嶺東科技大學
LING TUNG UNIVERSITY
資訊管理系

大數據分析應用-以中華郵政為例

指導教授： 王曉玫 教授
組員名單： 陳劭泓 A58C010
桑文濤 A58C012
鍾 戎 A58C036
許丞嘉 A58C042

中華民國一〇八年十二月

嶺東科技大學

資訊管理系

大數據分析應用–以中華郵政為例

中華民國一〇八年十二月

嶺東科技大學
LING TUNG UNIVERSITY

資訊管理系專題口試委員審定書

大數據分析應用-以中華郵政為例

指導教授：王 曉 玫 教授

組員名單：陳 劭 泓 學號 A58C010

桑 文 濤 學號 A58C012

鍾 戎 學號 A58C036

許 丞 嘉 學號 A58C042

指導教授：_____

口試委員：_____

_____

中華民國 年 月 日

謝　　誌

本專題報告得以順利完成，首先要感謝恩師王曉玫老師細心引導我們，耐心的協助我們，克服研究過程中所面臨的困難，給予我們最大的協助，使本專題得以順利完成。

研究報告口試期間，感謝李靜怡老師、陳明華老師不辭辛勞細心審閱，不僅給予我們指導，並且提供寶貴的建議，使我們的專題內容以更臻完善，在此由衷的感謝。

最後，感謝系上諸位老師在各學科領域的熱心指導，增進資訊管理知識範疇，在此一併致上最高謝意。

陳劭泓、桑文濤、許丞嘉、鍾戎

謹誌

中華民國一〇八年十二月於嶺東

摘　要

現今各個企業的數據量逐日增加，如何透過大數據分析的方式，從中取得有價值的資訊，並將數據轉化為商機，幫助企業在轉型或創新的道路上，發揮更大的空間，亦或者提升營運績效、尋找潛在客戶等，成為目前各個企業追逐的目標。

中華郵政為台灣最早的郵務中心，橫跨百年的歷史，最早為軍事用途之郵局，爾後引入郵票制度讓民間也能夠進行信件與文書之寄送，除了郵政業務領域外，其業務亦拓展到儲蓄、人身保險、金融服務等等；近年來，中華郵政引入大數據分析系統以及人臉辨識，希望能夠經由大數據分析來提升銷售業績以及提升服務品質。

本專題藉由參加「2019 年中華郵政大數據競賽」所提供之一季營運數據，進行資料整理、視覺化分析、以及永久戶預測之實作分析。結果顯示，全台以新北市郵件量最多，占全部郵件量的 29.99%，其次為臺北市和臺中市，各占全台的 17.15%和 13.27%，一般縣市中，郵件量以彰化縣為最大宗，占一般縣市的 32.38%，其原因為彰化縣為六都之外第一大縣，人口和公司行號數量僅次於六都，故彰化縣有較大的發展空間。在產業別方面，中華郵政特約戶多為製造業、零售業、金融業，建議可針對這三種產業推出優惠方案。此外，是否為特約戶主要會受專業議價、是否為混合交寄議價戶、和地區的影響，不同的產業也會影響特約戶的簽訂。

關鍵詞：中華郵政、大數據、資料分析

目錄

圖目錄

表目錄

第壹章　緒論

1.1 研究動機

現代企業有許多會擁有無數多的客戶或業務相關資料，而企業本身若能對所擁有的客戶資料進行習慣上的分析，進而幫助企業本身進行業務推廣或拓展業務範圍，便能達到提升營運效益。

中華郵政為台灣老字號之郵政機構，其核心價值為【以客為尊、提供誠信效率的服務】，近年來積極的進行數位轉型和邁向智慧物流，整合郵政物流、金流及資訊流功能，提供普遍優質之郵儲壽服務，以便帶給台灣民眾更好、更便利的郵政服務，以突破既有的經營限制、並增加市場競爭力。中華郵政於 2019 年 2 月舉辦大數據賽，尋求各方面的意見，以加速推動郵政「智慧物流」及邁向「數位轉型」發展，帶給臺灣民眾更好、更便利的郵政服務。因此，本專題藉由參加「2019 年中華郵政大數據競賽」所提供之一季營運數據進行實作分析。

1.2 研究目的

由中華郵政提供之營運資料(2018 年 1 月 1 日~2018 年 3 月 31 日)得知，全台的產業最多為製造業、零售業、金融保險業，若與政府統計的資料(圖 1.1)進行對比，發現中華郵政尚有許多潛在客戶可以發掘(圖 1.2)；此外，可藉由觀察客戶的特性，尤其是永久特約戶(本專題之後簡稱為特約戶)，探討如何提升中華郵政業務量。因此，本專題將以資料視覺化和決策樹模型預測的方法對中華郵政進行分析，並透過分析結果給予中華郵政一些決策上的建議。主要研究目的如下：

1. 分析中華郵政在各地營運上的郵務負擔狀況和其郵件種類的收受狀況；
2. 探討各地區不同產業特約戶合約的簽訂狀況；
3. 以模型預測未來特約戶的辨識方式，探討該如何在業務上提升簽約效率。

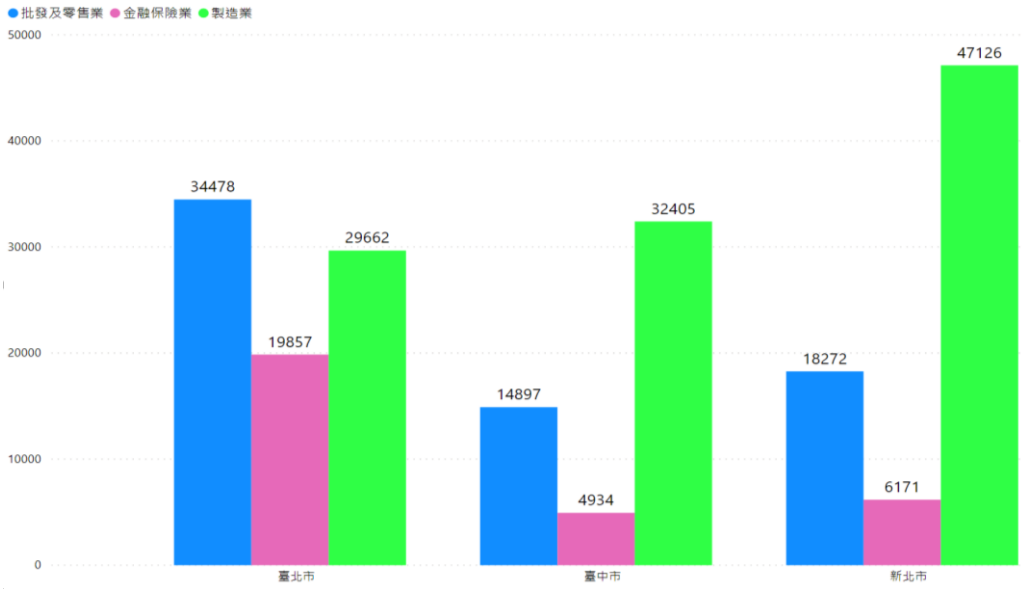

圖 1.1 官方簽約戶數

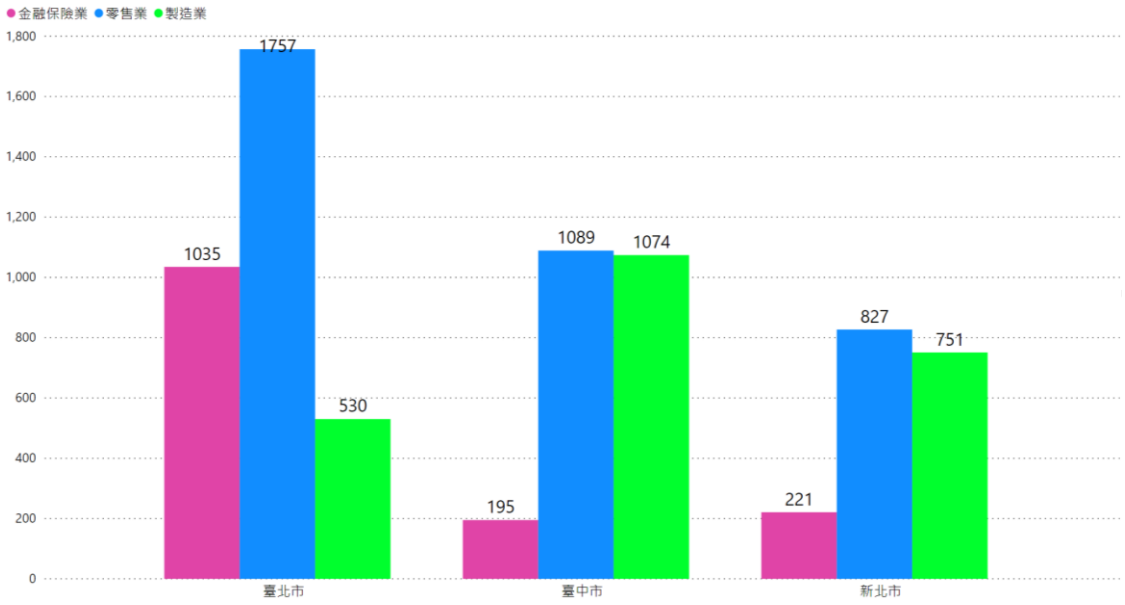

圖 1.2 簽約戶數

第貳章　文獻回顧與探討

2.1 何謂大數據

大數據又稱為巨量資料，主要以龐大的資料量(volume)、速度(velocity)與資料多樣性(variety)為主，除了結構化資料外，巨量資料中還包含許多半結構化或非結構化資料，巨量資料通常具有時效性，一旦取得必須盡快分析，以取得分析結果才能發揮其最大價值。一般而言，大數據主要使用的分析方法有以下幾種：資料視覺化、決策樹、隨機森林、預測性分析，本專題僅針對資料視覺化及決策樹作介紹[1]。

