國立雲林科技大學資訊管理系

資料探勘

作業二

使用迴歸演算法預測 Adult 與 Bank 資料集

M11023044 沈俊良

M11023006 謝媄衣

M11023061 王成綱

D10630002 陳冠臻

指導教授：許中川

2022 年 11 月 22 日

摘要

本研究先採用 Adult 測試資料集準確率後，由於 Adult 測試集準確率(accuracy)較低，準確率為 0.88 之準確率;接著透過 bank-additonal 資料集進行驗證，使用 MS SQL Server 進行資料清洗、正規化資料，訓練集與測試集切割為 7:3，透過 Python 訓練演算法(SVR、KNN、Random Forest、XG Boost)、迴圈找出訓練集之最佳特徵屬性，並計算出 MAE、RMSE、MAPE，最後測試資料集準確率(accuracy)為 0.99。

關鍵字: SVR、KNN、Random Forest、XG Boost、accuracy、MS SQL Server、MAE、RMSE、MAPE

一、 緒論

1.1. 動機

烏俄戰爭使得歐洲經濟衰退，2022 年經濟合作暨發展組織（OECD）最新報告「付出戰爭的代價」，直指全球經濟正在為烏俄戰爭付出高昂代價。OECD 預測全球經濟成長率將在 2022 下半年持續低迷，2023 年進一步減速至 2.2%。報告中描繪黯淡的全球經濟前景：商業信心、可支配所得和家庭支出均直線下降，而燃料、食品和交通成本持續飆升。在經濟不確定的大環境底下，又同時留意新冠疫情有可能在冬天又再度捲土重來，因此銀行業為了維持業務開發，必須選擇適合的行銷方法，一邊能提高成交率，持續與客戶維持關係，另一邊則保護銀行從業人員能與減少與客戶見面接觸次數，減低染疫的風險。而電話行銷就是一種透過電話、傳真等通訊技術，來實現有計劃、組織，並且高效率地擴大顧客群，同時能提高顧客滿意度和維持顧客等市場的行銷手法之一，而在當前新冠病毒肆虐全球時，就可以藉由電話的通訊技術來減低染疫的風險。電話行銷並非是隨機打出大量電話，或是靠運氣去推銷幾樣產品的行銷手段，當商品的選擇或是獲取商品管道越來越多元時，消費者就會漸漸開始重視產品的附加價值，並在選擇上更願意著重自己所認同的價值上，而非僅僅只關注商品的基本功能與價格，因此行銷決策者對於消費者的相關脈動與不同層次的客戶的需求都必須有所掌握。本研究挑選一個西班牙銀行營銷數據集，希望以當前歐洲正深陷經濟走低與疫情風險可能再次升高的情況之下，幫助銀行業透電話進行業務開拓。在開始聯繫客戶之前，必須積極且用心的了解市場信息和對於相關數據進行分析與吸收，並找從客群資料中尋找明顯的消費習慣與屬性特質，再根據其結果設計出吸引客群的行銷產品與理財規劃服務，而這也是本次研究的首要目標。

1.2. 目的

傳統大量行銷策略逐漸轉向顧客導向的行銷方向，與顧客進行更直接、互動關係更頻繁方式進行溝通。而多樣化與個人化之溝通方式，可使個別的消費者發展長期互惠的關係，來達到客戶滿意度的提升和創造更多顧客價值，而非僅發展成買與賣的單項式交易關係。資料庫分析結合電話來進行行銷活動，是一種被動與主動行銷的戰略組合，這種行銷方式比起其他單一方式的數位行銷更具個性化。而過往行銷策略主要拜科技的進步和網際網路的發達，因此數據大多都是從網路來取得，但在資料搜集的過程中容易犯“垃圾進，垃圾出”的錯誤，但本研究數據是根據與電訪後的資料分析，而非僅是網路數據，因此在數據品質上更可貼近消費者真實行為，就可在相關的資料中找到明確擬定更精準的行銷商品，來打動已原本已是固定的客群，將穩固的關係轉化成具體投資了行動。

1.3. 方法

將原始之 bank-additonal 與 Adult 資料集匯入至 SQL Server，透過 SQL 語法編譯，將名目資料進行正規化處理，再依亂數抽出訓練集與測試集。透過 Python 計算各演算法(SVR、KNN、Random Forest、XG Boost)，排序各特徵屬性，針對各演算法找出最佳準確率及最佳 MAE、RMSE 及 MAPE 值，依據迴圈寫入各演算法訓練方法，再依迴圈測試，進而在此資料集中找出最佳演算法。

