國立雲林科技大學

資訊管理研究所

資料探勘專案作業二

指導教授：　許中川　教授

學生：M11223038 陳品佑

M11223032 張祥恩

M11223033 鍾季衡

M11223036 魏冠宇

目錄

表目錄

圖目錄

Data Mining Homework: The Adult, Boston Housing Dataset Mining

摘要

本研究採用機器學習算法對 Adult 和 Boston Housing 兩個資料集進行數值預測。在 Adult 資料集中，本研究預測每週工作時數，並利用 KNN、SVR、Random Forest、XGBoost 等演算法進行比較；而在 Boston Housing 資料集中，本研究以 XGBoost 進行房屋中位價格的預測。實驗結果顯示，在 Adult 資料集中，Random Forest 表現最佳，而在 Boston Housing 資料集中，特徵選擇後的模型績效有所提升。本文不僅展示各演算法的預測效能，還探討特徵重要性對模型性能的影響。

關鍵字：Boston Housing、Adult 資料集、KNN、RandomForest、XGBoost、SVR

Abstract

This study employs machine learning algorithms to perform numerical predictions on two datasets, Adult and Boston Housing. In the Adult dataset, the study predicts the weekly working hours and compares the performance using algorithms such as KNN, SVR, Random Forest, and XGBoost. In the Boston Housing dataset, the study utilizes XGBoost to predict the median house prices. The experimental results indicate that Random Forest performs the best in the Adult dataset, while in the Boston Housing dataset, the performance of the model improves after feature selection. This paper not only demonstrates the predictive performance of each algorithm but also explores the impact of feature importance on model performance.

Keywords：Boston Housing, Adult dataset, KNN, Random Forest, XGBoost, SVR

一、緒論

1.1 研究動機

隨著大數據時代的到來，機器學習在資料分析領域的重要性日益顯著。有效的數值預測不僅對學術研究具有重要價值，同時也在商業、醫療等領域發揮著關鍵作用。因此，本研究旨在運用公開資料集並比較不同機器學習算法在數值預測方面的效能。

1.2 研究目的

本研究的目的是透過實驗來評估和比較不同機器學習算法在數值預測上的表現。透過對 Adult 和 Boston Housing 資料集的分析，本研究希望深入理解各演算法的適用場景及其性能差異。

二、研究方法

本研究使用 KNN、SVR、Random Forest 和 XGBoost 等演算法。針對 Adult 資料集，本研究比較四種演算法在預測每週工作時數上的效能；對於 Boston Housing 資料集，本研究採用 XGBoost 演算法進行房屋中位價格的預測。所有模型的評估均基於 MAPE、RMSE 及 R^2 三項績效指標。

三、研究實驗

3.1.1 Adult 資料集

Adult 資料集包含多種人口統計特徵，本研究將使用其中的年齡、教育程度等作為特徵來預測每週工作時數。下表 1, 2 為 Adult 資料集屬性簡介。

- Training Dataset
- 資料集名稱：Adult
- 資料筆數：32560
- 屬性數量：14

表 1 Adult 資料集屬性簡介

屬性名稱	型態	尺度
age	Int	Ordinal
workclass	String	Nominal
fnlgwt	Int	Nominal
education	String	Ordinal
educationnum	Int	Ordinal
marital-status	String	Nominal

(續下表)

(呈上表)

occupation	String	Nominal
relationship	String	Nominal
race	String	Nominal
gender	String	Nominal
country	String	Nominal
capital-gain	Int	Ratio
hours-per-week	Int	Ordinal
income	String	Ordinal

- Testing Dataset
- 資料集名稱：Adult
- 資料筆數：16281
- 屬性數量：14

表 2 Adult 資料集屬性簡介

屬性名稱	型態	尺度
age	Int	Ordinal
workclass	String	Nominal
fnlgwt	Int	Nominal
education	String	Ordinal
educationnum	Int	Ordinal
marital-status	String	Nominal
occupation	String	Nominal
relationship	String	Nominal
race	String	Nominal
gender	String	Nominal
country	String	Nominal
capital-gain	Int	Ratio
hours-per-week	Int	Ordinal
income	String	Ordinal

