第一屆 東森盃*BigData*校園爭霸戰

東森購物網客戶行為分析與商品推薦報告

團隊成員：黃伊、陳曦、王羽柔

2014年5月

Contents

東森購物網是一個綜合性的購物平台，為客戶提供數萬種不同類別的商品。本報告以 2013 年 11、12 月與 2014 年 1 月三個月的東森購物網瀏覽記錄、2013 年 11、12 月兩個月的交易記錄、客戶輪廓資料和商品分類資訊為基礎，涵蓋 18 萬個會員和 1 萬個商品分類，分析這些會員的瀏覽購買習慣和商品偏好，建置客戶購買預測模型和商品推薦模型，從而協助網站更好的進行精準行銷，創造出更多的交易機會。本報告主要分為兩部份，第一部份為瀏覽量/購買量分析，第二部份為建置購買預測模型和商品推薦模型的過程說明。分析部份使用軟體為 SAS EG，建模部份使用軟體為 SAS EM。

1 瀏覽量/購買量分析

在這一部份，我們提取 2013 年 11 和 12 月份的瀏覽記錄及交易記錄，分析不同時間區段和不同種類客戶的瀏覽量/購買量差異，以發現客戶的瀏覽購買習慣和商品偏好。分析部份又分為基礎分析和交岔分析，基礎分析主要針對單個變數進行瀏覽量/購買量差異分析，交岔分析則結合多個變數進行瀏覽量/購買量的交岔分析。

1.1 瀏覽量/購買量基礎分析

瀏覽量/購買量基礎分析的目的在于發現不同時間區段，如一週七天內或一天 24 小時內瀏覽人次和購買人次的變化情況，以及不同群體客戶，如不同年齡段或不同地區的客戶瀏覽量和購買量的差異。

1.1.1 24 小時內瀏覽人次/購買人次分析

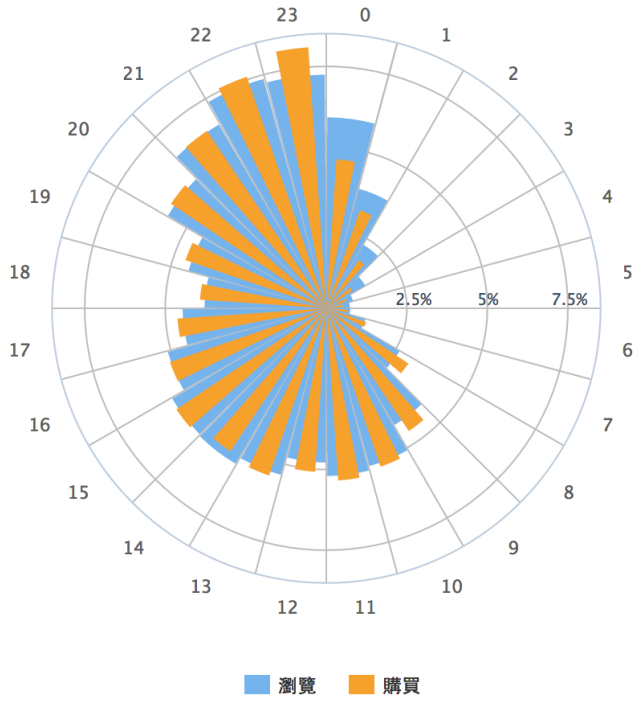

Figure 1: 24 小時流量分佈

瀏覽人次和購買人次在 24 小時內有明顯的變化，圖 1 所展示的是某小時的瀏覽量/購買量占全天 24 小時瀏覽量/購買量的比例，其中藍色的部份代表瀏覽的比例，黃色的部份代表購買的比例。總體來看，瀏覽量和購買量隨著日常生活作息習慣變化，主要集中于早上 8 點至晚上 24 點。瀏覽量和購買量在早上 8 點后開始逐漸上升，在午間用餐時間有細微下降，用餐時間結束后，瀏覽量和購買量回升至下午 3 點，三點后，瀏覽量和購買量逐漸下降，在晚間通勤和用餐時降至一個低谷，晚餐結束后，瀏覽量和購買量又開始逐步上升，并在晚間的 22、23 點到達一天的最高點，零點后，瀏覽量和購買量開始顯著下降，在凌晨 5、6 點到達一天中的最低點。從早上 8 點到晚上 24 點，絕大多數的時段購買量所佔比大於瀏覽量所佔比，這說明在這些時段，會員在瀏覽之後也會產生一定的購買行為，特別是在晚間 23 點，購買量的所佔比遠大於瀏覽量，會員可能在之前已選定偏好商品，在這一時段不需要更多的瀏覽而傾向于直接購買商品。而在零點過後到凌晨 6 點這段時間里，會員更傾向於只看不買，購買量所佔比明顯低於瀏覽量所佔比。

1.1.2 一週內瀏覽人次/購買人次分析

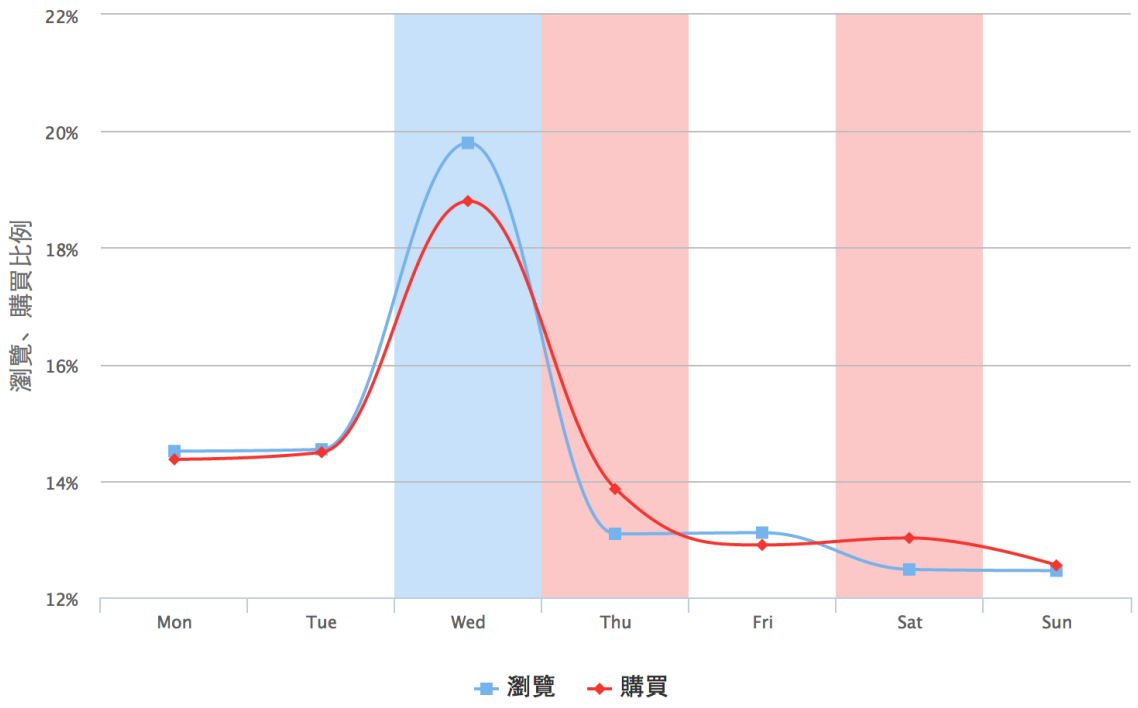

Figure 2: 一週流量分佈

在一週七天中，瀏覽人次/購買人次所佔總瀏覽人次/購買人次在週三有一個明顯的高點 (見圖 2)，週三過後，瀏覽人次/購買人次明顯下降至週末，隨著新的一週開始，瀏覽人次/購買人次在週一有一定的回升，一週的瀏覽人次/購買人次最低點出現在週日。在一週中，週三的瀏覽人次雖然最高，但瀏覽所佔比明顯高於購買所佔比，相對更多的客戶傾向于瀏覽而不是購買，而在一天後的週四，購買所佔比明顯高於瀏覽所佔比，相對更多的客戶傾向于購買商品，購買所佔比高於瀏覽所佔比的還有週六。

