 1일 ~ '23년 9월 30일의 부동산 데이터
- Feature 수 : 102개 (기존 데이터셋 외에 50개 피쳐 추가)
- 데이터 수 : 1,118,822

목표

- 23년 7~9월 시점의 실거래가를 예측하는 Regression 대회

✓ 평가 지표

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

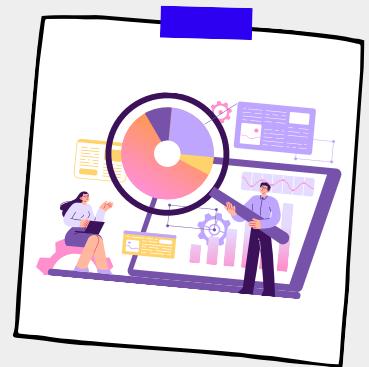
PROS

- 지표 자체가 직관적이며 예측변수와 단위가 같다
- 잔차를 제곱하기 때문에 이상치에 민감
- 잔차를 제곱해서 생기는 값의 왜곡이 MSE에 비해 좀 덜하다.

CONS

- 실제 값에 대해 under/overestimates 인지 파악하기 힘듦
- 스케일에 의존적 (MAE, MSE, RMSE와 통일)

EDA & Feature Engineering



3

EDA & Feature Engineering



건축물대장

건물검색 자치구선택 동선택 지번주소, 도로명주소, 건물명 검색
주용도 선택 승강기유무 선택 지상층수 선택 지하층수 선택 검색

전체 일반 집합

총 599,566건 엑셀다운로드

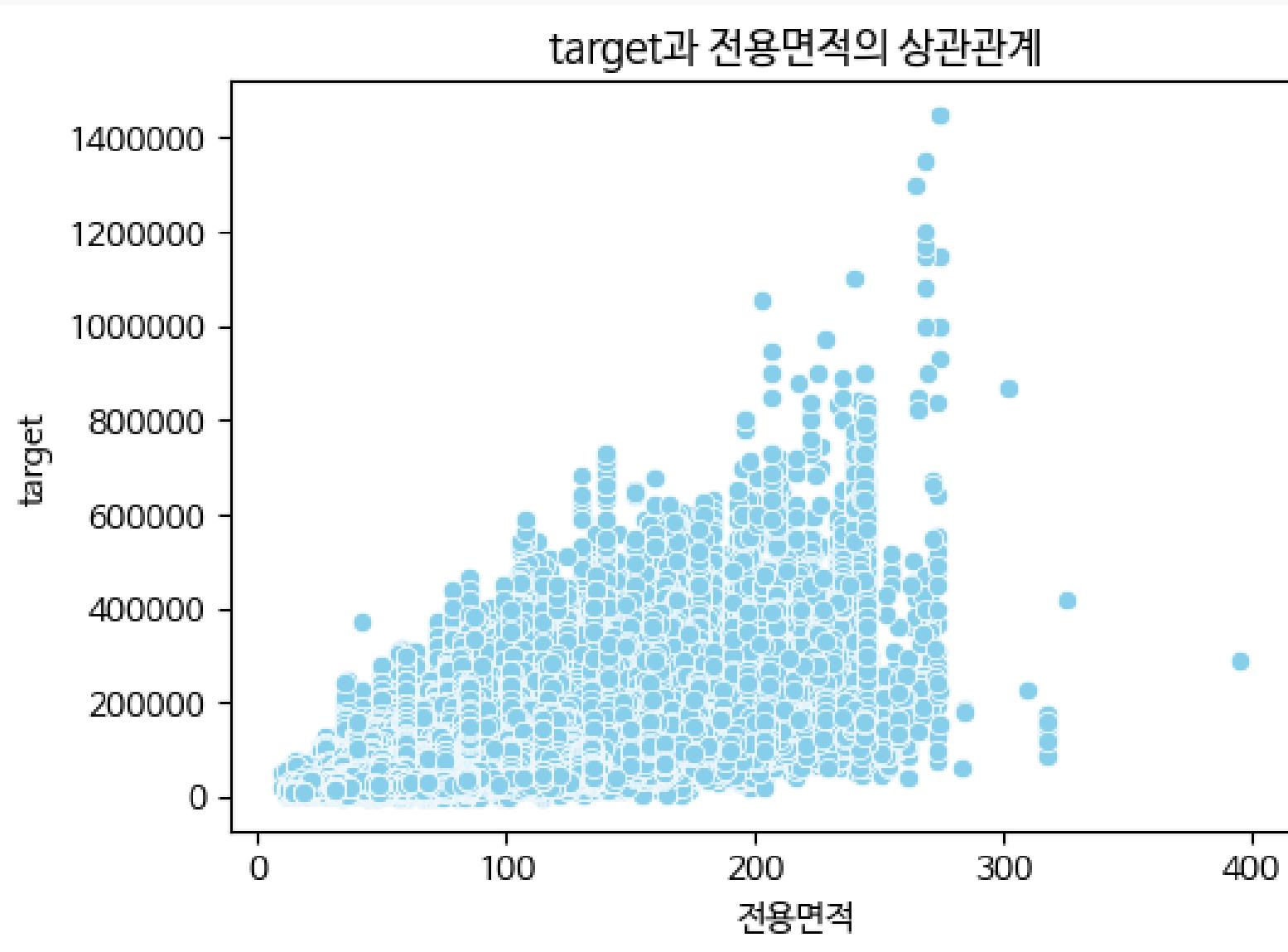
주용도	건물위치	건물명	지상층수	승강기수	연면적(m ²)	건축면적(m ²)
공동주택	서울특별시 송파구 석촌동 16-11번지		5층	0대	474m ²	143m ²
단독주택	서울특별시 송파구 석촌동 16-12번지		3층	0대	530m ²	143m ²
공동주택	서울특별시 송파구 석촌동 16-13번지		5층	0대	608m ²	141m ²
단독주택	서울특별시 송파구 석촌동 16-14번지		3층	0대	371m ²	105m ²
공동주택	서울특별시 송파구 석촌동 16-15번지		6층	0대	417m ²	105m ²
제2종근린생활시설	서울특별시 송파구 석촌동 17번지		3층	0대	637m ²	0m ²
공동주택	서울특별시 송파구 석촌동 17-1번지		5층	0대	493m ²	120m ²

