國立雲林科技大學資訊管理系

資料探勘

National Yunlin University of Science and Technology

Department of Information Management

專案作業二

以三種演算法預測成人人口普查資料集與首爾共享自行車需求資料集

M11123048 陳韋彤

M11123060 毛俐蘋

M11123019 張芸婷

M11123026 林宥昇

指導老師：許中川

Advisor：Chung-Chain Hsu，Ph.D.

中華民國 111 年 11 月

November 2022

摘要

本研究透過不同迴歸演算法進行資料探勘，對 adult 資料集以及 Seoul Bike Sharing Demand 資料集進行探討，為了看出各項數據分別對每周的工作時數和首爾共享自行車租借數量的影響。本研究經計算不同演算法下的重要特徵值後，以刪除較重要的特徵值的方式，以 RMSE、MAE、MAPE 作為評估的績效指標，評估刪除重要績效指標前與刪除後的績效差異。Adult 資料集中，多數演算法中，三項績效指標皆呈下降，僅 Random forest 的 MAPE 上升；Seoul Bike Sharing Demand 資料集中，三種演算法的 RMSE、MAE 皆下降，但 KNN 及 Random forest 的 MAPE 上升，僅 SVR 的 MAPE 績效呈下降結果。

關鍵字：回歸分析、資料探勘、工作時數、自行車需求

一、 緒論

1.1 動機

透過大數據去分析及進行預測是現在的趨勢，預測的方法有很多，其中本次研究選擇使用三種演算法建構數值預測模型，並比較不同演算法之預測績效。本研究選用了 adult 資料集與 Seoul Bike Sharing Demand 資料集進行探勘。adult 資料集提供了人口普查的各項數據，使我們能夠探討各項因素對於一週的工作時數的影響；Seoul Bike Sharing Demand 資料集是有關於首爾租借共享自行車的案例，本研究藉此探討天氣的各項數據對於人們租借自行車數量的關聯性。

1.2 目的

本次研究旨在透過三種演算法建構數值預測模型對 adult 的資料集的大數據進行迴歸分析，並透過數據中的職位、教育、國籍、婚姻狀況、年齡等因素去判斷對於一週工作時數的影響；對 Seoul Bike Sharing Demand 資料集進行迴歸分析，根據過去溫度、濕度、能見度等資料作為判斷依據，進而分析各項因素對於共享自行車租借數量的影響。

