國立雲林科技大學資訊管理系

資料探勘

作業一

使用 gini、entropy 預測比較 Adult 與 Coupon 資料集

M11023044 沈俊良

M11023006 謝媄衣

M11023061 王成鋼

D10630002 陳冠臻

指導教授：許中川

2022 年 10 月 23 日

摘要

本研究採用 Coupon 資料集，使用 MS SQL Server 進行資料清洗、轉換與資料及訓練集切割(7:3)，透過 Python 計算各特徵屬性 Gini Split 值、Entropy Split 值、迴圈找出訓練集之最佳特徵屬性與決策樹(Decision Tree)深度，進行決策樹(Decision Tree)訓練，最後測試資料集準確率(accuracy)為 0.69;由於 Coupon 測試集準確率(accuracy)較低，本研究再透過 Adult 資料集進行驗證，最終獲得 0.85 之準確率(accuracy)。

關鍵字: Decision Tree、accuracy、MS SQL Server、Gini Split、Entropy Split

一、 緒論

1.1. 動機

當美中較勁成為新常態，也促使全球局勢更加不穩定，而動盪所引起的波瀾打亂既有的經貿規則與供應鏈，使的「通貨膨脹」儼然成為全球最重要的金融課題。身處在高物價時代，當個人所得沒變動，而物價卻節節高升時，人們也越來越會精打細算，此時廠商必須利用各種促銷方法與活動，來刺激顧客的購買慾望與滿足顧客消費需求；消費券是用來刺激消費的行銷手法之一，但廣泛的運用消費券來刺激消費的同時，如何能夠找到正確的消費族群以及積極開發潛在消費者，勢必成為各家廠商當務之急的任務。本研究希望可以改善廣泛發放消費券的行銷方式，以更聚焦的行銷手法進行，因此用資料探勘的方式對資料進行探索性研究，從資料中挖掘更多消費行為的潛在脈絡，探究資料不同面向的可能性。我們挑選駕駛在不同情境下使用優惠券行為的數據集進行資料的分析，希望能透過數據的分析能夠達到精準行銷，減少無謂的人力、物力、時間與空間成本的浪費，並可針對潛在消費者未達成消費之因素加以改善，期望能夠調整銷售策略進而達成交易與提升交易的次數，為本研究的首要目標。

1.2. 目的

對於優惠券相關研究主要大多是探討使用優惠卷的消費者之行為模式、消費者購買意願的態度、優惠券如何影響購買意圖等面向，而對於過去的研究重點主要都放在消費者的行為和對於消費券的認知與態度上，但今日身處於分眾時代的我們，除了解消費者與行為模式之外，我們更希望可以將消費者族群進行分眾，並且在其中尋找潛在的消費客群以及討潛在影響消費者在駕駛時，因使用情境中的潛在因素而阻擾消費的行為，若當我們可以掌握到相關訊息，在未來我們可以擬訂更明確的優惠券的行銷策略與行銷對象，排除潛在客群受阻礙的原因，促使潛在消費族群達到消費的目的，進而提升優惠券的使用率。

1.3. 方法

將原始的 Coupon 跟 Adult 資料集匯入到 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 Gini Split Value & Entropy Split Value，排序各特徵屬性，針對 Gini 及 Entropy 調深度找出最佳準確率，依據迴圈寫入決策樹訓練方法，再依迴圈測試決策樹深度，進而找出最佳決策樹。