아파트 정보

- k-복도유형, k-난방방식, 시군구, 전용면적, 계약일, 층, 건축년도, k-전체동수, k-전체세대수 등 53개 column이 존재
- 대부분의 결측치가 존재하는 열은 약 100만개 데이터 중 약 87만개가 결측치
- 서울시 **건축물 대장 데이터**에서 도로명 주소를 기반으로 결측치를 채움
- 이외 여러 새로운 feature를 추가

3

EDA & Feature Engineering

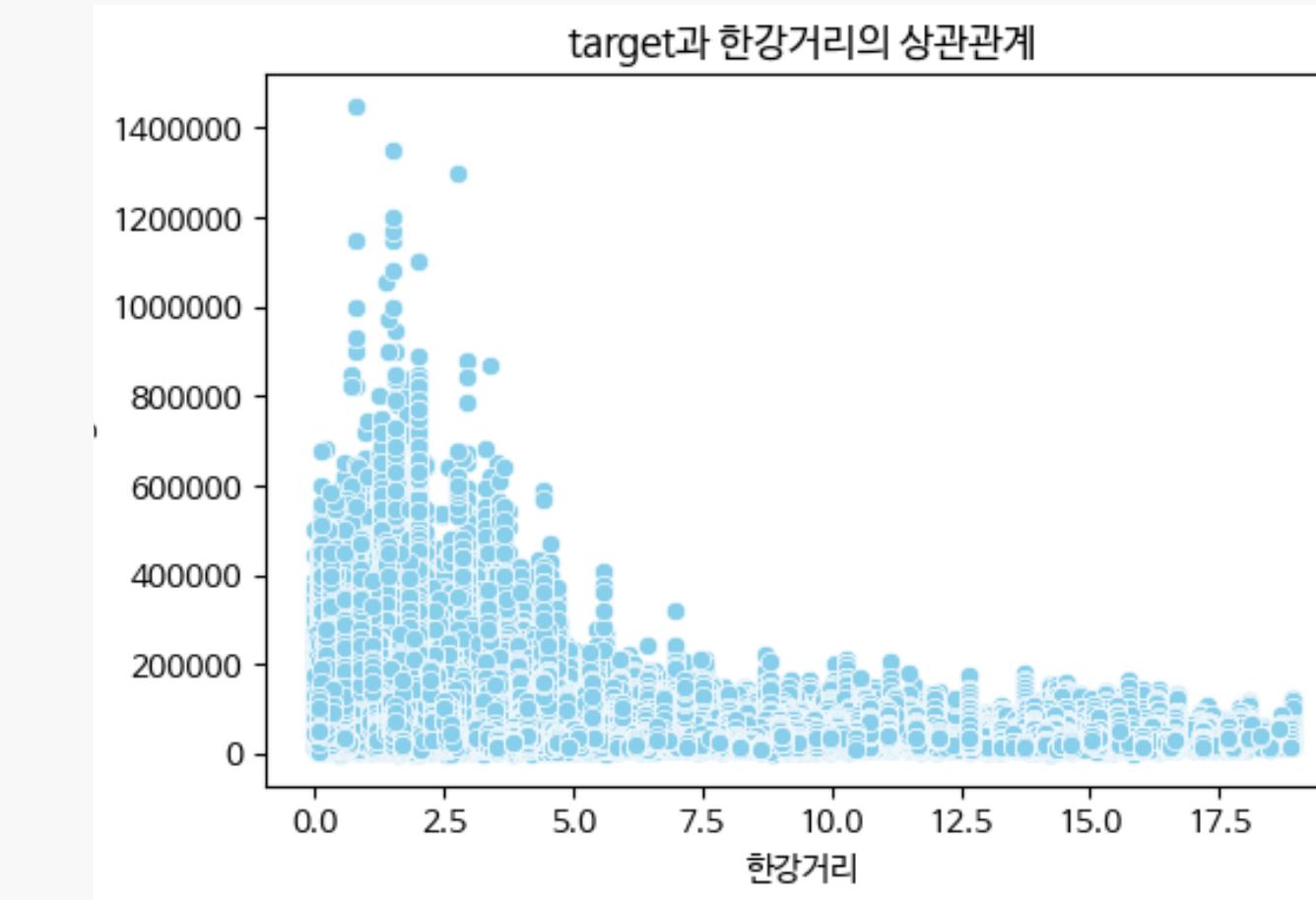


전용면적과 target

- 전용 면적이 클수록 아파트 가격이 상대적으로 높음
- 전용 면적이 작을수록 아파트 가격이 상대적으로 낮음
- 전용 면적과 아파트 가격인 **target**은 서로 양의 상관관계를 가지고 있음

