中華郵政大數據競賽 |

分析方案說明書

組名 | 消失的章魚

代碼 | 140006

成員 | 蕭妤安 施安隆 張耀仁 俞又瑄

指導教授 | 國立臺灣大學 朱致遠教授

目錄

01

郵局起訖關聯性分析

郵局起訖關聯性分析

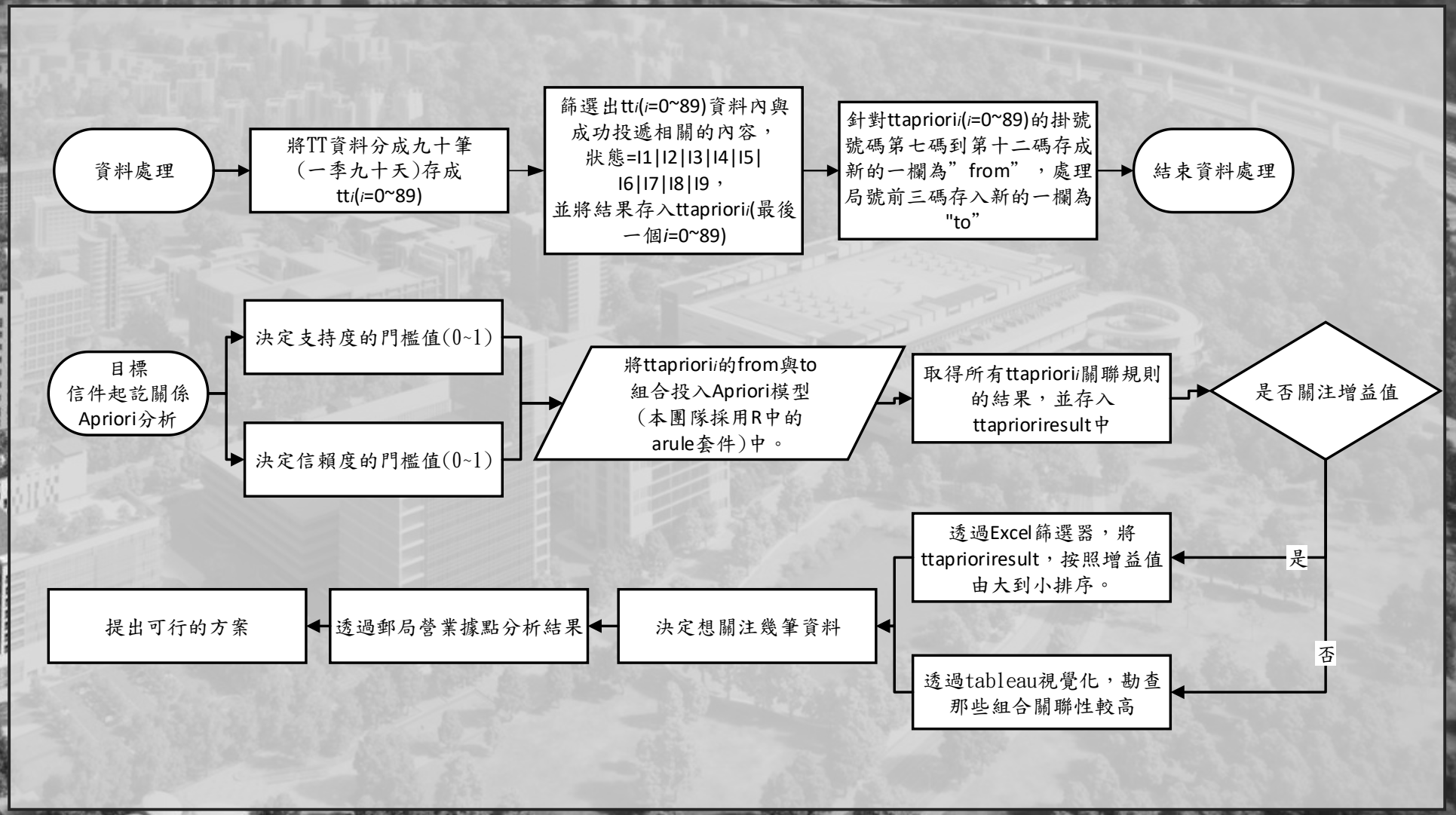
數據規劃
資料流程圖

資料處理

使用資料：TT
取用欄位：
成功投遞相關內容

數據分析

使用方法：
- Apriori 分析



Proposal 1

郵局起訖關聯性分析

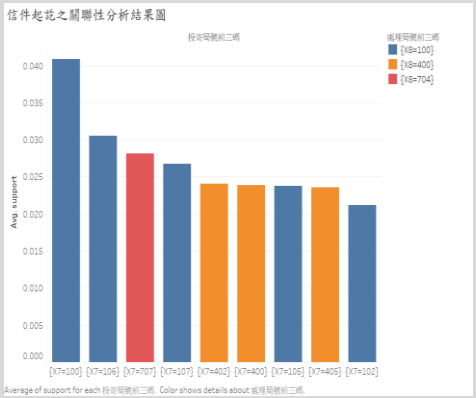
Proposal 1

問題 1

想要了解所有信件起迄點的關聯性

將TT數據中，有成功狀態"I"取出，並將掛號號碼的第七到第九碼(投寄局號的前三碼，LHS)，與處理局號前三碼(RHS)，以天為單位(將資料切成90筆)，做關聯性分析。

投寄局號	100 台北	707 台南	100 台北	400 台中	400 台中	100 台北	400 台中	100 台北
	↓	↓	↓	↓	↓	↓	↓	↓
處理局號	106 台北	704 台南	107 台北	402 台中	400 台中	105 台北	405 台中	102 台北



LHS		RHS	Support	Confidence	Lift	Count
{X7=903}	→	{X8=900}	0.004235	0.944646	38.88139	4270
{X7=224}	→	{X8=220}	0.007243	0.828722	8.99864	7192