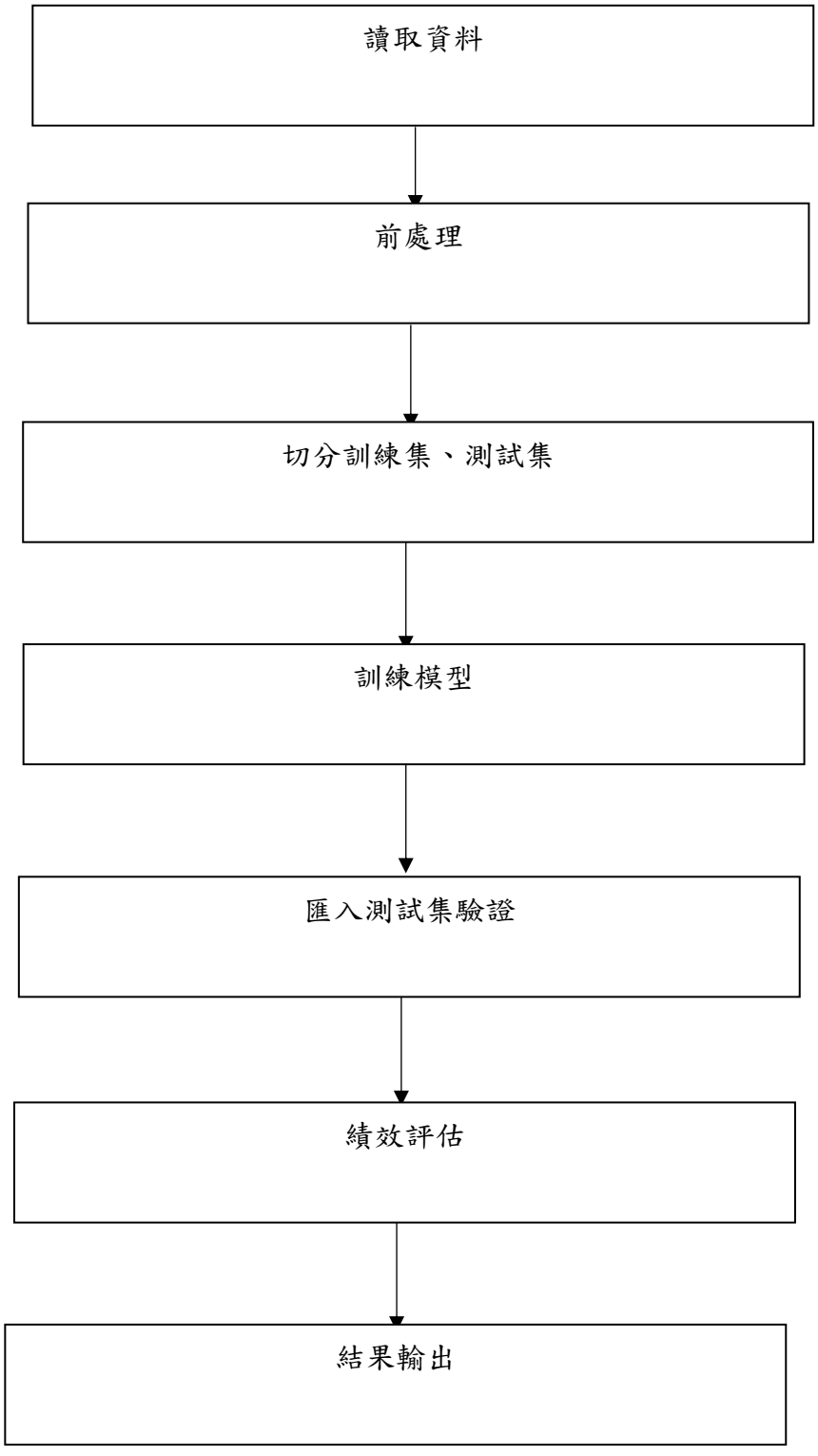

圖 1 bank-additonal 資料集預測流程圖

二、 方法

2.1. 資料集

表 1

bank-additonal 資料屬性一覽

欄位名稱	譯名	型態
age	年齡	Integer
job	工作	String
marital	婚姻狀況	String
education	教育	String
default	違約	String
housing	住房	String
loan	貸款	String
contact	聯繫人	String
month	聯繫月份	String
day_of_week	一周最後聯繫日	Integer
duration	最後一次聯繫持續時間	Integer
campaign	活動	String
pdays	最後一次聯繫客戶天數	Integer
previous	以前客戶聯繫次數	Integer
poutcome	以前營銷結果	Integer
emp.var.rate	就業變化率(季)	Integer
cons.price.idx	消費者價格指數(月)	Integer
cons.conf.idx	消費者信心指數(月)	Integer
euribor3m	三個月利率(日)	Integer
nr.employed	雇員人數(季)	Integer
樣本數		屬性個數
45211		20

表 2

Adult 資料屬性一覽

欄位名稱	譯名	型態
age	年齡	Integer
workclass	工作類別	String
fnlwgt	fnlwgt	Integer
education	教育程度	String
education-num	受教育學齡	Integer
marital-status	婚姻狀況	String
occupation	職業	String
relationship	關係	String
race	種族	String
sex	性別	String
capital-gain	資本收益	Integer
capital-loss	資本損失	Integer
hours-per-week	每周小時	Integer
native-country	國籍	String
Listing of attributes	屬性	String

樣本數	屬性個數
32652	15

2.2. 前置處理

將 Adult 資料集之原始資料所記錄非連續性資料，轉換為 One Hot encoding，但 One Hot encoding 無法直接對字串進行編碼，必須先透過 One Hot encoding 將字串以數字取代後再進行 One Hot encoding 處理，接著刪除缺失資料與轉換格式，最後再進行正規化。bank-additonal 資料集由於未切分出訓練集與測試集，則需進行資料分割，接著將原始資料做正規化進行預測。

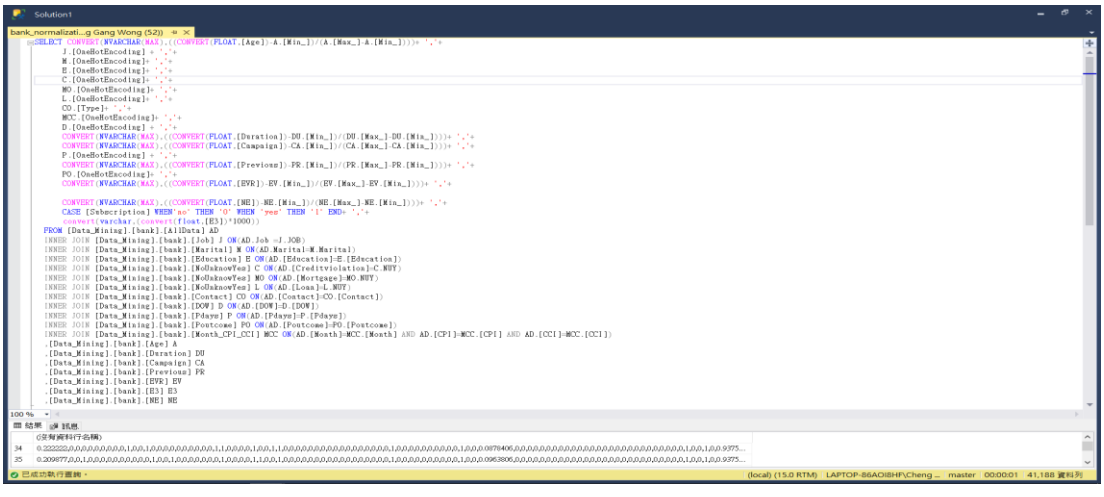

圖 2 SQL Server 前置處理 (詳情請見 github)