1.1.3 每日瀏覽人次/購買人次分析

兩個月中，11 月每日的瀏覽人次/購買人次所佔兩個月總瀏覽人次/購買人次的比例沒有較大的波動 (見圖 3)。而在 12 月中旬瀏覽量和購買量都有一個較為明顯的起伏，在十二月末，瀏覽量和購買量有一個非常明顯的上升，在 12 月 31 號瀏覽量和購買量達到兩個月的最高點，這一天的瀏覽量和購買量是平日的 3 倍左右，東森購物網可能在這些時段進行了一定的促銷活動。兩個月中的大部份日期瀏覽所佔比高於購買所佔比。(在所提供的瀏

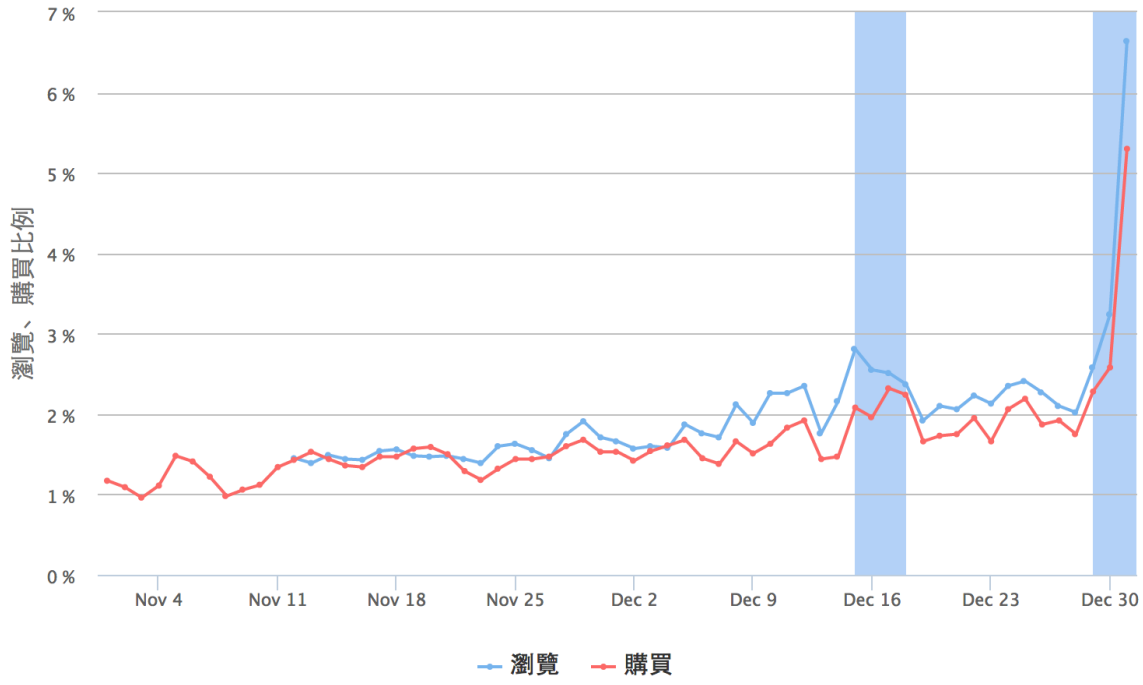

Figure 3: 每日流量分佈

覽記錄中我們發現 11 月 12 日前的數據出現了整日遺失或部份遺失的情況，所以在每日瀏覽人次/購買人次的分析中，并沒有包含 11 月 12 日之前的瀏覽數據。)

1.1.4 不同瀏覽器瀏覽量/購買量分析

不同瀏覽器所帶來的瀏覽量和購買量的分佈非常類似 (見圖 4 左圖)。無論是瀏覽量還是購買量，所佔比最大的是 IE 瀏覽器，均超過 50%，其次是 Chrome 瀏覽器，所佔比在 25% 左右。我們將原始數據中的 iPhone、iPad、Android 合併成一個大類，統稱為 Mobile，瀏覽量和購買量所佔總體比例均在 6% 左右。圖 4 右圖所展示的是不同瀏覽器的點擊購買轉化率。去除樣本數較小的 Others 分類，在剩下的 6 類瀏覽器中，點擊購買轉化率超過平均水平的共有 3 類，分別是 IE、Linux 和 Mobile，特別是 Mobile 類的轉化率非常突出，可以考慮推出東森購物手機 app。

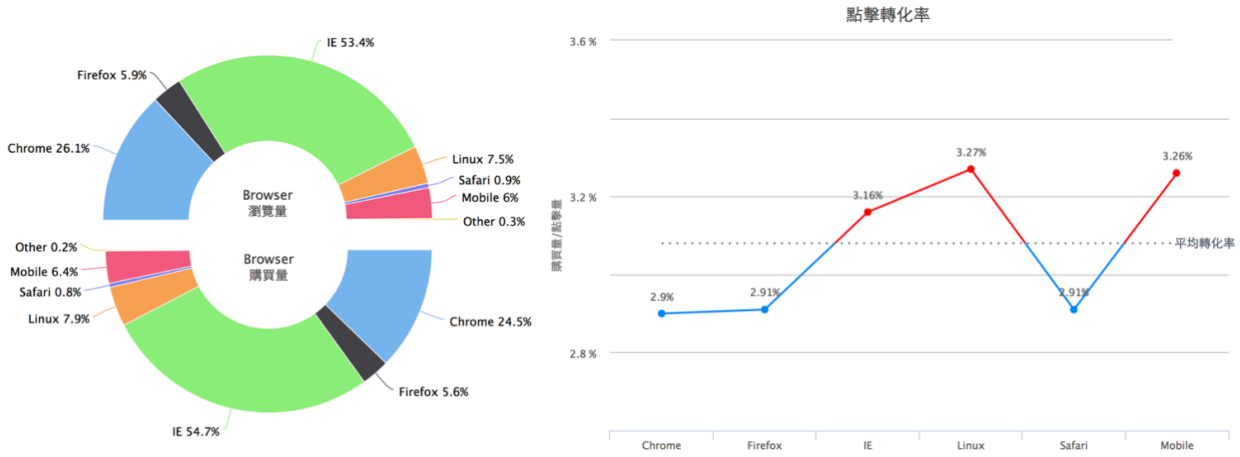

Figure 4: 不同瀏覽器瀏覽量/購買量分佈

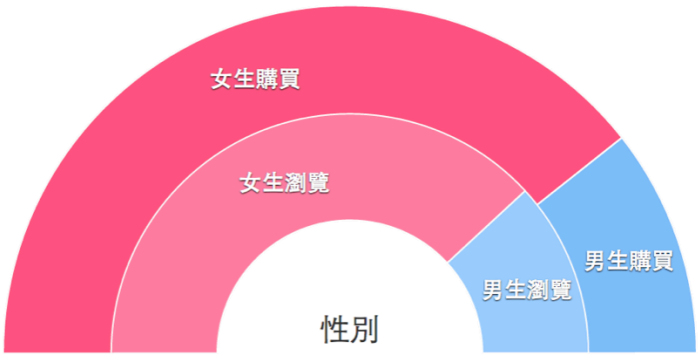

Figure 5: 不同性別瀏覽量/購買量分佈

1.1.5 不同性別瀏覽量/購買量分析

從圖 5 可以看出，女性的瀏覽量和購買量都遠高于男性，所佔總體瀏覽量和購買量的比例均超過 70%。同時，女性購買所佔比略高于女性瀏覽所佔比，說明女性在瀏覽之後更傾向于購買，而男性在瀏覽之後，購買的可能性要低於女性。

