國立雲林科技大學資訊管理系

資料探勘-作業二

Department of Information Management

National Yunlin University of Science & Technology

Assignment

成人資料集和波士頓房價資料集分析

Adult Datasets and Boston House Price Dataset Analysis

楊欣蓓、陳怡君、鄭皓名

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國112年11月

November 2023

摘要

在成人資料集中，由於具有豐富的特徵屬性加上工時過長的議題在臺灣越來越受到重視，故本研究欲利用該資料集針對每週工作時長做迴歸預測分析，以方便了解工作時長與資料集各特徵屬性之間的關係。主要的預測目標是hours-per-week這項數值。研究首先將成人資料集做前處理，再使用KNN、SVR、RandomForest、XGBoost 四種演算法進行迴歸預測，並透過不同超參數組合做測試，最後，使用 R2、RMSE、MAPE 三種指標及執行時長評估績效，實驗結果顯示 RandomForest 最佳預測模型。

臺灣物價日益上漲，貧富差距大，房價亦逐年增漲，在這樣的時代下，人民想要買到一棟大小適宜的房子愈趨困難，本研究主要欲利用波士頓房價資料集找出與房價相關聯的特徵，研究主要預測的目標為房價(medv)，使用現在機器學習演算法預測最佳的XGBoost模型做房價的迴歸預測。在使用K-fold Cross-Validation技術與否的實驗結果中，發現使用K-fold交叉驗證的平均預測績效(使用三種績效評估指標)比未使用 K-fold 交叉驗證還要差，但能縮小一般化誤差(Generalization Gap)，以提高泛化能力。在刪除資料集特徵欄位的實驗中，依據實驗結果發現刪除特徵欄位後比未進行刪減前還要差，實驗推估可能與特徵欄位的篩選無直接關係，也許與欄位的正規化有關，未來研究可以進一步針對欄位的正規化做進一步研究。

關鍵字：hours-per-week、medv、KNN、XGBoost、SVR、RandomForest、Generalization Gap、K-fold Cross-Validation

一、緒論

1.1 研究動機

本研究透過 Adult Dataset 與 Boston House Price Dataset 來進行研究，Adult Dataset 具有豐富的特徵屬性，可藉由這些資料來進行模型的建構，進而做相關預測，並探索其規律和趨勢。 Boston House Price Dataset 是和波士頓房價相關的資料集，透過此資料集可分析波士頓郊區的生活品質與房價之間的關係，進而豐富此領域相關的知識。

1.2 研究目的

Adult Data Set 是一個常用於深度學習和數據分析的資料集，其中包含了對成年人進行的調查和測試的數據。這個資料集通常用於分析性別、年齡等個人屬性與收入之間的關係，除了可預測一個人的收入水平(分類預測)，也可以進行工作時長的預測分析(迴歸預測)。本研究旨在利用該資料集針對每週工作時長做迴歸預測分析，以方便了解工作時長與資料集各特徵屬性之間的關係。

Boston House Price Dataset 是一個包含波士頓郊區資訊和評估房價的資料集，本研究想要透過此資料集來進行機器學習模型訓練，主要任務是針對波士頓的房價做迴歸預測，而這些資料還可以用於進行特徵分析，即透過對特徵之間的關係進行探索，從13個特徵中找出對目標屬性(房價) 房價影響較大的特徵，從而更好地了解生活郊區和房價之間的關係。

二、實驗方法

2.1 實作說明

本研究對成人資料集使用 KNN、SVR、RandomForest、XGBoost 進行工作時長的預測分析；波士頓房價資料集則是使用 XGBoost、K-fold cross validation 進行房價預測。在進行模型訓練前，做了資料前處理，其中包括刪除重複資料、將名目資料做轉換和數值資料做正規化等，也嘗試的修改模型裡面的架構，像是調整超參數，觀察修改前後的績效有何變化。

2.2 操作說明

本研究執行環境採用 Python 3.11.6，以 Visual Studio Code 作為開發工具，使用 Pandas、Numpy、sklearn、Matplotlib…等，函式資料庫讀取資料、資料前處理以及將模型訓練的績效以視覺化的形式呈現，透過將資料集中的無序性名目資料使用獨熱編碼(One-hot encoding)技術及有序性名目資料用 Label Encoding 技術轉為數值型資料、正規化技術來降低模型過度擬合，以利模型訓練時可以得到較佳的泛化程度(Generalization)，再使用兩種資料集分別做模型訓練與評估績效，最後觀察模型泛化程度。其中，由於波士頓資料集的筆數較小，因此使用 K 折交叉驗證(K-fold Cross-Validation)技術評估績效。

