SAS/R 商業資料分析期末報告

電子商務：顧客購買意願分析及後續成效追蹤

指導教授：周珮婷 教授

組員：

105304021 統計四 陳人寧

105304028 統計四 方品謙

105304035 統計四 彭湘翎

105304046 統計四 游述宇

目錄

壹、動機與目的

由於近年網路普及，使消費者對於電子商務的使用量驟增，造就了無可限量的市場潛力。在過去實體零售中，銷售人員可以透過顧客購買時的反應及自身經歷，提供多種誘人的促銷方案；如今商業模式轉戰電商平台後，只能透過網路數據追蹤消費者的購買行為。

此報告希望能透過不同機器學習方法預測顧客最終是否完成交易，找出影響最終購買結果的重要變數，作為廣告投放、促銷方案參考及電商平台優化的依據。

貳、資料介紹、整理與變數選取

一、 資料介紹

在此份數據的預測變數為二元資料，顯示顧客有無完成交易，FALSE（0）為沒有完成交易，TRUE（1）為有完成交易。目的為分析有無完成交易之顧客與各變數間的關係，數據一共有 12,330 筆資料與 18 個變數。其中 10 個數值型變數與 8 個類別型變數。蒐集一年內不同顧客的消費紀錄，並且避免特殊假期，節日，或用戶個人慶祝活動（如生日等），在資料中有 84.5%（10,422 筆）為沒有完成交易資料，其餘 1908 筆為完成交易資料。下表為變數名稱及變數的定義。

二、 變數介紹

我們將資料中的變數依資料類型分為數值型與類別型兩類：

（一）數值型變數

變數名稱	變數定義
Administrative	帳戶管理（用戶瀏覽有關帳戶管理的頁數）
Administrative duration	帳戶管理時間長度（用戶瀏覽有關帳戶管理的秒數）
Informational	資訊搜尋（用戶瀏覽有關購物網站的網址、聯絡方式、地址等的頁數）
Informational duration	資訊搜尋時間長度（用戶瀏覽購物網站相關資訊的秒數）
Product related	產品搜尋（用戶瀏覽產品相關網頁的頁數）
Product related duration	產品搜尋時間長度（用戶瀏覽產品相關網頁的秒數）
Bounce rates	跳出率（用戶進入網站後只瀏覽了一個網頁就離開的訪客百分比）
Exit rates	離開率（用戶瀏覽的網頁為最後一頁的比例）
Page Values	頁面價值（用戶在完成交易之前訪問過的網頁帶來的平均價值）

Special day	特殊日子（用戶訪問頁面的時間距特殊日子的天數）

（二）類別型變數

變數名稱	變數定義
Operating Systems	用戶的操作系統（共 8 種）
Browser	用戶的瀏覽器（共 13 種）
Region	用戶使用網頁的地理區域（共 9 區）
Traffic type	用戶跳接至網站方式（橫幅廣告、SMS、直接等 20 種）
Visitor type	訪客類型（新用戶、舊用戶、其他）
Weekend	是否為週末（有以 TRUE 表示，無則 FALSE）
Month	瀏覽時的月份（2 月、3 月、5 至 12 月）
Revenue	是否完成交易（有以 TRUE 表示，無則 FALSE）

三、 資料觀察

首先對目標變數是否完成交易（Revenue）做敘述統計，圖（一）為 Revenue 分布情形，可以發現有購買的比例為 15.5%，無購買則為 84.5%。

接著，將所有變數對 Revenue 做圖，若變數為連續型，做直方圖，即附錄中的圖（七）至圖（十二）；若變數為離散型做長條圖，即附錄中的圖（十三）至圖（十六）；若變數為類別型，做馬賽克圖，即附錄中的圖（十七）至圖（二十三）。

由圖（七）至圖（九）和圖（十三）至圖（十五），我們觀察到是否完成交易和瀏覽頁數和瀏覽時間長度有明顯分布差異，有完成交易的直方圖和長條圖較為右偏，表示各類型瀏覽頁數及時間長度和是否完成交易之間有某種程度的關聯。

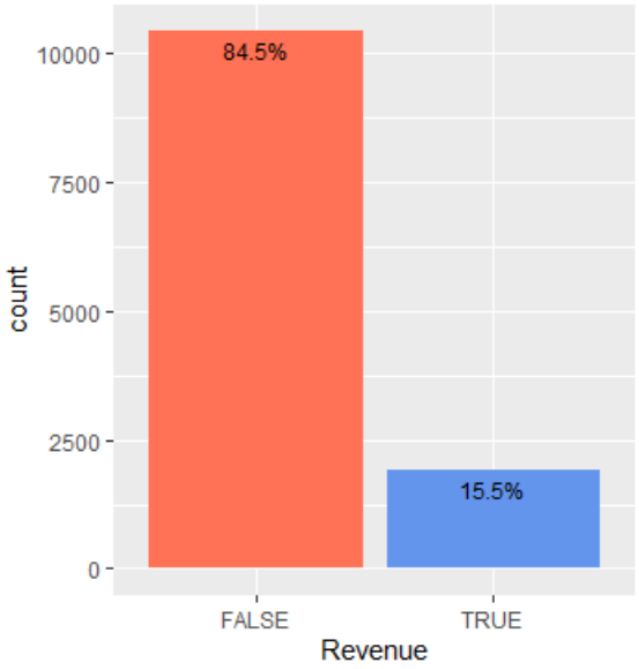
由圖（十一）至圖（十二），我們可以觀察到頁面的離開率以及頁面價值和是否完成交易有較明顯的分布差異，在有完成交易的直方圖中，頁面的離開率較為左偏，而頁面價值中為右偏，表示頁面的離開率以及頁面價值和是否完成交易之間有某種程度的關聯。

由圖（二）可觀察到變數「Browser」以第 12 及 13 種完成交易的比例最高，分別為 30%、26.2%。

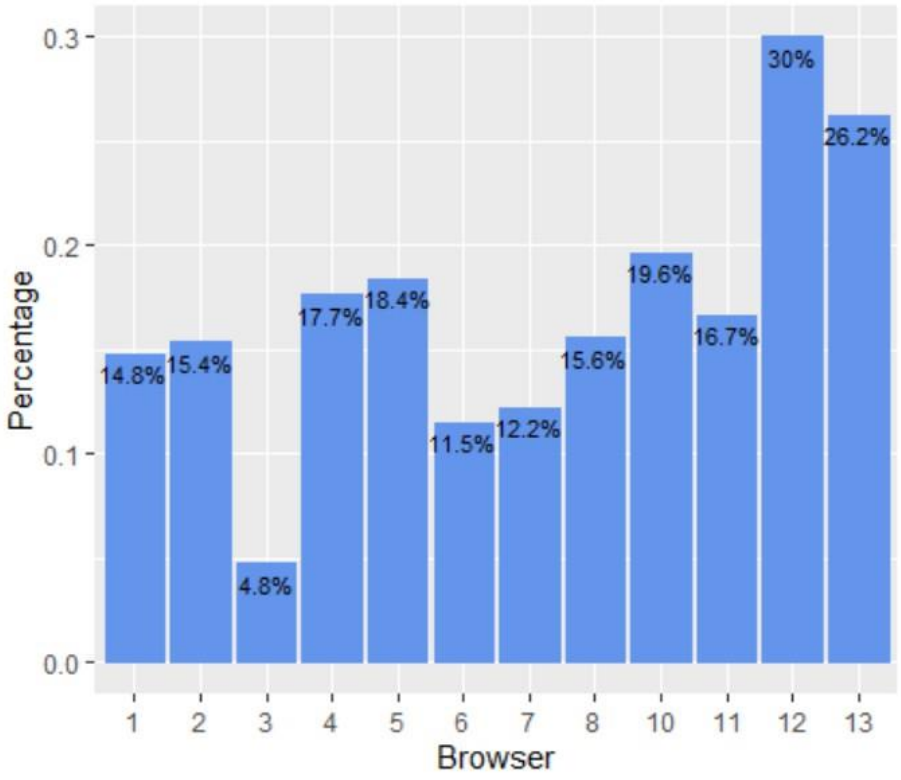
由圖（三），可觀察到變數「Visitor Type」中，新用戶完成的交易比例較高為 24.9%，而舊用戶比例最低為 13.9%。

由圖（四）可觀察到變數「Month」以 5 月、11 月、3 月為消費者使用電商平台的高峰，而由圖（四）可發現完成交易比例較高的月份為 11 月、10 月、9 月，分別為 25.4%、20.9%、19.2%。