2.2 資料視覺化

資料視覺化是借助軟體的功能將數據製作成完整的圖表，以圖形化手段清晰有效地傳達與溝通訊息，可視化分析能夠直觀的呈現數據，並透過視覺化的方式從不同的圖形中找出不同的隱藏資訊[2]；一般而言，視覺化可使用的軟體有 Power BI、Tableau、QlikView、Kibana……等。

資料視覺化在各個領域都有著較好的輔助效果，張志偉(2016)將高齡者運動數據上傳網站並以 Tableau Desktop 做視覺化呈現，讓醫護人員在網頁中觀看數據，進而輔助人員進行分析和成效檢討[3]，羅意智(2018)使用網路爬蟲取得銷售資料，以 Kibana 將資料視覺化，並建置出幫助買家尋找 出推薦賣家的系統[4]；本專題則使用 Power BI 對郵政資料進行資料視覺化。

2.3 決策樹

決策樹是用於分類和預測的主要技術之一，利用像樹一般的圖形或決策模型的決策支持工具，決策樹的目的是找出屬性和類別間的關係。它是利用詢問一系列問題，將數據進行分割，並根據不同屬性值判斷節點向下的分支，後於決策樹的葉節點得到結論[5]。一般而言，可依據正確率(Accuracy)、精確率(Precision)、及召回率(Recall) 判斷決策樹模型預測好壞的，其定義如下(表 2.1)；本專題使用 RapidMiner 內的決策樹來尋找客戶特約戶的特徵。

表 2.1 2X2 列聯表

	實際值為 1	實際值為 0	合計
預測值為 1	a	b	a+b
預測值為 0	c	d	c+d
合計	a+c	b+d	n

$$正確率(\text{Accuracy}) = 全部資料中有多少筆被正確預測 = \frac{a+d}{n} \text{ x } 100\%$$

$$精確率(\text{Precision}) = 被預測出的資料有多少是正確的 = \frac{a}{a+b} \text{ x } 100\%$$

$$召回率(\text{Recall}) = 實際資料中有多少筆被正確預測出來 = \frac{a}{a+c} \text{ x } 100\%$$

決策樹除了可以利用於商業上的分析外，亦可用於許多領域上的研究，如呂泇都(2011)運用決策樹、類神經網路、羅吉斯迴歸，建立淋巴癌預測模型，用以預測淋巴癌病人的預後 2 年存活狀況[6]，李語嫣(2010)使用健檢、生活習慣資料建立糖尿病預測模型，研究中使用了關聯規則、決策樹、時間序列等資料探勘技術，用以協助醫療人員對於糖尿病的診斷，並達到提早預防的成效[7]。許智翔(2018)使用決策樹於臺灣地層含水量預測，分析超抽地下水對地層之影響並給予建議[8]，張瀚文(2014)使決策樹、遺傳演算法進行水庫防洪減淤研究，協助災害來臨時，如何規劃最佳操作模式，達到提升水庫排砂量，延長水庫壽命等目的[9]。因此，本專題使用決策樹對中華郵政資料進行特約戶的預測，盼能提升郵政業務。

第參章　研究方法

3.1 研究步驟

本專題流程先從界定研究方向，確定研究方向後，將資料下載並進行清理和整合，再將資料進行分析，最後取得資料分析結果，如圖 3.1。

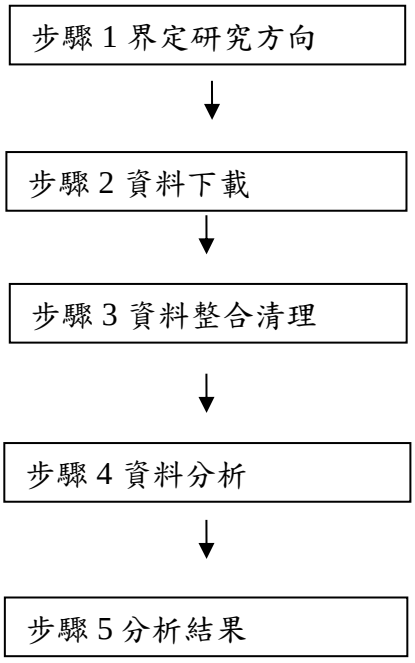

圖 3.1 研究步驟流程圖

3.2 研究工具

3.2.1 Power BI

本專題使用 Power BI Desktop (Power BI 桌機版)進行資料分析和資料視覺化，Power BI 是一套商務分析工具，它能夠對資料進行獲取與清理、資料建模、資料視覺化，可以將得到視覺化的資料以即時儀表板和報表方式呈現[10](圖 3.2)。

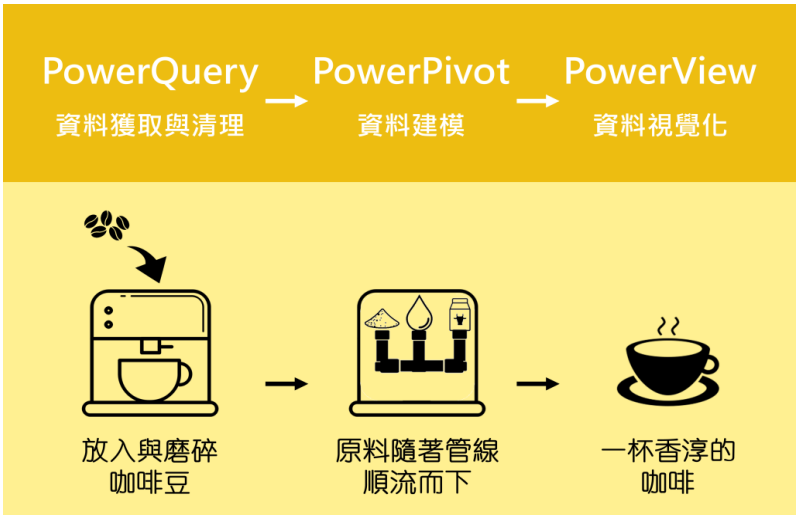

圖 3.2 Power BI 處理流程圖

3.2.2 SPSS

圖 3.3，SPSS 為 IBM 推出的進階統計分析軟體，可用於統計學分析運算、數據挖掘、預測分析等，亦可用來支持決策，之所以不使用 Excel 進行資料整理原因為，Excel 資料筆數上限為 1,048,576 筆，而中華郵政提供之資料超越 Excel 上限；因此，本專題使用 SPSS V.26 進行數據清理及整合[11]。

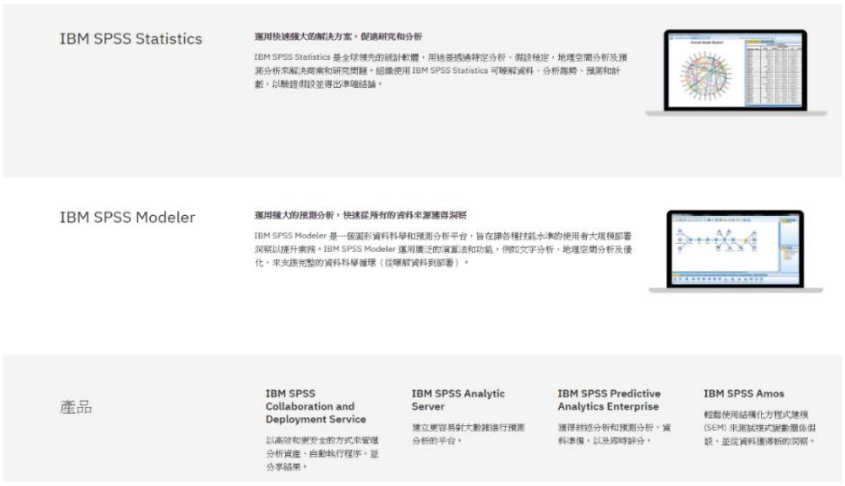

圖 3.3 SPSS 特色

3.2.3 RapidMiner

RapidMiner 為數據科學軟體，可用於機器學習、深度學習、預測分析等，並支援結果可視化、模型驗證、優化，本專題使用 RapidMiner 進行決策樹分析[12]。

