Predicting Consumers' Purchase Decision by Clickstream Data: A Machine Learning Approach

經濟碩三 陳伯駒

Graduate Student
Department of Economics
National Taiwan University

2018/01/31

Outline

Motivation

- Main Question:
  根據網站上訪客的瀏覽行為，分類出什麼類型的訪客可能會購買？
- Why do we want to answer this question?
  - 對電子商務而言，訪客的回流率一般不高; 即使擁有很高的回流率，最終真正會使用網站平台購買的人也不多。若能夠從瀏覽行為中準確的預測訪客是否是潛在客戶，對廠商而言可以對此類訪客有進一步動作（廣告、產品推薦）、進而加速交易的進行。
  - 從模型預測中，歸納出訪客在網站上的瀏覽 pattern，從中分析可能的 incentives。

Main Page


超快速找酒
波爾多列級酒
布根地專區
精釀啤酒
日本清酒
ROBERT MONDAVI PRIVATE SELECTION
兩瓶$1,550
加贈原廠皮革雙入保冷袋乙只
我的清單
近期瀏覽
分類找酒
Calon-Segur
進階搜尋
轉為原文選單
找酒莊
找年份
熱門搜尋
頂級名酒 (844)
快速到貨 (1417)
超值預訂 (2)
複選
類型
紅酒 (4978)
白酒 (1386)
粉紅酒 (55)
氣泡酒 (470)
甜酒 (571)
加烈酒 (124)
烈酒 (10)
複選
產區
法國 (4265)
義大利 (667)
西班牙 (422)
鑽石、白銀
會員專屬優惠
Pol Roger
年份香檳三瓶組 NT $8,200
快速到貨
初夏推薦商品
VIP快訊
MEMBERS ONLY
飲・自然
2017仟元內值得喝的葡萄酒
飲食書
不要輕易翻開扉頁，因為飲食書既誘人又罪惡！知識帶來好奇，懂越多越想多方嘗試，往往停不下好奇心驅使地，一杯接一杯、一勺又一勺，在作者高潮迭起的敘述下，嘴邊的酒與餐居然更美味了！
More
N/A 2007
$221
N/A 2017
$458
N/A 2017
$300
{ 本週特價 }
飲酒過量有礙健康、禁止酒駕、未滿十八歲禁止飲酒
Responsible Drinking, Never Drink and Drive, No Underage Drinking

Main Page


西班牙 (422)
美國 (488)
澳洲 (315)
複選
展開更多
品種
梅洛 (2015)
卡本內蘇維濃 (1716)
卡本內弗朗 (1355)
黑皮諾 (990)
夏多內 (716)
複選
展開更多
單價
低於 $500 (184)
$500–1000 (1020)
$1001–2000 (1328)
$2001–3000 (1122)
$3001–4000 (659)
$4001–5000 (420)
$5000 以上 (1397)
輸入價格
例：0~5000
元之間
評價
Robert Parker (3099)
Jancis Robinson (2645)
Wine Spectator (1037)
Stephen Tanzer (1005)
Wine Enthusiast (656)
複選
展開更多
等級
預購專區
絕美粉紅香檳送原廠香檳杯2個
Laurent Perrier Cuvée Rosé Brut "Rose Jacket"
>> 看更多
Le Vieux Donjon
Châteauneuf du Pape
南隆河七大名莊 80歲祖父級老藤
$1,550
JOH. JOS. CHRISTOFFEL ERBEN
Mosel 歷史名園 Ürzig Würzgarten
$660
Pol Roger NV香檳
+
布裘粉紅酒
驚爆組合$2,000
我的清單
近期瀏覽
飲酒過量有礙健康、禁止酒駕、未滿十八歲禁止飲酒
Responsible Drinking, Never Drink and Drive, No Underage Drinking

Wine List


目前搜尋條件
紅酒
低於 $500
全部刪除
熱門搜尋
快速到貨 (40) 三點下單隔天到
複選
產區
法國 (14)
義大利 (6)
西班牙 (16)
美國 (9)
澳洲 (3)
複選
展開更多
品種
卡本內蘇維濃 (18)
梅洛 (13)
格那希 (8)
希哈 (8)
JS 91
750ml 2015
$473
750ml 2014
$499
搜尋結果 我們找到了80款
依 紅酒 / 低於$500
縮小搜尋範圍
我想換看 酒莊、年份 排列
我只想看有庫存的品項
< 上一頁 [1] 2 下一頁> 最末頁 >>
酒款 ▼ 年份 ▼ 評價 ▼ 容量 省最多 ▼ 單價 ▼ 預訂價 ▼
瓦特里酒莊，紅唇摩爾多瓦蜜思嘉甜紅酒
Château Vartely, Muscat Moldovenesc
2011 750ml 8折 360
摩爾多瓦
多瑪士嘉薩酒莊，磨坊經典紅酒
Mas de Daumas Gassac, Moulin de Gassac Classic Rouge
2011 750ml 9折 405
2013 750ml 9折 405
我的清單
近期瀏覽