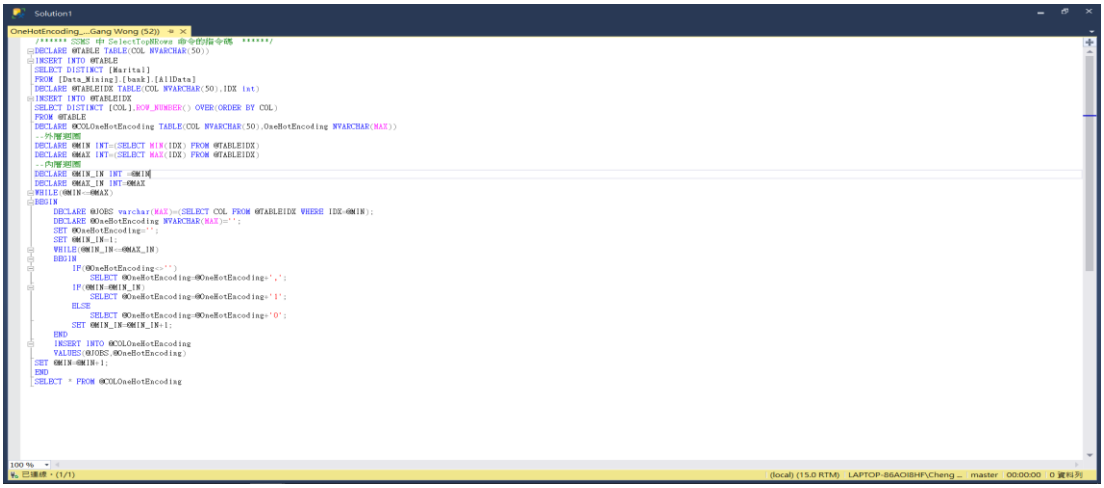

圖 3 SQL Server 前置處理 (詳情請見 github)

2.3. 實驗設計

本研究是在 Anaconda3_spyder(Pythonversion=3.10)環境，以及 Visual Studio Code 下進行開發，使用套件 K-Nearest Neighbor Regressor、Random Forest Regressor、SVR、XG Boost Regressor 調整各演算法線性之參數。並找出在各演算法下之最佳屬性。

2.4. 實驗結果

本研究 Adult 資料集中之準確率，利用 KNN 預測準確率(accuracy)，Accuracy:0.999，KNN 之 n_neighbors 為 3。另預測出 KNN 之 MAE、RMSE 與 MAPE 之值。

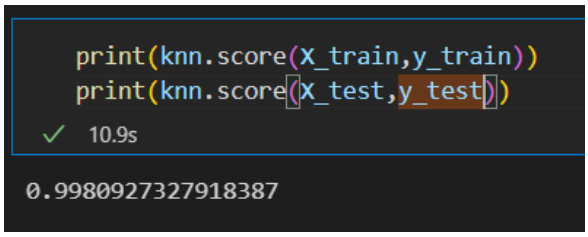

圖 4 KNN 預測結果 n_neighbors 為 1

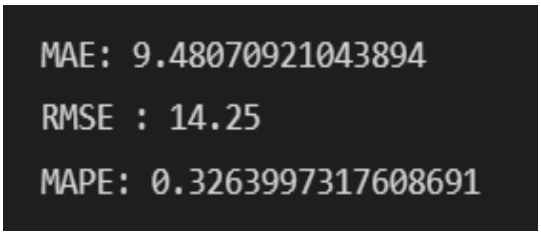

圖 5 KNN 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_neighbors 為 1。另針對訓練集與測試集為 KNN 做準確率預測。

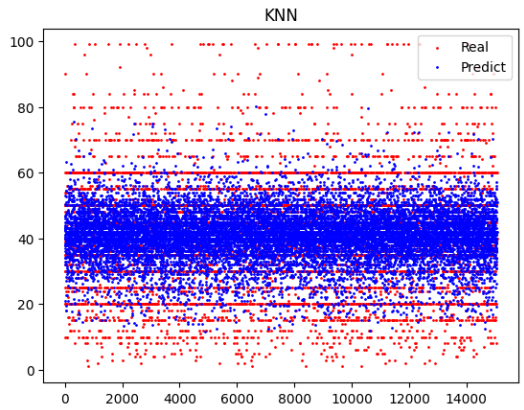

圖 6 KNN 分散圖全預測結果

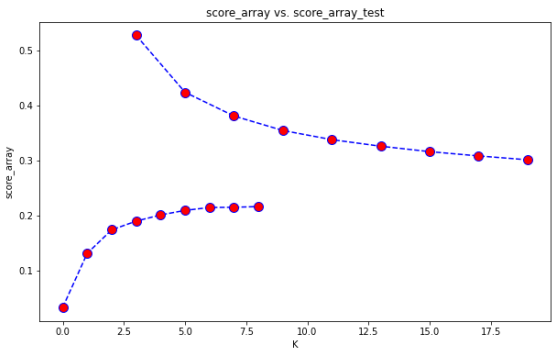

圖 7 KNN 分散圖重要屬性預測結果

接著展示 Adult 資料集中之準確率，利用 SVR 預測準確率(accuracy)，Accuracy:0.116，SVC 之分類器為 linear，C 為 10。另有預測最重要之屬性為 MSRS(Never-married Own-child)、Over50K、Workclass(Self-emp-inc(4))。

