台北市房價預測報告

(一) 實驗方法

1-1 問題描述

本次實驗目的是根據房屋的特徵（如建築面積、建材、環境因素等）及周邊環境（如地鐵站、超商、公園等），預測每平方公尺的單價，並以 RMSE（Root Mean Square Error）作為評估模型準確性的主要指標。

1-2 數據處理

1. 資料加載與檢視：
   (1)使用訓練集（X_train.csv 和 y_train.csv）和測試集（X_test.csv），檢視特徵的數據類型和缺失情況。
   (2)特徵分為類別型與數值型特徵，並進行相應的處理。

2. 類別型特徵編碼：
   (1)對類別型特徵（如「鄉鎮市區」、「交易標的」、「主要用途」等）進行 LabelEncoder 數值化處理，確保這些特徵能被模型接受。
   (2)將相似的特徵值合併，例如將「陽臺」和「陽台」進行統一處理，減少類別數量。

3. 日期特徵處理：
   (1)提取建築完成年份，計算距今的年數，作為新特徵「建築完成年份」。
   (2)使用中位數填補無效值或缺失值，保證數據的完整性。

4. 新增衍生特徵：
   (1)計算「土地建物面積比例」，表示土地面積與建物面積的關係。
   (2)引入「環境密度」特徵，例如地鐵站、超商、公園的數量除以土地面積，用以反映周邊設施的密集程度。
   (3)計算每筆資料到市中心的距離，使用高斯核公式生成「到市中心距離」特徵。

5. 數據標準化：
   對數值型特徵進行標準化處理，使用 StandardScaler 將特徵轉換到均值為 0、標準差為 1 的分布，消除不同特徵量綱的影響。

1.3 模型訓練與驗證

1. 數據分割：
   將訓練集分割為 80% 的訓練子集和 20% 的驗證集，用於模型訓練和效果評估。
2. 模型選擇：
   選擇 LightGBM 回歸模型，該模型能高效處理結構化數據，並支持大規模數據訓練。
3. 超參數設置：
   設定 LightGBM 的核心參數，包括樹深（max_depth=7）、葉節點數量（num_leaves=31）、學習率（learning_rate=0.05）等。
4. 訓練與評估：
   使用驗證集計算 RMSE，作為模型在未見數據上的準確性指標。

1.4 測試集預測

將訓練好的模型應用於測試集，預測測試集每筆資料的單價，並將結果保存為 CSV

(二) 實驗結果

2.1 模型評估結果

在驗證集上的 RMSE 為 30215.68112，表明模型能捕捉房價與特徵之間的關係。

2.2 測試集預測

測試集的預測結果生成的提交文件格式符合要求，並預測了每筆測試資料的單價（元/平方公尺）。

(三) 分析與發現

3-1 關鍵特徵的影響

1. 環境特徵：

地鐵站、超商、公園的密度特徵顯著提升了模型的解釋能力，表明周邊設施的密集程度對3-2房價有重要影響。

2. 到市中心距離：

到市中心的距離特徵顯示出房價與地理位置的關聯性，距市中心越近的房屋，其單價越高。

3. 建築完成年份：

房屋的新舊程度影響房價，特別是較新的建築物通常具有更高的單價。

3-2 模型的優點與局限性

1. 優點：

(1)LightGBM 高效地處理了多樣的數值與類別特徵，且訓練速度快。

(2)特徵工程提取的衍生特徵，如比例特徵和高斯距離特徵，有助於提升模型表現。

2. 局限性：

(1)缺乏更深層次的數據，例如內部裝修情況和業主需求等主觀因素，這些可能對房價影響顯著。

(2)環境特徵的影響未進一步分層分析，例如不同鄉鎮的環境特徵可能影響不同。

(四) 結論

本次實驗成功構建了一個基於 LightGBM 的房價預測模型，並通過一系列特徵工程顯著提升了模型的準確性。驗證集 RMSE 約為 32,765，展示了該模型在房價預測任務上的潛力。

未來可通過以下方法進一步改進：

1. 引入更多數據特徵，例如房屋內部裝修和周邊交通便利程度。
2. 探索其他模型（如 XGBoost 或深度學習）對於特徵非線性關係的表現。
3. 進一步優化超參數配置，通過自動化調參方法如 Optuna 提高模型表現。

這次實驗為台北市房價的智能預測提供了有價值的基線模型，具有良好的應用前景。