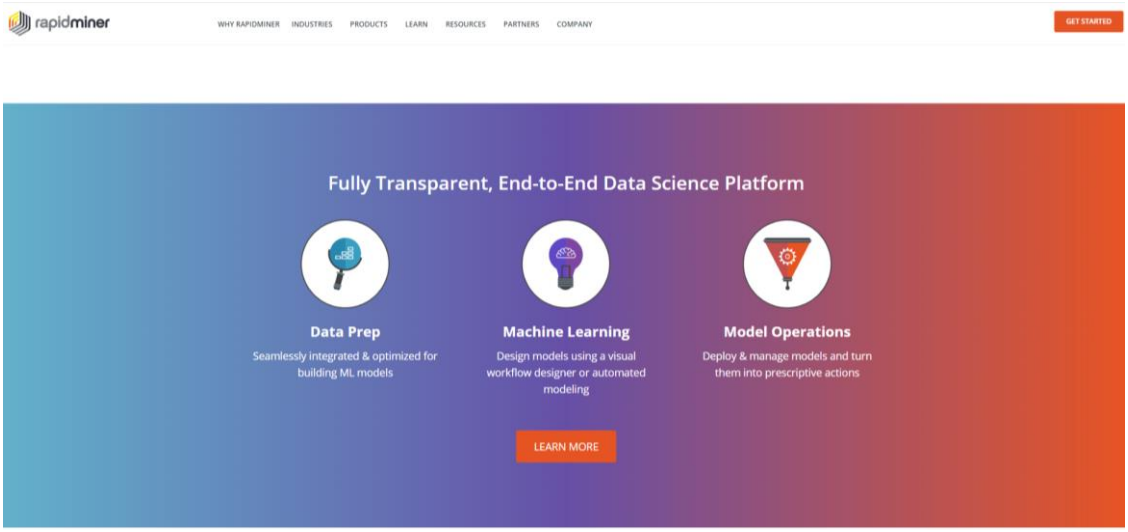

圖 3.4 RapidMiner 特色

3.3 研究資料

本專題研究的資料為中華郵政所提供之一季的郵政相關資料，以及自行取得之開放資料兩個方面。

3.3.1 中華郵政提供資料

中華郵政提供競賽資料【收寄資料明細檔】及【特約客戶主檔】，檔案大小共為 4.5GB ，資料收集時間為 107 年 1 月 1 日到 3 月 31 日，合計筆數為 2 千 8 百 02 萬 5420 筆。

3.3.2 開放資料

政府的開放資料平台下載多個資料檔，以輔助資料分析，如下所示。

(1)中華郵政全國營業據點 ,Post_All 檔，資料網址:
https://scidm.nchc.org.tw/en/dataset/315830000m- 000003/resource/845f4372-950b-4bd3-a8a6-67b8bb2378da。

(2)臺灣地區郵遞區號前 3 碼一覽表,臺灣地區郵遞區號前 3 碼一覽表檔，資料網址:
https://www.post.gov.tw/post/internet//Download/all_list.jsp?ID=2201#dl_link_178。

(3)108 年 09 月底存活公司行業別家數統計 108 年 09 月底存活公司行業別家數統計檔，資料網址:
https://serv.gcis.nat.gov.tw/StatisticQry/cmpy/StaticFunction1.jsp。

3.4 資料清理流程

圖 3.5 為本專題資料清理流程圖。

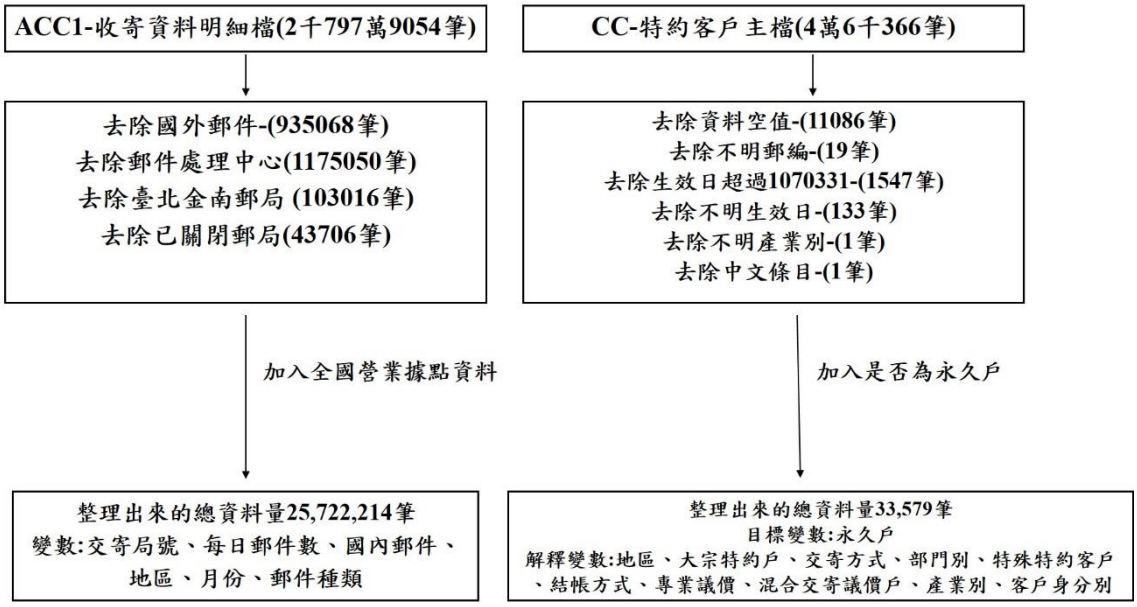

圖 3.5 資料清理流程圖

3.4.1 收寄資料檔

在收寄資料檔中有 27,979,054 筆資料，我們將符合以下情況之資料去除。

(1). 國外郵件(935,068 筆)，因本專題分析郵件範圍為國內，故不使用國外郵件。

(2). 郵件處理中心郵件(1,175,050 筆)，無法得知郵局編號及地區所在，故將其去除。

(3). 臺北金南郵局(103,016 筆)，相當於臺北地區的國內外郵件集散地，因有大量郵件聚集於此，故將其去除。

(4). 已裁撤的郵局(43,706 筆)。

再加入全國營業據點資料，並將各地郵局進行地區分類，最後整理出來的總資料量一共有 25,722,214 筆資料，表 3.1 為收寄資料檔中使用的變數。

表 3.1 收寄資料明細檔變數表

變數名稱	變數解釋
交寄局號	收寄郵件之郵局代號
每日郵件數	每日收受郵件量
國內郵件	臺灣包含外島之郵件
地區	臺灣地區分類
月份	收寄月份
郵件種類	各郵件之分類(1 函件 2 包裹 3 快捷)

3.4.2.特約客戶主檔

特約客戶檔中有 46,366 筆資料，去除空值(11,086 筆)，郵局內沒有的郵編代號(19 筆)，日期不明(1547 筆)及不符的資料一共(133 筆)去除，中華郵政代號查詢不到的產業資料(1 筆)，中文條目(1 筆)，再加入永久戶的資料，合計 33,579 筆資料。表 3.2 為特約客戶主檔中所使用的變數。

表 3.2 特約客戶主檔變數表

變數名稱	變數解釋
永久戶	是否為永久客戶(0 否 1 是)
地區	臺灣地區分類
大宗特約戶	是否有申請大宗回傳服務(0 否 1 是)
交寄方式	郵件交寄的方式
部門別	是否有部門別(0 否 1 是)
特殊特約客戶	是否為特殊特約客戶(0 否 1 購物平台業者)
結帳方式	結帳方式(0 當場繳清 1 記帳)
專業議價	是否有專業議價(0 否 1 是)
混交寄議價戶	是否有混合交祭議價(0 否 1 是)
產業別	客戶的產業別
客戶身分別	客戶的身分別

3.5 資料分析

本專題以敘述統計之次數分配、交叉表等方式進行郵政資料分析， 並以 Power BI Desktop 作視覺化方式呈現；此外，以 RapidMiner 進行決策樹分析建立預測模型。

3.6 硬體配置

表3.3為整理資料和分析時使用的硬體配置。

表3.3硬體配置表

CPU	Intel Core I7-8750H 4.1GHz
RAM	DDR4 32GB 2666MHz
SSD	M.2 PCIE 512GB
HDD	1TB 5400RPM (SSHD)

第肆章　研究結果與討論

4.1 儀表板架構

圖 4.1 儀表板(1)

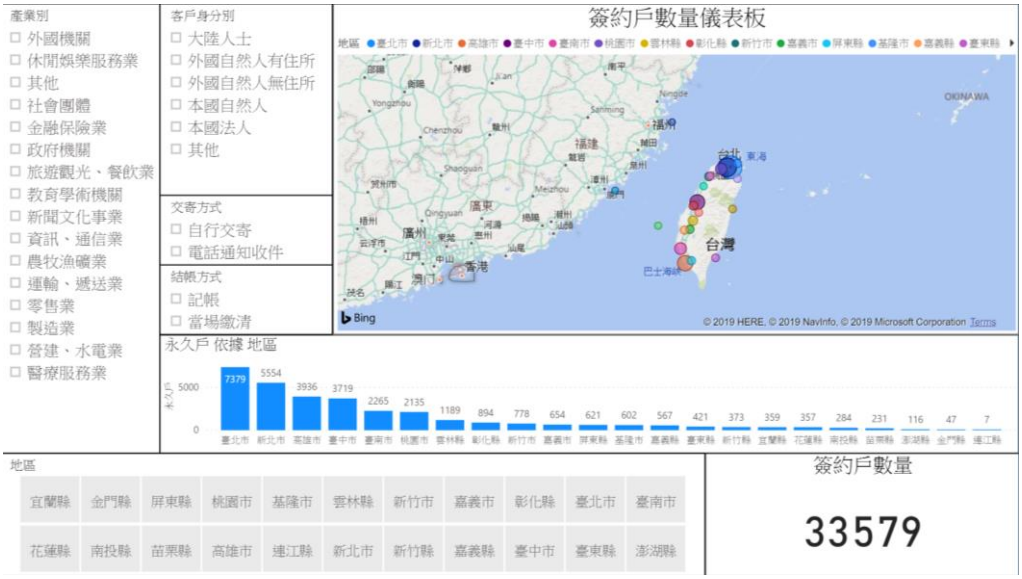

圖 4.2 儀表板(2)

