

資料科學期末專題
第十二組

Santander Product Recommendation
資料分析及消費者未來行為預測

組員

魏廷宇	r11942104
陳泳源	r11942163
李鎮宇	r11921078
陳奕舟	r10942061
黃湛元	r11942180

民國112年1月5日

摘要

本篇報告將說明分析 Santander 銀行提供的客戶購買各式貸款服務之資料結果，並透過機器學習模型預測客戶可能購買之產品。實作主要分為三個階段，分別為資料前處理，資料特徵分析，以及模型結果預測。首先我們會將資料進行清洗以及轉換，接著會採用 ANOVA 進行特徵選擇，並根據選擇後的特徵進行數據分析和交叉比對，比較數值化特徵選擇的結果。最後我們會利用 Catboost Model 進行客戶購買商品的預測，並根據參數以及使用的訓練資料上的差異進行比較。

1. 導論

為了支持客戶財務決策的需求，Santander銀行通過分析客戶行為推薦客戶各種貸款服務。然而第一階段的系統設計的並不完整，使某些用戶會一次收到大量的推薦，某些用戶卻鮮少收到推薦，造成不均勻的客戶使用體驗。因此本次實驗的目的是根據Santander Product Recommendation競賽的資料預測下一個月客戶當月會購買什麼商品，並期望預測成績能越高越好。此外，本實驗同時希望清楚分析出購買某個商品的原因是什麼。

2. 材料與方法

2.1 實驗資料集

本次實驗的資料是在kaggle上的Santander Product Recommendation競賽的資料。此資料一共包含一年半的用戶資料。此資料記錄從2015-01-28開始，且每個月記錄一次每個用戶擁有的金融商品。此測驗特別註記其提供的資料沒有包含任何真實的Santander銀行西班牙用戶的資料，因此這些資料不能用來代表任何西班牙用戶資料。

2.2 資料前處理

觀察資料中每一個特徵的分布情況做出不同的調整，主要是資料型態的轉換以及缺項補值，如表1所示，最後把非數字的label利用sklearn裡的LabelEncoder轉換成數字，其餘的保持不動。

Features	Preprocessing
'ind_emptado'	其中N占多數，因此nan的部分我都替換成N
'ult_fec_cli_1t'	這部分的重點是他是否為primary customer，因此nan我都先換成0，然後非0的部分再換成1（代表是primary customer）

'conyuemp'	這個column有太多的nan，不好做填補，因此我drop掉整個column
'renta'	把nan換成占比最多的數字
'cod_prov', 'segmento', 'canal_entrada', 'sexo', 'ind_nom_pens_ult1', 'ind_nomia_ult1'	因為nan的欄位占比很少，因此我把有nan的整個row從資料中抽掉
'indrel_lmes'	把nan換成1，因為1出現次數最多，然後把P換成0
'fecha_alta'	改成timestamp

表1、資料前處理

有了完整的資料後，我們觀察到商品會以累加的方式記錄，好幾個月前購入的商品，會在之後每一個月的資料中出現，但我們希望能找到每一位顧客在每個月當中會新購入的品項，藉此了解是哪些特徵會影響購入新商品的行為。因為資料是從2015-01-28開始，因此先建立一群假資料，時間是2014-12-28，全部品項皆為0個，接著再把每個月往前一個月減，就能得到所有新購入商品的紀錄。

2.3 特徵選擇

最一開始，我們希望先分析各個feature是否帶有足夠的資訊量，也就是彼此間的variance是否足夠大。因為若variance過小甚至為零，都會導致分類器訓練容易出問題，或是不容易收斂。接著，我們希望透過Univariate feature selection，選出其中幾個分數最高的feature進行接下來的訓練。本次我們採用的方法是ANOVA。ANOVA會考慮feature間的variance以及feature內的variance。且透過F-test判斷兩個分布間的平均值是否有顯著的差異。最後取F value最大的幾個feature。

根據[1]的結論，在multiclass的情況下，每個feature在不同class的得到的分數，最後取平均或是最大值，都能有效地取出富含最多資訊的feature，使得訓練結果較其他feature selection方法優秀。因此針對每一個label我都有進行一次ANOVA test，並將每次得到的F value紀錄，最後取平均值得到每個feature的score。最後將score取Log並scale到0~1之間，繪出下圖1。結果整理為，表1。

