Upstage AI Lab
ML Advanced 경진대회

팀 이름 **ML_4조**
팀원 이름 : 유현지,
김정헌, 박주혁, 지수영,
김민혁

Content

01

팀원 소개

팀원 소개 (1)

이름 : 김정헌

약력

컴공 2학년 수료 후 휴학

팀원 소개 (2)

이름: 지수영

약력

에너지자원공학과 학사

팀원 소개 (3)

이름 : 유현지

약력

경영 & 인공지능융합 학사

팀원 소개 (4)

약력

팀원 소개 (5)

약력

02

대회 소개

대회 개요

대회

서울시 아파트 실거래가 매매 데이터를 기반으로 아파트 가격을 예측하는 대회

목표

목표
House Price Prediction 경진대회는 주어진 데이터를 활용하여 서울의 아파트 실거래가를 효과적으로 예측하는 모델을 개발

개요

소개 및 배경 설명
UpstageAiLab 에서 배운 MachinLearning 을 실제 대회에서 사용해보는 대회

기간
2024. 03. 19 ~ 2024. 04.02

평가지표

사진 및 내용을 입력해주세요.

평가지표

사진 및 내용을 입력해주세요.

제출 파일

사진 및 내용을 입력해주세요.

03

Data Description

데이터 설명

- 데이터 형태: csv 파일
- 데이터 기간: train(2007.01.01~2023.06.30), test(2023.07.01~2023.09.26)
- 데이터 개수: train(1118822개), test(9272개)
- 아파트 정보에 대한 변수: 52개(+ 거래시점에 대한 변수)
- 서울시 지하철역, 버스 정류장에 대한 데이터 추가

	시군구	번지	본번	부번	아파트 명	전용면적 (m²)	계약년월	계약 일	층	건축년 도	...	건축면 적	주차대 수	기타/의무/임대/임의 =1/2/3/4	단지승인일	사용허가 여부	관리비 업로드	좌표X	좌표Y	단지신청일	target
0	서울특별시 강남구 개포동	658-1	658.0	1.0	개포6차 우성	79.97	201712	8	3	1987	...	4858.0	262.0	임의	2022-11-17 13:00:29.0	Y	N	127.057210	37.476763	2022-11-17 10:19:06.0	124000
1	서울특별시 강남구 개포동	658-1	658.0	1.0	개포6차 우성	79.97	201712	22	4	1987	...	4858.0	262.0	임의	2022-11-17 13:00:29.0	Y	N	127.057210	37.476763	2022-11-17 10:19:06.0	123500
2	서울특별시 강남구 개포동	658-1	658.0	1.0	개포6차 우성	54.98	201712	28	5	1987	...	4858.0	262.0	임의	2022-11-17 13:00:29.0	Y	N	127.057210	37.476763	2022-11-17 10:19:06.0	91500
3	서울특별시 강남구 개포동	658-1	658.0	1.0	개포6차 우성	79.97	201801	3	4	1987	...	4858.0	262.0	임의	2022-11-17 13:00:29.0	Y	N	127.057210	37.476763	2022-11-17 10:19:06.0	130000
4	서울특별시 강남구 개포동	658-1	658.0	1.0	개포6차 우성	79.97	201801	8	2	1987	...	4858.0	262.0	임의	2022-11-17 13:00:29.0	Y	N	127.057210	37.476763	2022-11-17 10:19:06.0	117000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1118817	서울특별시 은평구 구산동	382	382.0	0.0	갈현현대	59.94	200707	12	11	1998	...	0.0	366.0	의무	2013-06-04 16:18:51.0	Y	N	126.905638	37.612962	2013-03-07 09:46:27.0	20000
1118818	서울특별시 은평구 구산동	382	382.0	0.0	갈현현대	59.94	200708	25	10	1998	...	0.0	366.0	의무	2013-06-04 16:18:51.0	Y	N	126.905638	37.612962	2013-03-07 09:46:27.0	20000
1118819	서울특별시 은평구 구산동	382	382.0	0.0	갈현현대	84.83	200708	31	20	1998	...	0.0	366.0	의무	2013-06-04 16:18:51.0	Y	N	126.905638	37.612962	2013-03-07 09:46:27.0	28000
1118820	서울특별시 은평구 구산동	382	382.0	0.0	갈현현대	84.83	200709	15	8	1998	...	0.0	366.0	의무	2013-06-04 16:18:51.0	Y	N	126.905638	37.612962	2013-03-07 09:46:27.0	29000
1118821	서울특별시 중구 묵정동	11-67	11.0	67.0	묵정	52.46	200701	10	5	1981	...	7354.0	45.0	임의	2020-07-10 00:00:00.0	Y	Y	127.000071	37.560706	2017-09-05 20:06:39.0	13250