三、實驗設計

3.1 資料集

名稱: 成人資料集

原始資料集: 32561筆(訓練資料)+16281筆(測試資料)=48842筆

資料前處理後: 32537筆(訓練資料)+16276筆(測試資料)=48813筆

表1

成人資料集欄位介紹

欄位	屬性	內容
0	age	continuous
1	workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
2	fnlwgt	continuous
3	education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
4	education-num	continuous
5	marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
6	occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
7	relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
8	race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
9	sex	Female, Male
10	capital-gain	continuous
11	capital-loss	continuous
12	hours-per-week	continuous

（續下表）

（續上表）

13	native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

名稱: 波士頓房價資料集

原始資料筆數：506 筆

前處理後的資料筆數：506 筆

表2

波士頓房價資料集欄位介紹

欄位	屬性	內容
0	crim	per capita crime rate by town
1	zn	proportion of residential land zoned for lots over 25,000 sq.ft.
2	indus	proportion of non-retail business acres per town.
3	chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
4	nox	nitrogen oxides concentration (parts per 10 million).
5	rm	average number of rooms per dwelling.
6	age	proportion of owner-occupied units built prior to 1940.
7	dis	weighted mean of distances to five Boston employment centres.
8	rad	index of accessibility to radial highways.
9	tax	full-value property-tax rate per \$10,000.
10	ptratio	pupil-teacher ratio by town
11	Black(B)	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
12	lstat	lower status of the population (percent)
13	medv	median value of owner-occupied homes in \$1000s

3.2 資料前處理

3.2.1 成人資料集

- 資料前處理：
  - 將資料欄位為「?」的部分，替代成「Nan」，判斷「Nan」的欄位為名目資料或數值資料，前者取眾數做替補，而後者則使用平均值做填補。
  - 刪除意思相近的特徵欄位，如：'education'、'education-num' 取其一。
  - 刪除對於預測年收入較無高度關聯性的欄位 ' fnlwgt '。
  - 刪除資料集空白的筆數。
  - 名目資料的欄位，如：'workclass','education','maritalstatus','occupation', 'relationship','race','sex','native-country','class'，透過 One hot encoding 轉成數值資料。
  - 將名目資料採獨熱編碼方式處理，經 One-hot encoding 後，訓練集的欄位比測試集多出了'native-country_Holand-Netherlands'欄位，故在測試集新增該欄位，並將其值都設為 0，讓兩個資料集欄位數相同。
  - 將有順序性的欄位資料採用 Label Encoding 技術，本次研究針對欄位 Income 做處理，將>50K 設為 1；<=50K 設為 0。
  - 刪除資料集中資料重複的筆數。
  - 數值資料使用 Normalization 技術(z-score)，將欄位 'age','education-num',capital-gain','capital-loss'做處理，降低模型發生 Overfitting 的狀況。

表 3

部分經資料前處理後的成人資料集

資料 / 特徵	No.0	No.1	No.2	No.3	No.4	No.5
Age	0.783411	-0.09358	1.002659	-0.8244	-0.16666	0.710329
capital-gain	0.148292	-0.14598	-0.14598	-0.14598	-0.14598	-0.14598
capital-loss	-0.22971	-0.22971	-0.22971	-0.22971	-0.22971	-0.22971

(續下表)

(續上表)

			. . . .			
sex_ Female	0	0	0	0	1	1
sex_ Male	1	1	1	1	0	0
hours-per-week	13	40	40	40	40	16

3.2.2 波士頓房價資料集

- 資料前處理：
  - 刪除資料集中資料重複的筆數。
  - 檢查資料集中有無缺失值，再依據資料的欄位屬性做補值的處理。由於本資料集的特徵欄位皆為數值型欄位，故取特徵欄位中所以資料的平均數當作補值。
  - 數值資料使用 Normalization 技術(z-score)，將欄位{'CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT'}做處理，以降低模型發生 Overfitting 的狀況。
- 資料分割：