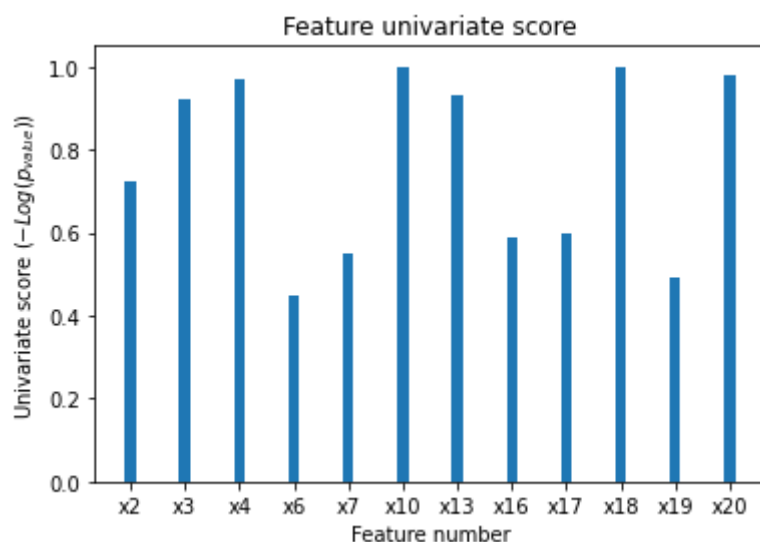


圖1 各個feature分數圖

表1 Feature Selection結果

Column Number	Feature Name	Description
2	sexo	Customer's sex
3	age	Age
4	fecha_alta	The date in which the customer became as the first holder of a contract in the bank
10	tiprel_lmes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer), R (Potential)
13	canal_entrada	channel used by the customer to join
18	ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
20	segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated

3. 結果與討論

3.1 數據分析與推論

本節希望對原始資料的特徵提出一些解釋，並嘗試與數值化方法選出的特徵結果做交叉比對。

根據整理完的資料，我們計算了每個月每個產品的銷售量，如圖2。也計算了各個商品在每個月份被購買的比例，如圖3。因此可清楚的觀察到各個產品的買氣，以及購買趨勢。依照趨勢以及購買情況我們可以分成幾個類別：

1. 穩定低落型：ind_ahor_fin_ult1 (Saving Account)、ind_aval_fin_ult1 (Guarantees)、ind_cder_fin_ult1 (Derivada Account)、ind_viv_fin_ult1 (Home Account)和ind_ctju_fin_ult1 (Junior Account) 這幾項產品的銷售額非常的差，且並沒有隨著時間成長。
2. 瞬間成長型：ind_cco_fin_ult1 (Current Accounts) 產品，在2015年的六、八、九、十、十一以及十二月銷售量特別高，然而其他月份並沒有明顯增長。ind_valo_fin_ult1 (Securities)在十月與一月銷售量特別高。ind_nomina_ult1 (Payroll)的比例僅在十月特別突出。ind_reca_fin_ult1 (Taxes)在六月使用量特別高，有可能是報稅的季節。
3. 穩定型：ind_cno_fin_ult1 (Payroll Account)則沒有特別哪個月份的銷售量最高，與現實情況相比對似乎也非常合理，因為薪轉戶的需求理論上應該每個月份的差不多。ind_tjcr_fin_ult1 (Credit Card)和ind_recibo_ult1 (Direct Debit)、ind_hip_fin_ult1、ind_plan_fin_ult1和ind_pres_fin_ult1 (Mortgage、Pensions和Loans)的需求同理。
4. 穩定成長型：ind_ctma_fin_ult1、ind_ctop_fin_ult1和ind_ctpp_fin_ult1 (Más particular Account、particular Account和particular Plus Account) 三者的需求有逐年成長的趨勢。ind_ecue_fin_ult1 (e-account)也有來越多人使用的趨勢。
5. 穩定衰退型：ind_deco_fin_ult1、ind_deme_fin_ult1和ind_dela_fin_ult1 (Short-term deposits、Medium-term deposits和Long-term deposits)的買氣有逐月下降的趨勢，顯然儲蓄似乎越來越不流行。ind_fond_fin_ult1 (funds)也是。

雖然都有嘗試對於結果提出解釋，然而由於資料集無法直接代表西班牙民眾，因此很難與當地民情直接比較，故以上都只是根據資料揣測。以上的觀察都有助於我們設計模型時，調整hyperparameters，例如觀察結果時，就應該挑選預測穩定成長型與穩定型為主的模型。

