出國報告(出國類別：進修)

赴美國加州大學爾灣分校進修
「數據科學及預測分析」
(Data Science & Predictive Analytics)

服務機關：國家發展委員會
姓名職稱：林奎后專員
派赴國家：美國
出國期間：107 年 4 月 2 日至 6 月 21 日
報告日期：107 年 9 月

摘要

為強化本會對數據科學在總體規劃應用與潛力的研究能量，本次赴美國加州大學爾灣分校參與「數據科學及預測分析」專業證書課程。該校在數據科學領域頗富名聲，校園內建有「UCI Machine Learning Archive」，為國際間機器學習的重要資料庫，另師資學經歷豐富，均具有長期實務經驗，每年均吸引眾多學生前往就讀。

本此進修課程中，充分學習到數據科學及預測分析的基礎理論與實務應用的最新進展。在基礎理論方面，透過數學模型的推導，理解迴歸、k 近鄰法、單純貝氏法等演算法的運作原理，亦瞭解常用統計評估指標的概念與採用時機；在實務應用方面，透過操作專業統計軟體 R，及商用智慧軟體 KNIME 與 Tableau 等，並依「跨行業數據挖礦標準程序」，學習如何進行數據瞭解、數據準備與模型建置等工作，並提出可行之因應對策。

目次

第一章 目的

壹、本案緣起

數位經濟時代，數據科學在政府總體經濟規劃的角色益顯重要，為當前全球關注議題。數據科學對數據源的整合運用，有助提升總體規劃品質及決策視野。近年來，政府積極運用公務統計，推動大數據分析與應用，揭示資料經濟(Data Economy)的重要性。例如：財政部的財稅大數據、衛福部的健保大數據等。

為因應數據科學對傳統總體計量模型的衝擊與挑戰，本處將進行大數據納入經濟預測模型的相關研究，以提升總體經濟應用評估層面與價值。另數據科學在國發計畫編擬的應用，亦強化複雜總經議題的視覺化效果，提高國發計畫的可讀性與可近性。準此，本項進修與研究項目與本會國家發展業務直接相關。

貳、學校簡介

美國加州大學爾灣分校(UC Irvine)為全美排名前十公立大學，與同屬加州大學的柏克萊分校(UC Berkeley)、洛杉磯分校(UCLA)齊名。加州大學爾灣分成立於 1965 年，校風開放且注重創新，學術成就及研究能量亦相當充沛，創校至今已有 3 位學者獲頒諾貝爾學獎，顯見其年輕且旺盛之活力。

加州大學爾灣分校提供 192 個學位課程，計有超過 30,000 名學生。進修教育部門(Division of Continuing Education)設有國際課程(International Programs)，提供國際學生進修管道。本次參與 2018 年春季班的「專業證書課程(Accelerated Certificate Programs)」，本期共計 181 位學員，主要來自法國、日本、巴西等；課程為「數據科學與預測分析(Data Science and Predictive Analytics)」，學員來自巴西、南韓、日本與我國等共計 10 名，背景十分多元，包括新創事業家、資通訊高階行銷主管、醫生等。

第二章 過程

壹、課程大綱

隨著全球進入大數據時代，各界紛紛採用數據科學與預測分析的研究成果，來制定發展戰略與預測未來發展。在此一趨勢下，加州大學爾灣分校積極強化此領域的研究能量，建構「UCI Machine Learning Archive」，並聘請重量級學者(如預測分析開創者 Eric Siegel 教授等)；另為培育數據科學與預測分析的人才，開設「數據科學與預測分析」專業證書課程，共計 7 門科目，包含：

1. **使用預測分析的商務戰略分析**(Strategic Business Analysis using Predictive Analytics)
2. **預測分析應用**(Applications of Predictive Analytics)
3. **數據挖礦的數據準備**(Data Preparation for Data Mining)
4. **預測模型的建置、部署與改進**(Modeling Methods, Deploying, and Refining Predictive Models)
5. **數據科學的商業應用**(Business Applications of Data Science)
6. **數據探勘、分析與視覺化**(Data Exploration, Analytics, and Visualization)
7. **大數據視覺化與分析**(Big Data Visualization and Analytics)

貳、課程內容

一、使用預測分析的商務戰略分析(Strategic Business Analysis using Predictive Analytics)

本課程授課老師 Ash Pahwa 博士，擁有超過 30 年的學術研究與業界工作經驗，專長包括搜尋引擎最適化、網頁分析與設計、數位圖像處理、資料庫管理、數位影音及資料存儲技術。曾任職於奇異(General Electric)、AT&T 貝爾實驗室(AT&T Bell Laboratories)、全錄公司(Xerox Corporation)及甲骨文(Oracle)，亦在加州理工學院(California Institute of Technology)、加州大學洛杉磯分校(UCLA)、聖地牙哥分校(UC San Diego)授課。

課程內容主要為**數據科學(Data Science)的基礎入門講解**。數據科學是指透過各種分析方法(analytics)，從大量的結構化及非結構化數據(structured and unstructured data)中擷取資訊的研究領域。傳統之統計數據多屬結構化數據，但隨著數位時代的快速演進，非結構化資料(如文字、照片、影片等)大量產生，

透過觀察新型態資料，有助瞭解人類消費、投資等各種經濟行為。

(一)數據科學的名稱與常用工具

近年來數據科學已被廣泛應用在不同領域，如資料挖礦(data mining)、機器學習(machine learning)，以及預測分析(predictive analytics)等，三者都是數據科學一部份，但強調重點稍有不同。

－**資料挖礦**：強調運用分析工具，瞭解資料的態樣(pattern)。

－**機器學習**：強調透過模型的建構與分析，歸納資料的特性。

－**預測分析**：強調運用分析方法，針對給定的新資料進行預測，以解決商業問題。

數據科學與預測分析常用的重要軟體有三：**R、Python、KNIME**，三者均可免付費下載。R 為專業統計軟體，已廣泛應用在預測分析；Python 由 Google 開發並大力推廣的程式語言；KNIME 為近期開發之互動式預測分析軟體，因操作較為容易且直觀，使用人數成長快速。

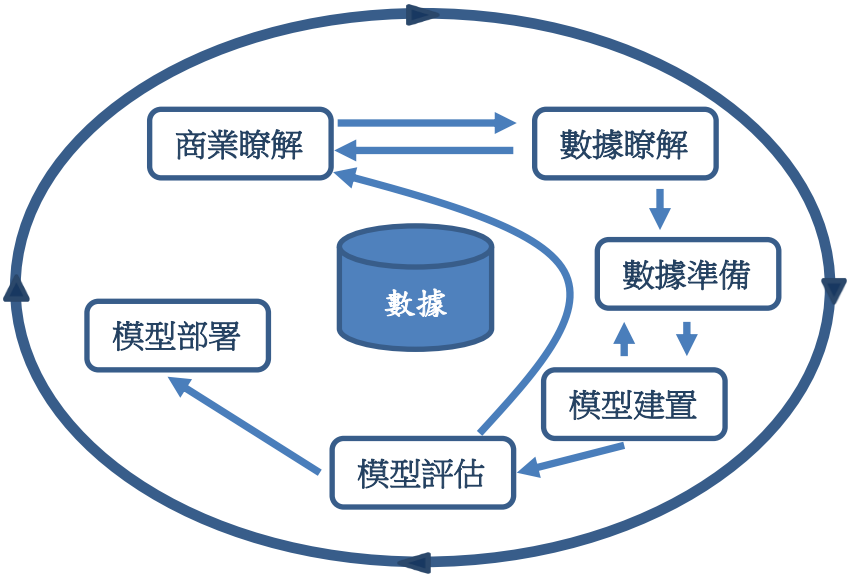
(二)預測分析方法與目標

數據科學與預測分析領域中具有多種分析方法，其中較廣泛使用的是「**跨行業數據挖礦標準程序**(cross-industry standard process for data mining，縮寫為 CRISP-DM)」，包括 6 項步驟：

1. **商業瞭解**(business understanding)；
2. **數據瞭解**(data understanding)；
3. **數據準備**(data preparation)；
4. **模型建置**(Modeling)；
5. **模型評估**(Evaluation)；
6. **模型部署**(deployment)。

步驟間的流程並非單向固定，必要時可在步驟間來回移動，進行相互驗證，以確保結果的有效性。

圖 1　跨行業數據挖礦標準過程(CRISP-DM)流程圖



預測分析目標可分為**預估**(estimation)與**分類**(classification)。預估是指分析目標為數值(number)，如房價、次季經濟成長率等；分類則是指分析目標為某類別(category)，如電子郵件是否為垃圾郵件等。

(三)**數據的特性與準備**

數據是數據科學與預測分析中最重要的基礎。數據務求可靠(reliable)及有效(valid)，但仍常發生數據遺漏(missing data)、數據不足(sparse data)、數據不正確(inaccurate data)、數據不一致(inconsistent data)、數據過剩(redundant data)等情況。為提高資料品質，須進行**數據準備**(data preparation)，4 項重要步驟如下：

1. **特性選取與數據移植**(feature extraction and data portability)：選取具備可預測性的應變數(dependent variable)與相互獨立的自變數(independent variable)等；並對變數進行移植，如將數字變量透過離散化(discretization)，轉換為分類變量。
2. **資料清理**(data cleaning)：對數據遺漏、數據不正確、數據不一致、數據過剩等問題進行處理。
3. **資料轉換**(data transformation)：對變數取對數或倒數等。
4. **資料縮減**(data reduction)：對樣本數進行分群(sampling)，分為訓練樣本

(training sample)與測試樣本(testing sample)。

(四)預測分析的應用

根據 E. Siegel (2013)研究，預測分析應用的九大領域包括：

1. **家庭與個人生活**(family and personal life)；
2. **行銷、廣告及網頁**(marketing, advertisement, and the web)；
3. **金融風險與保險**(financial risk and insurance)；
4. **健康照護**(healthcare)；
5. **執法與詐欺識別**(law enforcement and fraud detection)；
6. **錯誤檢測、安全與物流效率**(fault detection, safety and logistical efficiency)；
7. **政府、政治、非營利組織與教育**(government, politics, nonprofit, and education)；
8. **人類語言理解、思維與心理**(human language understanding, thought, psychology)；
9. **勞動力：員工與僱員**(workforce: staff and employees)。

二、預測分析應用(Applications of Predictive Analytics)

本課程授課老師 Robert Nisbet 博士為資料科學家，亦為專業經理人，曾多次帶領團隊為零售、銀行、電信業等提供大數據解決方案，其主筆之「統計分析與數據挖礦手冊(Handbook of Statistical Analysis & Data Mining Applications)」，曾獲得美國書界奧斯卡之稱的美國出版協會 2009 年獎項，為預測分析重量級人物。

課程內容主要是透過 **KNIME 軟體**的實際操作，深入介紹**跨行業數據挖礦標準程序及其應用**。預測分析的主要目的在於，透過機器學習的演算法找尋資料間的關聯態樣(pattern of relationship)，以作為決策時的重要依據。機器學習與傳統參數統計(parametric statistics)的運作原理相似，但對於資料的標準化處理及資料間獨立性的假設較寬鬆，因此可應用範圍更為廣泛。

(一)基礎演算法

由於機器學習領域發展快速，各類演算法不斷推陳出新，且應用領域各有不同。歸納而言，重要的基礎演算法有三：

1. **決策樹**(decision tree)：使用樹狀結構來分析資料，每個節點表示一個評估欄位，而每個分支代表一種可能的欄位輸出結果，而每個最終節點則是不同的類別標記。常用的決策樹演算法有隨機森林(random forest)及樹集成(tree ensemble)。
2. **神經網路**(neural networks)：模擬人類大腦神經元的運作方式來分析資料，將變數中每筆資料視為對神經元的一個輸入值(input)的訊號，透過特定函數的運算，形成輸出值(output)，再傳入下一個神經元，直到最後一層形成最後結果。
3. **支持向量機器**(Support Vector Machine, SVM)：運用統計學習理論(statistical learning theory)及複雜的數學函數，在高維度的資料空間中，找尋非線性的超平面(hyper-plane)，以達成困難的分類工作。常見的支持向量機器演算法有 LIBSVM 等。

(二)**演算法的評估**

由於演算法並無適用準則，必須透過多次反覆的試誤(try and error)，輔以基礎統計指標，檢視及評估各演算法的配適性。

1. **R 平方**(R-square)：若變數的原始資料或經資料處理後符合常態分配，則可觀察該演算法的 R 平方值，該值越高，意謂該演算法所得分析結果越佳。
2. **平均絕對誤差**(Mean Absolute Error)：若變數的原始資料不符合常態分配，則無法採信 R 平方值，必須改採平均絕對誤差評估，該值較低，意謂該演算法所得分析結果較佳。

(三)**演算法的應用**

1. **消費者購物籃**(market basket)：為預測分析的早期著名應用之一，透過分析消費者購物行為，改變商品陳列方式，以刺激消費，經典案例如啤酒與尿布。
2. **客群流失**(churn)：最早應用於電信業，評估公司客戶是否可能轉換至其他電信公司。目前以廣泛應用在其他非營利事業，例如：學校評估學生是否可能輟學或公益團體評估捐款人是否持續捐款等分析。
3. **科學**(science)：機器學習可協助進行專業的科學探索。例如應用機器學習