  將資料集切分成70%訓練資料、30%測試資料。

表 4

部分經資料前處理後的的波士頓房價資料集

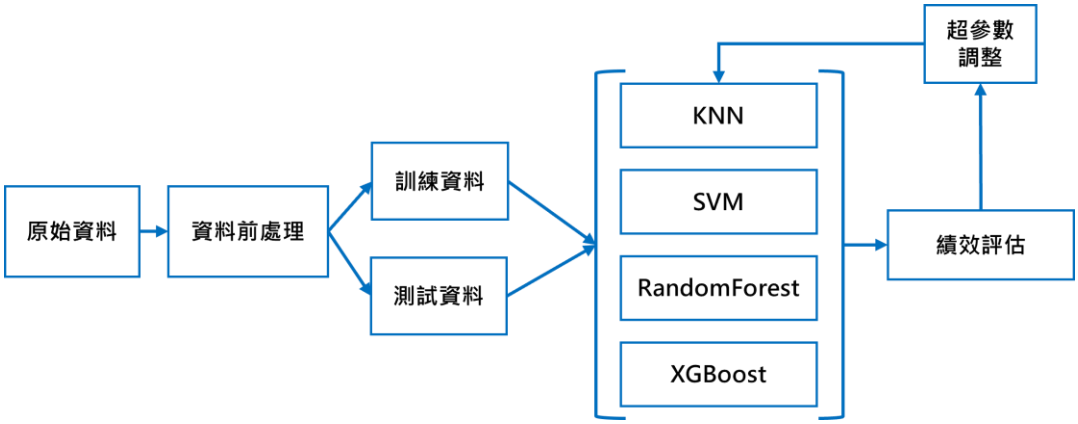
資料 / 特徵	No.1	No.2	No.3	No.4	No.5	No.6
CRIM	-0.42233	-0.41987	-0.41987	-0.41928	-0.41928	-0.41957
ZN	0.296443	-0.48964	-0.48964	-0.48964	-0.48964	-0.48964
INDUS	-1.31101	-0.59977	-0.59977	-1.33043	-1.33043	-1.33043
. . . .						
B	0.441052	0.441052	0.396427	0.416163	0.441052	0.410571
LSTAT	-1.10415	-0.51035	-1.23975	-1.39533	0	-1.07132
MEDV	24	21.6	34.7	33.4	36.2	28.7

3.3 實驗設計

3.3.1 成人資料集

圖 1

成人資料集實驗設計流程圖

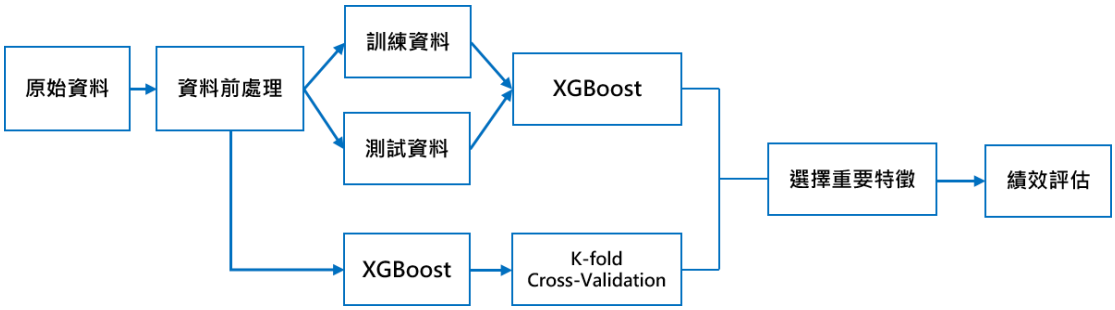


以成人資料集執行的實驗，實驗設計如圖1所示。首先將成人資料集做資料前處理，得到乾淨的資料集後，再分成訓練資料和測試資料，分別使用 KNN、SVR、RandomForest、XGBoost 四種模型進行訓練，接著藉由測試資料來進行績效評估，最後根據績效的結果，對各模型進行超參數調整，再重新評估績效。

3.3.2 波士頓房價資料集

圖 2

波士頓房價資料集實驗設計流程圖



以波士頓房價資料集所進行的實驗，實驗設計如上圖 2 所示， 以兩個部分進行。首先，將切割好的訓練資料丟入 XGBoost 模型中，用測試資料做績效評估，再依據特徵的重要程度，刪除重要性較低的特徵，再次評估績效。由於該資料集資料筆數較少，因此本研究亦使用 K-fold 交叉驗證方法作為績效評估，以提高評估的準確性。

