

ML 5조

팀 이름 입력
팀원 이름 입력

Content

- 01. 팀원 소개
- 02. 대회 소개
- 03. Data Description
- 04. Modeling
- 05. 결과
- 06. 경진대회 진행 소감

01

팀원 소개

팀원 소개

팀원 이름	역할
김창희	데이터 전처리, EDA, 모델링
김형수	데이터 전처리, EDA, 모델링
윤수인	데이터 전처리, EDA, 모델링
이소영	데이터 전처리, EDA, 모델링
장은혁	데이터 전처리, EDA, 모델링
장준규	데이터 전처리, EDA, 모델링

02

대회 소개

대회 개요

- 주어진 아파트 매매 데이터를 활용하여 서울의 아파트 실거래가를 효과적으로 예측하는 모델을 개발하는 (regression) 대회

House Price Prediction | 아파트 실거래가 예측

서울시 아파트 실거래가 매매 데이터를 기반으로 아파트 가격을 예측하는 대회

#비공개대회 #UpstageAILab1기 #MLAdvanced

D-Day | 2024.01.15 ~ 2024.01.25 19:00

19팀

개요

데이터

서버

제출

리더보드

게시판

팀 관리

☒ 대회 참여

소개

House Price Prediction 경진대회는 주어진 데이터를 활용하여 서울의 아파트 실거래가를 효과적으로 예측하는 모델을 개발하는 대회입니다.

<https://next.stages.ai/competitions/276/overview/description>

RMSE (Root Mean Squared Error)

- $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
- 회귀 모델이 실제 거래 가격의 차이를 얼마나 잘 잡아내는지 측정

제출 파일

- 9,272개의 input에 대한 예상 아파트 거래금액(단위 만원)을 담은 csv 확장자 파일

	A	B	C	D	E	F	G	H	I	J	K
1	target										
2	128108										
3	154749										
4	133011										
5	139851										
6	157270										
7	146353										
8	142269										
9	145193										
10	134475										
11	142960										
12	141412										
13	135734										
14	146433										
15	138875										
16	131291										
17	131540										
18	160122										
19	144545										
20	147128										
21	131997										

03

Data Description

데이터 설명

학습 데이터

- 거래 금액을 포함한 거래 아파트 관련 52개의 변수를 포함한 1,118,822개의 데이터
- 학습 데이터의 거래 기간은 2007년 1월 1일부터 2023년 6월 30일까지

지하철역 데이터

- 역사명, 호선, 역의 x좌표, y좌표를 포함한 서울시 지하철역에 대한 정보

버스정류장 데이터

- 정류소번호, 정류소명, 정류소의 x좌표, y좌표를 포함한 서울시 버스 정류소에 대한 정보

평가 데이터

- 예측해야 되는 거래금액을 제외한 51개의 변수를 포함한 9,272개의 데이터
- 평가 데이터의 거래 기간은 학습 데이터 거래 기간 이후 3개월인 2023년 7월 1일부터 2023년 9월 26일

데이터 전처리 (1-1)

- ‘서울시 공동주택 아파트 정보’(외부 데이터)를 사용하여 ‘도로명주소’를 기준으로 일부 결측치 대체
- 서울시 공동주택 아파트 정보: ‘k-연면적’, ‘k-건설사(시공사)’와 같이 학습데이터와 일부 공통된 열을 가진 공공 데이터.