演算法分析先進的衛星與攝影設備拍攝的高解析度衛星圖片，以深入研究與分析地表森林種類。

4. **醫療與健康照護**(medical and healthcare)：機器學習透過分析醫療相關資料庫，提供醫療院所或健康照顧機構更客觀的分析，減少依賴對人為經驗與判斷，提高診斷準確度。
5. **社群媒體**(social media)：社群媒體的蓬勃發展帶來大量人們情緒的重要訊息，例如：證券業運用情緒分析(sentimental analysis)，預測股市為牛市(bullish)或熊市(bearish)等。
6. **運動**(sports)：機器學習目前已廣泛應用於棒球、籃球等各類運動，透過分析球員過去表現預測可能的明日之星，協助補強陣容，提升球隊整體表現。
7. **休閒**(entertainment)：最廣為人知的應用案例為人知的是美國線上影音平台 Netflix 從消費者在搜尋引擎(search engine)上的搜尋歷程，分析其喜好並推薦下一部電影。
8. **保險**(insurance)：此類應用亦稱為關聯銷售分析(cross-sell analysis)，主要是從現有資料庫中，找尋高關聯度的保險產品，以增加銷售及營收。

三、數據挖礦的數據準備(Data Preparation for Data Mining)

本課程授課老師為 Robert Nisbet 博士，使用「統計分析與數據挖礦手冊」一書及 KNIME 軟體，介紹與實際操作介紹**跨行業數據挖礦標準程序**(CRISP-DM)中的**數據準備**相關工作。數據準備是預測分析中最為耗時的工作，平均而言占整體工作時間至少六成，最多可能高達 90%，原因在於現行資料庫眾多，且使用者的方式與習慣亦不相同，造成資料格式相當不一致。為確保資料與分析結果的品質，必須統整數據資料，使演算法可正確判讀。大致而言，數據準備有下列重要步驟：

(一)資料整併(data integration)

在取得各個資料庫的資料後，首先必須進行的就是資料的整併，原則在於必須依預測分析的目的及需要，盡量維持資料的完整性與可利用性。因此，大多數的資料整併過程是以一個主要資料庫作為中心，再將其他資

料庫資料的交集部分進行整併，其他情況則可能採單純的交集或聯集。資料整併後，必要時可再透過行列互換(transposition)、交叉表(cross tabulation)或樞紐分析(data pivoting)等，進行資料的爬梳工作。

(二)**資料清理**(data cleansing)

資料整併完成後，必須針對明顯的錯誤資料進行調整，即資料清理(data cleaning)工作。首先必須先觀察同一變數內是否有人為錯誤或其他錯誤，造成資料數值的不合理或資料格式的不一致等問題，進行資料重編碼(recoding)，例如郵遞區號變數應為正數且應統一為五碼等。另外，若變數內若有離群值(outliers)，則可直接剔除離群值，或透過適當數學公式予以調整，如用變數的平均數或中位數取代離群值等。

(三)**遺漏值填補**(missing value imputation)

由於機器學習的演算過程是採取逐筆(case-by-case)計算，遺漏值的存在將會大幅降低演算法的預測精準度，因此遺漏值的填補是相當重要的步驟。常見方式有三：

1. 用常數(constants)填補，如變數的平均數或中位數。
2. 利用數學公式(formula)，以資料內其他變數資料計算填補值。
3. 利用演算法模型(model)推估填補值。

(四)**特殊變數推導**(deriving special variables)

由於資料庫內部變數的蒐集目的與使用資料的研究目的多有差距，因此在多數情況下，自行運用資料內變數再推導的解數變數，多對目標變數能有更好的解釋能力。

1. 若模型採取回歸演算法，則必須將資料內的分類變數轉為虛擬變數(dummy variables)或數值(number)，必要時需進一步標準化(standardize)變數。
2. 若變數數值變動區間過大，可能造成模型解釋效果不佳，原因在於該解釋變數的區間值對目標變數更具有解釋力，因此可透過離散化將數值轉換為分類變數。
3. 若變數為時間序列資料，可推導滯後變數(lag variables)，並納入模型中，以提高模型解釋力，此概念與傳統參數統計分析的 ARIMA 模型

(Autoregressive Integrated Moving Average)相似。

(五)**變數挑選**(feature selection)

對於機器學習而言，模型中的解釋變數數量並非越多越好，而是有一最適個數，才能最適化模型的預測能力。因此，在完成所有變數資料的調整與推導後，必須進行變數的挑選。理論上，剔除對目標變數影響較低的解釋變數為重要準則，但實務上不易執行，需謹慎評估。評估的方法大致有二：第一是觀察每個變數及的統計特質，剔除具低變易數(low variance)或是與其他變數具高相關性(high correlation)的變數；第二則是利用軟體中的相關預設模型直接進行挑選。

(六)**演算法挑選**(algorithm selection)

如何挑選適當的演算法，以符合資料型態與研究目的是資料準備的重要關鍵步驟。主要常用的演算法及其適用情況有三：

1. 邏輯迴歸(logistic regression)：適用於當解釋變數資料為數值型態，且符合常態分配，資料挖礦的目的在於釐清目標變數與解釋變數間的關係。
2. 神經網絡：同時適用於目標變數是數值或是分類變數，惟其背後演算邏輯複雜，如同黑盒子(black box)，無法進一步瞭解目標變數與解釋變數間的關係。
3. 決策樹：適用於同時適用於目標變數是數值或是分類變數，但與神經網路相比，其背後演算邏輯具可解釋性，且可用解釋變數的不同數值描述目標變數，已被商業領域廣泛應用。

(七)**資料調節**(data conditioning)

在進行演算法之前，必須進行資料分割(data partitioning)及數據平衡(data balancing)兩大工作：

1. 資料分割：將資料分為建模資料(modeling data)及測試資料(testing data)，其中建模資料可在分為訓練資料(training data)及驗證資料(validating data)。其流程為先用訓練資料建立模型，再用驗證資料調整模型，最後用測試資料設算模型準確性。
2. 數據平衡：若目標變數為分類變數，且資料出現某一結果出現次數多過另一結果的情況，則必須進行數據平衡，以提升演算法預測能力。平衡

方式有減少多數法(under-sampling)及增加少數法(over-sampling)兩種。

(八)**初步模型建立**(preliminary model building)

當資料已全數調整完成，且已挑選可能適當的演算法後，即可建立多組的初步模型，並可從中挑選預測能力最佳的模型。值得一提的是，在資料準備的步驟中，並非是單向直線的依序進行，實務上往往需要重回先前步驟，不斷地檢視、調整與篩選，才能完備整個資料準備的過程。

四、預測模型的建置、部署與改進(Modeling Methods, Deploying, and Refining Predictive Models)

本課程授課老師為 Ash Pahwa 博士，透過推導數學模型並實際操作 R 及 Python，介紹**跨行業數據挖礦標準程序**(CRISP-DM)的**模型建構**(Modeling)相關基礎概念與應用。

(一)模型分類

常見的機器學習演算法包含**迴歸分析**(regression)、**k-近鄰法**(k nearest neighbor)、**單純貝氏法**(Naïve-Bayes)、**群聚法**(clustering)及**神經網路**(neural networks)，並可依下列特性分類：

1. **反應變數**：可分為**數值**(numerical)或**類別**(categorical)變數。
2. **學習方式**：可分為兩類。
   －**監督式**(supervised)為預估或分類某特定目標；
   －**非監督式**(unsupervised)為找尋存在於母體的既有特性。
3. **學習策略**：可概分為四類。
   －**基於誤差的學習**(error based learning)；
   －**基於相似性的學習**(similarity based learning)；
   －**基於機率的學習**(Probability based learning)；
   －**模仿人類大腦**(mimicking the human brain)。

表 1：機器學習演算法分類

演算法	反應變數	學習方法	學習策略
迴歸	數值變數	監督式	基於誤差的學習
k-近鄰法	類別變數	監督式	基於相似性的學習
單純貝氏法	類別變數	監督式	基於機率的學習
群聚法	-	非監督式	-
神經網路	數值變數 類別變數	監督式	模仿人類大腦

(二)**模型介紹**

1. **迴歸**(regression)

迴歸一種廣泛使用的資料分析方法，目的在於建立反應變數與預測變數間線性關係的模型，其數學一般式如下：

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n$$

迴歸採監督式學習方法，反應變數為數值，預測變數多為數值，亦可為類別，但需轉為虛擬變數(dummy variables)，以納入模型。由於在推估模型的線性關係時，即推導截距項(β_0)與各預測變數的係數(β_1、β_2、…、β_n)，採用最小平方法(ordinary least square)，讓模型的預測數值與真實資料的誤差平方項總和為最小，因此稱為基於誤差的學習策略。

2. **k-近鄰法**(k nearest neighbor)

k-近鄰法是一種相當直觀的資料分類方法，透過計算欲分類的資料與現有分類資料的距離，並觀察其最靠近 k 個鄰居的多數屬性，藉以推論其屬性。計算距離的常見公式有二：

$$Euclidean\ distance=\sqrt{\sum_{i=1}^{k}(p_i-q_{1i})^2}$$

$$Manhattan\ distance=\sum_{i=1}^{k}|p_i-q_{1i}|$$

k-近鄰法為監督式學習，反應變數為類別變數，並透過計算計算兩點間的距離來觀察其相似性，因此稱為基於相似性的學習策略。k-近鄰

法的優點在於簡單有效，且可彈性的選擇 k 值，惟實務上，k 值多以資料總數之平方根做為基準。

3. **單純貝氏法**(Naïve-Bayes)

單純貝氏法是另一種資料分類方法，以貝氏定理(數學公式如下)為基礎，透過機率的計算，用以判斷未知類別的資料應該屬於哪一個類別。

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

單純貝氏法為監督式學習，反應變數為類別變數，透過原有資料計算機率，即P(A)、P(B)、P(B | A)，再推估欲特定事件下欲觀察事件的機率，即P(A | B)，據以判斷屬於哪一種類別，因此稱為基於機率的學習。此外，在機率推估過程中，若特定事件為兩件以上，模型僅簡單假設特定事件均為獨立事件，因而稱為單純。雖然假設較為簡化，但單純貝氏法仍是相當重要且計算上相當有效率的分析方法。

4. **群聚法**(clustering)

群聚法是將資料分類的方法，為非監督式學習，即僅就資料進行分群，而不針對任何未知類別的資料進行識別。常用的群聚法有二：**k 平均數群聚法**(k-means clustering)以及**階層群聚法**(hierarchical clustering)，兩者最大的差異在於前者是先給定欲分群的數量後再行分群，後者則是新進行分群再決定欲分的群數。雖然方法有所差異，但目標均為極小化類內變異(within-cluster-variation)，數學公式如下：

$$minimize \sum_{1}^{k} W(C_k)$$

$$W(C_k) = \frac{1}{|C_k|} \sum \sum \left(x_{ij} - x_{i'j}\right)^2$$

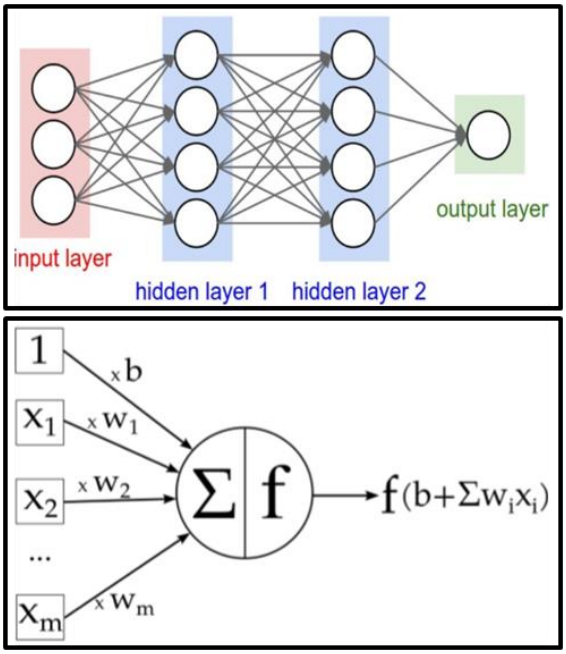
實務上，群聚法的用途在於可將資料先進行分類，再運用迴歸等其他演算法進行預測或再分類。集群後的資料同質性較高，且排除離群值的干擾，可有效提升模型準確度。

5. **神經網路**(neural networks)

神經網路為近期成長相當快速的監督式演算法，其特點在於可以

進行數值的預測，也可以進行資料的分類。模型背後的運作原理是一種模仿人類神經網路的結構與功能所產生的數學模型，首先將資料放置在輸入層(input layer)，再透過隱藏層(hidden layer)進行加權計算與函數轉換，最後在輸出層獲致模型結論。

圖 2　神經網演算流程及數學模型



隨著神經網路的應用不斷趨向複雜化，隱藏層層數隨之增加的結果，即是深度學習(deep learning)，而深度學習的主要應用也都在於圖像辨識與分類、字跡辨識或聲音識別等高難度工作。

(三)**模型評定**(assessment)

模型評定是指資料透過模型計算後，必須針對模型的預測或分類能力進行客觀的評估。常見的**預測能力評估**可觀察 **R 平方值**(R square)與**均方根誤差**(Root Mean Square Error, RMSE)；**分類能力評估**則為**混淆矩陣**(confusion matrix)與 **ROC 曲線**(receiver operating characteristic curve)以及**曲線下面積**(Area Under Curve, AUC)。