1118822 rows × 52 columns

	역사_ID	역사명	호선	위도	경도
0	9996	미사	5호선	37.560927	127.193877
1	9995	강일	5호선	37.557490	127.175930
2	4929	김포공항	김포골드라인	37.562360	126.801868
3	4928	고촌	김포골드라인	37.601243	126.770345
4	4927	풍무	김포골드라인	37.612488	126.732387
...	...	...	...	...	...
763	154	종로5가	1호선	37.570926	127.001849
764	153	종로3가	1호선	37.570406	126.991847
765	152	종각	1호선	37.570161	126.982923
766	151	시청	1호선	37.565715	126.977088
767	150	서울역	1호선	37.556228	126.972135

768 rows × 5 columns

지하철역 추가 데이터

	노드 ID	정류소번호	정류소명	X좌표	Y좌표	정류소 타입
0	100000001	1001	종로2가사거리	126.987752	37.569808	중앙차로
1	100000002	1002	창경궁.서울대학교병원	126.996566	37.579183	중앙차로
2	100000003	1003	명륜3가.성대입구	126.998251	37.582581	중앙차로
3	100000004	1004	종로2가.삼일교	126.987613	37.568579	중앙차로
4	100000005	1005	혜화동로터리.여운형활동터	127.001744	37.586243	중앙차로
...	...	...	...	...	...	...
12579	124000334	25995	우성아파트	127.139338	37.550386	일반차로
12580	124000333	25996	우성아파트	127.140046	37.550643	일반차로
12581	124000332	25997	조일약국	127.123596	37.533630	일반차로
12582	124000331	25998	성내시장	127.125497	37.536155	일반차로
12583	124000330	25999	천호우체국.로데오거리	127.127337	37.540343	일반차로

12584 rows × 6 columns

버스정류장 추가 데이터

데이터의 기초 통계 및 정보 요약

```
#   Column                    Non-Null Count    Dtype
--  ------                    --------------    -----
0   시군구                      1118822 non-null  object
1   번지                       1118597 non-null  object
2   본번                       1118747 non-null  float64
3   부번                       1118747 non-null  float64
4   아파트명                     1116696 non-null  object
5   전용면적(㎡)                  1118822 non-null  float64
6   계약년월                     1118822 non-null  int64
7   계약일                      1118822 non-null  int64
8   층                        1118822 non-null  int64
9   건축년도                     1118822 non-null  int64
10  도로명                      1118822 non-null  object
11  해제사유발생일                    5983 non-null     float64
12  등기신청일자                   1118822 non-null  object
13  거래유형                     1118822 non-null  object
14  중개사소재지                    1118822 non-null  object
15  k-단지분류(아파트,주상복합등등)       248131 non-null   object
16  k-전화번호                    248548 non-null   object
17  k-팩스번호                    246080 non-null   object
18  단지소개기존clob                68582 non-null    float64
19  k-세대타입(분양형태)              249259 non-null   object
20  k-관리방식                    249259 non-null   object
21  k-복도유형                    248932 non-null   object
22  k-난방방식                    249259 non-null   object
23  k-전체동수                    248192 non-null   float64
24  k-전체세대수                   249259 non-null   float64
25  k-건설사(시공사)                247764 non-null   object
26  k-시행사                     247568 non-null   object
27  k-사용검사일-사용승인일             249126 non-null   object
28  k-연면적                     249259 non-null   float64
29  k-주거전용면적                  249214 non-null   float64
30  k-관리비부과면적                 249259 non-null   float64
31  k-전용면적별세대현황(60㎡이하)       249214 non-null   float64
32  k-전용면적별세대현황(60㎡~85㎡이하)  249214 non-null   float64
33  k-85㎡~135㎡이하              249214 non-null   float64
34  k-135㎡초과                 327 non-null      float64
35  k-홈페이지                    113175 non-null   object
36  k-등록일자                    10990 non-null    object
37  k-수정일자                    249214 non-null   object
38  고용보험관리번호                   205518 non-null   object
39  경비비관리형태                   247834 non-null   object
40  세대전기계약방법                  240075 non-null   object
41  청소비관리형태                   247644 non-null   object
42  건축면적                     249108 non-null   float64
43  주차대수                     249108 non-null   float64
44  기타/의무/임대/임의=1/2/3/4       249259 non-null   object
45  단지승인일                    248536 non-null   object
46  사용허가여부                    249259 non-null   object
47  관리비 업로드                   249259 non-null   object
48  좌표X                      249152 non-null   float64
49  좌표Y                      249152 non-null   float64
50  단지신청일                     249197 non-null   object
51  target                    1118822 non-null  int64
dtypes: float64(18), int64(5), object(29)
memory usage: 443.9+ MB
```