圖 4.1，圖 4.2 為進行視覺化分析時所使用的儀表板，圖 4.1 用於全台縣市郵件量的分析，圖 4.2 用於簽約戶數量分析。

4.2 特約戶分析

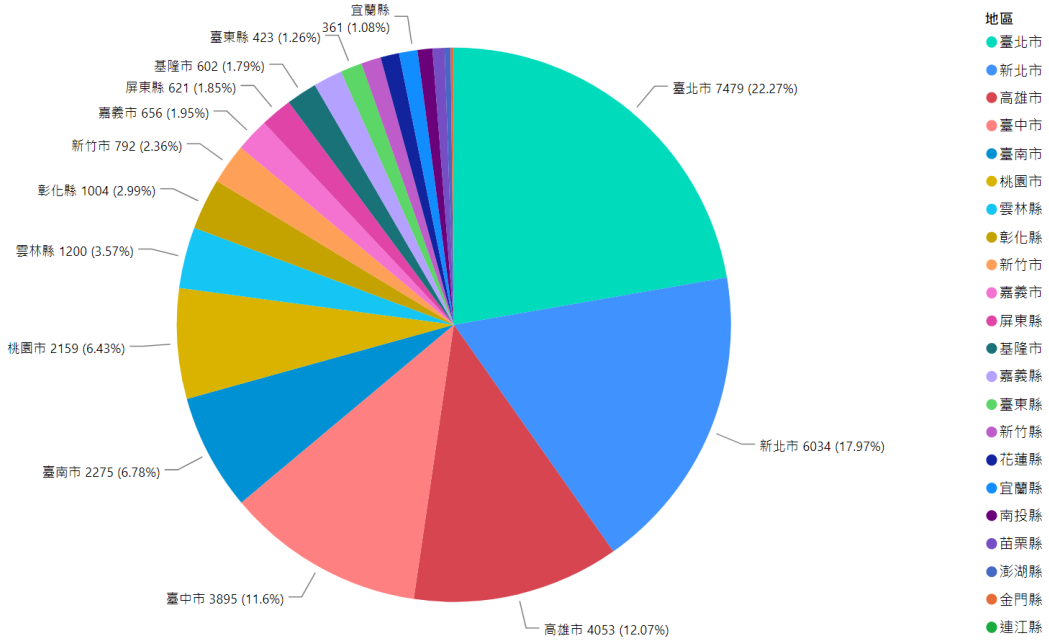

圖 4.3 特約戶地區分布圖

圖 4.3，全台各縣市的特約戶多集中在六都地區，且客戶都有在兩千名以上，並有 77.12%的特約戶在六都。六都中以臺北市的特約戶最多，數量有 7479 位，占全台總數的 22.27%，桃園市是六都中最少的，特約戶數為 2159 位，占全台 6.43%，連江縣的特約戶是最少的，只有 7 位，占 0.02%。

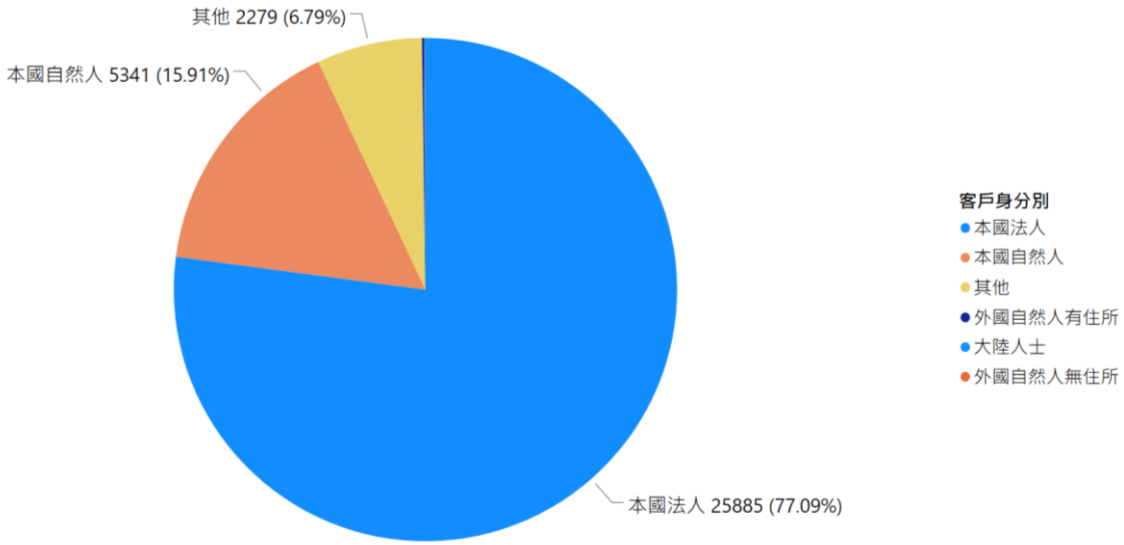

圖 4.4 特約戶客戶別分布圖

圖 4.4，在全台的客戶中，以本國法人簽訂特約戶為最大宗，有 25885 位，並且占整體的 77.09%，其次為本國自然人，有 5341 位客戶，占 15.91%，最少的為外國自然人無住所，只有 7 位，占 0.02%。

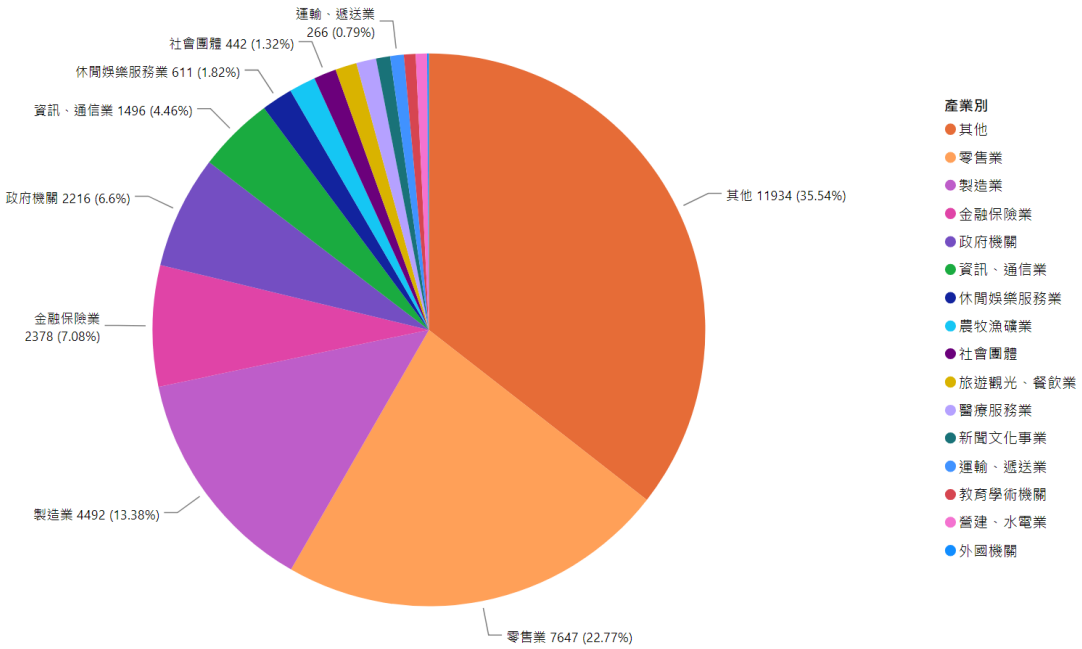

圖 4.5 特約戶產業分布圖

圖 4.5，在特約戶的產業別中，以其他產業為最大宗，有 11934 位，占整體 35.54%，第二多的簽約戶是零售業，客戶數為 7647 位，占總體 22.77%，再者是製造業第三多，客戶為 4492 位，占 13.38%，最少的是外國機關，只有 35 位客戶，占 0.1%。

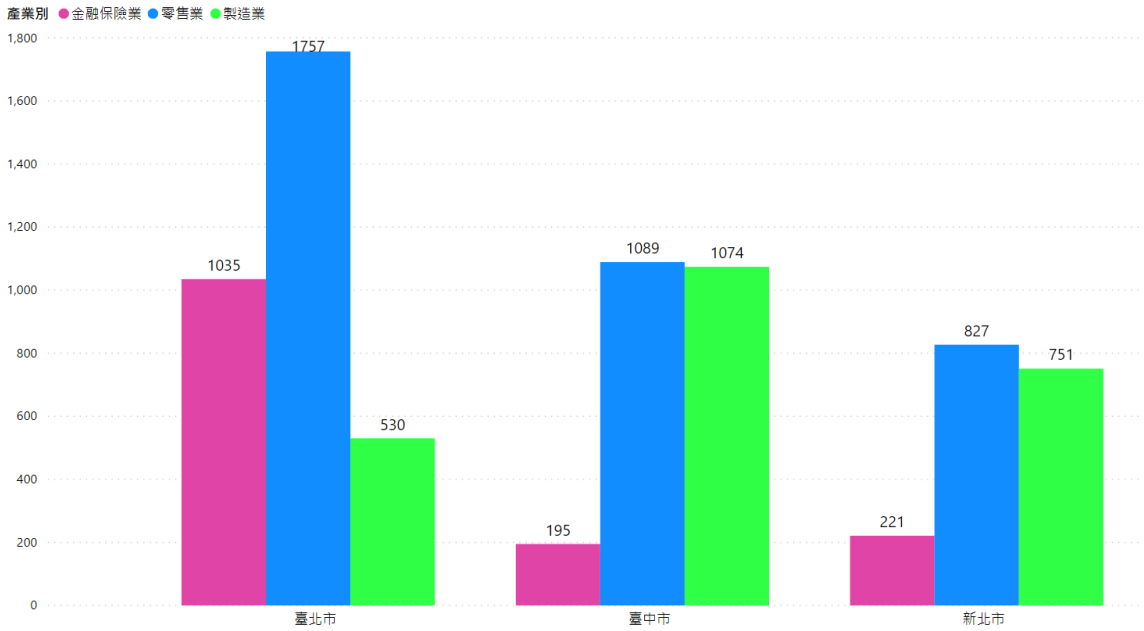

圖 4.6 前三多城市產業分布圖

從特約戶前三多的城市(臺北市、臺中市、新北市)分析(圖 4.6)，臺北市的特約戶以零售業為最大宗，有 1757 個客戶，占臺北市 52.89%，第二多為金融保險業有 1035 個客戶，占有 31.16%，第三為製造業，有 530 個客戶，占 15.95%。臺中市最大宗是零售業，有 1089 個客戶，占 46.18%，第二為製造業，有 1074 個，有 45.55%，第三是金融保險業，有 195 個，占 8.27%。新北市最大宗為零售業，有 827 個客戶，占 45.97%，第二多為製造業，有 751 個，占 41.75%，第三為金融保險業，221 個客戶，占 12.28%。