3.4 實驗結果

3.4.1 成人資料集－四種模型的績效比較

實驗一使用四種模型：KNN、RandomForest、SVR、XGBoost 對成人資料集進行每週工時的預測。進行模型訓練時，本研究分別對 KNN、RandomForest、SVR 進行超參數做調整，調整後 R2、RMSE、MAPE 分別最好的績效為 0.26(RandomForest)、10.72(RandomForest)、0.30(SVR)、RunTime(s)最好的為 0.029(KNN)。透過上述綜合數據績效的結果 RandomForest 為表現最好的模型，針對程式執行時長則是 KNN 表現最佳。

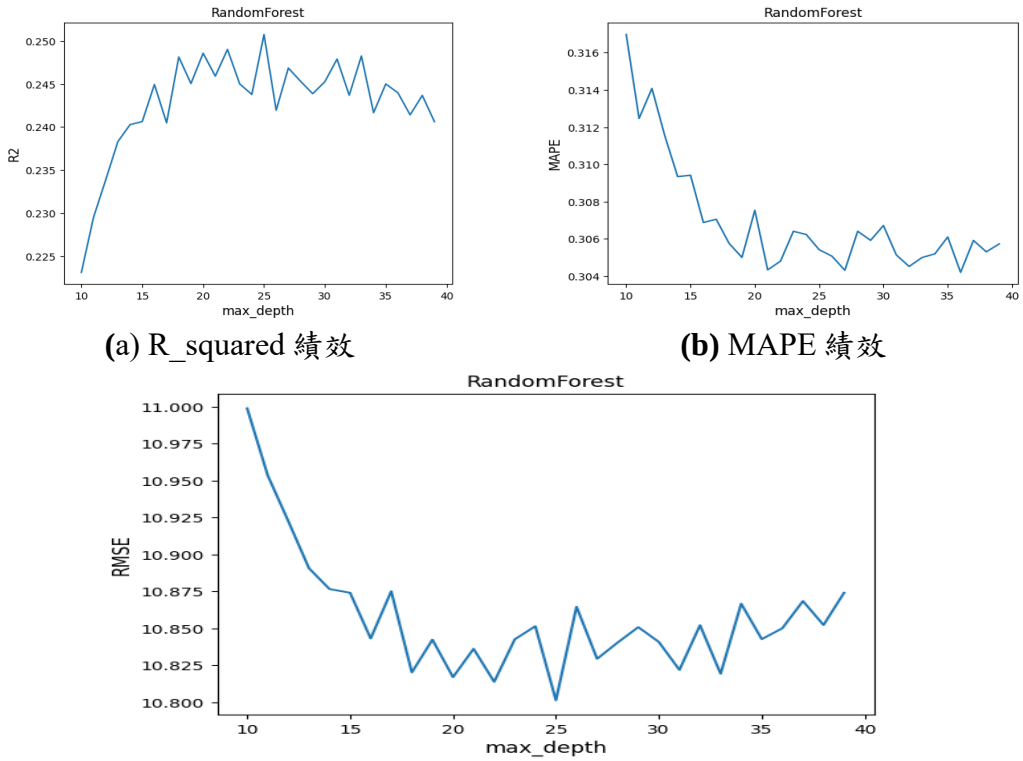
在 RandomForest 的超參數調整中，本研究透過改變深度的方式分析各指標表現，如圖3所示，最後發現當 max_depth = 24時各指標表現最好。min_samples_split 的部分同理，當分裂內部節點所需的最小樣本數為25時，各績效增長最高。另外，當 max_features='log2'，模型尋找最佳分割時要考慮的特徵數量減少，績效並沒有明顯增長；當 max_features=' sqrt'，模型尋找最佳分割時要考慮的特徵數量增加，各個績效表現明顯提升，因此選擇 sqrt。

在 SVR 的超參數調整中，本研究透過設定多組 C、max_iter、cache_size 指標做績效評估，其中 C 為懲罰項，當數值設定越高，容錯率就越低，反之亦然；max_iter 為最高次數、cache_size 為記憶體快取大小。實驗結果為當 C = 200、max_iter = 1000000、cache_size = 1000時，績效表現最佳。

最後，XGBoost 皆使用預設值，由於實驗測試的超參數組合，如：max_depth、eta、gamma，比預設的超參數設定的預測績效還要低，因此在本實驗中以預設的為主。