{X7=972}	→	{X8=970}	0.002272	0.767263	75.49142	2100

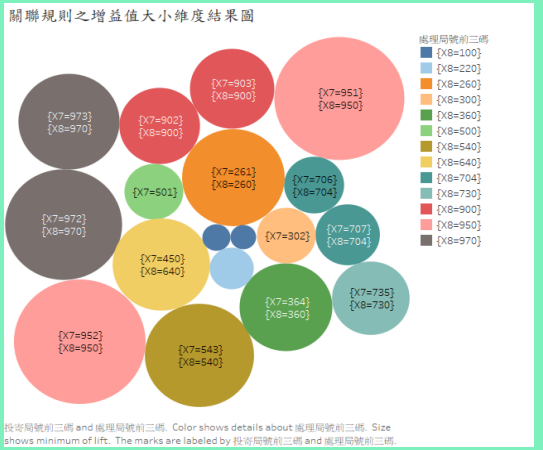
問題 2

想要找出較稀少但是卻又有穩定關係的信件起訖組合

此關聯規則的結果，可以推測LHS與RHS組合數佔總信件組合較少(支持度均低於0.01)，但從LHS出發的信件有很高的機率會送往RHS(信賴度均高於0.7)，其中信賴度最高的為903->900，由結果可以推測，903寄出郵件雖然較少，但只要一有郵件被寄出時，有很高的機率會送往以900為前三碼的處理局號。

信賴度第二高為224(瑞芳區)->220(板橋土城區)，此組合雖然信賴度沒有第一組高，但其組合數卻明顯較高，故不容忽視。

信賴度第三高，增益值最高的組合為972->970，會在問題三進行更深入的講解。

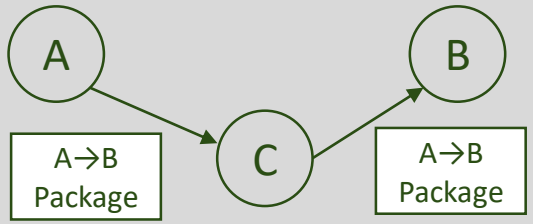


郵局起訖關聯性分析

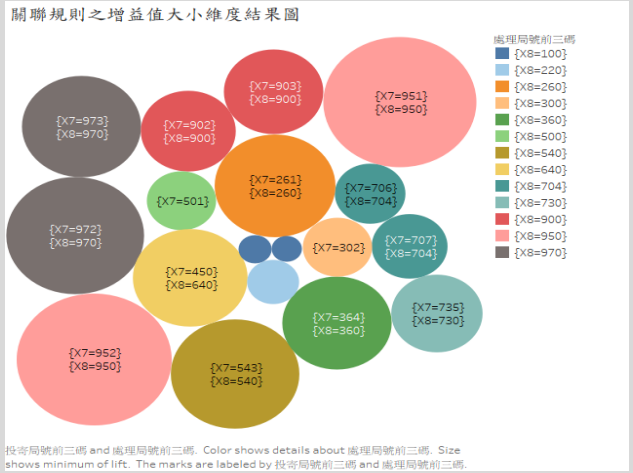
Proposal 1

問題 3

呈問題二，近一步看其增益值與信件數量，判斷是否能有信件轉乘運送的可能性，以降低不必要的物流移動。其概念取自飛機航班的轉乘。



由圖可以推測，增益值前五大分別為，952->950、951->950、972->970、543->540以及261->260，其中處理局號前三碼為950佔兩筆，可推測送往950的信件數並不多，但若是送往950，有很大的機率其信件來自投寄局號前三碼為951或是952的郵局。