4.3 郵件量分析

圖 4.7，全臺各縣市中，以新北市的郵件量為最大宗，占全臺 29.99%，第二多為臺北市，占 17.15%，再者是臺中市，占 13.27%，郵件量最少的是連江縣，占全臺灣 0.02%。

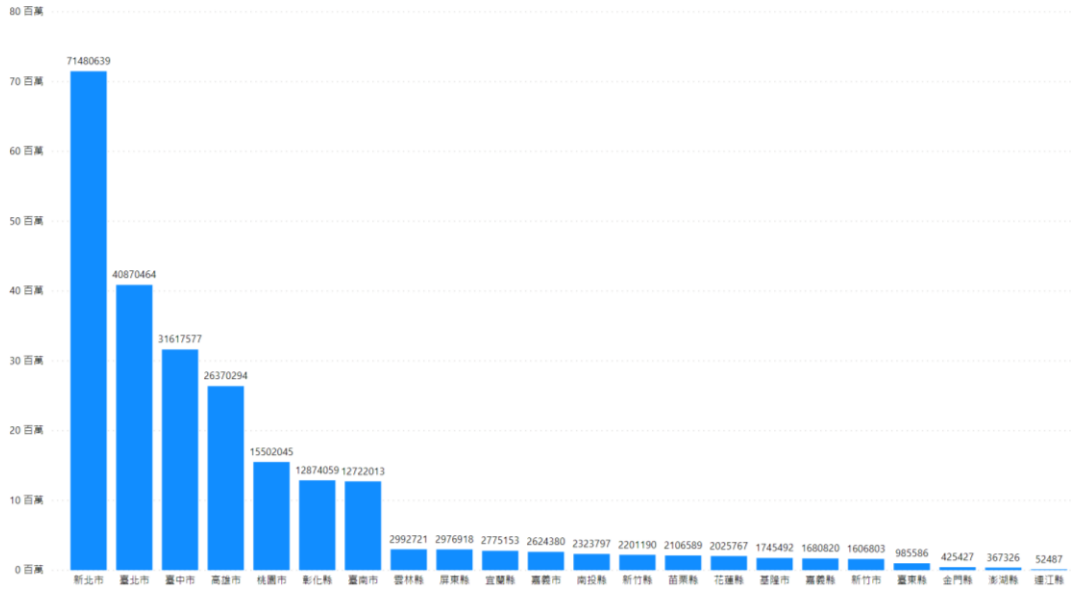

圖 4.7 各縣市郵件量分布圖

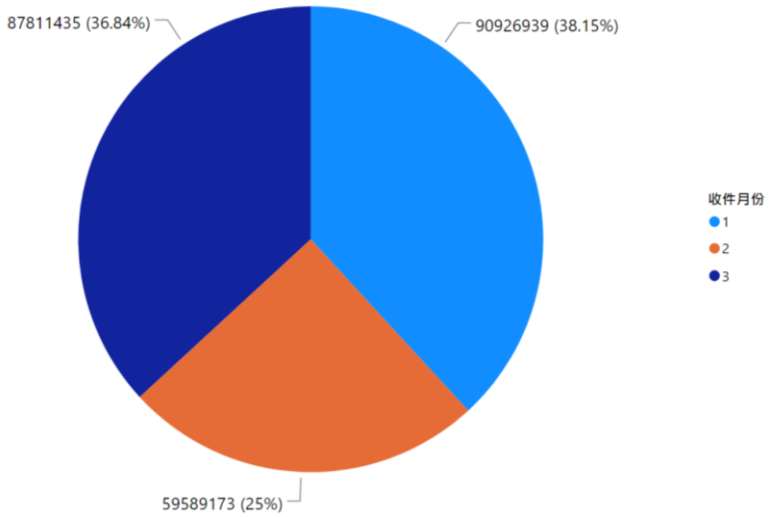

圖 4.8 三個月郵件量分布圖

圖 4.8，在三個月中，以一月的郵件量是最多的，達 9 千萬件郵件，占全部的 38.15%，而三月的郵件量第二多，為 8 千 7 百萬件左右，占 36.84%，最少的是二月，郵件量為 5 千 9 百萬件左右，占全部的 25%。

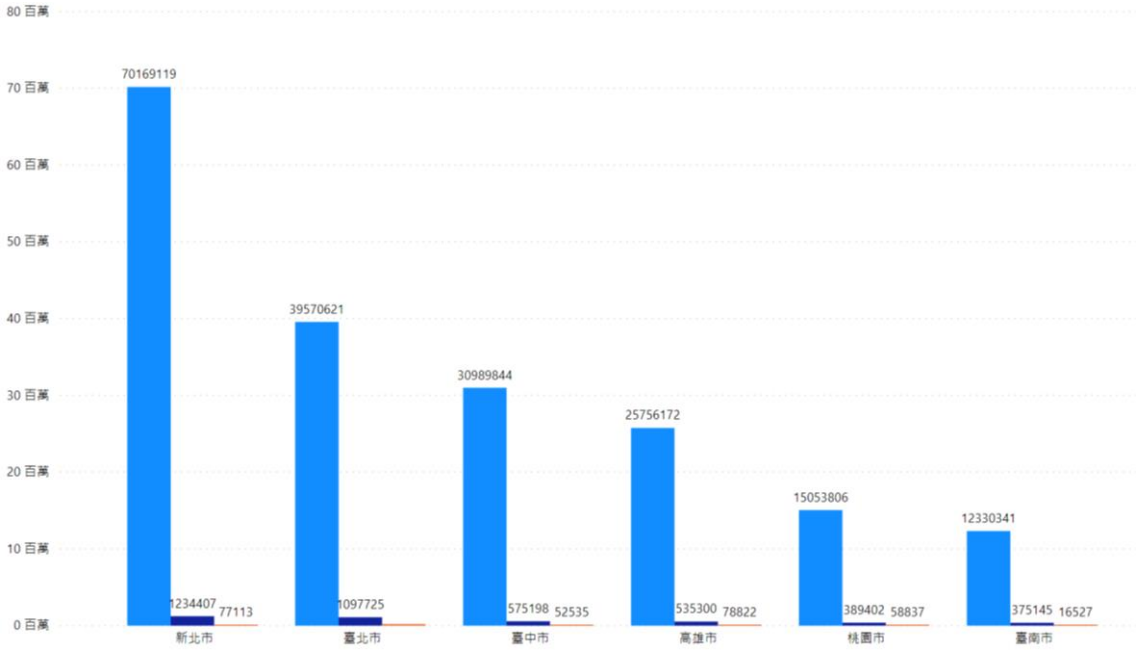

圖 4.9 六都郵件量分布圖

圖 4.9，在六都中，新北市的郵件總量是最多的，占六都郵量的 36%，第二名臺北市占 20.58%，第三多的臺中市占 15.92%，最少的臺南市只占有 6.41%。

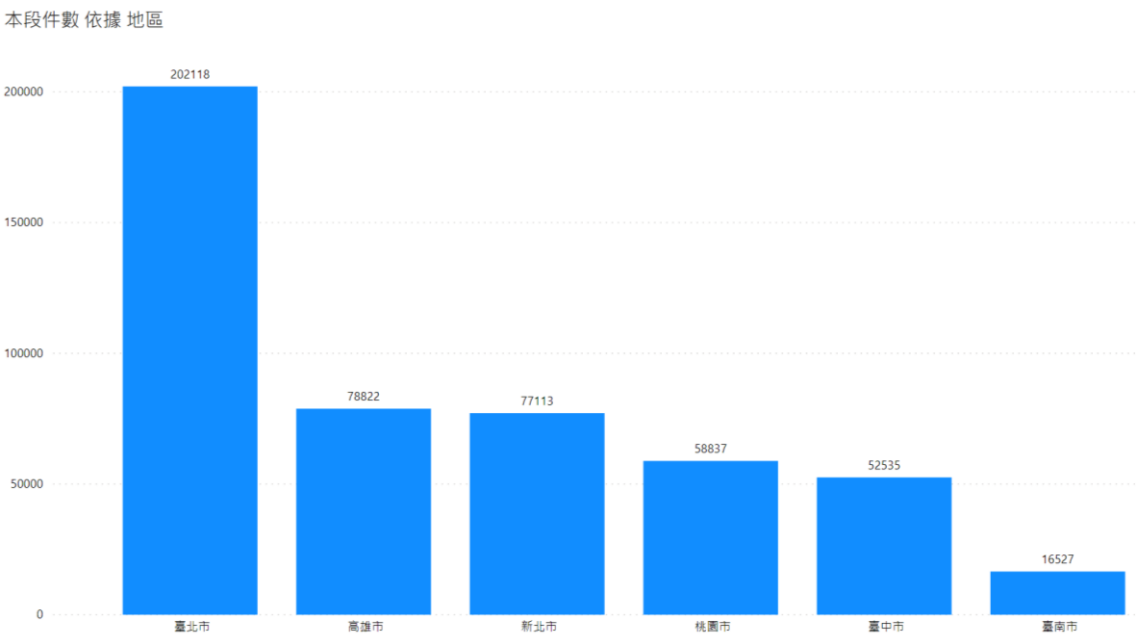

圖 4.10 六都快捷郵件量分布圖

圖 4.10，六都中的快捷郵件中，臺北市的快捷郵件量是最多的，占六都郵量的 41.59%，第二名是高雄市占 16.22%，第三多的新北市占 15.87%，最少的臺南市只占有 3.40%。