二、 方法

本研究實驗流程如圖 1 所示。

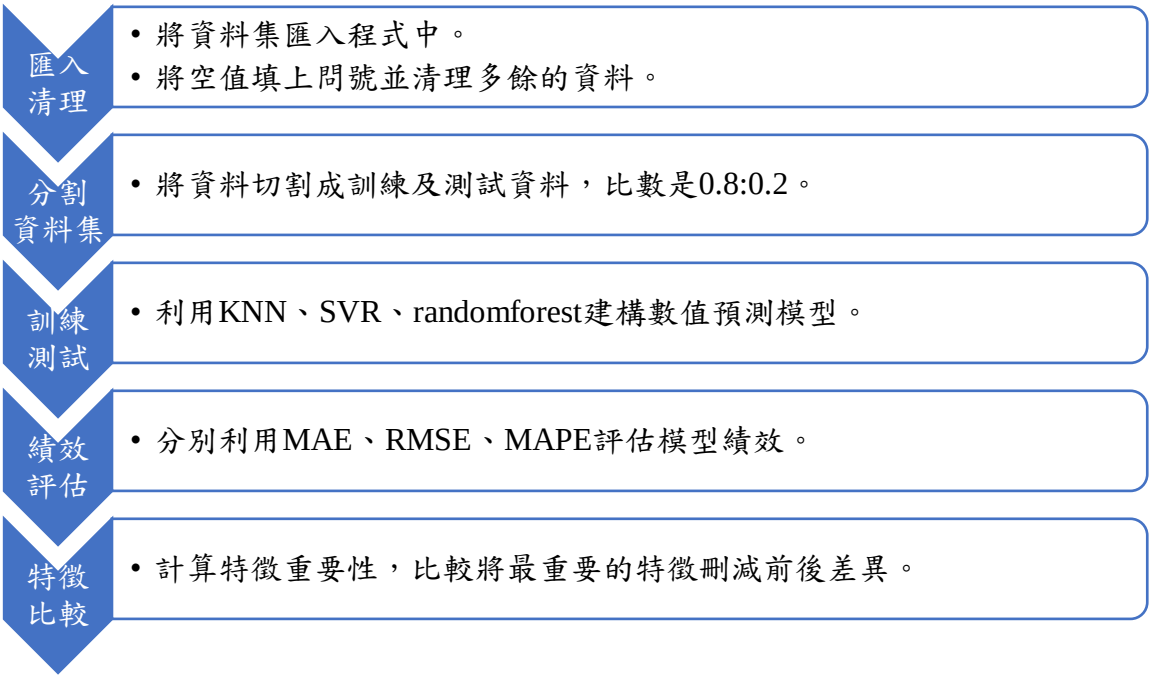

圖 1 實驗方法流程圖

三、 實驗

3.1 資料集

Adult 資料集是從 1994 年的人口普查數據庫中做提取，用以預測各項因素對於一週的工作時數的影響。共有 14 種屬性，詳細屬性資訊如表 1 所示。

表 1

Adult 資料集屬性資料

屬性名稱	詳細資料描述
age	continuous.
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Fnlwgt	continuous.
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.
education-num	Continuous.
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
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.
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Female, Male.
capital-gain	Continuous.
capital-loss	Continuous.
hours-per-week	Continuous.
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.

Seoul Bike Sharing Demand 資料集針對首爾自行車的租借數量進行預測，分析各項因素對於預測欄位之重要性。共有 10 種屬性，詳細屬性資料見表 2 所示。

表 2

Seoul Bike Sharing Demand 資料集屬性資料

屬性名稱	詳細資料描述
Rented Bike count	Count of bikes rented at each hour
Hour	Hour of he day
Temperature	Temperature in Celsius
Humidity	%
Windspeed	m/s
Visibility	10m
Dew point temperature	Celsius
Solar radiation	MJ/m2
Rainfall	mm
Snowfall	cm

3.2 前置處理

Adult 和 Seoul Bike Sharing Demand 資料集都用「?」代替缺失的資料，並且刪除空值。Adult 資料及還有再利用 Label Encoder 進行資料轉換，針對職位、教育程度、婚姻狀況、關係、種族、性別、國籍、收入等非數值屬性的資料進行資料轉換，將字串資料轉換成有順序大小之分的數字類別資料，以利後續的分析。

3.3 實驗設計

在進行 Adult 資料集的分析時，本研究以「一週工作時數」作為預測結果進行線性迴歸分析預測，Seoul Bike Sharing Demand 資料集以「自行車租借數量」作為預測結果進行預測。KNN 根據每個值的最鄰近的 k 個數據點的均值作為預測值，本研究將 k 值設計從 1 到 30 做測試；Random forest 透過改變森林的樹木數量進而評估績效指標，本研究設計以 50、100、150 棵探討預測績效。Adult 資料集下，利用 SVR 演算法探討績效，在 Seoul Bike Sharing Demand 資料集則是以 XGBoost 探討績效。兩種演算法皆以 RMSE、MAE、MAPE 作為指標進行績效評估。

3.4 實驗結果

3.4.1 Adult 資料集

在 Adult 資料集中，KNN 演算法下是根據每個值的最鄰近的 k 個數據點的平均值作為預測值。經過測試 1 到 30 的績效後得出，k 值使用 30 可獲得最好的績效，如圖 2 所示，故選擇 30 作為 k 值進行預測。