진 및 내용을 입

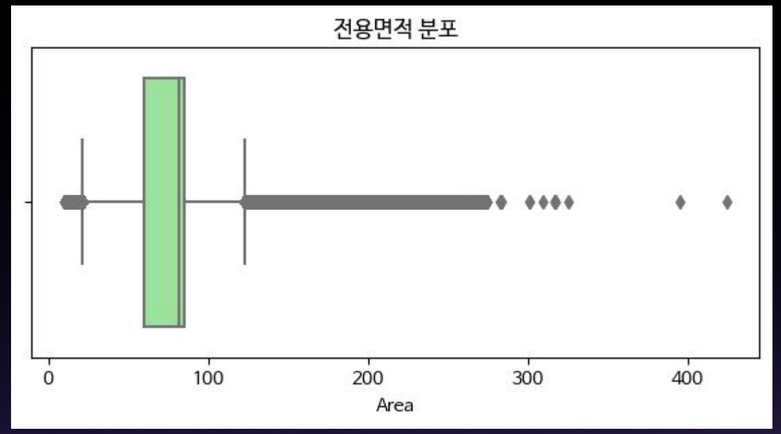
```
Data columns (total 51 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   시군구                       9272 non-null   object
 1   번지                        9270 non-null   object
 2   본번                        9272 non-null   float64
 3   부번                        9272 non-null   float64
 4   아파트명                      9262 non-null   object
 5   전용면적(㎡)                   9272 non-null   float64
 6   계약년월                      9272 non-null   int64
 7   계약일                       9272 non-null   int64
 8   층                         9272 non-null   int64
 9   건축년도                      9272 non-null   int64
 10  도로명                       9272 non-null   object
 11  해제사유발생일                     212 non-null    float64
 12  등기신청일자                    9272 non-null   object
 13  거래유형                      9272 non-null   object
 14  중개사소재지                     9272 non-null   object
 15  k-단지분류(아파트,주상복합등등)        2690 non-null   object
 16  k-전화번호                     2696 non-null   object
 17  k-팩스번호                     2666 non-null   object
 18  단지소개기존clob                 554 non-null    float64
 19  k-세대타입(분양형태)               2710 non-null   object
 20  k-관리방식                     2710 non-null   object
 21  k-복도유형                     2708 non-null   object
 22  k-난방방식                     2710 non-null   object
 23  k-전체동수                     2695 non-null   float64
 24  k-전체세대수                    2710 non-null   float64
 25  k-건설사(시공사)                 2693 non-null   object
 26  k-시행사                      2692 non-null   object
 27  k-사용검사일-사용승인일              2709 non-null   object
 28  k-연면적                      2710 non-null   float64
 29  k-주거전용면적                   2710 non-null   float64
 30  k-관리비부과면적                  2710 non-null   float64
 31  k-전용면적별세대현황(60㎡이하)        2710 non-null   float64
 32  k-전용면적별세대현황(60㎡~85㎡이하)   2710 non-null   float64
 33  k-85㎡~135㎡이하               2710 non-null   float64
 34  k-135㎡초과                  2 non-null      float64
 35  k-홈페이지                     1396 non-null   object
 36  k-등록일자                     718 non-null    object
 37  k-수정일자                     2710 non-null   object
 38  고용보험관리번호                    1819 non-null   object
 39  경비비관리형태                    2699 non-null   object
 40  세대전기계약방법                   2630 non-null   object
 41  청소비관리형태                    2699 non-null   object
 42  건축면적                      2707 non-null   float64
 43  주차대수                      2709 non-null   float64
 44  기타/의무/임대/임의=1/2/3/4        2710 non-null   object
 45  단지승인일                     2704 non-null   object
 46  사용허가여부                     2710 non-null   object
 47  관리비 업로드                    2710 non-null   object
 48  좌표X                       2710 non-null   float64
 49  좌표Y                       2710 non-null   float64
 50  단지신청일                      2710 non-null   object
dtypes: float64(18), int64(4), object(29)
memory usage: 3.6+ MB
```