1.1.6 不同星座瀏覽量/購買量分析

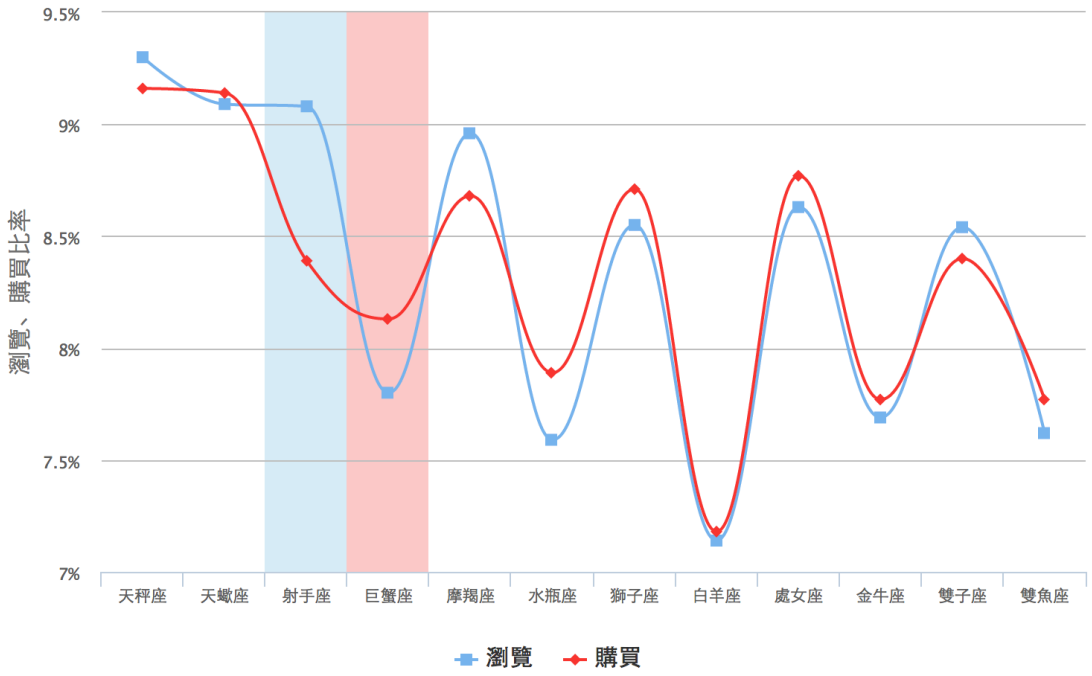

Figure 6: 不同星座瀏覽量/購買量分佈

不同星座客戶的瀏覽量和購買量所佔總體比例的分佈較為類似 (見圖 6)。在十二星座中，天秤座的瀏覽量/購買量所佔總體瀏覽量/購買量的比例最高，白羊座的所佔比最低。大部份的星座瀏覽量/購買量所佔總體比例的差異都不大，瀏覽量/購買量所佔比差異最大的星座為射手座和巨蟹座，其中射手座的瀏覽量所佔比明顯高於購買量所佔比，說明射手座客戶的每一次購買前可能需要高於平均次數的瀏覽或更傾向于只逛不買，而巨蟹座客戶購買量所佔比明顯高於瀏覽量所佔比，說明這些客戶可能在瀏覽網站前就已有較明確的購買目標或更傾向于衝動消費。

1.1.7 不同地區瀏覽量/購買量分析

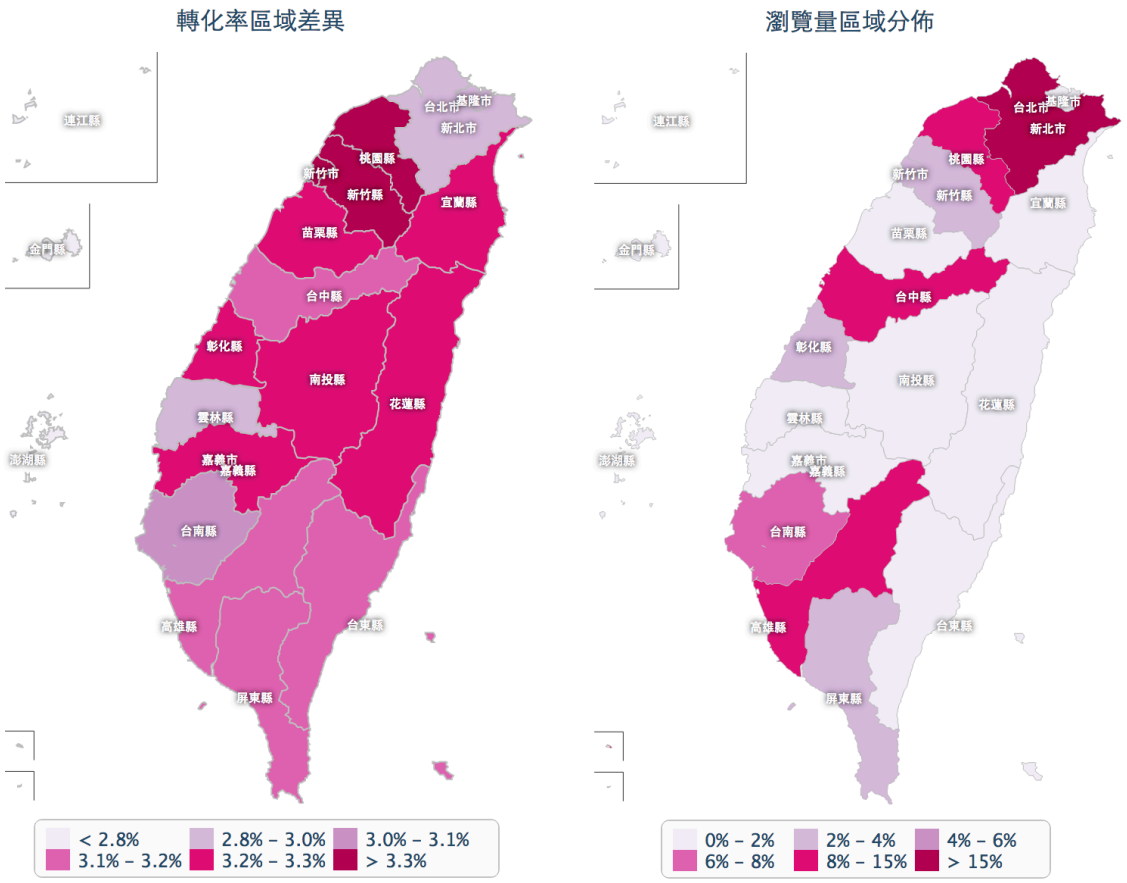

Figure 7: 不同地區瀏覽量/購買量分佈

不同地區的客戶瀏覽量所佔總瀏覽量的比例有較為明顯的差異，從圖 7 左圖可以看出，經濟較為發達的地區對應的網站瀏覽量也相對較大，台北市和新北市是全台瀏覽東森購物網最多的地區，其次是台中縣、桃園縣和高雄縣。按地理位置看，北部的瀏覽量高於南部，西部的瀏覽量明顯高於東部，離島的瀏覽量非常少。那麼瀏覽量高地區的點擊購買轉化率也相對應較高嗎？不一定。從圖 7 右圖可以發現，瀏覽量最大的台北市和新北市並不是點擊購買轉化率最大的地區，東部的宜蘭花蓮由於瀏覽量較少，點擊購買轉化率都高於台北市和新北市。點擊購買轉化率最高的地區為桃源縣、新竹縣和新竹市，最低的是離島地區，在本島點擊購買轉化率較低的地區則是雲林縣、台北市、新北市和基隆市。

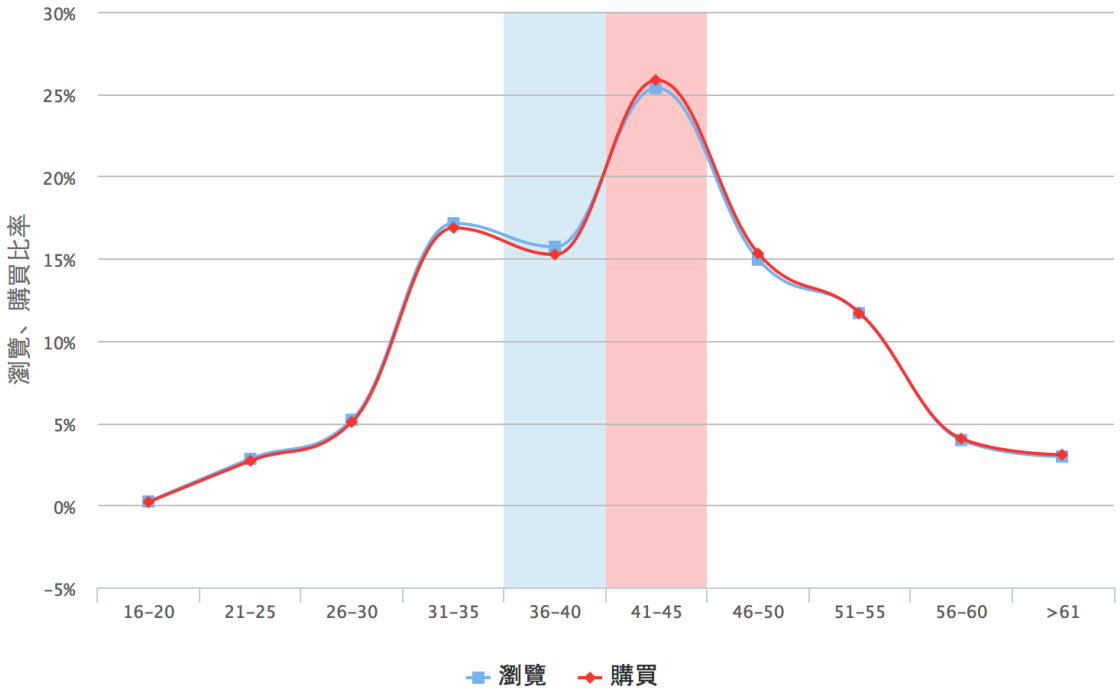

Figure 8: 不同年齡段瀏覽量/購買量分佈

1.1.8 不同年齡段瀏覽量/購買量分析

從圖 8 可以看出，不同年齡段的瀏覽量所佔總體比例和購買量所佔總體比例高度相似。瀏覽量和購買量隨著年齡的增長大致呈現先上升后下降的趨勢，處於 41-45 歲這一年齡層的客戶瀏覽量和購買量最大。值得關注的是，36-40 歲這一年齡段的客戶瀏覽量和購買量出現了細微的回落，可能原因是這一年齡段的人消費減少或者更傾向于線下購物。

1.1.9 商品瀏覽量/購買量 **top10**

東森購物網上的商品分類採用層次結構，本報告在進行商品分析時選擇細化到中分類。客戶瀏覽最多的商品集中於視聽家電、服飾鞋包和民生日用這幾個館別 (圖 9)。從瀏覽量 top10 的清單中可以看到，視聽家電和民生日用館別的瀏覽量較為成功的轉化為了購買行為，購買量 top10 的清單中也有這些商品的蹤影，而瀏覽量 top10 的清單中服飾鞋包這個館別下有四個中分類上榜，但只有一個上榜購買量 top10，說明客戶在服飾鞋包這個館別下更傾向于多瀏覽少購買。值得一提的是，珠寶精品館別中也有一類商品上榜瀏覽量 top10，但購買量排行和瀏覽量排行差異非常明顯。

	類別瀏覽 TOP 10	類別購買 TOP 10
1	視聽家電 - 廚房家電	視聽家電 - 廚房家電
2	內著塑衣 - 流行內衣褲	內著塑衣 - 流行內衣褲
3	服飾鞋包 - 流行女鞋	美妝保养 - 乳霜/凝膠
4	服飾鞋包 - 女裝上著	視聽家電 - 生活家電
5	服飾鞋包 - 機能運動服	美妝保养 - 藥妝
6	服飾鞋包 - 女裝外套	美妝保养 - 身體保養
7	視聽家電 - 生活家電	美妝保养 - 美髮造型
8	珠寶精品 - 國際精品	服飾鞋包 - 女裝上著
9	民生日用 - 鍋具/碗盤	民生日用 - 鍋具/碗盤
10	民生日用 - 淨水/杯瓶	風味美食 - 零嘴餅乾

Figure 9: 商品瀏覽量/購買量 top10

客戶購買最多的商品集中於視聽家電、美妝保養等館別。購買量 top10 的清單中大多數商品的購買量排行和瀏覽量排名差異不大，排名差異較大的是風味美食館別下的堅果/果乾分類，點擊購買轉化率比較高，客戶在購買時普遍瀏覽次數較少。

1.1.10 客戶每次上站瀏覽平均頁面數/平均停留時間分析

我們將原始數據中每個 session 所包含的行為定為每次上站所產生的行為，在 11 和 12 兩個月中，客戶每次上站瀏覽的平均頁面數為 13 頁，標準差為 18 頁，每次上站瀏覽的平均停留時間為 13 分鐘，標準差為 38 分鐘。在交岔分析部份有更多關於瀏覽平均頁面數和平均停留時間的分析。

1.2 瀏覽量/購買量交岔分析

結合不同的變數進行交岔分析，能幫助我們更深入的分析不同群體客戶的瀏覽習慣、購買習慣和商品偏好，從而更好的建置客戶預測模型。

1.2.1 不同性別商品瀏覽量/購買量 **top5**

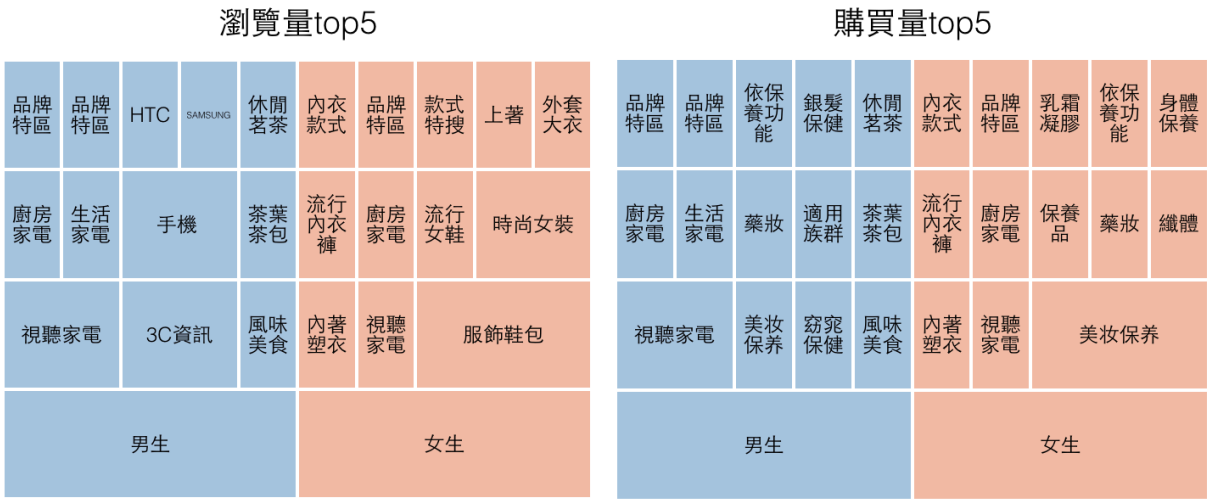

Figure 10: 不同性別商品瀏覽量/購買量 top5

從圖 10 可以看出，不同性別客戶的商品瀏覽量/購買量 top5 清單有明顯的差異。圖 10 從下往上依次為性別分類，不同性別商品瀏覽量/購買量 top5 所屬館別，不同性別商品瀏覽量/購買量 top5 所屬大分類和不同性別商品瀏覽量/購買量 top5 中分類。女性客戶瀏覽最多的為內著塑衣館和視聽家電館下的商品，購買最多的也是這兩類商品。女性客戶也非常喜愛瀏覽服飾鞋包館下的商品，但更喜愛購買美妝保養館下的商品。由於女性的瀏覽量和購買量都遠大于男性，女性客戶的瀏覽量/購買量 top5 清單與所有客戶的瀏覽量/購買量 top10 清單重疊。

相較於女性，男性更喜愛瀏覽視聽家電館和 3C 資訊館下的商品，同時，男性也經常瀏覽風味美食館下的茶葉類別。從男性購買量 top5 清單中可以發現，視聽家電館別下和風味美食館下的瀏覽行為更好的轉化成購買行為，而在 3C 資訊館更傾向于只逛不買。

1.2.2 不同年齡段商品瀏覽量/購買量 **top5**

不同年齡段的客戶的商品瀏覽喜好不同，這裡我們將 10 個年齡段合併為 5 個以便更好的描述年齡對瀏覽商品喜好的影響 (見圖 11)。在 45 歲前，客戶瀏覽最多的是內著塑衣館別下的商品，而在 45 歲之後，瀏覽量 top5 的清單中完全沒有這個館別下的商品。所有年齡段的客戶都喜歡瀏覽廚房家電這一類型的商品，而對於同在一個館別下的生活家電類商

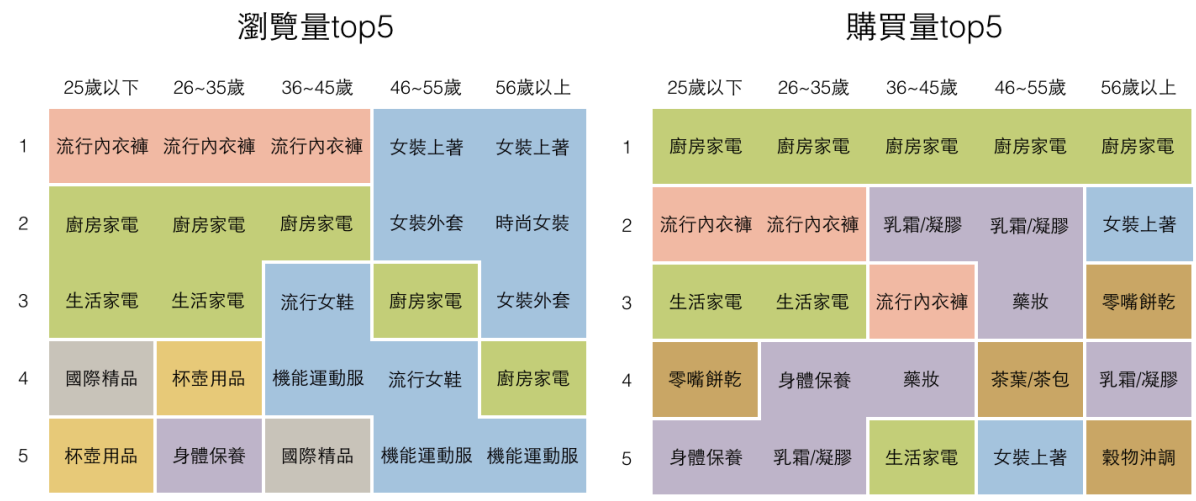

Figure 11: 不同年齡段商品瀏覽量/購買量 top5

品，35 歲以下的客戶更喜歡瀏覽。35 歲之後，客戶才逐漸開始喜好瀏覽服飾鞋包館別下的商品，服飾鞋包館別下的上榜商品數量的逐漸增加。

雖然不同年齡段的客戶購買最多的都是視聽家電館下的廚房家電類商品，但年齡的變化也在影響著購買商品的喜好。16 – 25 歲的年輕客戶喜歡購買風味美食館下糖果和巧克力類的商品，25 歲以上 45 歲以下的客戶相對來說沒有那麼喜好風味美食館下的商品，風味美食再次出現在購買量 top5 的清單上是在 46 – 55 這一年齡段，但具體的商品種類已經轉向相對健康的茶葉類，特別是到 56 歲以上這個年齡段，他們更傾向于購買堅果類和穀物沖調類的商品。雖然每個年齡段的客戶都喜好購買美妝保養館別下的商品，但購買量的排名有所不同，美妝保養館商品的排名大致隨著年齡的上升呈現先升後降的趨勢。

1.2.3 不同性別 **24** 小時瀏覽量/購買量分析

從圖 12 很明顯可以看出，無論是哪個時段，女性客戶的瀏覽量和購買量都遠遠大於男性。在清晨 5 點到 7 點這段時間，男性客戶的瀏覽量占總體比例相對于其他時段有明顯的提升，而在凌晨 3 點到 5 點，男性客戶的購買量占總體比例相對于其他時段有所提升，不知是男性客戶起得比較早還是睡得比較晚呢？

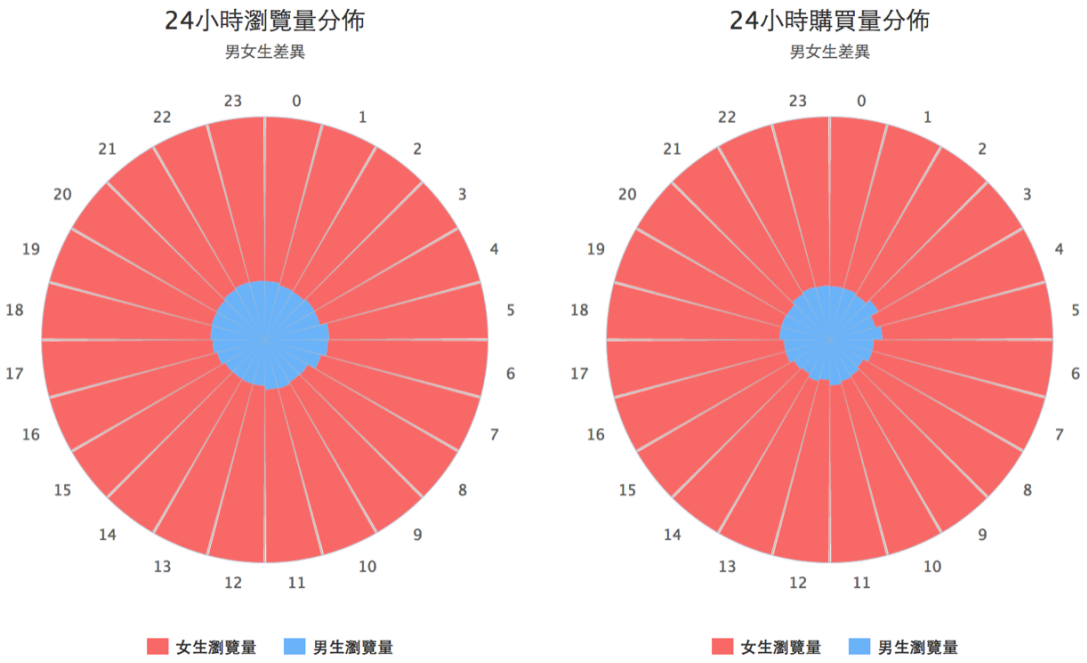

Figure 12: 不同性別 24 小時瀏覽量/購買量分佈

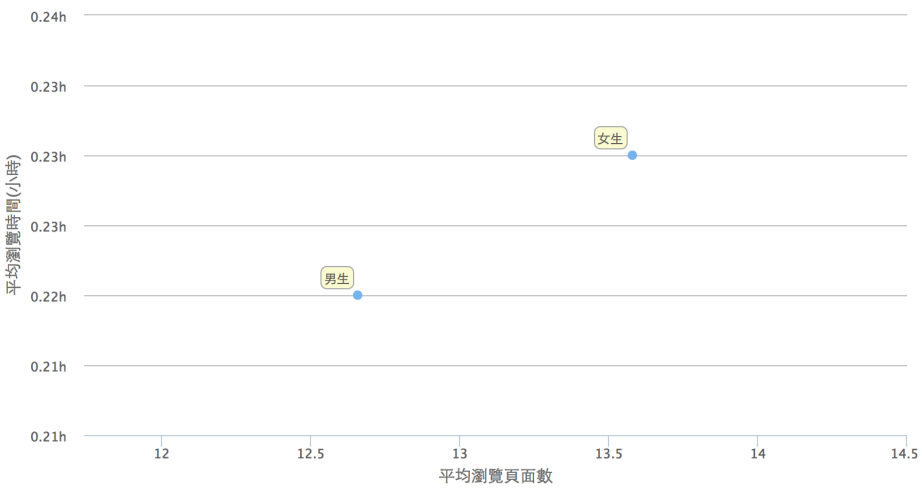

Figure 13: 不同性別客戶每次上站平均瀏覽頁面數/平均停留時間分佈

1.2.4 不同性別客戶每次上站平均瀏覽頁面數/平均停留時間分析

女性客戶的上站平均瀏覽頁面數和平均停留時間均大於男性，但從圖 13 也可以看出，這樣的差異並不明顯，上站平均瀏覽頁面數女性只比男性多了一頁，平均停留時間只比男性多了 0.6 分鐘。

1.2.5 不同地區客戶每次上站平均瀏覽頁面數/平均停留時間分析

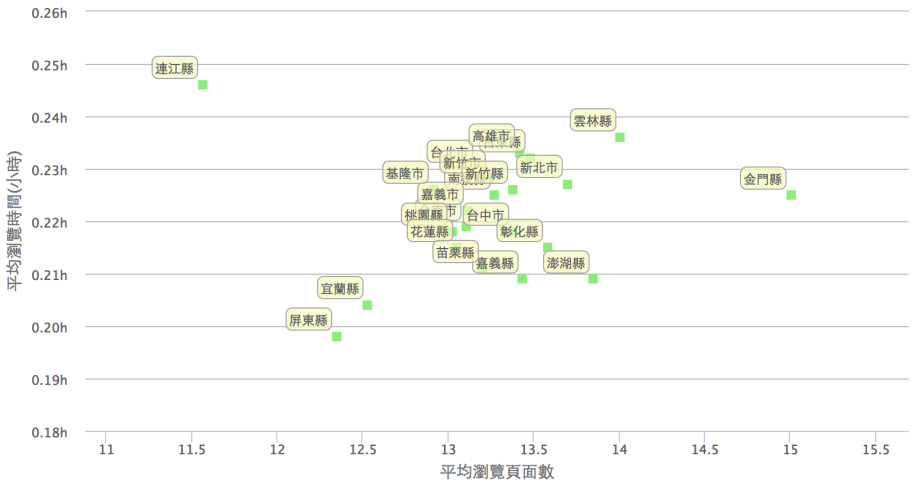

Figure 14: 不同地區客戶每次上站平均瀏覽頁面數/平均停留時間分佈

不同地區客戶每次上站平均瀏覽頁面數和平均停留時間的分佈較為集中（圖 14），大部份地區的客戶每次上站瀏覽平均頁面為 13 頁左右，平均停留時間為 13 分鐘左右。連江縣和金門縣雖然與其他地區差異較大，但由於這兩個縣樣本量少，不具有代表性。

1.2.6 不同年齡層客戶每次上站瀏覽平均頁面數/平均停留時間分析

不同年齡層在每次上站平均瀏覽頁面數和平均停留時間有一定的差異 (見圖 15)。16-25 歲的年輕客戶和大於 55 歲的客戶每次上站瀏覽的頁面數明顯少於其他年齡層，差異在 2-3 頁左右。25 歲以下的年輕客戶在平均瀏覽時間上也明顯少於其他年齡層，其他年齡層的平均瀏覽時間在 13 分鐘左右，而年輕客戶則在 10 分鐘左右。

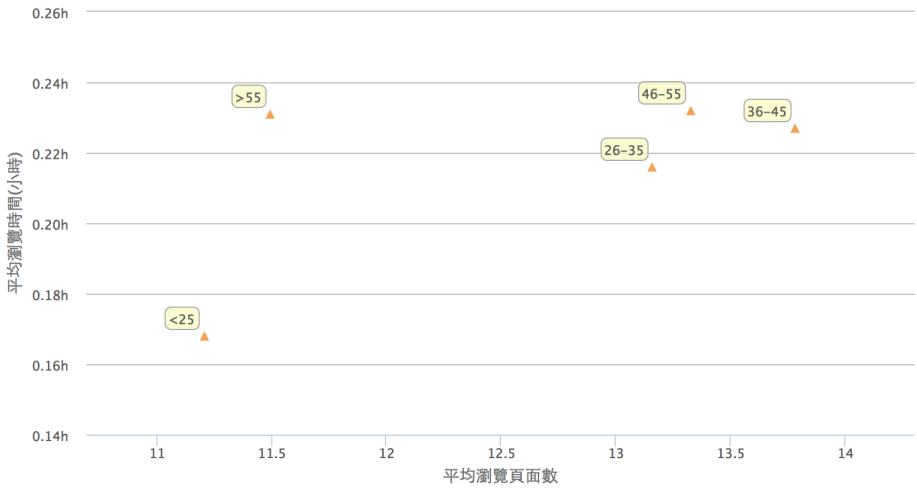

Figure 15: 不同年齡層客戶每次上站平均瀏覽頁面數/平均停留時間分佈

1.2.7 不同星座客戶每次上站平均瀏覽頁面數/平均停留時間分析

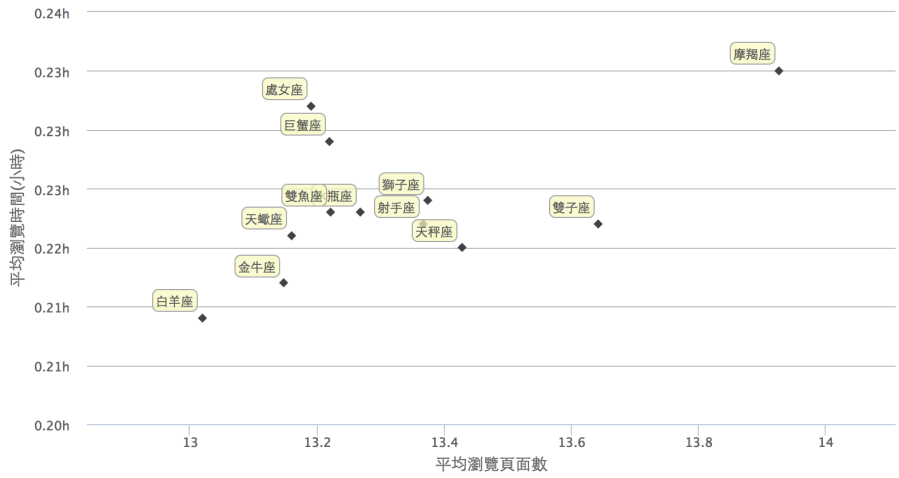

Figure 16: 不同星座客戶每次上站平均瀏覽頁面數/平均停留時間分佈

十二星座中，平均瀏覽頁面數和平均瀏覽時間最大的均為摩羯座，最小的均為白羊座，但差異也不大，頁面數的差異在一頁左右，瀏覽時間的差異在 1 分鐘左右，因此，總體來說，不同星座客戶每次上站平均瀏覽頁面數和平均停留時間的分佈較為集中，沒有非常明顯的差異 (見圖 16)。

2 購買預測模型

根據第一部份的分析結果，下面將針對資料採礦步驟 SEMMA 詳述本次客戶購買預測模型的建立過程。

2.1 樣本抽樣

在模型建置階段，我們主要使用 2013 年 11、12 月與 2014 年 1 月三個月的東森購物網瀏覽記錄和 2013 年 11、12 月兩個月的交易記錄這兩份資料，以 2013 年 11 月、12 月的瀏覽記錄和交易記錄為建模資料，2014 年一月的瀏覽記錄為預測資料。在建模資料中我們採用隨機抽取的辦法，抽取其中的 80% 作為訓練集 (Train)，剩餘的 20% 作為測試集 (Test)。

2.2 變數觀察

在建置購買預測模型階段，我們從 11、12 每個月的瀏覽數據中，提取某個用戶該月的瀏覽次數和搜索次數作為輸入變量 (input), 從 11、12 每個月的交易記錄中，提出中用戶在該月的購買次數作為目標變量 (target), 也就是說以月份為單位，我們希望通過一個月的瀏覽數據去預測當月的購買量。

2.3 模型建置

由於目標變數屬於連續性，我們選擇線性迴歸和決策樹這兩種模型。

2.3.1 線性迴歸結果

從圖 17 線性迴歸的結果來看，瀏覽次數與搜索次數與購買次數呈現正相關關係，我們選取的兩個變數和截距的 P 值都非常小，通過顯著性檢驗。

Parameter	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	1.2432	0.00701	177.42	<.0001
n_view	1	0.00246	0.000500	4.93	<.0001
n_search	1	0.00609	0.000067	90.82	<.0001

Figure 17: 線性迴歸結果表

2.3.2 決策樹結果

由於決策樹的結點過多，在此不做結果展示。

2.4 模型評估

Model	Train	Test
決策樹	2.108964	1.865408
線性迴歸	2.172711	1.952975

Figure 18: 模型 Average Squared Error 結果表

在模型評估階段，我們選擇 Average Squared Error 作為評估標準。由於決策樹模型在訓練集和測試集都有較小的誤差，所以選取決策樹作為我們最終的模型並將模型運用至 1 月份的資料上，最終選取預測購買量最大的三千客戶作為最終的客戶預測名單。

3 商品推薦模型

在建置客戶購買預測模型之後，下面我們將詳述商品推薦模型。本次商品推薦模型的目標是為客戶推薦 1 個最可能購買的商品中分類。

3.1 變數觀察

為了預測用戶最後可能購買的商品，我們選擇預測在特定日期某個客戶是否會購買特定的商品分類，也就是說，我們通過分析客戶對特定商品分類在特定時間點上的狀態，來預測客戶在這個時間點上是否會購買該商品分類。為了完整的概括客戶對特定商品分類在特定時間點上的狀態，我們將客戶、商品分類和時間點這三個因素做交岔分析，延伸出了以下新的變數 (圖 19 第三行)，這些欄位根據性質不同，分別屬於六張不同的表。這些延伸變數為商品推薦模型的輸入變數 (input)，目標變數則通過統計客戶在某天對某商品分類是否有交易行為得出，目標變數的屬性為二元變數。下面將詳述延伸表及其對應的延伸變數的意義。

name	**key**	**variables**
1. 客戶行為表	UID,Cate_ID,Date	瀏覽次數，是否搜索，累積瀏覽次數，累積瀏覽天數，占當天瀏覽所屬大分類的比例
2. 客戶商品類別喜好表	UID,Cate_ID	瀏覽次數，瀏覽天數
3. 客戶行為變化表	UID,Date	瀏覽次數，瀏覽商品類別個數
4. 商品類別熱度變化表	Cate_ID,Date	被瀏覽次數，被瀏覽客戶數
5. 客戶描述表	UID	瀏覽次數，瀏覽天數，購買商品數
6. 商品類別描述表	Cate_ID	被瀏覽次數，被瀏覽客戶數，被購買次數

Figure 19: 延伸變數表

3.1.1 客戶行為表

客戶行為表是對每個客戶在每一天對其有過交互行為的所有商品中分類的行為統計，所延伸變數有瀏覽次數（某個客戶在某天共瀏覽某中分類的次數)、是否搜索（某個客戶在某天在瀏覽某中分類的同一個 session 內是否有過搜索行為)、累計瀏覽次數 (某個客戶在這一天前對某中分類的累計瀏覽次數)、累計瀏覽天數（某個客戶在這一天前對某中分類的累計瀏覽天數）和占當天瀏覽所屬大分類的比例。

3.1.2 客戶商品類別喜好表

客戶商品類別喜好表統計的是客戶對特定商品類別的喜好程度，所延伸變數有瀏覽次數（某個客戶對某個商品類別在三個月內的總瀏覽次數）和瀏覽天數（某個客戶對某個商品類別在三個月內的總瀏覽天數）。

3.1.3 客戶行為變化表

客戶行為變化表統計客戶每日的行為，主要反映客戶在不同時間點上是否有活躍度的變化，所延伸變數有瀏覽次數（某個客戶在某天的總瀏覽次數）和瀏覽商品類別數（某個客戶在某天所瀏覽的中分類個數）。

3.1.4 商品類別熱度變化表

商品類別熱度變化表統計商品類別在每天的熱度，主要反映商品類別在不同時間點上是否存在熱度的變化，所延伸變數有被瀏覽次數（某個中分類在某天被瀏覽的次數）和被瀏覽客戶數（某個中分類在某天在多少個客戶瀏覽）。

3.1.5 客戶描述表

客戶描述表統計的是客戶自身的行為，以衡量客戶的購買習慣和瀏覽習慣，所延伸變數有瀏覽次數（某一客戶在三個月內對所有中類別的總瀏覽次數）、瀏覽天數（某一客戶在三個月內對所有中類別的總瀏覽天數）和購買商品數（某一客戶在 11 和 12 這兩個月中購買的商品總數）。

3.1.6 商品類別描述表

商品類別描述表統計的是商品類別的資訊，以衡量商品類別的總體熱度，點擊購買轉化比等訊息，所延伸變量有被瀏覽次數（某一中分類在三個月內被瀏覽的總次數）、被瀏覽客戶數（某一中分類在三個月內總共被多少個客戶瀏覽）和被購買次數（某一中分類在 11 和 12 這兩個月里被購買的總次數）。

3.1.7 最終資料表

最終用於建模和預測的資料表通過合併以上六張表得到，最終資料表 (如圖 20 所示) 中的一條記錄 (UID,Cate_ID,Date,…,buy) 的含義為某個客戶在某一天和…的狀態下, 對某個商品類別是否產生購買行為。