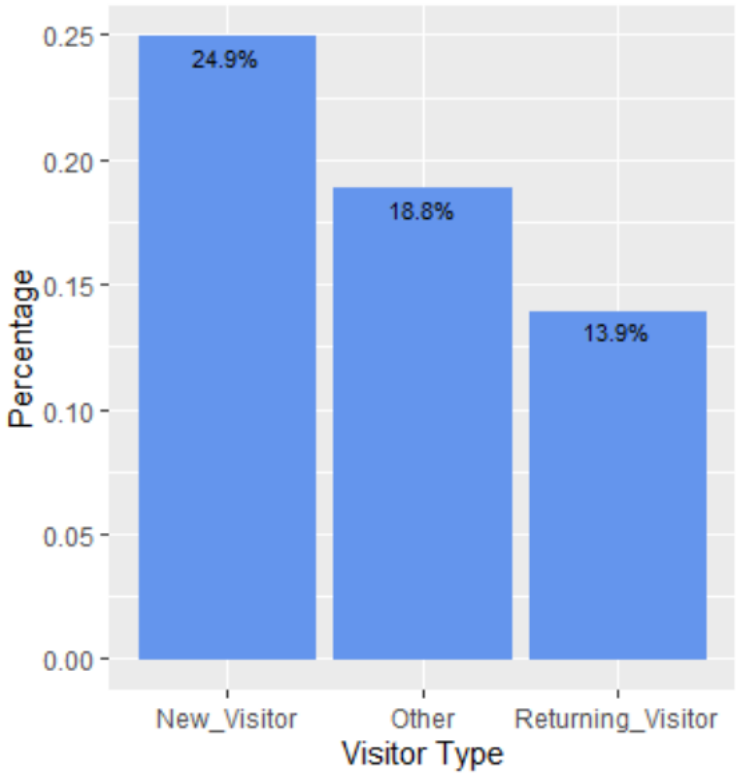
另外，在圖（五）中，根據相關係數可以看出該筆資料中是否完成交易與離開率為負相關；而與頁面價值則是正相關。



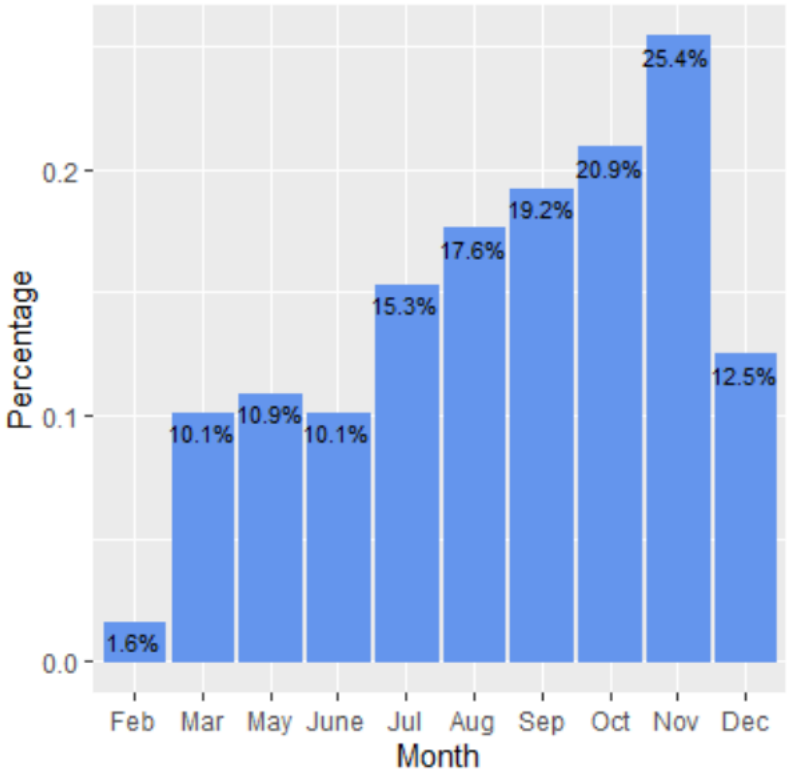
圖（一）：是否完成交易的比例



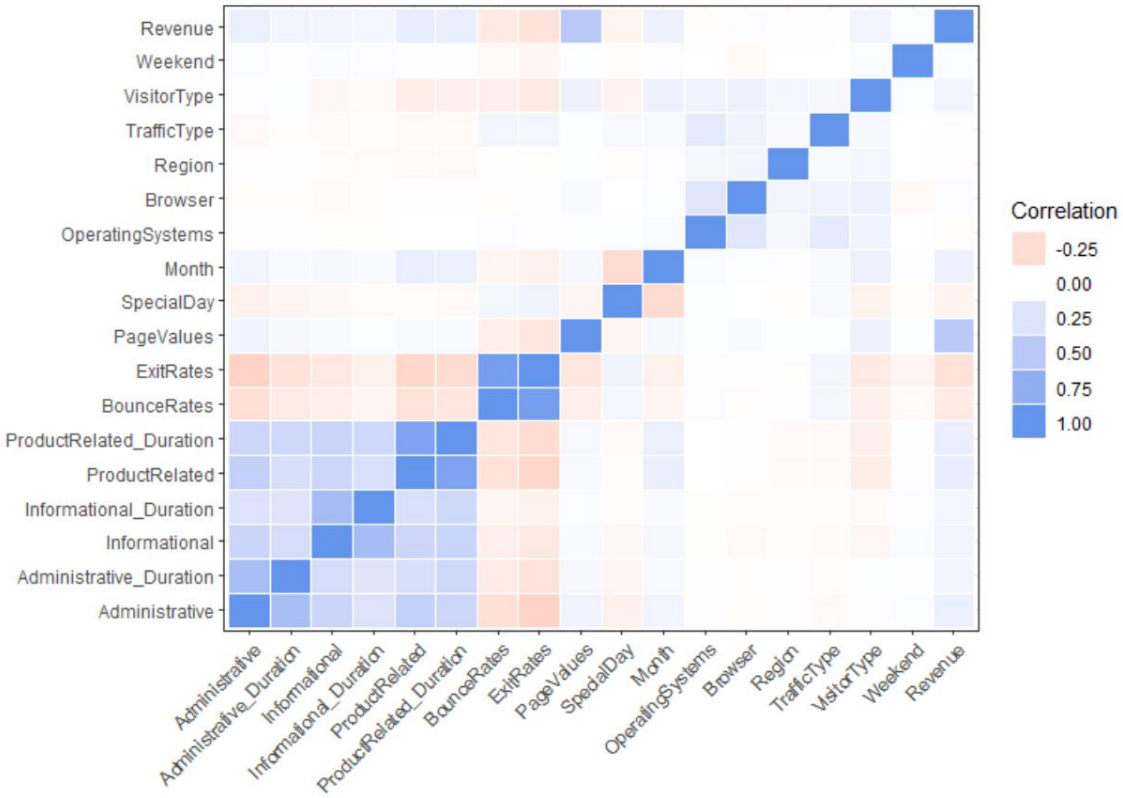
圖（二）：使用各瀏覽器完成交易的比例長條圖



圖（三）：各類型訪客完成交易的比例長條圖



圖（四）：各月份完成交易的比例長條圖



圖（五）：相關係數熱圖

參、建模與分析

採羅吉斯迴歸、支援向量機（SVM）、決策樹以及隨機森林四種監督式機器學習模型訓練模型，並利用十折交叉驗證（10-fold cross-validation，簡稱 10-fold CV）的方式評量，切割資料為十等分，每次取一份當成測試集、一份為驗證集，其餘全作為訓練集，共進行十次的預測並求得平均值。

由於此資料預測值是否完成交易（Revenue）的比例約為 2:11，為了避免預測產生誤差，將進行資料平衡。

- 資料平衡：

  由於是否完成交易的比例較為懸殊，因此先處理資料不平衡的問題，首先利用 Clara 演算法分群，該演算法是分群演算法的一種，使用 PAM 算法為特定大小的一小部分數據樣本生成最佳的中位數，並以每個觀測值和離他最近的中位數之間的平均距離來衡量該分群的優劣，保留有最小距離的子資料並重複上述步驟直到子資料達到樣本量為止。接著在各群集中生成或壓縮相等比例的資料，即可在保留資料特徵的情況下，將未完成交易的資料壓縮為 0.9 倍，並將完成交易的資料放大為 5 倍，使得是否完成交易（Revenue）調整成 1:1。