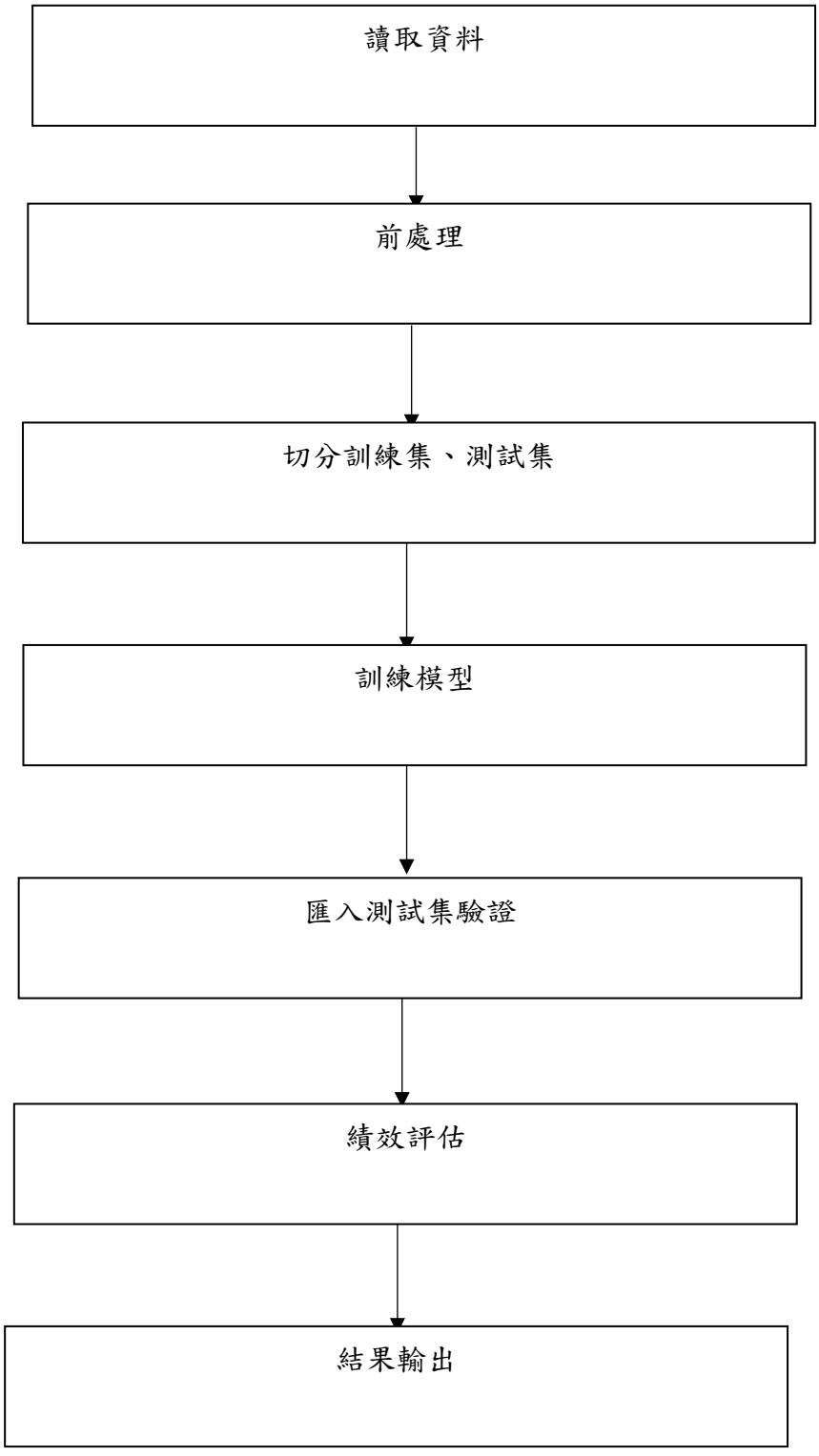

圖 1 coupon 資料集預測流程圖

二、 方法

2.1. 資料集

表 1

Seoul Bike Sharing Demand 資料屬性一覽

欄位名稱	譯名	型態
Destination	目的地	String
Passanger	乘客	String
Weather	天氣	String
Temperature	溫度	String
Time	時間	String
Coupon	優惠卷	String
Expiration	有效期	String
Gender	性別	String
Age	年齡	String
Occupation	狀況	String
Income	收入	String
Car	車種	String
Bar	到酒吧的次數	String
CoffeeHouse	到咖啡館次數	String
CarryAway	外賣次數	String
RestaurantLessThan20	花費在餐廳(低於 20)	String
Restaurant20To50	花費在餐廳(20~50)	String
Coupon_GEQ5min	行駛距離(5 分鐘)	String
Coupon_GEQ15min	行駛距離(15 分鐘)	String
Coupon_GEQ25min	行駛距離(25 分鐘)	String
DirectionSame	方向是否相同	String
DirectionOpp	方向是否相反	String
UseCoupon	是否使用優惠卷	String
樣本數		屬性個數
12684		23

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	祖國	String
Listing of attributes	屬性	String