Variables	**BelongTo**	**Name**
UID	客戶行為表	客戶 id
Cate_ID	客戶行為表	商品分類 id
Date	客戶行為表	產生行為之日期
View	客戶行為表	瀏覽次數
Searched	客戶行為表	是否搜索
Cumview	客戶行為表	累計瀏覽次數
Cumday	客戶行為表	累計瀏覽天數
LRatio	客戶行為表	占當天瀏覽所屬大分類比例
UserCate_total_view	客戶商品類別喜好表	瀏覽次數
UserCate_total_days	客戶商品類別喜好表	瀏覽天數
UserDate_total_view	客戶行為變化表	瀏覽次數
UserDate_total_cate	客戶行為變化表	瀏覽商品類別個數
CateDate_total_view	商品類別熱度變化表	被瀏覽次數
CateDate_total_user	商品類別熱度變化表	被瀏覽客戶數
User_total_view	客戶描述表	瀏覽次數
User_total_days	客戶描述表	瀏覽天數
User_total_buy	客戶描述表	購買商品數
Cate_total_view	商品類別描述表	被瀏覽次數
Cate_total_user	商品類別描述表	被瀏覽客戶數
Cate_total_buy	商品類別描述表	被購買次數
buy		當天是否購買該商品分類

Figure 20: 最終資料表及來源說明

3.2 樣本抽樣

與購買推薦模型相同，我們同樣以 2013 年 11、12 月與 2014 年 1 月三個月的東森購物網瀏覽記錄和 2013 年 11、12 月兩個月的交易記錄為基礎，2013 年 11 月、12 月的瀏覽記錄和交易記錄為建模資料，2014 年一月的瀏覽記錄為預測資料。在建模資料中我們根據目標屬性是否購買採用分層抽樣的辦法，分別在有購買和未購買的資料中抽取 80% 作為訓練集 (Train)，剩餘的 20% 作為測試集 (Test)。

3.3 資料修正

新產生的變數在進行模型訓練前，還有進行標準化的步驟以消除某些極端值對模型的影響。我們選用如下公式將變數的值控制在 0 到 1 的區間內。

$$x_i\_nor = \frac{log(x_i+1)-min(log(x+1))}{max(log(x+1))-min(log(x+1))}$$

3.4 模型建置

由於模型的目標是預測客戶在某一天購買某商品類別的機率，所以挑選羅吉斯迴歸、類神經網路和決策樹為主要模型。

3.4.1 羅吉斯迴歸（逐步迴歸）結果

圖 21 為羅吉斯迴歸（逐步迴歸）最後選擇的變數及權重，其中，與預測結果是否購買有正相關關係的是 Cate_total_buy、Cumview 和 User_total_buy，也就說當一個商品類別被購買的次數增多，客戶對這個商品類別的累計點擊次數增多，用戶購買商品數增多，都會提高商品類別被購買的機率。而與預測結果是否購買負相關的變數是 Cate_total_user、User_total_days、User_total_view 和 UserDate_total_view，說明當一個商品類別被更多的人瀏覽，當用戶長期活躍于購物網，當用戶的瀏覽量增大或在特定時間點瀏覽量增大，都會降低商品類別在特定時間點被購買的機率。

Parameter	Estimate	Pr>ChiSq	BelongTo	Variables
Intercept	0.0844	<.0001		截距
Cate_total_buy	3.8297	<.0001	商品類別描述表	被購買次數
Cate_total_user	-4.5473	<.0001	商品類別描述表	被瀏覽客戶數
Cumview	0.8381	<.0001	客戶行為表	累計瀏覽次數
User_total_buy	4.5540	<.0001	客戶描述表	購買商品數
User_total_days	-2.0699	<.0001	客戶描述表	瀏覽天數
User_total_view	-1.4904	<.0001	客戶描述表	瀏覽次數
UserDate_total_view	-0.6003	<.0001	客戶行為變化表	瀏覽次數

Figure 21: 羅吉斯迴歸（逐步迴歸）結果表

3.4.2 羅吉斯迴歸（逐步迴歸 **&** 交叉驗證誤差）結果

圖 22 結果表中為羅吉斯迴歸，以逐步迴歸的方式選取模型并以交叉驗證誤差為選取模型標準所選取的變數及權重值。結果中，權重較大的變數為 User_total_buy、Cate_total_buy 和 Cate_total_user，其中 User_total_buy 和 Cate_total_buy 這兩個變數與預測結果有正相關關係，用戶購買總數上升以及商品種類被購買次數上升，都會提高商品種類被購買的機率，而 Cate_total_user 這個變數與目標變數呈負相關關係，越多客戶瀏覽該商品總類，反而會降低該商品總類被購買的機率。

3.4.3 羅吉斯迴歸（逐步迴歸 **&** 交叉驗證誤分類）結果

使用逐步迴歸的方式選取模型並以交叉驗證誤分類為選取模型的標準所選取的羅吉斯迴歸模型與 3.4.1 羅吉斯迴歸 (逐步迴歸) 所選取的模型相同，在此不做重複敘述，請參考 3.4.1。

3.4.4 類神經網路結果

由於類神經網路為多層次模型，結點過多，在此不做結果展示。

Parameter	**Estimate**	**Pr>ChiSq**	**BelongTo**	**Variables**
Intercept	-0.3253	<.0001		截距
LRatio	0.3008	<.0001	客戶行為表	占當天瀏覽所屬大分類比例
Cate_total_buy	3.5000	<.0001	商品類別描述表	被購買次數
Cate_total_user	-3.7244	<.0001	商品類別描述表	被瀏覽客戶數
Cate_total_view	-0.5205	<.0001	商品類別描述表	被瀏覽次數
CateDate_total_user	-0.4763	<.0001	商品類別熱度變化表	被瀏覽客戶數
CateDate_total_view	0.6413	<.0001	商品類別熱度變化表	被瀏覽次數
Cumday	0.1116	<.0001	客戶行為表	累計瀏覽天數
Cumview	0.5746	<.0001	客戶行為表	累計瀏覽次數
Searched	-0.0283	<.0001	客戶行為表	是否搜索
User_total_buy	4.5725	<.0001	客戶描述表	購買商品數
User_total_days	-2.0919	<.0001	客戶描述表	瀏覽天數
User_total_view	-1.3469	<.0001	客戶描述表	瀏覽次數
UserCate_total_view	-0.0338	0.0125	客戶商品類別喜好表	瀏覽次數
UserDate_total_cate	-0.1906	<.0001	客戶行為變化表	瀏覽商品類別個數
UserDate_total_view	-0.4975	<.0001	客戶行為變化表	瀏覽次數
View	0.2437	<.0001	客戶行為表	瀏覽次數

Figure 22: 羅吉斯迴歸（逐步迴歸 & 交叉驗證）結果表

3.4.5 決策樹結果

由於決策樹的結點過多，在此不做結果展示。

3.5 模型評估

Model	Train	Test
羅吉斯迴歸 (逐步)	0.83	0.83
羅吉斯迴歸 (逐步 & 交叉驗證誤差)	0.808	0.808
羅吉斯迴歸 (逐步 & 交叉驗證誤分類)	0.83	0.83
類神經網路	0.846	0.847
決策樹	0.839	0.838

Figure 23: 模型 ROC 結果表

在模型評估階段，我們選取 ROC 作為模型評估標準，因此，選取在訓練集和測試集中表現最好的類神經網絡作為最終模型。將模型運用至 1 月份的資料之上，預測在 1 月份的每一天用戶可能購買某個中分類下商品的機率，最終選取每個客戶預測購買機率最高的中分類作為商品推薦結果。

4 總結

本報告以東森購物網所提供的網站瀏覽記錄、客戶交易記錄、客戶輪廓資料和商品分類資料為基礎，使用 SAS EG & EM 深入分析客戶的瀏覽習慣、購買習慣和商品偏好，并以此為出發點，詳述建置客戶購買預測模型和商品推薦模型的過程，最終得出最可能購買的三千客戶名單和對應的商品推薦。

大數據浪潮來襲，如何分析數據、尋找數據中蘊藏的訊息成為市場的焦點，東森集團主辦這次的 BigData 校園爭霸戰，慷慨的把寶貴的網站資料交由我們分析，讓我們有了實際處理大數據的機會，感謝東森集團、SAS 和精誠資訊！

敬 上

2014年5月