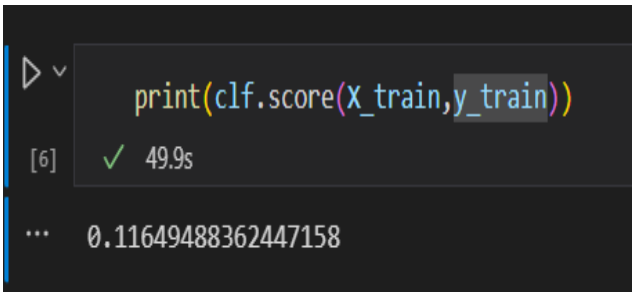

圖 8 SVR 預測結果，預測 C 為 10

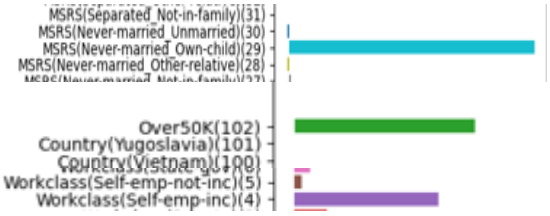

圖 9 SVR 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 C 為 10。另單獨針對重要屬性做測試。

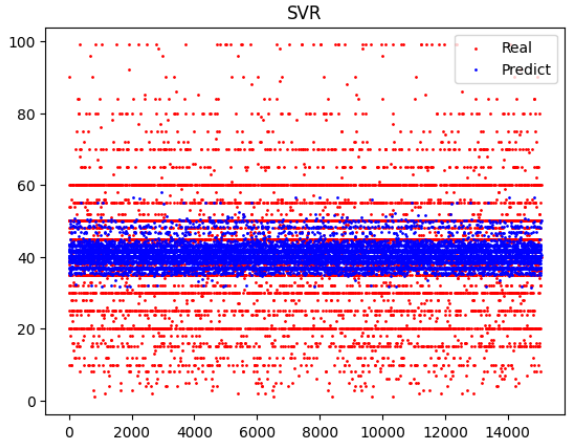

圖 10 SVR 分散圖全預測結果

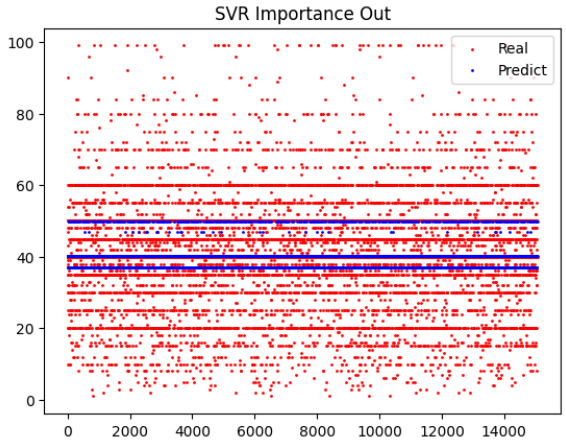

圖 11 SVR 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 SVR 之 MAE、RMSE 與 MAPE 之值。

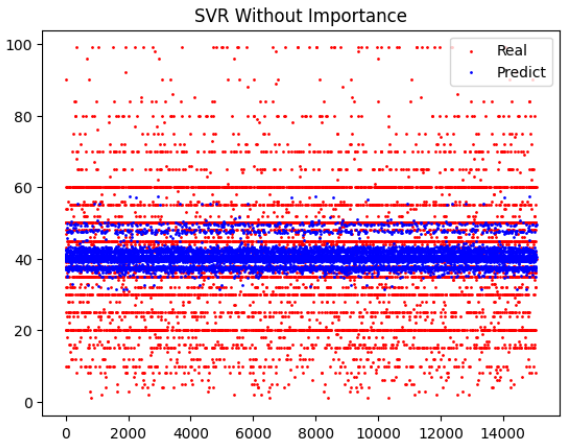

圖 12 SVR 分散圖無重要屬性預測結果

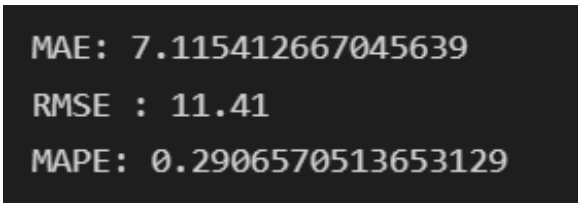

圖 13 SVR 衡量標準預測結果

接著展示 Adult 資料集中之準確率，利用 Random Forest 預測準確率(accuracy)，Accuracy:0.88，Random Forest 之 n_estimators 為 100。另有預測最重要之屬性為 NE、EVR、Month_CPI_CCI_20(54)。

```
print(randomForestModel.score(X_train,y_train))
print(randomForestModel.score(X_test, y_test))

0.883271589103296
0.19650958796288753
```

圖 14 Random Forest 預測結果

n_estimators 為 50

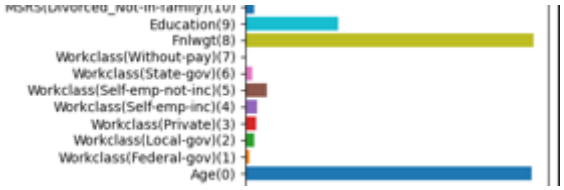

圖 15 Random Forest 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_estimators 為 50。另單獨針對重要屬性做測試。

