






# Upstage AI Lab 3기 Regression 대회

9조: Team 클라우드 9 (행복의 절정)

# 팀원 소개

   				
<u>조용중(팀장).</u>	<u>김나리</u>	 <u>최윤설</u>	<u>한지웅</u>	<u>남상인</u>
전반적인 의견조율, 전처리와 모델링의 기초코드 제공, 발표	데이터 결측치 처리, catboost 모델링, 발표자료	데이터 전처리, EDA, 모델링, 발표자료		

# 1. 대회 개요

- 목표: 서울의 아파트 실거래가 예측 모델 개발
- 평가 지표: RMSE (Root Mean Squared Error)
- 기간: 2024년 7월 09일 ~ 7월 19일

## Timeline

- 2024년 7월 09일 (화) ~ 7월 14일 (월) - 온라인 수업, 각자 데이터 EDA
- 2024년 7월 15일 (월) - 회의 후, 회의 결과를 바탕으로 데이터 전처리
- 2024년 7월 16일 (월) - 각자 EDA 및 Feature Engineering
- 2024년 7월 17일 (수) - 최종 데이터셋 설정 및 Modeling
- 2024년 7월 18일 (목) - Feature Selection 및 Modeling

## Hyper-parameter tuning

- 2024년 7월 19일 (금) - 최고 성능 모델 추가 처리 및 최종 제출 기한

## 2. 데이터 설명

### 학습 데이터

- 1,118,822개 샘플, 52개 변수
- 기간: 2007년 1월 1일 ~ 2023년 6월 30일
- 주요 변수: 시군구, 아파트명, 전용면적, 건축년도 등

### 평가 데이터

- 9,272개 샘플, 51개 변수 (타겟 변수 제외)
- 기간: 2023년 7월 1일 ~ 2023년 9월 26일

# 3. EDA 및 전처리

## 주요 발견사항

- 70% 이상의 결측치를 가진 특성들이 다수
- '번지', '본번', '부번', '아파트명'의 결측률은 0.2% 이내
- '전용면적'에 따른 가격에서 이상치 존재

## 전처리 과정

### 1. 결측치 처리

- 서울시 공동주택 아파트 정보 활용
- 카카오 API를 이용한 좌표 정보 보완

### 2. 파생변수 생성

- 지리적 정보, 시공사 정보, 거리 점수 등

### 3. 이상치 처리

- '층'의 음수값을 1로 대체

## 4. 모델링

### 사용 모델

- RandomForest
- XGBoost
- LightGBM (결정모델)
- CatBoost (양상블 모델)

### 주요 기법

- Feature Importance를 이용한 변수 선택
- K-Fold 교차 검증
- 하이퍼파라미터 최적화 (wandb, optuna)
- 양상블 기법
- [Wandb 결과](#)

# 5. 결과

## 최종 모델

- LightGBM + 하이퍼파라미터 최적화

## 리더보드 순위

- Public: 5위
- Private: 2위

## 6. 결론 및 향후 계획

- 단순 모델이 더 좋은 성능을 보임
- 메타 모델 등 추가 실험 필요
- 특성 공학에 더 집중할 필요성 확인



# 참고 자료

- [서울 열린데이터 광장](#)
- [한국은행경제통계시스템](#)
- [Wandb 결과](#)