3

EDA & Feature Engineering



한강거리

- 한강 지나는 곳의 사각형의 중앙을 기준으로 한강의 위도(Y) 계산
- 좌상단 위경도(Y/X) : 난지한강공원
- 우하단 위도(Y) : 반포수난구조대
- 우하단 경도(X) : 고덕수변생태공원
- 각 아파트 X, Y를 기준으로 한강의 Y좌표만을 가지고 한강과의 거리 계산
- **한강거리와 target의 분포를 보면,** 서로 **음의 상관관계**를 가짐

3

EDA & Feature Engineering



행정구역(시군구)별(1)	행정구역(시군구)별(2)	2021			
		급여총계		과세대상근로소득(총급여)	
		인원 (명)	금액 (백만원)	인원 (명)	금액 (백만원)
전국	소계	19,959,148	807,198,885	19,907,727	803,208,612
서울	소계	3,952,583	184,858,796	3,944,244	184,066,506
	강남구	215,632	17,477,665	215,441	17,417,230
	강동구	192,080	8,488,374	191,800	8,449,977
	강북구	105,489	3,274,497	105,266	3,259,552
	강서구	252,396	9,964,209	251,917	9,911,518
	관악구	224,352	7,550,177	223,714	7,517,118
	광진구	150,149	6,121,949	149,746	6,096,454
	구로구	174,004	6,504,808	173,678	6,478,283
	금천구	104,225	3,388,533	103,985	3,374,171
	노원구	194,250	7,784,253	193,811	7,745,559
	도봉구	118,133	4,009,934	117,908	3,991,429
	동대문구	129,684	4,913,041	129,267	4,891,094
	동작구	167,538	7,403,320	166,747	7,368,294
	마포구	160,890	8,245,601	160,441	8,210,265
	서대문구	124,643	5,565,047	124,209	5,536,145
	서초구	169,405	13,638,923	169,251	13,591,150
	성동구	118,248	6,207,800	118,079	6,182,497
	성북구	162,378	6,852,822	161,948	6,817,860



구별 1인당 평균급여

- 국가통계포털 KOSIS의 시군구별 근로소득 연말정산 신고현황 데이터 활용
- 금액 / 인원 = 1인당 평균급여(구 별)
- 16~21년도의 서울시 구 별 1인당 평균급여 계산한 피처 추가
- 22~23년 Linear Regression으로 추론

3

EDA & Feature Engineering

**10단위 m²(제곱미터) 환산표**

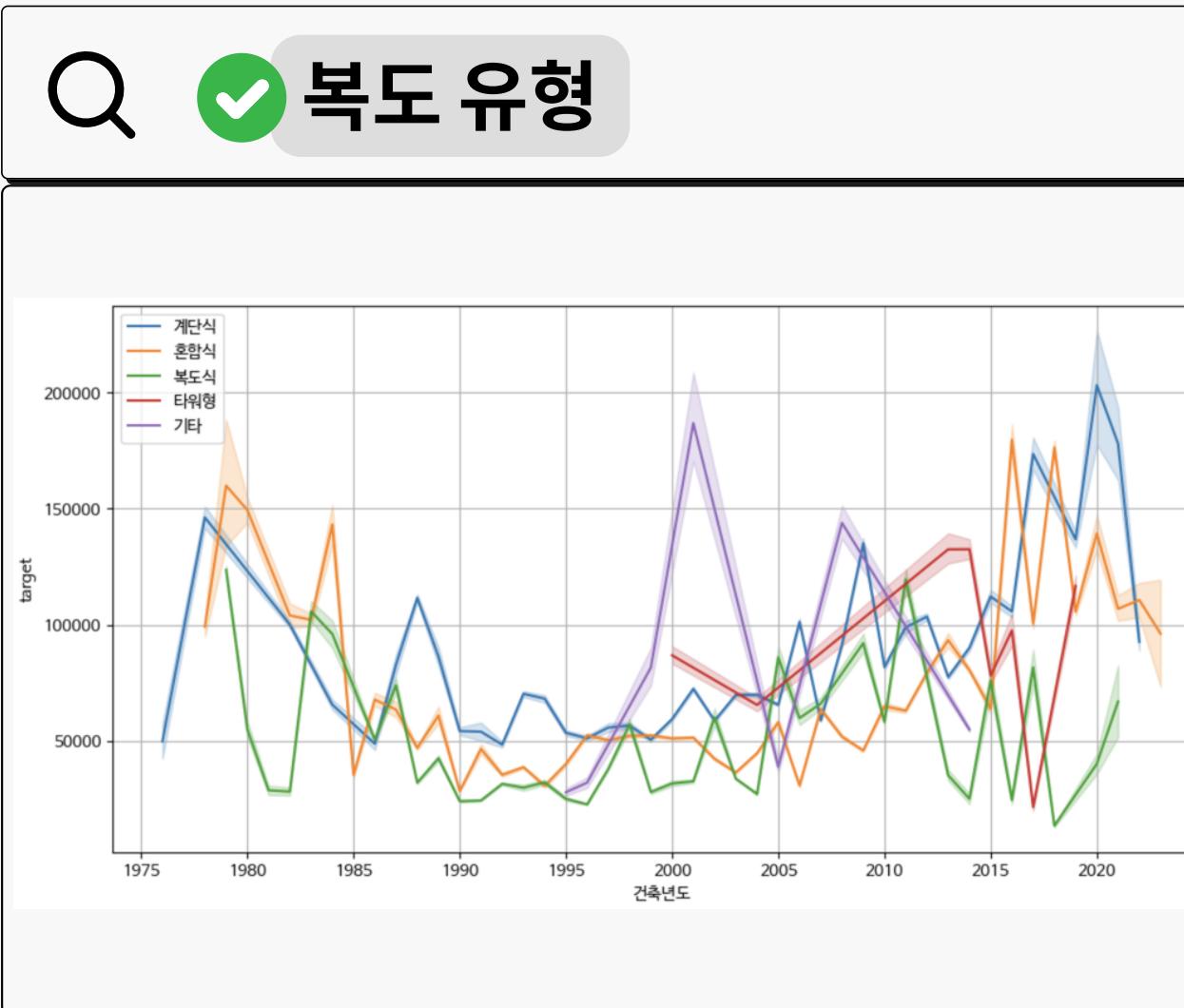
30m ²	약 9평	90m ²	약 27평
40m ²	약 12평	100m ²	약 30평
50m ²	약 15평	110m ²	약 33평
60m ²	약 18평	120m ²	약 36평
70m ²	약 21평	130m ²	약 39평
80m ²	약 24평	140m ²	약 42평
30m ²	약 9평	90m ²	약 27평