Wine Detail


超快速找酒
波爾多列級酒
布根地專區
精釀啤酒
日本清酒
主題高分酒
分類找酒
Calon-Segur
進階搜尋
轉為原文選單
找酒莊
找年份
熱門搜尋
頂級名酒 (728)
快速到貨 (1182) 三點下單隔天到
複選
類型
紅酒 (3522)
白酒 (1098)
粉紅酒 (48)
氣泡酒 (427)
甜酒 (466)
加烈酒 (120)
烈酒 (9)
複選
產區
750ml
MUSCAT
瓦特里酒莊 紅唇摩爾多瓦蜜思嘉甜紅酒
Château Vartely,Muscat Moldovenesc 2011
#1061
定價 $ 450 / 酒窩價
$ 360
快速到貨
✔有庫存
產區：摩爾多瓦
生產商：Château Vartely
類型：紅葡萄酒 / 半甜型
品種：100 % Muscat
建議適飲期：2011-2014
搭配折扣活動：初夏搭餐不敗白中白氣泡酒加價購
《自然酒》第一本完整介紹「自然酒」的著作
我的清單
《
近期瀏覽
加價優惠商品：

Introduction

- Transaction process:

1. 將商品加入「我的清單」（購物車）
2. 在購物車中，點選「下一步」並「確認訂單」
3. 輸入配送地址及購買者資訊
4. 確認付款及配送
5. 送出訂單（order-id）

- 不需加入會員也可以購買; 但加入會員有好處（coupon, EDM, VIPepaper）
- 購買酒或加入會員需填入個人資料（性別、年齡、居住地、信箱）

Literature Review

- Verheijden (2012). "Predicting Purchase Behavior throughout the Clickstream."
- Walter (2013). "Web-shop Order Prediction Using Machine Learning."
- Ricardo (2015). "Using Clickstream Data to Analyze purchase intention."
- Kumar & Guruprasad (2015). "Clustering of Web Usage Data using Hybrid K-means and PACT Algorithms."
- Gabadinho, Ritschard and Mueller (2011). "Analyzing and Visualizing State Sequences in R with TraMineR."

Literature Review

- Solberg (1996). "A Large-Scale Evaluation of Features for Automatic Detection of Oil Spills in ERS SAR Images."
- Lunardon (2014). "ROSE: A Package for Binary Imbalanced Learning."
- Menardi & Torelli (2014). "Training and assessing classification rules with imbalanced data."
- Chawla et al (2002). "SMOTE: synthetic minority over-sampling technique."

Data

- Consumers' browsing log file on the website.
- Duration: Jan.06.2015 to Jan.01.2017
- Unique visitors: 667,068
- Visits(# of unique session): 1,348,855
- Pages viewed: 4,983,067

Data

Terminology

- Unique visitors: someone who visits site
  - IP + user agent + cookie + email account
- Visits(sessions): group of interactions on site within given time duration(30 min)
- Pages viewed: request to load HTML of site

Composition of Page Type

Page Type	Frequency	Percent	Cum
list_wine	1,798,236	36.09	36.09
wine_detail	1,217,020	42.42	60.51
main_wine	445,982	8.95	69.46
winery_intro	322,926	6.48	75.94
edm	236,998	4.76	80.70
spirit_detail	218,460	4.38	85.08
wine_cart1	129,780	2.60	87.68
wine_cart1_coupon	110,717	2.22	89.91
list_spirit	107,641	2.16	92.07
vipepaper	84,661	1.70	93.77
mem_login	39,264	0.79	94.55
wine_recommend	26,527	0.53	95.09
mem_member	24,832	0.50	95.58
main_spirit	22,488	0.45	96.04
...	...	...	...
spirit_variety	70	0.00	100.00
spirit_partial_pop	31	0.00	100.00
spirit_recommend	25	0.00	100.00
Total	4,983,067	100.00	

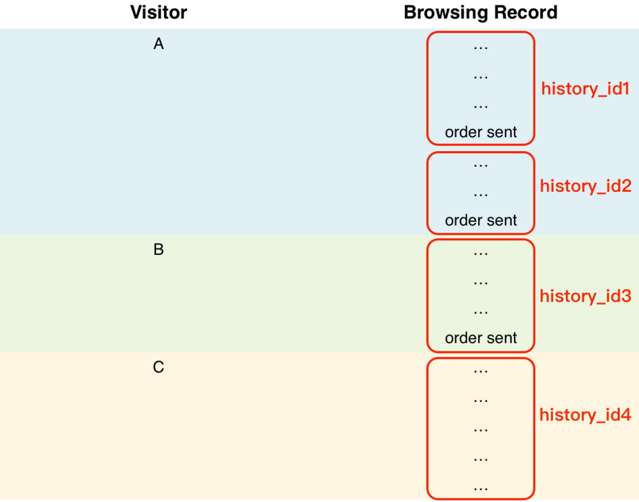
Browsing History - History_Id

Reason of using history_id as observation level instead of visitor_id

1. 假設使用 visitor_id 當做 observation level，對於那些購買很多次的訪客，無法有效分離出個別的瀏覽行為。
2. 資料結果使用上比較合乎預想。因為如果使用 visitor_id，所做出的結果是根據每個訪客的瀏覽紀錄所得出的結論。

- 直觀上：history_id 就是把 visitor_id 再切得更細！

Browsing History - History_Id



Data

Table: Order Conversion among History Record

	Register State no	Register State yes	Total
Order Sent			
no	664,068	2,105	666,173
yes	4,320	2,263	6,583
Total	668,388	4,368	672,756

Data

Table: Browsing Information

	Mean	S.D	Min	Max	Skewness	Kurtosis	Obs.
Duration (min)	10,290.48	1,462,931	0	8.31E+08	419.203	198,312.6	672,756
Page	5.48	57.94	1	29851	412.856	209,010.2	672,756

- Duration is the total time spent on website before an order sent.
- Page is the total viewed pages on website before an order sent.