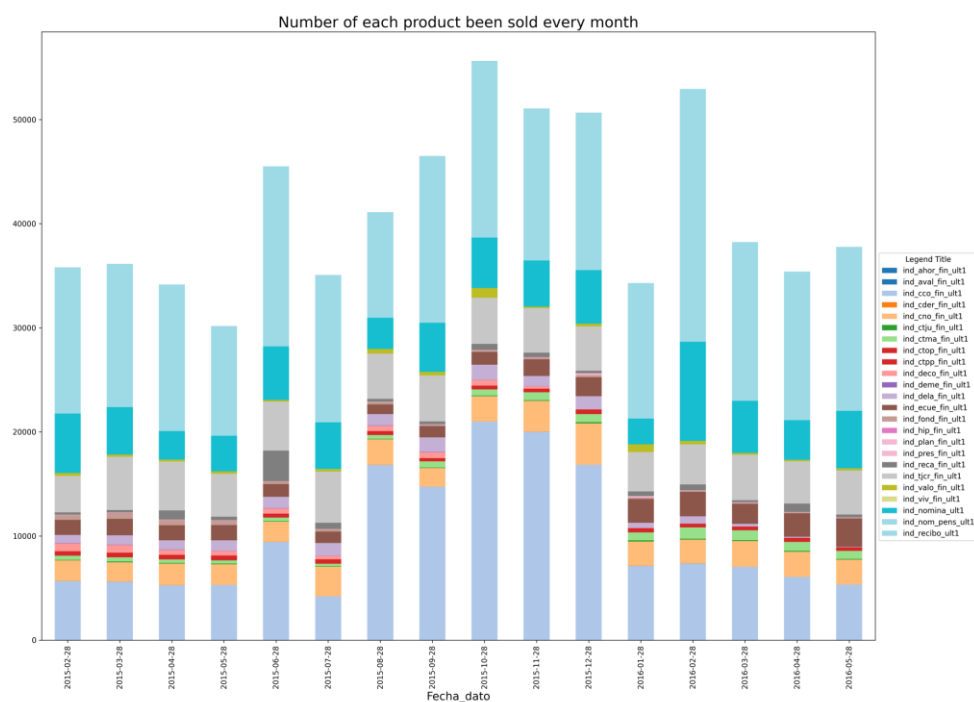


圖2 每個產品每個月的銷售量。

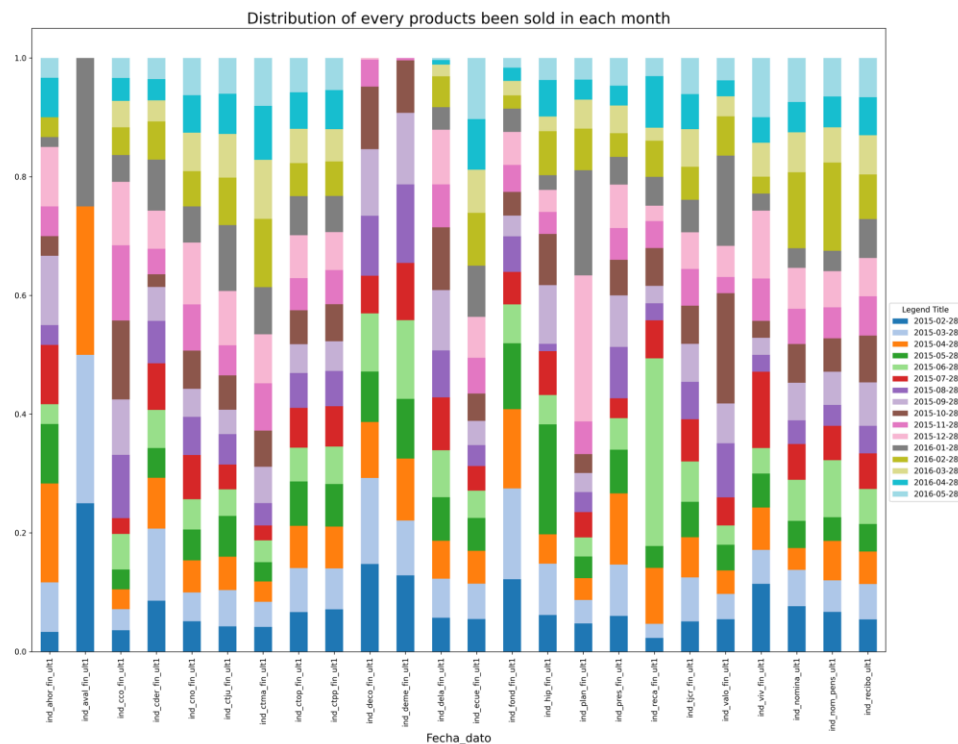


圖3 每個商品不同月份購買的比例

根據整理完的資料，我們可以獲得購買新商品的月份為何。經過處理後，可以得到兩次購買新商品的時間間隔長度。我們依照此資訊繪製了圖4。由圖三可以看出大部分的消費者連續購買商品的時間間隔主要分布在一個月到四個月之間。平均大約是落在三個月購買一次左右。由這項資訊可以推估消費者大致上的購買習性，在正確的時段推薦消費者適當的商品資訊，達到事半功倍的效果。