✓ 동일 아파트 내 전용면적 탑입비율 계산

- 특정 면적 탑입의 세대 수를 총 세대 수로 나누어 아파트당 탑입 비율을 계산
- (0, 30), (30, 60), (60, 90), (90, 120), (120~) 57개의 범위로 나누어 해당 아파트의 탑입 비율 계산
- 해당 아파트가 속하는 전용면적 비율 추적 가능
(넓은 평수가 많이 해당되어 있는 아파트인지)



3 EDA & Feature Engineering



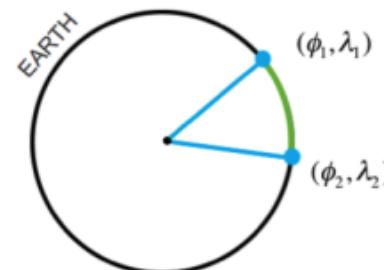
- '복도유형' 카테고리별로 년도에 따라 실거래가 추이가 다르게 나타남
- 이를 통해 '복도유형' 카테고리가 실거래가를 예측하는데 주요한 변수로 작용할 것이라 예상
- 복도유형 별 전용면적의 평균값으로 구간을 나누어서 각 구간에 따라 복도유형의 결측치 값을 채우는 방식으로 진행

3 EDA & Feature Engineering

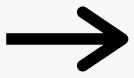


일정 거리 내의 버스 정류장 개수의 총합

$$\text{haversine}\left(\frac{d}{r}\right) = \text{haversine}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{haversine}(\lambda_2 - \lambda_1)$$



- 주변의 교통시설 여부에 따라 집값에 영향이 있을 것이라는 가설
- 외부 버스 데이터와 위도, 경도 값을 이용해서 각 데이터마다 일정 거리 이내에 버스 정류장이 몇 개가 있는지를 기준으로 버스 정류장의 총합을 계산한 피처 생성
- '역세권'의 기준이 500m 이내에 지하철역이 있는가 이므로 버스 정류장의 개수를 세는 기준도 500m 단위로 설정