- 評估指標：

  評估模型指標除了準確率（Accuracy）之外，也將計算精準率（Precision）、召回率（Recall）及 F1 Score，得到較客觀的成果。

一、羅吉斯迴歸（Logistic Regression）

1. 模型一：全模型（Full Model）

首先使用全部變數進行羅吉斯迴歸，並利用 10-fold CV 交叉驗證，模型的效能指標如表（一）。

2. 模型二：最終模型

透過逐步迴歸的方式進行變數篩選，將不顯著的變數踢除，最後挑選的變數依資料型態分為數值型變數：「Administrative」、「Region」、「Informational」、「Product Related」、「Bounce Rates」、「Exit Rates」、「Page Values」、「Operating Systems」、「Browser」及類別型變數：「Visitor Type」、「Month」，共 11 個變數。

模型建置完成後進行共線性檢查，發現所有變數的 VIF 值皆落在 1 到 3 之間，以經驗法則而言 VIF 值小於 4，模型便不易受到線性干擾，由此判斷所選變數間不存在共線性。模型之效能指標如表（二）所示。

3. 模型比較

由表（一）及表（二）可發現變數篩選後，模型二之準確率與召回率皆有小幅度提升，而精準率僅下降 0.0031，由此可知經過變數篩選後的預測結果較為優異。

表（一）：羅吉斯迴歸 模型一效能指標

Accuracy	Precision	Recall	F1 Score
0.8255	0.821	0.8399	0.8294

表（二）：羅吉斯迴歸 模型二效能指標

Accuracy	Precision	Recall	F1 Score
0.8262	0.8179	0.8452	0.8303

二、支援向量機（Support Vector Machine，SVM）

1. 模型一：全模型（Full Model）

將訓練集放入支援向量機後，得模型效能指標如表（三）。

2. 模型二：調參後最終模型

利用 mlbench package 的 tune function 進行參數調整，找出最佳的 Cost 與 gamma 值，其值分別為 10 及 2。將參數放入 SVM 模型得以效能指標表（四）。

3. 模型比較

由表（三）及表（四）可發現調整參數後，模型二之各項效能指標皆有大幅度提升，由此可知經過變數篩選後的預測結果較為優異。

表（三）：支援向量機 模型一效能指標

Accuracy	Precision	Recall	F1 Score
0.8519	0.8411	0.872	0.8558

表（四）：支援向量機 模型二效能指標

Accuracy	Precision	Recall	F1 Score
0.9889	0.9927	0.9852	0.989

三、決策樹（Decision Tree）

將運算時間 cp 訂在 1e-4 得到最佳結果如下表（五），決策樹的分類圖如附錄圖（二十四）。

表（五）：決策樹 效能指標

Accuracy	Precision	Recall	F1 Score
0.8975	0.8744	0.9305	0.9014

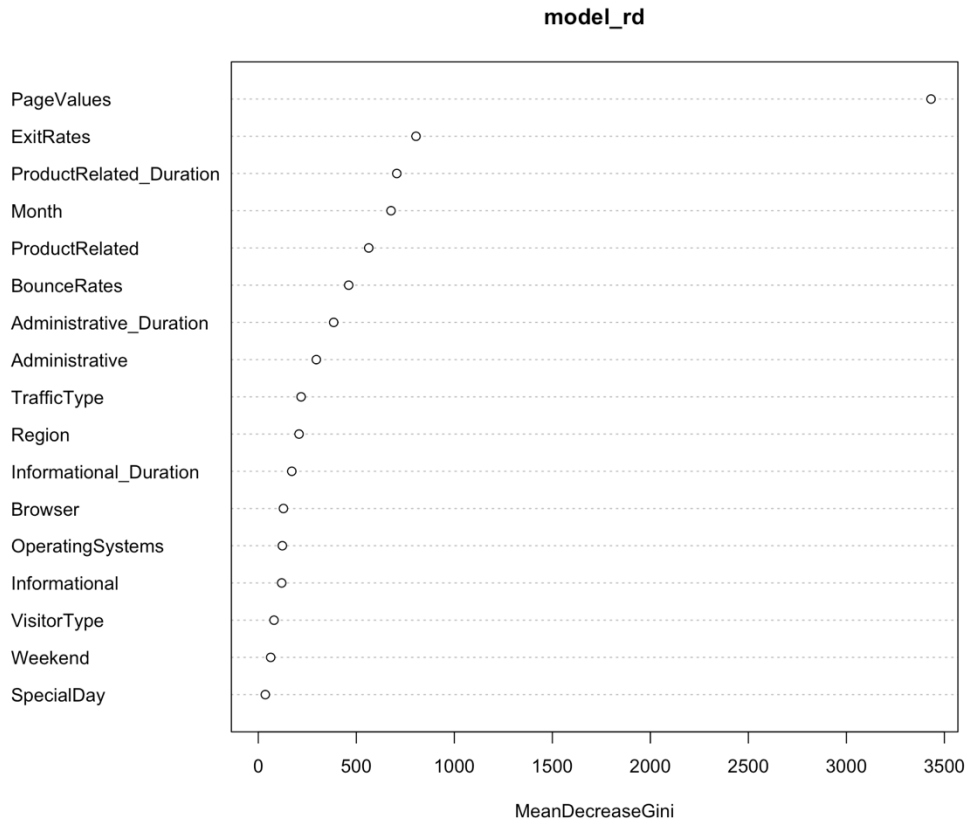
四、隨機森林（Random Forest）

1. 檢視參數：

我們將各個變數抽出進行隨機森林模型建置，並觀察抽出哪個變數會造成的誤差最多，意即對於整體模型的影響力最大，影響力大小如圖（六）所示。根據 Mean Decrease Accuracy 影響力大小，我們決定選擇「Page Values」、「Exit Rates」、「Product Related Duration」、「Month」、「Product Related」、「Bounce Rates」、「Administrative Duration」、「Administrative」「Traffic Type」、「Region」、「Informational Duration」、「Browser」、共 12 個變數來進行最終隨機森林模型的建置。

2. 模型建置：

將篩選後的變數放入隨機森林，並進行交叉驗證，經過以 500 倍數為基礎的樹數量測試後，發現分裂 3,500 棵決策樹可得到最好的預測，再增加或減少數量所得出的效果都不如 3,500 棵決策樹理想。以 3,500 棵樹建置模型之效能指標如表（六）。



圖（六）：隨機森林重要變數

表（六）：隨機森林 效能指標

Accuracy	Precision	Recall	F1 Score
0.9637	0.9363	0.9958	0.9651

肆、結論

消費者對於電子商務的大量使用造就了龐大的市場，本研究透過找出重要變數、分析過去顧客購買的行為以預測顧客是否會完成交易。藉由各種分析電商消費者的行為，可以了解顧客和電商的互動模式，以制定更加吸引消費者購買的策略。

根據模型預測，各個模型的衡量效能指標如表（七），由表（七）中的 F1 Score 以支援向量機 SVM 0.989 的預測效果最好。因為有先處理資料不平衡與觀察變數間的多元共線性問題，所以透過 SVM 得到最佳的預測效果也在可以預期的範圍，至於隨機森林較為擅長處理資料不平衡的狀態，因此在未來預測新資料時，此模型中仍有一定參考價值。在該模型中「Page Values」、「Exit Rates」、「Product Related Duration」、「Month」、「Product Related」五個變數最有影響力，而其中「Page Values」、

「Product Related」兩個變數分別代表「用戶在完成交易前訪問過的網頁帶來的價值」、「瀏覽商品的頁數」，因此該二變數無論於直觀或以模型而言，皆對該資料中客人「是否購買」有高度解釋力。

在變數「Month」中，完成交易最多的月份依序為 9、10、11 月，本研究推測為節慶所致，例如 11 月為感恩節。電商業者可在特殊節慶期間推出行銷活動吸引消費者，增加買氣，或是利用本身平台週年慶制定促銷方案吸引流量。藉由變數「Exit rates」，可以獲得每個頁面的品質，若過高代表頁面沒有引起讀者的興趣，由此證明網站介面的美觀性會直接影響消費者購買意願，因此電商業者可根據依此對內容與品質進行改善，以增加顧客繼續瀏覽的意願。

由於本研究發現，跳出率（Bounce rates）並不是消費者最終是否購買的重要變數之一，這也表示目前廣告投放的族群並不能真正觸及到有購買慾望的顧客。因此未來預期能將跳出率作為精準行銷成效的衡量指標，使跳出率與消費者最終是否購買有關聯，以確保廣告投放的效益。

