House Price Prediction

# Zero to One

2024. 11. 15

Avengers | 김예승, 김도경, 김지식, 은지영, 이홍록

Fast campus × upstage

어벤져스

# Avengers

김 예 승

*ENTP(J)*

김 도 경

*ESTJ*

김 지 식

*ENFJ*

은 지 영

*INTJ*

이 홍 록

*ENFJ*

*‘따로 또 같이’*

*— 학습도, 도전도 함께 하는 가치*

# CONTENTS

# 1 프로젝트 개요

목적 / 목표

## 경진대회를 통해 얻고 싶은 것은 무엇인가?

# 경진대회 목적 및 목표

**House Price Prediction** | 서울시 아파트 실거래가 매매 데이터를 기반으로 아파트 가격을 예측

- 데이터 수집 (Data Collection)
- 데이터 정제 및 전처리 (Data Cleaning and Preprocessing)
- 피처 엔지니어링 (Feature Engineering)
- 모델 구축, 모델 훈련 및 평가 (Model Building, Training and Evaluation)
- 모델 튜닝 및 최적화 (Model Tuning and Optimization)

(A to Z) 팀원 모두 ML 프로세스를 처음부터 끝까지 학습 및 실습하기

(TOP) 경진대회 1등

# Time Line

| 기간 | 1W | | | | | 2W | | | | | 3W | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 28 | 29 | 30 | 31 | 1 | 4 | 5 | 6 | 7 | 8 | 11 | 12 | 13 | 14 | 15 |

**활동**

"개념과 실전을 동시에 "

ML Advanced 강의 학습 + ML 실전 프로젝트 실습

- (활동) 오전 10분 스크럼, 오후 4시 그룹 스터디 진행
- (주제) '부동산 가격 예측 및 원인 분석' 강의 학습 등 논의
- (심화) 팀원 간 Kaggle private 대회 도전
  — Boston House Prices

**경진대회 Start**

**S1. 문제정의/분석**

- APT. 개별적으로 A to Z 탐색 및 분석

Two Track

**S2. 데이터 수집 및 결측치 복구**

**S2. 데이터 전처리 및 피처 엔지니어링**

**S3. 모델 구축/훈련**

- 적절한 알고리즘 선택 및 모델 훈련/평가 결과 공유

**S4. 튜닝**

# 2 핵심 아젠다

주요 관점

## 각 단계에서 어떤 문제해결과정을 거쳤는가?

# 프로세스별 핵심 아젠다

## 데이터 수집 및 결측치 복원

### 유사/동일 데이터 수집

- 대회에서 제공된 데이터 검토
- 데이터의 결측값을 토대로 유사/동일 데이터 API로 수집

### 데이터 매칭 및 결측값 복원

- 아파트 코드를 기준으로 아파트 단지 정보 복원
  - #1. 도로명 주소와 아파트명 동일 시 매칭
  - #2. 지번 주소와 아파트명 동일 시 매칭
  - #3. 매칭되는 도로명 주소 유일할 경우 매칭
  - #4. 매칭되는 지번 주소가 유일할 경우 매칭
- 매칭—검토—복구의 과정을 거쳐 결측치 비율 77% → 18% 감소

## 데이터 전처리 및 피처 엔지니어링

### 데이터 전처리

- 시계열 특성 활용
- 인코딩
- 스케일링
- 결측치 보간

### 피처 엔지니어링

- (전제조건) 상관분석을 토대로 데이터 간의 유효성 분석
- 서울시 한정으로 아파트 가격에 대한 도메인 지식 활용
- 내부 데이터를 활용한 유의미한 파생변수 도출
- 제공된 외부 데이터(지하철, 버스) 간의 유효성 분석
- 추가적으로 외부데이터(학군, 한강 등)를 수집하여 유효한지 분석

## 모델

- (지역적 특성을 고려) 서울시 25개의 '구'를 기준, 각각의 RMSE 값을 토대로 5개 그룹으로 분류, 그룹별 훈련, 예측
- (다양한 모델 적용) XGBoost, Rasso, LightGBM, Random Forest으로 훈련, 예측