1. **模型預測能力評估**：概念為計算反應變數預測值與實際值間的差距，主要指標有二。
   (1) **R 平方值**：依據數學公式(如下)，值介於 0 與 1 之間，越接近 1 代表模型預測能力越佳。

$$r^2 = 1 - \frac{\sum(f(x_i) - y_i)^2}{\sum(y_i - ybar)^2}$$

(2)**均方根誤差**：依據數學公式(如下)，值越小代表型預測能力越佳。

$$RMSE = \sqrt{\frac{\sum(f(x_i) - y_i)^2}{N}}$$

2. **模型分類能力評估**：概念為觀察反應變數預測分類與實際分類間的異同，主要使用統計圖表有二。

(1)**混淆矩陣**：又稱為**誤差矩陣**(error matrix)，並可計算**準確性**(accuracy)、**敏感性**(sensitivity)及**特異性**(specificity)。其中，準確性顯示模型的整體分類能力、敏感性顯示模型正確預測為陽性的能力、特異性則顯示模型正確預測為陰性的能力。三者越高，代表模型分類能力越佳。

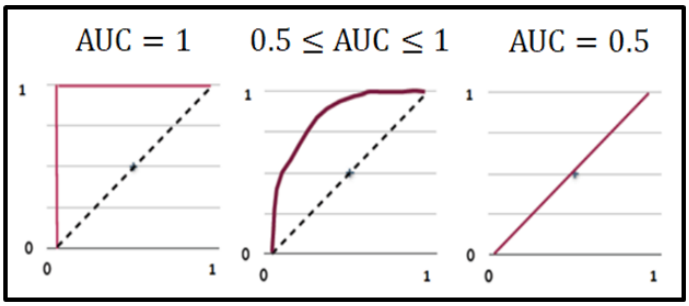
次數		預測	
		陽	陰
實際	陽	真陽 (True Positive)	偽陰 (False Negative)
	陰	偽陽 (False Positive)	真陰 (True Negative)

- 準確性$(Accuracy) = \frac{TP+TN}{TP+TN+FP+FN}$
- 敏感性$(sensitivity) = \frac{TP}{TP+FN}$
- 特異性$(specificity) = \frac{TN}{TN+FP}$

(2)**ROC 曲線及曲線下面積**：以圖像的方式呈現模型的分類能力，圖形的縱軸為「敏感性」，橫軸為「1-特異性」，即曲線上的一點代表一組「敏感性」與「1-特異性」的組合。判別 ROC 曲線時，以對角線為參考基準，若 ROC 曲線為對角線，則代表模型不具鑑別度，若越往左上方移動，表示模型正確預測為陽性的機率越高，同時錯誤預測為陽性的機率越低，則代表模型鑑別度佳，在圖形左上角，即座標點為(0,1)者為最佳模型。除直接觀察曲線外，亦可利用曲線下面積來判別模型的鑑別度。曲線下面積介於 0 與 1 間，判別標準如下：

- AUC = 1，模型分類能力最佳
- 0.5 ≤ AUC ≤ 1，模型具分類能力
- AUC = 0.5，模型不具分類能力

圖 3　ROC 曲線與對應之曲線下面積



五、數據科學的商業應用(Business Applications of Data Science)

本課程授課老師為 Hasan Hboubati 數據分析顧問，目前任職於美國加州的數據分析公司 Alteryx。該公司宗旨在打造一站式的數據統計分析軟體，讓使用者能夠在操作簡便的介面上，完成數據輸入、模型建構及數據圖像化，並曾於 2016 年榮登富比士(Forbes)最佳雲端服務公司排行榜(Cloud 100 List)全球第 24 名。本課程由管理科學角度切入，並透過分組討論與報告，探究數據科學在商業上的實務應用重點及重要基礎工具。

(一)實務應用重點

在網際網路及科技的進步下，過去各種無形的資訊(invisible information)已逐漸資料化(datafication)，例如：社群軟體的蓬勃發展帶動人際網絡的資訊化，讓資料數量呈現爆炸型的成長；另外，使用者在網際網路上進行各種活動所產生的資料廢氣(data exhaust)，例如瀏覽網頁或閱讀電子書每一頁的停留時間，亦是重要數據科學的重要資訊來源。

- **數據科學應用的價值在於「發掘潛在的有價資訊」**

數據科學的價值是從資料中找出尚未被發掘的有價資訊，以增加營收。舉如：美國跨國零售企業沃瑪特(Walmart)分析颶風侵襲前的銷售資料發現，家樂氏草莓吐司餅乾(Kellogg's strawberry Pop-Tarts)與啤酒銷量大增，因而增加這兩項產品在颶風來襲前的存貨；飯店將數據科學與平均客房收益(Revenue Per Available Room)模型結合，提供每位房客不同的房

價。另外，資訊也可以對重大風險提出示警，降低可能的損失。舉如：海上鑽油平台分析近似錯誤(near miss)與其他即時資料，避免爆炸意外等。

- **數據科學應用的工作在於「正確地解讀資訊」**

數據科學的工作是讓數據說話(Let’ s data speak)，但是數據必須資料科學家進行解讀，因此在解讀過程中容易有誤判或加諸過多因果關係(causation)的情形。舉如：在著名的夫妻婚前與婚後簡訊內容分析案例中，若僅透過觀察文字雲，可能無法正確推論夫妻情感變化；另外，簡訊使用的文字雖然相同，但也可能隨著時間代表不同的意義，例如：家(home)在婚前是代表各自的家庭，但婚後則是代表共組的家庭。

圖 4　夫妻婚前與婚後簡訊內容比較

- **數據科學應用的關鍵在於「擴大資料範圍」**

運用數據科學的關鍵在於必須善用各種內部與外部資料，思考拓展可能運用的資料範圍。舉如：沃瑪特使用企業內部銷售資料預測颶風侵襲前的銷售，也可嘗試運用民眾在網路上搜尋商品的外部資料；飯店業則可顧客在社群軟體的反饋意見，調整價格或提供客製化服務，以留住客群等。但在使用資料時，也必須在擴大營收與保障使用者隱私間取得平衡，例如歐盟近期實施的一般資料保護規則(GDPR)，即是對消費者隱私保護的重要規範。

- **數據科學成功的關鍵在於「善用外部力量」**

面對機器學習可能的瓶頸與障礙，數據科學家應巧妙結合外部力量，尋求解決最佳的方案。舉如：Google 面對街景圖片過於模糊，導致機器無法有效辨識圖片中的文字或數字時，採用 reCAPTCHA 機制，要求使用者登入時，驗證「我不是機器人」(I'm not a robot)，並判讀圖片中模糊的文字或數字。透過每位使用者僅花費數秒的時間，無須企業額外人力與研究經費，就足以有效解決機器判讀大量無法判讀的困境。

(二)重要基礎工具：Hadoop 及其生態系

隨著科技快速發展，資料呈現量(volume)「大」、型態(variety)「多元」、產生速度(velocity)「快」、資料真實性(veracity) 「不一致」等特性，讓傳統存儲與技術架構難以運作，因而產生新型數據基礎工具。**Hadoop 是目前廣受各種組織與產業青睞與採用的儲存及管理資料的雲端平台**，透過分散式架構的 HDFS 檔案系統(Hadoop Distributed File System)以及可分散運算的 MapReduce 程式演算方法，組合多個伺服器(節點)成為單一儲存及運算叢集系統(cluster system)，以達成處理巨量資料的任務。

Hapood 的基礎核心機制有二：

- HDFS 檔案系統(Hadoop Distributed File System, HDFS)

  HDFS 為 Hadoop 的底層架構，其中主要包含 NameNode、Secondary NameNode 及 DataNode。HDFS 會將資料分割成為檔案大小固定(通常為 64MB)的區塊(block)，且將每個區塊複製(預設值為 3 份)，分散交由不同的 DataNode 保管，並統一由 NameNode 監管所有 DataNode 上的資料。若某個 DataNode 上的資料遺失或損壞，NameNode 就會複製另一個 DataNode 上的資料；若是 NameNode 失效，則可由 Secondary NameNode 取代。透過 HDFS 的分散式存儲設計，不僅可突破資料大小的限制，也能確保資料的完整性。

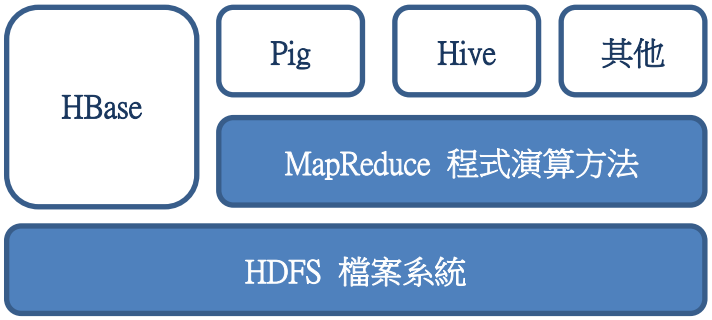
- MapReduce 程式演算方法

  MapReduce 為 Hadoop 的運算架構，以數學概念中 Map(映射)與 Reduce(歸納)為基礎的應用程式。MapReduce 會將需要處理的資料進行分拆，並分送給各節點運算，此階段稱為映射；再將各節點運算的結果整

合，此階段則稱為歸納。整個流程中，由單一的 JobTracker 分派任務及監管作業執行流程，並由 TaskTracker 管理各個任務在節點上的執行情況。透過 MapReduce 的分散式計算設計，讓各節點可平行處理資料，大幅節省資料處理時間。

在 HDFS 與 MapReduce 的基礎上，與 Hadoop 相關的擴充技術也相繼被開發，例如：Hadoop 專用資料庫系統「HBase」、可用來撰寫 MapReducet 程式的「Pig」，以及資料庫轉換系統「Hive」等，形成日益完備的 **Hadoop 生態系**(Hadoop Ecosystem)。

圖 5　Hadoop 生態系

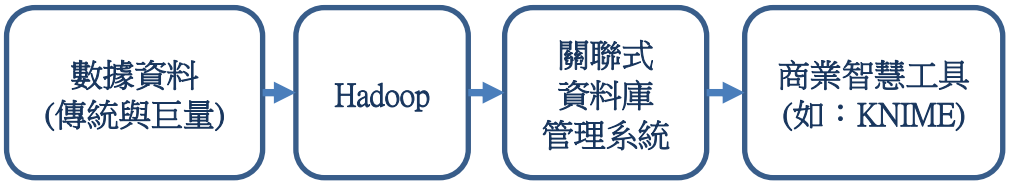


實務上，Hadoop 的常見應用方式有二：

- **資料提取**(data ingestion)

  此方法是將所有資料均先經由 Hadoop 處理後，轉入關聯式資料庫管理系統(Relational DataBase Management System, RDBMS)，再透過 KNIME 等商業智慧(Business intelligence, BI)工具分析。

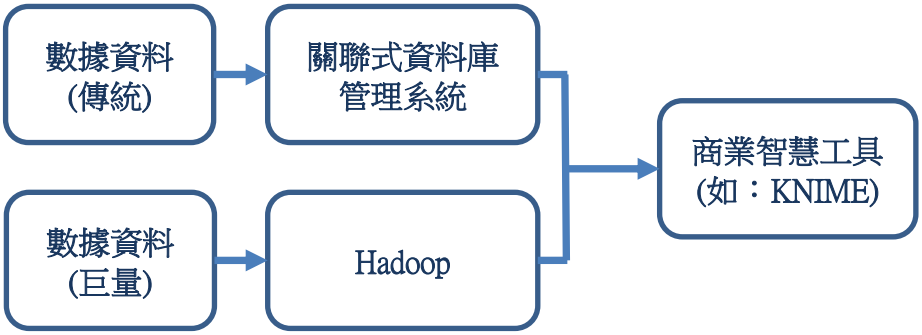
圖 6　Hadoop「資料提取」應用流程



- **資料管理**(data management)

  此方法是僅將巨量資料交由 Hadoop 處理，傳統資料則經由資料庫管理系統處理；再透過商業智慧工具分析。

圖 7　Hadoop「資料管理」應用流程



六、數據探勘、分析與視覺化(Data Exploration, Analytics, and Visualization)

本課程授課老師為資料科學家 Alfred Nigl 博士，研究領域為應用認知科學(applied cognitive science)，將高等統計應用於神經科學、消費者研究、市場研究等相關經驗超過 40 年。曾為惠普(Hewlett Packard)、微軟(Microsoft)、威訊無線(Verizon)等公司建構預測分析模型，並在辛辛那提大學(University of Cincinnati)、聖地牙哥州立大學(San Diego State University)等學校任教。

課程主要為分組製作報告，**嘗試應用商業智慧軟體**(如 Tableau 或 KNIKE 等)**及專業統計軟體**(如 SPSS 或 R 語言等)，**進行完整的行業數據挖礦標準過程**(CRISP-DM)，包含各種基本**統計分析**及**預測分析**，並透過**數據視覺化**將抽象的數據分析結果轉化為具體的各類新型態圖像，釐清面臨的問題與挑戰，並清楚傳遞數據背後所隱藏的有價資訊。另本課程指定參考書目為「用數據說故事：商業人士數據視覺化指南(Storytelling with Data: a Data Visualization Guide for Business Professionals)」。

(一)統計分析(statistical analysis)

為從龐大的數據庫中萃取有用的資訊，傳統基本的統計分析為重要的第一步驟，透過描述統計分析與推論統計分析，可對數據庫有基本的瞭解與認識。

1. **描述統計分析**(descriptive statistical analysis)