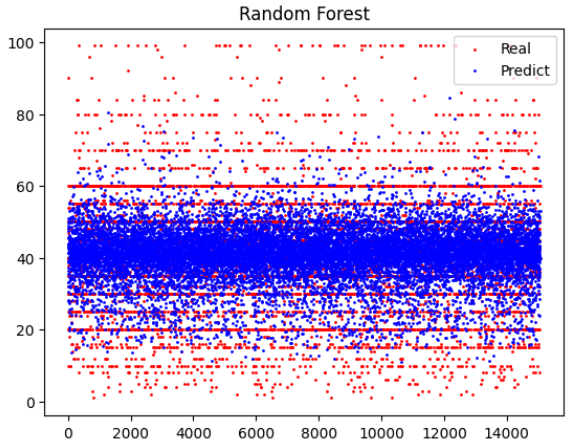

圖 16 Random Forest 分散圖全預測結果

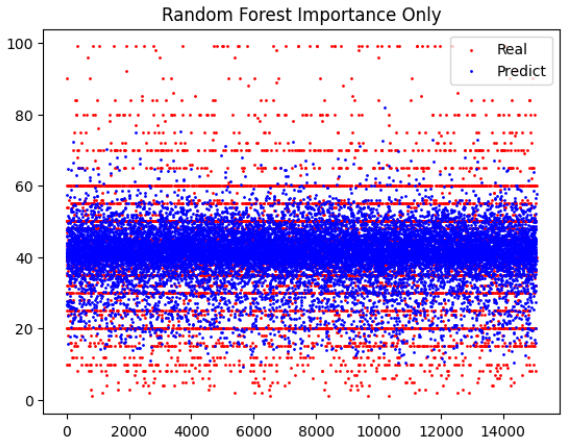

圖 17 Random Forest 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 Random Forest 之 MAE、RMSE 與 MAPE 之值。

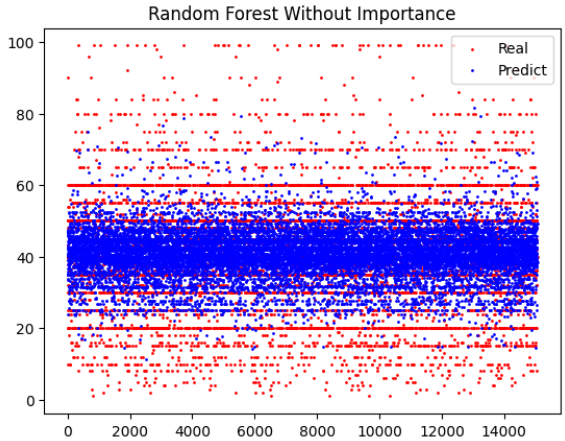

圖 18 Random Forest 分散圖無重要屬性預測結果

```
MAE: 7.451135728041134
RMSE : 10.81
MAPE: 0.27014146571064573
```

圖 19 Random Forest 衡量標準預測結果

接著展示 Adult 資料集中之準確率，利用 XG Boost 預測準確率(accuracy)，Accuracy:0.999，Random Forest 之 n_estimators 為 100。另有預測最重要之屬性為 NE、EVR、Month_CPI_CCI_20(54)。

```
print(xgbc.score(X_train,y_train))
print(xgbc.score(X_test, y_test))
```

```
0.542000758908503
0.22988515442344992
```

圖 20 XG Boost 預測結果 n_estimators 為 200

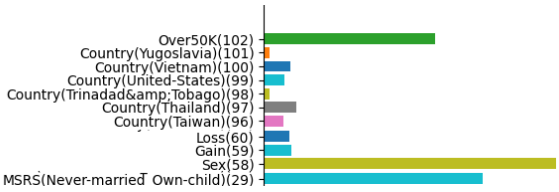

圖 21 XG Boost 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_estimators 為 200。另單獨針對重要屬性做測試。

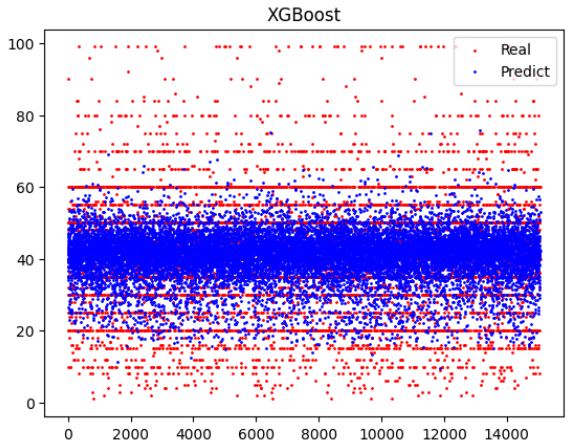

圖 22 XG Boost 分散圖全預測結果

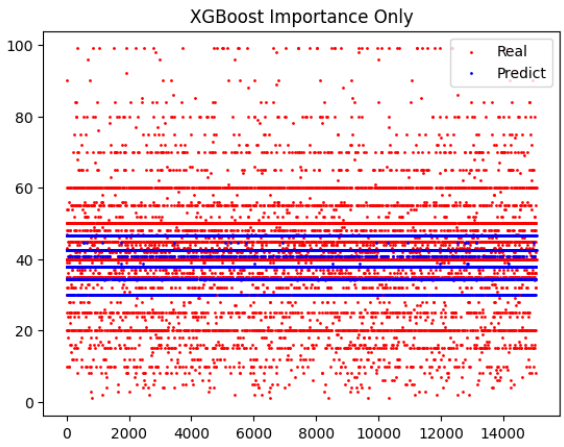

圖 23 XG Boost 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 XG Boost 之 MAE、RMSE 與 MAPE 之值。

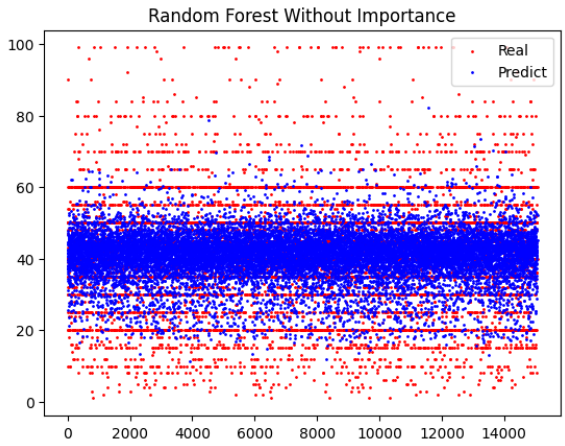

圖 24 XG Boost 分散圖無重要屬性預測結果

```
MAE: 7.2498206861133605
RMSE : 10.59
MAPE: 0.26512252886810234
```

圖 25 XG Boost 衡量標準預測結果

3.4. bank-additonal

接著展示 bank-additonal 資料集中之準確率，利用 SVR 預測準確率(accuracy)，Accuracy:0.999，SVC 之分類器為 linear，C 為 2。另有預測最重要之屬性為 NE、EVR、Month_CPI_CCI_18(52)。

```
print(clf.score(X_train,y_train))
print(clf.score(X_test, y_test))
```

✓ 0.4s

```
0.9993100533534547
0.9992884069380691
```

圖 26 SVR 預測結果，預測 C 為 2

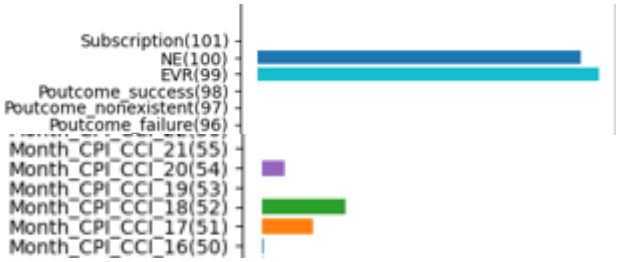

圖 27 SVR 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 C 為 2。另單獨針對重要屬性做測試。

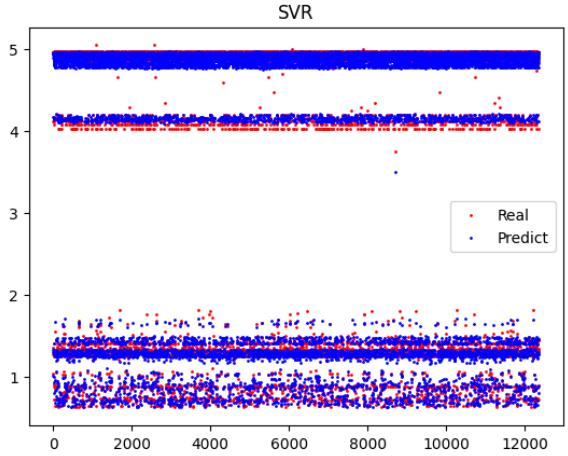

圖 28 SVR 分散圖全預測結果

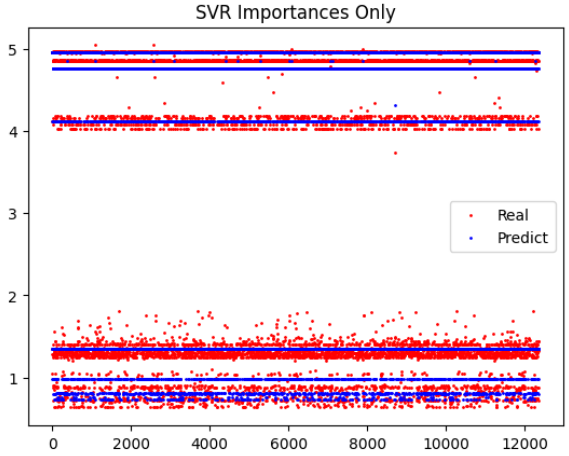

圖 29 SVR 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 SVR 之 MAE、RMSE 與 MAPE 之值。

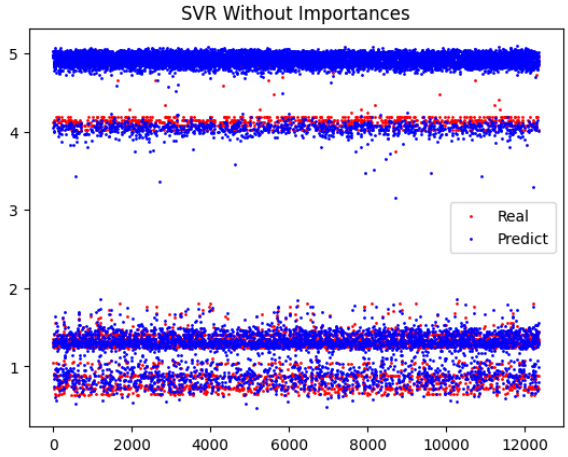

圖 30 SVR 分散圖無重要屬性預測結果

```
MAE: 0.021378888372260187
RMSE : 0.03994
MAPE: 0.010391008461660791
```

圖 31 SVR 衡量標準預測結果

接著展示 bank-additonal 資料集中之準確率，利用 Random Forest 預測準確率(accuracy)，Accuracy:0.999，Random Forest 之 n_estimators 為 100。另有預測最重要之屬性為 NE、EVR、Month_CPI_CCI_20(54)。

```
print(randomForestModel.score(X_train,y_train))
print(randomForestModel.score(X_test, y_test))
✓ 0.4s

0.9999996405870385
0.9999925877708667
```

圖 32 Random Forest 預測結果
n_estimators 為 100

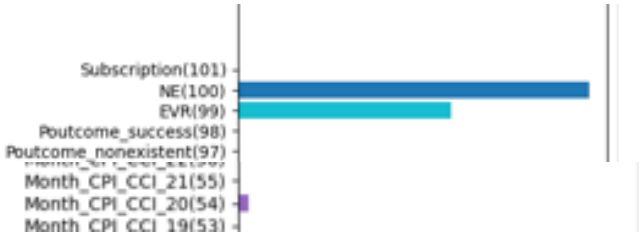

圖 33 Random Forest 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_estimators 為 100。另單獨針對重要屬性做測試。

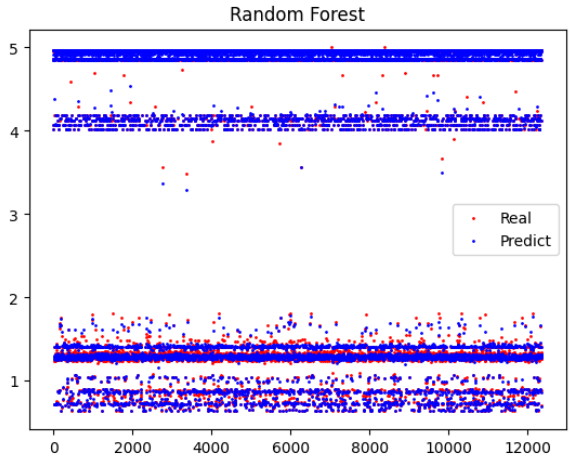

圖 34 Random Forest 分散圖全預測結果

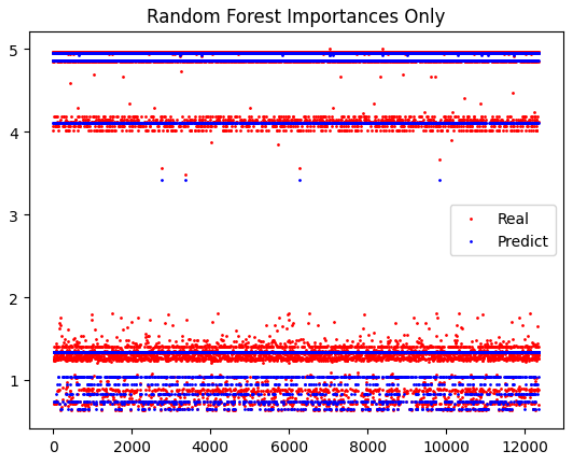

圖 35 Random Forest 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 Random Forest 之 MAE、RMSE 與 MAPE 之值。

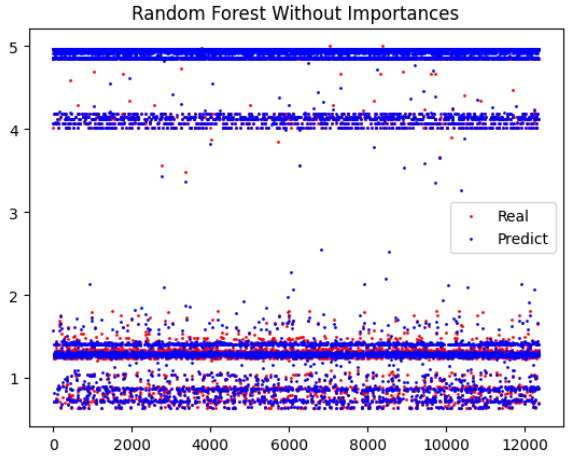

圖 36 Random Forest 分散圖無重要屬性預測結果

```
MAE: 9.567370722565437e-05
RMSE : 0.004714
MAPE: 3.519675912408643e-05
```

圖 37 SVR 衡量標準預測結果

接著展示 bank-additonal 資料集中之準確率，利用 XG Boost 預測準確率(accuracy)，Accuracy:0.999，Random Forest 之 n_estimators 為 100。另有預測最重要之屬性為 NE、EVR、Month_CPI_CCI_20(54)。

```
print(xgbc.score(X_train,y_train))
print(xgbc.score(X_test, y_test))
✓ 0.1s
0.999999962000224
0.9999984486489452
```

圖 38 XG Boost 預測結果 n_estimators 為 200

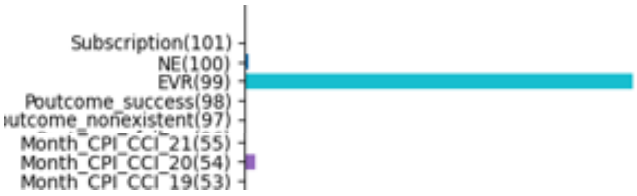

圖 39 XG Boost 之最重要屬性

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_estimators 為 200。另單獨針對重要屬性做測試。

