서울 아파트 실거래가 예측 5조

팀 이름 입력
팀원 이름 입력

Content

01

팀원 소개

팀원 소개 (2)

마서연

역할

- 도메인 지식 획득 및 가설검정
- 탐색적데이터분석(EDA)
- 모델링
- 검증

팀원 소개 (3)

김정우

역할

- 도메인 지식 획득 및 가설검정
- 탐색적데이터분석(EDA)
- 모델링
- 검증

팀원 소개 (4)

주남정

역할

- 탐색적데이터분석(EDA)
- 외부 데이터 수집
- 데이터 전처리
- 데이터베이스 저장 및 관리

팀원 소개 (4)

김윤환

역할

- 탐색적 데이터 분석(EDA)

02

대회 소개

(모든 조가 공통적으로 발표하는 내용이니 간단하게 해주시면 됩니다.)

대회 개요

서울 아파트 실거래가 예측 대회

대회목표

- Data와 Baseline_Code를 바탕으로 서울의 아파트의 해당 시점의 매매 실거래가를 예측

- 예측된 값과 실제 값 간의 평균 편차를 최소화

- 회귀모델이 예측값과 실거래가과의 차이를 잘 설명하는지 측정

평가지표

평가 지표란?

- 모델의 성능을 평가하기 위해 사용하는 수단.
- 예측값과 실제값 간의 차이를 수치로 표현하여 모델의 정확도를 판단.

왜 중요한가?

- 적절한 평가 지표를 선택해야 모델의 성능을 정확히 평가하고 개선 방향을 제시할 수 있음.

회귀 모델의 대표적인 평가 지표

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)

평가지표

$$Target_1 = 10, \quad Pred_1 = 12$$
$$Target_2 = 15, \quad Pred_2 = 14$$
$$Target_3 = 20, \quad Pred_3 = 18$$
$$Target_4 = 25, \quad Pred_4 = 22$$
$$Target_5 = 30, \quad Pred_5 = 25$$

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

$$RMSE = \sqrt{\frac{1}{5}\left((10-12)^2 + (15-14)^2 + (20-18)^2 + (25-22)^2 + (30-25)^2\right)}$$

$$RMSE = \sqrt{\frac{1}{5}(4+1+4+9+25)}$$

$$RMSE = \sqrt{\frac{43}{5}}$$

$$RMSE \approx 2.93$$

RMSE란?

• 예측값과 실제값의 차이를 제곱한 값의 평균을 구한 후 제곱근을 취한 값.

특징

•오차에 민감하여 큰 오차에 더 큰 페널티를 부여함.
•값이 작을수록 모델 성능이 우수함을 의미.

•해석

•예측값과 실제값 사이의 평균적인 오차 크기를 나타냄.
•큰 오차에 민감해 세밀한 평가 가능.

03

Data Description

데이터 설명

목적

아파트 실거래가 예측

구성

총 데이터 수 :1,118,821건

변수 수: 54개 변수

주요 변수: 시군구, 아파트명, 전용면적, 계약년월, 층, 건축년도, 도로명, 거래유형, K_난방방식, K_단지 종류 등

특징

1. 결측치 다수 존재

2. 데이터 분포 다양

데이터의 기초 통계 및 정보 요약

전용면적

- 평균: 71.5㎡, 중위수: 59.9㎡
- 최소 - 최대: 14.0㎡ - 273.9㎡
- 표준편차: 30.2㎡

•거래 금액(target):

- 평균: 7.5억원, 중위수: 6.2억원
- 최소 - 최대: 0.5억원 - 65억원
- 표준편차: 5.6억원

건축년도:

- 평균: 1999년, 중위수: 2003년
- 최소 - 최대: 1966년 - 2021년
- 표준편차: 10.5년

층수:

- 평균: 9.7층, 중위수: 8층
- 최소 - 최대: 지하 6층 - 지상 80층
- 표준편차: 7.8층

•건축년

• 결측치 현황

- 'k_난방방식' 결측치 비율: 약 78%
- 주요 변수들의 결측치 비율 표기

•데이터 상관관계:

- 전용면적과 거래 금액: 양의 상관관계
- 층수와 거래 금액: 약한 양의 상관관계
- 건축년도와 거래 금액: 최근일수록 가격 상승

EDA

층수 영향 분석

- 지하층 데이터는 가격대가 낮음
- 지하여부 변수 추가로 RMSE 소폭 감소

이상치 처리

- 전용면적 상한을 122.92㎡ → 280㎡으로 수정하여 RMSE 감소

재건축 조건 분석

- 재건축 연한만으로는 가격 상승 영향 미미

Feature Importance 분석 결과

- 'k_난방방식'이 가격 예측에 중요한 변수로 확인

데이터 전처리

DB 활용

MariaDB/MySQL에 데이터 적재하여 관리

결측치 처리

- 외부 데이터와 조인하여 'k_난방방식' 보완

변수 조정

- 전용면적 이상치 상한 수정으로 모델 성능 향상
- 지하여부 변수 추가로 성능 개선

이상치 제거

- IQR 사용하여 실거래가 이상치 제거 시 RMSE 증가로 제외

외부 데이터 활용

서울시 **Open API**

- 공동주택 아파트 정보를 통해 'k_난방방식', 'k_단지분류' 결측치 보완

경제 지표 추가

- 금리와 1인당 총국민소득 데이터를 신규 변수로 추가

효과

추가된 변수들이 모델의 예측 성능 향상에 기여

Feature Engineering

신규 변수 생성

- 지하여부: 지하층 여부를 나타내는 변수
- 재건축연한여부: 재건축 가능 연한 충족 여부

변수 조정

- 전용면적 이상치 상한 수정으로 RMSE 절반 감소

효과 분석

- 지하여부 변수는 RMSE 소폭 감소에 기여
- 최고층 변수는 영향도 낮아 제외

Feature Select

주요 변수 선정

- Feature Importance 기반으로 영향도 높은 변수 선택

변수 제거

- 모델 성능에 부정적 영향 또는 영향도 낮은 변수 제거

결과

- 'k_난방방식' 등 중요한 변수 중심으로 모델 구성
- 불필요한 변수 제거로 모델 효율성 향상

강사님께 받은 피드백 및 의견을 정리해주세요.

EDA의 중요성

멘토님 조언: 머신러닝에서 **좋은 성과를 위해서는 EDA라는 기초공사가 중요**

- 데이터에 대한 깊은 이해가 모델 성능 향상의 **핵심 요소**

피드백 적용 사항

- **EDA 수행**으로 데이터의 특성 파악
  - 데이터 분포, 결측치, 이상치 등 상세 분석
- **변수 탐색과 선택에 시간 투자**
  - 중요한 변수 식별 및 불필요한 변수 제거
- **외부 데이터 활용 방안 모색**
  - 추가적인 인사이트를 얻기 위한 **다양한 데이터 소스 검토**

얻은 인사이트

- **EDA를 통한 문제 정의 명확화**
  - 데이터의 문제점을 사전에 파악하여 **효과적인 전처리 가능**
- **모델링 이전 단계의 중요성 인식**
  - 데이터 이해와 전처리가 **모델 성능에 직접적인 영향**을 미침

04

결과

최종 순위 및 평가지표 결과

Leaderboard [mid] Leaderboard [final]

The final ranking of the competition may change because the remaining evaluation dataset that was not used for the remaining scoring will be used.

Refresh

Last update: 2024.09.13 13:37:15

Rank	Team Name	Team Member	RMSE	Entries	Final
1	ML 6조	S 민석 성범 동호	10624.0663	25	12h
2	ML 4조	S i J	11929.9130	41	3h
3	ML 1조	··· 서현 J	12518.1396	58	37m
4	ML 7조	J 김 재현 최수	12917.0588	51	48m
5	ML 2조	J 문기	13150.1735	47	1h
6	ML 3조	민수 승민	13684.9428	16	47m
7	ML 5조	S N	18373.8542	11	3d
8	남진우(운영진)		34223.5885	1	1w

Gold Silver Bronze

06

그룹 스터디 진행 소감

그룹 스터디 진행 소감

협업 경험
- 팀원들과의 협업을 통해 다양한 관점 공유
- 협업을 통한 문제 해결 시도

시간 부족으로 인한 아쉬움
- 각자의 일정으로 프로젝트에 충분히 몰입하지 못함
- 이로 인한 성적 저조 발생

향후 목표
- 효과적인 시간 관리와 우선순위 설정으로 성과 향상 도모
- 다음에는 더 적극적인 참여로 더 나은 결과 달성 목표

Q&A

감사합니다.