Basic Conception

- Dealing the classification problem
- Train "Machine"(computer) to "Learn" how to classify correctly.
- Supervised learning(監督式學習) v.s Unsupervised learning(非監督式學習)

Why choosing ML?

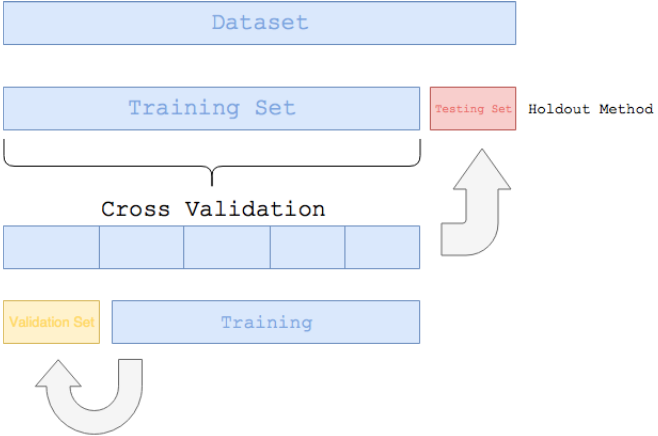
- High-Dimensional, nonparametric method
- Non-necessary requirement for assumption of relation between X & Y, especially for a non-polynomial model.
- High predictive power

Term

- **Training Set:** 用來學習的數據集，建立一種分類方式來訓練模型。
- **Testing Set:** 用來測試訓練模型之分類能力的數據集。
- **Validation Set:** 用來做 Cross–Validation(交叉驗證) 的數據集。

Term

Figure: Visualization of train/test split and cross-validation



k-folds Cross Validation

- It is used to **qualify** the model, instead of choosing the model.
- 初始的 train set 被分成 k 個子樣本，其中某一個子樣本會被當作驗證模型的數據集，剩下的 $k - 1$ 個樣本數據集會被拿來訓練。
- 上述動作重複 k 次以後，最後平均 k 次的 Accuracy rate，來衡量模型最終的精準率。
- k-folds CV 可以有效地避免 overfitting 與 under-fitting 的學習狀態發生，最終得到的結果直觀上也較有說服性。

Training Set & Testing Set

- Random sample 80% of whole data as the training set.
- Training Set: 538,205. Testing Set: 134,551. Total: 672,756.
- 10-folds cross validation.

Features: $\mathbf{X}$, Classification: Y

- Y: whether the history-id send the order(Binary)
- $\mathbf{X}$: possible features that may affect the classification of Y, here we use the vector: $(DP, VP, cluster, reg - state, month)$
  - dur: Duration spent on website before purchase session
  - pages: Pages viewed on website before purchase session
  - cluster: cluster result for the page-type sequence before purchase session
  - reg-state: whether this visitor is registered until the purchase session
  - month: the month which the order is sent.

Decision Tree

- 演算法 (CART) 中，一開始 Training Set 中所有的樣本都會在第一個根節點 (node)、計算各屬性所提供的資訊量後，挑選最好的屬性當作分節點，並反覆地將樣本分開來。
- 若以下條件發生，則會停止分裂:
  1. 所有的屬性皆被使用完畢。
  2. 選取某個屬性後，發生某個分支完全沒有樣本被歸在裡面的情況。

Decision Tree

parameters tuned in code

- **Complexity parameter(複雜性參數):** 主要功用是修剪樹，決定樹某個節點要不要分裂的閾值：直觀上如果 cp(閾值) 越大，樹越不容易分裂、分枝越少; 反之 cp 越小，樹越容易分裂、分之越多。
- Rule: if any split does not increase the overall prediction rate of model by at least **"cp"**.
- 當複雜度過高時，可能會發生 over–fitting 的問題，若要建立一棵能精準分類、又不會過度擬合的樹需要調整一個最適的 cp。(透過 cross-validation 來修正 cp)

Random Forest: 一群隨機產生出來的分類器 (樹)

- **Random:**
  1. 假設 Train Set 的大小為 N，對於森林中的每棵樹而言，採 bootstrap 的方式從訓練集抽取 N 個樣本作分類 (重複且放回地抽出 N 個，因此每棵樹的訓練集可能不一樣)。
  2. 假設每個樣本的特徵屬性維度為 M，則指定一個 $m << M$、隨機地從 M 個特徵中選取 m 個屬性子集。每棵樹進行分裂時，從這 m 個屬性裡挑選最好的進行分裂。
- **Forest:** 由 T 棵決策樹所構成，而每棵決策樹都是一種分類器。輸入一個樣本後，T 個分類器會有 T 個分類結果，這些分類器進行簡單多數決的投票選出最終的分類結果。

Random Forest

parameters tuned in code

- **mtry:** 對於每棵樹、個別的分裂點 node，從所有變數中隨機挑選 m 個來選擇分裂 (因此選擇過的 feature 在後續分裂點也可能被選到)，停止的準則為隨機所選取的屬性中都無法將樣本繼續分類的情況。隨機森林的分類效果與另外下列兩個因素有關：
  - 森林中任兩棵樹的相關係越大、錯誤率越大。
  - 森林中每棵樹個別的分類能力越強、整體森林錯誤率越小。