描述統計分析主要是用來觀察數據分布的特性，其中又以觀察數據的集中趨勢及離散程度為主。

(1)**集中趨勢**(central tendency)：用來初步瞭解數據分布的中間位置，常見衡量方式為算術平均數(arithmetic mean)、幾何平均數(geometric mean)、

中位數(median)、眾數(mode)等。

(2)**離散程度**(dispersion)：用來瞭解數據在中間值附近的分布狀態，若離散程度越小，數值越集中，常見衡量方式為全距(range)、變異數(variance)、標準差(standard deviation)。

2. **推論統計分析**(inferential statistical analysis)

推論統計分析主要是根據樣本的數據去推論真實情況，但其結果僅是機率性的推論，而非確定性的推論。本課程主要應用變異數分析及事後檢定。

(1)**變異數分析**(analysis of variance, ANOVA)：用來檢定數據內各分群的平均數是否有顯著差異(significant differences)，即研究自變數對依變數平均數的影響。若僅考慮一個自變數，則稱為單因子變異數分析(one-way ANOVA)；若同時考慮多個自變數，則稱為多因子變異數分析(factorial analysis of ANOVA)。

(2)**事後檢定**(post hoc test)：若變異數分析結果為各分群平均數有顯著的差異，則須進行事後檢定，以探討各分群間的差異。事後檢定又稱為多重比較(multiple comparison)，常見的方法有 Tukey's 檢定法及 Scheffé 檢定法等。

(二)預測分析(predictive analysis)

統計分析的目的在於對於數據與當前情況的瞭解；而預測分析的目的則在於對未來情況的推估，以協助決策單位做出正確的判斷及研擬最佳的對策。在預測分析的過程中，演算法選擇與設定將依個案的不同而有所差異，例如在進行數據的分群時，必須應用不同的集群演算法，同時也必須考量分群的數量，以確保預測結果的簡潔與有效性。

(三)數據視覺化(visualization)

數據的視覺化為數據分析報告相當重要的工作。一份好的數據分析報告，除了須要進行完整的統計分析與預測分析外，也需要運用視覺化方式呈現數據，讓繁雜的數據轉化為易懂的內容，以清楚地傳遞研究成果。為與進行預測分析相似，數據的視覺化也依個案的不同而有相當大的差異，

例如在討論各地區差異時，或可配合地圖標示，提高內容的可讀性與記憶性。

七、大數據視覺化與分析(Big Data Visualization and Analytics)

本課程授課老師為 Kenneth Reed 博士，具有超過 30 年業界經驗，專長在於分析關聯式資料庫。曾為居全球領先地位的商業軟體公司天睿(Teradata Corporation)開發零售預測系統，亦曾為全球最大企業管理顧問公司埃森哲(Accenture)的副合夥人(associate partner)，與多間世界大型企業合作建立全球客戶分析典範(practice)，為企業提供具洞察力的資料庫分析方法。

課程主要是採用全球第三大、歐洲第一大的電腦軟體公司 SAP 開發之**預測分析軟體**，**分析與視覺化企業機構內部的大數據資料**，以**提高**客戶滿意度及**企業競爭力**的方法，即為企業進行客戶關係管理(Customer Relationship Management, CRM)。對於企業機構而言，一個完整的商業分析團隊(business analytics team)必須要有決策者(executive)的直接參與以及專案經理(project manager)的規劃與管控，才能夠確保專案的如期如質推動，而資料庫建置與管理人員及數據科學家(data scientists)則是提供必要的資料建模與分析技術，最後則是必須要有專業的簡報與報告人員(communicator)，將分析結果以簡單易懂的方式向各界說明。

課程以真實線上旅遊網站(Zumber)的客戶資料，首先介紹資料庫所採用的**邏輯資料模式**(Logical Data Model, LDM)，後續再使用 SAP 預測分析軟體對客戶做**集群分析**(clustering analysis)，了解不同客群的消費習性，以提供客製化商品與服務，最終目的在於提高企業的利潤。

(一)邏輯資料模式

在資料建模(data modeling)過程中，邏輯資料模式是目前最為普遍的資料模式之一。此模式是由台灣學者陳品山(Peter Chen)教授於 1976 年提出，將資料建立在實體關係(entity-relationship)，組成的元件包含實體(entity)與關係(relationship)，其中實體是資料內重要的標的，而所有實體間屬性與關係則是由主鍵(primary key)及連結實體關係的外來鍵(foreign keys)所構成，以簡單且精確的方式，描述真實世界的語意與關係。

(二)集群分析

集群分析是大數據視覺化與分析最重要的工作之一，透過集群分析可以將資料進行分群，深入理解各群的重要特性，並對不同族群提出有效策略。惟實務上，在集群分析仍有兩點事項須特別注意：第一，分群個數雖然可以由操作者自由設定，但不應過多，避免需提出的策略過多，增加後續執行的難度；第二，在決定分群個數後，仍須對各群的特性進行理解，且須統整歸納出一明顯的特性，才能使集群分析具有意義與效果。

參、與講師交流

為擴大本次進修效益，特別針對數據科學及預測分析在總體經濟規劃的應用可能等與本處業務推動相關之議題，於課堂或課後向授課老師請益。

一、傳統參數統計分析與機器學習的差異

問：過去總體經濟常用傳統參數統計進行預測分析，該方法與機器學習最大概念上的差異為何？

答：R. Nisbet 博士表示，傳統參數統計與機器學習在**本質上**有相當的不同。第一是**資料運用方式**的不同，傳統參數統計是透過「整體」的運算，讓資料顯現出資訊，如計算平均數或標準差等；但機器學習是透過「逐筆」的運算，讓資料產生價值。第二是**模型假設**的不同。傳統參數統計在進行預測時，模型必須符合許多的假設，如預測變數必須符合常態分配且相互獨立等；但是機器學習算法並未要求這些假設。在商業的實務應用上，多數資料並無法滿足傳統參數統計的假設，這也是機器學習目前廣泛被商業應用的原因。

二、機器學習在總體規劃的應用可能

問：既然機器學習已經廣泛應用在商業上，未來是否能應用在總體經濟規劃？可能的應用方法？

答：R. Nisbet 博士表示，就渠所知，經濟領域已開始思考如何應用機器學習演算法，但就目前而言，**傳統參數統計在總體經濟規劃上仍有很好的表現**，且機器學習在總體經濟的應用仍有許多困難需克服。因此，短期間機器學習尚難以完全取代傳統模型。雖然如此，仍可將**機器學習作為傳統模型的**

輔助工具，藉由比較兩者的預測結果，強化分析的準確性。未來嘗試應用機器學習在總體規劃時，或可嘗試廣義線性模型(generalized linear model)，同時也必須留意前期的資料準備，例如：遺漏值的處理與資料的平衡等。

答：A. Nigl 博士表示，渠未有應用機器學習或預測分析在總體經濟規劃的相關經驗，但相信未來在嘗試應用的過程中，只要**掌握數據科學的重要原則**，相信亦能有所成果。舉如：在建置模型時，必須要掌握**簡要原則**，**避免過度複雜化模型**，導致可解釋性的降低；另外，數據科學在實務操作上則必須掌握**重複試驗原則**，必須透過反覆多次的嘗試，並對結果進行分析與調整，永不放棄才能獲致好的成果。

答：K. Reed 博士認為，在大多數的預測應用實務上，**迴歸模型仍具有相當的解釋能力**，相較而言，機器學習的神經網絡模型雖然可提高對原有資料的配適度(fitness)，但也有可能有過度訓練(over-training)的疑慮，並造成預測能力的下降，因此如要運用機器學習於總體規劃時，需特別留意相關的問題。

三、應用 R 在機器學習在總體規劃的可能

問：目前總體經濟多採用 R 等統計軟體，若未來如要應用機器學習在總體經濟，R 是否仍能夠勝任，或是必須使用其他程式語言？另外，有無推薦之相關書籍？

答：A. Pahwa 博士認為，**R 語言在未來機器學習仍是重要工具**。雖然目前在 Google 的推展下，Python 使用者越來越多，且已經開發出許多與機器學習與深度學習相關的開放原始碼程式庫，例如 TensorFlow 等，但 R 為統計軟體，現階段在進行傳統統計分析上仍具優勢。未來軟體工具可能有不同的變化與演進，但只要清楚瞭解機器學習演算法背後的概念與邏輯，其餘僅是操作介面的轉換，技術上應均能克服。另外，總體經濟規劃或可參考時間序列(time series)模型，相關書籍包括：「Time Series Analysis: With Applications in R」、「Time Series Analysis: Forecasting and Control」及「Introduction to Financial Forecasting in Investment Analysis」等。

四、政府部門應用數據科學的策略與實例

問：數據科學為商業帶來新的商機與發展，政府部門在嘗試應用時的策略建議為何？有無成功實行的案例？

答：H. Hboubati 顧問表示，數據科學在商業應用的目的是提高營收，與政府應用數據科學目的或有不同，但是就**應用策略**來說並無二致。第一，**必須確認擁有的內部資料**，這部分政府部門往往有許多私部門無法取得、與民眾切身相關的重要資料，是數據科學的應用題材。第二，**必須善用外部的資料，並與內部資料結合**。舉例而言，**美國國家稅務局**(Internal Revenue Service, IRS)**嘗試應用數據科學進行查稅**。長久以來，監測民眾低報所得的欺騙行為很難有效實行，但隨著近年來電子報稅的普及，提供完備的內部數據基礎，另外國稅局也運用社群軟體的外部資料，對有逃漏稅嫌疑的民眾進行其所得申報的評估，有效提高查稅效率。

肆、課程成果報告

數據科學與預測分析是一門理論理解與實務操作並重的課程，因此為強化學習效果，課程均需完成課後作業、進行期中考試或製作成果報告，成果報告重點摘要如下：

－「**預測分析應用**」課程成果報告，應用 KNIME 軟體分析「UCI machine learning archive」中美國所得普查(census income)資料，預測可能的高所得族群(詳見附錄 1 之一)。

－「**數據挖礦的數據準備**」課程成果報告，應用 KNIME 軟體分析分析「UCI machine learning archive」中的美國癱瘓退伍軍人協會捐款資料，預測未來可能的捐贈者(詳見附錄 1 之二)。

－「**數據科學的商業應用**」課程成果報告，主要介紹 Hadoop 中可提供重要演算法的工具 Mahout，以及商業智慧工具 Pentaho(詳見附錄 1 之三)。

－「**數據探勘、分析與視覺化**」課程成果報告，應用各種商業智慧軟體及專業統計軟體，分析 Tableau 軟體內建之 Global Superstore 在歐洲市場中的家具銷售情況，並提出增加收益之可能因應策略(詳見附錄 1 之四)。

第三章 心得與建議

壹、心得

從實務流程觀察，數據科學的應用是結合數學、統計、機器學習、資料探勘與資料視覺化等，分析組織既有內部與外部資料，發掘與解決組織潛在的各種問題。數據科學在總體規劃的最大優勢在提升政策分析的洞察力及加快政策形成與回應速度。數據科學應用在總體規劃的應用有三：

1. 大數據與經濟預測

由於大數據具有數量大、種類多、產生速度快的特性，可大幅減少數據的蒐集成本，並提供即時性高、精準度高、細緻度高、涵蓋度高的資料。整體而言，大數據可以協助個體經濟學更精準的掌握個體的動態行為，包括：網路平台或金融市場的交易行為、社群媒體的訊息傳遞行為，及搜尋引擎的資訊蒐集行為等；另方面，大數據亦可以提供總體經濟學更具個體基礎的衡量指標，包括：失業率、通貨膨脹率等。

2. 機器學習與計量經濟

經濟學家經常運用各種統計方法分析經濟議題，值得注意的是機器學習與大數據的應用已開始在經濟研究中出現。機器學習與計量模型均可應用在預測分析，但對於計量經濟而言，應用機器學習的「資料分割」與「資料交叉驗證」方法，將可提升計量模型預測能力；另方面，面對高維度資料(high-dimensional data)，「特徵篩選」(即在資料集中選取具代表性的子集合)與「特徵提取」(即將高維度資料集映射在低維度空間)則是計量模型求解時必要的降維作為。

3. 數據科學與決策思維

數據科學的價值在於可用客觀的數據輔佐主觀的決策。近來，由於數據儲存與分析技術的進步，使政府部門使用數據進行決策的門檻降低，推動政府數據治理的加速實現。政府應用數據科學的優勢在於擁有大量的行政資料(administrative data)，這些資料的特點是涵蓋範圍廣泛，包含社會各群體與階層，且通常為長期追蹤資料，對政策的評估與形成至關重要。若能在保護個資的前提下，強化政府跨資料庫的連結與資料的開放，可為政策制定提供更完善的數據基礎。

貳、建議

國際貨幣基金(IMF) 2018 年 3 月發布「數位時代數據與統計總戰略(Overarching Strategy on Data and Statistics at the Fund in the Digital Age)」，為 IMF 首次發布關於數據與統計的白皮書，在「整合(Integration)」、「創新(Innovation)」、「智能(Intelligence)」三大關鍵要素上，強化數據的蒐集、應用與分享，並將進行組織與作業的微調，及適時採用數據科學分析方法。另 OECD、世界銀行指出，適時應用數據科學，可找出數據中潛在有用的關聯性資訊與規律，創造數據新價值，落實「循證決策(evidence-based policy making)」的目標。因此，政府宜持續關注數據科學在規劃整體戰略的國際推動進展。

從數據科學角度，國發計畫總體規劃的研擬步驟有四：第一步是建構數據庫，據以作為決策資源系統；第二步是數據挖掘，搜尋及掌握有價值的資訊；第三步是進行數據的整合運用及模擬情境分析(scenario analysis)；第四步是根據數據評估，制定政策、行動方案(action plan)及項目分配(project allocation)。目前國發計畫總體規劃作業亦是如此，惟所應用數據多以結構性數據為主，對於大量非結構性與半結構性相對不足;另在分析模式方面,亦多使用傳統計量方法，而非與人工智慧發展密切相關的機器學習理論。

準此，未來總體規劃可從「擴充數據來源」及「轉化分析工具」兩方面著手。在擴增數據來源方面，政府行政資料是政策制定的重要數據來源，惟社會網絡、商業系統及物聯網上的數據亦值得關注；在轉化分析工具方面，數據科學的快速發展，對大量非結構化與半結構化數據的儲存、運算與分析能力已大幅提升，不僅可作為官方統計的創新來源，亦可彌補官方統計的滯後，對現有指標進行預測，更將有助總體規劃回答新問題、產生新指標。

附錄

附錄 1：課程成果報告

一、「預測分析應用」課程成果報告

“Applications of Predictive Analytics”

Final Report

Who Can Afford High-End Products?

Using Census Income to Predict Consumers’ Income

Kwei-Ho Lin and Hung-Ling Pan

Apr./28/2018

Abstract

The purpose of this report is to identify who are the potential customers can afford high-end products. By carefully examining “Census Income” data set with “Cross Industry Standard Processing for Data Mining (CRISP-DM)”, we can predict consumers’ income with high overall accuracy rates, which are 0.816 and 0.819 under the “Random Forest” algorithm and “Tree Ensemble” algorithm. The report suggests that people who have capital gain or have child are more likely to have higher income and can afford high-end products.

Introduction

Generally, people at the top of the pyramid have more disposal income to buy high-end products or luxury goods. Therefore, companies offer not only ordinary products or service, but also high-end ones to cater those consumers. For example, Apple Inc. released both IponeX and Iphone8 smart phones in 2017. Undoubtedly, IphoneX is designed for the premium end of the market. It equipped with up-to-date high-tech, such as a new all-screen design and with face ID, and its sale price is 40% higher than Iphone8 on average.

Thus, companies are eager to know whether high-end market exits and to find out who could afford them. In this report, “Census Income” data set, downloaded from UCI Machine Learning Repository, is under examination in order to predict whose income exceeds $50,000 per year.

Methodology

To ensure quality of the data set, Cross Industry Standard Processing for Data Mining (CRISP-DM) is adopted. Also, In order to precisely predict consumers’ income, 2 algorithms (Random Forest and Tree Ensemble) and 2 balancing methods (unbalanced and equal size balancing) are adopted to build 4 models.

1. Data input

In this dataset, there are 32,650 rows and each has 15 attributes, including age, work class, final weight, education, education-number, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and whether his/her income exceeds $50,000 per year.

2. Data understanding

a. Learning the data type of each variable and their basic statistical features: 8 of 15 attributes are categorical, including work class, education, marital-status, occupation, relationship, race, sex.

b. Confirming which variables have missing value or outlier value: 2,399 rows have missing value.

3. Data preparation

a. Dealing with missing values: Removing rows which have missing values.

b. Feature selection: Only excluding irrelevant attribute “final weight.”

c. Data partitioning: Dividing the dataset into modeling data (including training data and testing data) and validating data, with a 70:30 ratio.

d. Data balancing: (I) No balancing; (II) Balancing via Equal Size Sampling.

4. Modeling/Algorithm

a. Random Forest: Doing a random search along various parameter path of change.

b. Tree Ensemble: Doing an optimized search along various parameter path of change.

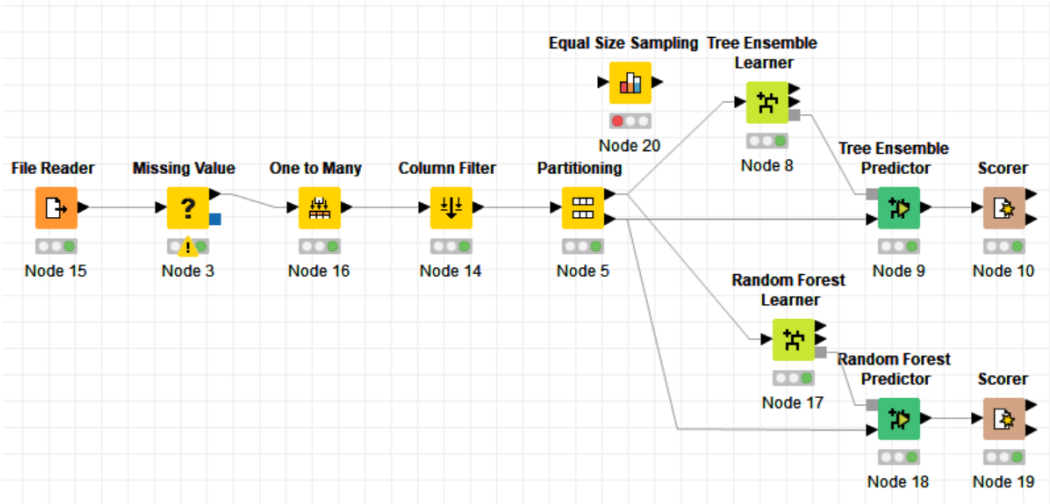


Figure 1 CRISP-DM of this report (KNIME)

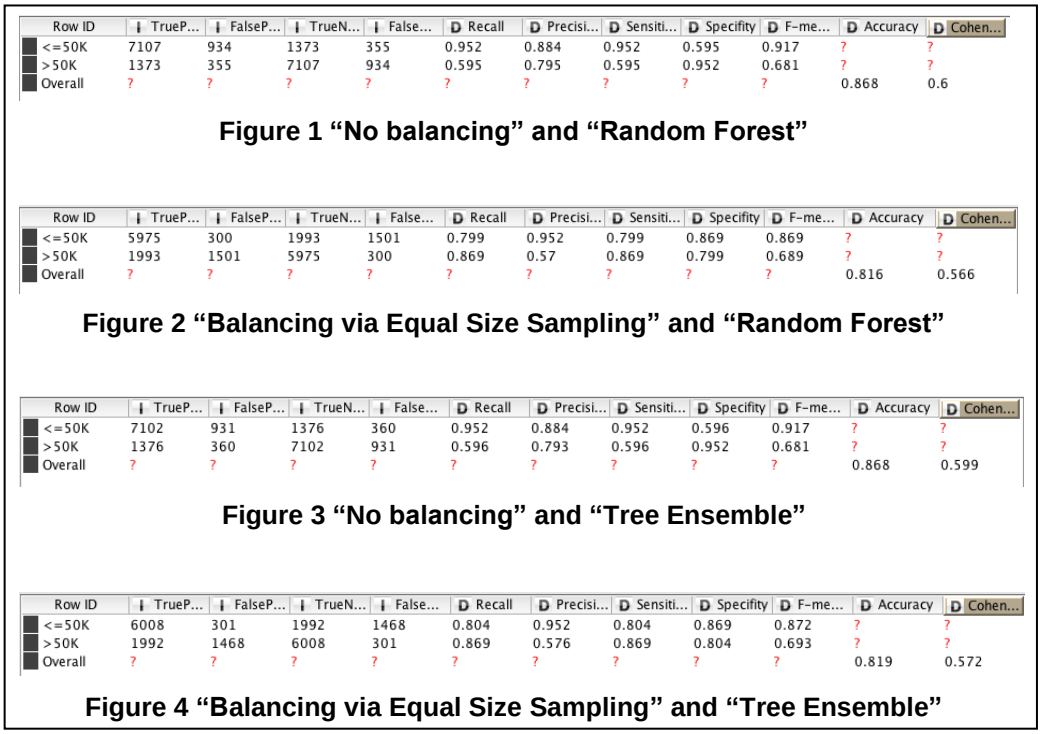
Results and Discussion

1. **“Equal Size Sampling” is preferred in this report.**

   a. Using “No balancing” method, the accuracy value is always higher than “Equal Size Sampling.”

      I. For Random Forest, the result is 0.868>0.816.

      II. For Tree Ensemble, the result is 0.868>0.819.



Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy	Cohen...
<=50K	7107	934	1373	355	0.952	0.884	0.952	0.595	0.917	?	?
>50K	1373	355	7107	934	0.595	0.795	0.595	0.952	0.681	?	?
Overall	?	?	?	?	?	?	?	?	?	0.868	0.6

Figure 1 “No balancing” and “Random Forest”

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy	Cohen...
<=50K	5975	300	1993	1501	0.799	0.952	0.799	0.869	0.869	?	?
>50K	1993	1501	5975	300	0.869	0.57	0.869	0.799	0.689	?	?
Overall	?	?	?	?	?	?	?	?	?	0.816	0.566

Figure 2 “Balancing via Equal Size Sampling” and “Random Forest”

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy	Cohen...
<=50K	7102	931	1376	360	0.952	0.884	0.952	0.596	0.917	?	?
>50K	1376	360	7102	931	0.596	0.793	0.596	0.952	0.681	?	?
Overall	?	?	?	?	?	?	?	?	?	0.868	0.599

Figure 3 “No balancing” and “Tree Ensemble”

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy	Cohen...
<=50K	6008	301	1992	1468	0.804	0.952	0.804	0.869	0.872	?	?
>50K	1992	1468	6008	301	0.869	0.576	0.869	0.804	0.693	?	?
Overall	?	?	?	?	?	?	?	?	?	0.819	0.572

Figure 4 “Balancing via Equal Size Sampling” and “Tree Ensemble”

   b. However, the gap between recall values of target classes is greater than “Equal Size Sampling.”

      I. For Random Forest, the result is 0.357(=0.952-0.595)>0.07(=0.869-0.799).

      II. For Tree Ensemble, the result is 0.356(=0.952-0.596)>0.065(=0.869-0.804).

   c. Using “Equal Size Sampling” method, its overall accuracy is high enough, so are its recall values. Thus, the following discussion will base on it.

2. **Both “Random Forest” algorithm and “Tree Ensemble” algorithm perform well.**

   a. Their overall accuracy values are very similar, i.e. 0.816 and 0.819.

   b. There is only a slight difference in their recall values of target classes.

      I. In Random Forest algorithm, the recall values for <=50K and >50k are 0.799 and 0.869.

      II. In Tree Ensemble algorithm, the recall values for <=50K and >50k are 0.804 and 0.869.

c. Both “Random Forest” algorithm and “Tree Ensemble” algorithm are adopted in the following discussion.

3. **“Capital Gain” and “Owned Child” are the most important attributes in predicting whose income exceeds 50,000 per year**.

   a. Partitioning the data set by the rule of “Draw Sampling,” rather than “Linear Sampling,” in order to make sure the flexibility of our model. In addition, running this process three times with an aim to gaining a more robust attribution results.

      I. In Random Forest algorithm, “capital gain” and “own-child” are the only two appearing in these three top 5 attribution tables.

      II. In Tree Ensemble algorithm, “capital gain” and “own-child” are also the only two appearing in these three top 5 attribution tables.

   b. Thus, “capital gain” and “own-child” are for sure the most important attributes.

Conclusions

1. By applying CRISP-DM with “Equal Size Sampling” and two algorithms (“Random Forest” and “Tree Ensemble”), **the company** could be quite confident to predict customers’ income and **could aim at those who have capital gain or have children to promote the high-end products.**

2. **In the future, there are some actions can be taken in order to enhance the model.**

   a. Applying new algorithms, such as Support Vector Machine (SVM).

   b. Adding new variables into the datasets, such as IQ test scores.

二、「數據挖礦的數據準備」課程成果報告

"Data Preparation for Predictive Analysis"

Final Report

Kwei-Ho Lin

May/6/2018

Abstract

The purpose of this report is to predict which Paralyzed Veterans of America (PVA) donor will continue to donate. Among 13 data mining models, this report suggests that using over-sampling method to balance data set and using Random Forest or Tree Ensemble algorithm is the best model, with high model accuracy (0.995), to do the prediction. This accuracy can be further enhanced to 0.996 when take the feature selection into the best model. In this paper, the top 3 decision rules are also provided by the data mining process in order to offer more well-defined information to PVA.

Introduction

This report highlights the results of the analysis based on the Paralyzed Veterans of America (PVA) data set. PVA is a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. With over 13 million donors, PVA is also one of the largest direct mail fund raisers in the US. For maintaining its operation, precise prediction of which donor is still willing to donate next year will be crucial. In this report, a data mining method is proposed to carefully examine the PVA data set.

The first part of this report is the introduction. The second part is to explain the steps to prepare the dataset for analysis. The third part is the results and discussion of the experimental analysis using an array of models. The fourth part is the conclusion.

Methodology

There are 10 steps in the data preparation. In order to precisely predict which donor will donate next year, 4 algorithms (Logistic Regression, Decision Tree, Random Forest, Tree Ensemble) and 3 balancing methods (unbalanced, under-balanced, over-balanced) are adopted to build 12 models. Then, the best algorithm and balancing method is filtered by the short-list to enhance performance of the model (Table 1).

	Unbalanced	Under-sampling	Over-sampling
Logistic Regression	Model 1	Model 5	Model 9
Decision Tree	Model 2	Model 6	Model 10
Random Forest	Model 3	Model 7	Model 11
Tree Ensemble	Model 4	Model 8	Model 12
The best balancing method and algorithm + Short-list			Model 13

Table 1 13 Models with different experimental designs

1. **Data integration:** Joining data from 4 source data files, including Donors2.csv, Demographics3.csv, Donor_Hist2.csv, and Promo_Hist.csv.
2. **Data cleansing:** Transforming those zips which are negative to positive ones and transform those zips which are consisted of only 4 figures to 5. Also, substituting outliers into second-high number in those variables.
3. **Missing values imputation:** Missing values of Wealth1 variable can be imputed with a constant, a formula or a model. Here, the Simple Regression model is applied.
4. **New variables deviation**
    a. **Deriving dummy variable:** Transforming RFA_2A variable to 4 dummy variables
    b. **Replacing alphanumeric codes with numbers:** Converting categorical variables to numbers, then deleting the original categorical variables.
    c. **Deriving new variable:** Deriving DOCTOR and GIFT_PER_PROMO.
5. **Data normalization**: This step is necessary only for Logistic Regression algorithm.
6. **Feature selection:** Using "Feature Elimination" node, which is a meta node, to create the short-list.
    e. Using "Backward Feature Elimination Start" node and "Backward Feature Elimination End" node to set up a processing loop to delete unimportant variables one at a time.
    f. "Decision Tree" algorithm is adopted in this process.

g. Using “Backward Feature Elimination Filter” node to decide which combination between error level and variable number is preferable.

7. **Algorithm selection:** Logistic Regression, Decision Tree, Random Forest, Tree Ensemble.

8. **Data Conditioning**

   a. **Data partitioning:** Dividing the dataset into modeling data and validating data, with a 70:30 ratio.

   b. **Data balancing:** Unbalanced, Under-sampling, Over-sampling.

9. **Preliminary model building**: Using step 1 to 8 to understand which data preparation technique to use for the final model.

10. **Feedback to earlier data preparation**

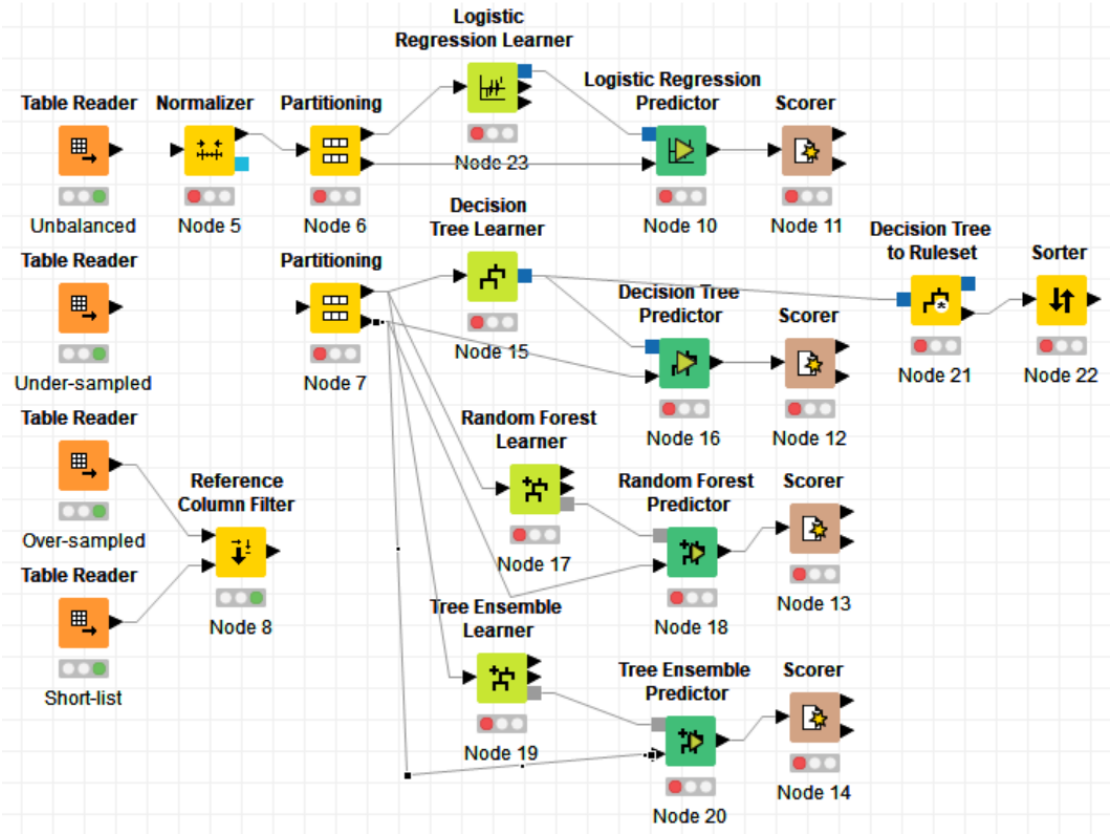


Figure 1 CRISP-DM of this report (KNIME)

Results & Discussion

There are 4 important results of the experimental models.

1. The effect of both balancing method are significant on each algorithm

d. Table 2 shows that Under-sampling method decreases overall accuracy when modeling with all algorithms. However, it narrows down the gap between sensitivity and specificity.

e. Table 2 also shows that Over-sampling method increases overall accuracy and narrows down the gap between sensitivity and specificity.

Unbalanced method	Under-sampling method	Over-sampling
Logistic Regression ● Accuracy: 0.952 ● Sensitivity: 0 ● Specificity: 1	Logistic Regression ● Accuracy: 0.506 ● Sensitivity: 0.502 ● Specificity: 0.51	Logistic Regression ● Accuracy: 0.612 ● Sensitivity: 0.604 ● Specificity: 0.62
Decision Tree ● Accuracy: 0.904 ● Sensitivity: 0.079 ● Specificity: 0.946	Decision Tree ● Accuracy: 0.513 ● Sensitivity: 0.544 ● Specificity: 0.485	Decision Tree ● Accuracy: 0.937 ● Sensitivity: 0.966 ● Specificity: 0.908
Random Forest ● Accuracy: 0.951 ● Sensitivity: 0 ● Specificity: 1	Random Forest ● Accuracy: 0.541 ● Sensitivity: 0.566 ● Specificity: 0.518	Random Forest ● Accuracy: 0.995 ● Sensitivity: 0.991 ● Specificity: 1
Tree Ensemble ● Accuracy: 0.951 ● Sensitivity: 0 ● Specificity: 1	Tree Ensemble ● Accuracy: 0.557 ● Sensitivity: 0.558 ● Specificity: 0.555	Tree Ensemble ● Accuracy: 0.995 ● Sensitivity: 0.991 ● Specificity: 1

Table 2 Comparison b/w each balancing method and algorithm

2. **Random Forest algorithm performs best with oversample balancing method, so does Tree Ensemble algorithm with oversample balancing method.**

Table 2 shows that under over-sampling method, the performance of Random Forest and Tree Ensemble algorithm is the same, both with highest accuracy (0.995), sensitivity (0.991) and specificity (1).

3. **"Decision Tree to Ruleset" node is adopted to examine the top 3 decision rules.**

The results of each data set are as following Table 3 to 5.

a. Unbalanced Data

Top 1 (Record count=7)	**Top 2** (Record count=6)	**Top3** (Record count=6)
AGE<=75.5 DOMAIN (Num) >12.5 CARDPROM<= 34.5	LOCALGOV<=6.5 STATEGOV<=6.5 CLUSTER<=1.5	B_GEOCODE<=0.5 STATE (Num)<=11.5 AVGGIFT<=7.43

Table 3 Top 3 decision rule from unbalanced data and Decision Tree

b. Under-sampled Data

Top 1 (Record count=29)	**Top 2** (Record count=25)	**Top3** (Record count=22)
MAGFEM<=0.5 CLUSTER<=39.5 STATE (Num)>15.5	AVGGIFT<=40.625 CLUSTER2>3.5 STATE (Num)<=14.5	RAMNTALL>71.75 AVGGIFT<=5.408 AGE<=87.5

Table 4 Top 3 decision rule from under-sampled data and Decision Tree

c. Over-sampled Data

Top 1 (Record count=3,802)	**Top 2** (Record count=1,415)	**Top3** (Record count=576)
X_PERSTRFL<=1 X_PERSTRFL>0	C_GEOCODE2<=1 C_GEOCODE2>0 MALEMILI>0	RAF_2F<=2 RAF_2F>1 X_PERSTRFL<=0

Table 5 Top 3 decision rule from over-sampled data and Decision Tree

4. **When short-list is adopted into model, it can slightly improve the outcomes, especially with Tree Ensemble algorithm (Table 6).**

Random Forest algorithm	Tree Ensemble algorithm
With short-list ● Accuracy: 0.995 ● Sensitivity: 0.992 ● Specificity: 0.999	With short-list ● Accuracy: 0.996 ● Sensitivity: 0.992 ● Specificity: 0.999
Without short-list ● Accuracy: 0.995 ● Sensitivity: 0.991 ● Specificity: 1	Without short-list ● Accuracy: 0.995 ● Sensitivity: 0.991 ● Specificity: 1

Table 6 Comparison best model with and without short-list

Conclusions

This report uses the data mining approach to analyze PVA data set and to predict which donor will donate next year. In the whole data mining process, data preparation counts an important role, since the method to balance the data set and the algorithm to fit the data set are crucial to determine the model accuracy. Also, the feature selection is necessary in data preparation step to further enhance the best model.

The main purpose of predictive analytics is to provide information from data set and then to help people to make a better decision. Therefore, the top 3 decision rules have been listed in this report. For PVA, these rules are valuable when contacting potential donors.

三、「數據科學的商業應用」課程成果報告

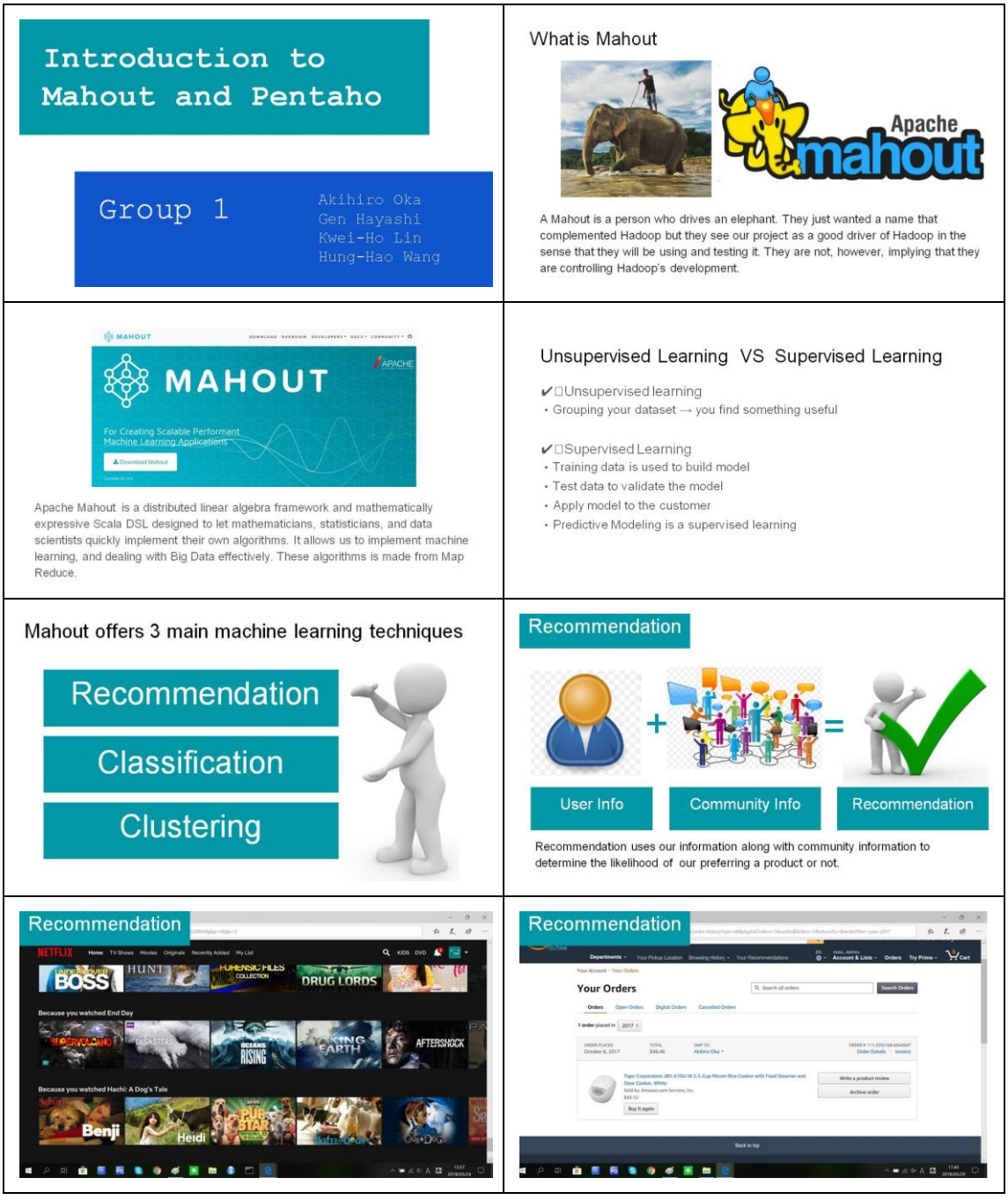
“Business Applications of Data Science”

Final Report

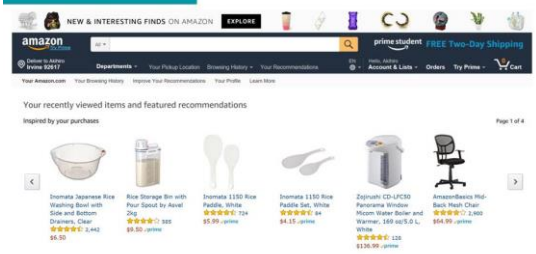
Introduction to Mahout and Pentaho

Akihiro Oka, Gen Hayashi, Kwei-Ho Lin, Hung-Hao Wang

May/30/2018



Recommendation



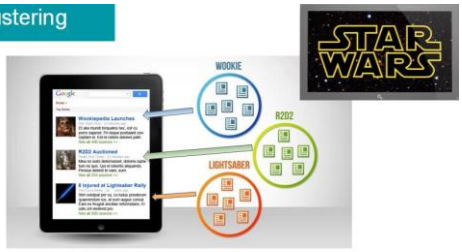
Classification

How does your favorite email system mark them as spam will they use the second technique classification?
Classification uses known data to determine how new data should be classified into a set of existing categories, so every time we mark or unmarked an email as spam we directly influence or emails classification engine for flagging future spam

Clustering

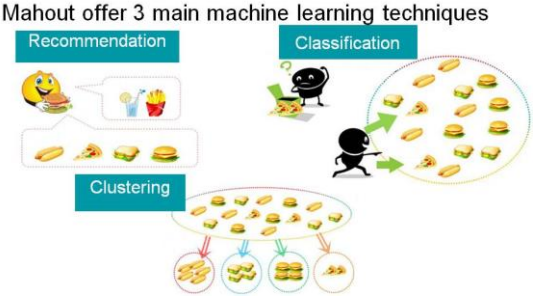
Clustering forms groups of similar data based on common characteristics unlike classification. Clustering does not group data into an existing set of known categories. This is particularly useful when you aren't sure how to organize your data in the first place

Clustering



Google News uses this powerful technique to make sense of the ever-changing stream of news articles from around the world enabling you to keep up with the latest events

Mahout offer 3 main machine learning techniques



pentaho®
A Hitachi Group Company

Pentaho

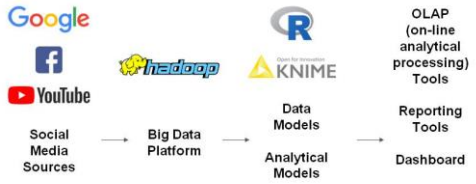
- Pentaho is the commercial open source software for Business Intelligence (BI)
  - Original author: Pentaho Corporation, 2004
  - Developer: Hitachi Vantara, 2015
  - Website: www.pentaho.com
- Running well under multi platform
  - Windows, Linux, Mac OS X,...
- Pentaho's Applications are all build under Java platform
- License: Community Edition(CE), Enterprise Edition(EE)

Leaders Use Pentaho

CAT
VEIKKAUS
OPOWER
Nasdaq
BT
edo
randstad
Marketo

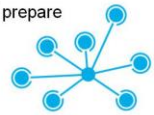
Why Pentaho?

- One stop solution for all the business analytics need

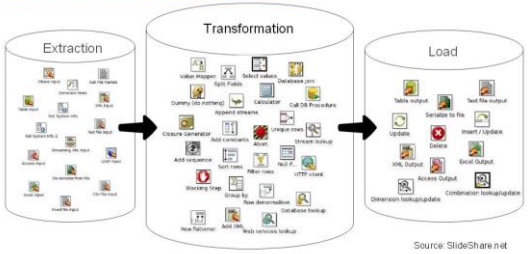


Pentaho Data Integration

- Enables users to ingest, blend, cleanse and prepare diverse data from any source.
- Some of the Pentaho reporting features
  - Rich transformation library
  - Advanced data warehousing support
  - Enterprise class performance and scalability
  - Data integration enterprise console
  - Modern, standard based architecture

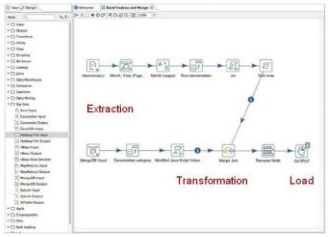


Pentaho Data Integration



Source: SlideShare.net

Pentaho Data Integration

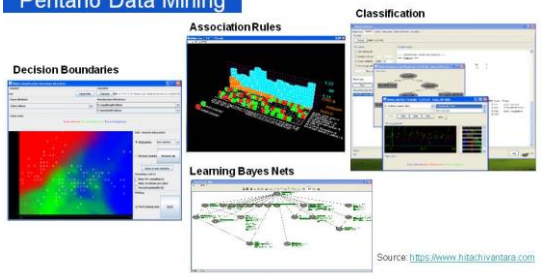


Source:

Pentaho Data Mining

- Provides insight into hidden patterns and relationships in data and also provides indicators for future performance based on those historical patterns
- Some of the Pentaho data mining features
  - Wide range of algorithms
  - Powerful graphical design tools
  - Provides powerful data engine machine
  - Enables embedding of recommendations in applications

Pentaho Data Mining



Source: https://www.hitachivantara.com

Pentaho Reporting

- Allows organizations to easily access, format and deliver information to employees, customers and partners
- Some of the Pentaho reporting features
  - Flexible reporting
  - Broad data source report
  - Pentaho report designer
  - Pentaho User console
  - Ad hoc reporting interface
  - Comprehensive role based security

Pentaho Reporting

Source: https://www.hitachivantara.com

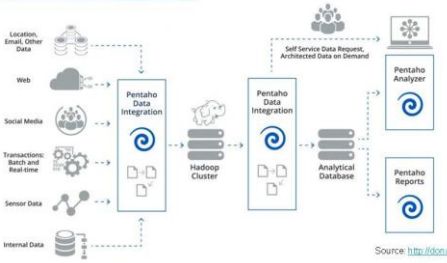
Pentaho Dashbord

- Gives business users the critical information they need to understand and improve organizational performance
- Some of the Pentaho dashboards features
  - Flexible deployment options
  - Easy information access by subjects or roles
  - Security compliance
  - Rich graphical displays
  - Pentaho user console
  - Pentaho dashboard designer

Pentaho Dashbord

Source: https://www.hitachivantara.com

Pentaho is a useful tool



Source: http://donatz.info

thank you!

四、「數據探勘、分析與視覺化」課程成果報告

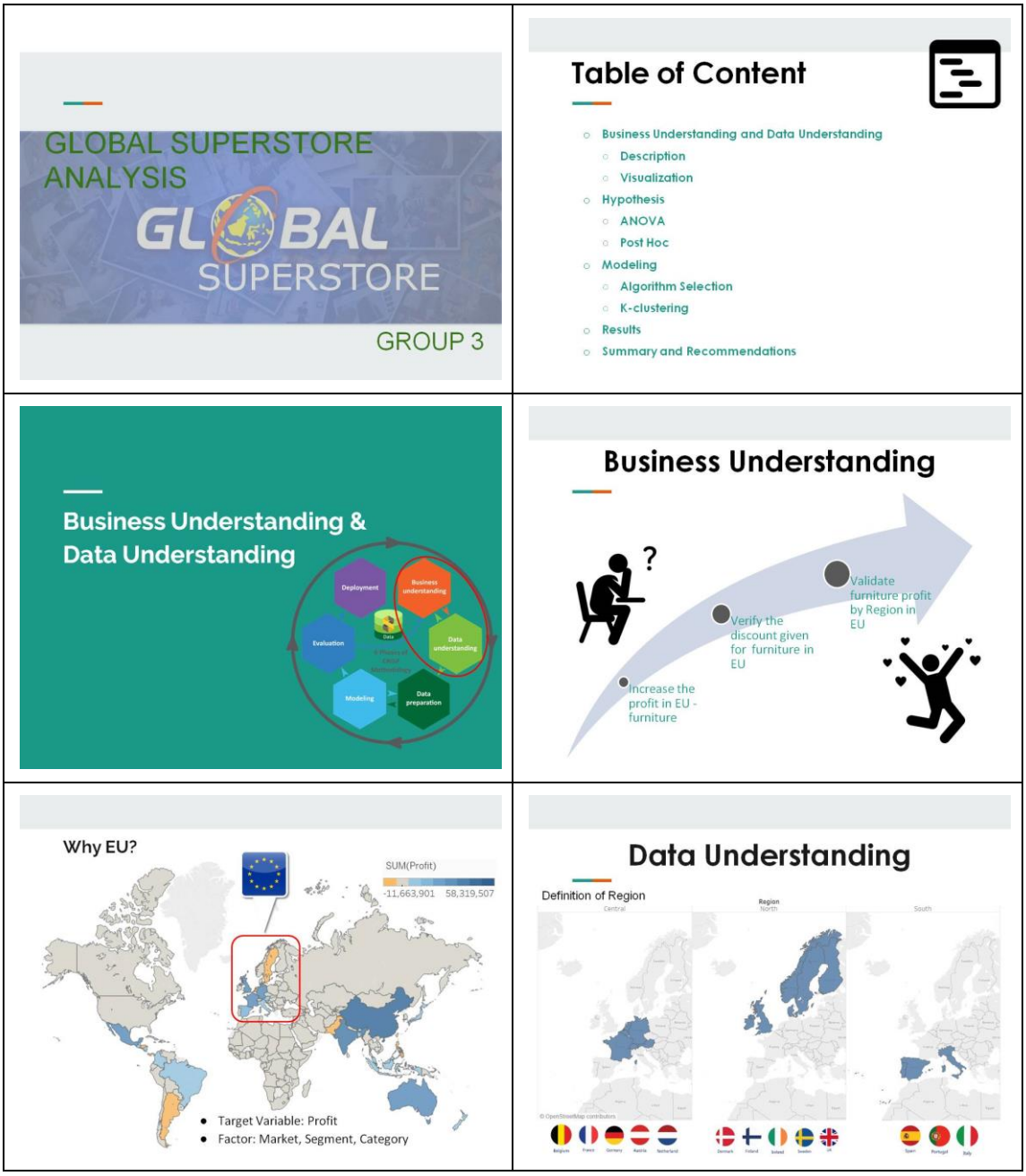
“Data Exploration, Analytics, and Visualization”

Final Report

Global Superstore Analysis

Akihiro Oka, Kwei-Ho Lin, Hung-Hao Wang

June/11/2018



Data Understanding

Data Understanding

Hypothesis

- **Hypothesis 1**
  H0: Shipping cost on furniture has no significant effect between EU Regions
  H1: There is a significance effect
- **Hypothesis 2**
  H0: Profit on furniture has no significant effect between EU Region
  H1: There is a significance effect
- **Hypothesis 3**
  H0: Sales on furniture has no significant effect between EU Regions
  H1: There is a significance effect
- **Hypothesis 4**
  H0: Discount on furniture has no significant effect between EU Region
  H1: There is a significance effect

Methodology

- **Hypothesis 1**
  H0: Shipping cost on furniture has no significant effect between EU Regions
  H1: There is a significance effect

ANOVA

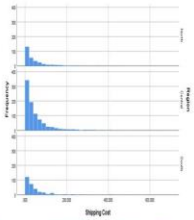
Shipping Cost

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5499.550	2	2749.775	.502	.605
Within Groups	8206957.035	1498	5478.610		
Total	8212456.585	1500			

Descriptives

Shipping Cost

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	Upper Bound	Minimum	Maximum
North	304	52.73850	72.105396	4.135526	44.60051	60.87649	.310	449.910
Central	894	54.73472	74.915222	2.505674	49.81703	59.65242	.400	652.980
South	303	49.86894	73.224904	4.206664	41.59086	58.14703	.710	593.910
Total	1501	53.34819	73.993047	1.909856	49.60192	57.09446	.310	652.980



One Way ANOVA (F(2,1498)=0.502, p>0.01) ----→ Do not reject the null hypothesis

Methodology

Hypothesis 1

- H0: Shipping cost on furniture has no significant effect between EU Regions
- H1: There is a significance effect

Shipping Cost

Tukey HSD[a,b]

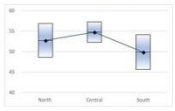
Region	N	Subset for alpha = 0.05 1
South	303	49.86894
North	304	52.73850
Central	894	54.73472
Sig.		.630

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 389.187.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Multiple Comparisons

Dependent Variable: Shipping Cost
Tukey HSD

(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
North	Central	-1.996220	4.914259	.913	-13.52525	9.53281
	South	2.869556	6.008574	.882	-11.22678	16.96589
Central	North	1.996220	4.914259	.913	-9.53281	13.52525
	South	4.865776	4.920307	.584	-6.67745	16.40900
South	North	-2.869556	6.008574	.882	-16.96589	11.22678
	Central	-4.865776	4.920307	.584	-16.40900	6.67745



Methodology

- **Hypothesis 2**
  H0: Profit on furniture has no significant effect between EU Region
  H1: There is a significance effect

ANOVA

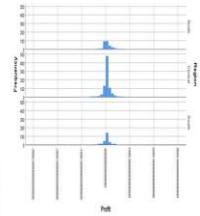
Profit

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	199338.855	2	99669.428	1.389	.250
Within Groups	107480913.5	1498	71749.608		
Total	107680252.4	1500			

