房地產價格預測：資料分析與模型建置

運用 vibe coding 完成資料清理與模型建置

動機與目標

本研究旨在運用vibe coding 快速建立完整的資料清理流程及房價預測模型，並將產製相關視覺化成果。

知識整合

結合過往房價知識與資料庫架構，掌握區域房價特性，提升模型預測精度。

明確目標

透過資料探索與模型比較，找出最佳預測方法，支持區域房價決策。

方法與素材 - 程式架構與實作

資料取得

收集歷史房價與地理資訊

清理/標準化

處理缺失值與異常值：**pandas, numpy**

可視化分析

探索資料特性與關聯：**matplotlib, seaborn**

模型訓練

多種演算法實作比較：**sklearn**

超參數調整

優化模型表現：**sklearn**

結果比較

評估模型效能：**plotly.express、mpimg, os**

方法與素材 — 系統架構圖

資料源與資料庫層

歷史房價資料、地理位置資訊、區域經濟指標、房屋基本資料

ETL與特徵工程層

數據清洗、特徵選擇與轉換、缺失值處理、異常值偵測

可視化分析層

相關性分析、分佈圖、地理熱力圖、時間序列趨勢

模型訓練與評估層

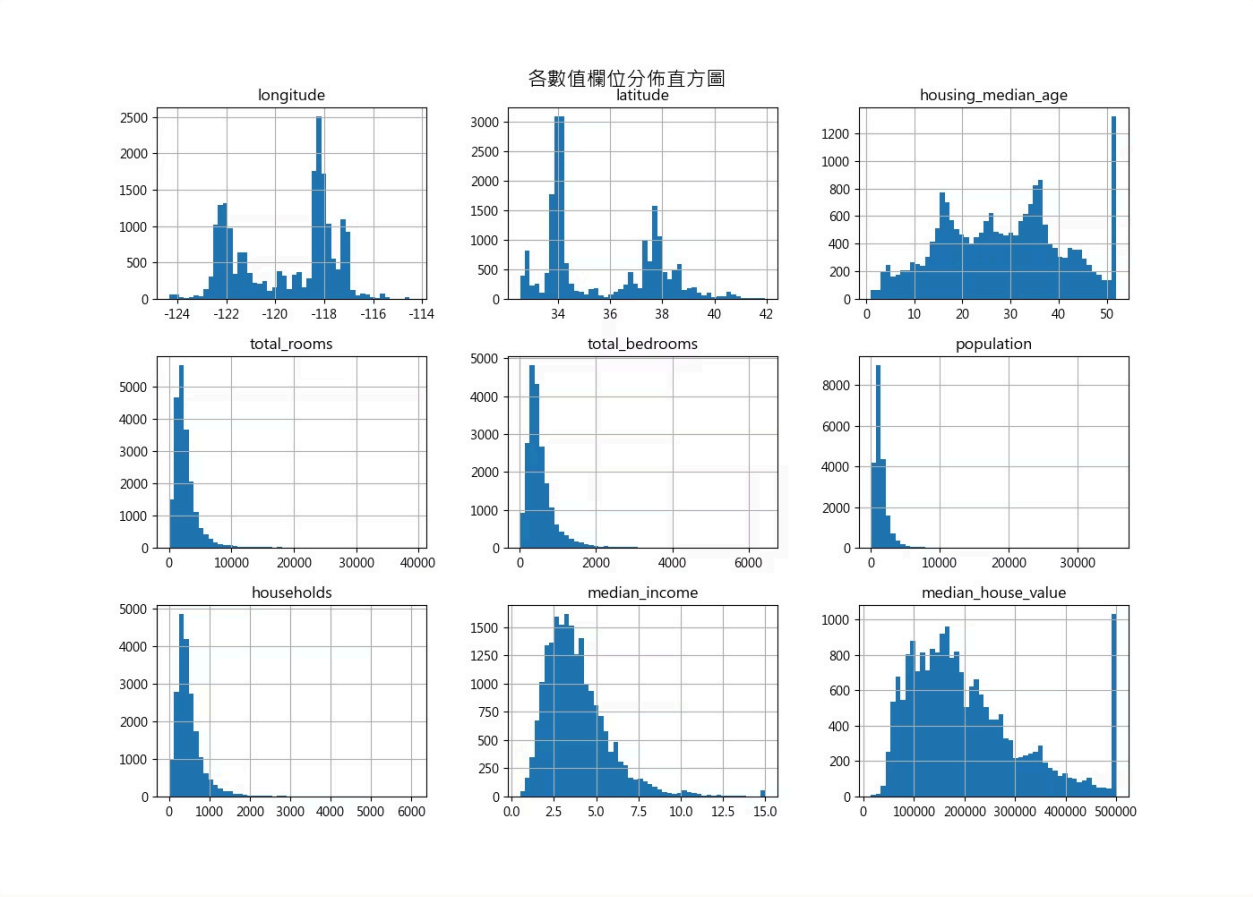
多模型訓練、交叉驗證、性能指標計算、模型解釋

決策支持層

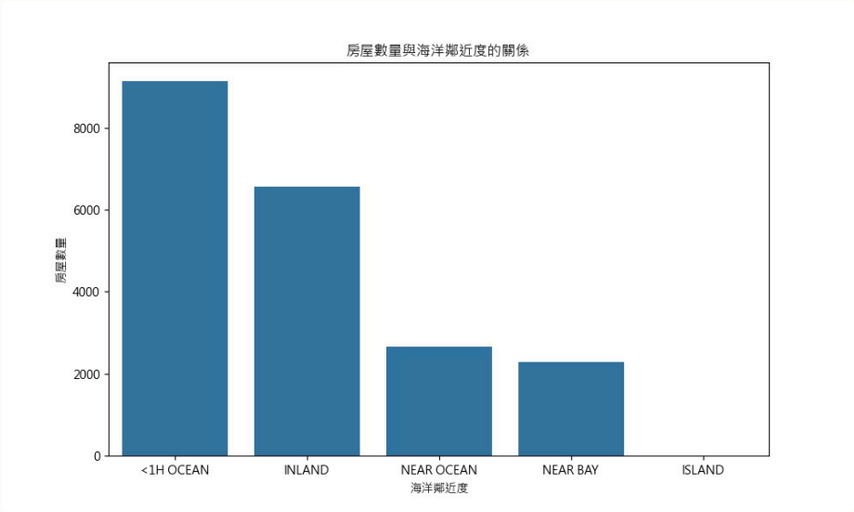
預測結果視覺化、風險評估、價格區間估計、投資建議

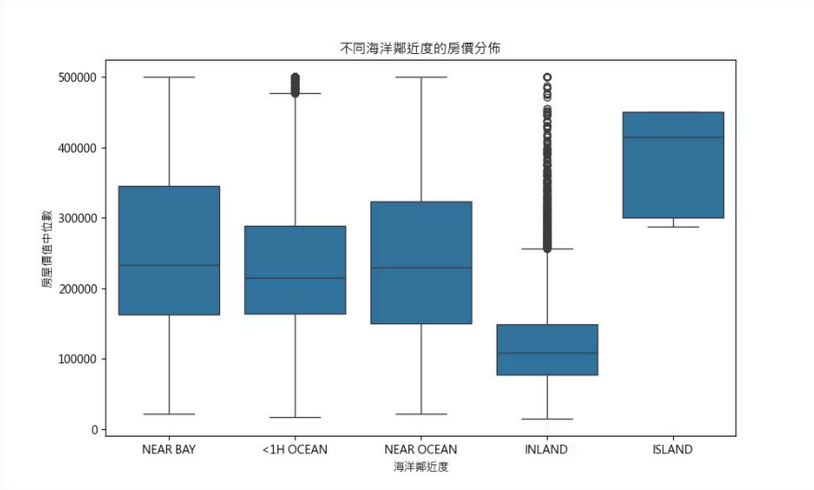
具體實作方法 — 資料前處理與樣態分析

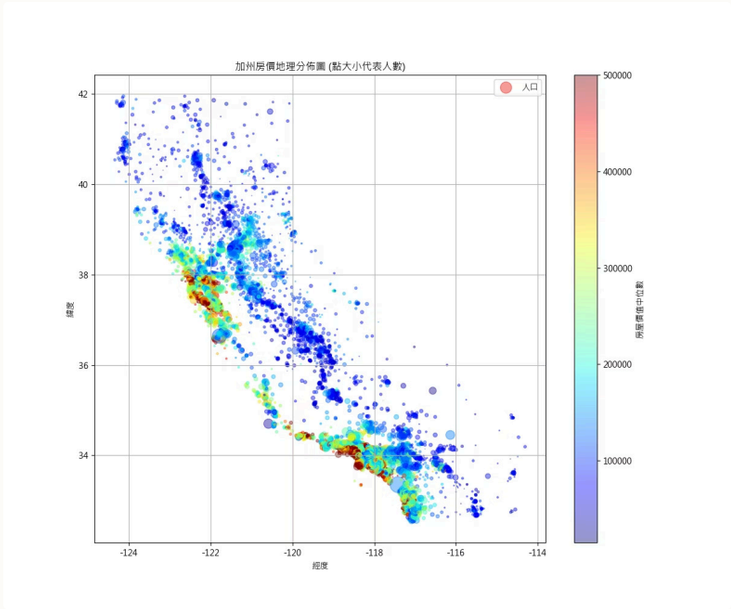
- **數據上限：**許多特徵（如 housing_median_age, median_house_value）的分佈在最右側有突然的截斷，這表明數據可能經過人為設限（Capped），例如房價最高只記錄到 50 萬美元。
- **長尾分佈：**`total_rooms`, total_bedrooms, population, households 等特徵呈現明顯的右偏（長尾）分佈，表示大多數地區的這些數值較低，但有少數人口密集的地區數值極高。
- **收入分佈：**`median_income` 的單位似乎不是美元，且分佈也偏右。在進行模型訓練前，對這些偏態分佈的特徵進行轉換（如取對數）可能會有所幫助。

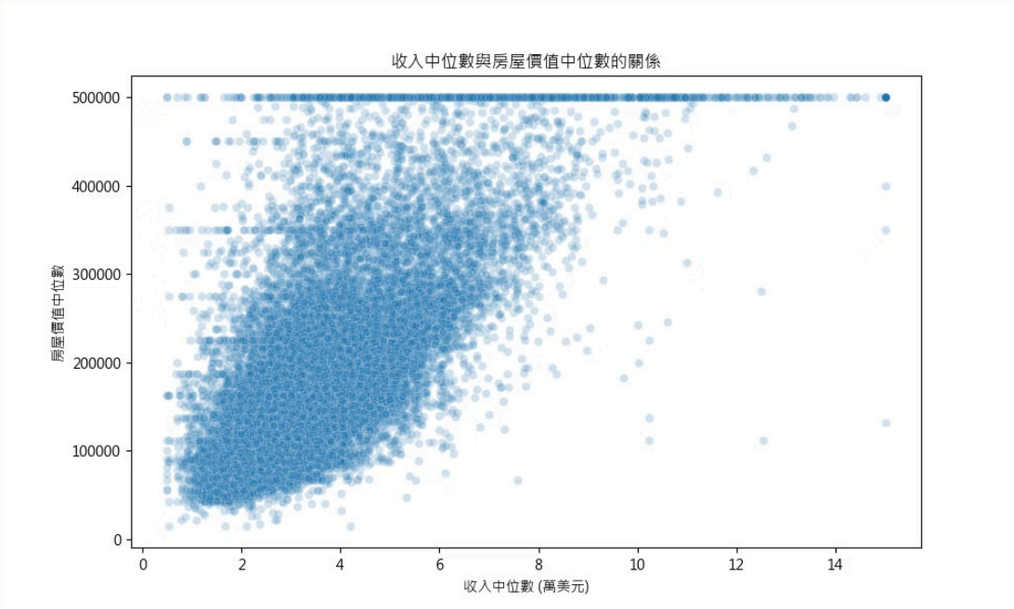


資料樣態分析

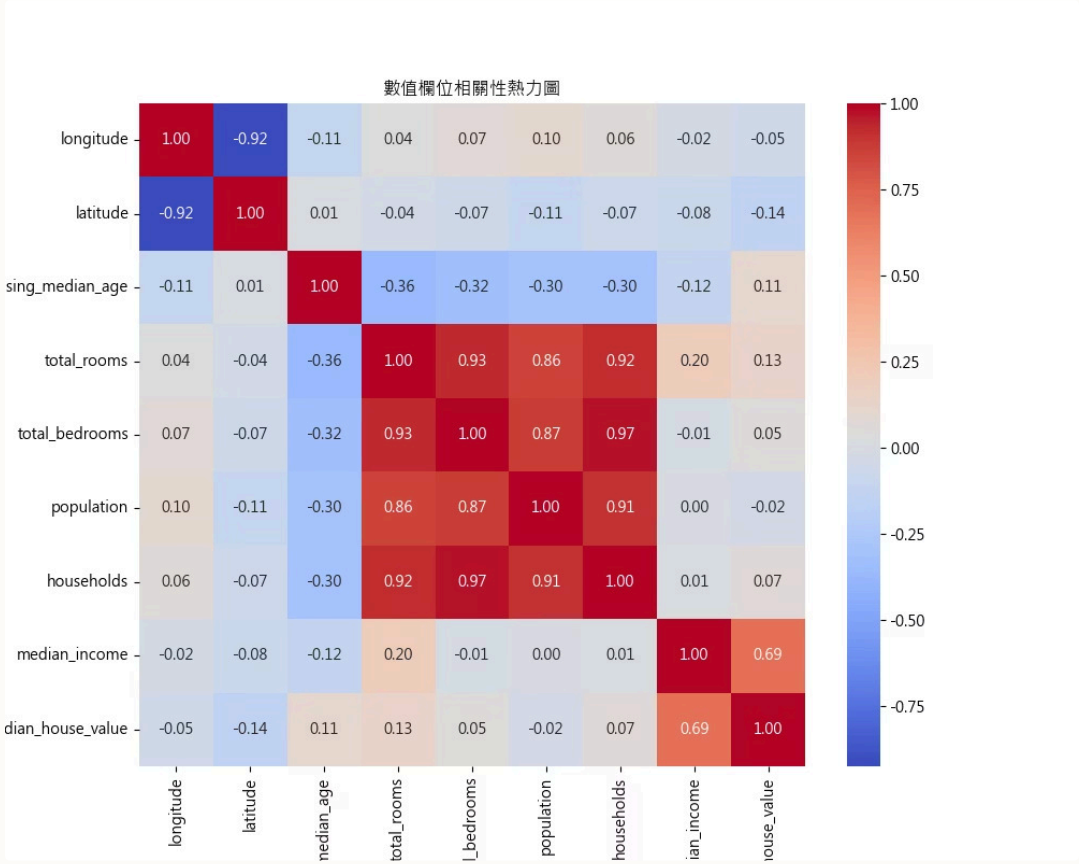








關鍵影響因子分析



- **收入是核心：**從熱力圖可見，`median_house_value` 與 `median_income` 呈現最強的正相關（相關係數為 0.69）。這在下方的散佈圖中也得到清晰的驗證。
- **收入與房價的關係：**散佈圖顯示，收入中位數越高的地區，房價中位數也越高。同時，圖中再次確認了房價在 50 萬美元的上限。
- **其他因素：**`latitude`（緯度）和房價也有一定的正相關，可能反映了北加州（如灣區）房價較高的趨勢。而房間總數 (`total_rooms`) 等與房價的相關性反而較弱，這提示我們人均指標（如我們在模型訓練中創建的 `rooms_per_household`）可能比總量指標更具解釋力。

模型訓練管道

評估指標：RMSE（均方根誤差）、MAE（平均絕對誤差）、R^2（決定係數）、訓練時間

1

線性迴歸 (Linear Regression)

基礎模型，建立特徵與房價間的線性關係，用於模型比較基準。

- 優點：解釋性強、訓練快速
- 缺點：難以捕捉非線性關係

2

隨機森林 (Random Forest)

集成決策樹，減少過擬合風險，適合處理複雜特徵間關係。

- 優點：穩定性高、特徵重要性明確
- 缺點：調參複雜、計算資源消耗大

3

支援向量回歸 (SVR)

利用核函數將資料映射到高維空間，處理非線性關係。

- 優點：泛化能力強、抗噪性好
- 缺點：大資料集訓練緩慢

4

梯度提升機 (Gradient Boosting)

迭代提升模型，逐步改進預測結果，通常表現最佳。

- 優點：預測精度高、可處理多種資料類型
- 缺點：過擬合風險、參數調整較繁瑣

超參數優化 — 針對 Gradient Boosting Regressor

關鍵超參數

學習率 (learning_rate)：控制每次迭代的步長，範圍0.01-0.3

樹的深度 (max_depth)：控制模型複雜度，範圍3-10

子樣本比率 (subsample)：隨機取樣比例，範圍0.5-1.0

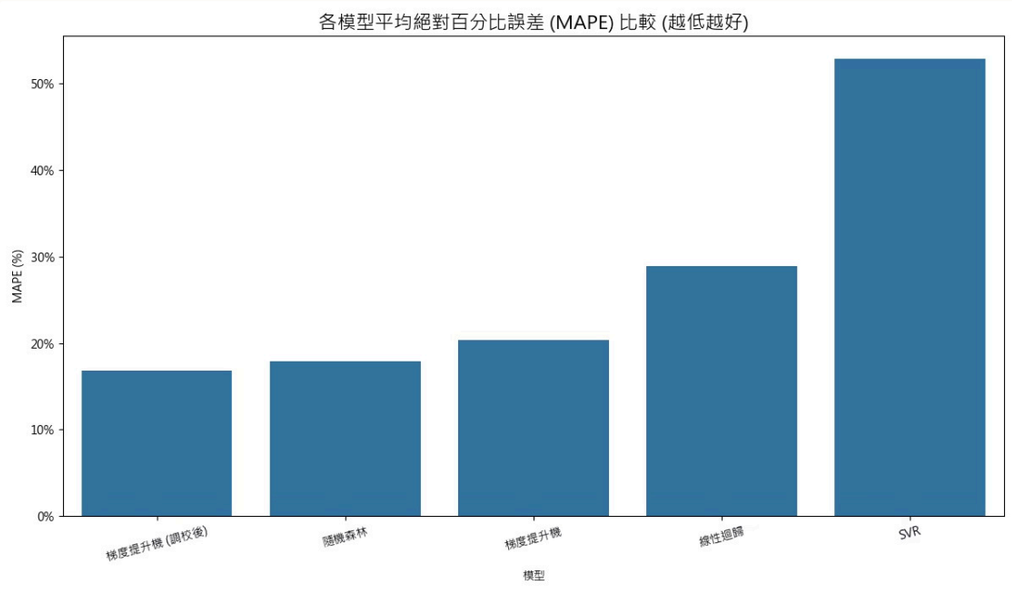
樹的數量 (n_estimators)：控制集成規模，範圍100-1000

優化方法

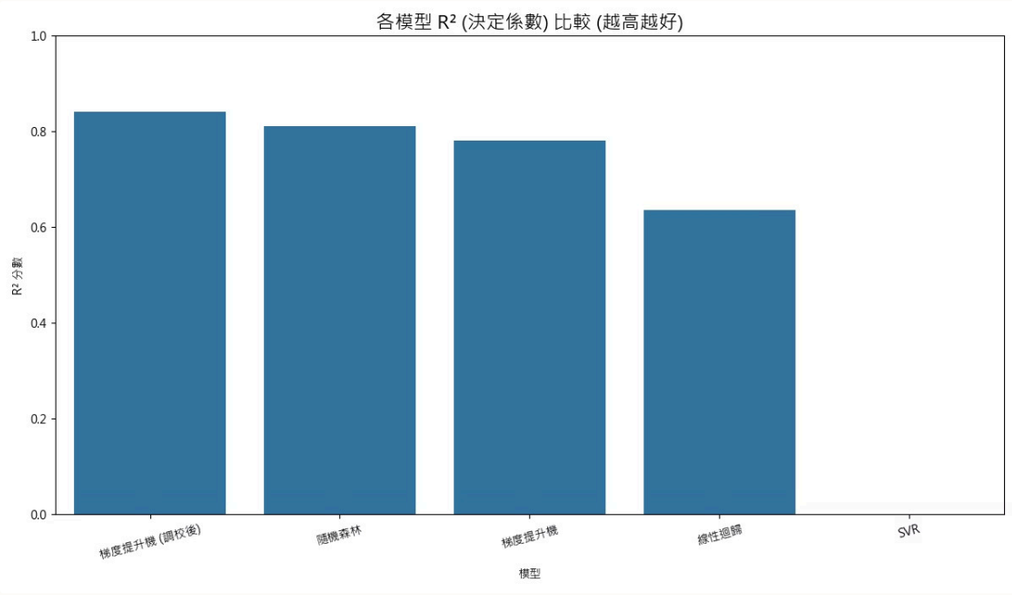
- **Grid Search**：窮舉所有參數組合，尋找最佳設定