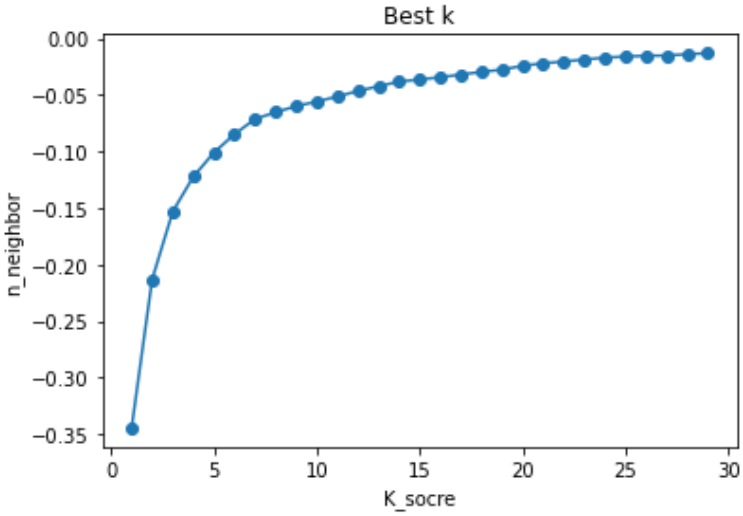

圖 2 Adult 資料集下不同 K 值的績效

Adult 資料集中，KNN 根據重要特徵成績結果，如圖 3 所示，將重要的特徵刪除，再重新進行分析、了解差異，刪除的特徵為「capital-gain」，刪除績效指標前後比較如表 3 所示。

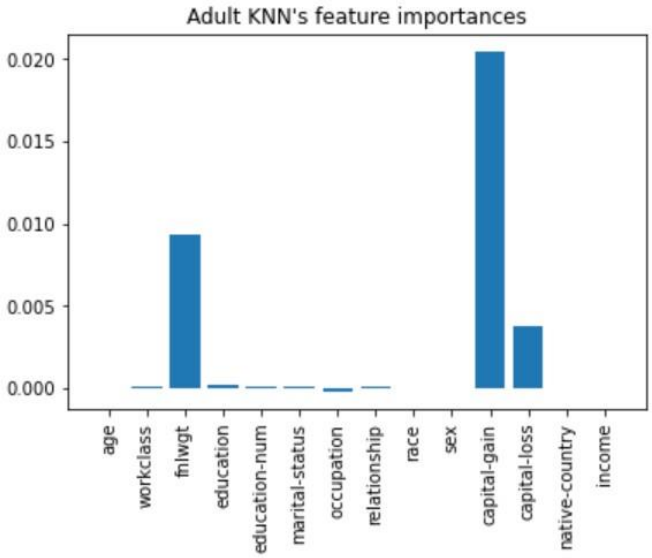

圖 3 Adult 資料集下 KNN 的特徵重要性

表 3

Adult 資料集 KNN 刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	12.56	12.64
MAE	8.27	8.32
MAPE	0.37	0.38

在 Random forest 演算法中，本研究以 50 棵、100 棵、150 棵樹建立森林比較績效指標。根據 RMSE、MAE、MAPE 指標比較結果，如表 4 所示，決定以績效較好的 150 棵樹作為後續的研究數值。

表 4

Adult 資料集 Random forest 下不同棵樹的績效指標比較表

績效指標	50 棵樹	100 棵樹	150 棵樹
RMSE	4.22	4.13	4.10
MAE	2.89	2.85	2.83
MAPE	11.33%	11.19%	11.19%

根據 Random fores 演算法下重要特徵成績結果，如圖 4 所示。隨機森林演算法中 150 棵樹下，將重要的特徵刪除，再重新進行分析，了解差異。刪除的特徵為「age」、「fnlwgt」，刪除前後績效指標比較如表 5 所示。