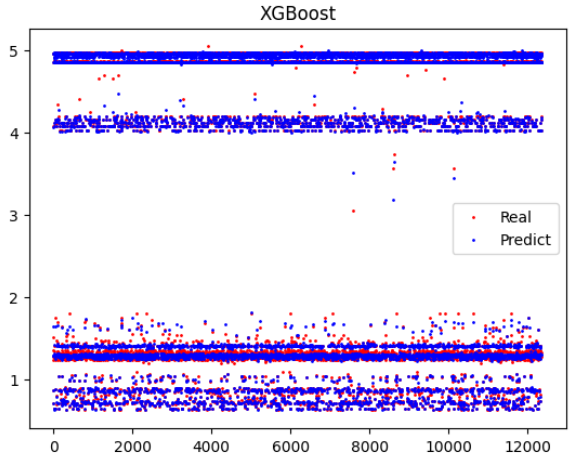

圖 40 XG Boost 分散圖全預測結果

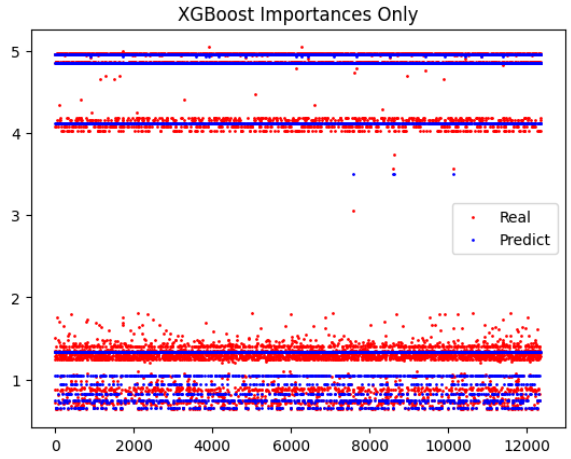

圖 41 XG Boost 分散圖重要屬性預測結果

接著將資料集重要屬性刪除後得到真實值與預測值之間關係。另預測出 XG Boost 之 MAE、RMSE 與 MAPE 之值。

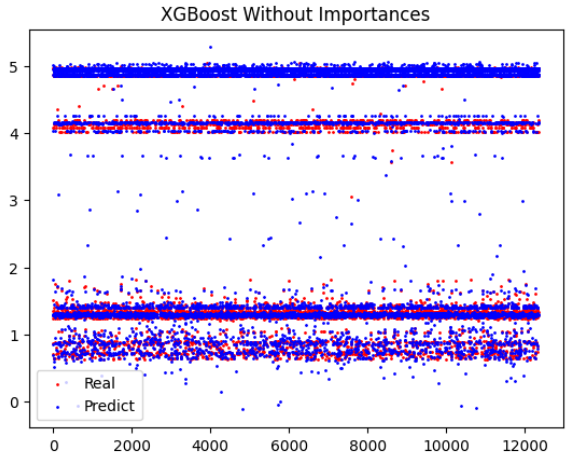

圖 42 XG Boost 分散圖無重要屬性預測結果

```
MAE: 0.00016321283260034306
RMSE : 0.002156
MAPE: 0.00010237528198354835
```

圖 43 XG Boost 衡量標準預測結果

接著展示 bank-additonal 資料集中之準確率，利用 KNN 預測準確率(accuracy)，Accuracy:0.999，KNN 之 n_neighbors 為 3。另預測出 KNN 之 MAE、RMSE 與 MAPE 之值。

```
print(knn.score(X_train,y_train))
print(knn.score(X_test, y_test))
✓ 14.6s

0.9993272369742893
0.998509857171078
```

圖 44 KNN 預測結果 n_neighbors 為 3

```
MAE: 0.0220516576299533317
RMSE : 0.06683
MAPE: 0.0192064597040143777
```

圖 45 KNN 之衡量標準預測結果

本研究利用 One Hot Encoding 這個套件，算出真實值與預測值之間關係，預測 n_neighbors 為 3。另針對訓練集與測試集為 KNN 做準確率預測。

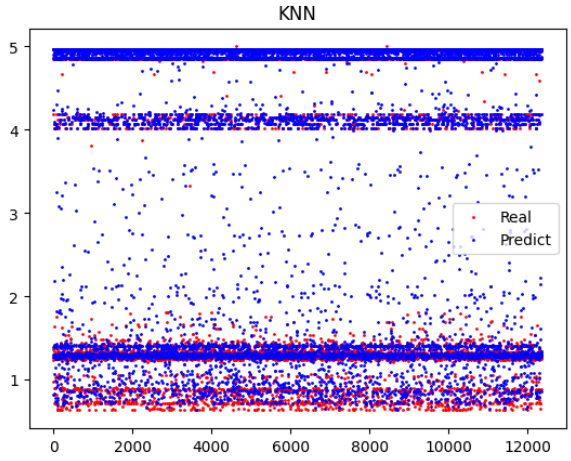

圖 46 Random Forest 分散圖全預測結果

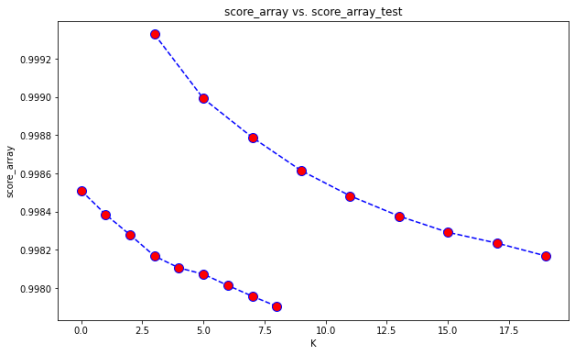

圖 47 Random Forest 分散圖重要屬性預測結果

三、 結論

本研究使用 SQL Server 進行前置處理，後利用 python 進行各演算法(SVR、KNN、Random Forest、XG Boost)分析，計算出兩資料集不同重要特徵屬性及調參數，結果得知 bank 資料集之準確率較高，而 adult 成人資料集雖準確率較低，原先較低資料集(準確率<0.5)最後將測試資料集透過 KNN 提升至 0.9 以上。

參考文獻

PyInvest (2020,Nov,04)。Python 實作支援向量機 SVM，PyInvest。
https://pyecontech.com/2020/04/11/python_svm/
10 程式中 (2021,Sep,23)。核模型 - 支持向量機 (SVM)，2021 iThome 鐵人賽。
https://ithelp.ithome.com.tw/articles/10270447
Jose Portilla (2019,Apr,10)。Python 學習筆記#14：機器學習之 KNN 實作篇，Liz's Blog。
。https://medium.com/@search.psop/python%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-14-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E4%B9%8Bknn%E5%AF%A6%E4%BD%9C%E7%AF%87-64071fbd0ac8
PyInvest (2020,Mar,05)。Python 實作 K-近鄰演算法 KNN，PyInvest。
https://pyecontech.com/2020/05/03/python_knn/
Poul_henry (2019,Mar,16)。python_matplotlib 分別使用 plot()和 scatter()画散点图，以及如何改变点的大小，CSDN。https://blog.csdn.net/Poul_henry/article/details/88602806
scikit-learn (2022)。sklearn.ensemble.RandomForestRegressor，scikit-learn。
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
Maximus (2022,Jul,17)。Feature Importance with SVR，Stackoverflow。
https://stackoverflow.com/questions/70467781/feature-importance-with-svr
Machine Learning Repository。bank，Machine Learning Repository。
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
Ronny Kohavi & Barry Becker (1996,May,01)。adult，Machine Learning Repository。
https://archive.ics.uci.edu/ml/datasets/Adult