綜上所述，商家可以透過本研究了解消費者對於瀏覽電子商務頁面的行為模式以及預測會進行購買的客戶，並且根據分析的結果增進廣告投放的效率、改善頁面對於客戶的吸引力與行銷策略的制定。

表（七）：各模型衡量效能指標

效能指標 / 模型	Accuracy	Precision	Recall	F1 Score
羅吉斯迴歸	0.8262	0.8179	0.8452	0.8303
SVM	0.9889	0.9927	0.9852	0.989
決策樹	0.8975	0.8744	0.9305	0.9014
隨機森林	0.9637	0.9363	0.9958	0.9651

伍、問題延伸與討論

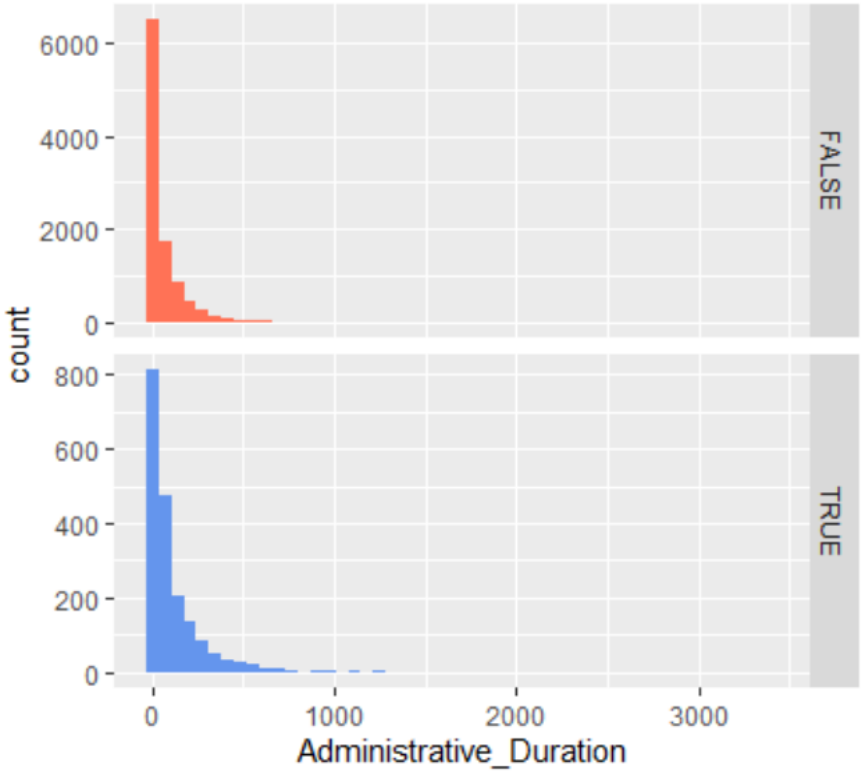
由上述模型的建立，能夠獲得消費者是否完成交易的預測結果，以利電商業者制定行銷策略，然此資料所蒐集的變數大多為電商端的網頁資訊，無法進一步分析消費者行為並追蹤客戶，未來希望取得更多客戶端資料（如：年紀，性別等），了解顧客基本面資料與購物習性，做出客製化的推薦系統，達到精準行銷。此外，若未來能取得後續的線上消費者購物資料，便能加入 A/B test，分析介面更動與客戶是否完成交易間的關係，提供電商平台檢視成效並衡量頁面質量。

陸、附錄

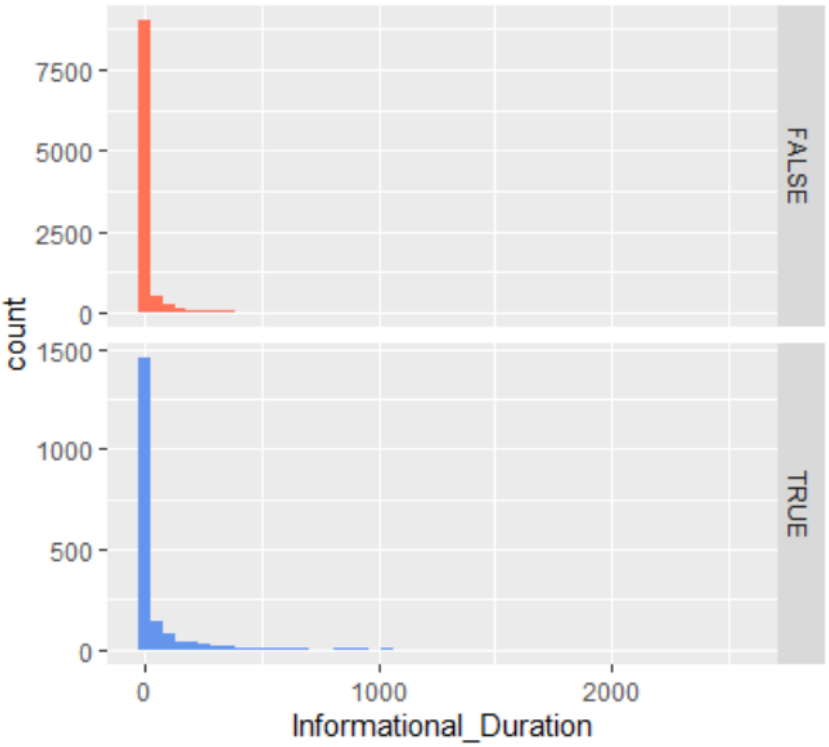
一、原始資料

UCI Machine learning repository - Online Shoppers Purchasing Intention Dataset Data Set (https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#)