<https://data.seoul.go.kr/dataList/OA-15818/A/1/datasetView.do>

데이터 전처리 (1-2)

번호	k-아파트코드	k-아파트명	k-단지분류(아...	kapt도로명주소	주소(시도)k-ap...	주소(시군구)
1	A15679103	우리유엔미	아파트	서울특별시 동작구 서달로 83	서울	동작구
2	A13876112	송파파인타운13단지	아파트	서울특별시 송파구 송파대로8길 10	서울	송파구
3	A13873701	오금현대백조(임대)	아파트	서울특별시 송파구 양재대로72길 20	서울	송파구
4	A15275101	개봉건영	아파트	서울특별시 구로구 고척로21나길 8...	서울	구로구
6	A13991016	월계동원베네스트	아파트	서울특별시 노원구 월계로53길 21	서울	노원구
7	A13789201	양재우성KBS(113동)	아파트	서울특별시 서초구 바우뒀로 91	서울	서초구
8	A13486701	천호삼익	아파트	서울특별시 강동구 상암로12길 13	서울	강동구
9	A14319001	자양경남아너스빌	아파트	서울특별시 광진구 자양로3길 55	서울	광진구
10	A12208101	신사성원	아파트		서울	은평구
11	A15809001	신월대방샤인힐	아파트	서울특별시 양천구 신월로9길 37	서울	양천구
12	A13820001	문정3차푸르지오	아파트	서울특별시 송파구 송이로 257	서울	송파구
13	A13790701	잠원한신그린	아파트	서울특별시 서초구 신반포로33길 71	서울	서초구
14	A13402202	천호e-편한세상	아파트	서울특별시 강동구 올림픽로81길 15	서울	강동구
15	A11007001	경희궁의아침3단지	아파트	서울특별시 종로구 사직로8길 34	서울	종로구
16	A13403001	성내2차e-편한세상	아파트	서울특별시 강동구 풍성로37길 55	서울	강동구
17	A15609001	사당유니드	아파트	서울특별시 동작구 동작대로29길 1...	서울	동작구
18	A13186801	신내다우헤밀리	아파트	서울특별시 중랑구 송림길 16	서울	중랑구
19	A13187801	명인하이츠11차	아파트	서울특별시 강동구 명인로 836	서울	강동구

<https://data.seoul.go.kr/dataList/OA-15818/A/1/datasetView.do>

데이터 전처리 (2)

- 'geopy' 라이브러리를 사용하여 좌표(좌표x, 좌표y) 결측치 대체
- 'geopy' 라이브러리를 사용하면 '도로명'주소를 위도와 경도 좌표로 변환할 수 있음

```
# '도로명'이 같은 관측치끼리 한 번만 좌표값을 가져오도록 처리
unique_roads = df_concat_select[df_concat_select['좌표x'].isnull()]['도로명'].unique()

for road in tqdm(unique_roads):
    mask = (df_concat_select['도로명'] == road) & (df_concat_select['좌표x'].isnull())

    if mask.any():
        coordinates = get_coordinates(road)

        # 가져온 좌표값을 결측치가 있는 관측치에 채우기
        df_concat_select.loc[mask, ['좌표x', '좌표y']] = coordinates

# 좌표x, 좌표y 결측치 대체 후 남은 결측치 확인
df_concat_select[['좌표x', '좌표y']].isnull().sum()
```

데이터 전처리 (3)

- 남은 결측치에 대해서 연속형(수치형) 변수일 경우 회귀모델을 기반으로 결측치를 추론하여 보간하고, 범주형 변수일 경우 'NULL'로 결측치를 대체

기준금리 데이터

- 아파트 거래날짜를 기준으로 적용되는 기준금리를 원본데이터에 추가

변경일자		기준금리
2023	01월 13일	3.50
2022	11월 24일	3.25
2022	10월 12일	3.00
2022	08월 25일	2.50
2022	07월 13일	2.25
2022	05월 26일	1.75
2022	04월 14일	1.50
2022	01월 14일	1.25
2021	11월 25일	1.00
2021	08월 26일	0.75