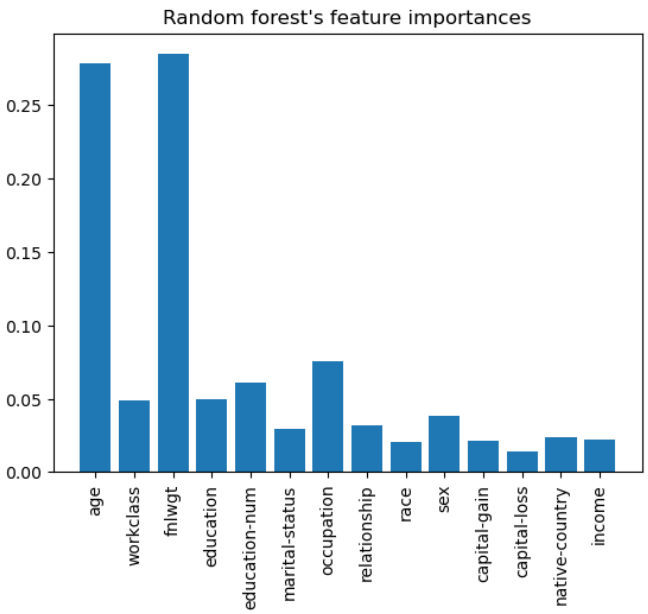

圖 4 Adult 資料集 Random forest 下的重要特徵值長條圖

表 5

Adult 資料集 Random forest 下刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	4.10	8.70
MAE	2.83	5.77
MAPE	11.19%	0.23%

Adult 資料集中，在 XGBoost 演算法下，計算出重要特徵屬性值，如圖 5 所示。本研究決定刪除「sex」、「income」、「age」等較重要的特徵，並根據 RMSE、MAE、MAPE 指標比較結果，刪除前後績效指標比較如表 5 所示。

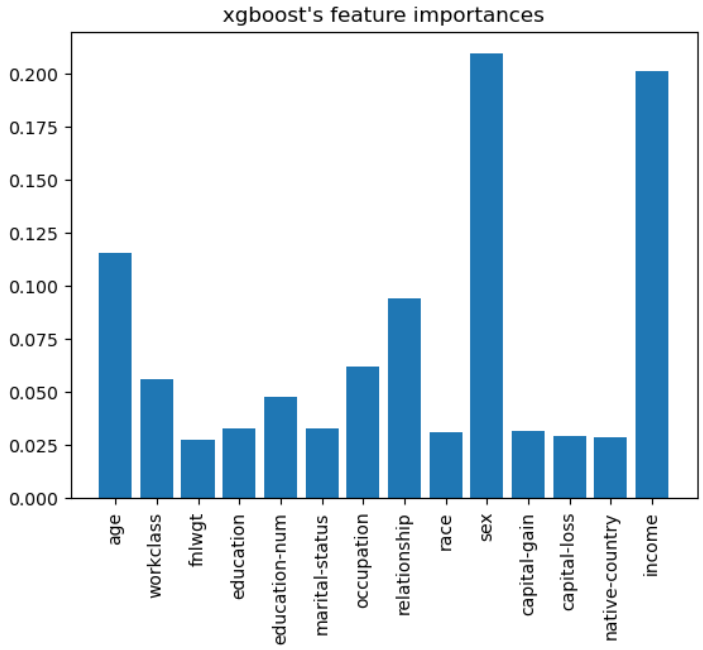

圖 5 Adult 資料集 XGBoost 下的重要特徵值長條圖

表 5

Adult 資料集 XGBoost 下刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	10.86	11.26
MAE	7.44	7.69
MAPE	0.29	0.326

3.4.2 Seoul Bike Sharing Demand 資料集

在 Seoul Bike Sharing Demand 資料集中，KNN 演算法下根據每個值的最鄰近的 k 個數據點的均值作為預測值，經過測試 1 到 30 的績效後得出，k 值使用 7 可獲得最好的績效，如圖 6 所示，故選擇 30 作為 k 值進行預測。

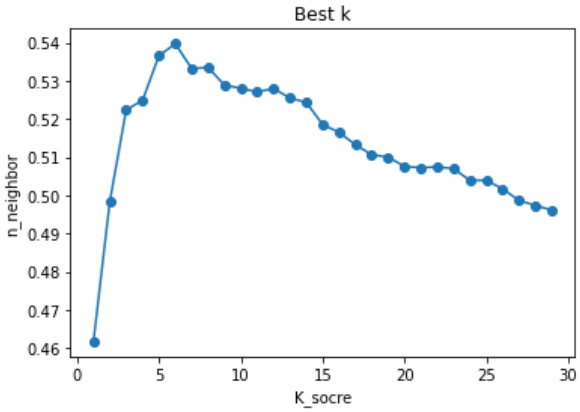

圖 6 Seoul Bike Sharing Demand 資料集下不同 K 值的績效

在 Seoul Bike Sharing Demand 資料集中，KNN 演算法根據重要特徵成績結果，如圖 7 所示。本研究將重要的特徵刪除，刪除的特徵為「Humidity」，刪除特徵後再重新進行分析，了解差異。刪除績效指標前後比較如表 6 所示。