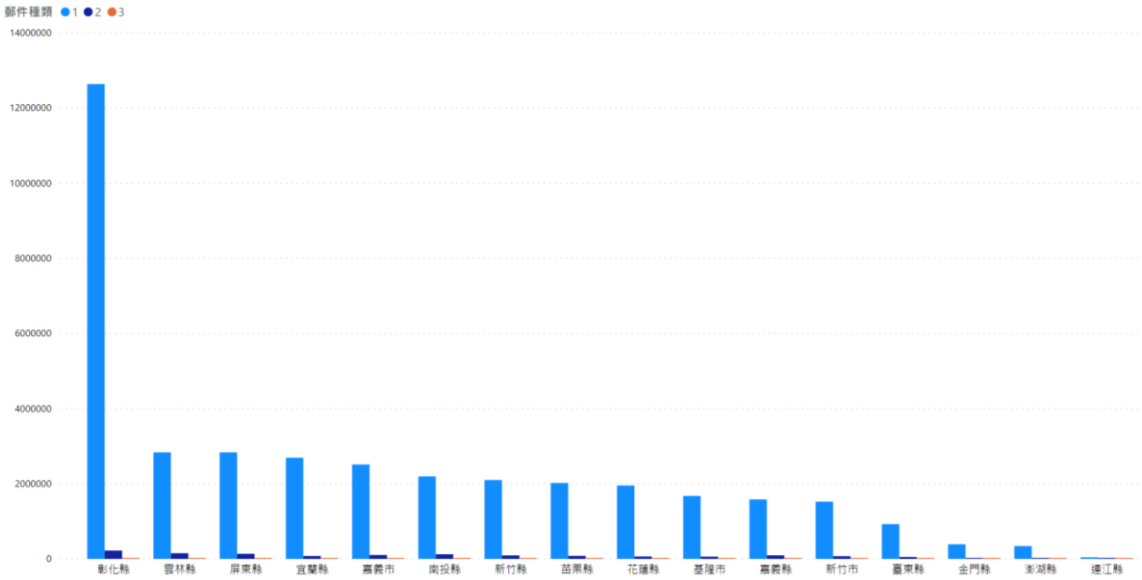

圖 4.11 一般縣市郵件量分布圖

圖 4.11，一般縣市中，彰化縣的郵件總量是最多的，占整體郵量的 32.38%，第二名雲林縣占 7.53%，第三多的屏東縣占 7.49%，最少的連江縣只占有總體的 0.13%。

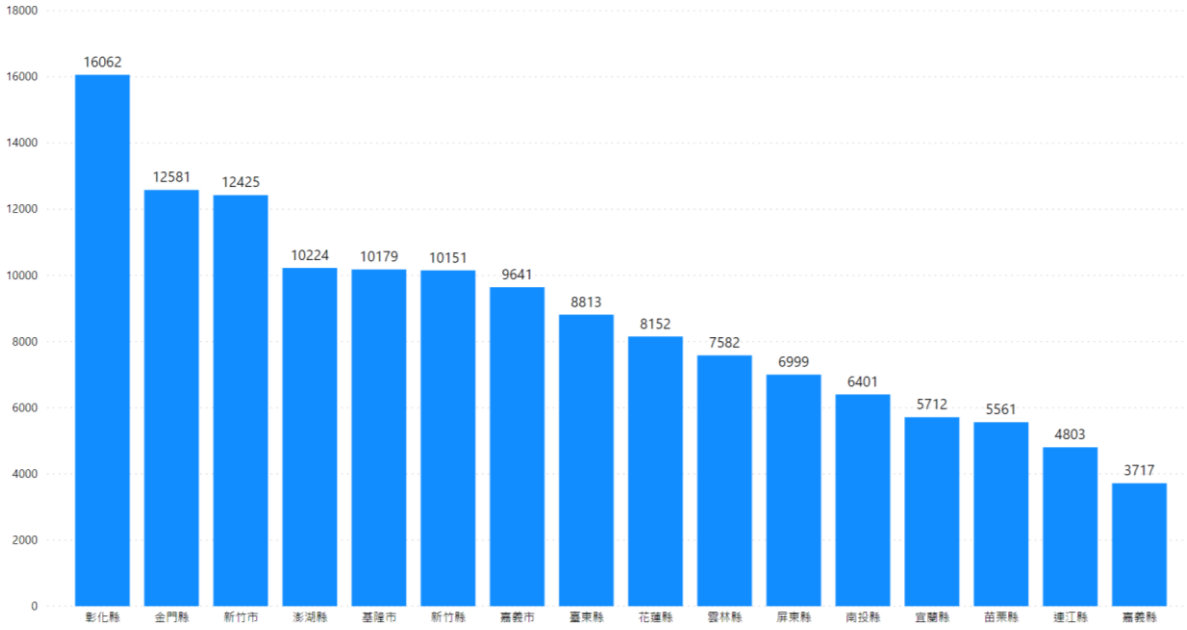

圖 4.12 一般縣市快捷郵件量分布圖

圖 4.12，一般縣市的快捷郵件中，彰化縣的快捷郵件量是最多的，占總體的 11.56%，第二名是金門縣占 9.05%，第三多的新竹市占 8.94%，最少的嘉義縣只占有 2.67%。

