# 1. 引言

作業目的為預測臺北市房地產市場中房屋的單價元平方公尺，資料包括房屋的基本特徵（如建築完成年、總樓層數）及環境特徵（如鄰近設施數量）。為達到最佳預測效果，採用 XGBoost 和 LightGBM 模型進行建模，並結合特徵工程、超參數調整及交叉驗證優化模型性能。本文將說明實驗方法、結果及分析。

# 2. 實驗方法

## A. 資料處理

a) **資料整合**：
   - 使用 Id 合併 X_train 和 y_train。
   - 針對時間特徵 建築完成年月，提取 建築完成年，並計算房齡（2024 年減去建築完成年）。
   - 缺失值處理：房齡缺失值填補為 0（有進步空間，但需要更多元的資料來進行準確的推斷）。

b) **特徵交互**：
   - 增加兩個交互特徵：
   - 房齡_總樓層：房齡與總樓層數的乘積。
   - 房齡_建物移轉總面積：房齡與建物移轉總面積的乘積。

c) **類別變數編碼**：
   - 使用 OneHotEncoder 將類別特徵（如鄉鎮市區、建物型態等）進行 One-Hot編碼，生成多個二元特徵。

d) **資料集劃分**：
   - 將處理後的特徵集分為 X_train_final 和 X_test_final，並移除無關欄位（如 Id）。

## B. 模型選擇與訓練

a) **模型選擇**：
   - 使用 XGBoost 和 LightGBM 兩種基於梯度提升的樹模型，它們在處理非線性和高維特徵方面表現優異。
   - XGBoost 和 LightGBM：為 Tree-based Models 的擴展版本，基於梯度提升決策樹（GBDT）的實現。
   - 採用交叉驗證（K-Fold，k=5）進行性能評估。

b) **超參數調整**：

   使用 RandomizedSearchCV 對 XGBoost 和 LightGBM 模型進行超參數搜索，調整的超參數包括：
   - n_estimators：樹的數量。

- max_depth：樹的最大深度。
- learning_rate：學習率。
- 正則化參數（reg_alpha 和 reg_lambda）。
- 子樣本比例（subsample 和 colsample_bytree）。

c) **預測與提交：**

- 使用調優後的 XGBoost 和 LightGBM 模型對測試集進行預測，並將兩模型的預測結果取平均作為最終輸出。

# 3. 實驗結果

a) **超參數調整結果**

- XGBoost 最佳超參數：{'n_estimators': 1000, 'max_depth': 8, 'learning_rate': 0.05, 'subsample': 0.9, 'colsample_bytree': 0.7, 'reg_alpha': 0.1, 'reg_lambda': 0.5}
- LightGBM 最佳超參數：{'n_estimators': 1000, 'max_depth': 10, 'learning_rate': 0.05, 'subsample': 1.0, 'colsample_bytree': 0.8, 'reg_alpha': 0.5, 'reg_lambda': 0.1}

b) **模型分析**

**優點：**

- XGBoost 和 LightGBM 能夠有效處理類別特徵、非線性關係及高維數據，適合本次房價預測任務。
- 超參數調整提高了模型的表現，使模型更具穩健性。

**限制：**

- 測試集中可能存在模型未見過的數據模式，導致泛化性能受限。
- 地理特徵（如房屋與捷運站距離）的缺乏可能影響準確性。

# 4. 結論

本次實驗通過 XGBoost 和 LightGBM 模型進行房價預測，並使用超參數調整提升模型性能。模型融合的結果對測試集進行了準確預測 (顯著優於 strong baseline)。

# 5. References

https://medium.com/jameslearningnote/資料分析-機器學習-第 5-2 講-kaggle 機器學習競賽神器 xgboost 介紹-1c8f55cffcc
https://chwang12341.medium.com/machine-learning-給自己的機器學習筆記-kaggle 競賽必備-86305848a0c4
https://medium.com/數學-人工智慧與蟒蛇/smote-enn-解決數據不平衡建模的採樣方法-cdb6324b711e