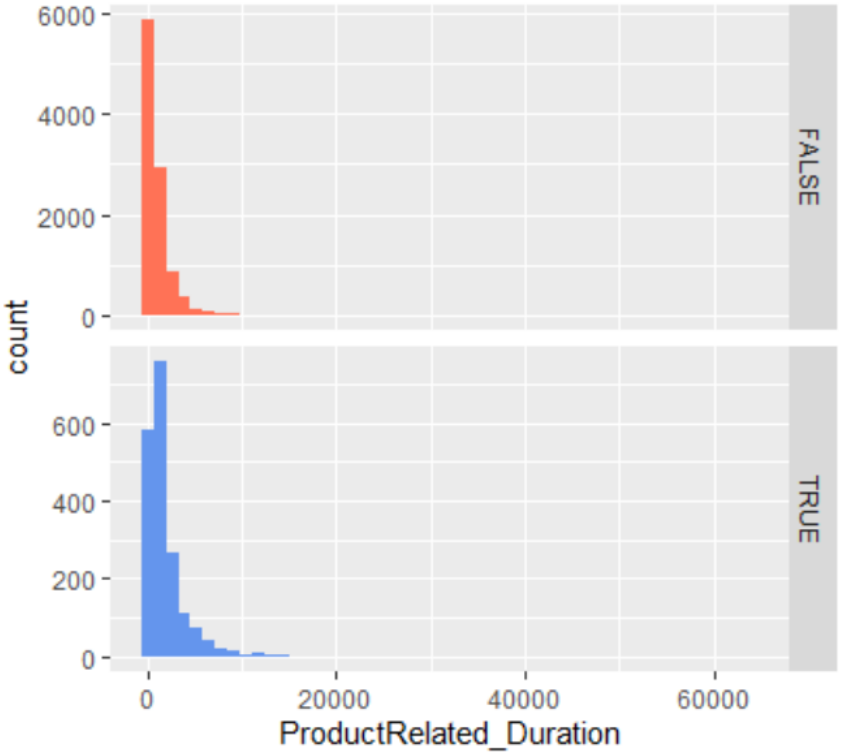
二、連續型變數與 Revenue 的直方圖



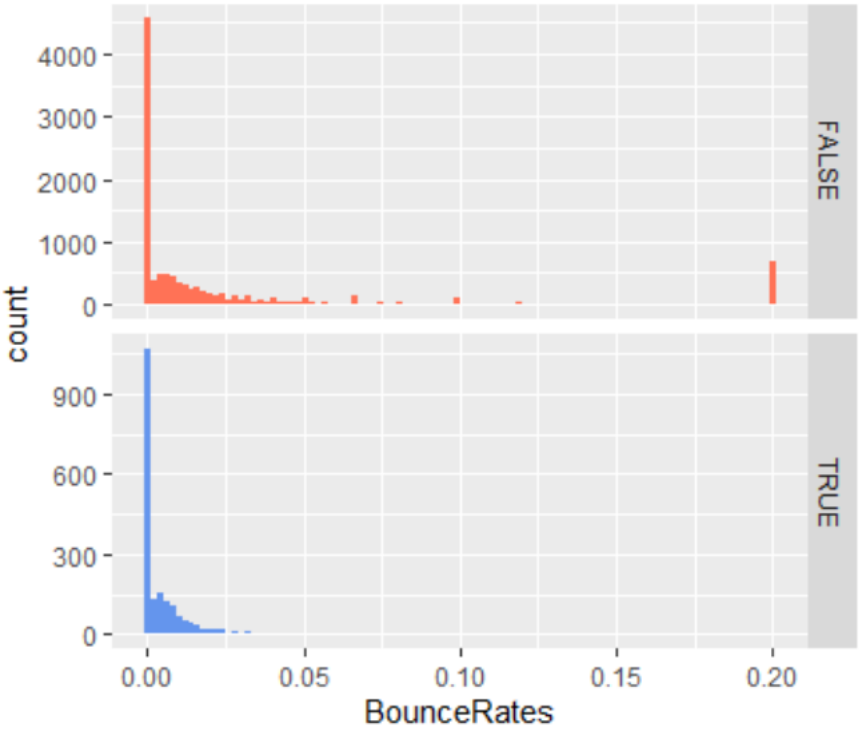
圖（七）：Administrative Duration 與 Revenue 的直方圖



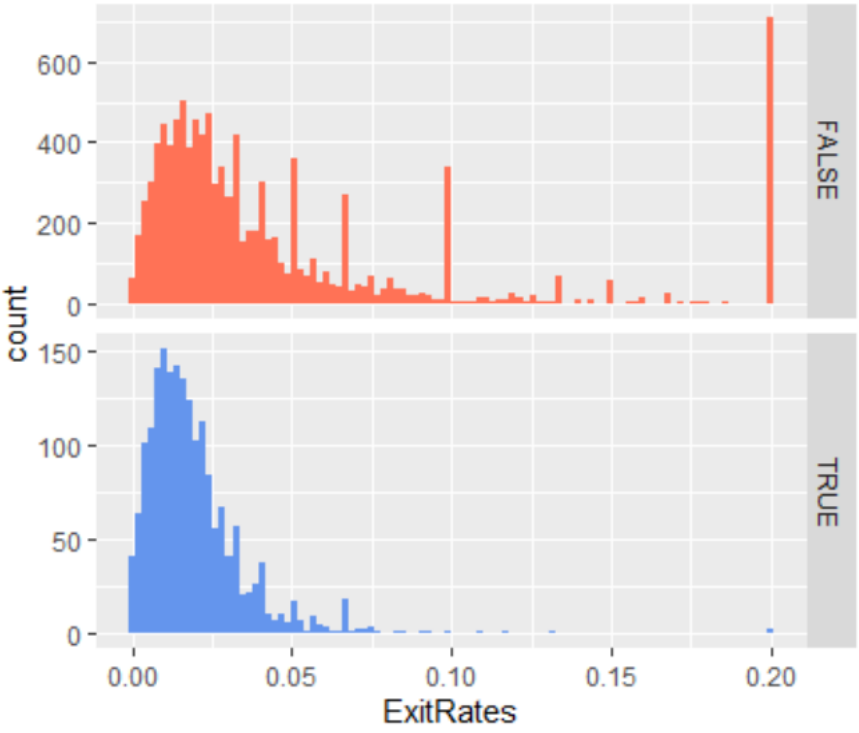
圖（八）：Informational duration 與 Revenue 的直方圖



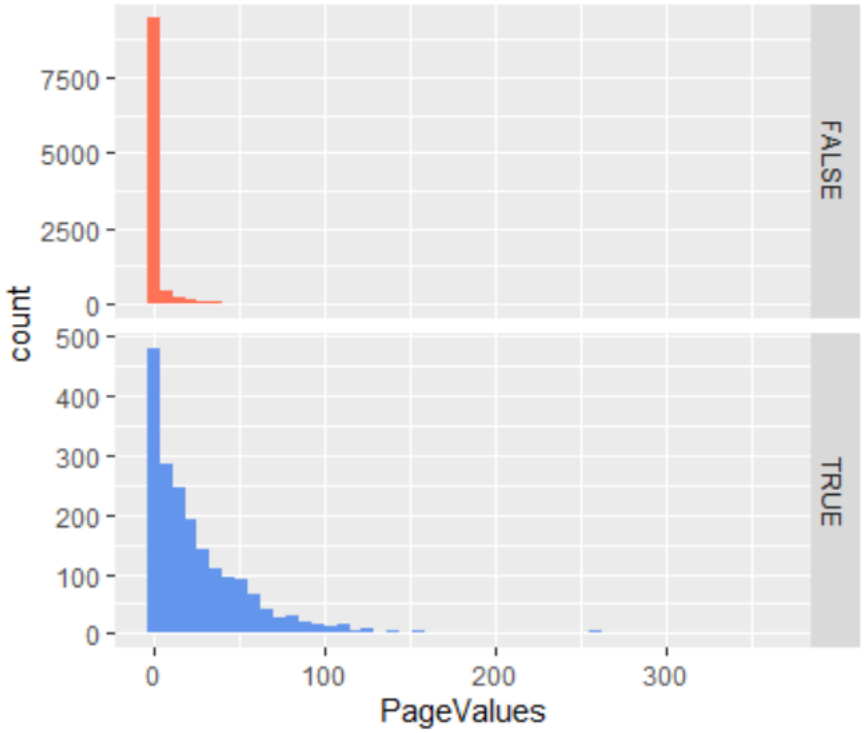
圖（九）：Product related duration 與 Revenue 的直方圖