  減少挑選的 m(feature set 不盡相同、隨機性強)、樹之間的相關性與個別樹分類能力都會降低; 增加選取的 m(feature set 都一樣、隨機性弱)，兩者也會同時提升，因此 m 的選取是一種 trade–off，需要選一個使錯誤率降到最低的 m。

Random Forest

parameters tuned in code

- **ntree:** 欲生成之決策樹木個數 T(韓信點兵、多多益善)。
  樹的數目越多、對於預測率會越準（產生越多分類器進行投票），但樹的數目若太多將增加無意義的運算時間，且在樹木生長到一定的數目時，袋外錯誤率會趨於穩定，因此只須選擇一個合適的生長數目：亦即使袋外錯誤率趨於穩定的成長數目即可 (nTree 夠大即可)。

Sequence Clustering

- R packages: WeightedCluster & TraMineR
- Observation Level: history-id, identification of browsing history.

Weighted Cluster & TraMineR

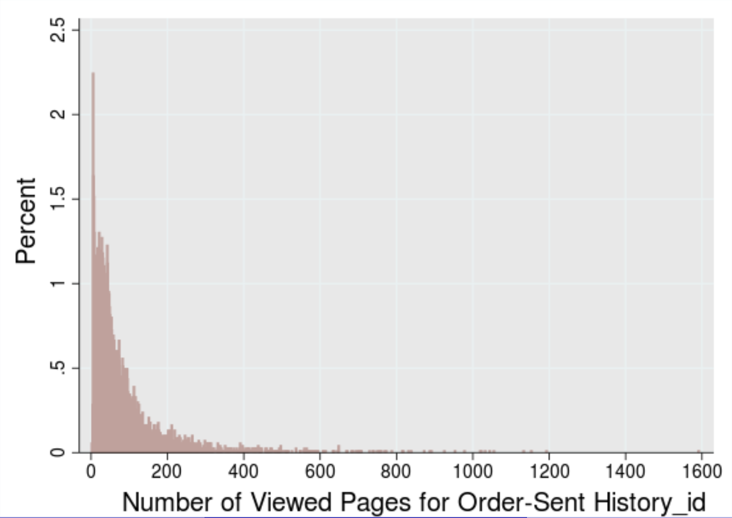
- 先將相類似的 sequence，aggregate 起來後再做分類。(避免運算過於龐大)
- 使用 LCP 運算，造出每個數列之間的distance matrix。
- LCP(Longest Common Prefix) distance: one of the measure similarity/distance between distances, which is based on the length of the longest common prefix. 直覺上，兩個數列重複部分的長度越長、之間的距離就越短。
- $\sqrt{2^{31}} \approx 46,340$

Distribution of Pages Viewed for order-sent history

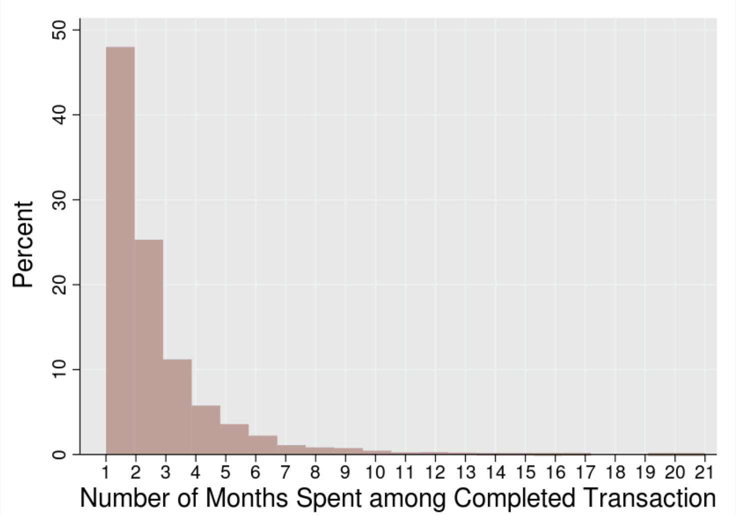
Table: Summary of Viewed Pages for Order-Sent History_id

Variable	Mean	Std. Dev.	min	Max	N
Number of Viewed Pages	83.59	108.01	2	1,592	6,583

Distribution of Pages Viewed for order-sent history



Distribution of transaction time



Rule

1. drop if mon>2(因大部分有購買的 history 紀錄都在本月及上個月之間完成，我們也假設說一個人會不會購買、在他前幾個月前的紀錄比較會影響。)
   - order==0: 抓一開始進入網站後 2 個月內的瀏覽紀錄。
   - order==1: 抓完成送出訂單當下那個月及前一個月份 (總共 2 個月) 的瀏覽紀錄。
2. 對於 order==0: drop if obs<2(因 order==1 最少看 1 頁) or 1000 以上的頁面。
3. keep if page-type=winelist, wine-detail, mainwine 。
4. observation level: each purchase history

結論：沒有被納入歸類的 history-id 為：只有看 1 頁紀錄的 + 頁面組成不包含「Main-Wine、Wine-Detail、Wine-List」，直覺上來想這些瀏覽紀錄比較不是我們所關心的，因其代表進來只看 1 頁就離開，或是整個瀏覽紀錄都跟產品酒的頁面無關。

Total: 229,574. Unique: 41,433 $(< 46,340 \sim \sqrt{2^{31}})$

Imbalanced Data Modification

- Imbalanced label data: 有購買紀錄的 history-d 佔全部的比例過低。
- Hazard of unbalanced data in machine learning:
  The machine learning model could make the classification that predict the whole history-id are un-order-sent and still get a high accuracy rate.