關聯規則關係圖		
A	C(LHS)	B(RHS)
其他郵局	952/951	950
	972	970
	543	540
	261	260

建議分別針對送往950、970、540、260(示意圖中的B)四個處理局號前三碼的信件，可以先分別送至其相對應的投寄局號前三碼的郵局(示意圖中的C，LHS)，再與其他信件合送至950、970、540、260(示意圖中的B)。

提案未來方向

短期目標 |

1. 確認分析資料的正確性
2. 確認目前郵局運送信件的物流模式
3. 確認轉乘信件的想法與可行性

中期目標 |

1. 將分析結果與智慧郵局政策做連結，例如：改善i郵箱的運送模式。
2. 延伸關聯性分析模型的應用，例如：郵局與全家合作後，改變運送模式的探討。
3. 提出其他能與智慧郵局政策做連結的方案。

長期目標 |

1. 為不同方案做成本效益評估。
2. 將符合成本效益的方案落實。

02

i郵箱相關收寄資料明細檔之文字探勘

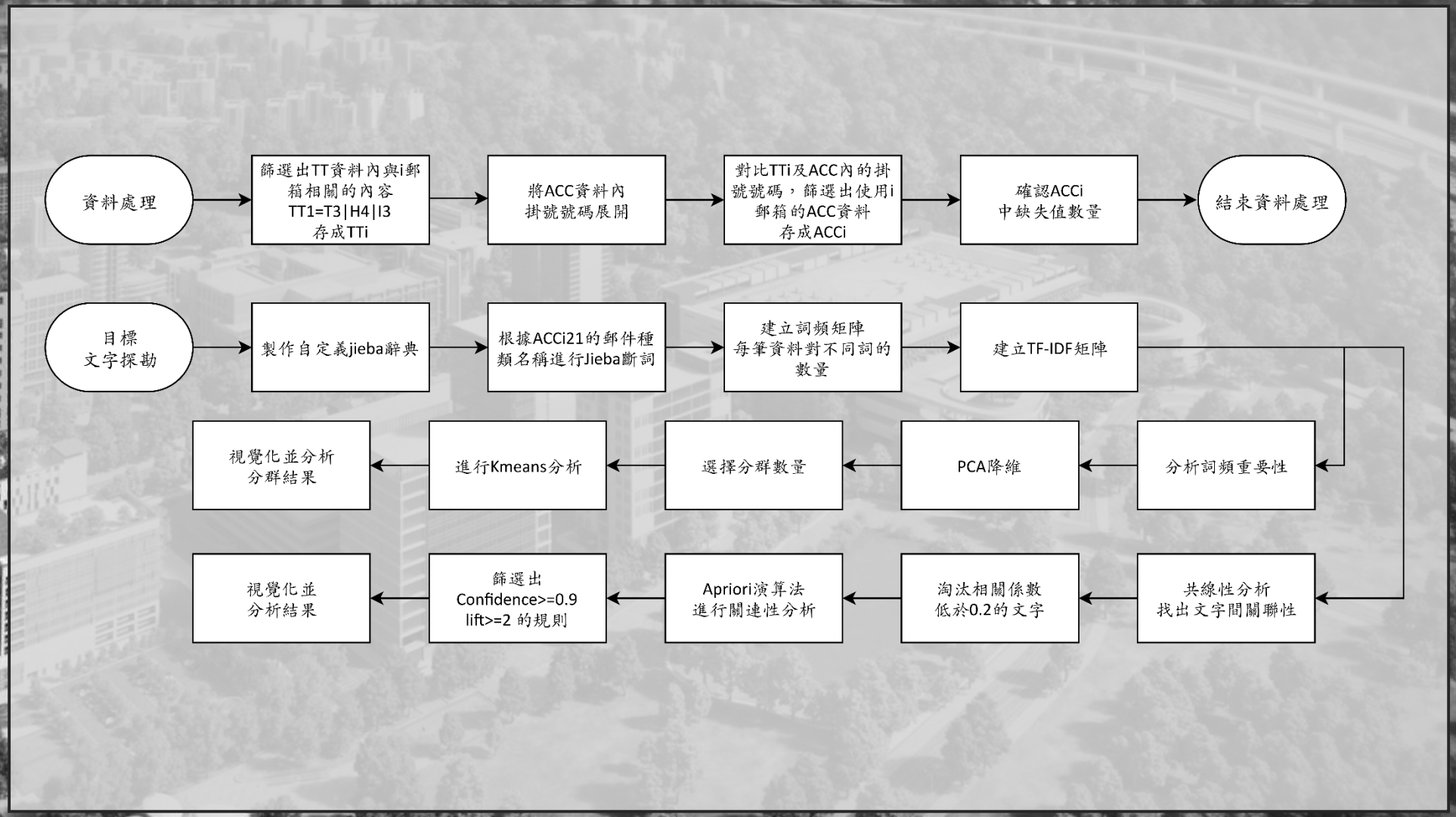
數據規劃
資料流程圖

資料處理

使用資料：TT, ACC
取用欄位：
i郵箱相關之內容

數據分析