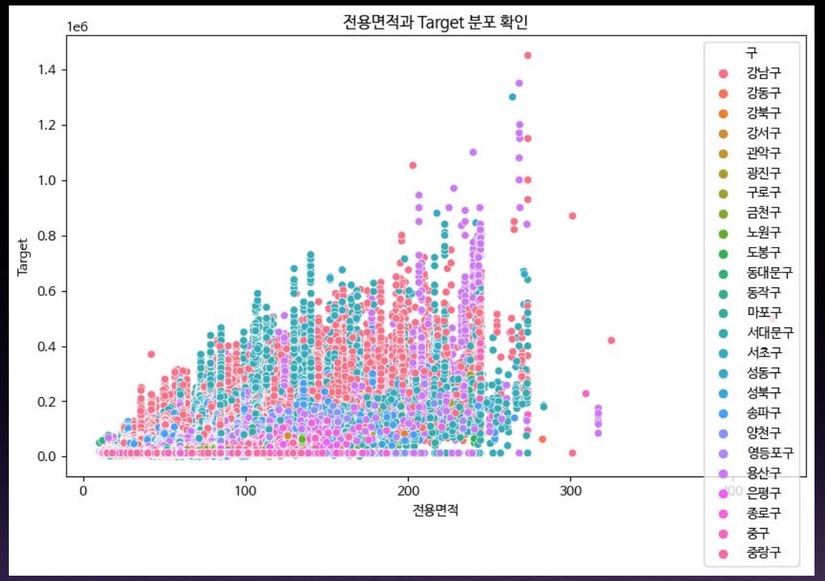
데이터의 기초 통계 및 정보 요약

사진 및 내용을 입력해주세요.

EDA

– '전용면적'





데이터 전처리

「결측치 처리」

ㅡ 특정 의미 없는 value np.nan 처리

- 기본적으로 결측치가 90% 이상을 차지하는 피처 → Deletion
  - 연속형 변수: 선형보간
  - 범주형 변수: NULL 임의의 범주로 대체

[본번/부번]

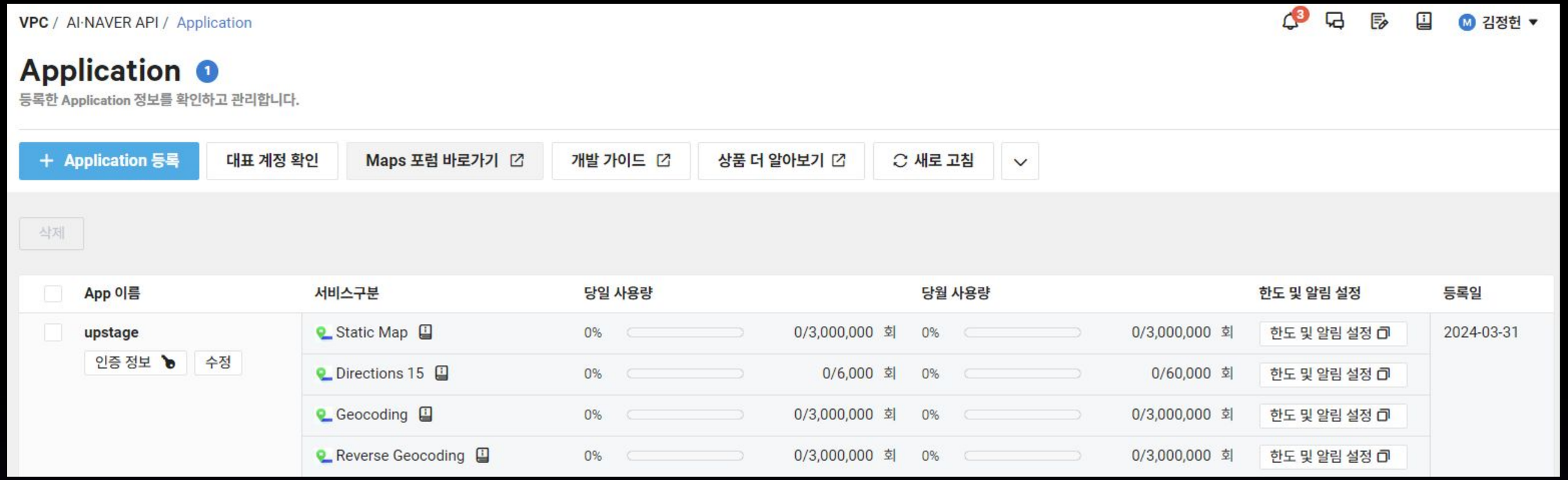
- 도로명의 경우, 결측치 X
- 본번, 부번의 결측치는 도로명을 기준으로 해당 번지를 찾아 Imputation 진행
  - '현릉로8길 10-22' → 특정 하나의 도로명에 대한 본번/부번 결측 확인

```
현릉로8길 10-12    75
Name: 도로명, dtype: int64
```