3.2.1 前置處理

在進行模型訓練之前，本研究對資料集進行處理，包括缺失值處理、雜訊處理、特徵轉換，進行刪除或填補缺失值，以及識別並處理異常值，將類別型特徵轉換為數值型特徵。

3.3.1 實驗設計

Adult 資料集中，根據每個演算法的特性設定適當參數，使用 MAPE、RMSE、R^2 績效指標評估四種演算法的性能。

3.4.1 實驗結果

根據分析並比較不同演算法在預測每周工作時數上的性能差異，並分別對訓練資料與測試資料做評估(如下表 3 及表 4)。實驗顯示，在 Adult 資料集中，Random Forest 的預測效能最佳。

表 3 Adult 四種演算法之訓練集績效

	KNN	SVR	Random Forest	XGBoost
MAPE	25.94%	30.40%	11.27%	24.78%
RMSE	9.36	11.11	4.15	9.19
R^2	0.43	0.22	0.89	0.45
Training Time(seconds)	0.09s	127.06s	298.41s	0.36s

表 4 Adult 四種演算法之測試集績效

	KNN	SVR	Random Forest	XGBoost
MAPE	32.77%	31.16%	31.02%	29.94%
RMSE	11.60	11.11	11.15	10.77
R^2	0.43	0.21	0.20	0.25

3.1.2 Boston Housing 資料集

Boston Housing 資料集提供波士頓地區房屋的各種特徵，如犯罪率、房屋平均房間數等，為預測房屋中位價格。下表 5 為 Boston Housing 資料集屬性簡介。

- Boston Housing Dataset
- 資料集名稱：Boston Housing
- 資料筆數：506
- 屬性數量：14

表 5 Boston Housing 資料集屬性簡介

屬性名稱	型態	尺度
crim	Float	Ratio
zn	Float	Ratio
indus	Float	Ratio
chas	Int	Nominal
nox	Float	Ratio

(續下表)

(呈上表)

rm	Float	Ratio
age	Float	Ratio
dis	Float	Ratio
rad	Int	Interval
tax	Int	Ratio
ptratio	Float	Interval
b	Float	Ratio
lstat	Float	Ratio
medv	Float	Ratio

3.2.2 前置處理

在進行模型訓練之前，本研究同樣對 Boston Housing 資料集進行前處理，包括處理缺失值、特徵縮放，對數值型特徵進行標準化或正規化。

3.3.2 實驗設計

本研究使用 XGBoost 進行預測，並採用 K-fold 進行交叉驗證，計算每個 fold 的績效，且通過特徵重要性分析以及相關性(如下圖 3, 4)來優化模型，並篩選出 lstat(低收入人群比例)和 rm(住宅平均房間數)兩特徵，測試篩選前後的 MAPE、RMSE 和 R^2 為何。

3.4.2 實驗結果

分析特徵相關係數(下圖 1)及重要性(下圖 2)並比較特徵篩選前後的模型績效，以及 K-fold 設為 5 的每一筆績效(表 6)，提供對特徵重要性的洞見和模型性能比較。在 Boston Housing 資料集中，在特徵選擇後的 XGBoost 模型表現優於未進行特徵選擇時的 XGBoost 模型(下表 7)。

表 6 Boston Housing 之每一個 fold 的預測績效

K-fold	K = 1	K = 2	K = 3	K = 4	K = 5	Mean
MAE	2.1803	2.2656	1.6750	2.4955	1.9288	2.1090
MSE	7.6913	11.7391	5.2357	15.9862	6.8729	9.5052
R^2	0.9222	0.8551	0.9282	0.7948	0.9221	0.8845