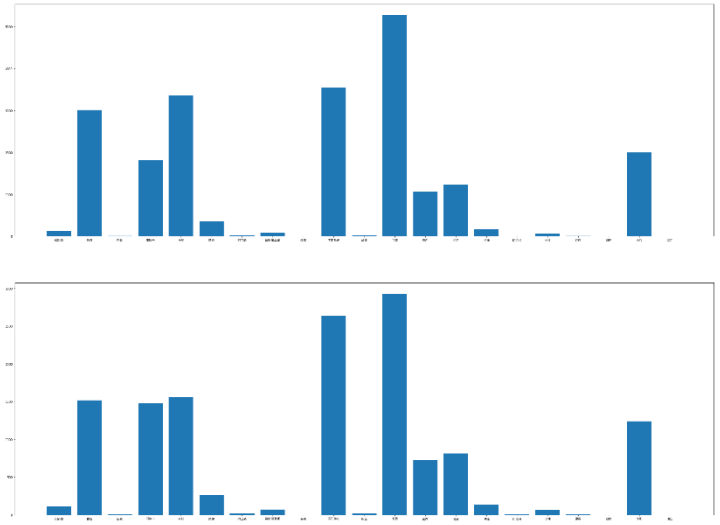
使用方法：
- TF-IDF矩陣
- PCA降維
- Kmeans分析
- 共線性分析
- Apriori 分析



分析方法選用與分析結果

TF-IDF

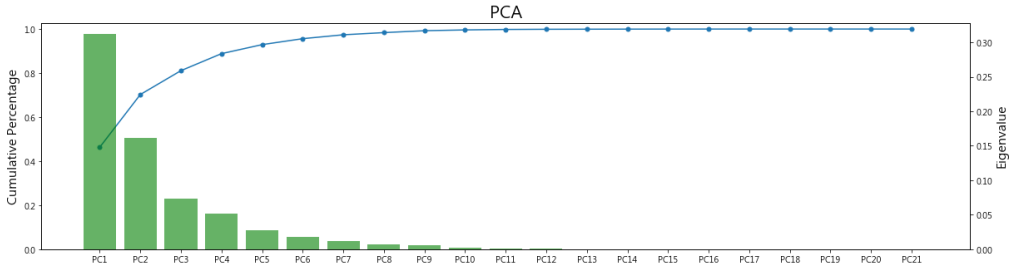
一種用於資訊檢索與文字探勘的常用加權技術，為一種統計方法，用來評估單詞對於文件的集合或詞庫中一份文件的重要程度。因此我們在此採用此方法，評估斷詞出的各詞彙在文件中的重要性，降低過於頻繁出現的詞彙的重要性，例如包裹。



上圖是未經TF-IDF處理過的詞彙特徵計數長條圖，下圖則是經TF-IDF，可以發現在比較所有郵件種類名稱時，雖然詞頻矩陣中小包出現頻率比同縣市高，但在TF-IDF矩陣卻是同縣市的重要性比小包還高，代表同縣市的重要性比小包更大。因此藉由TF-IDF處理，我們可以判斷各詞彙在文件中的重要性。

PCA降維

主成分分析被歸類成為降維時特徵擷取的一種方法，降維就是希望資料的維度數減少，但整體的效能不會差異太多甚至會更好。在這份文件中，我們有21個字彙作為特徵值，因此我們使用PCA降低資料維度，方便我們檢視資料之間的不同之處，進行分析。