# 3 역할 및 전략

세부전략

## 모델 성능 향상을 위한 고려요인은 무엇인가?

# #1. 결측치 복원

## 문제정의

- 데이터셋 결측치: 아파트 단지 정보가 90% 이상 누락되어 있는 상태
- 해결 필요성: 데이터의 정확성 및 완전성 확보 필요

## 접근방법

- 정부 데이터 포털에서 API를 활용하여 누락된 아파트 단지 정보를 복원

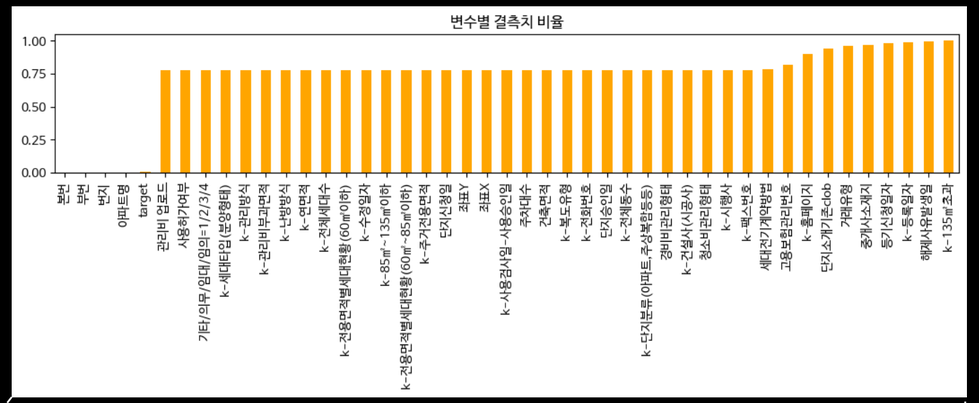
DATA 공공데이터포털 .GO.KR

## 매칭기준

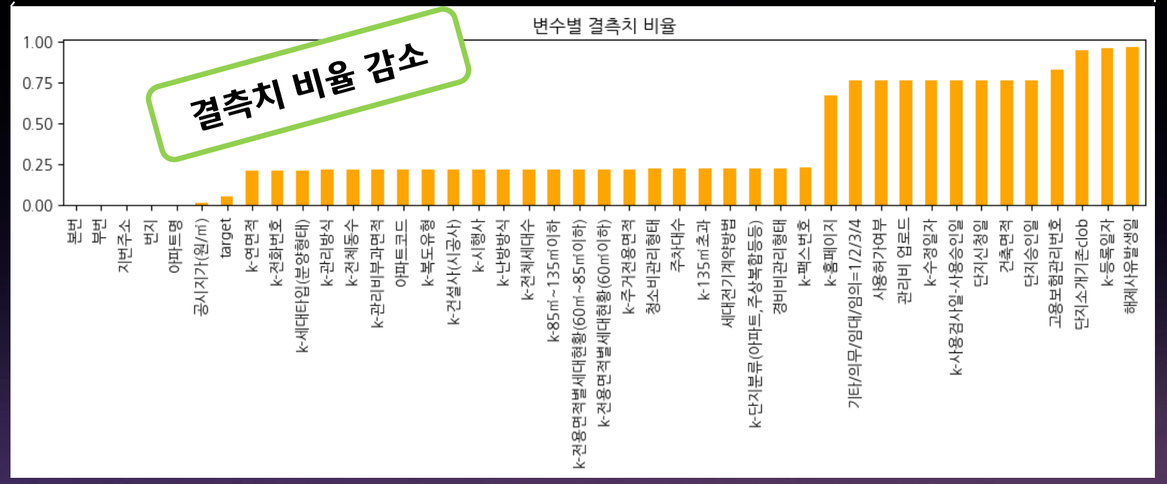
- 총 4가지 기준을 통해 아파트 코드를 매칭:
  1 . 도로명 주소와 아파트명이 동일할 경우 매칭
  2. 지번 주소와 아파트명이 동일할 경우 매칭
  3. 도로명 주소와 매칭되는 단지가 유일할 경우 매칭
  4. 지번 주소와 매칭되는 단지가 유일할 경우 매칭

## 결과

- 결측치 비율 감소: 77% → 18%
- 복원된 단지 정보: 매칭된 아파트 코드를 기준으로 정확한 단지 정보 복원




**“효율적인 복원 작업으로 데이터 신뢰성 및 활용성 향상”**



# #2. 데이터 전처리

## 시계열 특성 활용

- 데이터 시계열 특성과 코로나 기점 가격 급등을 확인
- 2020년 이후 데이터만 사용하여 약 13% 향상 (18800 -> 16400)

COVID-19 COMMUNICATION NETWORK

## 인코딩

- 원-핫 인코딩에서 레이블 인코딩으로 적용
  - 성능에 영향을 많이 끼치지 않음

레이블 인코딩

| color |
|---|
| red |
| green |
| blue |
| red |

→

| color |
|---|
| 0 |
| 1 |
| 2 |
| 0 |

## 스케일링

- 데이터 표준화 및 정규화를 수행하였으나, 성능 영향은 미비한 수준

## 결측치 보간 및 대체

- 결측치 복원 후, 결측값 선형보간법으로 적용
  - 도로명, 구, 순으로 그룹화하여 최빈값 or 평균값으로 대체
  - 선형보간법 대비 성능이 미비하여 배제

**Validation 데이터셋 분할**

Validation 데이터는 기간이 2023년 데이터와 아파트 정보가 누락된 데이터 중에 9271개 랜덤 샘플링

# #3. 피처 엔지니어링

## 내부 데이터

- (층수 정보) 도로명 주소 기준으로 전체 층수 및 층수 비율 생성
  - (추가) 고층 여부 데이터 생성
- (지하층 보정) 모든 데이터에 최소 층의 절댓값 추가
- (세대수 비율) 전용면적에 따른 세대수 비율 계산
- (아파트 연식) 계약년월과 건축년도 기준 연식 생성
  - (추가) 30년 이상 아파트 분리
- (지역 분류) 시, 구, 동으로 세분화
- (거래 횟수 집계) 동별 거래 횟수 생성
- (고가 지역 분리) 강남구, 서초구, 용산구 별도 분리
- (아파트 코드 생성) 도로명 기준 고유 식별코드 부여