	도로명	본번	부번	시군구
60194	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
60195	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
60196	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
60197	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
60198	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
...	...	...	...	...
720188	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
720189	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
720190	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
720191	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동
720192	현릉로8길 10-12	NaN	NaN	서울특별시 서초구 신원동

75 rows × 4 columns

데이터 전처리_ 결측치



x좌표, y좌표의 경우에는 네이버 api를 사용하여 결측치를 채우는 것이 가능하다.

-> 도로명 주소를 기입하면 x,y좌표를 반환한다.

외부 데이터 활용 - 금융 관련 데이터

- 가계 대출 규모 3개월 ~ 6개월가계 대출 증가는 주택 구매 여력 증가로 이어질 수 있지만, 실제 주택 구매까지는 3개월 ~ 6개월 정도의 시간이 소요될 수 있습니다.
- 한국 기준금리 6개월 ~ 1년 금리 변화는 주택 가격에 직접적인 영향을 미치지만, 그 영향은 즉각적으로 나타나지 않고 6개월 ~ 1년 정도의 시차를 두고 나타납니다.
- 미국 기준금리 6개월 ~ 1년 미국 금리 변화는 국내 금리 및 주택 가격에 영향을 미칠 수 있으며, 6개월 ~ 1년 정도의 시차를 두고 나타납니다.
- 본원통화량 6개월 ~ 1년 통화량 증가는 인플레이션으로 이어져 주택 가격 상승으로 이어질 수 있지만, 그 영향은 직접적이지 않고 다른 변수들에 의해 영향을 받으며, 6개월 ~ 1년 정도의 시차를 두고 나타납니다.
- 인허가 실적 6개월 ~ 1년 인허가 증가는 향후 주택 공급 증가를 나타내지만, 실제 주택 공급까지는 6개월 ~1개월의 시차를 두고 나타납니다.
- 아파트 미분양 현황 3개월 ~ 6개월미분양 주택 증가는 주택 공급 증가로 이어져 주택 가격 하락으로 이어질 수 있으며, 3개월 ~ 6개월 정도의 시차를 두고 나타납니다.
- 월별 아파트 거래량 1개월 ~ 3개월주택 거래량 증가는 주택 수요 증가를 나타내며, 1개월 ~ 3개월 정도의 시차를 두고 주택 가격에 영향을 미칠 수 있습니다.