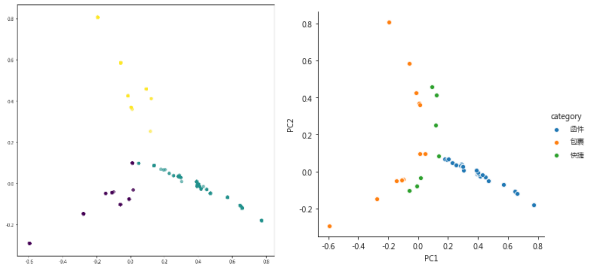


使用PCA降低維度，並且依據此圖可判斷出當選用兩個維度時，資料支持度已達到約0.7，足夠代表此資料，因此最後選擇二維進行後續分析。

分析方法選用與分析結果

K-means分析

因為K-Means適合處理分布集中的大型樣本資料，在此資料中繪出的二維圖形中可判斷他具有一定的密集度，因此我們採用K-Means進行分析，原理較簡單，收斂速度也快。

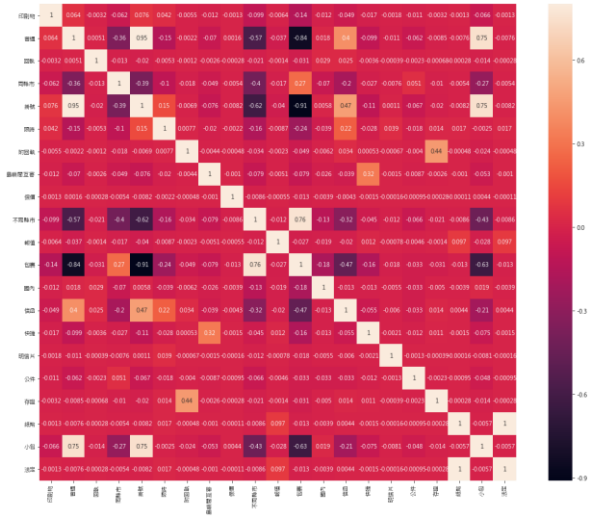


使用K-Means分析，可將資料分成三群，但我們發現此分群與掛號種類相關性不高，可能還需要依據其他資訊才能更好的解釋這三個群體。

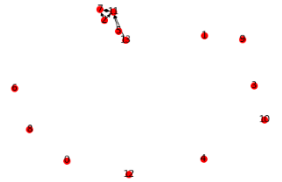
Apriori分析

Apriori是經典的挖掘資料關聯性演算法。並且此演算法簡單、易理解且對資料要求較低，符合此字彙資料的型態，因此選用Apriori進行資料分析。

共線性分析



進行共線性分析，剔除相關性低於0.2的特徵，包含：明信片、回執、公件、印刷物、國內、報值、保價。再用整理過的資料進行Apriori分析，篩選出confidence大於0.9且lift大於2的規則。



從左圖可以發現以下幾個關聯性:

1. 限時→掛號 2. 掛號→普通
3. 普通→掛號 4. 小包→普通
5. 信函→法定 6. 小包→掛號

因此我們可以推論投遞到i郵箱的信件其內容特性，像是如果有一個客人投遞小包到i郵箱，可能會使用普通或是掛號的方式寄件。

03

預設是否使用ｉ郵箱之分類模型及評估

提案動機與目標

根據中華郵政官網介紹，i郵箱為中華郵政公司提供一種新型態的收/寄郵件體驗，其特色在於可以配合收、寄件人用郵時間，因此不需要在營業時間到郵局領取郵件或是在家等待郵差。然而，在ACC資料檔(收寄明細資料)其總共2700萬筆資料中，僅有大約1萬筆有關於i郵箱的資料，顯示中華郵政公司在推動使用i郵箱的政策上還有努力改善的空間。此份作業期望能透過acc資料檔不同欄位的資料獲取關鍵的訊息，了解i郵箱使用者與非i郵箱使用者的背景與其各自考慮的因素。