樣本數	屬性個數
32652	15

2.2. 前置處理

將 Adult 資料集的原始資料所記錄非連續性的資料，轉換為 Label encoding，但 Label encoding 無法直接對字串進行編碼，必須先透過 Label encoding 將字串以數字取代後再進行 Label encoding 處理，接著刪除缺失資料與轉換格式，最後再進行正規化。coupon 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。

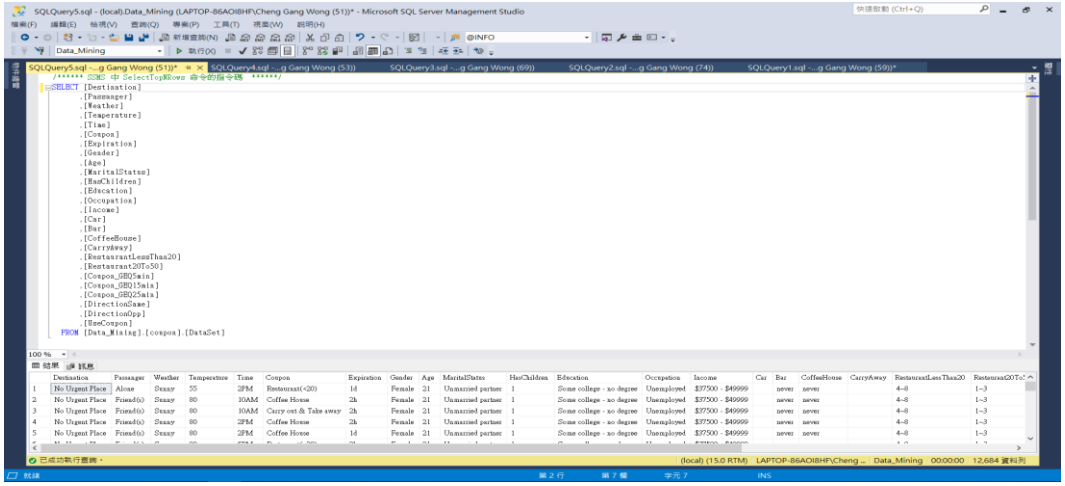

圖 2 SQL server 前置處理 (詳情請見 github)

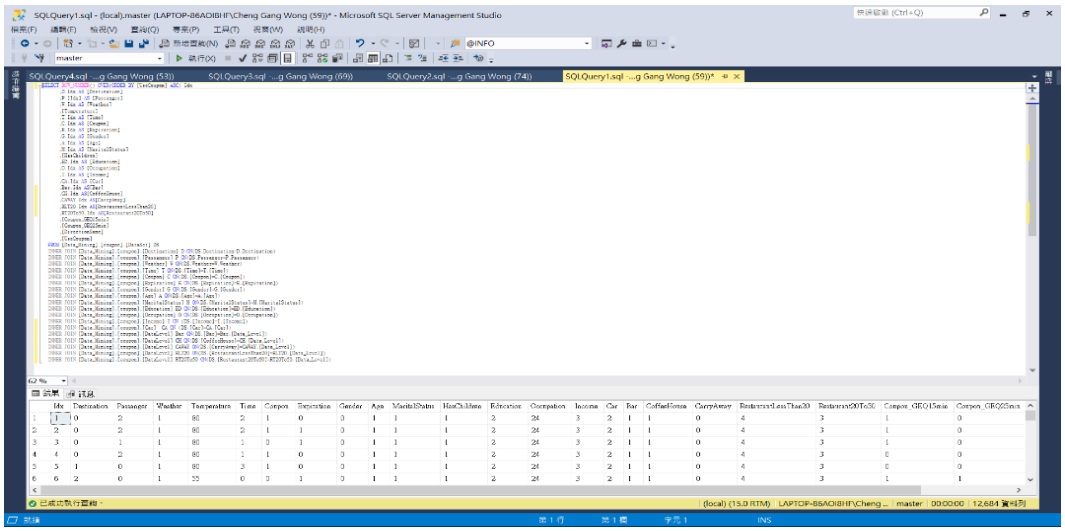

圖 3 SQL server 前置處理 (詳情請見 github)