我們也同步整理了，性別比例圓餅圖，如圖5。每個商品不同性別購買的比例，如圖6。我們觀察可以發現男性用戶數量比例大於女性用戶。根據此發現，與圖6比對，可推測男性在我的研究中購買力是大於女性的，因大部分商品性別購買比都大於56:44。此外觀察每個商品不同性別購買的比例，可發現男性相較於女性特別喜歡ind_ahor_fin_ult1 (Saving Account)和ind_cder_fin_ult1 (Derivada Account)；女性則無特別的喜好。因此我們認為Gender是個重要的特徵。

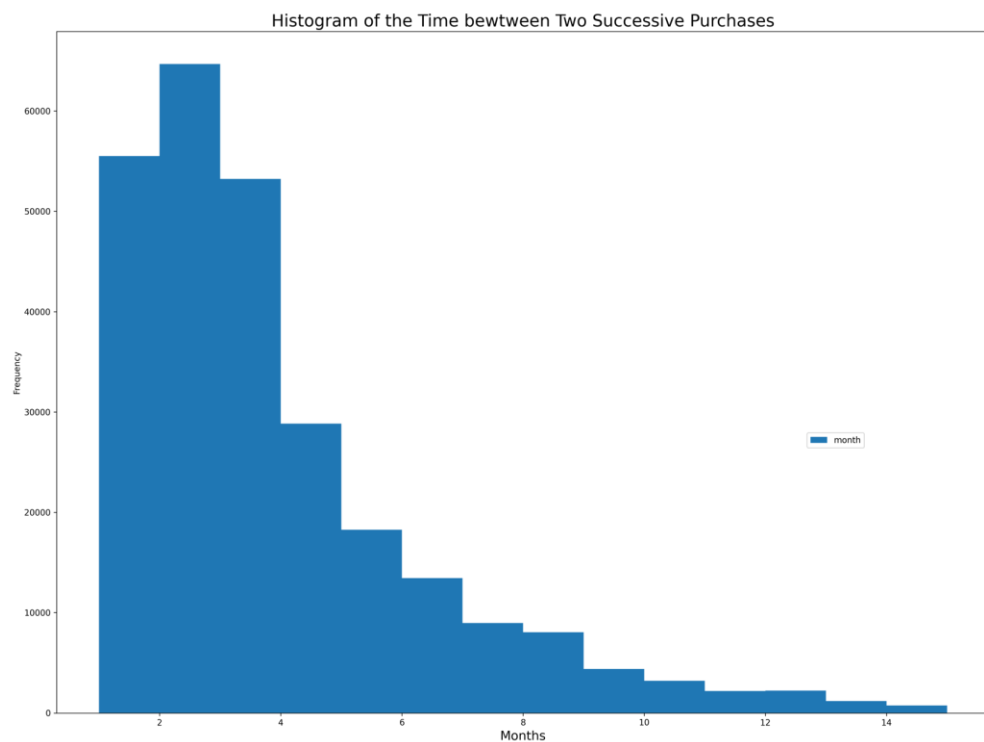


圖4 購買時間間隔直方圖

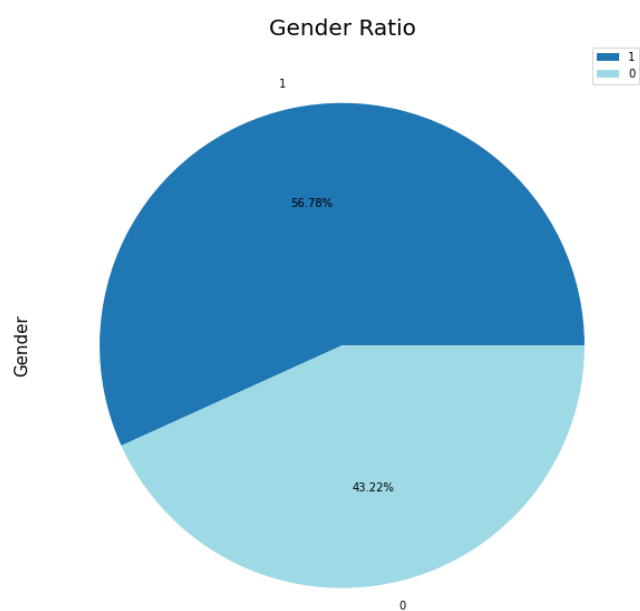


圖5 性別比例

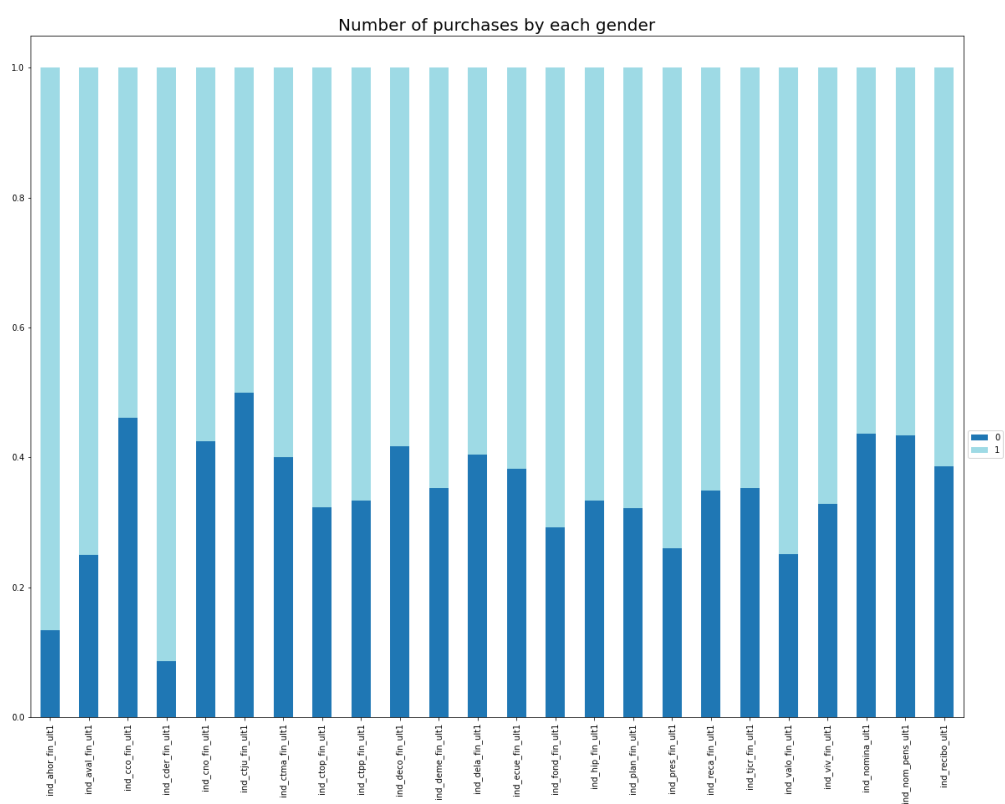


圖6 每商品不同性別購買比例

針對age這項特徵，我們有別深入討論。首先我們想知道每樣商品的購買年齡分佈，因此我們繪製了圖7。由圖六我們可以輕易觀察出每樣商品的購買平均年齡，以及其主要客群的分佈年齡。可發現ind_cco_fin_ult1 (Current Accounts)、ind_cno_fin_ult1 (Payroll Account)、ind_ctju_fin_ult1 (Junior Account)、ind_nomina_ult1 (Payroll)、ind_nom_pens_ult1 (Pensions)以及ind_recibo_ult1 (Direct Debit)的平均購買年齡，與其他商品的年齡有明顯差距。因此我們認為年齡是一個非常重要的特徵。

此外我們針對使用ind_ctju_fin_ult1 (Junior Account)與不使用此服務的族群年齡做了分佈圖，如圖8。可發現Junior Account的使用者年齡分佈，與其他服務使用者的年齡基本上並沒有太多重疊，因此可以發現使用年齡可以準確的將Junior Account的使用者預測出來。這也符合現實世界可以觀察出來的情況，並無太大爭議。