<https://www.bok.or.kr/portal/singl/baseRate/list.do?dataSeCd=01&menuNo=200643>

인구밀도 데이터

- 서울특별시 행정동 별 인구밀도 데이터를 활용하여 원본 데이터에 인구 밀도 변수 추가

동별(1)	동별(2)	동별(3)	2022		
			인구 (명)	면적 (km²)	인구밀도 (명/km²)
▲ ▼ ▢	▲ ▼ ▢	▲ ▼ ▢	▲ ▼ ▢	▲ ▼ ▢	▲ ▼ ▢
합계	소계	소계	9,667,669	605.21	15,974
	종로구	소계	152,211	23.91	6,365
		사직동	9,355	1.23	7,606
		삼청동	2,642	1.49	1,773
		부암동	9,536	2.27	4,201
		평창동	17,858	8.87	2,013
		무악동	8,052	0.36	22,367
		교남동	9,994	0.35	28,554
		가회동	4,065	0.54	7,528
		종로1,2,3,4가동	8,700	2.35	3,702
		종로5,6가동	5,817	0.60	9,695
		이화동	7,631	0.78	9,783
		창신1동	5,383	0.31	17,365
		창신2동	8,796	0.26	33,831
		창신3동	6,606	0.23	28,722
		송인1동	6,057	0.23	26,335
		송인2동	10,509	0.35	30,026
		청운효자동	11,818	2.57	4,598
		혜화동	19,392	1.12	17,314
	중구	소계	130,785	9.96	13,131
		소공동	3,578	0.95	3,766
		회현동	5,222	0.84	6,217
		명동	3,638	0.99	3,675
		필동	4,816	1.14	4,225

<https://data.seoul.go.kr/dataList/10790/S/2/datasetView.do>

그 외 사용한 외부데이터

- 전세가격지수
- 인근종합공원개수
- 인근병원개수
- 인근학교개수
- 한강지천생활지수

등의 외부 데이터를 사용하여 원본데이터에 새로운 feature 추가.

Feature Engineering (원본데이터 활용)

- 시군구 컬럼을 '시'와 '군'으로 분할
- '계약년월'을 '계약년', '계약월'로 분할
- '도로명'(전체 도로명 주소)에서 '도로'(도로 이름, 예: 삼성로)만 추출
- '부촌여부' 변수 추가: 실거래라 상위 아파트가 많이 위치한 동(청담동, 한남동, 성수동 1가) 여부
- '상위아파트여부' 변수 추가: 실거래가 top10 아파트 여부
- '대장아파트거리' 변수 추가: 지역구별 대장 아파트 기준 거리
- 'top아파트거리' 변수 추가: 동별 상위 아파트와의 거리
- '건물연식' 변수 추가: 계약년 - 건축년도
- '브랜드명' 변수 추가: '아파트명'이 주요 브랜드명을 포함하는 경우 해당 브랜드명 입력
등과 같은 방법으로 새로운 feature 추가 ('인근지하철역개수', '인근버스정류장개수')