2.3. 實驗設計

我們是在 Anaconda3_spyder(Pythonversion=3.10)環境，以及 Visual Studio Code 下進行開發，使用的套件 DecisionTreeClassifier 調整深度的參數。測試深度多少時的準確率為最高，該方法是預測分叉路徑可能的屬性。

Gini 公式:$1-\sum_{i=1}^{n} {p_i}^2$

Gini Split 公式:$\sum_{i=1}^{k} \frac{n_i}{n} GINI(I)$

entropy 公式:$-\sum_J P(j|t) log_2\, p(j|t)$

2.4. 實驗結果

首先我們先展示 Adult 資料集中的準確率，利用 Gini 來預測最佳準確率 (accuracy)，Accuracy: 0.8474388895713058，預測深度為 3

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
[101]  ✓ 0.3s
...  Accuracy: 0.8474388895713058
```

圖 4 Gini Accuracy 預測結果，預測深度為 3

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 3

```
score_array [0.759009009009009, 0.823710073710738, 0.8396396396396396, 0.8397215397215397, 0.849058149058149, 0.8538493038493038, 0.8567977067977068, 0.859541359541596,
0.8615888615888616, 0.8665028665028665, 0.8729320229320229, 0.8792792792792793, 0.8859131859131859, 0.892997542997543, 0.9008190008190008, 0.9092137592137592, 0.9161343161343162,
0.9237100737100737, 0.9303849303849304]

score_array_MAX: 0.9303849303849304
----------------------------------------------------------------------------------------------------
score_array_test [0.7597346763296893, 0.8183888404372927, 0.8470703844736519, 0.8473160545387545, 0.8556688367522417, 0.857142857142857, 0.8583712074683, 0.8602137329566393,
0.8591082176636777, 0.8544404864267289, 0.8571428571428571, 0.8522294558408058, 0.8522294558408058, 0.8481758997666135, 0.8451050239528314, 0.8421569831716006, 0.837612086967203,
0.8304876550792286, 0.8313475003070876]

score_array_test_MAX: 0.8602137329566393

[<matplotlib.lines.Line2D at 0x28108433e50>]
```

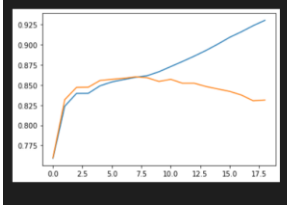

圖 5 Gini 折線預測結果，預測深度為 3

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 3

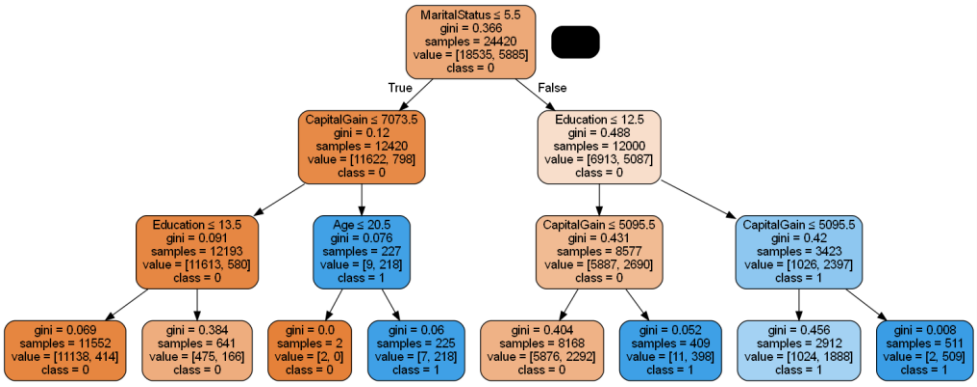

圖 6 決策樹分支圖，預測深度為 3

本研究利用 Gini 來預測準確率(accuracy)，Accuracy: 0.85124677558，預測深度為 6