表 7 Boston Housing 篩選特徵前後之績效比對

	篩選特徵前	篩選特徵後
MAPE	10.80%	19.01%
RMSE	3.08	5.18
R^2	0.89	0.67

圖 1 Boston Housing 特徵相關係數

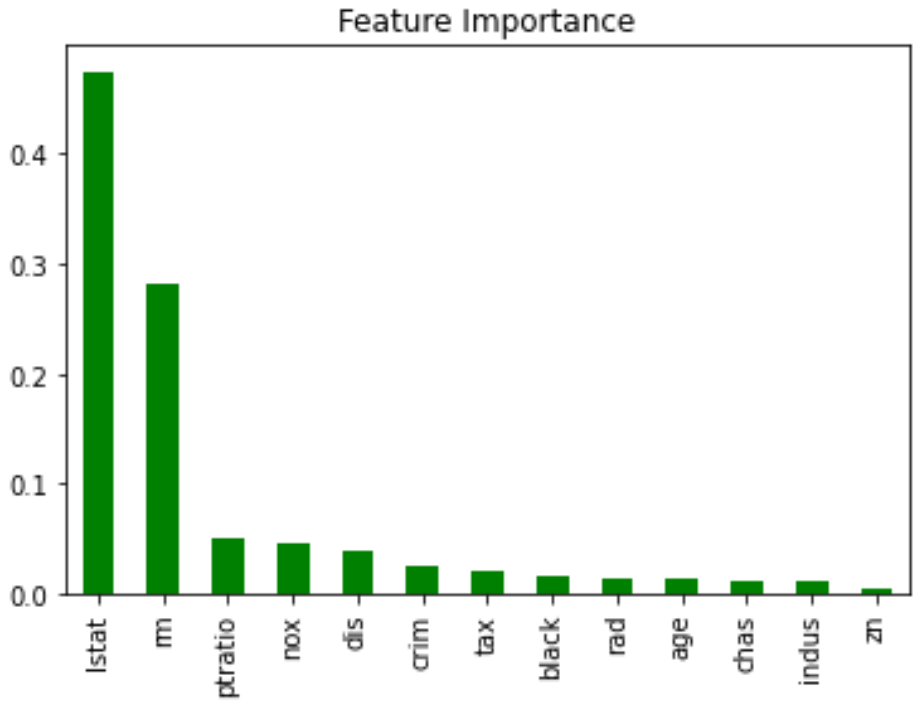

圖 2 Boston Housing 特徵重要性

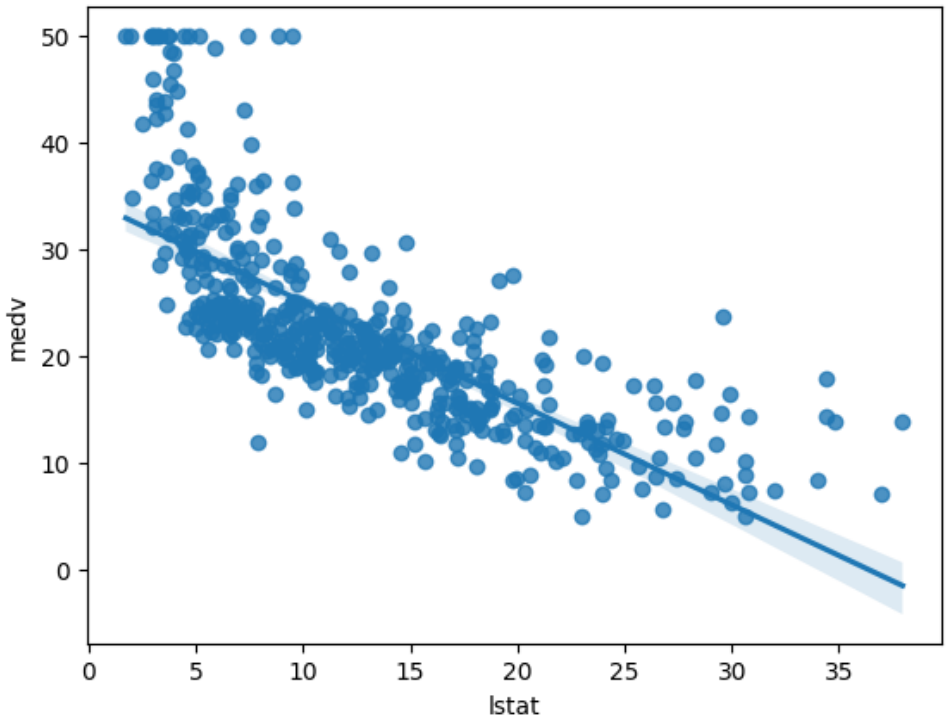

圖 3 lstat (低收入人群比例)對 medv (房價)呈現負相關

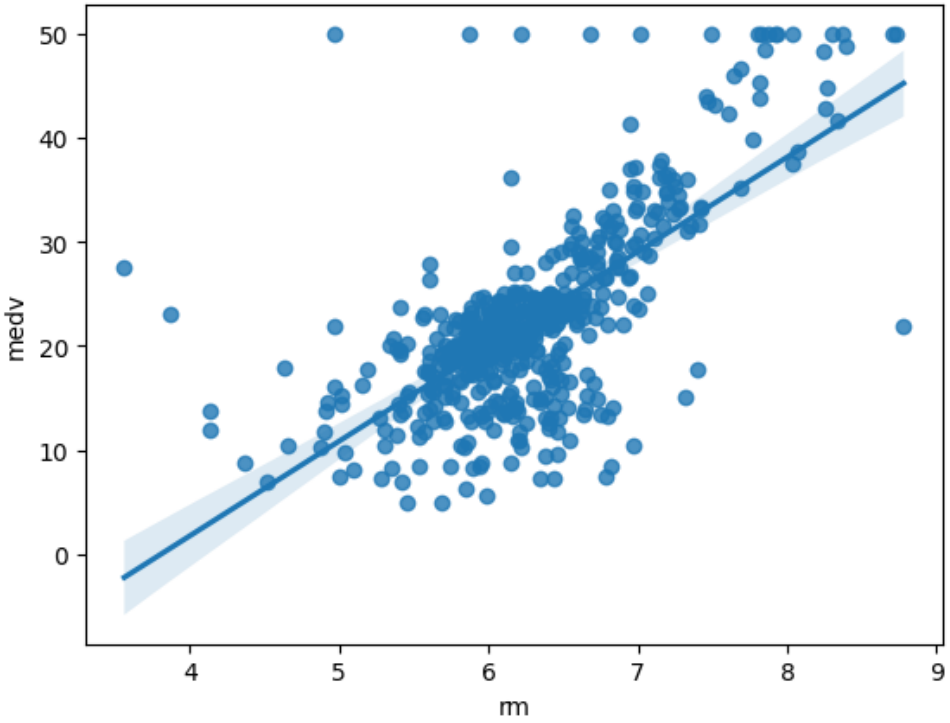

圖 4 rm (平均房間數)對 medv (房價)呈現正相關

四、結論

本研究中採用 XGBoost、SVR、KNN 和 Random Forest 演算法來分析和預測 Adult 資料集中的「每週工作時數」。透過對資料的前處理，包括特徵選擇和資料整理，本研究中確保資料的質量和模型訓練的有效性。在模型訓練過程中，本研究特別做了效率的計算，並透過記錄訓練時間來評估模型在計算時所需要耗費的時長。

XGBoost 作為一種非常高效的機器學習演算法，不僅在訓練速度上表現出色，而且在多種性能指標上均有著優異的預測能力。通過計算 MAPE、RMSE 和 R-square 三個指標，本研究對模型的預測準確度進行評估。這些指標顯示出模型在訓練集和測試集上均達到令人滿意的預測效果，這代表 XGBoost 在處理此類迴歸問題上的強大能力。此外，透過本研究中觀察到資料前處理對模型性能有著重要影響。適當的特徵選擇不僅可以提升模型的預測準確度，還可以減少計算資源的消耗。這一點對於處理大型資料集尤其重要，它能夠顯著提高資料處理的效率和可擴展性。

總結來說，本研究展現了 XGBoost 在解決實際資料問題中的應用潛力，同時也強調資料前處理在機器學習上的關鍵作用。未來的研究可以進一步探索不同類型的特徵工程技術，以及其他更厲害的機器學習演算法在類似問題上的應用，以提高預測模型的準確性和效率。

五、參考文獻

ADVIK MANIAR (2021 年 5 月 17 日)。XGBoost-Model Optimization(94%) Boston Housing。https://www.kaggle.com/code/advikmaniar/xgboost-model-optimization-94-boston-housing。