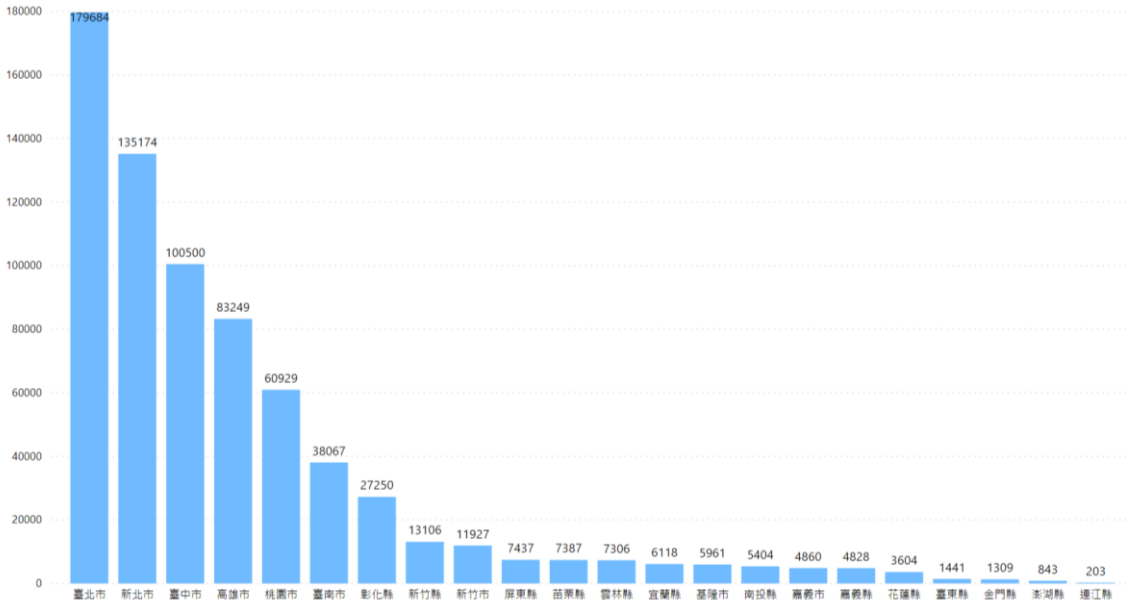

圖 4.13 全臺灣公司數量統計圖

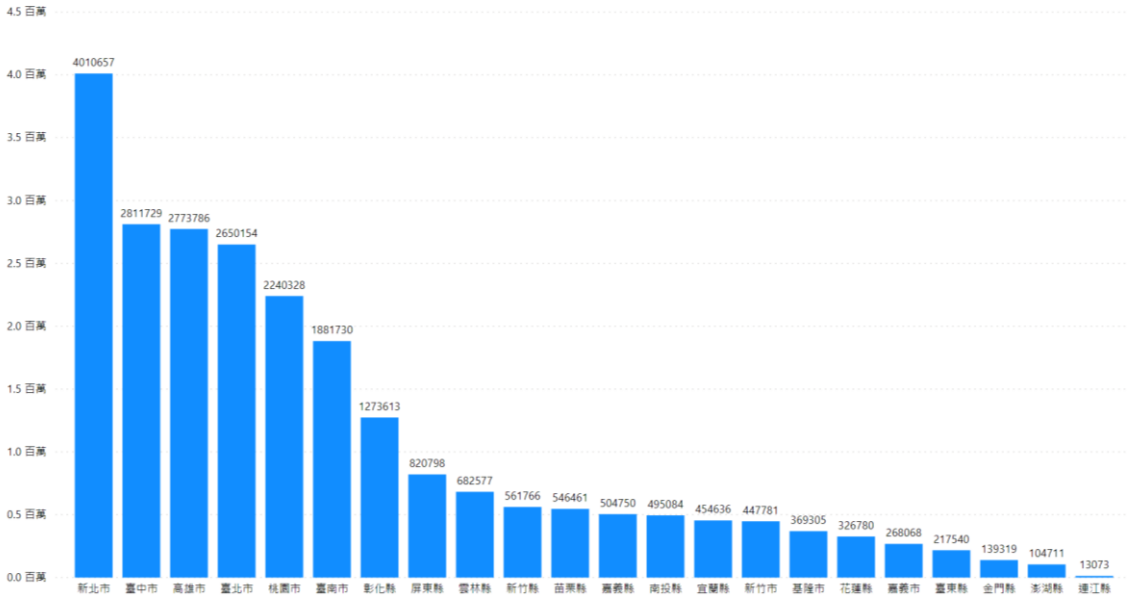

圖 4.14 全臺灣縣市人口統計圖

圖 4.13 顯示，除了六都之外，公司行號第一多的是彰化縣，有 27,250 間占全臺灣的 3.86%。由圖 4.14 得知，除了六都之外，人口最多的縣市是彰化縣，有 1,273,613 人占全臺灣的 5.40%，故可以驗證為何圖 4.11 及圖 4.12 彰化縣的郵件量是超越其他一般縣市的。

4.4 特約戶預測

本專題使用 RapidMiner 以隨機的方式將特約戶主檔分割為兩個檔案，分別為訓練資料和測試資料。

4.4.1 訓練資料及測試資料

表 4.1 為 RapidMiner 分割資料時所使用的模塊，首先透過 Retrieve 的功能將整理好的資料導入模塊中(圖 4.15)。圖 4.16，接著進入 Split Data 的設定中，設定分割資料比例，我們將資料設定為分割 60%(訓練資料)和 40%(測試資料)。將訓練資料、測試資料等檔案和原檔案進行敘述統計並進行對比，可以得知訓練資料、測試資料和原檔案具有一致性(表 4.2~表 4.4)。

表 4.1 RapidMiner 使用模塊

模塊名稱	說明
Retrieve	導入資料
Generate ID	將導入的資料編上 ID
Split Data	設定資料的分割比例
Write Excel	將分割好的資料寫入 Excel 中

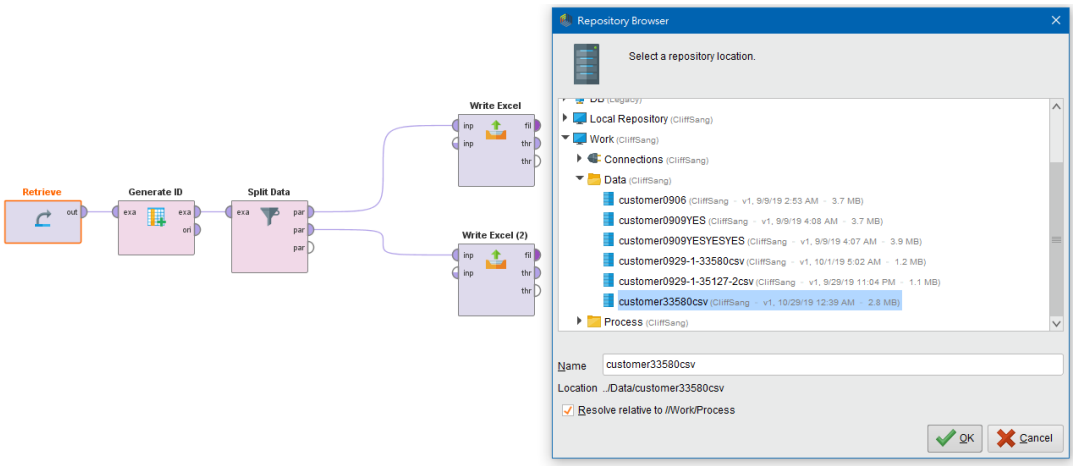

圖 4.15 分割資料流程圖(1)

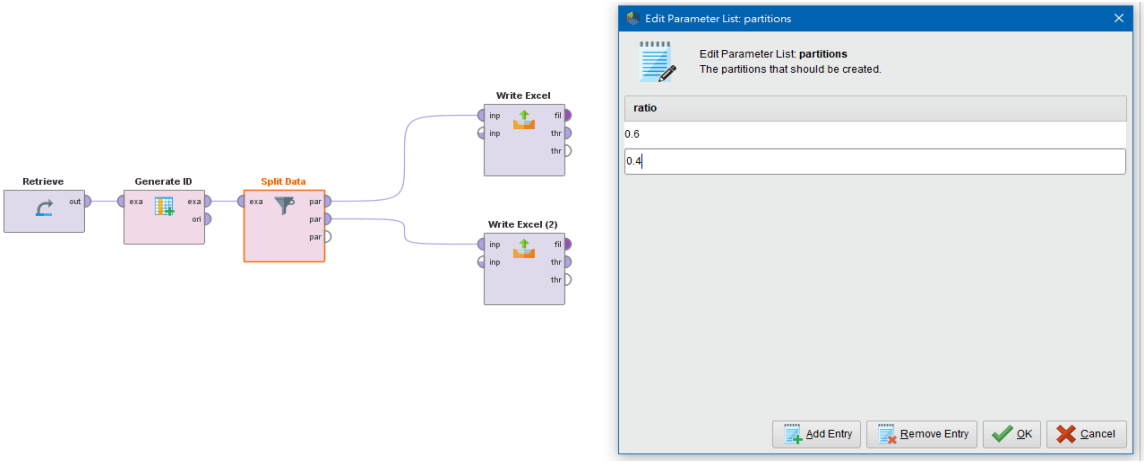

圖 4.16 分割資料流程圖(2)

表 4.2 次數統計對比表(1)

地區								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
臺北市	7479	22.27%	臺北市	4429	21.98%	臺北市	3050	22.71%
新北市	6034	17.97%	新北市	3646	18.10%	新北市	2388	17.78%
高雄市	4053	12.07%	高雄市	2401	11.92%	高雄市	1652	12.30%
臺中市	3895	11.60%	臺中市	2375	11.79%	臺中市	1520	11.32%
臺南市	2275	6.78%	臺南市	1383	6.86%	臺南市	892	6.64%
桃園市	2159	6.43%	桃園市	1302	6.46%	桃園市	857	6.38%
雲林縣	1200	3.57%	雲林縣	733	3.64%	雲林縣	467	3.48%
彰化縣	1004	2.99%	彰化縣	589	2.92%	彰化縣	415	3.09%
新竹市	792	2.36%	新竹市	482	2.39%	新竹市	310	2.31%
嘉義市	656	1.95%	嘉義市	399	1.98%	嘉義市	257	1.91%
屏東縣	621	1.85%	屏東縣	390	1.94%	基隆市	238	1.77%
基隆市	602	1.79%	基隆市	364	1.81%	嘉義縣	234	1.74%
嘉義縣	568	1.69%	嘉義縣	334	1.66%	屏東縣	231	1.72%
臺東縣	423	1.26%	新竹縣	243	1.21%	臺東縣	194	1.44%
新竹縣	388	1.16%	臺東縣	229	1.14%	宜蘭縣	161	1.20%
花蓮縣	364	1.08%	花蓮縣	216	1.07%	花蓮縣	148	1.10%
宜蘭縣	361	1.08%	宜蘭縣	200	0.99%	新竹縣	145	1.08%
南投縣	290	0.86%	南投縣	191	0.95%	南投縣	99	0.74%
苗栗縣	236	0.70%	苗栗縣	139	0.69%	苗栗縣	97	0.72%
澎湖縣	117	0.35%	澎湖縣	67	0.33%	澎湖縣	50	0.37%
金門縣	55	0.16%	金門縣	32	0.16%	金門縣	23	0.17%
連江縣	7	0.02%	連江縣	3	0.01%	連江縣	4	0.03%

表 4.3 次數統計對比表(2)

大宗特約戶								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	33438	99.58%	0	20065	99.59%	0	13373	99.56%
1	141	0.42%	1	82	0.41%	1	59	0.44%
永久戶								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	1091	3.25%	0	661	3.28%	0	430	3.20%
1	32488	96.75%	1	19486	96.72%	1	13002	96.80%
交寄方式								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
1	19824	59.04%	1	11875	58.94%	1	7949	59.18%
2	13755	40.96%	2	8272	41.06%	2	5483	40.82%
有部門別								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	33526	99.84%	0	20114	99.84%	0	13412	99.85%
1	53	0.16%	1	33	0.16%	1	20	0.15%
特殊特約客戶								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	33543	99.89%	0	20122	99.88%	0	13421	99.92%
1	36	0.11%	1	25	0.12%	1	11	0.08%
結帳方式								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	9993	29.76%	0	6047	30.01%	0	3946	29.38%
1	23586	70.24%	1	14100	69.99%	1	9486	70.62%
專案議價								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	28388	84.54%	0	16995	84.35%	0	11393	84.82%
1	5191	15.46%	1	3152	15.65%	1	2039	15.18%
混合交寄議價戶								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
0	33569	99.97%	0	20142	99.98%	0	13427	99.96%
1	10	0.03%	1	5	0.02%	1	5	0.04%