Re-sampling Method

- Over-Sampling：Randomly duplicating the samples from the minority so as to match the number of sample in each class.
  缺點：If the features are few, overfitting problem occurs.
- Under-Sampling：Randomly selecting the samples from majority so as to match the number of sample in each class
  缺點：Potentially losing relevant information from the left-out samples, which may lead to under-fitting.

Synthetic Data Generation

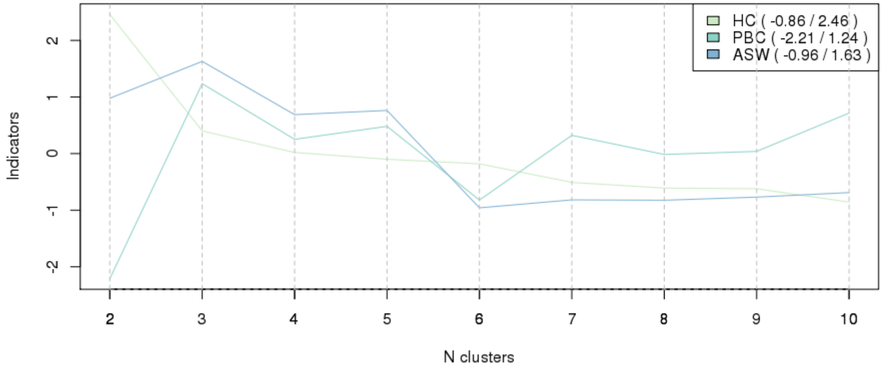
- ROSE： Using smooth bootstrap to create the articial samples from the feature space neighborhood around the minority class.
- SMOTE： Drawing artifcial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space.
- 直觀上，兩者的概念都是透過 minority 的 feature 間附近的距離、人工合成新的數據（創造更多類似的 minority）。

Weighted Cluster & TraMineR

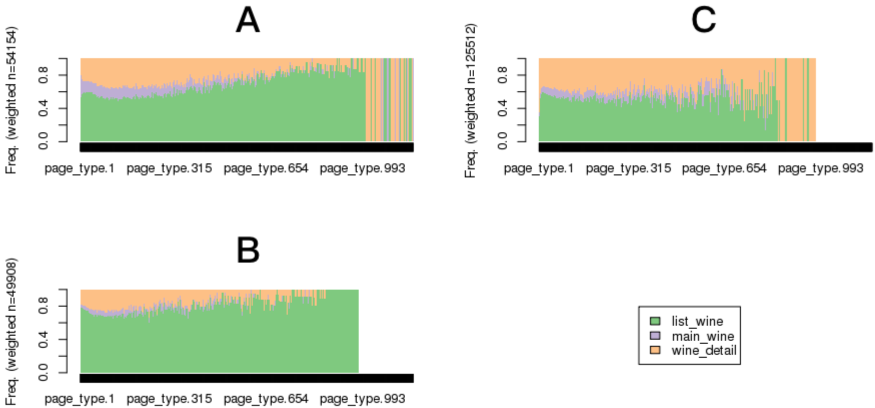
Clustering Quality Measure

- Point Biserial Correlation(PBC): Capacity of the clustering to reproduce the original distance matrix. **(argMAX)**
- Average Silhouette Width(ASW): Coherence of the assignments. A high coherence indicates high between groups distances and high intra group homogeneity. **(argMAX)**
- Hubert's C(HC): Gap between the current quality of clustering and the best possible quality for this distance matrix and number of groups. **(argMin)**

Weighted Cluster & TraMineR



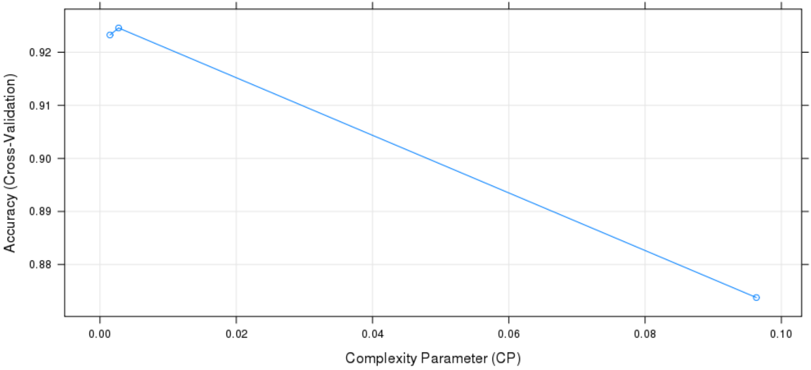
Weighted Cluster & TraMineR



Decision Tree

Hyperparameter Tuning

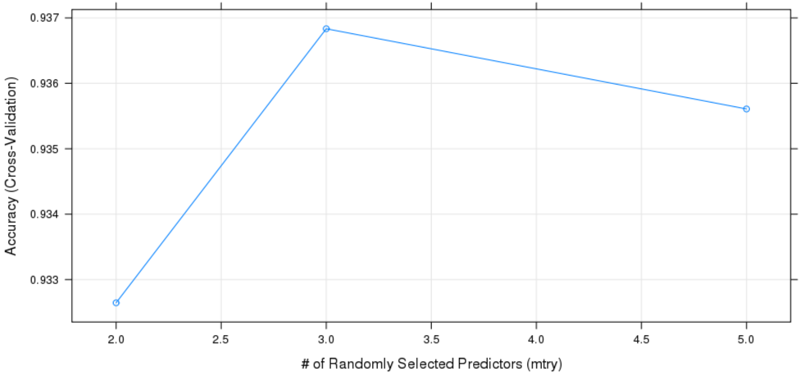
optimal $cp = 0.002736$



Random Forest

Hyperparameter Tuning

optimal $m_{try} = 3$



Model Evaluation

Confusion Matirx

	Predicted: No	Predicted: Yes
Actual: No	True Negative(TN)	False Positive(FP)
Actual: Yes	False Negative(FN)	True Positive(TP)

- False Positive(FP): " 錯誤地" 把樣本歸類成" 陽性"(所以代表樣本其實為陰性)
- False Negative(FN): " 錯誤地" 把樣本歸類成" 陰性"(所以代表樣本其實為陽性)

Model Evaluation

Terminology from Confusion Matrix

- "Positive" class(陽性): 研究問題主觀上所關心的類別，在本研究中陽性類即為" 有送出訂單的瀏覽紀錄"; 相對地，沒有送出訂單的瀏覽紀錄即為陰性類。
- Precision(準確率)= $\frac{TP}{TP+FP}$: 此模型預測樣本為陽性的準確率有多高?
- Sensitivity(敏感度)= $\frac{TP}{TP+FN}$: 此模型偵測到陽性類的機會有多大? 越高就代表越敏感。
- Specifity(特異度)= $\frac{TN}{FP+TN}$: 此模型偵測到陰性類的機會有多大? 就是對陰性類的敏感度。

Model Evaluation

Terminology from Confusion Matrix

- False Positive Rate= 1−Specifity: " 錯誤地" 把樣本歸類成" 陽性" 的比率有多高?
- False Negative Rate= 1−Sensitivity: " 錯誤地" 把樣本歸類成" 陰性" 的比率有多高?
- F1 score: Precision 與 Sensitivity 的調和平均數 (就是兩者平均的概念)。
- Overall Accuracy(整體準確率)= $\frac{TP+TN}{N}$: 此模型對於陽性與陰性的整體預測準確率有多高?

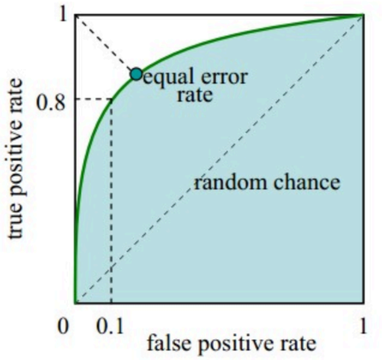
Model Evaluation

Receiver Operating Characteristic(ROC) curve and Area Under Curve(AUC)

- ROC curve 反映 Sensitivity(敏感度) 與 False Positive Rate(偽陽性率) 兩者關係的綜合指標。直觀上 ROC curve 曲線上每一個點代表著某一種 threshold(閾值) 下模型的感受性。
- 以一個使用 logistic regression 做二元分類的模型為例，模型最終會預測每個樣本為陽性的機率。那麼如果設定一個閾值、假設是 0.8，把為陽性機率大於 0.8 的才歸為陽性類、小於 0.8 的則歸為陰性類，則在此閾值下即可算出一組 (Sensitivity, True Positive Rate)。
- x-axis: False Positive Rate
- y-axis: Sensitivity(True Positive Rate)

Model Evaluation

Receiver Operating Characteristic(ROC) curve and Area Under Curve(AUC)



- 理想目標：找到某一個閾值、使得 TPR=1, FPR=0，因此 ROC curve 越往"「 " 移動越好。

Model Evaluation

Thresholds for ROC curve

- 以 logistic regression 做二元分類為例，ROC curve 的可能閾值即為模型最終對樣本預測機率的組合 (陽性機率, 陰性機率)。假設總共有 5 個樣本、每個樣本最後被估計出來的 (陽性機率, 陰性機率) 集合為 $\{(0.1, 0.9), (0.2, 0.8), (0.25, 0.75), (0.6, 0.4), (0.8, 0.2)\}$，則可能的閾值即為 $0, 0.1, 0.2, 0.25, 0.6, 0.8, 1$。
- 每個閾值下都會有一組 (TPR, FPR)。閾值最大時， $TPR = FPR = 0$，對應於圖中左下角原點 (0,0) 的位置; 閾值最小時， $TPR = FPR = 1$，對應於右上角的點 (1,1)。
- 45° 線即為 random select 的模型，所以每一種閾值下都是 TPR=FPR。

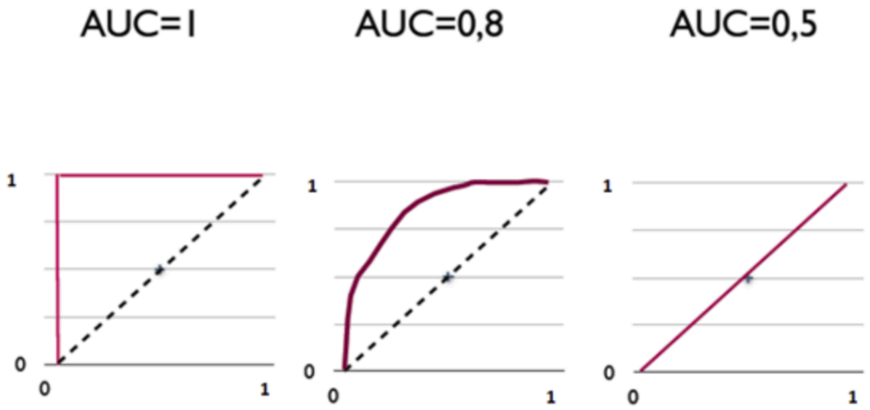
Model Evaluation

AUC: Area Under roc Curve

- $0.5 \leq AUC \leq 1$，可以把 AUC 想像成偏離 45° 線的程度，數值越高代表越偏離、越能夠最越好的分類。
- $AUC = 1$ 代表是完美的分類器，至少存在一種閾值使得樣本能完全分類正確。
- $0.5 < AUC < 1$ 代表模型優於隨機猜測。
- $AUC = 0.5$ 代表模型預測結果跟隨機猜測一樣。
- $AUC < 0.5$ 代表模型比隨機預測能力還要差。

Model Evaluation

AUC: Area Under roc Curve



Decision Tree

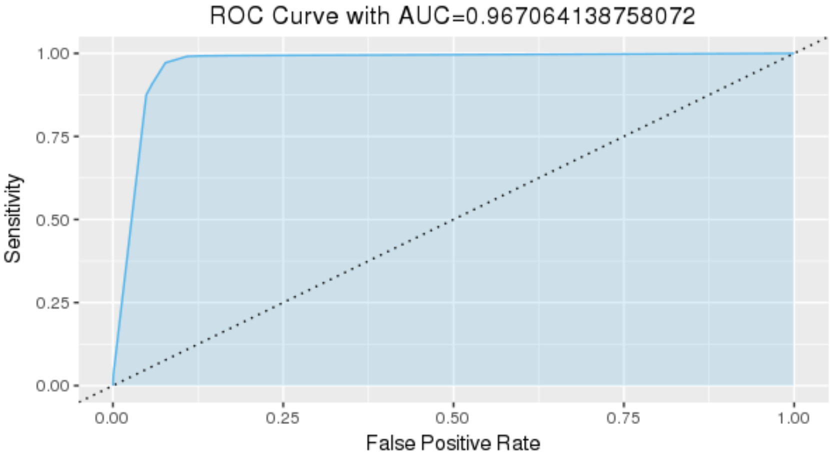
Confusion Matrix & Accuracy

Overall Accuracy: 92.35%, Sensitivity: 97.17%
Precision: 11.31%, F1 score: 20.25%

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	122,945	10,260
Actual: Yes	38	1,308

Decision Tree

ROC Curve & AUC



Random Forest

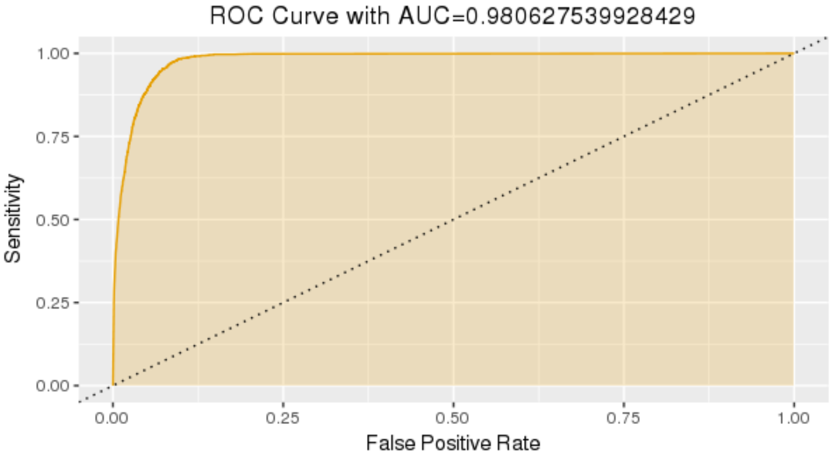
Confusion Matrix & Accuracy

Overall Accuracy: 93.69%, Sensitivity: 92.86%
Precision: 12.96%, F1 score: 22.75%

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	124,813	8,392
Actual: Yes	96	1,250

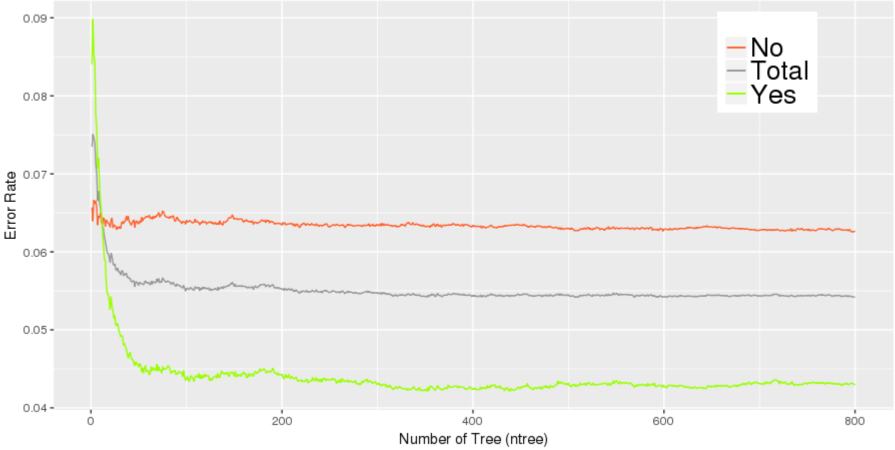
Decision Tree

ROC Curve & AUC

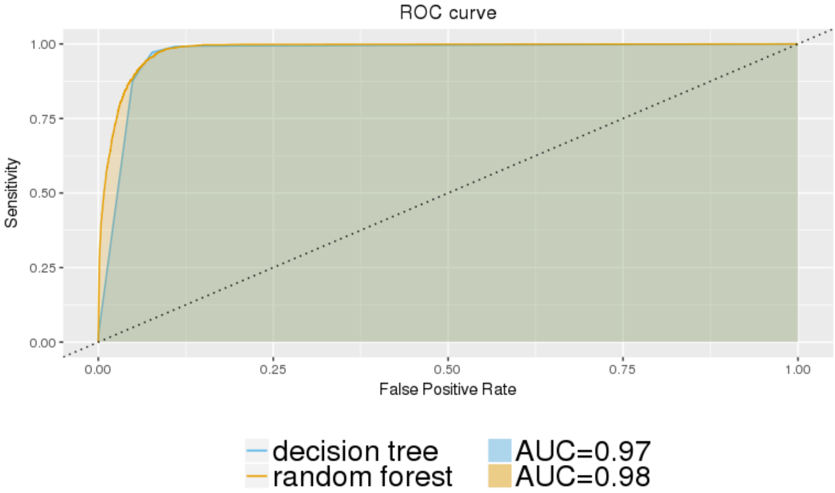


Random Forest

Learning Curve

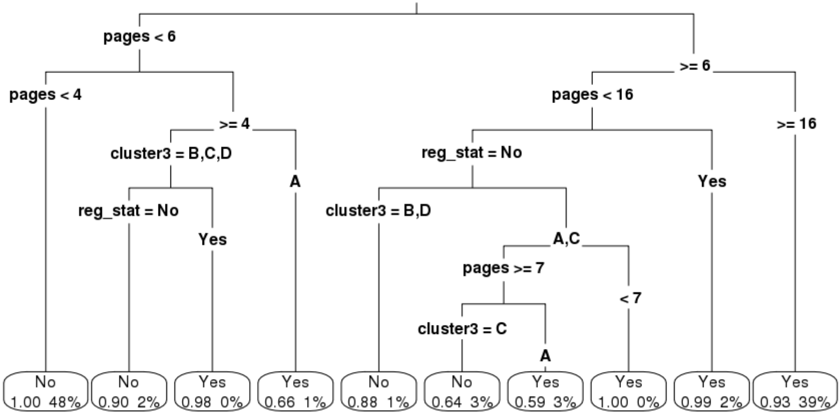


Synthesized ROC Curve

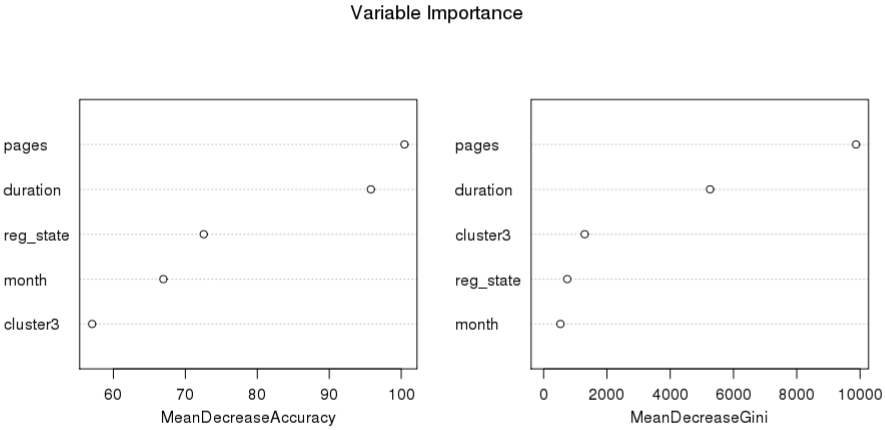


Decision Tree

Variable Importance



Random Forest



Conclusion

- Random Forest has higher overall predictive accuracy, precision and F1 score; Decision Tree has higher sensitivity.
- Both model has similar AUC of ROC curve, the value in Random Forest is slightly higher than the Decision Tree ones.
- Pages and Duration mainly contain the information of classification.

Further Work

- Extracting more features providing products' information.
- Feasible to store and compute larger matrix in R
- Applying other alternative classification models to verify the optimal prediction model.
- The model we implement may not be able to interpret the marginal effect; however, it's more like selecting a model which can predict the Y through X most precisely.

Decision Tree with Over-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	122,945	10,260
Actual: Yes	38	1,308

Decision Tree with Under-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	122,945	10,260
Actual: Yes	38	1,308

Decision Tree with ROSE-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	126,804	6,401
Actual: Yes	215	1,131

Random Forest with Over-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	132,571	634
Actual: Yes	864	482

Random Froest with Under-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	122,223	10,982
Actual: Yes	26	1,320

Random Forest with ROSE-Sampling Method

N=134,551	**Predicted: No**	**Predicted: Yes**
Actual: No	121,329	11,876
Actual: Yes	56	1,290