3 EDA & Feature Engineering

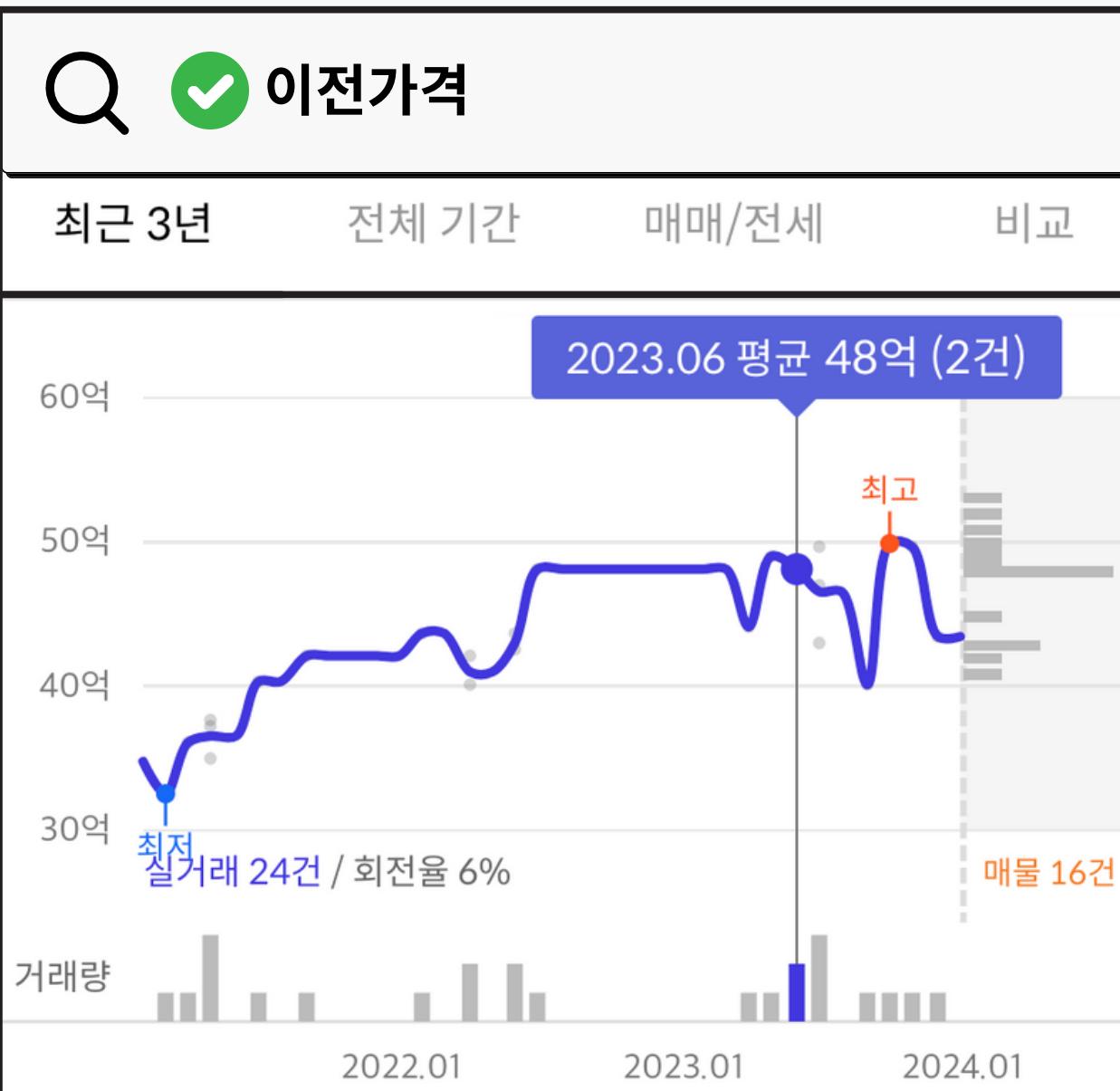
  가장 가까운 학교와의 거리

[Haversine Formula]

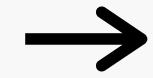
$$d = 2r \arcsin(\sqrt{\text{hav}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{hav}(\lambda_2 - \lambda_1)})$$


- 아파트에 가까운 학교(초,중,고)가 있다면 'target'값에 영향을 줄 것이라 가정
- 따라서 초,중,고등학교에 구분 없이 **가장 가까운 학교와의 거리 피처를 생성**
- 공공데이터의 학교 별 위도, 경도 데이터와 아파트의 위도, 경도 데이터를 haversine distance로 계산하여 dist_to_nearest_school에 저장

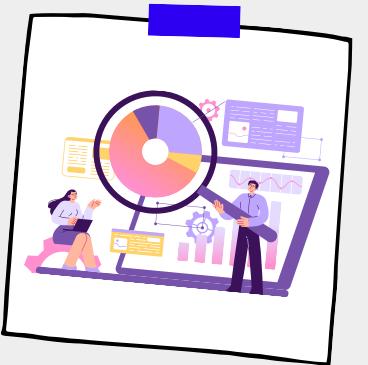
3 EDA & Feature Engineering



- 장기로는 다른 요인의 영향을 많이 받겠지만, 단기로는 **직전 시세**가 가장 중요합니다.
- 결국 우리가 학습하려는 데이터는 최근 3년간 데이터만 학습하기로 했기 때문에 동일 아파트, 동일 면적에 대한 직전 거래 가격을 **이전 가격으로 추가했습니다**.
- 해당 이전 가격과 **target**간의 상관관계는 0.98로 매우 높았습니다.
- 특히, 고가 아파트의 경우에는 저가 아파트와 다르게 비슷한 특성이 없기 때문에 해당 컬럼을 이용하는게 **매우 주요했습니다**.



Modeling





4 Modeling



Train-Validation Split

- 부동산 데이터는 자산 가격 특성상 외부적 요인에 영향을 매우 많이 받을 가능성이 높습니다.
 - 실제로 대통령이 누구인지에 따라서도 가격 상승에 차이가 많이 납니다.
 - 하지만 이런 비정형 데이터(자산 상승 호재, 재건축 짜라기)를 약 20년간에 걸쳐 찾거나 정의하는 것은 어렵습니다.
 - 따라서 자신의 데이터를 잘 맞추는 모델로 진행해보자라는 결론을 도출했습니다.
- Train-validation을 random split으로 하는것이 아니라 2023년 4~6월을 validation으로 선정하고, 이전 데이터를 Train dataset으로 지정했습니다.
- 또한, 학습 자체도 3년 이내의 부동산 데이터만 활용해서, 시기적 특성요소를 최대한 덜 학습하도록 했습니다.
- 진행하는 과정에서 Validation RMSE와 리더보드 RMSE 결과가 비례하지 않아 2023년 1월 ~ 3월을 Valid로, 2023년 4월 ~ 6월은 test 데이터셋으로 구축해서 모델간에 비교를 진행해서 최우수 모델로 선정했습니다.