```
Accuracy: 0.8512467755803955

array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

圖 7 Gini Accuracy 預測結果，預測深度為 6

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 6

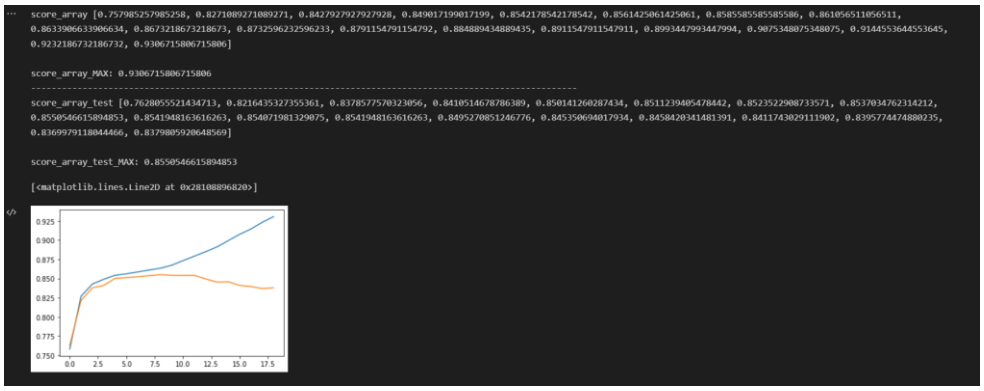

圖 8 Gini 折線預測結果，預測深度為 6

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 6

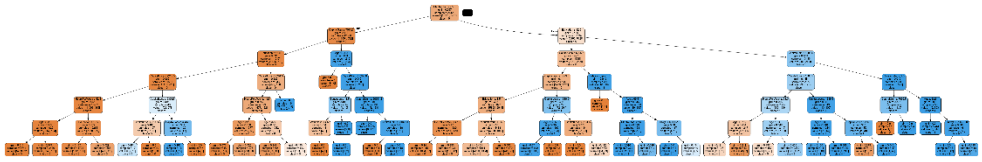

圖 9 決策樹分支圖，預測深度為 6

本研究利用 entropy 來預測準確率(accuracy)，Accuracy: 0.8512467755，預測深度為 6

```
Accuracy: 0.8512467755803955

array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

圖 10 Entropy Accuracy 預測結果，預測深度為 6

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 6

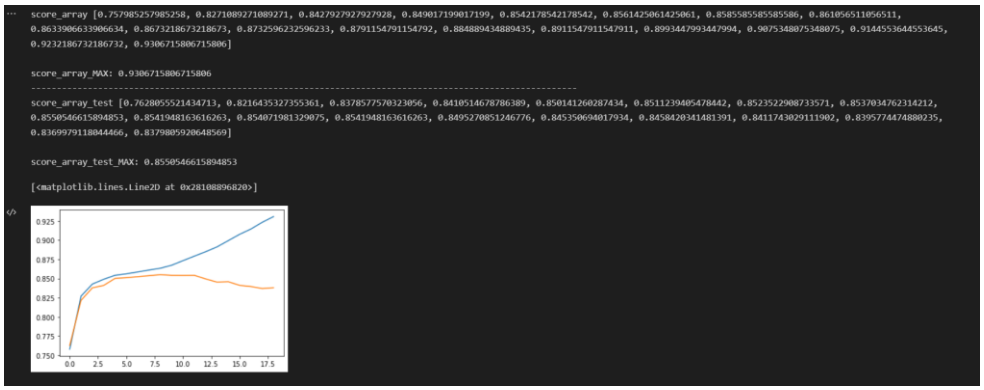

圖 11 Entropy 折線預測結果，預測深度為 6

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 6

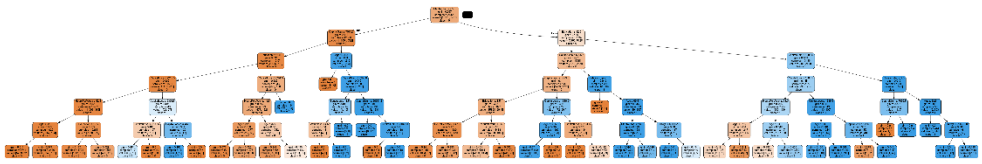

圖 12 決策樹分支圖，預測深度為 6

2.5. 測試集

本研究利用 entropy 來預測準確率(accuracy)，Accuracy: 0.852861704，預測深度為 5