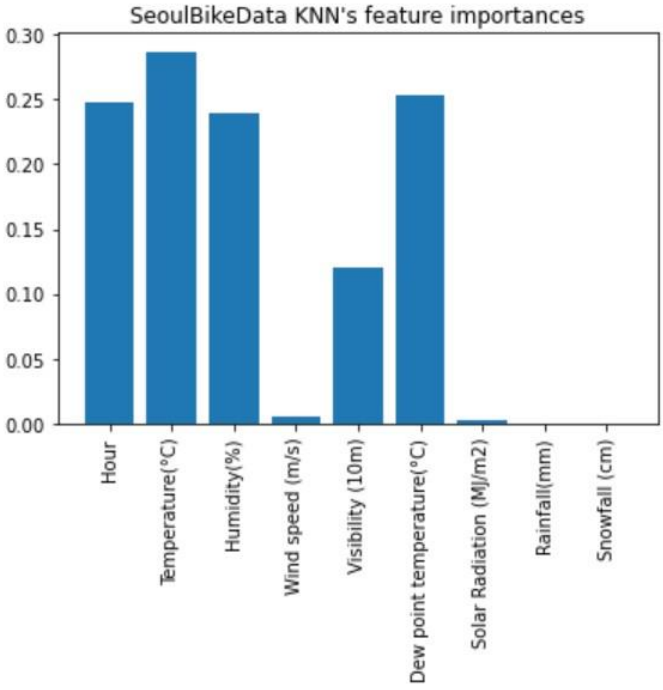

圖 7 Seoul Bike Sharing Demand 資料集下 KNN 的特徵重要值

表 6

SeoulBikeData 資料集 KNN 刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	437.92	451.17
MAE	294.89	301.69
MAPE	1.07	1.05

在 SeoulBikeData 資料集中，Random forest 演算法下，本研究以 50 棵、100 棵、150 棵樹建立森林比較績效指標。根據 RMSE、MAE、MAPE 指標比較結果，如表 7 所示，因此本研究決定以績效較好的 150 棵樹作為後續的研究數值。

表 7

SeoulBikeData 資料集 Random forest 下不同棵樹的績效指標比較表

績效指標	50 棵樹	100 棵樹	150 棵樹
RMSE	119.52	116.70	115.84
MAE	70.43	69.32	68.84
MAPE	46170173344246296.00 %	46150399405813904.00 %	46218165470412392.00 %

根據 Random fores 演算法下重要特徵成績結果，如圖 8 所示。隨機森林演算法中 150 棵樹下，將重要的特徵刪除，再重新進行分析，了解差異。刪除的特徵為「Hour」、「Temperature(℃)」，刪除前後績效指標比較如表 8 所示。

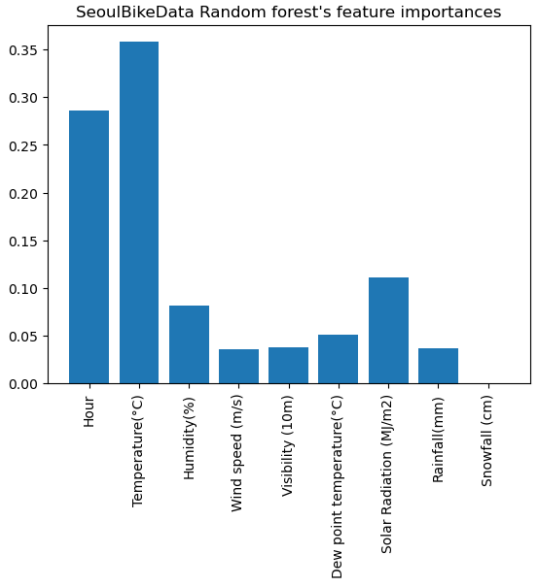

圖 8 SeoulBikeData 資料集 Random forest 下的重要特徵值長條圖

表 8

SeoulBikeData 資料集 Random forest 下刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	115.84	151.50
MAE	68.84	100.71
MAPE	46218165470412392.00 %	43533287343587600.00 %