圖（十）：Bounce Rates 與 Revenue 的直方圖

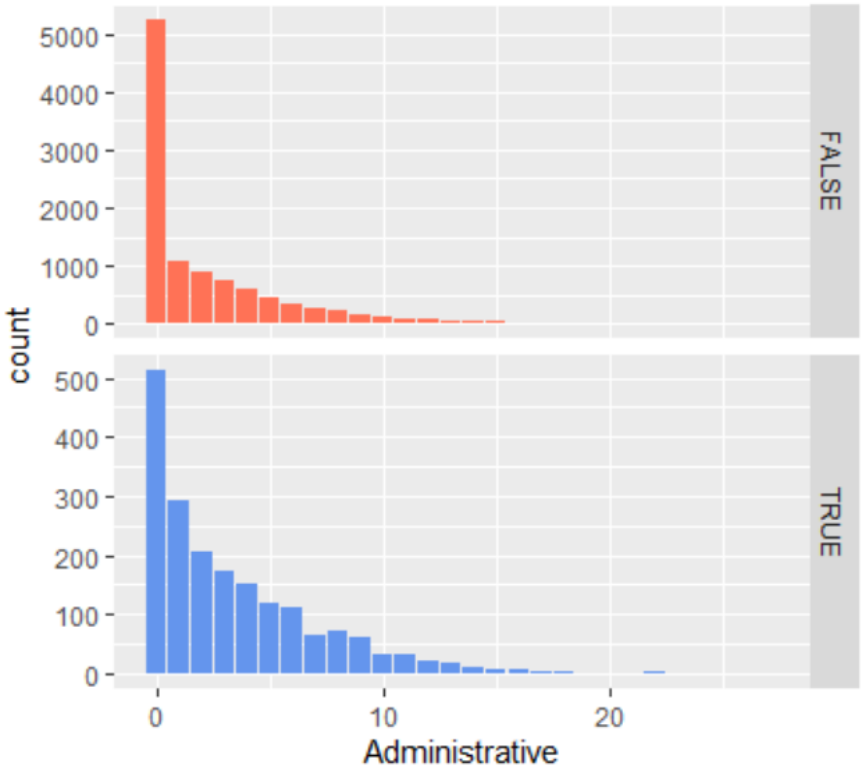


圖（十一）：Exit Rates 與 Revenue 的直方圖

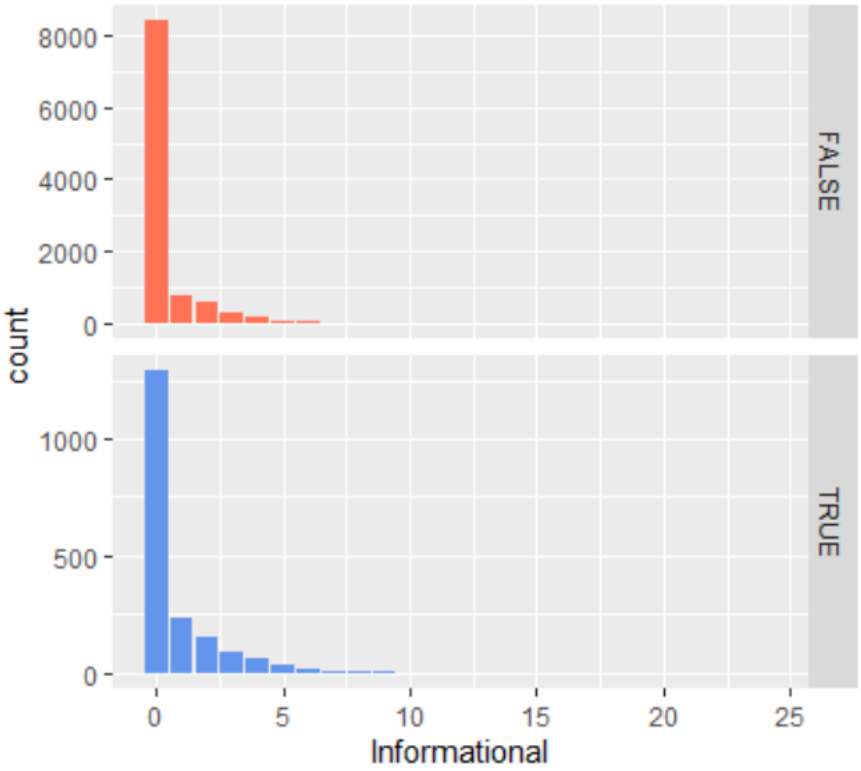


圖（十二）：Page Values 與 Revenue 的直方圖

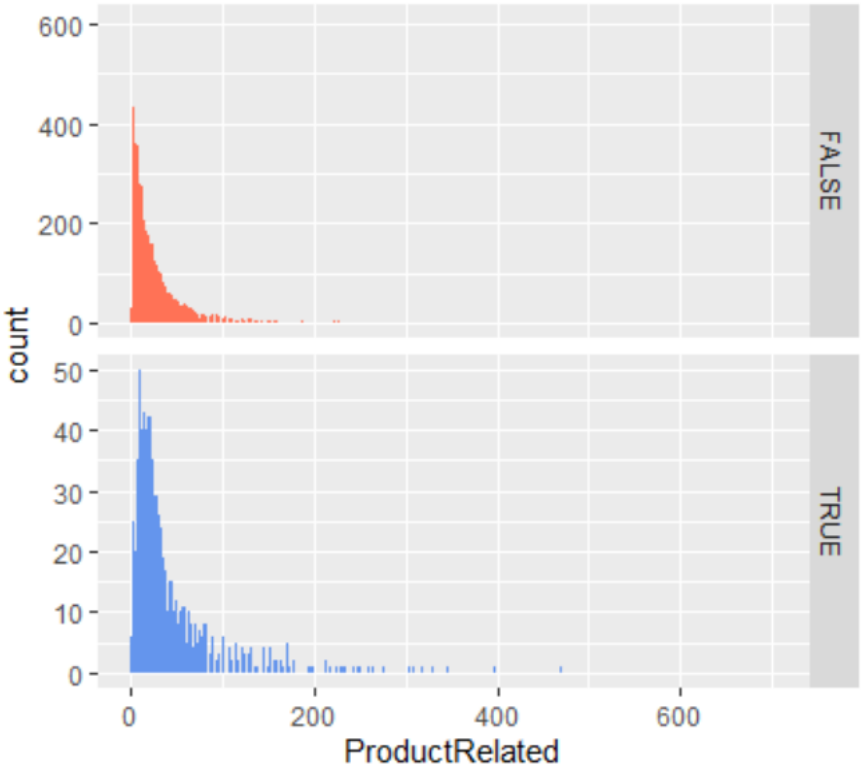
三、離散型變數與 Revenue 的長條圖



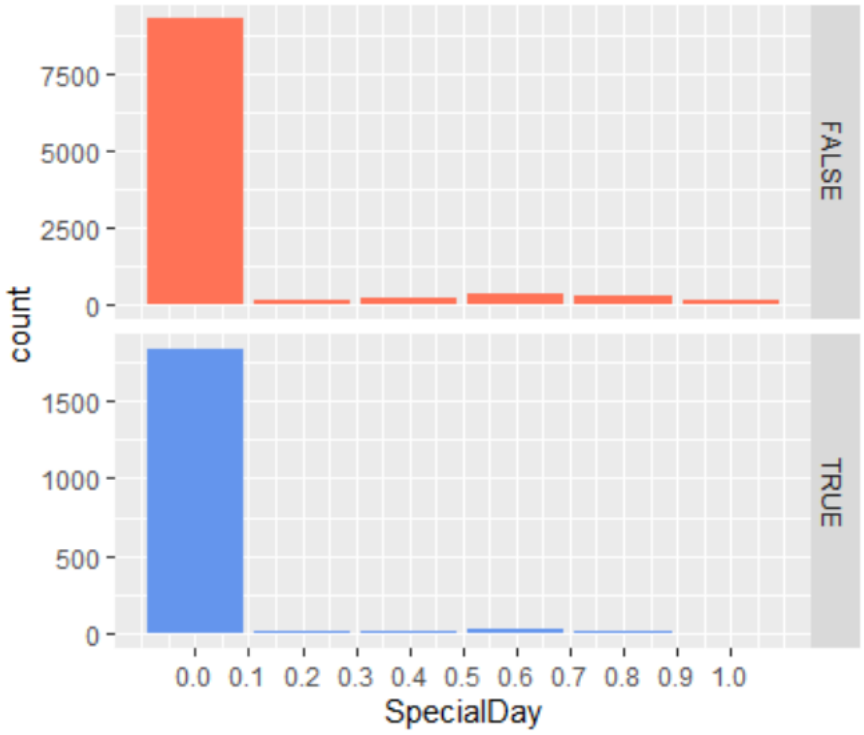
圖（十三）：Administrative 與 Revenue 的長條圖



圖（十四）：Informational 與 Revenue 的長條圖



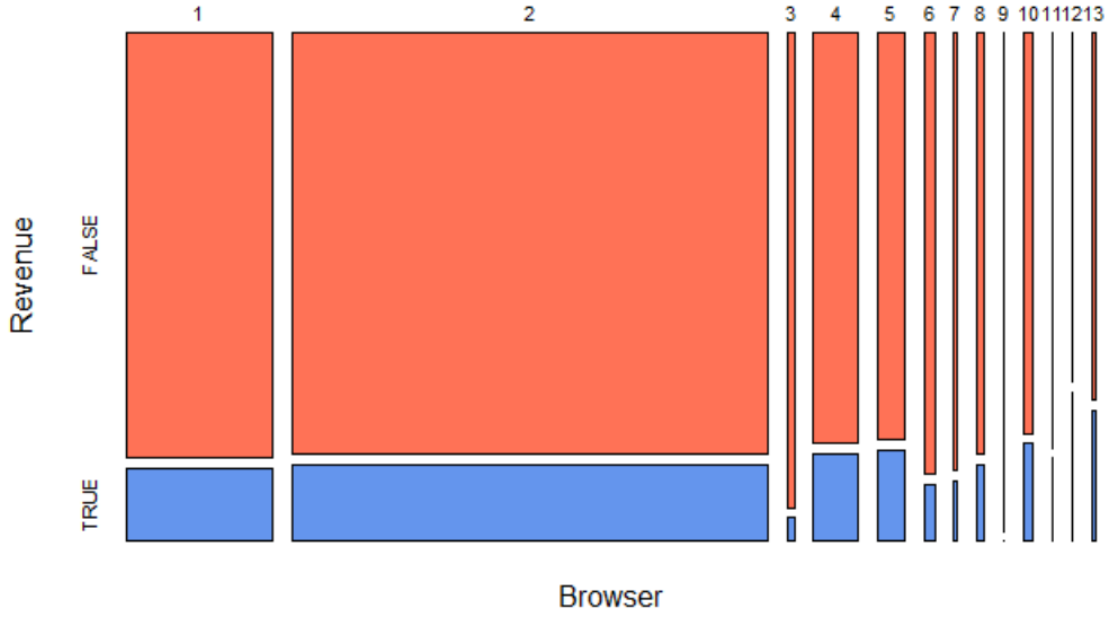
圖（十五）：Product related 與 Revenue 的長條圖



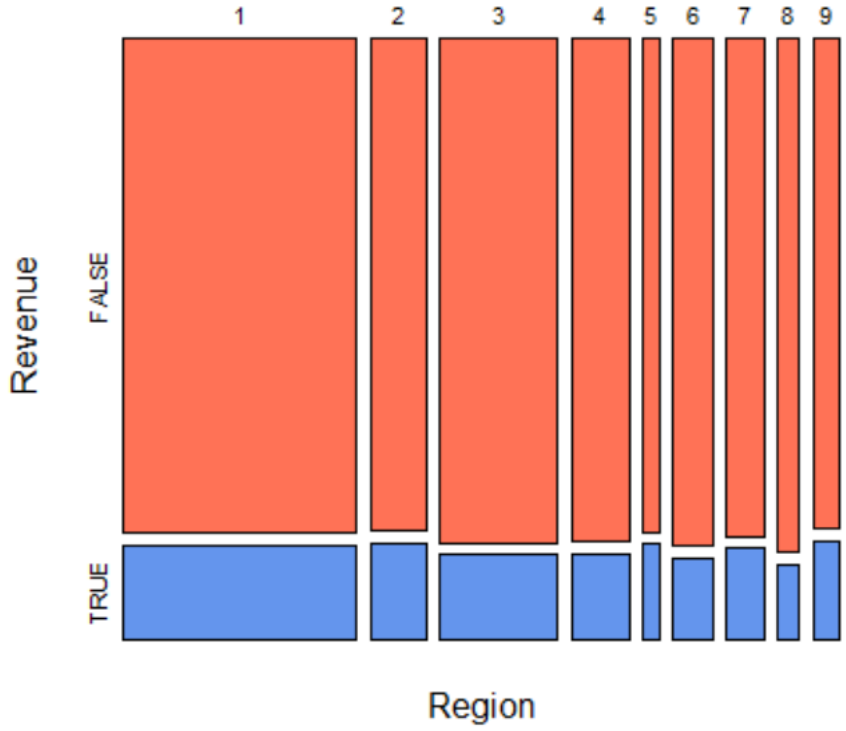
圖（十六）：Special Day 與 Revenue 的長條圖

四、類別型變數與 Revenue 的馬賽克圖