Ahmed Abdo (2021)。Knn classification for adult income dataset。https://www.kaggle.com/code/ahmedabdo85/knn-classification-for-adult-income-dataset。

Alakh Sethi (2023)。Support Vector Regression Tutorial for Machine Learning。https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/。

Desirahmaz(2022 年 6 月 1 日)。Boston-House-Price-Prediction-with-XGBoost-Model https://github.com/desirahmaz/Boston-House-Price-Prediction-with-XGBoost-Model/blob/main/Boston_House_Price_Prediction_with_XGBoost_Model.ipynb。

SANI KAMAL(2018 年 10 月 3 日)。XGBoost->Boston Housing Dataset https://www.kaggle.com/code/sanikamal/xgboost-boston-housing-dataset?scriptVersionId=6153573 。

Tan Phan(2021 年 6 月 7 日)。XGBOOST with Boston Dataset。https://www.kaggle.com/code/phanttan/xgboost-with-boston-house-dataset/notebook 。

Jack Twain (2014)。Getting training time in scikit。https://stackoverflow.com/questions/22210768/getting-training-time-in-scikit。

NITINESHWAR (2018)。Income prediction using Random Forest and XGBoost。https://www.kaggle.com/code/grayphantom/income-prediction-using-random-forest-and-xgboost。