預設是否使用i郵箱之分類模型及評估

Proposal 3

數據規劃
資料流程圖

資料處理

使用資料：TT, ACC

取用欄位：

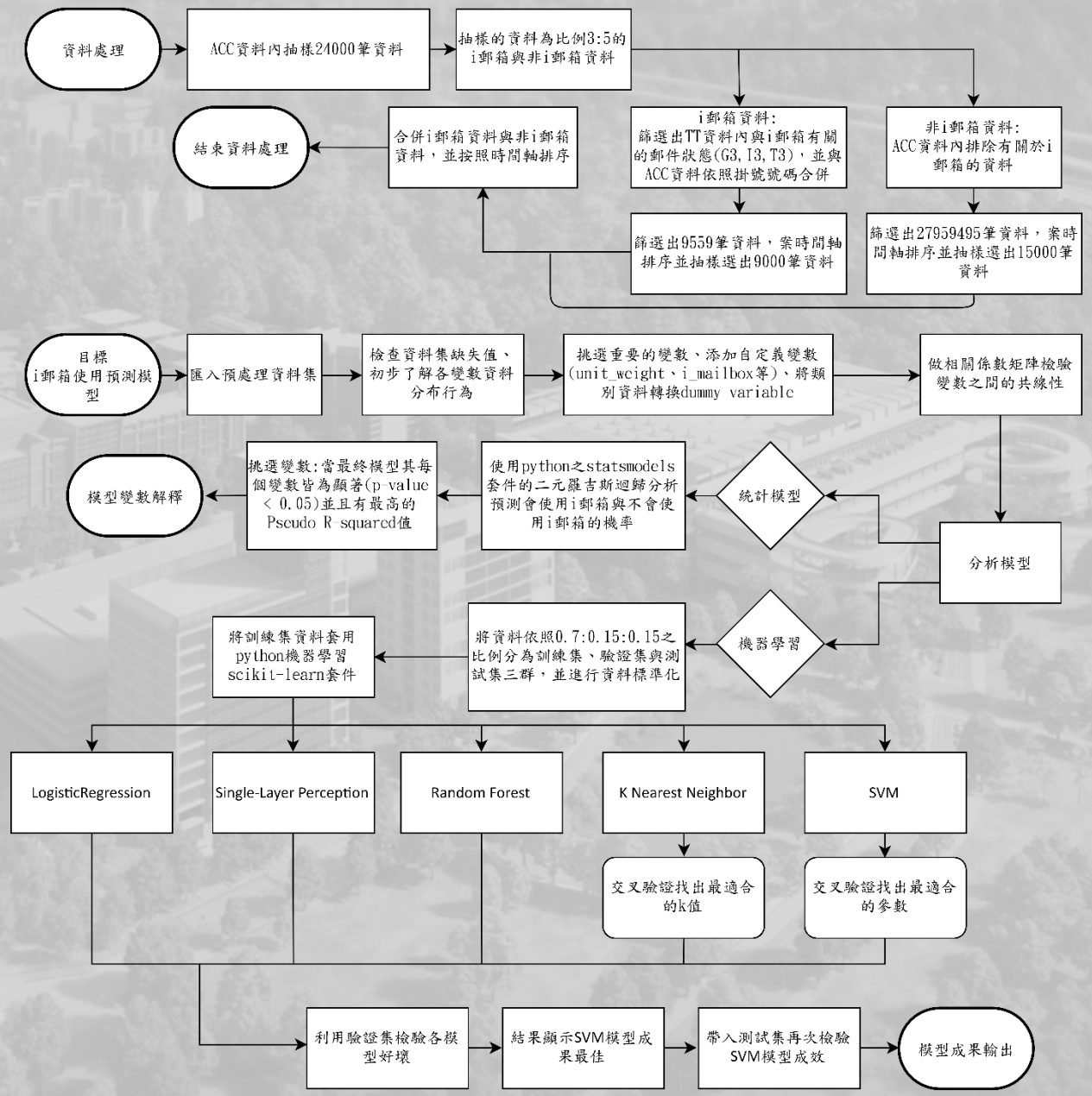
i郵箱相關之內容

與其他抽樣內容

數據分析

使用方法：

- 相關係數矩陣
- 二元羅吉斯回歸分析
- 機器學習



分析方法選用與程序

本研究採取[i_mailbox]欄位作為判斷資料是否屬於i郵箱，並透過傳統統計模型與機器學習之方法並行實作。之所以會使用兩種方式實作是因為兩種方式的目的並不同。對於統計模型而言，我們可以從中推論變數之間的關係，但預測結果不一定準；至於機器學習，我們可以很精準的預測結果，但不知道中間的計算過程，因此無法了解變數之間的關係。