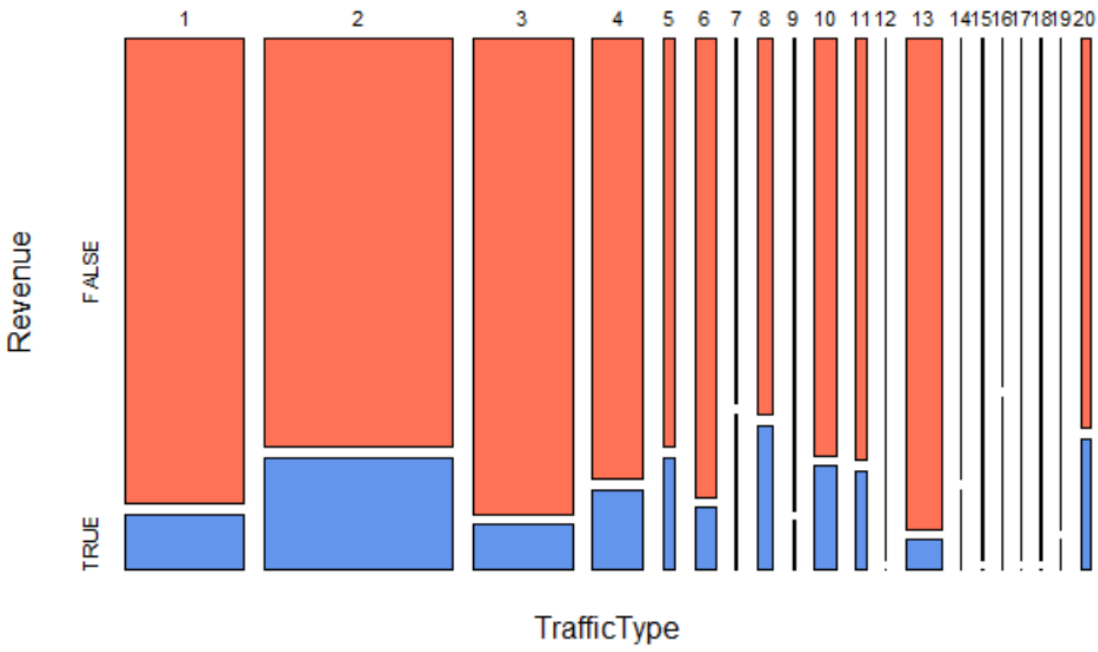
圖（十七）：Operating System 與 Revenue 的馬賽克圖



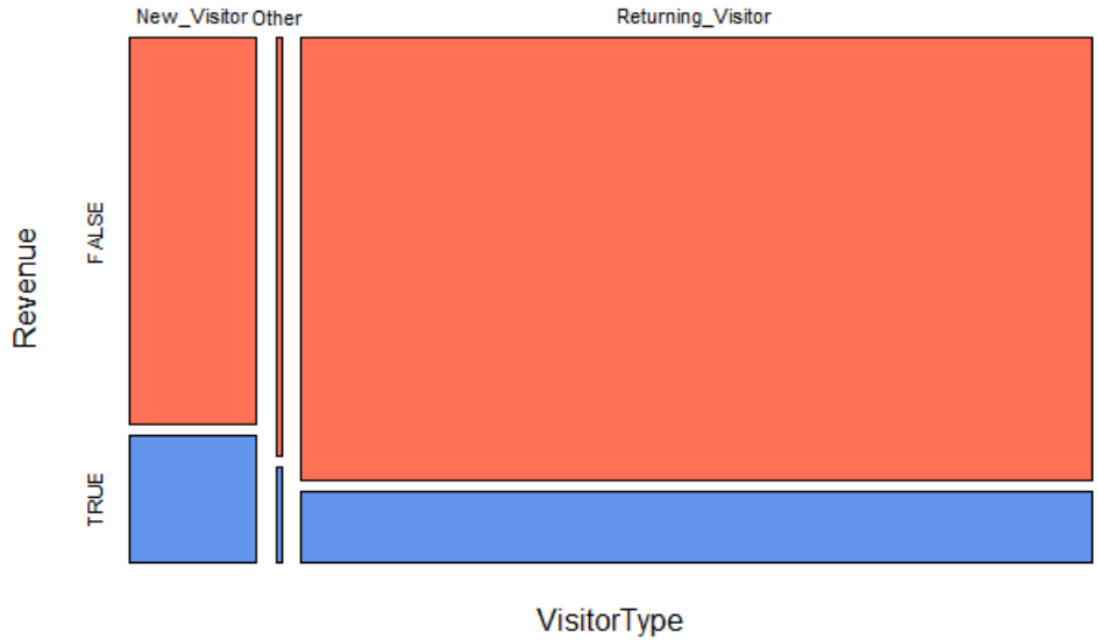
圖（十八）：Browser 與 Revenue 的馬賽克圖



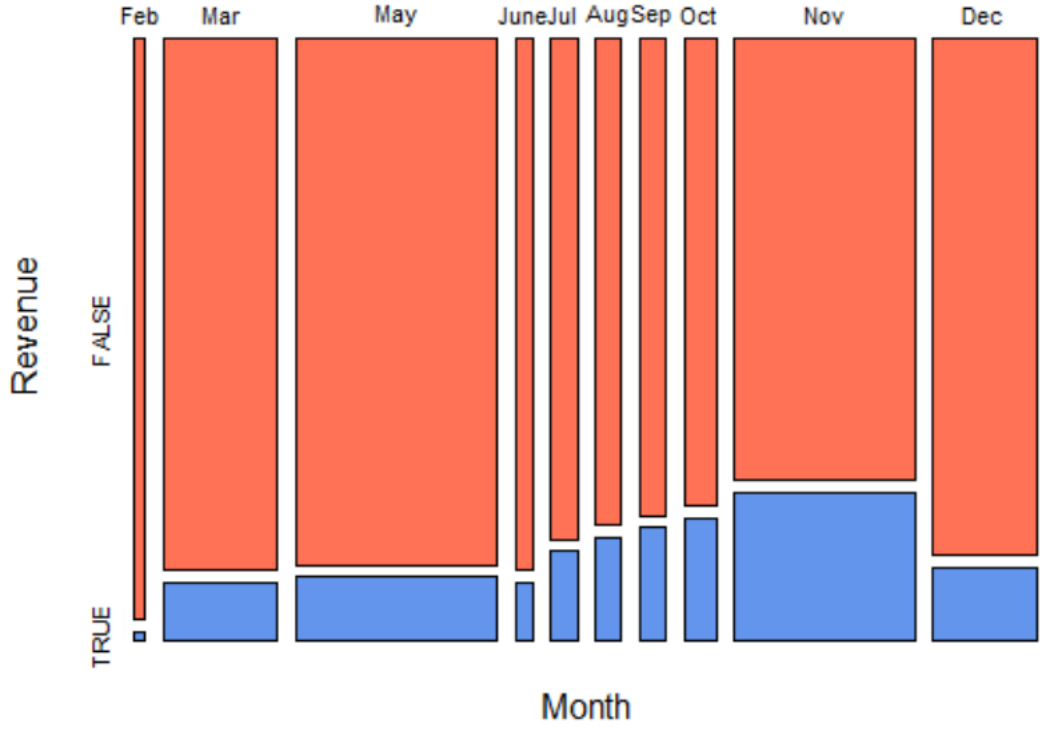
圖（十九）：Region 與 Revenue 的馬賽克圖



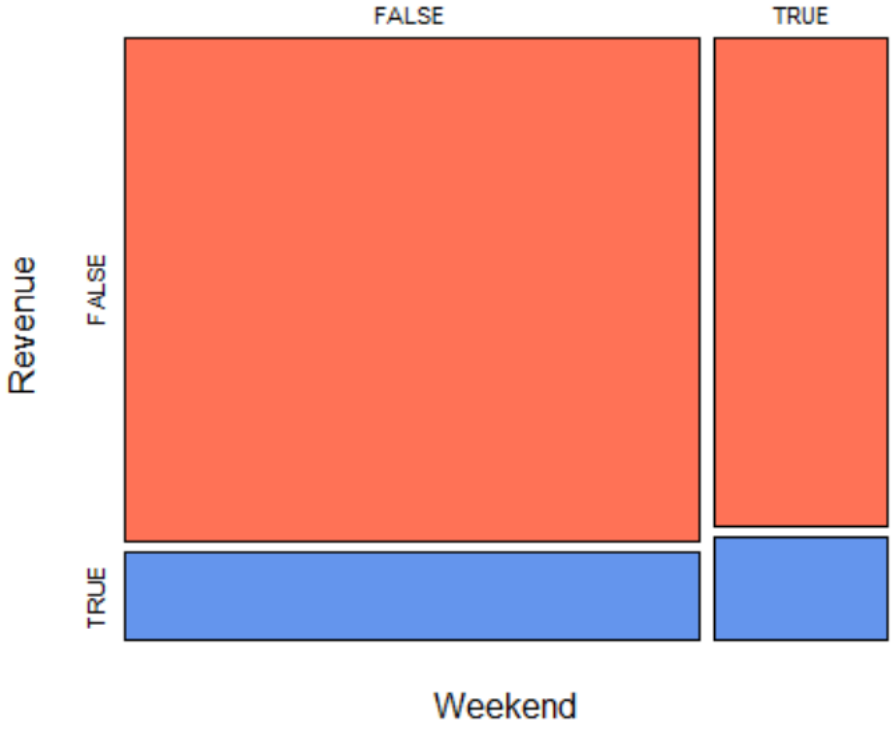
圖（二十）：Traffic Type 與 Revenue 的馬賽克圖



圖（二十一）：Visitor Type 與 Revenue 的馬賽克圖

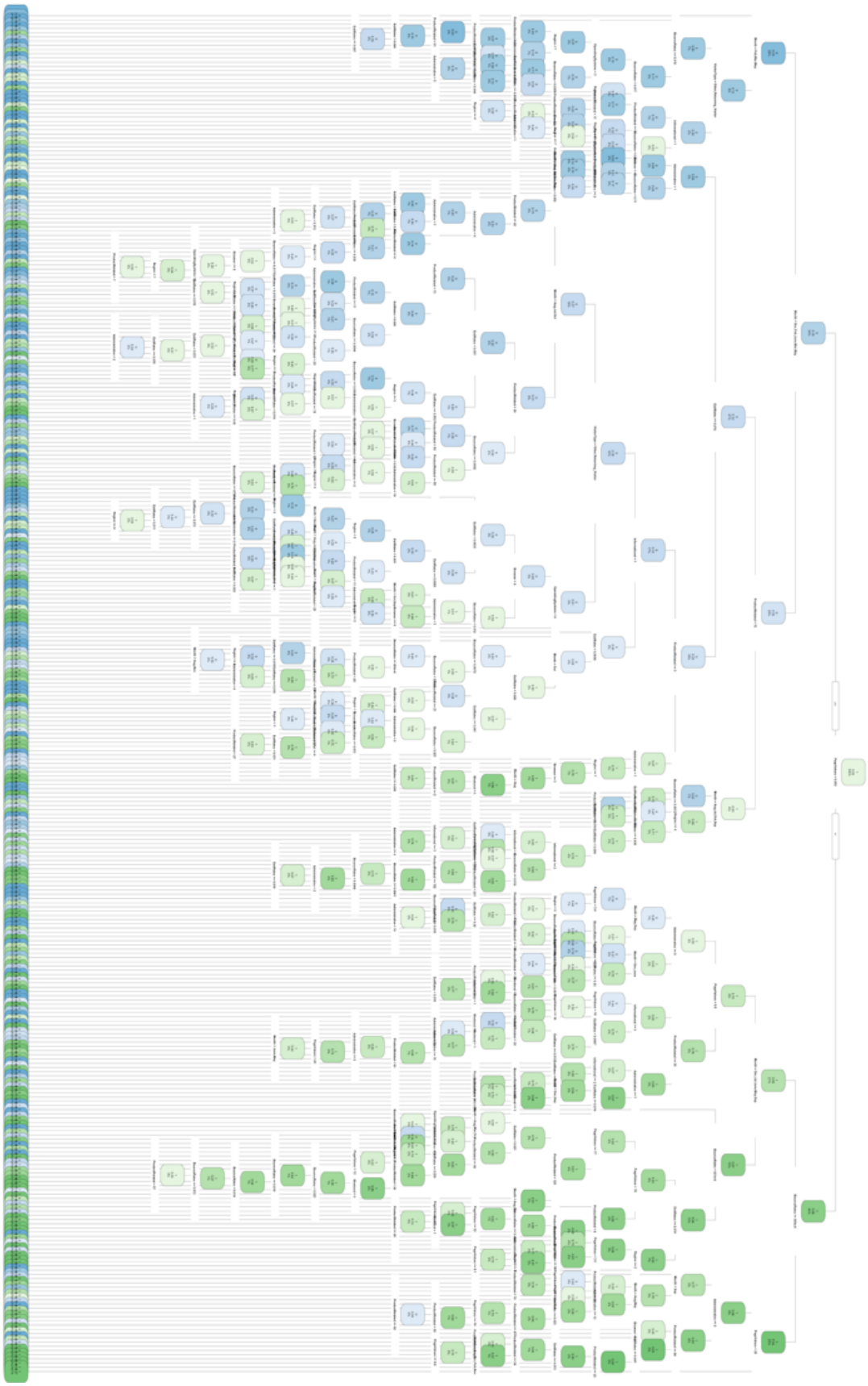


圖（二十二）：Month 與 Revenue 的馬賽克圖



圖（二十三）：Weekend 與 Revenue 的馬賽克圖

五、決策樹分類圖



圖（二十四）：決策樹分類圖

六、各模型的預測結果（混淆矩陣）

表（八）：羅吉斯迴歸模型混淆矩陣

預測\實際	False	True
False	1484	278
True	376	1646

表（九）：SVM 模型混淆矩陣

預測\實際	False	True
False	929	19
True	9	935

表（十）：決策樹模型混淆矩陣

預測\實際	False	True
False	824	78
True	114	876

表（十一）：隨機森林模型混淆矩陣

預測\實際	False	True
False	884	9
True	54	945

七、參考資料

（一）Data Visualisation with R: 100 examples by Thomas Rahlf

（二）Graphing Data with R by John Jay Hilfiger

（三）https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

（四）https://www.kaggle.com/roshansharma/online-shopper-s-intention

（五）https://github.com/sharmaroshan/Online-Shoppers-Purchasing-Intention

（六）http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf