國 立 臺 北 大 學 統 計 學 系 (所)

碩 士 論 文

指導教授：許玉雪 副教授

顧客消費行為分析及行動銀行使用預測
-決策樹、隨機森林與判別分析之比較

Analysis of Consumer Behaviors of Bank and Prediction of users in Mobile Banking-Comparison of Decision Tree, Random Forest and Discriminant Analysis

研究生：葉子維

中華民國一０七年六月

國 立 臺 北 大 學

統計學系（所） 碩士在職專班

106 學年度第 2 學期畢業生論文

研究生：葉子維撰

業經本委員會審議通過

題　　目：顧客消費行為分析及行動銀行使用預測-決策樹、隨機森林與判別分析之比較

論文考試委員：

召集人 林莉莉

委員 李宜峰

委員 許元聖

委員 _____

指導教授 許元聖

系（所）主任 黃怡婷

論文口試及格日期：

中　華　民　國　107　年　07　月　20　日

國立臺北大學 106 學年度第 2 學期碩士學位論文提要

論文題目：顧客消費行為分析及行動銀行使用預測-決策樹、隨機森林與判別分析之比較

論文頁數：52

所組別：統計學系碩士班（學號：710078908）

研究生：葉子維　指導教授：許玉雪

論文提要內容：

近年來，隨著智慧型手機、Bank3.0、Fintech 以及行動支付的崛起，越來越多人不帶現金出門即可完成一日生活，包含金融、消費、交通...等。因此金融業者越來越注重行動銀行的發展，什麼樣客戶會被行動銀行所吸引，是本研究想要探究的目的。本研究由客戶的刷卡消費特徵去預測行動銀行的潛在使用者，以國內某銀行 2017/2 月到 2017/7 月刷卡消費筆數大於 30 筆的客戶為研究對象;而行動銀行使用者則定義為刷卡消費後三個月有使用行動銀行者，再以簡單隨機抽樣抽出 7,700 位客戶為樣本，客戶的消費行為以 RFM 模式之消費特徵定義。本研究應用決策樹、隨機森林及線性判別分析預測潛在行動銀行使用者，並進行預測結果比較分析。研究結果顯示，使用隨機森林法的整體準確率最高，線性判別分析的靈敏度較好，決策樹的準確率最低。最後，再以模型預測結果提供分眾行銷建議。

關鍵字：行動銀行、RFM、決策樹、隨機森林、線性判別分析

ABSTRACT

Analysis of Consumer Behaviors of Bank and Prediction of users in Mobile Banking

-Comparison of Decision Tree, Random Forest and Discriminant Analysis

by

YEH, TZI-WEI

June 2018

ADVISOR(S): ASSOCIATE PROF. HSU, ESHER
DEPARTMENT: DEPARTMENT OF STATISTICS
MAJOR：STATISTICS
DEGREE: EXECUTIVE MASTER OF BUSINESS ADMINISTRATION IN STATISTICS

In recent years, with the rise of smartphones, banks 3.0, Fintech and mobile payments, more and more people can complete their daily lives without using cash to deal with finance, consumption, transportation and so on. Therefore, the financial industry is paying more and more attention to the development of mobile banking. What kind of customers will be attracted by mobile banking is the purpose of this study. This study aims to predict potential users of mobile banks based on the credit card consumption characteristics of customers.

The mobile banking users are defined as the customers who use mobile banking three months after credit card spending. Customers from a local bank for those who had more than 30 expenses with credit card during Feb. 2017 to Jul. 2017 were taken as the research object. A sample of 7,700 customers' transaction data selected by simple random sampling was used in this study. Customer's consumer behavior is defined by the consumption characteristics of the RFM model.

Random Forest, Decision Tree, and Linear Discriminant Analysis (LDA) are conducted to predict potential users of mobile banking and further for prediction performance comparison. Study results show that Random Forest has the largest prediction accuracy, LDA is more sensitivity, and Decision Tree has the lowest accuracy. Finally, market segmentation marketing was suggested by this study based on the prediction results.

Keywords: Mobile Banking, RFM, Decision Tree, Random Forest, LDA

目錄

表目錄

圖目錄

第一章 緒論

第一節 研究背景與動機

近年來，隨著智慧型手機、Bank3.0、Fintech 以及行動支付的崛起，越來越多人不帶現金出門即可完成一日生活，包含金融、消費、交通...等。行動裝置的普及更是改變現代人的生活方式，現今處於人手一機的世代，只要有網路隨時能使用線上服務，因此金融業者越來越注重行動銀行的發展，致力將臨櫃服務行動化，在這數位資訊爆炸的時代，除了能節省分行櫃檯的負擔，還能帶給客戶更便利的服務。客戶透過手機可以與往來銀行進行資金的查詢、轉帳、換匯、定存及提醒等多項服務。不像電話語音需要經過一層又一層的指示操作，也不像臨櫃辦理受限於時間及地點上。直覺的將金融服務顯示在智慧型手機上，因此越來越多民眾不再選擇到銀行臨櫃抽號碼牌等待，而寧願去學習如何使用數位金融。

吸引一位新的客戶，所花費成本是維持既有客戶的五到七倍(Berry & Parasurman,1991)。因此各銀行皆致力於提升顧客服務來提高客戶的忠誠度，而行動銀行是銀行提升競爭力的一項重要服務，它使得消費者不在受限於時間、地點、及空間的分布。只要金融服務能滿足民眾的需求，操作介面簡單易懂，就能讓顧客持續與該銀行往來，創造更高的價值。

行動銀行越來越重要，而銀行也越來越重視顧客關係管理，本研究將以顧客的消費行為去預測行動銀行的使用者，進而了解行動銀行使用者的特性。如此，銀行就能更確實的做好顧客關係管理，提供更好的服務，培養出有忠誠度的數位金融顧客。

第二節 研究範圍及目的

行動網路應用影響了金融業，行動銀行服務改變了使用者的生活模式，顛覆了傳統銀行的經營模式與理念，開啟了創新服務新世代。到底什麼樣的人會被行動銀行所吸引，是本文主要的研究目的。本研究由客戶的刷卡消費特徵去預測行動銀行的潛在使用者，以國內某銀行的資料做研究，由於有刷卡才有特徵，並限定必須是半年(2017/2 月-2017/7 月)刷卡消費筆數大於 30 筆的客戶，行動銀行使用者則是定義刷卡消費後三個月(2017/8 月-2017/10 月)有使用行動銀行者，再以簡單隨機抽樣抽出 7,700 位客戶為樣本。

研究銀行客戶的特質，我們可以由客戶的刷卡行為去蒐集，假如近一個月消費者在線上做高頻率的刷卡消費可能代表他對網路、3C、行動銀行等產品有很高的接受度。而我們又可定義出無數個諸如此類的刷卡行為，再搭配個人的基本屬性，如年齡、性別、收入等，就可以試圖畫出客戶輪廓。

然而，預測方法大致上可分為兩種：(1)統計方法；(2)機器學習。近年來，基於電腦設備不斷的進步提升，機器學習的演算不再那麼曠日耗時。而且能對大規模的資料進行探勘(Data Mining)，找出有價值的隱藏資訊，得出結構化模式的歸納，時常做為企業在進行決策時的參考依據。

隨機森林近幾年在機器學習中是非常熱門的演算法，由於可處理大量資料及變數，並評估變數的重要性，建造森林可同時在內部對於一般化後的誤差產生不偏估計，可以估計遺失的資料卻不失準確度，所以使用起來相當便利且快速，準確度卻不失傳統的一些模型。據此，本研究試圖應用隨機森林方法找出行動銀行愛用者的特性。

第三節 研究流程

圖 1：研究流程圖

第二章 文獻探討

本章主要在彙整本研究相關的文獻，包含行動銀行、客戶關係管理、RFM以及資料探勘相關的文獻，分述於下。

第一節 行動銀行

行動銀行並非新的金融產品，早在十幾年前，銀行業與電信業者合作STK(SIM Tool Kit)行動銀行服務，針對當時的 GSM 手機 SIM 卡做行動服務。當時的服務僅提供查詢無法交易，又受限於當時上網環境及手機並不風行，無全程加密的安全疑慮，無法大量推廣，很快地宣告失敗。2008 年起，因為 iPhone 帶起了智慧型手機的熱潮，連動帶起國內 3G 網路發展，造就了各種手機應用程式如雨後春筍般的出現。目前國內的行動銀行有多項方便的功能，除了查詢、轉帳、定存、匯率、基金等金融服務外，更結合了即時的生活資訊，舉凡天氣、車票、餐廳、娛樂、地圖等，已是各家行動銀行的標準配備。

行動銀行隨著無線科技與行動裝置的發達，不斷的演變與發展，主要可分為以下三階段的發展：

(1) STK 行動銀行：

STK(SIM Tool Kit，用戶識別模組開發工具)，是一種可以在 SIM 卡中開發應用程式的介面工具。行動銀行早期透過 GSM 手機的 SIM 卡，提供客戶各種金融服務，由於 STK 所提供的程式需事先燒錄於 SIM 卡中，必須搭配特定網路才能使用，客戶透過手機點選需要的服務，再以簡訊傳送交易訊息。當時行動銀行提供的服務僅有帳務查詢，無法進行即時交易，且無法安全加密，導致客戶在使用上產生疑慮，使用起來相當不便利，很快就遭到淘汰。

(2) WAP 行動銀行：

WAP(Wireless Application Protocol，開放式無線網路通訊協定之標準)，主要是提供無線網路通訊服務，銀行開始尋找多家電信業者合作，財金公司更在 2001 年結合銀行與電信業者，共同推出「行動銀行共用系統聯盟」。但在當時 2G 的網路環境中，由於連線速度過慢，無法提升客戶使用意願，行動銀行再次落寞。

(3) APP 行動銀行：

目前行動銀行應用程式分為行動版 APP 即網頁版 APP 兩類。隨著網路的發達，3G 到現在的 4G 以及即將到來的 5G 時代，智慧型動裝置的熱潮與 APP 程式興起，使得銀行重新重視行動銀行的重要性，目前已有超過 20 家銀行推出「APP 行動銀行」，主要的功能都以查詢、轉帳、繳費、基金下單、定存、換匯為主。近年來行動銀行使用已呈現快速成長趨勢，將會是電子金融非常重要的一環。

第二節 顧客關係管理

顧客關係管理(Customer Relationship Management, CRM)，最早發展客戶關係管理的國家是美國，這個概念最初由美國學者 Gartner Group 所提出，在 1980 年初便有所謂的接觸管理(Contact Management)，那個時期企業蒐集的資訊是與自身有關聯為主，即專門收集客戶與公司聯繫的所有信息，到 1990 年則演變成包括電話服務中心支持資料分析的客戶關懷（Customer Care），而現今進入了網路時代，運用科技的技術整合了行銷、銷售、管理與服務，使得顧客關係管理邁進了一個新紀元。

隨著關係行銷與資料庫行銷的觀念被廣泛的運用，顧客關係管理也越來越重要。顧客關係管理可區分為「資料庫行銷」、「一對一關係行銷」、「事件行銷」這三大部份。顧客關係管理的發展是源於企業對顧客價值的認知以及對顧客服務的重視，而企業在行銷觀念上的演進大致上經歷了四大階段：早期以交易為主的「大量行銷」；到以產品觀念為主的「目標行銷」， 再到以顧客為主的「顧客行銷」，後來演進至以一對一顧客關係為主的「一對一行銷」，而「關係行銷」則是強調有效行銷關係之建立。CRM 的核心是客戶價值管理，它將客戶價值分為既成價值、潛在價值和模型價值，通過一對一營銷原則，滿足不同價值客戶的個性化需求，提高客戶忠誠度和保有率，實現客戶價值持續貢獻，從而全面提升企業盈利能力。

Kalakota & Robinson (1999) 認為 CRM 是運用整合性銷售、行銷與服務下的一套系統，企業使用 CRM 發展出一致性行動，來滿足客戶需求。企業在結合整合行銷流程與科技發展下找出客戶需求，並且在產品與服務上求改進，致力於建立客戶忠誠度。Mulinder (1999) 認為 CRM 是與客戶建立終身關係，其內涵包含

四種要素：個體、誠實、熟悉與互動。現今幾乎大部分公司都有開發屬於自己的顧客關係管理流程與工具，CRM 已經成功的應用資訊技術，將客戶資料、銷售、服務、行銷等多方面功能整合，並可以針對個別客戶進行專有服務，可以有效提升企業營運成果與客戶忠誠度。

Kalakota & Robinson (1999) 認為 CRM 是運用整合性銷售、行銷與服務下的一套系統，企業使用 CRM 發展出一致性行動，來滿足客戶需求。企業在結合整合行銷流程與科技發展下找出客戶需求，並且在產品與服務上求改進，致力於建立客戶忠誠度。Mulinder (1999) 認為 CRM 是與客戶建立終身關係，其內涵包含四種要素：個體、誠實、熟悉與互動。現今幾乎大部分公司都有開發屬於自己的顧客關係管理流程與工具，CRM 已經成功的應用資訊技術，將客戶資料、銷售、服務、行銷等多方面功能整合，並可以針對個別客戶進行專有服務，可以有效提升企業營運成果與客戶忠誠度。

Kalakota & Robinson (2001) 認為了解顧客關係管理的重要性有下述幾點：

(1) 吸引一位新的客戶，所花費成本是維持既有客戶的五到七倍。

(2) 只要有一位顧客不滿意，這個不好的經驗通常會被通知給八至十個人。

(3) 如果企業顧客維持率若能每年成長 5%，則企業利潤將可提升 85%。

(4) 對新顧客行銷成功率只有 15%，但對舊顧客行銷成功率卻有 50%。

(5) 企業若能將服務缺失迅速解決，70%的抱怨顧客仍會再回頭與公司交易。

有關於顧客關係管理的定義，Linoff (1999)認為顧客關係管理是結合數種資訊科技的應用，目的在保留對企業有貢獻的客戶，也是一個持續改善的過程，從顧客生命週期中去了解顧客行為，進而提供所需的商品或服務。Alex(1999)認為顧客關係管理是滿足大部分對企業具有價值顧客的需要，同時也是許多技術與觀念的集合與發展，其涉及的技術與觀念包含行銷學、一對一行銷、資料倉儲、資料探勘、顧客區隔與忠誠度。Davids(1999)則認為顧客關係管理就是「關係管理」、「終身價值行銷」、「忠誠行銷」、「一對一行銷」，名詞雖不同，但其內涵是一樣的。

綜合以上這些學者的意見，顧客關係管理主要重點在於：企業在找新顧客的同時應花更多的心力維繫好舊客戶。因為尋找一位新顧客所消耗成本遠比維繫舊客戶多上許多，因此培養顧客忠誠度和加強顧客滿意度，是顧客關係管理的施行重點。

第三節 RFM 整理

Hughes(1994)的研究發現最近購買日(Recency)、購買頻率(Frequency)及購買金額(Monetary)是分析消費者終身價值的主要因素，因此建立了 RFM 模型。說明如下：

(1) 最近購買日(Recency)：最後一次購買某商品的日期距離現在越近，比起最近一次購買日較遠的客人更容易再次消費。

(2) 購買頻率(Frequency)：在一段區間內購買某商品的次數越高，表示對該商品或公司的忠誠度亦越高。

(3) 購買金額(Monetary)：在一段區間內購買某商品的金額越高，代表對公司的貢獻營收越高，也會被公司認定為高價值客戶。若以一段區間的總金額來看，客戶剛好此段區間的消費頻率較低，可能會低估客戶的消費能力，故通常以平均金額來代替。

Kahan(1998)認為 RFM 是研究顧客行為分析非常簡單卻有用的工具。行銷人員不需專屬軟體即可進行分析，且可因業務不同而自行定義 R、F 及 M 的加權權重，因此廣被各行各業所應用。由於 RFM 強調運用客戶過去的消費歷史資料來區隔客戶，能夠幫助企業經營管理，因此也常被企業用來衡量顧客的忠誠度、貢獻度及顧客價值。

表 1 可看出不同學者對於，R、F、M 區隔客戶的分法及三個項目的權重分配有不同的意見。Hughes(1994)認為 RFM 模型中的三個指標重要性相同，不應給予不同的權重。將客戶的購買日期、購買頻率、購買金額遞減排序，前 20% 的客戶給予「5」的標示，前 20%-40%的客戶給「4」，以此類推，RFM 這三個構面的組合為 5x5x5，可將客戶標示為 125 群。Stone(1995)則認為原始的 RFM 模型有其限制：各產業的 RFM 模型指標應該不同，應依其產業特性去調整指標的重要性；RFM 模型並非僅能分為 125 群，可視資料庫規模大小而有所調整；RFM 模型指標並不具備預測能力，僅針對顧客歷史消費行為去做區隔。因 Stone(1995)的實證資料為信用卡，為符合其產業特性，將 F 給予最高權重，R 為中，M 最低。區分客戶的組合數也有所調整。Miglautsch(2000)將 R 依照購買的日期區間均分為五等分，F 則將只購買一次的先區分出來，剩下依照購買頻率之平均值區分，M 則是沿用 Hughes 的方式。

表 1：FRM 模型指標整理

指標分數	R(Recency)	F(Frequency)	M(Monetary)
Hughes	前 20%客戶：5 分 前 20%~40%客戶：4 分 前 40%~60%客戶：3 分 前 60%~80%客戶：2 分 最後 20%客戶：1 分		
Stone	近 3 個月：24 分 近 4~6 個月：12 分 近 7~9 個月：6 分 近 10~12 個月：3 分 超過 12 個月：0 分	購買次數×4 分	購買金額×10% (上限為 9 分)
Miglautsch	近 3 個月：5 分 近 4~6 個月：4 分 近 7~12 個月：3 分 近 13~24 個月：2 分 超過 24 個月：1 分	先區分出購買 1 次者給 1 分，剩餘顧客>平均值給 5 分，依此類推，最後剩下的一群給 2 分	同 Hughes 作法

資料來源：本研究整理

第四節 資料探勘

資料探勘(Data Mining)，又稱為資料考古學(Data Archaeology)或資料挖掘(Data Dredging)，起源於二次大戰前美國政府用於軍事及人口普查，之後被廣泛地應用於各領域。近年來由於資訊科技的進步和顧客關係管理(CRM)的流行，資料探勘成為相當熱門的資料分析方法之一。資料探勘運用在商業用途上的主要價值，係藉由有效的工具來降低成本或增加收益，以提高公司的利潤。以下將大略介紹資料探勘的定義、技術及探勘流程。

一、資料探勘的定義

資料探勘是知識發現(Knowledge Discovery in Database, KDD)的步驟之一，資料探勘係透過演算法，將資料作一分析與應用，以找出其特徵(pattern)與模式(model)的過程，茲將各學者對資料探勘的定義整理如下表 2。根據表 2，本研究認為資料探勘不僅在強調發現的過程，而更應強調分析者的投入，故本研究將「資料探勘」定義為：從蒐集到的大量資料中，經過資料轉換、清理，並運用各種演算法賦予資料意義，找出未被發現，有用的樣式及資訊。

表 2：資料探勘的定義

學者	定義
Grupe & Owrang (1995)	資料探勘是從現存資料中取得以前從未得知的事實與未知曉的新關係。
Curt (1995)	資料探勘是一種資料轉換的過程，先從沒有組織的數字與文字集合而成的資料，轉換為資訊，再轉換成知識，最後產生決策。
Fayyad et al. (1996)	知識挖掘主要是透過選取適當的資料，進行處理與轉換，去挖掘未知但具有意義的資訊，而資料探勘只是知識挖掘過程中的步驟，要先瞭解資料的特性與其相關領域的專業知識與技術，透過特殊的演算法來獲得資料的特徵與型態。
Berry & Linoff (1997)	資料探勘是經由自動或半自動的方式，分析大量的資料，尋找有效的模型與規則。
Cabena et al. (1997)	資料探勘是將未知且有效的資訊從大型資料庫抽出的過程，並且將萃取出的有用資訊提供給主管做決定性的決策。
Hall (1998)	資料探勘是一種結合資料視覺化(data visualization)、機器學習(machine learning)、統計方法、以及資料倉儲(data warehousing)多種技術，以便從龐大資料量中，攫取以規則形式或其他模式所表達的知識。
McCluskey & Anand (1999)	資料探勘是一種發展中的科技，主要有機器學習、資料庫技術、統計學、演算法與數學等技術。透過半自動化的流程，可以從大量的資訊中取得有用且未知的知識。
Olmeda & Sheldon (2001)	資料探勘是人工智慧與統計的混合模型，從雜亂的資料庫中獲得有益的知識。
Berson (2001)	資料探勘是發現「有意義的新」(meaningful new)相互關係、樣式和趨勢的程序。

資料來源：本研究整理

二、資料探勘的技術方法

Berry & Linoff(1997)將資料探勘的技術分為六大類，這些功能的意義及相對應的演算法概述如下：

(1) 分類(Classification)

透過歷史資料中的各種屬性及特徵值，來建立分類法則，推導出存在資料中的事實。常用的方法有決策樹(Decision Tree)或類神經網路(Neural Network)等。

(2) 推估(Estimation)

根據既有連續性數值之相關屬性資料，以獲得某一屬性未知之值。例如：透過年收入來推估刷卡消費量。常用的方法有相關分析、迴歸分析(Regression Analysis)，及類神經網路等。

(3) 預測(Prediction)

預測是根據某特定對象屬性，觀察其過去行為或歷史資料建立模型，以預測未來的數值以及趨勢。相關技術包括迴歸分析、時間序列分析(Time Series Analysis)、類神經網路及案例庫推理(Case-Based Reasoning)等。

(4) 關聯法則(Association)

關聯法則主要用於了解龐大資料庫中某些資料項目彼此之間的關聯，用來判斷何種事件常常同時發生或存在。例如：常常購買啤酒的人同時也會購買尿布，表面上看不出任何關聯，但實際研究才發現，原來爸爸們在家顧小孩常常會喝著啤酒看球賽。相關的技術如購物籃分析(Market Basket Analysis)、模糊集合(Rough Set)。

(5) 集群(Clustering)

集群係根據資料間的相似性，將相似的資料分為一群，將不同屬性的資料區隔開來，使組間差異最大、組內差異最小。與分類方法的差異是，集群沒有事前明確的定義和規則可循。

(6) 時間序列(Time Sequential)

時間序列的重點在了解不同時間點上各事件的關聯性。主要分為順序性與週期性兩種型態，順序性用於了解事件發生之時間先後關係，週期性用於了解時間區段的變化，分析時間區段內所發生的事情，在其他相同的時間區段內是否亦會發生。常用的方法有類神經網路及時間序列分析等。

三、資料探勘的流程

完整的資料探勘或知識發現的流程分為以下五大類(Fayyad et al., 1996; Cabena et al., 1998; Tan et al., 2006)：

(1) 選取資料(Data Selection)

資料探勘第一步，是根據資料探勘的目的和目標，將需要的資料由龐大的資料庫中抽取出來，成為小的資料超市(Data Mart)做為分析基礎。資料來源不一定是同一個資料庫，可能從一個或多個資料庫中選取所需要的資料。

(2) 資料處理(Data Processing)

亦為資料淨化(Data Cleaning)，將有缺漏值、異常值、或矛盾的地方時加以修正，以獲得正確、乾淨的資料，是個非常繁瑣、卻常被忽略掉的重要工作。

(3) 資料轉換(Data Transformation)

根據欲分析的問題和方法，將目標變數由原始狀態轉換為可分析的形式；或是利用數學的運算，將資料以不同的維度(Dimension)呈現，以突顯出目標之特徵，如計算標準差、成長率、標準化(Normalization)等。

(4) 資料探勘(Data Mining)

使用資料探勘方法如分類、趨勢分析、分群、關聯及循序特徵等，從轉換後的資料中發掘存在的多種特徵及資訊。

(5) 解釋與評價結果(Interpretation)

經過資料探勘分析後的結果，透過文字或視覺化的圖形對擷取出來的資訊作一解釋與評價。可經由專家的審查將不適宜的規則進行修正或刪除，亦可將不同資訊探勘技術所獲得的知識進行整合，併入系統執行、評估成果。

第三章 研究方法

本研究使用的隨機森林(Random forest)方法，是以決策樹(Decision tree)為基底去做改良，兩個隨機觀念的加入：Leo Breimany 在 1996 年提出的 Bagging (Bootstrap aggregating)；Tin Kam Ho 在 1995 年提出的 Random subspace method，使得模型不易過度擬和(Overfitting)。因此在介紹隨機森林之前，我們必須要先了解決策樹的理論。而線性判別分析是傳統的統計方法，建立線性判別函數，可有效地切割不同類別資料。後續本研究將運用決策樹、隨機森林及判別分析去預測行動銀行的使用者，並比較其差異。

第一節 決策樹理論

決策樹常被使用於分類或預測的技術，屬於非線性資料分類法，可有效解決僅能線性區隔方法的缺點，如邏輯斯迴歸(Logistic Regression)。分析結果以樹狀圖呈現，其結構如圖 2 所示，由樹的根節點(Root Node)開始，按照分類的問題或屬性展開，樹的中間節點(Non-Leaf Node)代表測試的條件，每個節點代表一個屬性，分支(Branches)代表條件測試的結果，葉節點(Leaf Node)則代表分類的類別。

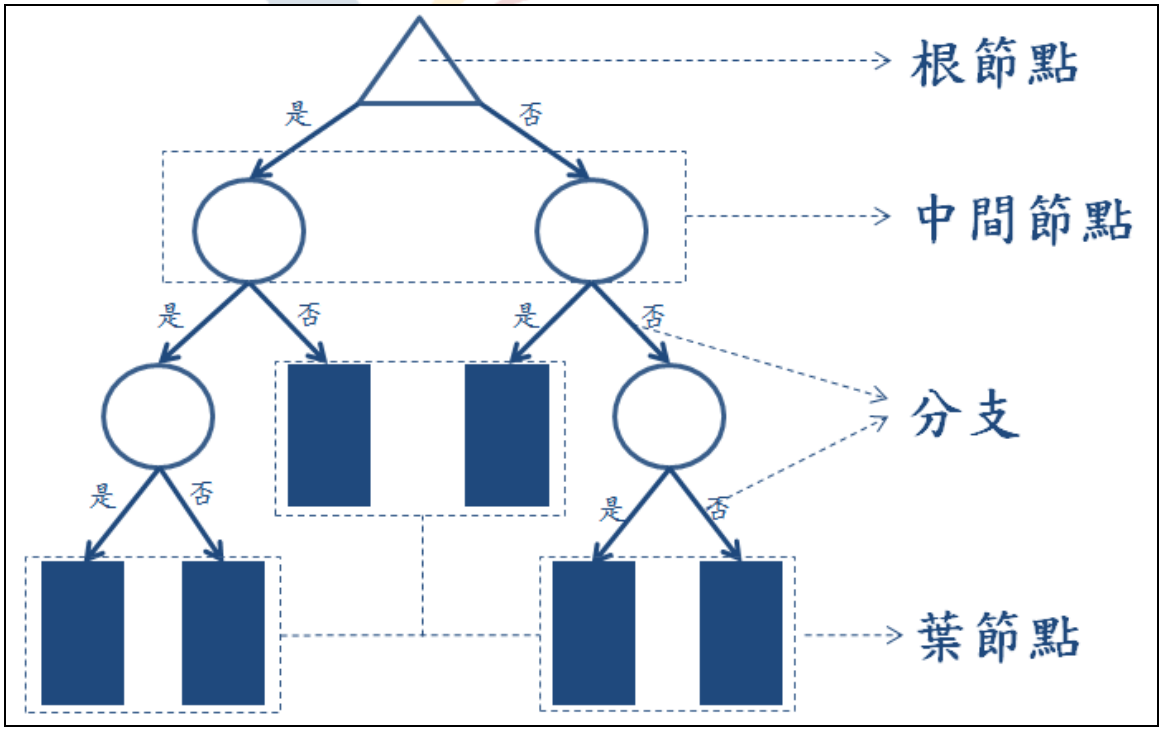

圖 2：決策樹結構示意圖

資料來源：本研究整理

決策樹有許多不同的方法，因建構的方法不同，步驟可能也有所不同。一般而言，要建立決策樹，先以訓練資料(Training Data)建立模式(Model)，再以測試資料(Testing Data)修剪決策樹，提高整體正確率後，才能以此模式去預測新的資料。建置決策樹的過程大致上可分為以下五個步驟：

(1) 決定分隔變數(Attribute Selection Measure)：

決策樹透過遞迴分割(Recursive Partitioning)建立而成，它是一種將資料分割成不同小的部分的疊代過程。如有下列情況，決策樹將停止分割：該群的每一筆資料都已歸類到同一類別；該群資料已無法再找到新的屬性做節點分割；該群已無任何尚未處理的資料。

(2) 培育完整的樹(Full Tree)：

將所有資料放置根節點，由根節點開始分隔出兩個或更多的節點，再以相同方式將每個節點分隔，不斷重複這個動作，直到無法找到可以顯著降低一個節點分散度的分隔，就將其標示為葉部節點，直到每一個分割都只剩下葉節點，就已培育出完整的樹。但完整的決策樹通常無法配適出最好的結果，很容易出現過度擬和(Overfitting)，所以這棵樹需要被評估與修剪。

(3) 評估每個節點及整體錯誤率：

決策樹建構完成後，訓練組的每一筆資料都被分配到各個葉節點，每個葉節點都有一個錯誤率，所有葉節點的錯誤率的加權總數就是整棵樹的錯誤率，可以用以評估及修正。

(4) 修剪決策樹(Tree Pruning)：

完全生長的決策樹容易出現過擬和(Overfitting)的狀況，像是特殊樣本這樣的離群值。例如：某一項分類經過了十個以上的節點，而最後留在葉節點的資料量極少，很有可能就是某種特例。此時，以測試組的資料，來測試修剪後的分支錯誤率，選擇最低留下來的值，此值經過修剪已去除掉 Overfitting 的狀況，又不會修剪掉有價值的資訊。修剪決策樹，其實在做的事就是刪除離群值(Outlier)，避免過度學習。

(5) 應用所建立模型套用在新資料上

決策樹學習的關鍵在於，在每個分裂節點處如何選擇最優劃分屬性。一般而言，隨着劃分過程不斷進行，我們希望決策樹的分支節點所包含的樣本儘可能屬於同一類別，即節點的「純度」越來越高。以下我們介紹三種決策樹最常使用到的特徵選擇準則：

(1) 信息增益(Information Gain)：

在訓練過程中決策樹會問出一系列的問題，像是年齡是否>30 歲，年收

入是否<100 萬之類的是非問題。由最上方的樹節點開始用資料的特徵將資料分割到不同邊，分割的原則是：這樣的分割要能得到最大的資訊增益(Information gain)。如公式(3.1)，可解釋為：獲得的資訊量=原本的資訊量-分割後的資訊量。

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{3.1}$$

D_p：整體資訊量，N_p：整體資料總數，

D_j：第 j 個分割資訊量，j=1,2,…,m(特徵總數)，

N_j：第 j 個資料總數，j=1,2,…, m(特徵總數)，

由於我們希望獲得的資訊量要最大，因此經由分割後的資訊量要越小越好，信息增益的資訊量最常使用的是熵(Entropy)，其公式如(3.2)：

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t) \tag{3.2}$$

$p(i|t)$：屬性的機率，c：分類類別，i：第幾個類別的資料總數，
我們用下面的例子來說明「熵」：如下這三筆資料集，何者只需要最少的資訊量便可清楚的說明：

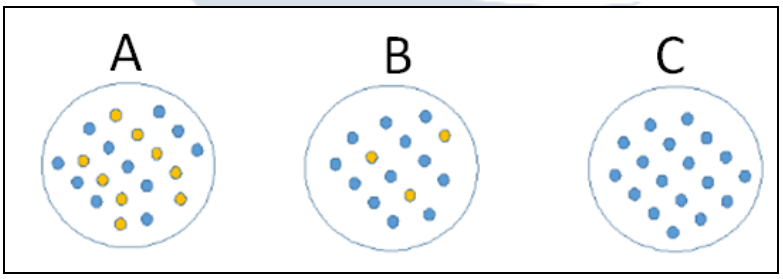

圖 3：Entropy 範例說明

很明顯是 C，我們一眼就看出 C 的資料都屬於同一類型不需多作說明，B 要稍加說明，才能解釋其中三個不同資料點的意義，而 A 則要解釋得相當多才行。因此，我們可以得到一個結論：越純淨的資料所需要的資訊量愈少，反之，越雜亂的資料需要的資訊量愈多。當所有的資料都一致，它們的 Entropy 就是 0，如果資料各有一半不同，那麼 Entropy 就是 1。

(2) 增益率(Gain Ratio)：

實際上，信息增益準則對可取值數目較多的屬性有所偏好，為減少這種偏好可能帶來的不利影響，C4.5 決策樹算法不直接使用信息增益，而是使用增益率(Gain Ratio)來選擇最優劃分屬性。因為增益率偏向選擇可取值數目較少的屬性，C4.5 算法並不是直接選擇增益率最大的候選劃分屬性，而是使用了一個啟發式的方法：先從候選劃分屬性中找出信息增益高於平均水平的屬性，再從中選擇增益率最高的。

(3) 基尼指數(Gini Index)：

基尼指數是另外一種數據的不純度的度量方法，採用基尼指數的代表是 CART 決策樹，它的概念與「熵」雷同，最大的差別在於「熵」一次可產生多個不同節點，基尼指數一次僅能產生兩個，即 True 或 False 的 Binary 分類。基尼指數也是決策樹衡量資訊量最常使用的指標之一，其公式如(3.3)：

$$Gini(D) = 1 - \sum_{i=1}^{n} p(i \mid t)^2 \tag{3.3}$$

D 包含 n 個樣本，$p(i \mid t)$ 為屬性的機率，當 $Gini(D)=0$ 時，在此節點處所有的資料都屬於同一類別，此時能獲得有用的資訊量最大；當 $Gini(D)=1$ 時，即此節點的資料屬性均勻分佈，能獲得有用的資訊量最小。如果將集合切分為 l 個部分，則進行分割的 Gini Index($Gini_{split}$)，如下列公式(3.4)：

$$Gini_{split}(D) = \sum_{i=1}^{l} \frac{n_i}{n} Gini(i) \tag{3.4}$$

l 為子節點個數，n 為節點母體樣本數，n_i 為子節點屬於 i 的樣本數，使用公式(3.4)找出最小的 $Gini_{split}(D)$，用以作為此節點處屬性分裂的標準。

決策樹有許多不同的方法，本節就以最常見的 CHAID、CART、C4.5 來做介紹，並且分別比較差異。由於隨機森林就是基於簡單且快速的 CART 決策樹理論上，因此後續本研究將使用 CART 決策樹與隨機森林方法去做預測及模型比較。

一、CHAID(Chi-square Automatic Interaction Detection)

由 Hartigan 在 1975 年提出，僅限於類別變數使用，CHAID 演算法的步驟是透過

卡方檢定去找出最顯著的分類屬性來進行分類，而分割後的資料再重複進行上述步驟去做分類，直到預先設定 P 值的大小為不顯著後而停止，所以此方法是屬於事前修剪，而不是等到決策樹完全成長後才做修枝。

二、CART(Classification and Regression Tree)

分類迴歸樹由 Breiman 於 1984 年提出，由於方法簡單且快速，並且能處理連續型與類別變數，是建構決策樹最常使用的方法之一。CART 主要是透過 Gini Index 作為屬性的分散度。若分散度越低，則表示每個節點下，資料越屬於同一種分類。而 CART 就是透過降低 Gini Index 來達到每個節點的分類純度提升，使得每個節點中資料的同質性較高，按照此規則分割，能生長出完整的決策樹，最後再進行修枝，屬於事後修剪。此演算法透過整體節點的誤判率(Entire Error Rate)去逐一檢視每一個葉節點，找出那些為無法有效降低整體節點錯誤率的分支，將其修剪。

三、C4.5

由 Quinlan 於 1993 年提出，為他先前提出之 ID3 的延伸，改良了 ID3 產生過多的子集合，而每個子集合僅有少數資料的問題，且更具有處理連續型變數、雜訊及決策樹修剪的能力。C4.5 主要是透過增益率(Gain Ratio)來劃分屬性，先從劃分屬性中找出信息增益高於平均水平的屬性，再從中選擇增益率最高的。完整生長出決策樹後，採取節點中錯誤分割的比例(Error Rate)去做修剪，屬於事後修剪。

以上三種演算法各有優缺點，需視資料特性及目的需求而使用。表 3 為這三種方法的差異比較：

表 3：決策樹整理表

演算法比較項目	CHAID	CART	C4.5
處理資料類型	離散	離散、連續	離散、連續
分裂數	不受限制	二元樹	不受限制
分類屬性選擇	使用卡方檢定找出 P 值最顯著者	找出 Gini 值最小的為分類屬性	先從候選劃分屬性中找出信息增益高於平均水平的屬性，再從中選擇增益率最高的為分類屬性
樹的修剪方式	倘若已無任何分類屬性能使資料達到顯著差異，即停止決策樹成長，屬於事前修剪	建立完全生長的決策樹，再以整體節點的誤判率進行修枝，屬於事後修剪	建立完全生長的決策樹，再以節點中錯誤分割的比例進行修枝，屬於事後修剪

資料來源：本研究整理

第二節 隨機森林法

隨機森林由 Breiman(2001)提出的新型決策樹演算法，採用 CART 決策樹作為元分類器，用隨機的方式建立一個森林，森林裡面由很多的決策樹組成，隨機森林的每一棵決策樹之間相互獨立。這個方法結合了 Breiman(1996)的「Bootstrap aggregating」的想法和 Tin Kam Ho(1995)的「random subspace method」去建造決策樹的集合。換言之，它使用拔靴法(bootstrap)隨機地取出放回重複抽取樣本資料、使用 random subspace method 隨機選取 m<M(訓練資料集特徵)個特徵變數來生成決策樹。Breiman(1996)提出 Bagging 的想法，如圖 4 所示。

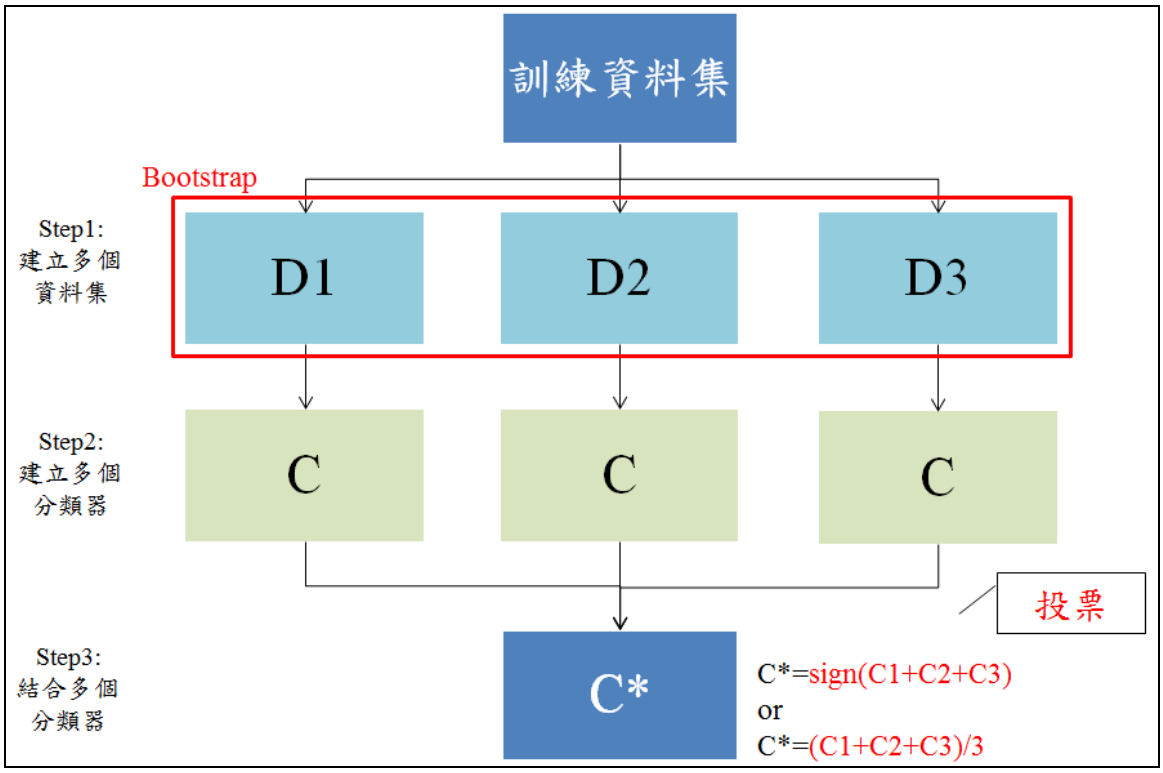

圖 4：Bagging 示意圖

資料來源：本研究整理

隨機森林在建立每一棵決策樹的過程中，有兩點需要注意：採樣與不剪枝分裂。首先，是兩個隨機採樣的過程，隨機森林中每一顆決策樹對訓練資料集要進行行、列的採樣，如圖 5 所示。關於行採樣，需與訓練資料集的筆數相同，採用有放回的方式，也就是在採樣得到的樣本集合中，可能會有重複的樣本。然後進行列採樣，決策樹的每一次分裂，從訓練資料集的 M 個特徵屬性中，隨機地選取 m 個，一般推薦 m 值的大小為 $\sqrt{M}$ ，再由 m 個特徵屬性中選取最重要的特徵去做分裂，直至完全生長決策樹。由於兩個隨機採樣的過程確保了隨機性，所以

就算不剪枝，也不會出現 overfitting。

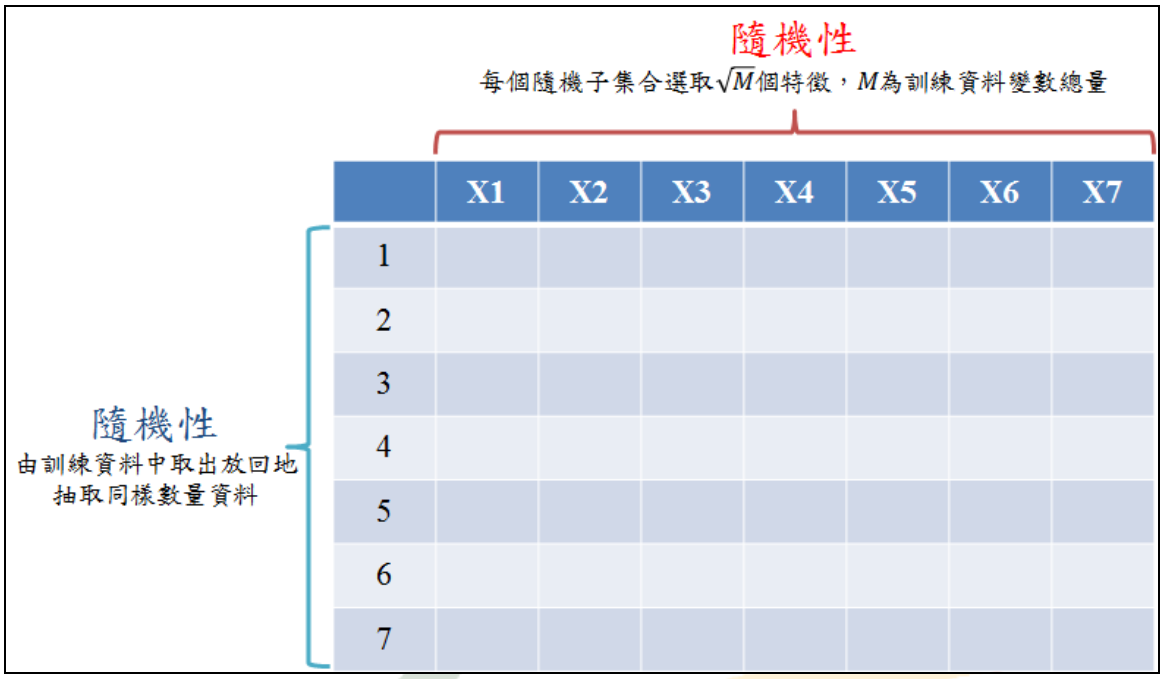

圖 5：隨機森林隨機採樣過程示意圖

資料來源：本研究整理

按這種演算法得到的隨機森林中的每一棵樹都是屬於弱分類器，但是經由大家組合起來就很變成厲害的強分類器。可以這樣比喻隨機森林演算法：每一棵決策樹就是一個精通於某一個窄領域的專家，而隨機森林中就有了很多個精通不同領域的專家，對一個新的問題（新的輸入資料），可以用不同的角度去看待它，最終由各個專家，投票得到結果。

隨機森林的演算步驟如下，其中的每一棵樹都是根據隨機向量值建立，而隨機向量是依據固定機率分配產生。圖 6 為隨機森林流程示意圖：

(1) 將資料切分為訓練資料(Training data)與驗證資料(Testing data)。

(2) 決定訓練資料 N 筆，分類特徵 M 種。

(3) 透過拔靴法(Bootstrap)，隨機由訓練資料重複抽樣出 T 組子集合(T 越大越好)，稱為袋內資料(InBag)。當|N|很大時，資料未被抽中的概率為 $(1-\frac{1}{|N|})^{|N|}\approx 0.368$。因此，抽中的袋內資料約占訓練資料集的 63.2%，剩下未抽中的樣本稱為 out-of-bag(OOB)約占 36.8%，可以用來估計模型的錯誤率(OOB-error)。當樹夠多時，隨機林可保證所產生的推論錯誤率的上限公式如(3.5)。由此可知：每顆樹的分類強度越大，則隨機森林的分類性能越好；樹之間的相關度越大，則隨機森林的分類性能越差。

$$\text{推論錯誤率} \geq \frac{\bar{\rho}(1-s^2)}{s^2} \tag{3.5}$$

$\bar{\rho}$：決策樹之間的平均相關程度

S^2：決策樹的分類能力

(4) 決策樹的每一次分裂，透過 random subspace method 隨機選取 $\sqrt{M}$ 個特徵去做分裂。

(5) 不剪枝，完全生長出 T 顆決策樹。

(6) 最後根據 T 顆決策樹進行投票(Vote)，選出最佳的結果。

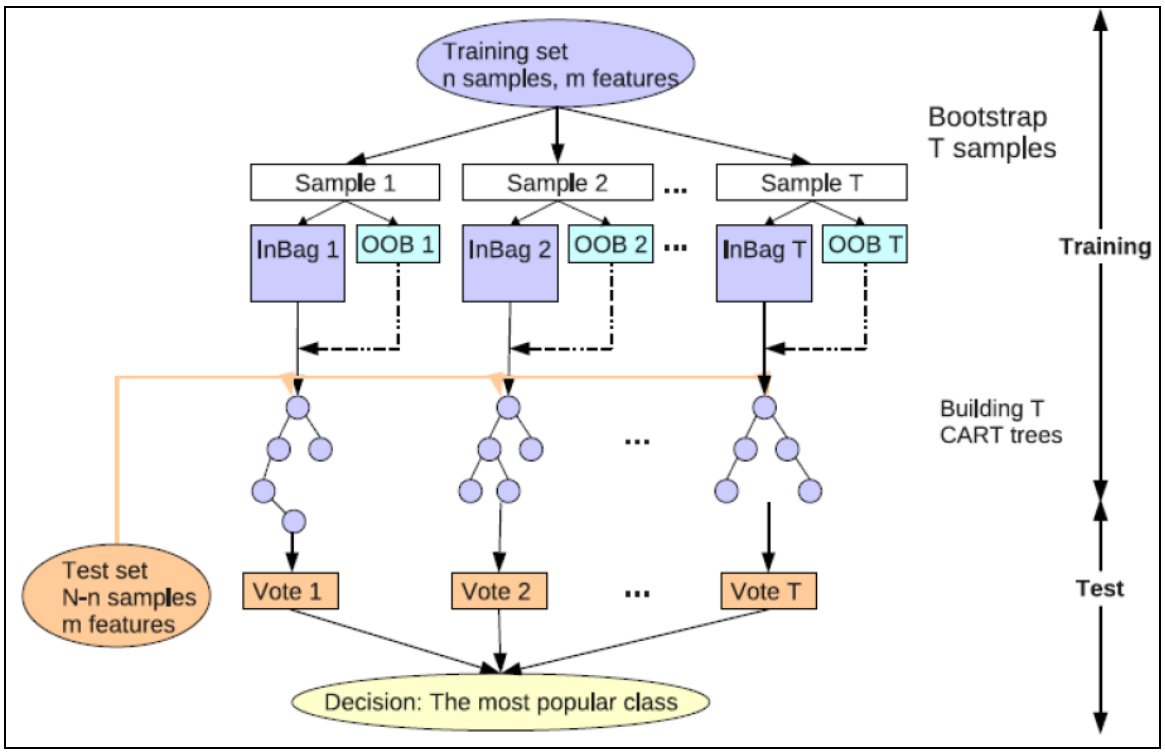

圖 6：隨機森林流程示意圖

資料來源：Guo(2011)

隨機森林的特徵選擇，使用的方法就是統計上常使用的 permutation test。Permutation test 用來計算特徵屬性的重要性，也就是對應變數的影響力。它的作法是用訓練資料建好模型，再將訓練資料中要檢測的變數資料打亂，以打亂的資料重新建構模型，再去比較新的模型與未打亂此變數前的模型去比較，若準確率較低則表示此變數對模型的影響能力高，其重要性也較高。依此方法去檢測每一個特徵屬性的重要性。

因此，隨機森林在做變數重要性的計算時，就必須一直重新建構決策樹，若樹木很大或是特徵變數很多時，將會非常耗時。所以隨機森林會檢測的是袋外資料中的變數，將其打亂再放入決策樹中預測，看預測的結果與打亂前的差別，便可知道變數的重要性。

以決策樹方法產生的規則具有結構簡單直觀、容易被理解、及計算效率高的特點，且能夠有效地抑制訓練樣本噪音和解決屬性缺失等問題。因此可以防止由於訓練樣本存在雜訊和資料缺失引起的精度降低。但決策樹也有與生俱來的缺點：

(1) 分類規則雜。

(2) 收斂到非全域的局部最佳解。

(3) 容易過度配適而導致 Overfitting。

為了克服以上的缺點，而引入了另一種預測模式：隨機森林，它具有以下特徵：

(1) 能夠有效地處理大的資料集。

(2) 可以處理成千上萬個特徵變數。

(3) 能夠在分類的過程中生成一個泛化誤差的內部不偏估計。

(4) 可以有效地處理遺失值卻不失準確率

(5) 可以處理不平衡的資料

(6) 抽樣未用到的樣本可做為驗證資料

但它也有缺點，結合多棵樹，無法清楚說明決策過程，模型無法被人所理解，因此被人們稱為黑盒子。

第三節 線性判別分析法

一、費雪判別分析(Fisher's Linear Discriminant Analysis, Fisher's LDA)

根據英國統計學家 Fisher(1936)所提出，是一種線性判別方法，其意圖是將 d 維空間中的數據點投影到 c-1 維空間上去，使得不同類的樣本點在這個空間上的投影儘量分離，同類的儘量聚集。在判別分析中，自變數為判別變數，透過判別變數的線性組合所建立之判別函數進行判別依據。而依變數為分類變數，根據判別函數的結果進行分類結果判別。在進行分類問題前，會先蒐集已知類別的資料，針對這群資料推導判別函數，再經由判別函數建立判斷準則，再利用此函數，對未知類別的資料進行分類動作。當資料不符合常態或是資料分佈未知時，則可使用費雪判別模型。假設兩母體判別變數具有相同之共變異數矩陣，將兩母體多變量的資料經過線性組合轉換成單變量資料，並透過此函數來判別母體之間的差異。最好的區別情況為兩母體之中心點差距越大越好，而母體間內差距越小越好。

判別分析主要的目的是瞭解群體的差異，先利用判別變數建立判別函數，再依此函數對群體作分類，預測每個個體屬於各群組的可能機率。判別分析之目的有下列四點：

(1) 找出判別變數的線性組合，使組間變異相對於組內變異的比值為最大，而每一個線性組合與先前已經獲得的線性組合均不相關。
(2) 檢定各組之重心是否有差異。
(3) 找出那些判別變數具有最大的判別能力。
(4) 根據新受試者的預測變數的數值，將該受試者指派到某一群體。

線性判別分析是由 Fisher 在 1936 年所創立，其理論不需要假設資料為常態分配就可以得到判別函數，此分析模式是以尋找判別變數 $x_1, x_2 ..., x_p$ 的線性組合 $y = w_1x_1 + w_2x_2 + ... + w_jx_j$ 之最佳權重 w_j，使檢定群體間 y 值平均數相等的組間變異數對組內變異數比值為最大。其中 x 為判別變數向量 $x_j = (x_1, x_2, .., x_j)$，$w$ 為判別係數(權重)向量 $w_j = (w_1, w_2, .., w_j)$，$j$ 為向量個數。

假設資料組數只有分兩群組時，我們可用 a_i 代表第 A 組資料，用 b_i 代表第 B 組資料。a_{ij} 為 A 群組第 j 個判別變數第 i 個觀察值，b_{ij} 為 B 群組第 j 個判別變數第 i 個觀察值，n_a 為 A 群組個數，n_b 為 B 群組個數。

(1) 計算兩組資料的差別向量 D：

$$d_j = \overline{A_j} - \overline{B_j} = \frac{\sum_{i=1}^{n_a} a_{ij}}{n_a} - \frac{\sum_{i=1}^{n_b} b_{ij}}{n_b} \tag{3.6}$$

(2) 計算其 A 群組及 B 群組之共變異數矩陣(Variances and Covariance Matrix)：

$$S_A = \begin{bmatrix} SS_A & SP_A \\ SP_A & SS_A \end{bmatrix} \text{及} \; S_B = \begin{bmatrix} SS_B & SP_B \\ SP_B & SS_B \end{bmatrix} \tag{3.7}$$

$$SS_{Aj} = \sum_{i=1}^{n_a} \left(a_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{n_a} \left(a_{ij}\right)^2\right)^2}{n_a}$$

$$SP_{Ajk} = \sum_{i=1}^{n_a} a_{ij} a_{jk} - \frac{\sum_{i=1}^{n_a} a_{ij} \sum_{i=1}^{n_a} a_{ik}}{n_a}$$

(3) 求出 A、B 兩組資料之綜合共變異數矩陣(Pooled Variances and Covariance Matrix)：

$$S = \frac{S_A + S_B}{n_a + n_b - 2} \tag{3.8}$$

(4) 求解下列方程式，以得到 W(判別係數)：

$$SW = D \Rightarrow W = S^{-1} D \tag{3.9}$$

(5) 解出矩陣解後，得到 W，可表示為：

$$W = \left(w_1, w_2, ..., w_j\right) \tag{3.10}$$

(6) 最後我們可以得到一組經判別分析計算過後的最佳線性模式，再計算其判別分數 y (Discriminant Score)：

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + ... + w_j x_{ij} \tag{3.11}$$

y_i：第 i 個觀察值的判別分數，i=1,…,n

x_{ij}：第 i 個觀察值第 j 個變數，i=1,…,n，j=1,…,k

(7) 計算判別指標 R_0 (Discriminant Index)來分類 A 組與 B 組資料，如果判別值 $y_i > R_0$ 就可以將資料區分為 B 群組，反之為 A 群組。

$$R_0 = \frac{\mu_B + \mu_A}{2} \quad (\mu_B：\text{B 群體之重心，} \mu_A：\text{A 群體之重心}) \qquad (3.12)$$

二、其他判別分析方法：以下介紹幾個其他較為著名的判別分析方法

(1) 馬氏距離法：

x、y 兩點的馬氏距離如公式(3.13)，S 為資料的共變異數矩陣

$$(x-y)' S^{-1}(x-y) \qquad (3.13)$$

分別計算每一個觀察樣本到每一組別重心(centroid)的距離，再將其分類到與某一重心距離最小的組別。由馬氏 P. C. Mahalanobis 所發展出來的距離法，使用傳統的歐幾里得距離公式套用聯合組內共變異數矩陣加以調整，計算觀察值到 A 群組重心的馬氏距離如果較到 A 群組重心距離較小，則分類為 A 群組，反之，則歸類為 B 群組。

(2) 最大概似法(maximum likelihood)：

此法依據事前機率(prior probability)及馬氏距離計算觀察樣本歸屬於某一組別的事後機率(posterior probability)，將其分類到機率最高的組別。

本研究將運用費雪判別分析函數，將客戶分類為「使用行動銀行」及「不使用行動銀行」這兩類人，並找出判別函數上 loading 較高的變數，用以判斷何種特徵將影響行動銀行使用與否。

第四章實證分析

本研究以國內某銀行客戶基本屬性及刷卡特徵資料進行實證分析。超過半年以上的刷卡特徵較為久遠，不具參考價值，刷卡頻率較低的客戶，沒有行為特徵可萃取，因此本研究資料範圍限定於半年(2017/2 月-2017/7 月)刷卡消費筆數大於 30 筆的客戶，行動銀行使用者則是定義在刷卡消費後三個月(2017/8 月-2017/10 月)有使用行動銀行者。本研究使用 R 軟體，受限於電腦硬體設備及建模效能，因此以簡單隨機抽樣抽出 7,700 位客戶為樣本。

第一節 變數的選取

本研究依據客戶的基本屬性及刷卡特徵，採納了共 30 個變數。其中，客戶基本屬性共 12 變數列如表 4。

表 4：客戶基本屬性變數定義彙整表

變數名稱	變數定義
性別	男；女
年齡	客戶年齡
婚姻	未婚；已婚；其他
教育	專科；大學；高中高職；碩士；博士；其他
職業	公務；企業；其他；教育；軍警；技術
居住地	台北市；新北市；基隆市；桃園市；新竹市；新竹縣；苗栗縣；台中市；彰化縣；南投縣；雲林縣；嘉義市；嘉義縣；台南市；高雄市；屏東縣；宜蘭縣；花蓮縣；金門縣
電子帳單記號	1：電子帳單；0：實體帳單
信用卡網站會員記號	1：信用卡網站會員；0：無
電子報記號	1：願意收到電子報；0：無
持有多卡或單卡	S：單卡；M：多卡；A：流失
卡齡	持有信用卡月份
信用卡額度	信用卡核定額度

Hughes(1994)認為衡量消費者行為應該考量購買日期、購買頻率及購買金額這三個面向，且三個指標重要性相同，不應給予不同的權重。然而，本研究則是要利用刷卡特徵去找出客戶性格或偏好，進而去預測誰會使用行動銀行，因此購買金額及購買日期相對就不那麼重要。所以我們採用 Stone(1995)提出重要性不

同的想法，重要性排序應為：購買頻率>購買日期>購買金額。本研究認為購買頻率幾乎解釋掉過半的客戶偏好，而購買日期及購買金額只是輔佐參考。參考 Stone 作法，將購買頻率的配分放大，將購買日期及購買金額配分降低，調整後的定義如表 5 所示。

表 5：調整後 RFM 指標定義

R(Recency)	F(Frequency)	M(Monetary)
近 36 天：15 分 近 36~72 天：12 分 近 72~108 天：9 分 近 108~144 天：6 分 近 144~180 天：3 分	購買次數×5 分	購買金額/1000 (上限為 9 分)

依據調整後的 RFM 指標，可將刷卡消費特徵區分為 18 類變數，每一類指標分數為 R+F+M，其定義及說明如表 6。

表 6：客戶刷卡特徵變數定義彙整表

變數名稱	變數定義
量販店消費 RFM	大潤發、家樂福...等超級市場量販店消費
百貨公司消費 RFM	新光三越、Sogo 百貨...等消費
加油站消費 RFM	中油、台塑...等加油站消費
台高鐵客運消費 RFM	高鐵、火車、客運...等消費
汽車相關消費 RFM	汽車零件、保養...等相關消費
餐廳消費 RFM	餐廳美食相關消費
國內旅遊消費 RFM	國內旅館、旅遊...等相關消費
國外旅遊消費 RFM	國外旅館、旅行社、航空公司、機票、外幣計價...等相關消費
醫療消費 RFM	醫療器材、醫療服務、藥品、醫院...等消費
保險消費 RFM	保險相關消費
分期消費 RFM	分期付款消費
電信費 RFM	台灣大哥大、遠傳、中華電信...等電信消費
電影消費 RFM	電影院消費
線上消費 RFM	網路刷卡消費
三 C 用品消費 RFM	三 C 相關產品消費
房屋傢具工程 RFM	房屋修繕、傢具、工程...等相關消費
美容按摩 RFM	化妝品、按摩院、整脊、理髮美容、整形...等消費
生活娛樂 RFM	舞蹈、遊戲、錄影帶、運動、藝術、音樂...等生活相關消費

第二節 樣本敘述統計

針對本研究 7,700 位樣本資料，初步進行結構描述。依客戶基本屬性，彙整如表 7。以「性別」來看，女性略高，人數共 4,089 人，占了樣本資料的 53%。以「年齡」來看，本研究採用刷卡資料，而銀行辦卡條件須年滿 20 歲以上，因此 0~19 歲的級距為 0。年齡較多集中在 40~49 歲的級距，占了 34%；次高集中在 30~39 歲，占了 29%；而最小(20 歲~29 歲)及最老(60 歲以上)的級距各僅占了 8%，這說明此樣本資料的年齡層集中在青壯年。以「婚姻」來看，未婚 42%，已婚 43%，差異不大，無填寫資料及離婚則歸類在其他，占 15%。以「教育程度」來看，其他部位最高，大部分為資料不齊全，國中以下學歷也歸類於此；剩餘項目以大學最高，占了 25%，其次為高中、高職或專科。以「職業」來看，無資料歸類在其他，剩餘項目以一般企業員工比例為 63%最高，而教育、技術、公務及軍警皆僅占個位數比例。以「居住地」來看，新北市最高，占了 32%，其次為台北市占了 24%。雙北地區人數占比已過半，這也跟本研究的國內某銀行的分行多半分佈在北部有關。

彙整信用卡相關狀態，描述如表 8，持有信用卡網站會員及寄送電子帳單的人數皆過半，表示 E 化程度較高；願意接收電子報僅占 22%，可能有 E 化意願但不希望收到行銷訊息打擾；持有多卡的比例比持有單張信用卡高出 5%；持有信用卡的時間超過 9 年的人數占了 59%，表示大多數的人皆為老客戶；信用卡額度 20 萬以上就占了 38%，15 萬~20 萬也占了 13%，這表示大多數的人可能經濟狀況不錯。

針對 7,700 位，在 2017/3 月~2017/8 月，這半年的 18 類消費特徵，初步進行描述。首先，以 RFM 的「最近一次消費日期」來看，依最近一次消費日期歸類各消費月份的人數及比例，如表 9 所示。消費人數前三名為線上消費、百貨公司消費及餐廳消費；線上消費 R 在消費最近的一個月，也就是 2017 年 8 月份，線上消費的人數占比很高，在最早的那一個月份(2017/3 月)很低，看的出線上消費的特性：會線上消費的人，消費的時機點很近，都集中在前一個月，如果超過 3 個月沒線上消費就幾乎不會來了；其餘消費屬性 R 皆為消費月份越近，人數越多。值得注意的是國外旅遊消費 R、房屋傢具工程 R、美容按摩 R 及電影消費 R 的最近一次消費日期分佈較為平均，這表示此種消費每次發生的時間距離不會太近。

表 7：樣本敘述統計-客戶基本屬性相關狀態

基本屬性相關變數	狀態	人數(人)	比例(%)
性別	女性	4,089	53
	男性	3,611	47
婚姻	未婚	3,219	42
	已婚	3,344	43
	其他	1,137	15
年齡	20 歲~29 歲	645	8
	30 歲~39 歲	2,269	29
	40 歲~49 歲	2,599	34
	50 歲~59 歲	1,585	21
	60 歲以上	602	8
教育	專科	829	11
	大學	1,906	25
	高中高職	913	12
	碩士	600	8
	博士	75	1
	其他	3,377	44
職業	企業	4,878	63
	教育	274	4
	技術	250	3
	公務	91	1
	軍警	51	1
	其他	2,156	28
居住地	新北市	2,426	32
	台北市	1,843	24
	高雄市	980	13
	桃園市	651	8
	台中市	535	7
	台南市	266	3
	新竹市	226	3
	新竹縣	166	2
	彰化縣	132	1
	基隆市	112	1
	宜蘭縣	91	1
	屏東縣	86	0
	苗栗縣	41	0

	嘉義市	37	0
	嘉義縣	24	0
	南投縣	23	0
	花蓮縣	22	0
	雲林縣	20	0
	金門縣	19	0

表 8：樣本敘述統計-信用卡相關狀態

信用卡相關變數	狀態	人數(人)	比例(%)
信用卡網站會員	有	5,227	68
	無	2,473	32
寄送電子帳單	有	4,308	56
	無	3,392	44
寄送電子報	有	1,683	22
	無	6,017	78
持有多卡或單卡	單卡	3,600	47
	多卡	3,979	52
	流失	121	2
卡齡	3 年內	671	9
	3 年~6 年	1,117	15
	6 年~9 年	1,399	18
	9 年~12 年	819	11
	12 年以上	3,694	48
信用卡額度	5 萬內	426	6
	5 萬~10 萬	1,928	25
	10 萬~15 萬	1,419	18
	15 萬~20 萬	1,027	13
	20 萬以上	2,900	38

表 9：樣本敘述統計-最近一次消費日期

人數比例	2017 3 月 人(%)	2017 4 月 人(%)	2017 5 月 人(%)	2017 6 月 人(%)	2017 7 月 人(%)	2017 8 月 人(%)	消費人數
線上消費 R	0	3	5	8	14	70	5,771
百貨公司消費 R	5	7	10	12	22	45	5,307
餐廳消費 R	3	6	9	11	21	50	5,253
國外旅遊消費 R	22	8	11	13	17	30	4,347
加油站消費 R	4	5	6	7	14	66	4,291
量販店消費 R	5	7	9	11	19	48	4,160
三 C 用品消費 R	7	9	11	13	20	41	4,154
保險消費 R	9	8	9	12	15	47	3,202
生活娛樂 R	6	8	8	10	17	51	3,138
分期消費 R	8	9	12	14	21	36	2,950
電信費 R	3	3	4	6	9	75	2,794
台高鐵客運消費 R	9	9	12	14	19	37	2,270
房屋傢具工程 R	11	12	14	15	21	27	2,203
汽車相關消費 R	8	10	12	15	18	36	2,197
國內旅遊消費 R	8	12	14	12	22	32	2,049
醫療消費 R	10	11	15	15	18	30	1,853
美容按摩 R	11	13	16	16	20	25	1,482
電影消費 R	13	15	11	17	23	20	839

如表 10 所示，以 RFM 的「消費筆數」來看，消費筆數最高為線上消費，次高為加油站消費，遠遠高於其他消費項目。加油為交通所需，消費頻率高可以理解；線上交易頻率高，表示民眾越來越喜歡在網路上購物，E 化的程度也越來越高，但其標準差也很高，這表示每個人在網路購物的金額相差甚大。以 RFM 的「消費金額」來看，消費金額最高為國外旅遊，次高為分期消費，第三名則為保險消費。國外旅遊、分期消費、保險消費，這三項交易金額高是可被理解的。

平均每筆消費金額則是分期消費最高，每筆已超過一萬元；次高為國外旅遊消費，平均每筆有 6,404 元；再來則是國內旅遊的 5,916 元。我們發現線上消費及加油站的平均筆數很高，但是每筆消費金額卻低；國內旅遊、分期消費的筆數較低，但每筆消費金額卻高。

表 10：樣本敘述統計-消費筆數及消費金額

消費項目	筆數 (平均數)	筆數 (標準差)	金額 (平均數)	金額 (標準差)	平均每筆 消費金額
線上消費	16	29	30,081	63,038	1,842
加油站消費	13	15	10,579	17,647	790
電信費	8	8	7,923	11,028	951
汽車相關消費	8	22	13,587	24,043	1,720
生活娛樂	8	31	10,687	69,340	1,381
國外旅遊消費	8	13	49,042	89,572	6,404
百貨公司消費	8	11	16,675	36,535	2,191
量販店消費	7	11	6,569	12,336	879
餐廳消費	7	9	10,420	21,249	1,531
保險消費	7	10	38,142	117,168	5,663
台高鐵客運消費	6	11	6,820	15,176	1,111
分期消費	5	8	46,123	69,545	10,070
三 C 用品消費	4	7	11,178	22,469	2,703
醫療消費	3	3	12,625	29,822	4,642
國內旅遊消費	3	3	15,537	84,644	5,916
房屋傢具工程	2	5	8,619	30,276	3,586
美容按摩	2	4	9,794	21,857	4,642
電影消費	2	2	1,576	4,552	787

透過前一節表 5 的 RFM 指標計算，呈現計算過的 18 類消費特徵指標的基本敘述，按指標平均數由大到小排序，如表 11。RFM 指標的平均數以線上消費為 103 最高、再來是加油站消費。因為指標權重以消費筆數最重，因此高低排序大致上與消費頻率相同。變異係數則是生活娛樂最高，表示不同人在生活娛樂上的指標差異較大；而在電影、國內旅遊、醫療及電信費上，差異較小。

我們將應變數，行動銀行使用的差異描述彙整如表 12，有使用行動的人占樣本 7,700 位的 33%。初步發現，使用行動銀行者偏年輕，卡的額度較低，其 E 化程度很高：像信用卡網站會員、電子帳單及電子報的人數占比皆比無使用行動銀行的人高出許多。在消費特徵上，較少在量販店消費，卻是在汽車相關、線上及生活娛樂上消費較高。

表 11：樣本敘述統計-RFM 指標

RFM 指標	最大值	最小值	平均數	標準差	變異係數
線上消費 RFM	5,794	11	103	145	1.41
加油站消費 RFM	724	8	86	77	0.89
電信費 RFM	534	8	61	41	0.67
汽車相關消費 RFM	2,154	8	57	110	1.91
百貨公司消費 RFM	834	8	57	58	1.02
國外旅遊消費 RFM	2,078	8	57	69	1.22
生活娛樂 RFM	7,789	8	56	154	2.76
量販店消費 RFM	904	8	55	60	1.10
保險消費 RFM	549	8	54	52	0.97
餐廳消費 RFM	989	8	53	49	0.94
台高鐵客運消費 RFM	739	8	48	56	1.18
分期消費 RFM	779	9	44	41	0.93
三 C 用品消費 RFM	989	8	38	35	0.91
國內旅遊消費 RFM	329	8	31	20	0.64
醫療消費 RFM	186	8	30	20	0.65
房屋傢具工程 RFM	1,116	8	28	29	1.04
美容按摩 RFM	614	8	27	20	0.73
電影消費 RFM	134	8	23	13	0.55

表 12：樣本敘述統計-行動銀行使用者

比較項目	是否使用行動銀行	
狀態	無	有
人數(人)	5,139	2,561
人數比例(%)	67	33
女性占比(%)	53	52
平均年齡(歲)	45	39
平均額度(元)	175,457	168,026
信用卡網站會員占比(%)	61	81
電子帳單占比(%)	43	81
電子報占比(%)	20	25
量販店消費 RFM(分)	59	46
汽車相關消費 RFM(分)	48	73
線上消費 RFM(分)	97	112
生活娛樂 RFM(分)	52	61

第三節 決策樹分析

本研究使用 R 軟體之 Rpart，為了處理分類問題，因此選定 method="class"，而對於 control(rpart.control object)常用到的重要參數解釋如下，
(1)與(2)的用意在於避免節點過度分支，造成某些特殊案例去影響到模型效力所造成的 overfitting，(3)的主要作用是通過修剪明顯不值得拆分的變數來節省計算時間。

(1) minsplit：每一個節點最少要有多少資料

(2) minbucket：在根節點最少要有多少資料

(3) cp(complexity parameter)：決定納入節點分裂的變數顯著性

(4) maxdepth：決策樹的深度

首先，我們將 7,700 筆資料切分 70%為訓練資料(Training Data)，30%為測試資料(Testing Data)，變數分割的顯著水準使用預設值 cp=0.01 去觀測分類規則，只可分出四條規則(圖 7)，其中只有一條規則對於反應變數為 1 的機率高於一半(0.58)，而這樣的規則對於商業應用的價值太低，因此我們再將顯著水準調整為 0.001，這樣卻產出了上百條規則(圖 8)，恐已陷入過擬合(Overfitting)。最後我們挑選顯著水準為 0.005 去做分裂，可產出 17 條規則(圖 9)，其中有 7 條規則對於反應變數為 1 的機率高於一半。

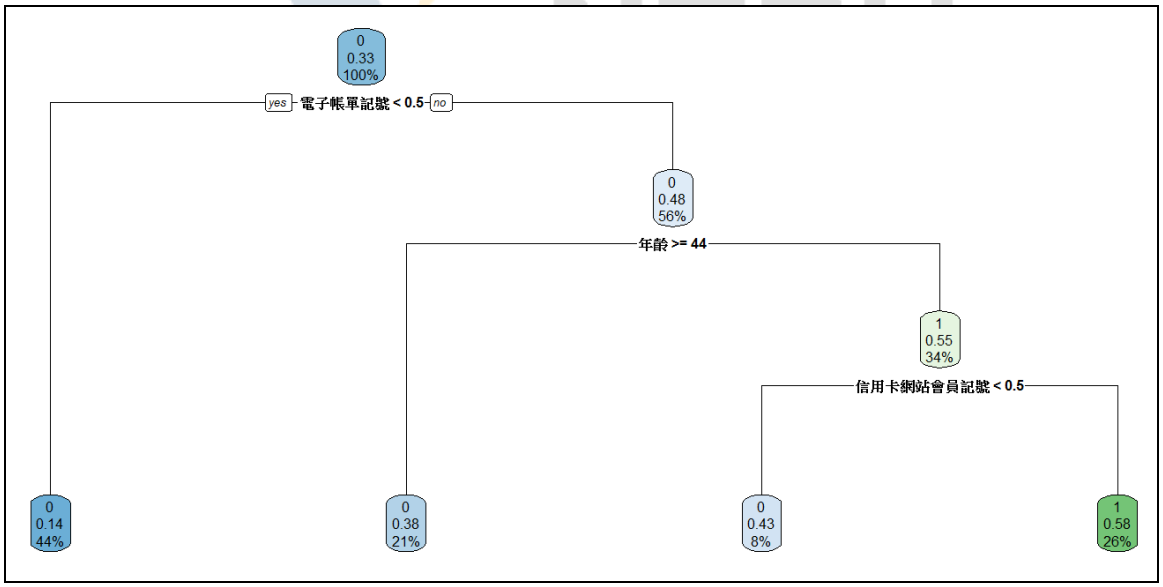

圖 7：決策樹分類結果(cp=0.01)

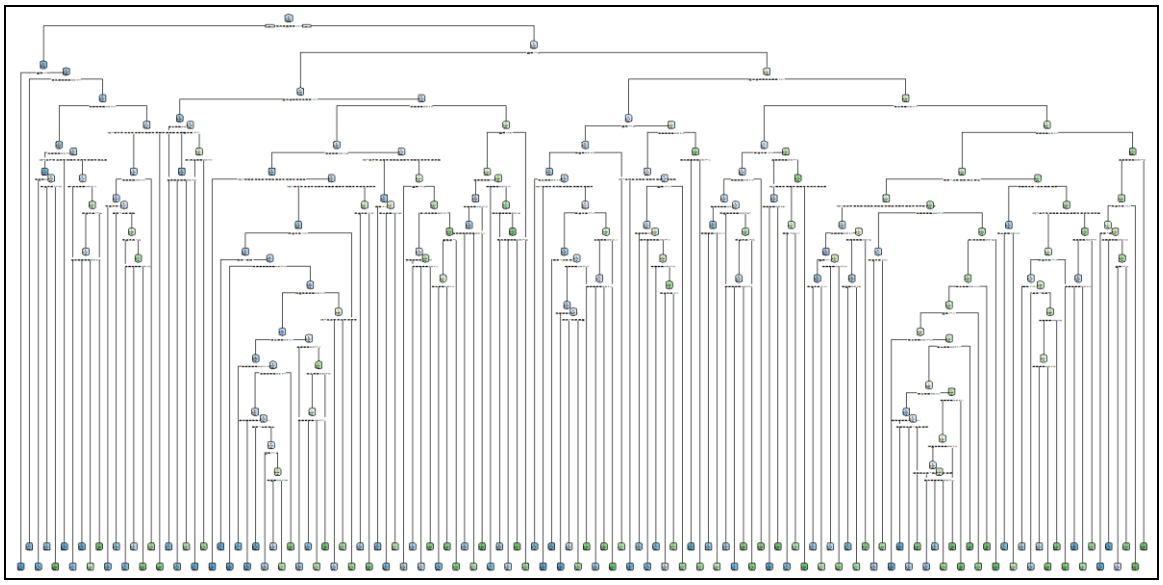

圖 8：決策樹分類結果(cp=0.001)

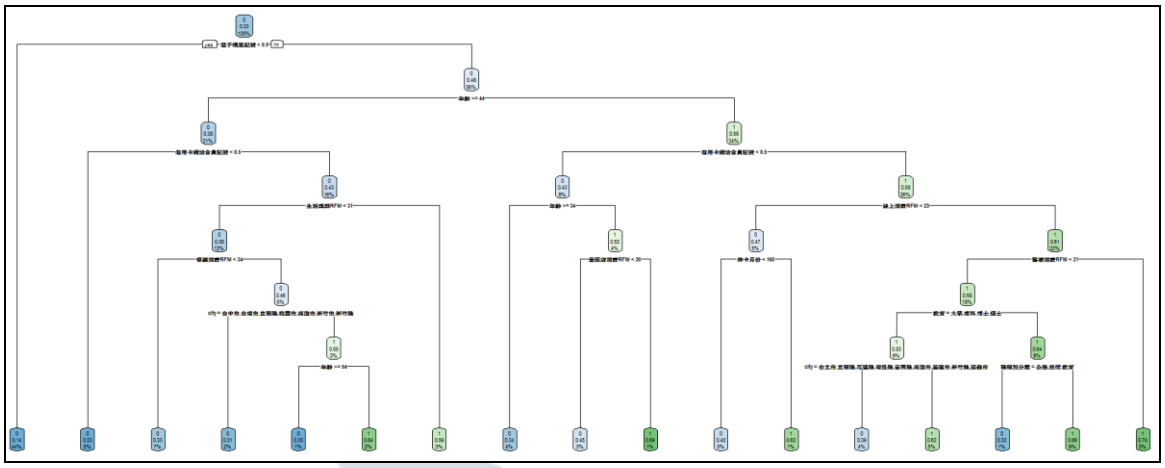

圖 9：決策樹分類結果(cp=0.005)

minsplit 的預設值為 20，但希望每個節點的資料最少有 500。使用調整前的參數去建模得到的整體準確率為 0.7049，調整後的準確率為 0.7157，提升了約一個百分比。而對於根節點的最少資料量，我們採用預設建議的 minsplit/3，且不再手動去調整樹的深度，任其依照顯著水準 α =0.005 及 minsplit=500 的條件下去生長。如圖 10，最後產出的規則共 7 條，其中有 4 條規則對於反應變數為 1 的機率高於一半。以下 4 條規則為潛在行動銀行使用者的特徵：

(1) 規則 1：

寄送電子帳單且年齡≥44 歲且為信用卡網站會員且生活娛樂 RFM 指標≥31 分，占整體人數的 3%，使用行動銀行的機率為 59%。

(2) 規則 2：

寄送電子帳單且年齡<44 歲，占整體人數的 34%，使用行動銀行的機率為 55%。

(3) 規則 3：

寄送電子帳單且年齡<44 歲且為信用卡網站會員，占整體人數的 26%，使用行動銀行的機率為 58%。

(4) 規則 4：

寄送電子帳單且年齡<44 歲且為信用卡網站會員且線上消費 RFM 指標≥23 分，占整體人數的 22%，使用行動銀行的機率為 61%。

本文使用一種對分類模型進行效果評估的方法：混淆矩陣(Confusion matrix)。混淆矩陣是機器學習對於分類方法準確率進行評估的工具，透過將模型預測的數據與測試數據進行對比，使用準確度、靈敏度和特異性等指標對模型的分類效果進行度量。我們以訓練資料建構模型，再將此訓練好的模型套用在測試資料上，就會產生每筆資料對於「是否使用行動銀行」所發生的機率，這時我們抓取會使用行動銀行的機率>0.5 的人將他定義為 1，由於 0.5 是一個中間值，對兩種結果都沒有任何的傾向性，因此我們選擇 0.5 為 cut point。再將測試的實際資料與前面定義的預測資料去做 2 維度交叉表的比對，就產生表 13 的混淆矩陣。

表 13：決策樹混淆矩陣

預測 / 實際	0	1
0	1,313(真陰性)	226(假陽性)
1	431(假陰性)	341(真陽性)

真陽性(True Positive，TP)：真實值是陽性且軟體預測輸出結果是陽性，正確。
真陰性(True Negative，TN)：真實值是陰性且軟體預測輸出結果是陰性，正確。
假陽性(False Positive，FP)：真實值是陰性且軟體預測輸出結果是陽性，不正確。
假陰性(False Negative，FN)：真實值是陽性且軟體預測輸出結果是陰性，不正確。

靈敏度(Sensitivity，SEN)：代表正確預測為陽性的命中率，亦被稱為真陽性率。

$$SEN=\frac{341}{431+341}=44.17\% \quad (4.1)$$

特異性(Specificity，SPE)：代表正確預測為陰性的命中率，亦被稱為真陰性率。

$$SPE=\frac{1{,}313}{1{,}313+226}=85.32\% \quad (4.2)$$

準確度(Accuracy，ACC)：代表預測正確的準確度。

$$ACC=\frac{1{,}313+341}{1{,}313+226+431+341}=71.57\% \quad (4.3)$$

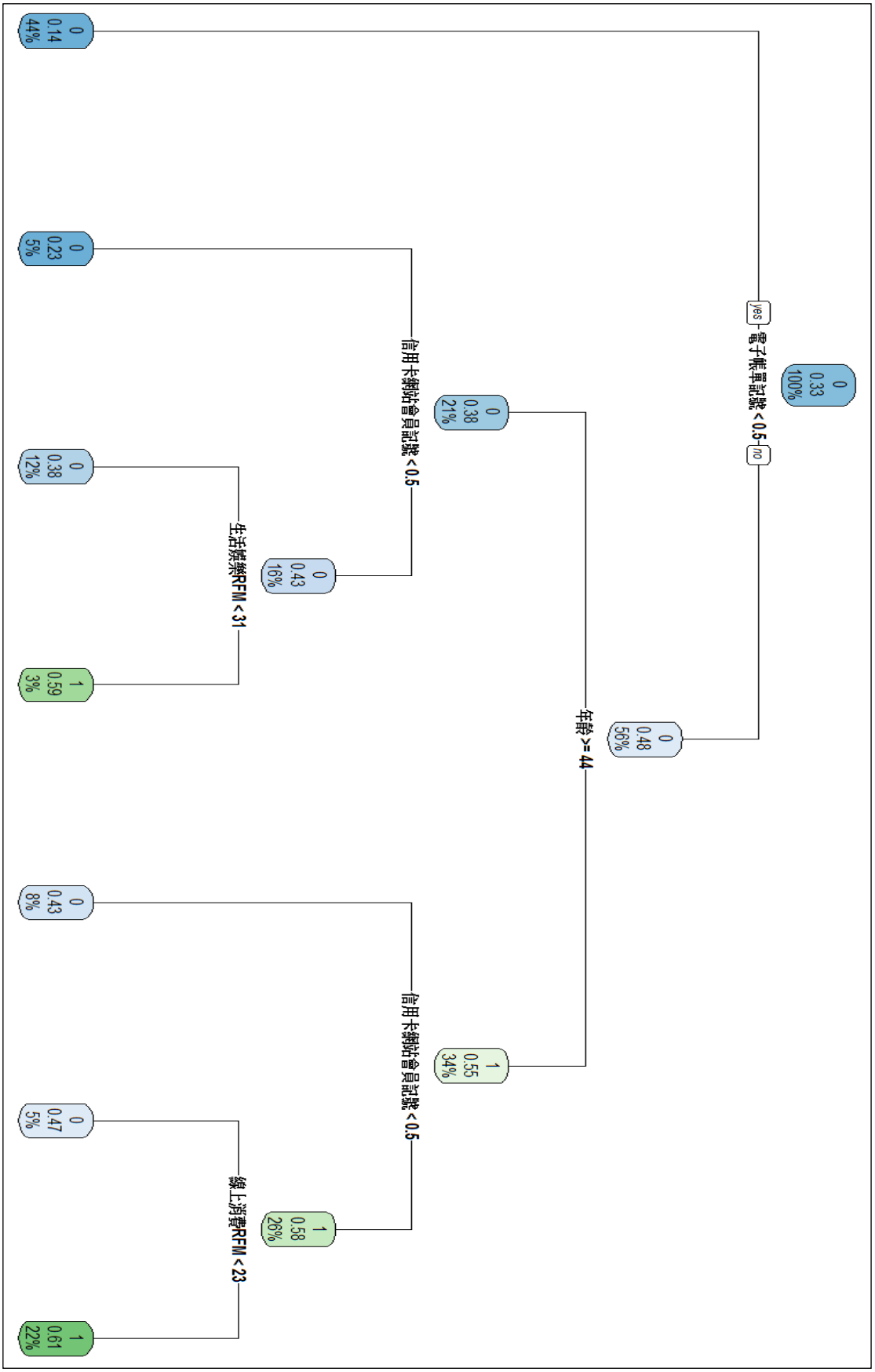

圖 10：決策樹最終分類結果

第四節 隨機森林分析

本研究使用 R 軟體之 RandomForest，任其不剪枝完全生長，所以我們需要控制的參數只有兩個：

(1) mtry：隨機選取 m 個特徵變數來生成決策樹

(2) ntree：決定樹的數量

首先，我們將 7,700 筆資料切分 70%為訓練資料(Training Data)，30%為測試資料(Testing Data)，mtry 建議不超過 $\sqrt{M}$ ，特徵變數有 30 個，因此 mtry 先取 4。ntree 是隨機森林中決策樹的數量，當 ntree 生長足夠時，能觀察隨機森林的 OOB-error 與 ntree 的變化，當袋外錯誤率趨於平穩時，就決定了樹木的數量。

袋外錯誤率與 ntree 的關係圖如圖 11，我們發現當 ntree=200 時，整體的袋外錯誤率趨於平穩。因此本研究將設定 ntree 七個不同值(200、400、600、800、1000、1200、1400)去建立隨機森林。

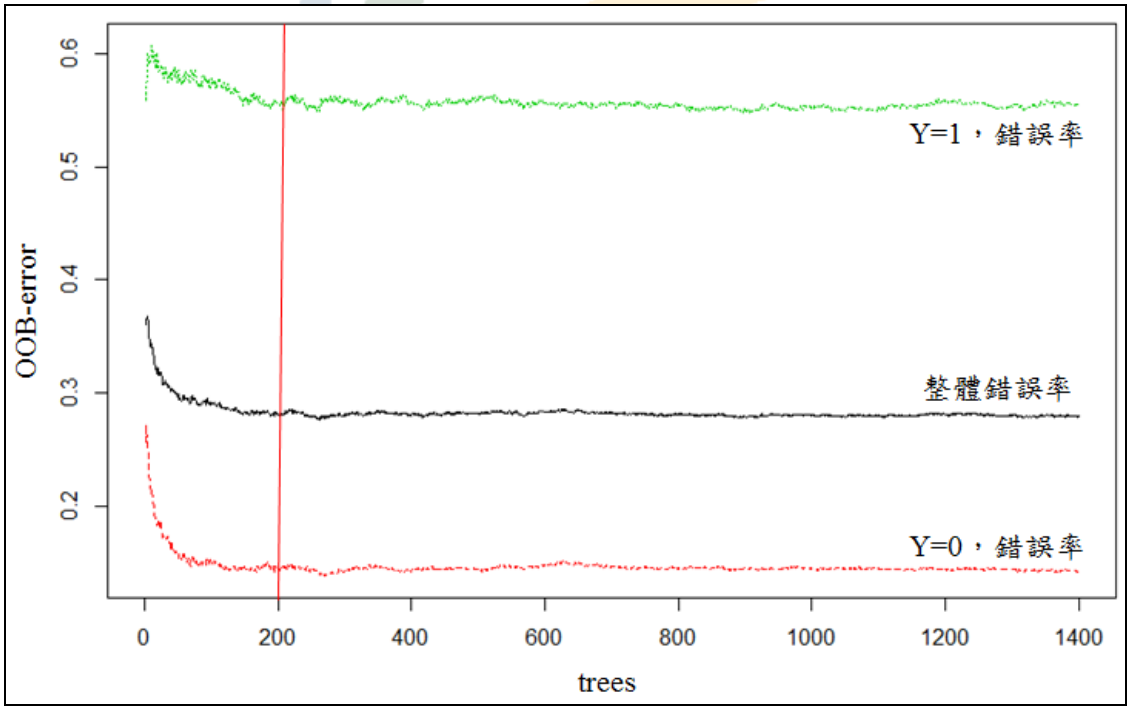

圖 11：袋外錯誤率與樹木數量關係圖

如表 14，ntree=200 後，OOB-error 趨於平穩，最後選定錯誤率最小的 ntree=600。再來是決定每次隨機子集合選取的特徵變數數量，如表 15 所示，最後選取錯誤率最小的 mtry=4，去建立隨機森林。

表 14：ntree 與 OOB-error

ntree	OOB-error
200	25.97
400	26.74
600	25.84
800	26.22
1000	26.20
1200	26.05
1400	26.00

表 15：mtry 與 OOB-error

ntree	mtry	OOB-error
600	1	30.64
600	2	27.22
600	3	26.57
600	4	25.84
600	5	26.70
600	6	26.27
600	7	26.66
600	8	27.26

前一章研究方法提及隨機森林對於變數重要性的判斷是使用 permutation test。如圖 12 所示，重要性前五高的變數依序為「電子帳單記號」、「年齡」、「居住地」、「卡齡」、「線上消費 RFM」。最後，我們使用混淆矩陣去判斷模型的準確率，如表 16 所示，無論在整體的準確度、靈敏度或特異性皆比決策樹要來的好。

靈敏度(Sensitivity，SEN)：代表正確預測為陽性的命中率，亦被稱為真陽性率。

$$SEN=\frac{368}{404+368}=47.67\% \tag{4.4}$$

特異性(Specificity，SPE)：代表正確預測為陰性的命中率，亦被稱為真陰性率。

$$SPE=\frac{1{,}346}{1{,}346+193}=87.46\% \tag{4.5}$$

準確度(Accuracy，ACC)：代表預測正確的準確度。

$$ACC=\frac{1{,}346+368}{1{,}346+193+404+368}=74.17\% \tag{4.6}$$

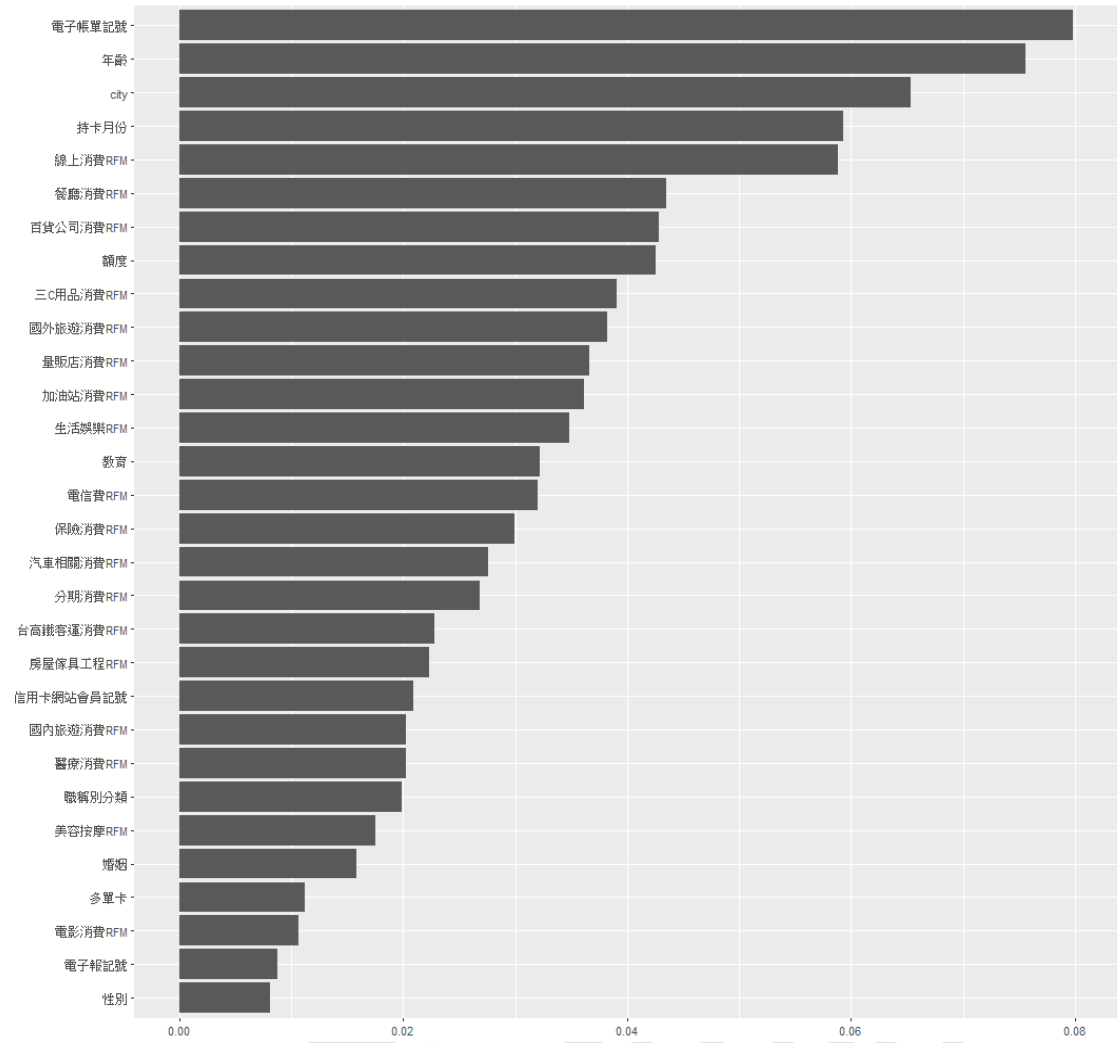

圖 12：變數重要性排序

表 16：隨機森林混淆矩陣

預測 / 實際	0	1
0	1,346(真陰性)	193(假陽性)
1	404(假陰性)	368(真陽性)

真陽性(True Positive，TP)：真實值是陽性且軟體預測輸出結果是陽性，正確。
真陰性(True Negative，TN)：真實值是陰性且軟體預測輸出結果是陰性，正確。
假陽性(False Positive，FP)：真實值是陰性且軟體預測輸出結果是陽性，不正確。
假陰性(False Negative，FN)：真實值是陽性且軟體預測輸出結果是陰性，不正確。

第五節 線性判別分析

本研究使用 R 軟體之 lda，採用費雪判別函數去分類行動銀行使用與否，其優勢在於對分布及方差並無任何限制。我們將 7,700 筆資料切分 70%為訓練資料，30%為測試資料。

判別係數如表 17，可組成判別函數 $y = w_1x_1 + w_2x_2 + ... + w_jx_j$ ，w 為係數權重，x 為判別變數，再以此函數將資料分類。「接受電子帳單」、「信用卡網站會員」、「年齡」、「接受電子報」為主要影響「是否會使用行動銀行」的變數。

表 17：判別係數表

判別變數	係數權重
接受電子帳單	1.3631
信用卡網站會員	0.5835
年齡	-0.0462
接受電子報	0.0287
醫療消費 RFM	0.0029
房屋傢具工程 RFM	0.0027
三 C 用品消費 RFM	0.0019
生活娛樂 RFM	0.0018
電信費 RFM	0.0018
汽車相關消費 RFM	0.0017
餐廳消費 RFM	0.0014
台高鐵客運消費 RFM	-0.0012
分期消費 RFM	0.0011
量販店消費 RFM	-0.0010
百貨公司消費 RFM	-0.0010
加油站消費 RFM	-0.0008
美容按摩 RFM	0.0007
電影消費 RFM	0.0005
國內旅遊消費 RFM	-0.0002
國外旅遊消費 RFM	-0.0001
保險消費 RFM	-0.0001
線上消費 RFM	0.0001
持卡月份	-0.0005
額度	0.0001

表 18：LDA 混淆矩陣

實際＼預測	0	1
0	1,292(真陰性)	247(假陽性)
1	377(假陰性)	395(真陽性)

真陽性(True Positive，TP)：真實值是陽性且軟體預測輸出結果是陽性，正確。
真陰性(True Negative，TN)：真實值是陰性且軟體預測輸出結果是陰性，正確。
假陽性(False Positive，FP)：真實值是陰性且軟體預測輸出結果是陽性，不正確。
假陰性(False Negative，FN)：真實值是陽性且軟體預測輸出結果是陰性，不正確。

靈敏度(Sensitivity，SEN)：代表正確預測為陽性的命中率，亦被稱為真陽性率。

$$SEN=\frac{395}{377+395}=51.17\% \tag{4.7}$$

特異性(Specificity，SPE)：代表正確預測為陰性的命中率，亦被稱為真陰性率。

$$SPE=\frac{1,292}{1,292+247}=83.95\% \tag{4.8}$$

準確度(Accuracy，ACC)：代表預測正確的準確度。

$$ACC=\frac{1,292+395}{1,292+247+377+395}=72.99\% \tag{4.9}$$

第六節 模型比較分析

模型效能比較，本研究使用機器學習最常使用的兩種方法：混淆矩陣、ROC 曲線。

一、混淆矩陣

在前兩節的決策樹分析中，我們已經介紹過混淆矩陣。簡單來說，就是拿預測與實際數據進行對比，可以得到準確度、靈敏度和特異性等指標對模型的分類效果進行度量。如表 19，隨機森林在整體的準確率為 74.17%，高出決策樹 2.6%，高出 LDA1.2%。LDA 在靈敏度的準確率較高，但在特異性的表現較差。

表 19：模型準確率比較表

混淆矩陣度量指標	隨機森林	決策樹	LDA
靈敏度	47.67%	44.17%	51.17%
特異性	87.46%	85.32%	83.95%
準確度	74.17%	71.57%	72.99%

二、ROC 曲線

接收者操作特徵（Receiver Operating Characteristic，ROC 曲線）是一種對於靈敏度進行描述的功能圖像。ROC 曲線可以通過描述真陽性率（TPR）和假陽性率（FPR）來實現。ROC 空間將 FPR 和 TPR 定義為 x 和 y 軸，這樣就描述了真陽性（獲利）和假陽性（成本）之間的博弈。而 TPR 就可以定義為靈敏度，而 FPR 就定義為「1-特異度」，最好的可能預測方式是一個在左上角的點，在 ROC 空間坐標軸(0,1)點，這個代表著 100%靈敏（沒有假陰性）和 100%特異（沒有假陽性）。而(0,1)點被稱為完美分類器。一個完全隨機預測會得到一條從左下到右上對角線上的一個點，例如：拋硬幣。而這條斜線將 ROC 空間劃分為兩個區域，在這條線的以上的點代表了一個好的分類結果，而在這條線以下的點代表了差的分類結果。AUC(Area Under Roc Curve) 的值就是處於 ROC 曲線下方的那部分面積的大小，較大的 AUC 代表了較好的 performance。

由圖 13 可得知，隨機森林的 ROC 曲線更靠近左上角的完美分類器，且 AUC 最高。線性判別分析的 sensitivity 高於隨機森林，specificity 低於隨機森林。決策樹模型效力皆為最低。

無論是混淆矩陣或是 ROC 曲線皆顯示隨機森林的模型效力最高，線性判別分析的 sensitivity 較好，決策樹的準確率最低。

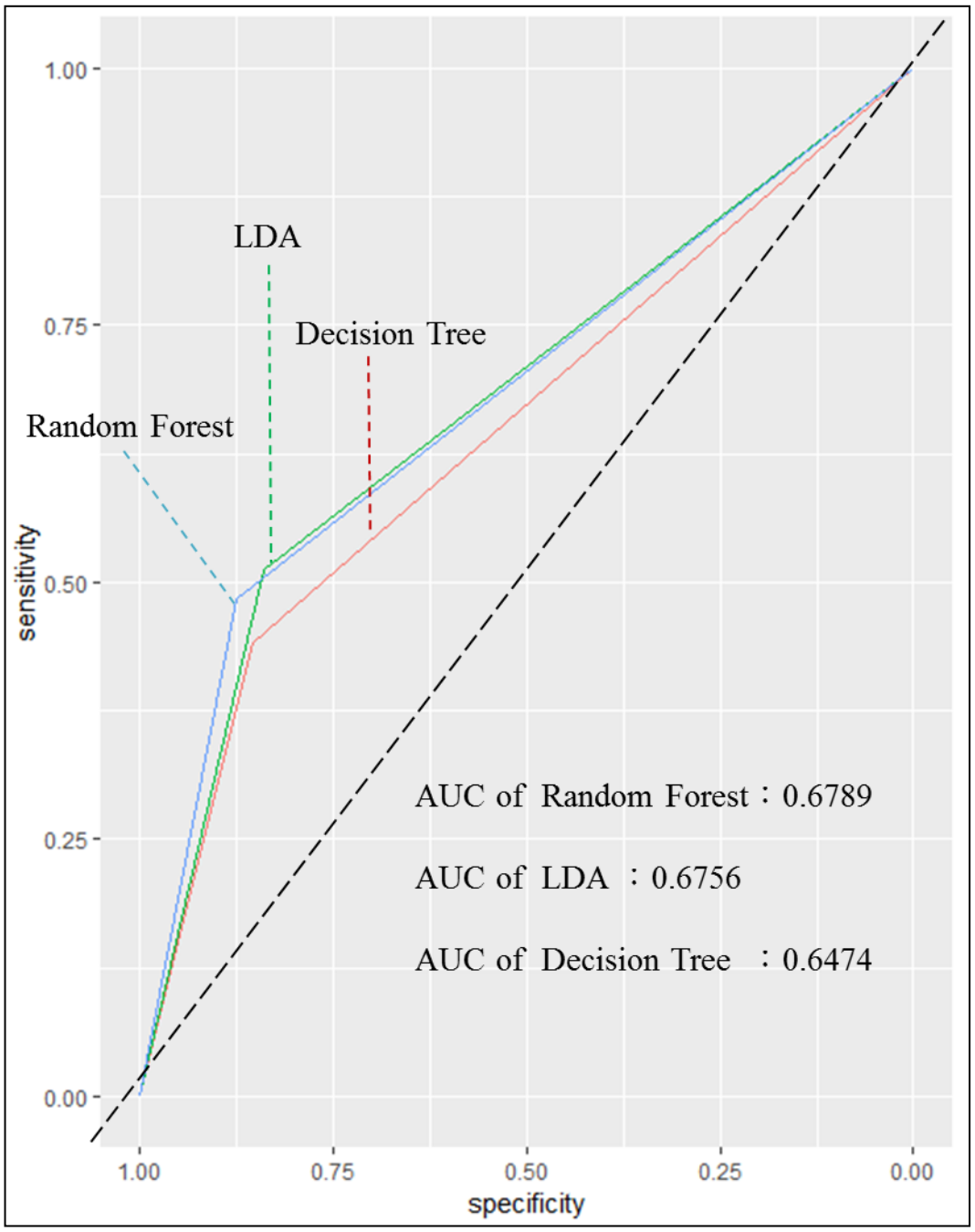
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
LDA
Decision Tree
Random Forest
AUC of Random Forest：0.6789
AUC of LDA ：0.6756
AUC of Decision Tree ：0.6474

圖 13：ROC 曲線比較

第五章結論與建議

第一節　結論

如果不講求解釋能力，只追求預測的準確度，那隨機森林會是一個很好的選擇。如果能接受少一點準確度卻多一點解釋能力，那決策樹的效果會比較好。若想使用線性函數分類，輕鬆計算出分類結果就選擇線性判別分析。