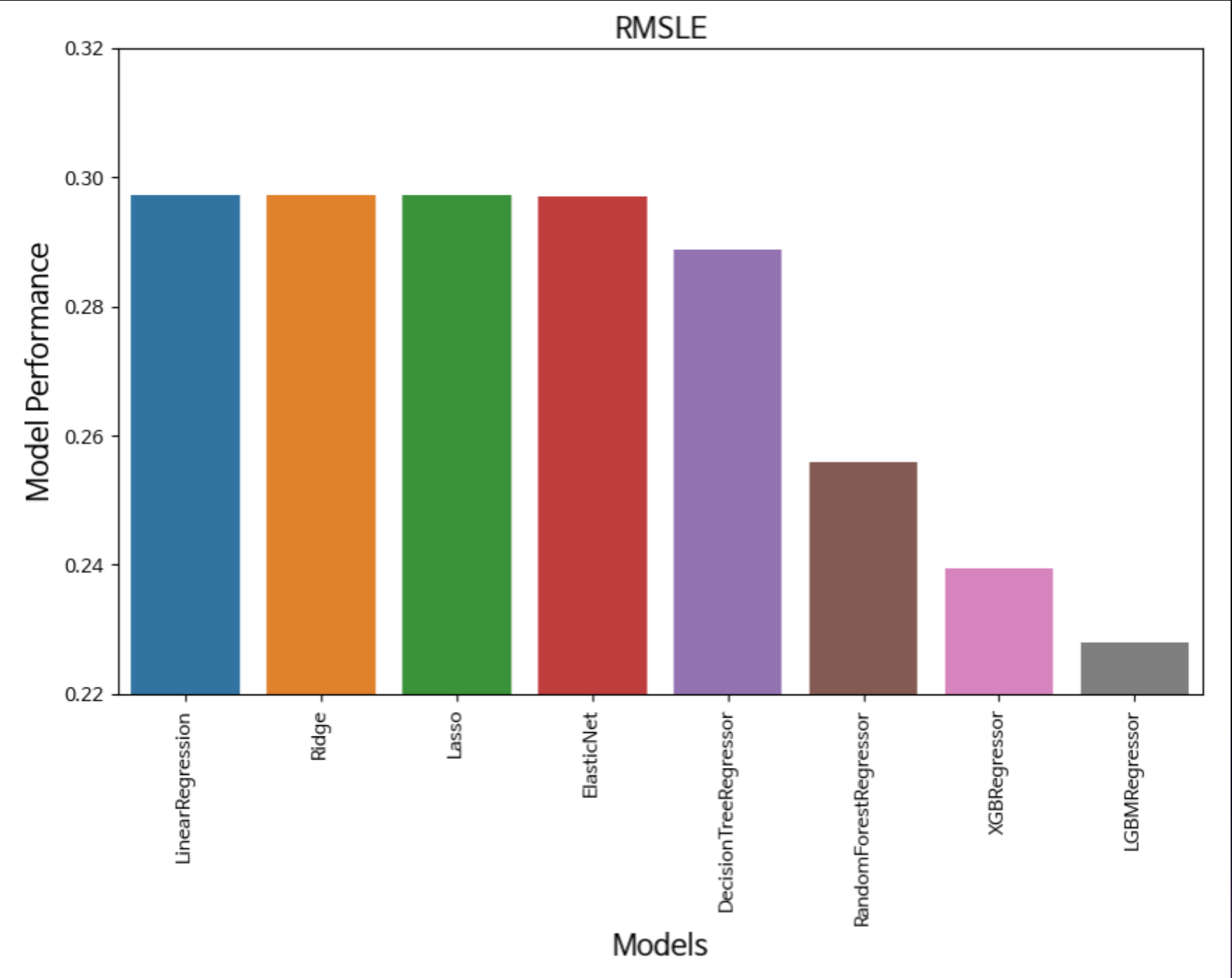
- 주어진 변수들 중 모델 학습 결과 중요도가 낮은 변수, A/B 테스트 결과 성능 향상에 도움이 되지 않는 변수 등을 제거하고 20개의 변수만 활용
- 사용한 변수 리스트:
 - '도로명_실거래가순위', '전용면적', 'k-복도유형', 'k-단지분류', '계약년', '계약월', '동_실거래가순위', '좌표X', '좌표Y', '건축년도', '부존여부', '상위아파트여부', '대장아파트거리', '도로_실거래가순위', '구', '주차대수', '인근지하철역개수', '브랜드명', '건물연식', 'top아파트거리'

04

Modeling

Model Select

여러 모델을 사용해본 후 그 중 가장 성능이 좋았던 LGBMRegressor를 사용.



1. 초반에는 K-fold Cross Validation 혹은 validation set 랜덤 추출 등을 통해 학습 성능을 검증하였으나, 신뢰할 수 있는 Validation Set 구축을 위해 '계약년월일'을 기준으로 최근 20%의 거래 데이터를 Validation Set으로 추출하였습니다.
 - 테스트 데이터가 학습데이터보다 미래의 데이터임을 반영.
2. 모델 학습 후 validation set의 예측값을 분석하여 실거래가가 높은 데이터에서 더 부정확한 예측을 하는 것을 확인. 이를 보완하기 위하여 두 가지 방법을 사용
 1. 실거래가가 높은 데이터의 불균형에 따른 것으로 판단하여 거래가가 높은(100억 이상인) 데이터의 개수를 늘림.
 2. $(Q1 - 1.5 * IQR, Q3 + 1.5 * IQR)$ 을 기준으로 밖에 있는 데이터를 이상치로 판단하여 제거