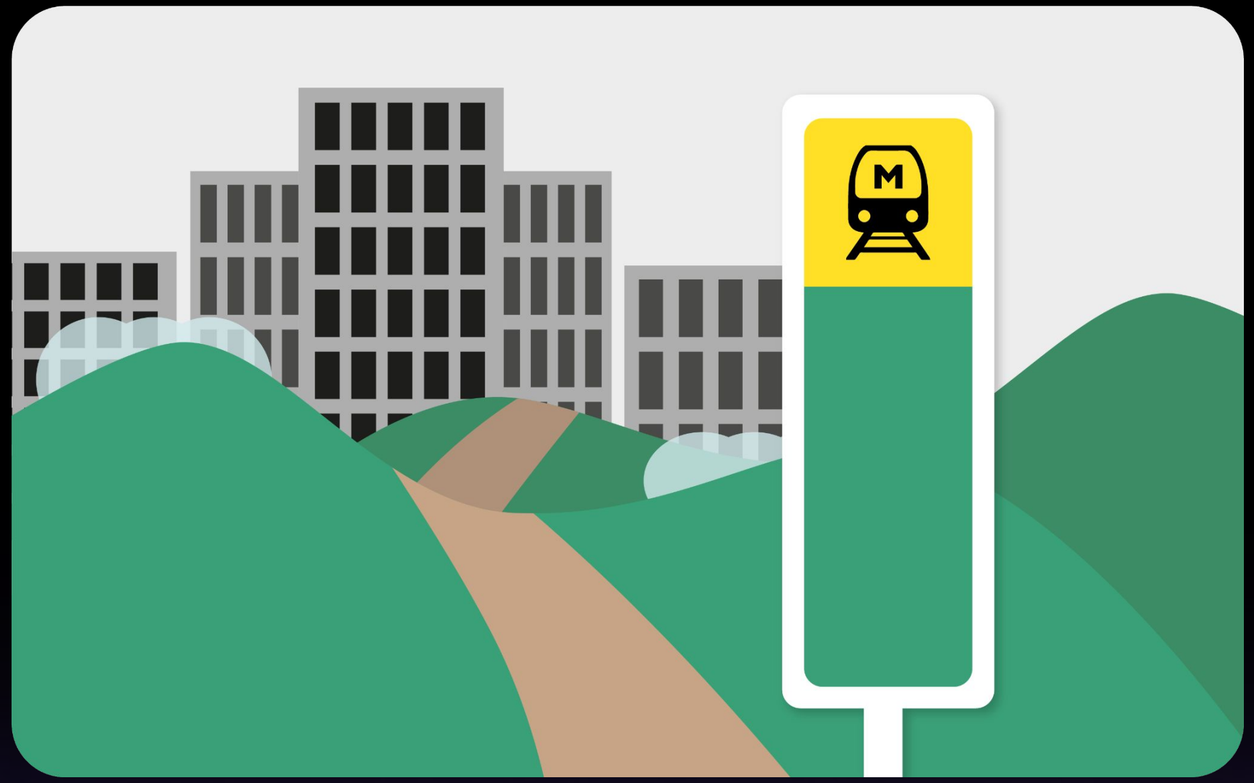
- 강남 지역 여부

https://www.gangnam.go.kr/contents/hello/3/view.do?mid=ID06_040401

- 역세권 여부

역세권이란) 철도역과 인근의 철도시설 및 그 주변지역 중 국토교통부장관이 지정한 지역을 말한다.

역세권의 범위에 대해서는 구체적으로 정의되어 있지 않으나
철도(지하철)를 중심으로 3km을 역세권으로 간주하였다.



https://www.honestfund.kr/blog/money/p2p/station-influence-area

Feature Engineering

- 신축 여부

  신축인지, 구축인지의 여부도 실거래가에 큰 영향을 줄 수 있습니다.

  2009년 이후에 지어졌으면 비교적 신축이라고 판단하였습니다.

https://www.mk.co.kr/news/realestate/10498777

04

Modeling

Model Select

사진 및 내용을 입력해주세요.

Various Models with Trial and Error

사진 및 내용을 입력해주세요.

Hyperparameter Tuning

사진 및 내용을 입력해주세요.

강사님께 받은 피드백 및 의견을 정리해주세요.

사진 및 내용을 입력해주세요.

05

결과

최종 순위 및 평가지표 결과

사진 및 내용을 입력해주세요.

06

그룹 스터디 진행 소감

그룹 스터디 진행 소감

Point 1

외부 데이터를 컨트롤하는 일이 생각보다 어렵다는 생각을 하였다.

이유 : 간단한 날짜에 대한 데이터여도 저장된 형태도 다르고 데이터셋의 구조가 다르기 때문에 합치는 과정에서 많은 어려움이 있었다.
향후 계획 : 실제 대회를 해봐야 알 수 있는 문제상황을 마주쳤다는 생각이 들고 그렇기에 유익한 경험이라는 생각이 든다.

Point 1

느낀점 2

이유 :
향후 계획 :

Point 1

느낀점 3

이유 :
향후 계획 :

Q&A

감사합니다.