Modeling



가격대별 별도 모델 구축

- RMSE가 8만 ~ 10만 정도로 나오는데 이렇게나 지표가 높게 나타난다는 것은 결국 비싼 아파트를 싸게 예측하는 경우밖에 없다고 판단했습니다.
- 실제로 valid data에서 예측을 잘 못하는 모델을 보아도 90억짜리 아파트를 28억으로 예측하는 것을 볼 수 있습니다.
- 따라서 최종 모델은 30억 이상 고가 아파트를 예측하는 모델과 그 이하 가격대의 아파트를 예측하는 모델을 분리하고 답을 합치는 형태로 진행했습니다.
 - 30억 이상 고가는 것을 테스트 데이터에서는 확인할 수 없기 때문에 2020.01~2023.06 동안 30억 이상으로 거래된 아파트들을 모두 가져와서 트레이닝 데이터셋으로 구축했고, 테스트 데이터셋에서는 이 아파트명을 가진 index만 예측했습니다.
 - 실제 프로덕션에서 이 모델을 활용하기 위해서는 별도로 고가 아파트라는 데이터셋을 구축해야한다는 한계가 있습니다.

4

Modeling



저가 아파트 모델링

- target 300,000 미만 데이터 사용
- 학습 데이터 : 2020년 01월 ~ 2022년 12월
- 검증 데이터 : 2023년 01월 ~ 2023년 03월
- 평가 데이터 : 2023년 04월 ~ 2023년 06월
- 모델 : LightGBM
- Hyper-Parameter tuning : Optuna



결과

- LightGBM 하이퍼파라미터 튜닝 전
 - Train Score : 11129.8078
 - Valid Score : 13101.2657
 - Test Score : 14701.9578
- LightGBM optuna 하이퍼파라미터 튜닝 후
 - Train Score : 4412.8290
 - Valid Score : 9704.2133
 - Test Score : 12333.9627
- 최종 학습 및 2023.04~2023.06 최종 테스트
 - Test Score : 10015.5281

4

Modeling



✓ 고가 아파트 모델링

- 30억 가격 초과 아파트 데이터만 사용
- 학습 데이터 : 2020년 01월 ~ 2022년 12월
- 검증 데이터 : 2023년 01월 ~ 2023년 03월
- 평가 데이터 : 2023년 04월 ~ 2023년 06월
- 모델 : LightGBM
 - num_estimators 수 비교로만 진행

✓ 결과

- 사용 변수
 - '꼭대기층 여부', '이전가격', '전용면적', '아파트 평균높이', '연GDP', '층', '계약년월', 'y', '한강거리', '500m이내 정류장 수', '건물나이'
- 최종 학습 및 2023.04~2023.06 최종 테스트
 - num_estimators 수 : 410개
 - Test Score : 30572.5281

4

Modeling



최종 제출 모델

- 저가 아파트 모델 예측값 & 고가 아파트 모델 예측값
- 저가 아파트 데이터 : 2020년 01월 ~ 2023년 03월
- 고가 아파트 데이터 : 2020년 01월 ~ 2023년 06월
- 평가 데이터 : 2023년 07월 ~ 2023년 09월
- 모델 : LightGBM
- Hyper-Parameter tuning : WandB

결과

- <저가 아파트 모델>
 - Train RMSE : 4958.61
 - Valid RMSE : 9761.01
- <고가 아파트 모델>
 - Train RMSE : 5493.75
 - Valid RMSE : 26855.08
- <Public 결과>
 - Public RMSE : 104262.6919
- <Private 결과>
 - Private RMSE : 104262.6919



5 마의 RMSE 10만...



RMSE 10만이란..?

- 저희 팀은 다양한 피쳐, 최적화 등으로 모델을 돌려도 항상 전용면적, 가격만으로 학습한 모델의 RMSE인 88000을 이기지 못하고 오히려 10만이라는 스코어를 깨지 못했습니다.
- Validation으로 88000을 기록한 모델과 비교해보더라도 훨씬 성능이 좋았지만 실제 리더보드만 보면 100000을 넘지 못했습니다.
- RMSE 10만은 모든 결측치에 대해서 10억만큼 틀린다는 의미입니다.
- 공교롭게도 서울특별시의 2023년 4~6개월 평균 아파트 실거래가는 10억입니다.
 - 물론 오피스텔이기 때문에 다르고 하지만 아파트 실거래가가 모두 10억이라고 했을 때 정답으로 모두 0원으로 예측하면 RMSE가 10만이 나온다고 볼 수 있습니다.
- 그래서 저희 팀은 테스트 데이터가 4600개로 적은편인데 최적화해도 10억씩 틀린다는 건 결국 엄청 비싼 아파트에 대해서 모델이 완전히 못맞추고 있다고 판단했습니다.(88000 스코어 모델도 가장 비싼 가격으로 30억을 예측하는데 만일 180억짜리 모델을 못맞춘다면? -> 20억을 10억으로 1개 잘못 예측하는 것 기준으로 약 150배 loss를 줌)

5

예상되는 원인 1: 비싼 아파트 가격을 잘 못맞추나?

Case 1: 고가 아파트 예측 모델이 40억 이상이라고 예측한 아파트들 96개 + 40억 미만은 리더보드상 Best Model로 예측

Open Leaderboard RMSE : 88377 → 94632 (오히려 개당 6억씩 못맞추는 걸로 결과가 나옴)

Case 2: 실제 고가 아파트 예측 모델이 40억 이상이라고 예측한 아파트들의 예측이 잘못되었나?