Hyperparameter Tuning

- Optuna를 사용해 최적의 파라미터 조합을 사용해 성능
- 이후 직접 값을 변경해가며 가장 좋은 성능을 보이는 값으로 설정

강사님께 받은 피드백 및 의견을 정리해주세요.


















































1. 전반적으로 모든 팀의 public score가 높은 편이다. 이는 모델들이 target 값의 이상치에서 부정확한 예측을 하고 있기 때문일 수 있다. 이를 해결하기 위해서
 1. 비싼 아파트를 잘 맞출 수 있는 변수 추가,
 2. 이상치 처리,
 3. valid-split을 교정,
 4. 시계열적인 피쳐 추가등을 시도해 볼 수 있다.
2. 모델에 변수를 너무 많이 추가하면 과적합을 야기할 수도 있다. 이를 방지하기 위해서 feature importance등을 참고하여 변수의 수를 줄여볼 수 있다.

05

결과

최종 순위 및 평가지표 결과

최종 점수는 96538.2074

1	ML 6조	    	85238.1922	61	2d
2	ML 7조	    	86148.2275	76	1d
3	ML 2조	     	88064.8558	32	6d
4	ML 1조	    	88359.1429	60	2d
5	ML_3조	    	88924.2337	38	2d
6	ML 9조	      	88924.2337	10	1d
7	ML 8조	    	89077.1826	73	3d
8	ML 5조	     	96538.2074	50	19h
9	ML 4조	    	106512.0996	65	16h

<https://next.stages.ai/competitions/276/leaderboard>

06

그룹 스터디 진행 소감

그룹 스터디 진행 소감

김형수: 2주 동안 대회를 진행하면서 모델의 성능을 높이기 위해 다양한 시도를 해볼 수 있었습니다. 다양한 시도가 항상 성능 향상으로 이어지지는 않았지만, 대회 진행의 감을 익힐 수 있는 좋은 기회였습니다.

장은혁: 사실 이번이 저에게 있어서는 두 번째 대회라 그래도 조금은 익숙해질 줄 알았는데 도메인 지식 부족이나 피쳐 엔지니어링 등과 같은 부분들에서 부족한 부분들이 많았던 것 같습니다. 대회를 진행하며 부족하다고 느꼈던 부분들은 더 공부하여 다음 대회 때는 좋은 결과 내볼 수 있도록 노력해보고 싶습니다.

장준규: 앞으로의 대회를 참여하면서 부족하다고 느껴졌던 실력을 좀 더 발전시키고 싶습니다.

윤수인: kaggle이나 데이콘에서 비슷한 대회를 보고 인사이트를 얻을 수 있어 좋았지만, 생각보다 인사이트를 적용해서 성능을 높이는게 쉽지 않다는 것을 느꼈고 아직 머신러닝에 대해 부족한 점이 많다는 것을 체감하였습니다. 그냥 이론을 공부하는 것보다 대회를 진행해보는 것이 더 많은 것을 얻을 수 있는 경험인 것 같습니다.

이소영: 실거래가라는 타겟이 훈련데이터에서 드러나지 않는 불규칙한 패턴을 가질 수 있는 미래의 정보이다보니 생각했던 것보다 성능을 향상시키는 데 많은 어려움이 있었고, 검증 데이터셋 설정과 검증 스코어 및 리더보드 스코어의 신뢰도를 판단하기가 까다로운 대회였던 것 같습니다. 하지만 팀원들과 문제에 대한 이해를 공유하고 모델 성능을 높이기 위한 다양한 시도를 하는 과정이 의미 있었습니다.

Q&A

감사합니다.

—