詳細分析流程請參考流程圖。這裡僅解釋針對機器學習重要的分析邏輯與理論架構。

抽樣

27,000,000筆資料抽樣24000筆資料在99%信心指數下誤差為1%

數據分類為三類

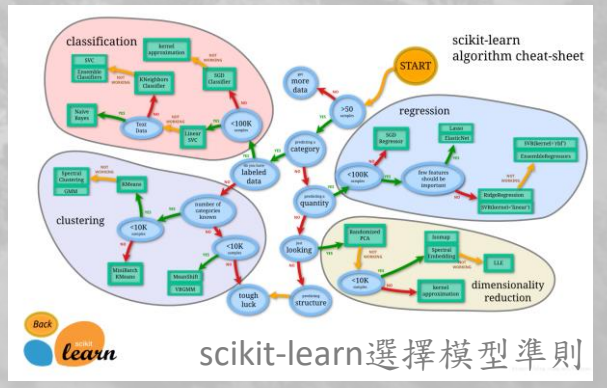
訓練數據(70%)：用來訓練模型的數據
驗證數據(15%)：用來檢驗模型準確率
測試數據(15%)：再一次確認驗證模型的好壞

模型選擇

參考scikit-learn官網簡略介紹選擇模型準則
sample>50筆 → 預測類別 → 有labeled data → sample<100K筆
最後選出Logistic Regression、Single-Layer Perception、Random-Forest、K Nearest Neighbor、SVM五種model。
需注意的是我並不知道資料是不是線性。

檢驗模型好壞

利用混淆矩陣（confusion matrix）
計算準確率作為評斷標準



scikit-learn選擇模型準則

分析方法選用與分析結果

統計模型｜二項羅吉特回歸模型

```
Optimization terminated successfully.
         Current function value: 0.536263
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:              i_mailbox   No. Observations:                24000
Model:                          Logit   Df Residuals:                    23986
Method:                           MLE   Df Model:                           13
Date:                Sun, 12 May 2019   Pseudo R-squ.:                  0.1894
Time:                        22:29:02   Log-Likelihood:                -12870.
converged:                       True   LL-Null:                       -15878.
                                        LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
acc29_5        0.8147      0.109      7.497      0.000       0.602       1.028
acc29_4        2.6832      1.049      2.559      0.011       0.628       4.738
acc29_3        2.4041      0.116     20.709      0.000       2.177       2.632
acc29_1        0.3901      0.107      3.662      0.000       0.181       0.599
acc24_2       -2.5950      1.008     -2.574      0.010      -4.571      -0.619
acc16_0        1.6279      0.219      7.429      0.000       1.198       2.057
acc12_2        1.6536      0.101     16.306      0.000       1.455       1.852
acc12_1       -0.5564      0.099     -5.639      0.000      -0.750      -0.363
office_time    0.2854      0.054      5.323      0.000       0.180       0.390
acc21         -2.5811      1.061     -2.433      0.015      -4.661      -0.502
acc32         -0.0059      0.001     -4.223      0.000      -0.009      -0.003
acc2           0.6626      0.061     10.897      0.000       0.543       0.782
unit_weight -9.957e-05   7.63e-06    -13.054      0.000      -0.000   -8.46e-05
const         -3.0185      0.261    -11.547      0.000      -3.531      -2.506
==============================================================================
```

結果顯示如上圖，最終得到的model其Pseudo R-squared 值為0.1894，且每一個變數之p-value皆小於0.05，顯示各變數具有顯著性。

$$f(x) = \ln\left(\frac{P_{\text{使用}i\text{郵箱}}}{P_{\text{不使用}i\text{郵箱}}}\right) = -3.0185 + 0.8147 * \text{acc29}_5 + 2.6832 * \text{acc29}_4 + 2.4041 * \text{acc29}_3 + 0.3901 * \text{acc29}_1 - 2.595 * \text{acc24}_2 + 1.6279 * \text{acc16}_0 + 1.6536 * \text{acc12}_2 - 0.5564 * \text{acc12}_1 + 0.2854 * \text{officetime} - 2.5811 * \text{acc21} - 0.0059 * \text{acc32} + 0.6626 * \text{acc2} - 0.00009957 * \text{unitweight}$$

我們可以從各變數的係數大小與正負號推論已知的事實或是假設

(1) 常數項為-3.0185顯示在不考慮其他變數影響下，民眾通常會選擇不使用i郵箱

(2) 計費方式皆為正數，表示當計費方式為1重量計費3單一4上收5首次使用便利箱袋都會提高使用i郵箱的機率，而其中上收的正面效應最強

(3) 法定紙幣報值會降低使用i郵箱的機率

(4) 包裹對於使用i郵箱的機率會提高

(5) 當超過郵局營業時間時，使用i郵箱機會會提高，符合i郵箱之設立目的

(6) 報價會降低使用i郵箱的機率

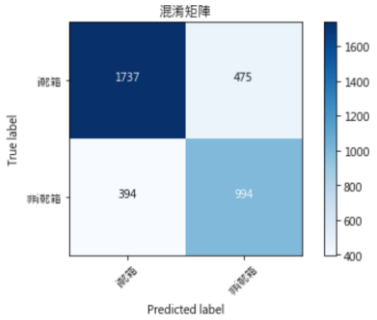
(7) 是否特約的係數(acc2)對於使用i郵箱為正面效應，可能原因是因為成為特約戶能得到的折扣優惠比較高，因此選擇使用i郵箱的意願會提高

