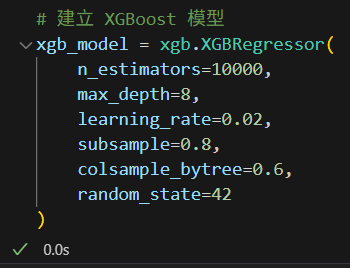
壹、方法設計  
本專案旨在預測台灣住宅物件的單價，任務提供包含建物類型、地點、樓層、格局與建築完成時間等特徵之資料，並需建立機器學習模型進行價格預測，採用的模型為 XGBoost。預測結果需提交至平台評分。

所採用的預測方法為 XGBoost（Gradient Boosting Tree 的優化版本），具備處理非線性資料、高維特徵交互、對異常值較不敏感的優勢。模型使用全部資料進行訓練，不拆分驗證集，並將資料進行標準化處理。

模型參數設定如下：



貳、引用方法與差異說明  
本專案參考過 Kaggle 平台類似房價預測競賽的最佳實踐（如使用 XGBoost、特徵工程設計），但我們自行設計以下關鍵特徵，並針對台灣住宅市場進行在地化優化：

(一) 使用台北車站座標（25.0478, 121.5171）計算距離作為區位特徵（distance\_to\_station）。

(二) 將建年（民國年）轉為西元年，正確計算屋齡。

(三) 從高價房與低價房的前 30 筆樣本中分析其共通特性，加入自訂特徵如下：

1. is\_large\_house：坪數超過 100

2. has\_parking：是否有車位

3. usable\_area\_ratio：可用空間佔總面積比例

4. is\_core\_town：是否位於高價區（大安、中正、新店等）

5. is\_fourth\_floor：是否銷售在四樓（文化避諱）

6. is\_top\_floor：是否為頂樓

7. is\_high\_without\_elevator：高樓層但無電梯

8. room\_density：房數與總面積比

此特徵設計並未直接複製其他資料來源，而是基於對台灣市場理解所做的擴充與轉換。

參、實驗結果與分析  
在使用完整訓練集進行訓練的設定下，模型於提交後獲得更佳的預測表現。與分割驗證集時相比，訓練結果更加穩定，推測原因為：

(一) 資料量較小，若切割 20% 作為驗證集將損失大量學習資訊。

(二) 台灣房市資料變異性高（格局、樓層、地區等），導致驗證集分布可能與整體不同，容易造成 early stopping 太早發生，限制模型學習。

(三) 模型學習傾向如下：

1. 高價樣本特徵：電梯大樓、格局完整、總面積大、區位佳、屋齡中等

2. 低價樣本特徵：老舊建物、小坪數、無車位、樓層 4 樓、非核心區

此分析結果可供未來房價評估與市場潛力區辨作為參考依據。

肆、參考資料與來源說明

(一) 資料來源：課程提供之 X\_train.csv、y\_train.csv 與 X\_test.csv

(二) 特徵靈感與想法來自實際觀察與分析台灣市場需求特性

伍、未來優化方向

(一) 實作交叉驗證（K-fold CV）提高穩定性

(二) 納入更多區域經濟指標（如公車距離、生活機能）

(三) 改進類別變數處理（目前使用 one-hot encoding，可改用 target encoding）

(四) 探索其他模型如 LightGBM 或 CatBoost 做比較