40억 이상이라고 예측한 아파트 96개에 대한 7~9월 실거래가 RMSE 계산시 : 46만 → 9.3만(오히려 내려감)

*아파트 96개에 대한 7~9월 실 거래가는 국토교통부 데이터를 제공하는 호갱노노, 네이버 부동산에서 전용면적, 계약일과 일치하는 데이터로 수동 조사

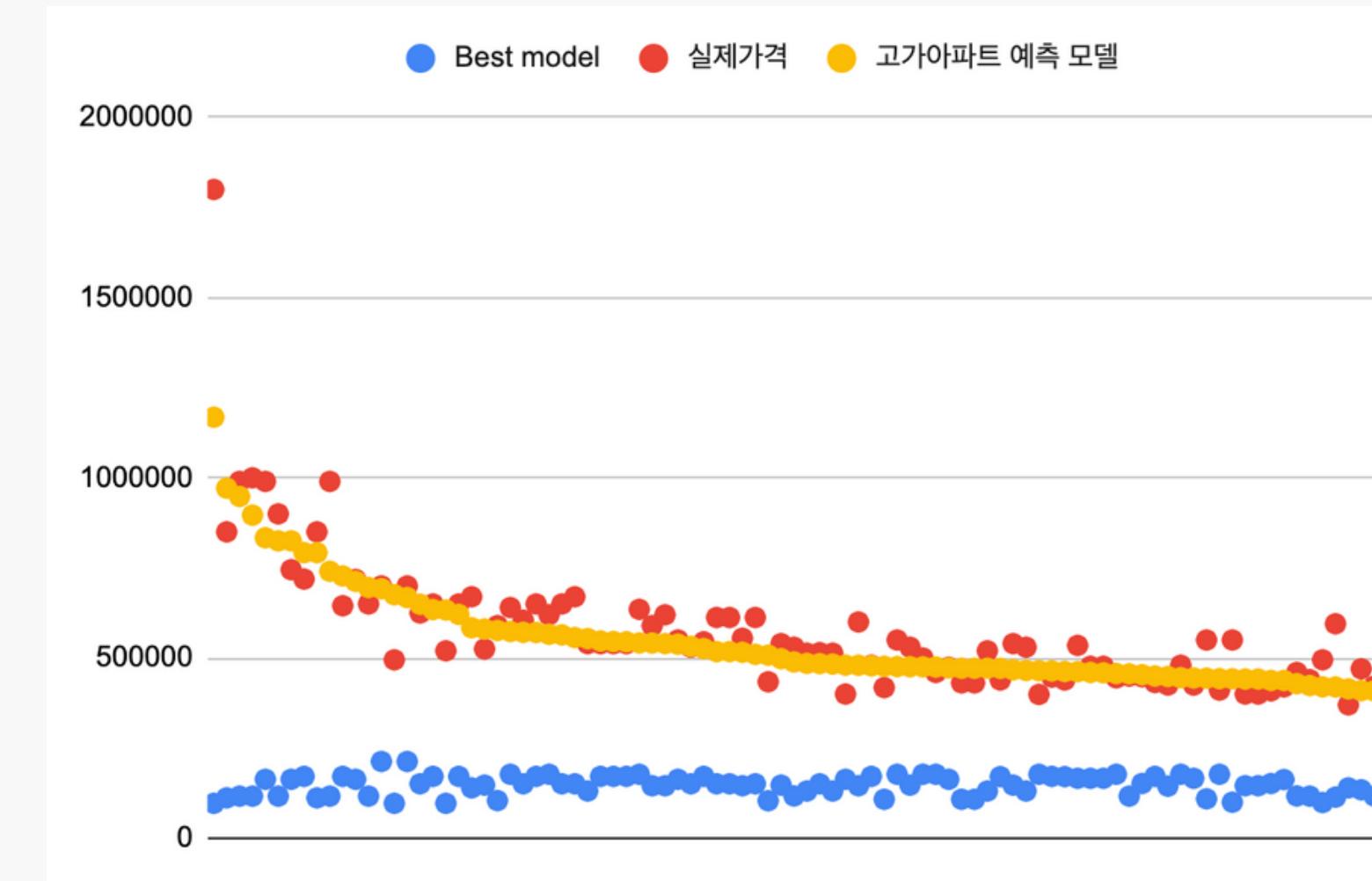
Case 3: 실제 고가 아파트 예측 모델이 40억 이상이라고 예측한 아파트들의 실거래가 96개 + 40억 미만은 Best Model로 예측

Open Leaderboard RMSE : 88377 → 96164 (오히려 개당 8억씩 못맞추는 걸로 결과가 나옴)

진짜 정답인 실거래가로 넣었는데 예측값보다 오픈리더보드보다.. 점수가 떨어진다...?

5

예상되는 원인 1: 비싼 아파트 가격을 잘 못맞추나?



아파트명	고가아파트 예측 모델	(리더보드 기준) Best model	실제가격
파크한남	116.8억	9.6억	180억
한남더힐	971366	112038	850000
아크로써울포레스트	948231	116794	990000
갤러리아포레	896390	116794	1000000
타워팰리스1	833655	163500	990000
갤러리아포레	825110	116794	900000
타워팰리스1	825101	163500	745000
반포자이	792410	171594	719000
한남더힐	792318	112038	850000



예상되는 원인 2 : 테스트 데이터셋의 인덱스가 꼬인걸까...?

Case 4 : 모든 정답에 2023년 1월 ~ 6월에 일어난 모든 실거래가의 평균을 넣어보자.

Open Leaderboard RMSE : 88377 → 86148 (오히려 감소!!! 등수도 4등에서 2등으로 증가!!)