圖 3

RandomForest 各項績效評估指標



(a) R_squared 績效 **(b)** MAPE 績效

(c) RMSE 績效

表 5

KNN 模型超參數設定

模型	n_neighbors	leaf_size
KNN	25	35

表 6

RandomForest 模型超參數設定

模型	min_samples_split	max_depth	max_features
Random Forest	25	24	'sqrt'

表 7

SVR 模型超參數設定

模型	C	max_iter	cache_size
SVR	200	1000000	1000

表 8

各模型績效表現

模型	R_squared	RMSE	MAPE	Run Time(s)
KNN	0.25	10.81	0.31	**0.029**
Random Forest	**0.26**	**10.72**	0.31	2.075
XGBoost	0.20	11.18	0.33	2.45
SVR	0.22	11.01	**0.30**	314.20

3.4.2 波士頓房價資料集

此實驗包含兩個部分，分為有使用 K-fold 交叉驗證與未使用 K-fold 交叉驗證。實驗使用的 XGBoost 演算法超參數設定兩個部份皆相同，由於實驗測試的超參數組合，如：max_depth、eta、gamma，比預設的超參數設定的預測績效還要低，因此在本實驗中以預設的為主。

在未使用 K-fold 交叉驗證的訓練中，模型訓練績效 RMSE=0.01、MAPE=0.0003、R_squared=0.99， 測試績效 RMSE=2.49、MAPE=0.088、R_squared=0.94，可以發現模型有過擬合 (Overfitting) 的狀況發生，實驗推估是因為資料筆數較少，導致在訓練階段模型的學習過於貼合訓練資料，導致一般化誤差 (Generalization Gap) 提高。因此，本實驗另外使用 K-fold 交叉驗證來避免此狀況的發生。使用 K-fold 交叉驗證的每個 fold 預測績效與平均預測績效如下表9所示，本實驗將 k 值設為5。

表 9

Performance of K-fold Cross-Validation

	RMSE	MAPE	R_squared
Fold 1	2.90	0.10	0.87
Fold 2	3.62	0.15	0.84
Fold 3	2.28	0.09	0.92
Fold 4	2.19	0.09	0.95
Fold 5	3.89	0.13	0.85
Average	2.97	0.11	0.89

本實驗為了找出影響預測目標(MEDV)較大的特徵屬性，以讓模型預測績效提高，實驗觀察了波士頓房價資料集中各特徵屬性的重要程度，如下圖4所示，可以看出重要程度較低的特徵前四名為：ZN、B、CHAS 及 INDUS， 透過逐一刪減特徵屬性後，其預測績效變化如下表10所示。可以看出在進行特徵欄位刪減後，預測績效比未進行特徵欄位刪減還要好，如下圖5所示，圖5為迴歸預測分布圖，從該圖可以得知，在未進行特徵欄位刪減時，預測值與實際值的差距較大，然而在進行特徵欄位刪減後，縮短了預測誤差，證明特徵欄位刪減確實能提高模型預測績效，但若刪除與目標欄位關聯性較大的特徵時，可能會導致預測績效比未做特徵欄位刪減時還要低。