Descriptives

Profit

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	Upper Bound	Minimum	Maximum
North	304	67.55936842	267.4214428	17.63185192	32.86268533	102.2560515	-1981.39500	2071.440000
Central	894	42.74069295	255.5233745	8.545969892	25.96914700	59.51223890	-2211.16500	1981.130000
South	303	33.33964356	260.7971559	14.98183865	3.857044705	62.82184242	-1629.54000	1472.210000
Total	1501	45.89691332	267.9305532	6.915634898	32.35447215	59.43489450	-2211.16500	2071.440000



One Way ANOVA (F(2,1498)=1.389, p>0.01) ----→ Do not reject the null hypothesis

Methodology

Hypothesis 2

- H0: Profit on furniture has no significant effect between EU Region
- H1: There is a significance effect

Profit

Tukey HSD[a,b]

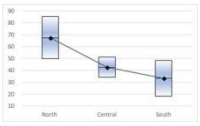
Region	N	Subset for alpha = 0.05 1
South	303	33.33964356
Central	894	42.74069295
North	304	67.55936842
Sig.		.176

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 389.187.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Multiple Comparisons

Dependent Variable: Profit
Tukey HSD

(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
North	Central	24.81867547	17.78413005	.343	-16.9035517	66.54090268
	South	34.21972486	21.74432884	.257	-16.7932753	85.23272498
Central	North	-24.8186755	17.78413005	.343	-66.5409027	16.90355175
	South	9.401049389	17.80601641	.858	-32.3725240	51.17462282
South	North	-34.2197249	21.74432884	.257	-85.2327258	16.79327526
	Central	-9.40104939	17.80601641	.858	-51.1746228	32.37252404



Methodology

- **Hypothesis 3**
  H0: Sales on furniture has no significant effect between EU Regions
  H1: There is a significance effect

ANOVA

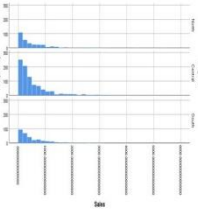
Sales

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	459601.115	2	229800.558	.640	.527
Within Groups	537714608.3	1498	358955.012		
Total	538174209.5	1500			

Descriptives

Sales

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	Upper Bound	Minimum	Maximum
North	304	532.6778224	687.0148881	39.40300618	455.0396373	610.3160075	11.21400000	5451.3000
Central	894	526.1286915	591.6630734	19.78917333	487.2922367	564.9651263	17.55000000	5729.3460
South	303	494.4329737	521.3767262	29.95227586	425.6791396	563.1868078	17.46000000	3893.2700
Total	1501	519.0559907	598.9848131	15.46056871	488.7293673	549.3826140	11.21400000	5729.3460



One Way ANOVA (F(2,1498)=0.640, p>0.01) ----→ Do not reject the null hypothesis

Methodology

Hypothesis 3

- H0: Sales on furniture has no significant effect between EU Regions
- H1: There is a significance effect

Sales

Tukey HSD[a,b]

Region	N	Subset for alpha = 0.05
		1
South	303	484.6207327
Central	894	526.1289815
North	304	532.5778224
Sig.		.504

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 389.187.

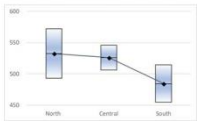
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Multiple Comparisons

Dependent Variable: Sales

Tukey HSD

(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
North	Central	6.448840825	39.77739317	.986	-86.8718031	99.76948480
	South	47.95708970	48.63582089	.586	-66.1443459	162.0585253
Central	North	-6.448840882	39.77739317	.986	-99.7694848	86.87180315
	South	41.50824887	39.82694667	.550	-51.9272418	134.9437396
South	North	-47.95708997	48.63582089	.586	-162.058525	66.14434592
	Central	-41.50824969	39.82694667	.550	-134.943740	51.92724184



Methodology

- **Hypothesis 4**

H0: Discount on furniture has no significant effect between EU Region

H1: There is a significance effect

ANOVA

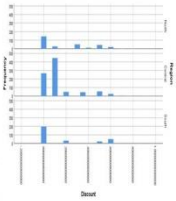
Discount

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.148	2	.574	15.266	.000
Within Groups	56.339	1498	.038		
Total	57.488	1500			

Descriptives

Discount

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	Upper Bound	Minimum	Maximum
North	304	.2019736842	.23064564 13	.01322643 48	.1759424515	.2280049169	.0000000000	.8000000000
Central	894	.1314876957	.15973964 28	.00534245 31	.1210024689	.1419729226	.0000000000	.8500000000
South	303	.1587458746	.24026973 96	.01380313 59	.1315833705	.1859083787	.0000000000	.6000000000
Total	1501	.1512658228	.19576788 12	.00505302 09	.1413540990	.1611775595	.0000000000	.8500000000



One Way ANOVA (F(2,1498)=15.266, p<0.01) ---→ Reject the null hypothesis

Methodology

Hypothesis 4

- H0: Discount on furniture has no significant effect between EU Regions
- H1: There is a significance effect

Discount

Tukey HSD[a,b]

Region	N	Subset for alpha = 0.05	
		1	2
Central	894	.1314876957	
South	303	.1587458746	
North	304		.2019736842
Sig.		.122	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 389.187.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

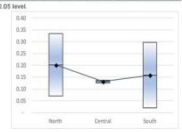
Multiple Comparisons

Dependent Variable: Discount

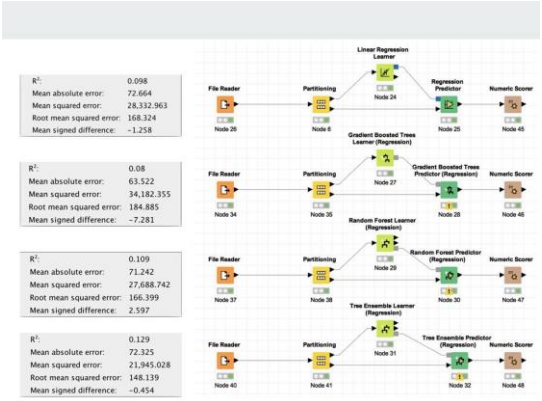
Tukey HSD

(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
North	Central	.070485988*	.012875749 6	.000	.0402790034	.1006929735
	South	.043227810*	.015742942 3	.017	.0062942844	.0801613349
Central	North	-.070485988*	.012875749 6	.000	-.1006929674	-.0402790003
	South	-.027258179	.012891595 4	.087	-.0575023390	.0029859810
South	North	-.043227810*	.015742942 3	.017	-.0801613351	-.0062942844
	Central	.027258188	.012891595 4	.087	-.0029859981	.0575023387

*. The mean difference is significant at the 0.05 level.



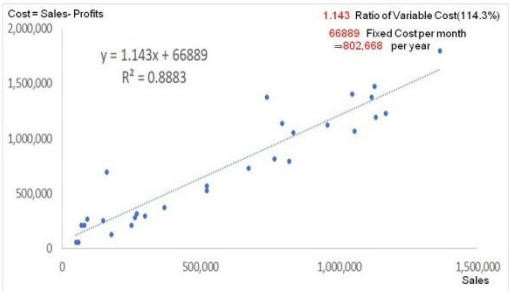
Modeling



Model Comparisons

		Linear Regression	Gradient Boosted Trees	Random Forest	Tree Ensemble
		Model 1	Model 2	Model 3	Model 4
Without Normalizer	R^2	0.098	0.08	0.109	0.129
	MAE	72.664	63.522	71.242	72.325
	MSE	28,332.963	34,182.355	27,688.742	21,945.028
	RMSE	168.324	184.885	166.399	148.139
	MSD	-1.258	-7.281	2.597	-0.454
		Model 5	Model 6	Model 7	Model 8
With Normalizer	R^2	0.086	0.097	0.108	0.105
	MAE	0.005	0.004	0.005	0.005
	MSE	0	0	0	0
	RMSE	0.012	0.011	0.011	0.011
	MSD	-0	-0	0	0

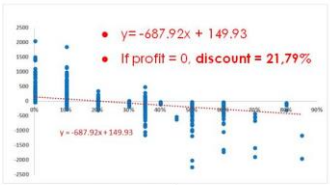
Example: Correlation between Sales and Cost of Central EU region Table



Discount and Profit on furniture in EU

SUMMARY

Regression Statistics	
Multiple R	0.5026382
R Square	0.2526452
Adjust R Square	0.2521466
St Error	231.70245
Observations	1501

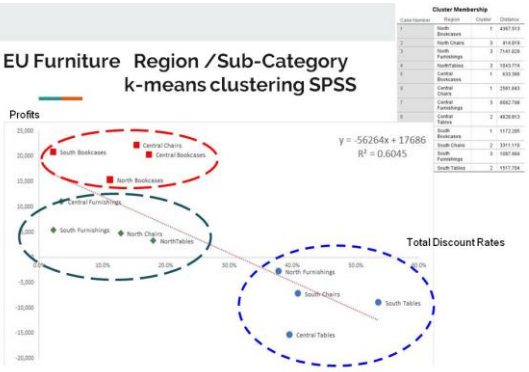


ANOVA

	df	ss	MS	F	Significant F
Regression	1	27204899	27204899.4	506.74079	6.6102E-97
Residual	1499	80475353	53686.026		
Total	1500	1.08E+08			

	Coefficient	St Error	Stat t	P-value	95% Lower	95% Upper
Intercept	149.92795	7.558775	19.8349538	6.19E-78	135.101054	164.754851
Discount	-687.91771	30.55931	-22.510904	6.61E-97	-747.861263	-627.974163
			|t|>2	p<0.05		

EU Furniture Region /Sub-Category k-means clustering SPSS



EU Furniture Region /Sub-Category

Results

Improve the Discount Rate about EU Furniture Profit Region / Sub-Category in the red convert to plus

Discount Rate

	Bookcases	Chairs	Furnishings	Tables
North	11.3%	12.9%	37.8%	18.0%
Central	17.3%	15.4%	3.5%	39.6%
South	2.2%	40.9%	2.2%	53.6%

Discount Rate After Improvement profit convert to plus

	Bookcases	Chairs	Furnishings	Tables
North	11.3%	12.9%	24.3%	18.0%
Central	17.3%	15.4%	3.5%	31.5%
South	2.2%	29.9%	2.2%	35.4%

Difference

	Bookcases	Chairs	Furnishings	Tables
North			13.6%	
Central				8.0%
South		11.1%		18.2%

Summary & Recommendation

Summary and Recommendation

- In the report, we found that ...
  - Discount rates on furniture are significantly different in EU regions and have impact on profit in EU
  - Discount rates are too high for Furnishings in North Europe, Tables in Central Europe, Chairs and Tables in South Europe
- So we suggest the Superstore to ...
  - Decrease the discount rates for Furnishings in North Europe, Tables in Central Europe, Chairs and Tables in South Europe by 13.6%, 8%, 11.1% and 18.2%, respectively

附錄 2：上課情形及合影

圖 1　Robert Nisbet 博士(後排左三)

圖 2　Ash Pahwa 博士(左)

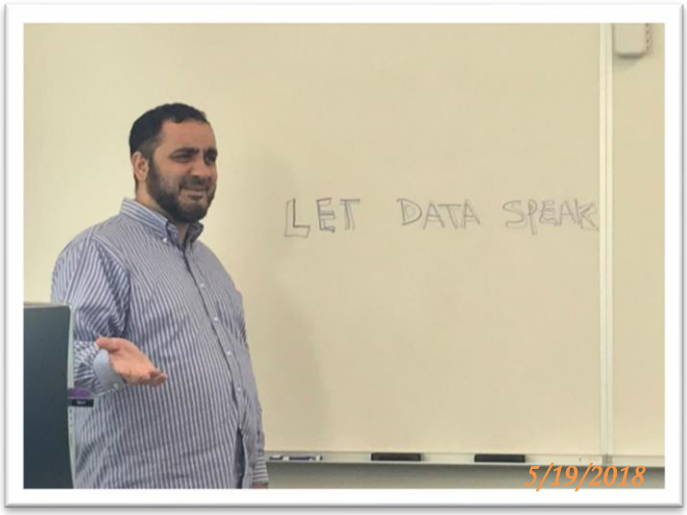


圖 3　Hasan Hboubati 顧問

圖 4　Alfred Nigl 博士(後排左四)

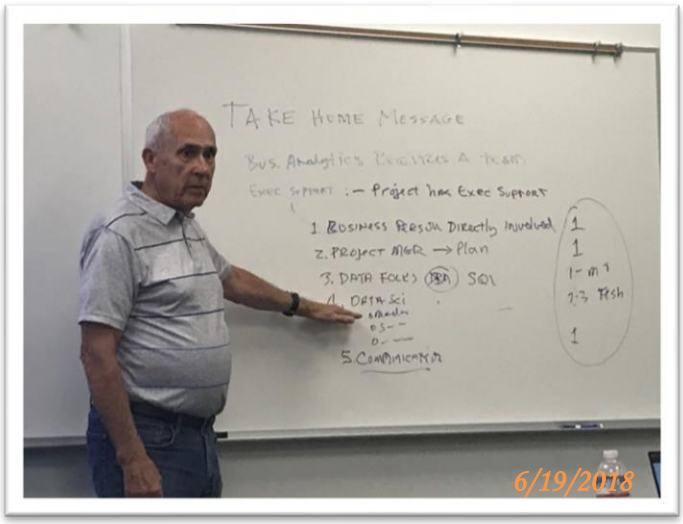


圖 5　Kenneth Reed 博士