(8) 寄件數越多、寄件單位重越重，會降低使用i郵箱的機率，可能和i郵箱的容量限制與價格有關

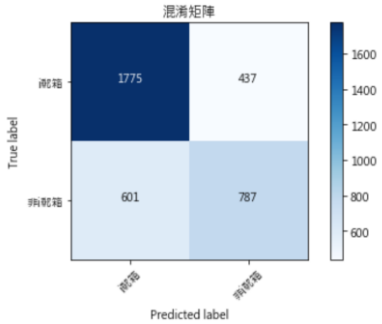
分析方法選用與分析結果

機器學習

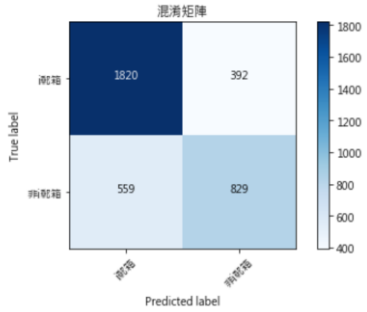
Logistic Regression



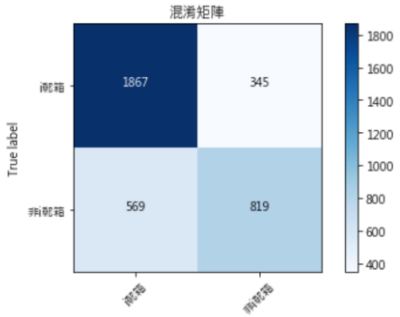
Single-Layer Perception



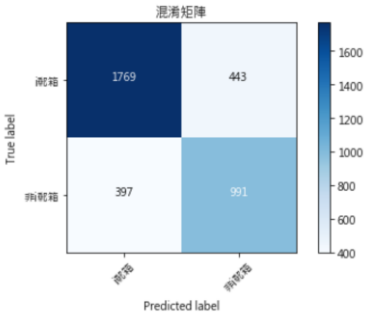
Random-Forest



K Nearest Neighbor

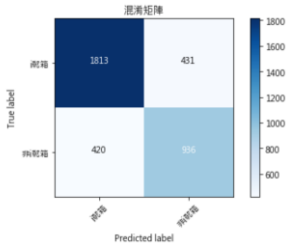


Logistic Regression



	Logistic Regression	Single-Layer Perception	Random-Forest	K Nearest Neighbor	SVM
準確率	0.7568	0.71	0.74	0.75	0.77

結果顯示SVM model 準確率最高，因此使用training後的SVM 模型帶入testing data再次驗證模型的好壞，結果如右圖，其準確率=0.76，可以看出差異不大，因此svm模型具有參考價值。



K值設定過小會降低分類精度；若設定過大，且測試樣本屬於訓練集中包含資料較少的類，則會增加噪聲，降低分類效果。

交叉驗證找出最適合的K值，結果顯示K=15準確率最高，因此使用K=15

交叉驗證得到參數為(C=1000, gamma=0.0001)準確率最高

提案未來方向

我們透過統計模型與機器學習方法找出使用i郵箱與不使用i郵箱的分類模型與可能影響民眾選擇的變數，有助於未來找出提高使用i郵箱使用率的策略。然而，ACC資料檔的資料包含的資訊仍不足夠，像是寄件費用、運送距離等重要資訊在未來必須考慮。此外，由於我們對於抽樣資料的不了解，可能會導致違背機器學習相關模型的背後假設，因此在未來仍必須針對理論架構修正。

參考資料

封面底圖 | https://www.logisticnet.com.tw/publicationArticle.asp?id=736

抽樣 | https://zh.surveymonkey.com/mp/margin-of-error-calculator/

機器學習 | https://scikit-learn.org/stable/

https://kknews.cc/zh-tw/news/b4n3bnm.html

i郵箱介紹 | https://www.post.gov.tw/post/internet/Postal/index.jsp?ID=1467188792821

機器學習與統計模型 | https://buzzorange.com/techorange/2019/05/02/difference-between-statistics-and-machine-learning/

Knn | https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/635163/

簡報結束
敬請指教