整體觀察下來，隨機森林確實是個不錯的分析模型。其一，若使用此模型也可不用考慮將資料分成訓練資料和測試資料，因為隨機森林所使用的 OOB-error 去評估錯誤率已十分精準；其二，當我們對於要分析的領域一無所知時，隨機森林可以處理眾多的特徵變數並找出影響力大的變數；其三，也是最重要的一點，它擁有較高的準確率卻不容易 overfitting。

使用決策樹，我們可以清楚明瞭的知道有四條規則，使用行動銀行的機率高於 50%：

(1) 規則 1：
寄送電子帳單且年齡≥44 歲且為信用卡網站會員且生活娛樂 RFM 指標≥31 分，占整體人數的 3%，使用行動銀行的機率為 59%。

(2) 規則 2：
寄送電子帳單且年齡<44 歲，占整體人數的 34%，使用行動銀行的機率為 55%。

(3) 規則 3：
寄送電子帳單且年齡<44 歲且為信用卡網站會員，占整體人數的 26%，使用行動銀行的機率為 58%。

(4) 規則 4：
寄送電子帳單且年齡<44 歲且為信用卡網站會員且線上消費 RFM 指標≥23 分，占整體人數的 22%，使用行動銀行的機率為 61%。

決策樹預測測試資料將有 24.53%的客戶會使用行動銀行，我們不難發現，其中電子帳單、年齡、信用卡網站會員、生活娛樂 RFM 及線上消費 RFM 為重要的特徵。而行動銀行的潛在使用者大致上可分為兩類：年紀較長者且 E 化程度高又偏向生活娛樂消費者、年紀較輕者且 E 化程度高又偏向線上消費者。年紀較輕且接受電子帳單者，占整體人數的 34%，使用行動行動銀行的機率為 55%；如再擁有信用卡網站會員身分，人數將下降 8%，但使用行動行動銀行的

機率上升 3%；如果線上消費 RFM 指標再超過 23 分，則人數再下降 3%，機率再上升 3%。

因此，銀行可利用這些規則，區分出四種類型客戶，對其分眾行銷。人數較多機率較小的客戶可透過 Email 行銷，不需花費過多成本；人數較少機率較高的客戶，可透過簡訊通知或專人聯絡，期望用較少的成本達到使用效益，做到精準行銷。

使用隨機森林，預測測試資料將有 24.23%的客戶會使用行動銀行，整體準確率高達 74.17%，也可知道以下這些變數為影響「是否使用行動銀行」的重要因素：電子帳單、年齡、居住地、卡齡、線上消費 RFM、餐廳消費 RFM、百貨公司消費 RFM。

使用線性判別分析，預測測試資料將有 27.78%的客戶會使用行動銀行，整體準確率高達 72.99%，以組成的判別函數將資料分類，而判別係數絕對值較高者為影響「是否使用行動銀行」的重要因素，依其重要性排序為：接受電子帳單、信用卡網站會員、年齡、接受電子報。

測試資料有 2,311 位客戶，隨機森林預測有 560 位會使用行動銀行，決策樹預測有 567 位會使用行動銀行，LDA 預測有 642 位會使用行動銀行。三種模型預測會使用行動銀行客戶的差異組合如表 20，1 代表是，0 代表否。三種方法皆預測會使用的人數為 370 位，占總人數 16%；隨機森林與決策樹皆預測會使用的人數為 370+45=415 位，占兩客群聯集的 58%；隨機森林與線性判別分析皆預測會使用的人數為 370+102=472 位，占兩客群聯集的 65%；決策樹與線性判別分析皆預測會使用的人數為 370+66=436 位，占兩客群聯集的 56%；因此我們知道使用隨機森林與線性判別預測使用者的重複比例最高。

表 20：三種模型預測行動銀行使用者差異組合

隨機森林	決策樹	線性判別	組合客戶數
1	1	1	370
1	1	0	45
1	0	1	102
0	1	1	66
1	0	0	43
0	1	0	86
0	0	1	104

本研究再以準確率最高之隨機森林，將回應率定義為「預測會使用行動銀行的機率」，觀察測試資料 2,311 位，回應率高於 50%的客戶差異，用以達到分眾行銷的效果。如表 21，回應率大致上與女性比例、信用卡網站會員比例、線上消費 RFM、三 C 用品消費 RFM 成正比，與年齡、額度、量販店消費成反比。依照各回應率級距，客戶特性的差異，我們可將行銷設計如下：

(1) 回應率 50%~55%：

191 位，占整體 34%，回應率相對較低，可透過 Email 行銷節省成本；而其客戶接近中年，信用卡額度較高，有點錢，量販店消費較高表示家庭花費較多，加油站消費較高表示時常需要用車，有 E 化意願但需要被促動，因此行銷內容可設計為：於期間內使用行動銀行，可獲得加油券或超級市場優惠券。

(2) 回應率 55%~60%：

170 位，占整體 30%，回應率尚可，可透過簡訊行銷，聯絡到本人的機率較 Email 高；而其客戶正值壯年，生活娛樂及餐廳消費較高，屬於有經濟基礎又懂得享受生活的階段，行銷內容可設計為：於期間內使用行動銀行，可獲得常去餐廳的優惠券。

(3) 回應率 60%~65%：

129 位，占整體 23%，回應率偏高，可透過 DM 專函寄送，聯絡到本人的機率較簡訊高；而其客戶為青壯年，E 化程度高，消費能力普通，花費在汽車相關及國外旅遊較高，行銷內容可設計為：於期間內使用行動銀行，可獲得國外旅遊租車折扣。

(4) 回應率 65%~100%：

70 位，占整體 13%，回應率最高，可透過 Email+簡訊+專人電訪行銷；而其客戶平均僅 31 歲，E 化程度很高，偏愛在網路及三 C 用品消費，行銷內容可設計的較為活潑生動：於活動期間內在購物或三 C 網站使用行動銀行轉帳付款者，可獲得該購物網站積分，積分可兌換網站商品。

表 21：隨機森林高回應率客戶差異

隨機森林回應率(%)	50~55	55~60	60~65	65~100
人數(人)	191	170	129	70
平均年齡(歲)	37	36	33	31
女性比例(%)	54	54	56	59
信用卡網站會員比例(%)	86	91	94	96
平均額度(元)	157,691	147,835	129,605	100,357
平均量販店消費 RFM(分)	27	24	18	13
平均汽車相關消費 RFM(分)	24	25	35	35
平均線上消費 RFM(分)	95	114	110	116
平均三 C 用品消費 RFM(分)	23	27	27	31
平均生活娛樂 RFM(分)	33	46	34	41
平均百貨公司消費 RFM(分)	38	38	37	34
平均加油站消費 RFM(分)	46	40	38	43
平均國外旅遊消費 RFM(分)	38	36	43	32
平均餐廳消費 RFM(分)	39	42	38	31

第二節 研究限制及建議

本研究使用 R 軟體之 RandomForest，要建立隨機森林就要建立非常多棵完整的決策樹，這樣很耗電腦效能，所以此套件會先評估使用者的設備，再決定是否給予建模。本研究因受限於個人電腦，只能被迫隨機抽取一成的資料來建立模型。雖然隨機森林的運算非常耗時，但在這個資料量越來越大的時代，誰都不願意丟失一點資料訊息。由於隨機森林是透過 bootstrap 獨立生成樣本資料再生長成多棵決策樹，所以我們可以利用多台電腦進行平行運算，就可解決此問題。

參考文獻

一、中文

丁振原(2003)，以資料探勘技術為基礎建構 PCB 客訴問題處理模式，元智大學工業工程與管理學系碩士論文。

白宸瑋(2009)，以資料探勘發掘行動加值服務市場之消費者特徵，長榮大學企業管理學系碩士論文。

吳正德(2004)，女性消費者購買行為與行銷策略之探討-以筆記型電腦為例，國立台北大學企業管理學系碩士論文。

吳昇洋(2004)，應用資料採礦技術評估客服中心顧客管理之績效，清華大學工業工程與工程管理學系碩士論文。

李章偉(2001)，資料庫行銷之顧客價值分析：以 3C 流通業為例，國立臺灣大學國際企業學研究所，碩士論文。

余豪(2011)，應用決策樹及類神經技術於車輛啟動系統之故障診斷，國立彰化師範大學車輛科技研究所碩士論文。

林郁珊(2012)，應用空載全波形光達資料於波形分析與地物分類，交通大學土木工程學系碩士論文。

林宸翊(2009)，應用於行為評等之 Random forests 及其變數選擇法，輔仁大學應用統計研究所碩士論文。

柳慧琴(1997)，資料庫行銷之顧客價值分析模式，國立暨南國際大學國際企業學系研究所碩士論文。

徐維志(2015)，以隨機森林為模式之美金/歐元匯率交易預測研究，輔仁大學統計資訊學系應用統計碩士論文。

陳宜欣(2006)，以資料探勘技術探討顧客忠誠方案-以某信用卡發卡銀行為例，國立中正大學行銷管理研究所碩士論文。

陳俊成(2002)，以實驗法探討網路互動對關係品質之影響-顧客關係管理觀點，國立屏東科技大學工業管理系碩士論文。

陳時仲(2015)，隨機森林模型效力評估，國立交通大學統計學研究所碩士論文。

陳麗秋(2013)，行動銀行系統成功模式之研究：以臺東地區行動銀行使用者為例，臺東大學資訊管理學系環境經濟資訊管理碩士論文。

郭律君(2015)，消費者對綠色產品的態度分析-以宏碁筆電消費者為例，國立台北大學統計學系碩士論文。

許智宇(2010)，整合 KMV 模型、約略集合及隨機森林應用於企業信用評等之研究，臺北科技大學商業自動化與管理研究所碩士論文。

張筱珍(2016)，消費者選擇行動銀行關聯因素探討，東吳大學國際經營與貿易學系碩士論文。

葉建良(2006)，利用 CART 分類與迴歸樹建立消費者信用貸款違約風險評估模型之研究-以國內 A 銀行為例，輔仁大學應用統計研究所碩士論文。

馮淑群(1998)，傳銷公司顧客價值分析與產品組合策略之研究，國立臺灣大學國際企業學研究所碩士論文。

湯勝隆(2014)，銀行財富管理客戶往來行為分析-多變量分析方法之應用，國立台北大學統計學系碩士論文。

黃靖紋(2014)，應用 RFM 顧客終身價值技術探討網路團購公司顧客消費型態-A 公司個案研究，致理技術學院企業管理系服務業經營管理碩士論文。

蔡韻菱(2013)，利用決策樹分析於六標準差管理架構以降低顧客抱怨-以台灣某服務業為例，輔仁大學國際經營管理碩士班碩士論文。

蕭宇伶(2016)，從交易成本觀點探討消費者使用行動銀行意圖，臺中科技大學企業管理系碩士論文。

盧瑜芬(2006)，使用三種資料探勘演算法-類神經網路、羅吉斯迴歸及決策樹-預測乳癌患者存活情形之效能比較，國防醫學院公共衛生學研究所流行病學組碩士論文。

魏瑞慧(2008)，層級貝氏模型應用於信用卡顧客消費行為分析之研究，國立台北大學統計學系碩士論文。

顏利憲(2013)，以決策樹分析鐵路誤點原因及解決方法，國立成功大學交通管理科學研究所碩士論文。

二、英文

Austen Mulinder (1999), “Hear Today ... or Gone Tomorrow? Winners Listen to Customers”,Texas A&M University, Vol. 11, No. 5.

Berry, L. L. and Parasuraman, A.(1991).The Free Press,Mareting Services:Competing Through Quality,New York.

Berry, M. J. A. & Linoff, G.,(1997). Data Mining Techniques: for Marking, Sales, and Customer Support. New York: John Wiley & Sons Inc.

Bucklin, Randolph E, Gupta,Sunil, and Siddarth,S. (1998), “Determining Segmentation in Sales Response Across Consumer Purchase Behaviors,” Journal of Marketing Research, 35, 2, pp.189-197.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.(1984), Classification and Regression Trees, New York, Chapman & Hall.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), pp.123-140.

Breiman, L. (2001). Random Forests. Machine learning, 45(1), pp.5-32.

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. Machine learning, 48(1-3), pp.287-297.

Cabena, P., Hadjinian, P. O., Stadler, R., Verhees, J. & Zanasi, A. (1997). Discovering Data Mining From Concept to Implementation. New Jersey: Prentice-Hall, Inc., pp.41-59.

Colombo, Richard and Jiang, Weina (1999), “A Stochastic RFM Model,” Journal of Interactive Marketing, 13, pp.2-12.

Fayyad, U. M., Shapi, G. P., Smyth, P. & Uthursamy, R. (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press/The MIT Press.

Fisher, R. A.(1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, pp.179-188.

Guo, L., Chehata, N., Mallet, C., Boukir, S., (2011). Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing, 66(1): pp.56-66.

Hughes, A. M.(1994). Strategic Database Marketing(4 ed.): McGraw-Hill.

Kahan, R.(1998). Using Database Marketing Techniques to Enhance Your One-to-One Marketing initiatives. Journal of Consumer Marketing, 15(5), pp.491-493.

Kalakota, R.,& Robinson, M. (1999). "Customer Relationship Management: IntegratingProcesses to Build Relationships",E-Business: Roadmap for Success, pp.109-135.

Kalakota, R., & Robinson,M. (2001)."E-business 2.0: Roadmap for Success",Addison -Wesley Professional.

Liang, T. P. and Huang, J. S. (1998). An empirical study on consumer acceptance of products in electronic markets: A transaction cost model.Decision Support Systems, pp.24, pp.29-43.

Lichung, Chien-Heng Chou and Allenby, Greg M. (2003) "A Bayesian Approach to Modeling Purchase Frequency," Marketing Letters, 14, 1, pp.5-20.

Miglautsch, J. R.(2000). Thoughts on RFM Scoring. Journal of Database Marketing, 8, pp.67-72.

Peacock, P. R. (1998). "Data mining in marketing: Part1," Marketing Management, 6 (4), pp.8-18.

Pal, M., & Mather, P. M.(2003),"An assessment of the effectiveness of decision tree methods for land cover classification" , Remote Sensing of Environment, 86(4) pp.554-565.

Quinlan, J. R.(1993),C4.5: programs for machine learning, Morgan:Kaufmann Publishers Inc.

Stone, B.(1995). Successful Direct Marketing Methods (Fourth ed.): McGraw-Hill/ Contemporary.

Tan, P., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining(1st ed.). Hoboken, N. J. : Addison-Wesley.

Teo, T. S. H. and Yu, Y. (2005). Online buying behavior: A transaction cost economics perspective. Omega, pp.451-465.

Williamson, O. E.(1975). Markets and Hierarchies: Analysis and Antitrust Implications. New York: Free Press.

著作權聲明

論文題目：顧客消費行為分析及行動銀行使用預測-決策樹、隨機森林與判別分析之比較

論文頁數：52 頁

系所組別：統計學系

研究生：葉子維

指導教授：許玉雪

畢業年月：107 年 6 月