- **Randomized Search**：隨機取樣參數空間，更高效率
- **Bayesian Optimization**：基於先驗知識調整搜索方向
- **K-fold交叉驗證**：確保模型穩定性與泛化能力

經過優化後，模型RMSE通常可降低15-25%，R^2提升0.05-0.15，實質提升預測精度。

最終模型表現成果



- **MAPE (平均絕對百分比誤差):** MAPE 衡量預測誤差的百分比。MAPE 越低表示預測的相對誤差越小。
- **梯度提升機 (調校後)** 在 MAPE 方面也表現出色，進一步證明其預測的準確性和穩定性。



- **R^2 (決定係數):** R^2 衡量模型解釋目標變異的程度。R^2 越接近 1 表示模型解釋能力越強。
- 同樣地，**梯度提升機 (調校後)** 模型的 R^2 最高，表明它能更好地解釋房價的變異。

結論

1. **地理位置為王：**鄰近海洋（特別是海灣地區）是房價最重要的驅動因素。沿海地區的房價遠高於內陸地區。
2. **收入是關鍵：**居民的收入中位數與房價有著極強的正相關性，是預測房價的核心指標。
3. **市場存在區隔：**加州房市可依據 `ocean_proximity` 明顯區分為數個次市場，各市場的房價水平與分佈特徵差異顯著。
4. **數據限制：**數據中房價中位數存在 50 萬美元的上限，這可能會影響模型對高價區的預測準確性，在解讀分析結果與模型預測時需將此納入考量。
5. **最佳模型：**梯度提升機模型在經過超參數調校後，展現了最佳的預測性能，可用於未來房價的預測。