```
Accuracy: 0.8528617047408499

array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

圖 13 Entropy Accuracy 預測結果，預測深度為 3

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 5

```
score_array [0.7621621621621621, 0.8254709254709255, 0.8384111384111385, 0.838574938574938, 0.8434889434889435, 0.852006552006552, 0.858067158067158,
0.8608517608517608, 0.8635544635544635, 0.8703521703521704, 0.8776412776412776, 0.8842751842751843, 0.8914004914004914, 0.8991809991809991,
0.9076167076167077, 0.9170352170352171, 0.9226863226863227, 0.9311220311220311, 0.937100737100737]

score_array_MAX: 0.937100737100737
--------------------------------------------------------------------------------------------------------
score_array_test [0.7686072218128224, 0.8393515106853353, 0.8528617047408499, 0.8531073446327684, 0.8543355440923606, 0.8582657823630557,
0.8597396217145664, 0.8594939881822648, 0.8548268238761975, 0.8531073446327684, 0.855072463768116, 0.8472119872267256, 0.847457627118644,
0.8418079096045198, 0.8422991893883567, 0.8327192336035372, 0.8332105133873741, 0.8305084745762712, 0.8371407516580692]

score_array_test_MAX: 0.8597396217145664

[<matplotlib.lines.Line2D at 0x269775372e0>]
```

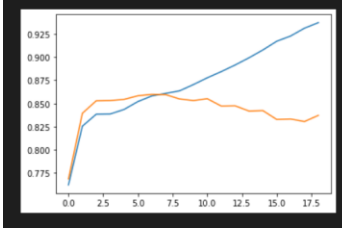

圖 14 Entropy 折線預測結果，預測深度為 5

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 5

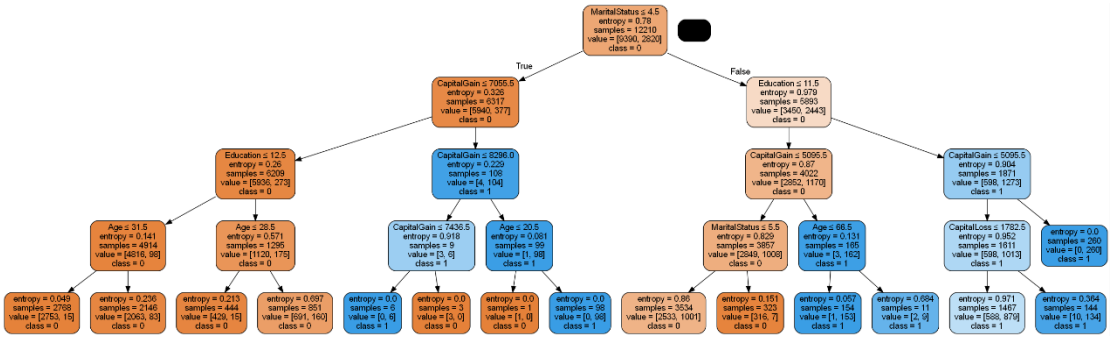

圖 15 決策樹分支圖，預測深度為 5

本研究利用 gini 來預測準確率(accuracy)，Accuracy: 0.8577，預測深度為 6

```
Accuracy: 0.8577745025792188

array([1, 0, 1, ..., 0, 0, 0], dtype=int64)
```

圖 16 Entropy Accuracy 預測結果，預測深度為 6

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 6

```
score_array [0.761998361998362, 0.8277641277641278, 0.8407862407862408, 0.8458640458640458, 0.8526617526617527, 0.8552006552006552,
0.8594594594594595, 0.8636363636363636, 0.8678951678951679, 0.872072072072072, 0.8783783783783784, 0.8850941850941851, 0.8945126945126946,
0.9025389025389026, 0.9117117117117117, 0.9200655200655201, 0.926945126945127, 0.936036036036036, 0.9428337428337429]

score_array_MAX: 0.9428337428337429
----------------------------------------------------------------------------------------------------
score_array_test [0.7690985015966593, 0.8324735937116188, 0.845738147875215, 0.8484401866863178, 0.855072463768116, 0.8575288626873004,
0.8577745025792188, 0.855072463768116, 0.8553181036600344, 0.8558093834438713, 0.8516335052812577, 0.843036109064112, 0.8398427904691722,
0.8462294276590518, 0.8413166298206829, 0.8378776713338246, 0.8361581920903954, 0.8273151559813313, 0.8329648734954557]

score_array_test_MAX: 0.8577745025792188

[<matplotlib.lines.Line2D at 0x26979af41c0>]
```