表 4.4 次數統計對比表(3)

產業別								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
1	2216	6.60%	1	1305	6.48%	1	911	6.78%
2	35	0.10%	2	24	0.12%	2	11	0.08%
3	230	0.68%	3	134	0.67%	3	96	0.71%
4	277	0.82%	4	158	0.78%	4	119	0.89%
5	442	1.32%	5	280	1.39%	5	162	1.21%
6	2378	7.08%	6	1420	7.05%	6	958	7.13%
7	4492	13.38%	7	2744	13.62%	7	1748	13.01%
8	611	1.82%	8	378	1.88%	8	233	1.73%
9	7647	22.77%	9	4515	22.41%	9	3132	23.32%
10	421	1.25%	10	268	1.33%	10	153	1.14%
11	1496	4.46%	11	917	4.55%	11	579	4.31%
12	390	1.16%	12	239	1.19%	12	151	1.12%
13	522	1.55%	13	320	1.59%	13	202	1.50%
14	266	0.79%	14	149	0.74%	14	117	0.87%
15	222	0.66%	15	133	0.66%	15	89	0.66%
99	11934	35.54%	99	7163	35.55%	99	4771	35.52%
客戶身分別								
全部資料			訓練資料 60%			測試資料 40%		
變數	次數	百分比	變數	次數	百分比	變數	次數	百分比
1	5341	15.91%	1	3182	15.79%	1	2159	16.07%
2	25885	77.09%	2	15543	77.15%	2	10342	77.00%
3	44	0.13%	3	23	0.11%	3	21	0.16%
4	2279	6.79%	4	1385	6.87%	4	894	6.66%
5	23	0.07%	5	10	0.05%	5	13	0.10%
6	7	0.02%	6	4	0.02%	6	3	0.02%

4.4.2 決策樹分析

(1)使用訓練資料建立預測模型

決策樹若分支越多，則容易產生越複雜的決策邊界，這就很容易導致決策樹過度適合，因此，需要將決策樹進行剪枝予以簡化[13]。由於時間因素，本專題並未針對決策樹過度適合的問題進行探討。圖 4.17 為以訓練資料進行 RapidMiner 決策樹分析所得之模型，模型顯示主要影響特約戶的的簽約為專業議價和混合交寄議價戶，其次則是各個地區著重的產業會對客戶的簽約產生影響。將結果帶入決策樹的判

斷公式進行計算得知(表 4.6)，正確率為 97.12%，精確率為 97.97%，召回率為 99.09%，表示訓練資料中有 97.12%被正確預測，97.97%的資料是正確的，實際資料中有 99.09%被正確預測出來。

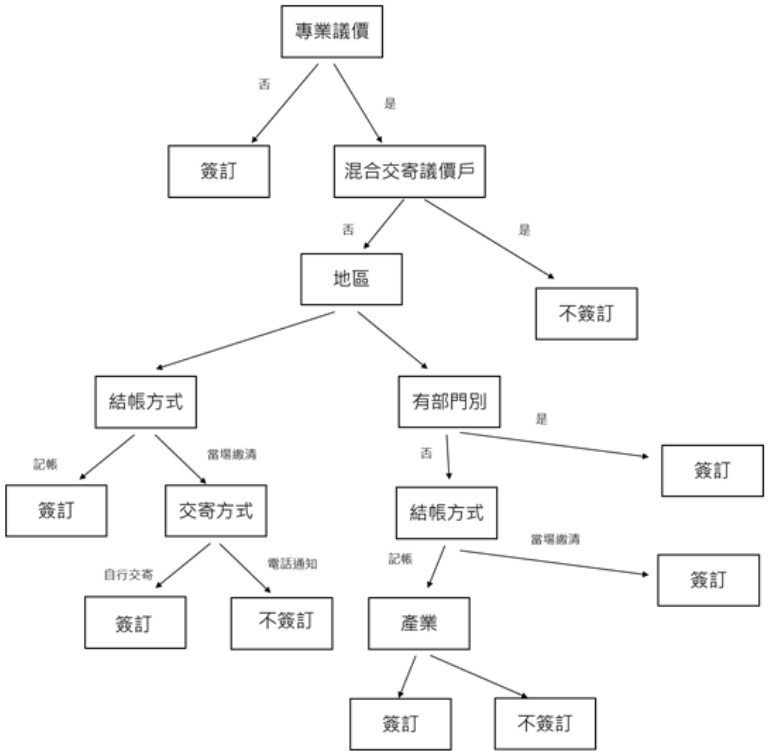

圖 4.17 決策樹預測模型

表 4.5 訓練資料預測結果

	實際值為 1	實際值為 0
預測值為 1	19316	403
預測值為 0	177	252

$$\text{正確率(Accuracy)} = \text{全部資料中有多少筆被正確預測}$$
$$= \frac{252 + 19316}{19316 + 177 + 403 + 252} \text{ x } 100\% = 97.12\%$$

$$\text{精確率(Precision)} = \text{被預測出的資料有多少是正確的}$$
$$= \frac{19316}{19316 + 403} \text{x } 100\% = 97.97\%$$

$$\text{召回率(Recall)} = \text{實際資料中有多少筆被正確預測出來}$$
$$= \frac{19316}{19316 + 177} \text{ x } 100\% = 99.09\%$$

(2)使用測試資料驗證預測模型

將前述(1)中所得模型結果帶入進行計算得知(表 4.7)，正確率為97.08%，精確率為 97.88%，召回率為 99.12%，表示測試資料中有97.08%被正確預測，97.88%的資料是正確的，實際資料中有 99.12%被正確預測出來；因此，可以得知該模型為最適模型。

表 4.6 測試資料預測結果

	實際值為 1	實際值為 0
預測值為 1	12881	278
預測值為 0	114	158

$$\text{正確率(Accuracy)} = \text{全部資料中有多少筆被正確預測}$$
$$= \frac{158 + 12881}{12881 + 114 + 278 + 158} \text{ x } 100\% = 97.08\%$$

$$\text{精確率(Precision)} = \text{被預測出的資料有多少是正確的}$$
$$= \frac{12881}{12881 + 278} \text{x } 100\% = 97.88\%$$

$$\text{召回率(Recall)} = \text{實際資料中有多少筆被正確預測出來}$$
$$= \frac{12881}{12881 + 114} \text{ x } 100\% = 99.12\%$$

第伍章 結論與建議

本專題研究結論與建議如下:

1. 將中華郵政各地的郵件量進行分析，可以得知六都之外，彰化縣為郵件量最多的縣市，公司行號和人口數也僅次於六都，彰化縣在一般縣是中有較大的發展潛力，外島地區則是快捷郵件居多，如金門縣在一般縣市中快捷郵件的總占量為 9.05%。因此，(彰化、新北)建議外島地區能夠對快捷郵件的部分進行推廣。

2. 經由中華郵政資料和政府資料進行交叉分析後得知，如臺中、新北地區的特約戶最大宗為零售業者，而官方資料中顯示，這兩個縣市以製造業為最大宗，故這兩個縣市的郵局可以推出諸如製造業優惠方案等，進而吸引更多的製造業者進行簽約，郵政所給的資料中顯示新北市特約戶占比最高的是零售業，但新北市的產業鏈中製造業的占比與零售業、金融業都高上許多，這代表新北市中製造業這一塊仍然有很大的潛在空間可以去挖掘新客戶。

3. 在預測特約戶之潛在客戶方面，發現是否成為特約戶主要受專業議價、混合交寄議價戶、和地區的影響，不同的產業也會影響特約戶的簽訂。因此，中華郵政可以對專案議價、混合交寄議價戶的特約戶進行業務加強，並針對每個地區所著重的產業進行業務推廣。

參考文獻

[1] 曾龍，「大數據與巨量資料分析」，科學發展，第 524 期，第 68 頁，2016 年 8 月。

[2] 黃昱霖、林豐正，「巨量資料視覺化之研究」，逢甲大學優質報告資料庫，2015 年 6 月。

[3] 張志偉、孫天龍，「數據資料視覺化應用於高齡健康促進活動成效評估與運動分析」，元智大學工業與工程管理學系，2016 年 7 月。

[4] 羅意智、楊朝棟，「運用 ELK Stack 於電子商務賣方價格分析與資料視覺化」，東海大學資訊工程學系，2018 年 6 月。

[5] 黃童宇、黃柏崴，「不懂程式也能學會的大數據分析術」，旗標，2019 年 2 月。

[6] 蘇裕傑、阮金聲、呂浚都，「應用資料探勘於淋巴癌病人存活預測之模式」，國立中正大學資訊管理系研究所，2011 年 8 月。

[7] 李語嫣、曾新穆、吳晉祥，「運用資料探勘技術由健康檢查與生活習慣資料建立疾病預測模型-以糖尿病為例」，國立成功大學醫學資訊研究所，2010 年 7 月。

[8] 許智翔、張良正，「應用決策樹與頻譜分析於含水層類別判識-以濁水溪沖積扇為例」，國立交通大學土木工程系所，2018 年 2 月 。

[9] 張瀚文、張良正，「應用決策樹於水庫防洪減淤 最佳操作規則之研究-以曾文水庫為例」，國立交通大學土木工程系所，2014 年 1 月 。

[10] PowerBI 官方網站
https://daxpowerbi.com/what-is-power-bi/

[11] SPSS 官方網站
https://www.ibm.com/tw-zh/analytics/spss-statistics-software

[12] RapidMiner 官方網站
https://rapidminer.com/

[13] 劉立民、吳建華，「Python 機器學習」，博碩，2019 年 6 月。