

基於 Boston Housing 資料 集之模型改進

任務概述

- 目標：預測波士頓房價（cmedv）並提升模型性能
- 使用模型：SVM、RandomForest、XGBoost
- 評估標準：基於提供的模板，通過調整模型的邏輯、特徵處理或調整參數等方式，改進 SVM、XGBoost 或 RandomForest 的基準性能。在簡報中詳細解釋改進過程並展示最終結果。模型評估以提升幅度為主，並根據報告與展示進行評分。

資料集介紹

- 資料集名稱：BostonHousing2
- 特徵描述：
 - 數值型特徵：lon, lat, crim, zn, indus, nox, rm, age, dis, rad, tax, ptratio, b, lstat,
 - 類別特徵：town, tract, chas
- 目標變數：cmedv
- 不能使用 medv

資料預處理

- 將資料打亂並建立一個 index 列表
- 使用 5-fold Cross Validation :
 - 將資料集分為 5 等分，例：
 - Fold 1: 第 1 到第 100 筆（測試集）
 - Fold 2: 第 101 到第 200 筆
 - 依此類推，最後一組（Fold 5）包含第 401 到第 506 筆
- 每一輪取一組作為測試集，剩下的 4 組作為訓練集

使用模型

- `random_state = 42`
- 使用以下模型作為基準：
 1. Support Vector Machine (SVM)
 2. XGBoost
 3. RandomForest
- 提升幅度以相同模型下的基準結果為標準，評估改進後的性能
 - 若使用 SVM 改進，則比較改進後 SVM 與基準 SVM 的R square等結果。同樣方法適用於 XGBoost 和 RandomForest。

基準評估指標

- 預處理步驟（label_encoding 和 StandardScaler）
- 基準模型整體 R^2
 - SVM : 0.6049
 - RandomForest : 0.8776
 - XGBoost : 0.8339
- 請盡量改善你的模型，使其分數高於基準模型，加油！