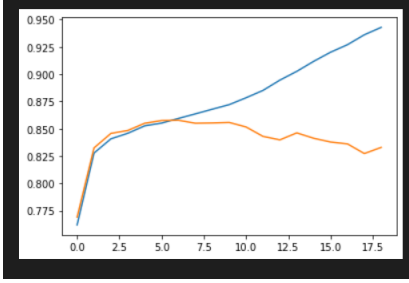

圖 17 Entropy 折線預測結果，預測深度為 6

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 6

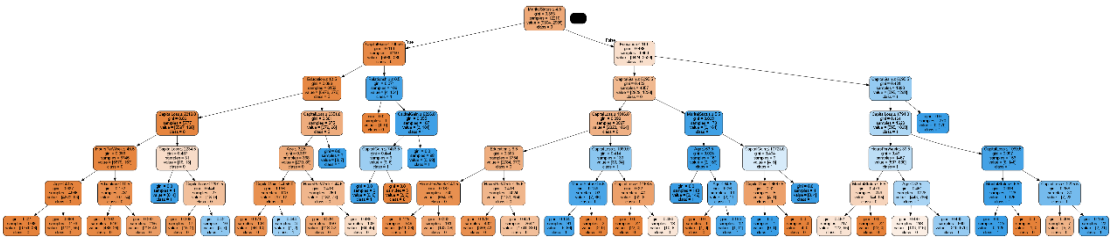

圖 18 決策樹分支圖，預測深度為 6

2.6. Coupon

接著展示 Coupon 資料集中的準確率，利用 Gini 來預測準確率(accuracy)，Accuracy:0.693，預測深度為 8 為最高準確率

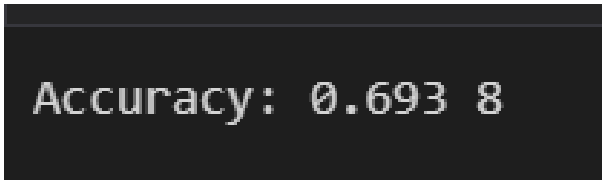

圖 19 Entropy Accuracy 預測結果，預測深度為 8

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 8

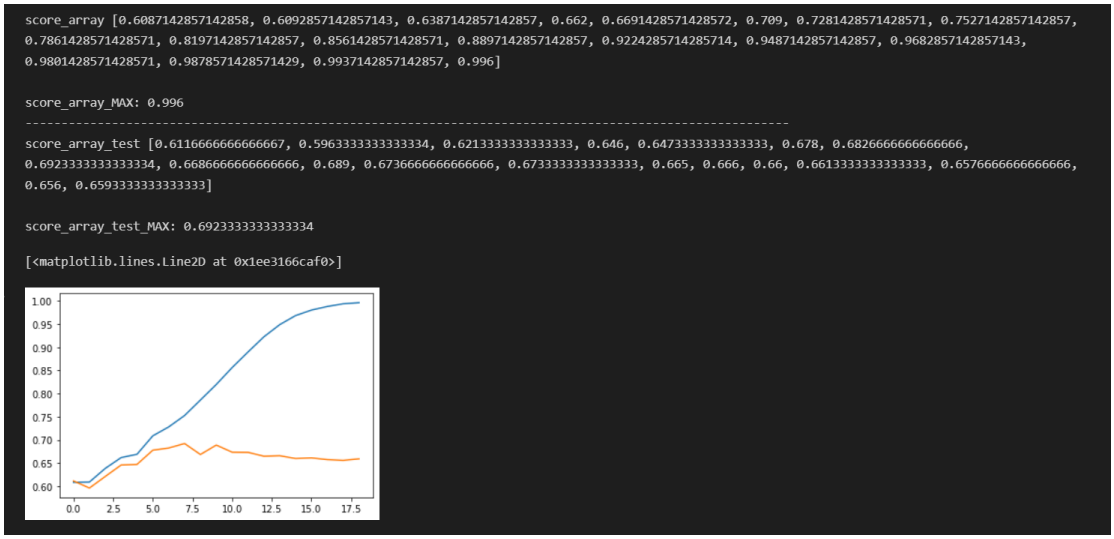

圖 20 Entropy 折線預測結果，預測深度為 8

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 8 (詳情請見 github)

圖 21 決策樹分支圖，預測深度為 8

接著展示 Coupon 資料集中的準確率，利用 entropy 來預測準確率(accuracy)，Accuracy:0.691，預測深度為 7 為最高準確率

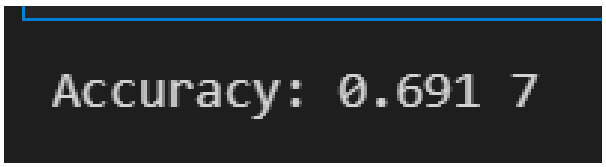

圖 22 Entropy Accuracy 預測結果，預測深度為 7

本研究利用 LabelEncoder 這個套件，算出訓練集與測試集是否會 overloading，預測深度為 7

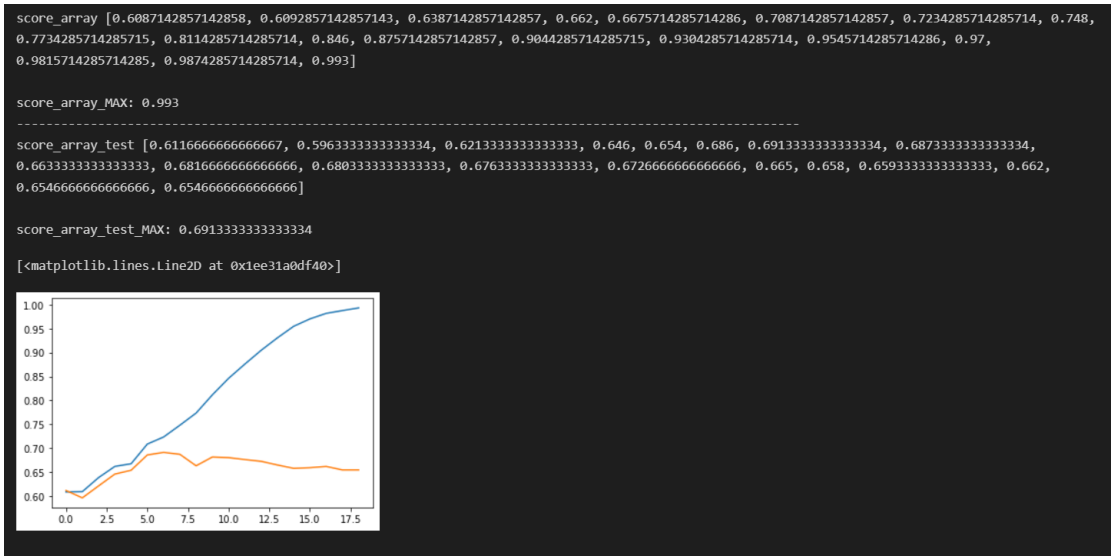

圖 23 Entropy 折線預測結果，預測深度為 7

本研究利用 Pydotplus 這個套件，利用訓練集跑出決策樹分類圖。預測深度為 7 (詳情請見 github)

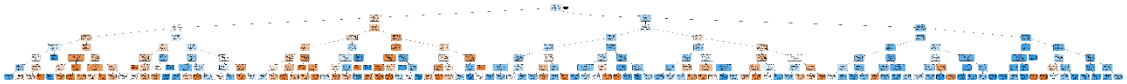

圖 24 決策樹分支圖，預測深度為 7

三、 結論

本研究使用 SQL server 進行前置處理，後利用 python 進行 Gini 及 Entropy 做分析，計算出兩資料集不同決策樹特徵屬性及深度，結果得知 adult 成人資料集之準確率較高，而 coupon 資料集雖準確率較低，原先較低資料集(準確率 <0.6)最後將測試資料集提升至 0.65 以上。

參考文獻

Avinash Navlani(Dec 2018)。Decision Tree Classification in Python Tutorial，datacamp。https://www.datacamp.com/tutorial/decision-tree-classification-python

Ronny Kohavi & Barry Becker (1996,May,01)。adult，Machine Learning Repository。https://archive.ics.uci.edu/ml/datasets/Adult

Tong Wang & Cynthia Rudin (2020,Sep,15)。coupon，Machine Learning Repository。

T. Hastie&R. Tibshirani& J. Friedman&J.R. Quinlan(1993)。Decision Trees，scikit-learn。https://scikit-learn.org/stable/modules/tree.html