圖 4

資料集的特徵重要性

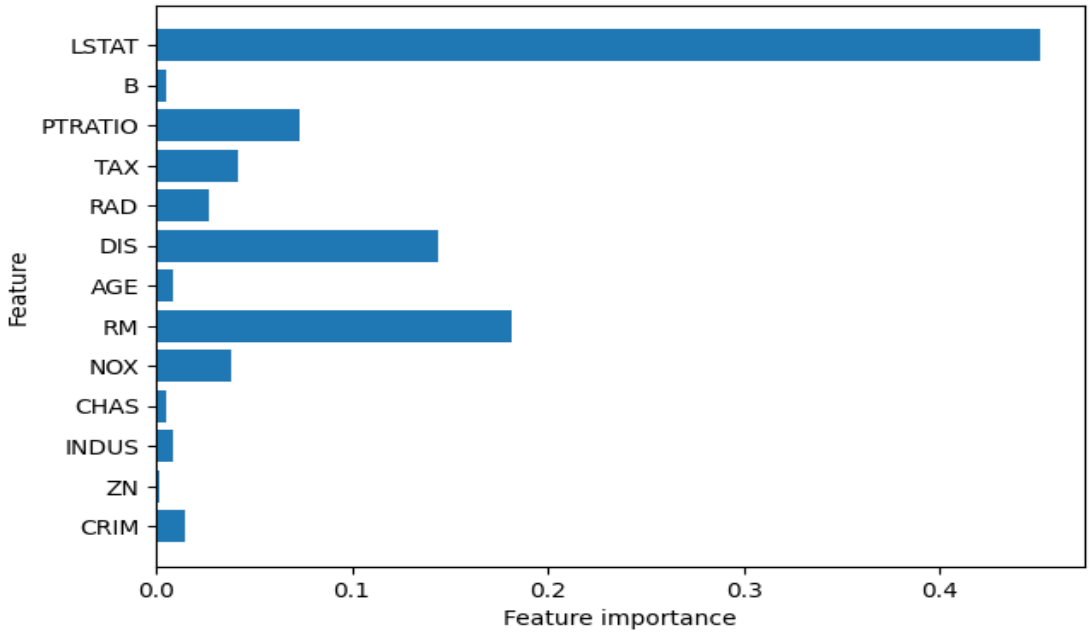


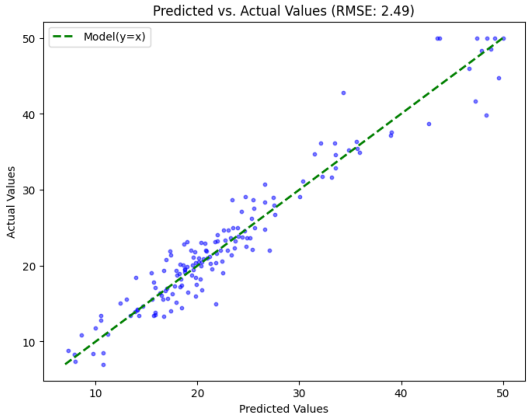
表 10

刪減特徵後的預測績效

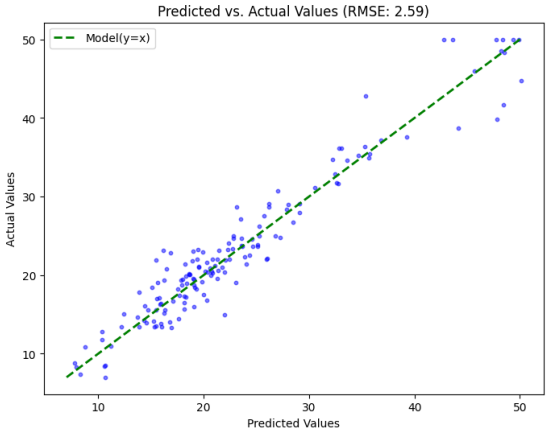
特徵	RMSE	MAPE	R_squared
ZN	2.51	0.09	0.934
B	2.59	0.10	0.930
CHAS	2.58	0.09	0.930
INDUS	2.59	0.09	0.930

圖 5

(a) 特徵刪除前迴歸預測 (b) 特徵刪除後迴歸預測



(a)



(b)

結論

在成人資料集中，主要是透過 KNN、RandomForest、XGBoost、SVR 四種演算法去各別預測 hours-per-week 的績效為何，在調整各個演算法的超參數後，調整後 R2、RMSE、MAPE、RunTime(s)分別最好的績效為0.26(RandomForest)、10.72(RandomForest)、0.30(SVR)、0.029(KNN)，透過上述本研究發現 RandomForest 是表現最好的模型。

在波士頓房價資料集中，主要透過 XGBoost 演算法預測房價，在使用 K-fold Cross-Validation 技術與否的實驗結果中，發現使用 K-fold 交叉驗證的平均預測績效(使用三種績效評估指標)比未使用 K-fold 交叉驗證還要差，但使用 K-fold 交叉驗證後，能縮小一般化誤差，以提高泛化能力。在刪除資料集特徵欄位的實驗中，依據實驗結果發現刪除特徵欄位後，不管是哪一種績效衡量指標都比未進行刪減前還要差，實驗推估可能與特徵欄位的篩選無直接關係，也許與欄位的正規化有關，未來研究可以進一步針對欄位的正規化做進一步研究。

參考文獻

Boston Housing. Kaggle. https://kaggle.com/competitions/boston-housing

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. *Economics and Management 5*, 81–102.

Ronny Kohavi、Barry Becker (1996)。成人數據集。
https://archive.ics.uci.edu/ml/datasets/adult

Ryan Lu (2018)。Preprocessing Data : 類別型特徵_OneHotEncoder & LabelEncoder 介紹與實作。
https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/preprocessing-data-onehotencoder-labelencoder-%E5%AF%A6%E4%BD%9C-968936124d59