Case 5 : 모든 정답에 2023년 1월 ~ 3월에 일어난 모든 실거래가의 평균을 넣어보자.

Open Leaderboard RMSE : 88377 → 86401 (약간 감소)

Case 6 : 모든 정답에 2023년 4월 ~ 6월에 일어난 모든 실거래가의 평균을 넣어보자.

Open Leaderboard RMSE : 88377 → 86200 (약간 감소)

앞서 고가 예측아파트를 더 잘 맞춘 모델보다.. 평균값을 넣은게 성능이 더 좋다..?



결론 : 대회의 테스트 데이터셋에 문제가 있는 것으로 판단했습니다 !

만일 인덱스와 정답이 잘 매칭되어 있었다면, 고가 아파트 데이터 96개에 정답을 바꿔서 넣었어도 지표가 떨어졌거나 변함이 없었을 것입니다.

4636(9272개 데이터 중 50%)개 중에 우연하게 고가 아파트 데이터 96개 모두 포함이 안되었다면?

오픈 리더보드 점수에는 변함이 없었을 것입니다.

하지만 점수에 변화가 일어났는데 그 점수가 오히려 나빠졌다면?

점수의 변화는 4636개에 고가 아파트가 포함되어있다는 것을 의미하는데 점수가 오르는 것은 수학적으로 불가능합니다.

전수조사 결과, 저희가 예측한 96개에 대해서 모두 best model 대비해서 RMSE가 모두 낮았으며 평균적으로 1/30배 낮았습니다. (9만, 46만)

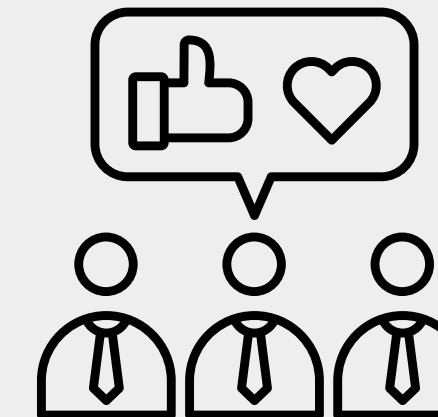
팀에서 확인하지 못한 다른 원인이 있을 수도 있으나, 이를 대회 마지막 날에 파악하게 되어서
위 가설 검증만 마무리하고 최종적으로는 고가 아파트 모델 혼합 & 리더보드 1등 모델로 제출했습니다.
최종 등수로 저희 팀은 2등을 차지했습니다. 그런데...

5 결론 : 대회의 테스트 데이터셋에 문제가 있는 것으로 판단했습니다 !

2등 제출 결과

프로젝트 진행 소감

Big
Honey
Jam





이준형 : 도메인에 대한 이해가 가장 중요하다는 생각이 드는 대회였습니다. 또한, 데이터 전처리 & 외부 데이터 연동을 하는데 프로젝트 대부분의 시간을 쓸았는데 그 과정에서 다음에는 좀 더 실수 없이 효율적으로 처리하는 방법에 대해서 연구해봐야겠다는 생각이 드는 대회였습니다

이영훈 : 결측치나 이상치 탐색 과정에서 '왜?'라는 부분에 집중해서 고민해 보았지만 생각보다 많은 결측치 때문에 쉽지 않았던 대회였던 것 같습니다. 다양한 피처를 생성하고 여러 모델을 실험하는 과정에서 좋은 팀원분들과 의견을 공유하며 많이 배우고 성장한 것 같아 의미있는 시간이었습니다.

서재현 : 시계열 데이터에 대해 시간적 특성을 어떻게 반영해야 하는지 고민하게 되는 대회였습니다. 또한 데이터에 대한 결측치와 이상치가 많아 데이터 분석이 어려웠고, 다양한 외부 데이터를 이용하기 위해 도메인에 대한 지식이 많이 요구되었습니다. 팀원들과 여러 의견을 주고 받으며 여러 인사이트를 얻을 수 있는 좋은 시간이었습니다.

신주용 : 피처엔지니어링에 집중하면서 팀원들의 인사이트를 보며 배운 점이 많았고, 결측치를 처리하기 위에 외부 데이터를 탐색하는 과정도 재미있었습니다. 다만 여러 피처들을 학습에 사용했으나, PB에서의 점수 향상으로 이어지지 않아 아쉬웠습니다.

이명진 : 먼저, 도메인 지식의 중요성이 강조되었던 대회였던 것 같습니다. 부동산 시장에 대한 이해가 없으면 데이터를 올바르게 해석하고 활용하는 데 어려움이 있었습니다. 또한, 많은 시간을 데이터 전처리와 외부 데이터 연동에 할애했는데, 이 과정에서 효율성과 정확성을 높이는 방법을 모색해야 했습니다. 이러한 데이터 문제들을 해결하기 위해 '왜?'라는 질문에 집중해야 했습니다. 또한, 다양한 피처를 생성하고 여러 모델을 실험하는 과정에서 팀원들과 의견을 공유하며 많은 것을 배웠습니다. 마지막으로, 시계열 데이터의 처리에 대한 고민으로 시간적 특성을 어떻게 반영할지에 대한 지식이 필요하였고 이와 관련해 팀원들과의 논의를 통해 여러 인사이트를 얻을 수 있었던 유익한 경험이었습니다.



Thank you!



경청해주세요
감사합니다!