在 SeoulBikeData 資料集中，SVR 演算法下，本研究以改變 C 的值比較績效指標。根據 RMSE、MAE、MAPE 指標比較結果，如表 9 所示，因此本研究決定以績效較好的 C=1000 作為後續的研究數值。

表 9

SeoulBikeData 資料集 SVR 下不同 R 值績效指標比較表

績效指標	C=950	C=1000	C=1500
RMSE	449.32	358.03	448.34
MAE	294.65	205.57	293.17
MAPE	127252890867923296.00 %	138503241759456608.00 %	126355574947756848.00 %

根據 SVR 演算法下重要特徵成績結果，如圖 9 所示。將重要的特徵刪除，再重新進行分析，了解差異。刪除的特徵為「Hour」、「Temperature(℃)」，刪除重要績效指標前後比較如表 10 所示。

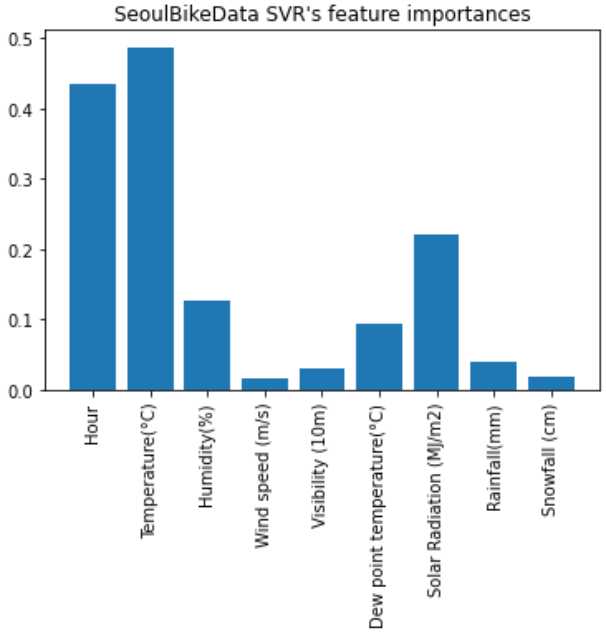

圖 9 SeoulBikeData 資料集 SVR 下的重要特徵值長條圖

表 10

SeoulBikeData 資料集 SVR 下刪除特徵前後比較

績效指標	刪除特徵前	刪除特徵後
RMSE	358.03	434.89
MAE	205.57	282.96
MAPE	138503241759456496.00 %	138503241759456496.00 %

四、 結論

根據結果所示，Adult 資料集的 KNN 演算法刪除重要績效指標，RMSE、MAE、MAPE 數值明顯下降，故可得知績效變差；Random forest 演算法下，RMSE、MAE 績效降低但 MAPE 增加；XGBoost 演算法下可看出，RMSE、MAE、MAPE 的數值皆明顯下降。

SeoulBikeData 資料集的中 KNN 演算法下可看出，RMSE、MAE 的績效下降，但 MAPE 績效上升；Random forest 演算法下，RMSE、MAE 績效下降，但 MAPE 績效上升；SVR 演算法下，RMSE 與 MAE 績效下降，但 MAPE 績效值無改變。

參考文獻

(2016, January 26). *default of credit card clients Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

(1996, May 1). *Adult Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Adult

(2007-2022). *1.10. Decision Trees*. scikit-learn. https://scikit-learn.org/stable/modules/tree.html

Tonykuoyj (2016 年 12 月 23 號)。 *[第 23 天] 機器學習（3）決策樹與 k-NN 分類器*。iT 邦幫忙。https://ithelp.ithome.com.tw/articles/10187191

CLY (2020 年 3 月 17 號)。*Python Pandas 匯出成 CSV 檔/Xlsx 檔*。CYL 菜鳥攻略。https://dotblogs.com.tw/CYLcode/2020/03/17/175255

古耕全 (2020 年 11 月 1 號)。*[Pandas 教學] 資料分析必懂的 Pandas DataFrame 處理雙維度資料方法*。LEARN CODE WITH MIKE。https://www.learncodewithmike.com/2020/11/python-pandas-dataframe-tutorial.html