## 외부 데이터

- 한강 인접 여부
  - Ver1: 12대교 경위도 기준으로 1km 이내 여부 체크
  - Ver2: 국토부 SHP 파일 활용하여 한강 및 지천(안양천 등) 기준 1km 이내 여부 체크
- (교통 접근성) 아파트와 지하철/버스 간 이동거리 파생변수 생성
- (학군 정보 추가) 아파트 코드를 기준으로 초, 중, 고등학교 학군 정보 추가
- (개별공시지가 추가) 지번 주소 기준으로 개별공시지가 추가
  - RMSE 개선: 19,400 → 18,800 (약 3% 감소)

# #4. 모델 (1) 지역별 특성 반영


1. 지역별 데이터 차이를 고려해 서울시 25개 구에 대해 각각 RMSE를 계산

2. RMSE를 기준으로 5개의 그룹으로 나눠 각 그룹별로 훈련 및 예측 진행

3. 점수 향상: RMSE 16,400 → 14,100

# #4. 모델 (2)





**모델 실험 결과**

- (사용 모델) LightGBM, XGBoost, Lasso, Random Forest
- (모델 성능 비교) 1등: LightGBM, 2등: XGBoost
  - XGBoost는 리더보드 퍼블릭 점수에서는 성능이 좋지 않았지만, 내부 테스트에서 우수한 결과 도출

**하이퍼파라미터 튜닝**

- Optuna, GridSearch, RandomSearch를 활용하여 최적의 하이퍼파라미터 도출

**앙상블 활용**

- LightGBM과 XGBoost의 장점을 결합하기 위해 앙상블 적용
  - 개별 모델의 강점을 활용한 성능 극대화

**결론**

- 지역별 그룹화와 최적 모델 선정 및 앙상블로 RMSE 크게 개선
- 다양한 모델 실험과 조합을 통한 효율적인 문제 해결 접근법

# 4 성과

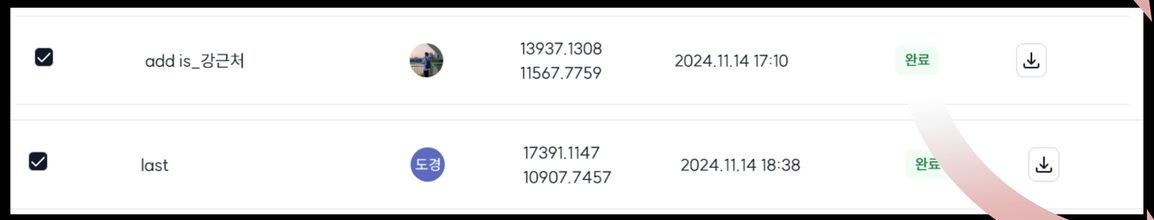
결과

## 그래서 모델 성능 향상이 향상되었는가?

# 최종 결과

**최종 모델 선택 기준**

- 퍼블릭과 내부 테스트 결과를 종합적으로 고려한 의사결정 모델 조합으로 신뢰성과 성능 극대화
  1. 퍼블릭 리더보드 성과 최상: LightGBM
  2. 내부 테스트 성과 최상: LightGBM + XGBoost 앙상블




*"최종 리더보드 1등 달성"*

| 순위 | 팀이 | 팀멤버 | RMSE | 제출횟수 | 최종 제출 |
|---|---|---|---|---|---|
| 내등수 1 | ML 3조 | 홍록 도경 | 10907.7457 | 50 | 19h |
| 1 | ML 3조 | 홍록 도경 | 10907.7457 | 50 | 19h |
| 2 | ML_5 | 성지 혜인 T A | 11809.6503 | 43 | 3d |
| 3 | ML_4 | Mj | 14844.3826 | 76 | 2d |
| 4 | ML 1조 | Y 다혜 종환 김동 | 24757.8335 | 33 | 5d |
| 5 | ML 2조 | 감자 지은 J 주은 | 31769.3940 | 9 | 19h |

골드 실버 브론즈

- 퍼블릭 리더보드: LightGBM 단일 모델 성과 최상
- 최종 리더보드: LightGBM + XGBoost 앙상블 모델로 최종 1위 달성

# 회고

“너무 훌륭한 팀원분들을 만나서, 경진대회중에 배운 점도 많았고, 즐거운 시간을 보냈습니다. 좀 더 기본지식을 보강해야하겠다는 생각을 많이 했습니다. 함께 전체적인 프로세스를 경험하게 되어 값진 기회였습니다. 앞으로 더 배우며 성장하고 싶습니다.”

# Q&A

# 감사합니다.