我們針對tiprel_lmes (Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer), R (Potential))做了分析，如圖9。發現此特徵在不同商品的分佈相對的有非常明顯的差距。可以清楚的觀察到ind_ctju_fin_ult1 (Junior Account)的Inactive消費者明顯的比其他商品的多，因此這個特徵同樣很適合用來分類。

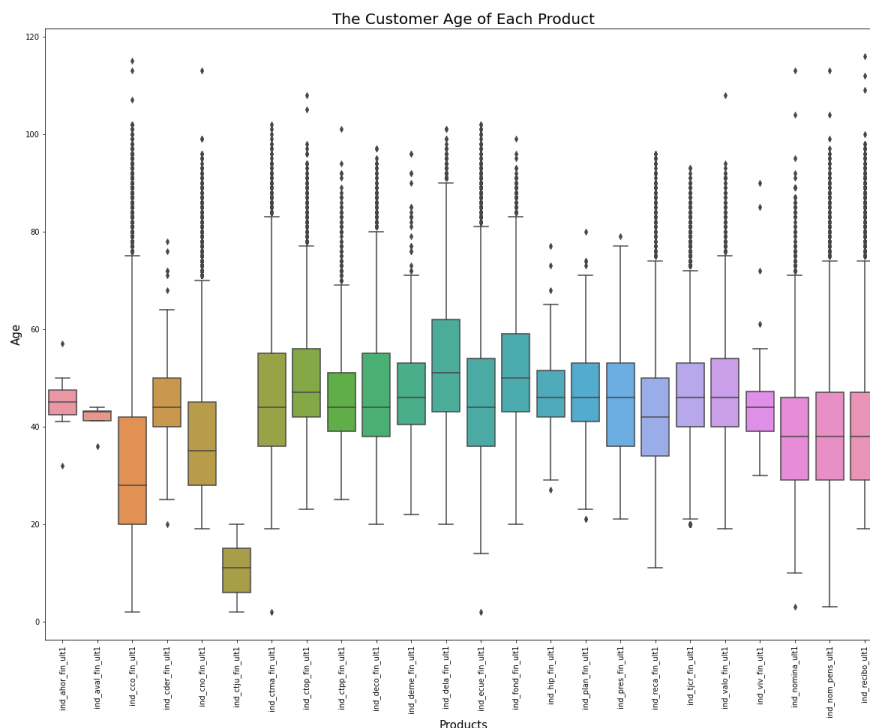


圖7 每個商品的購買年齡分佈

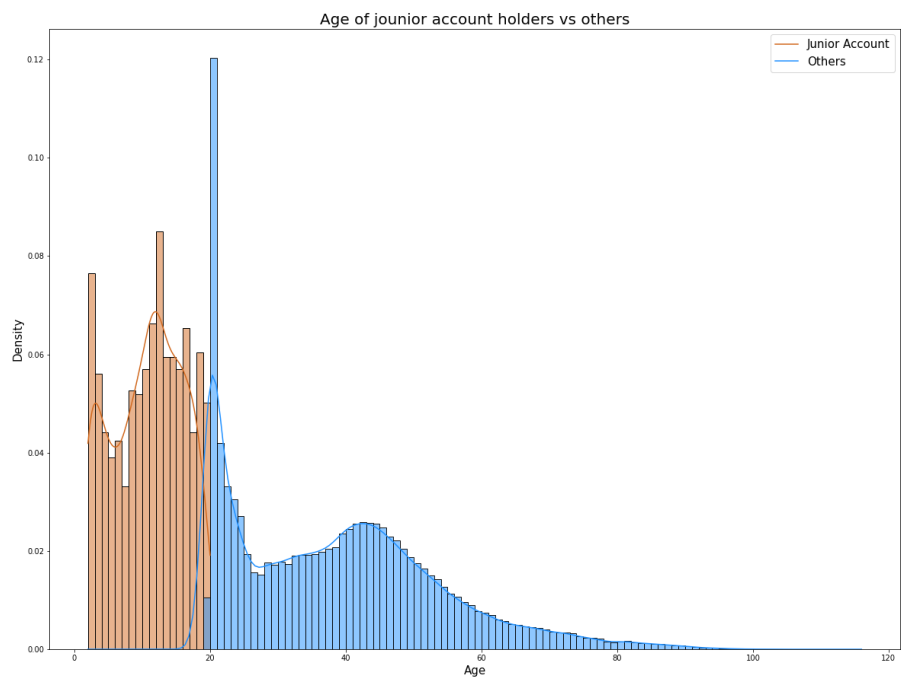


圖8 年齡與購買 junior account 與否的

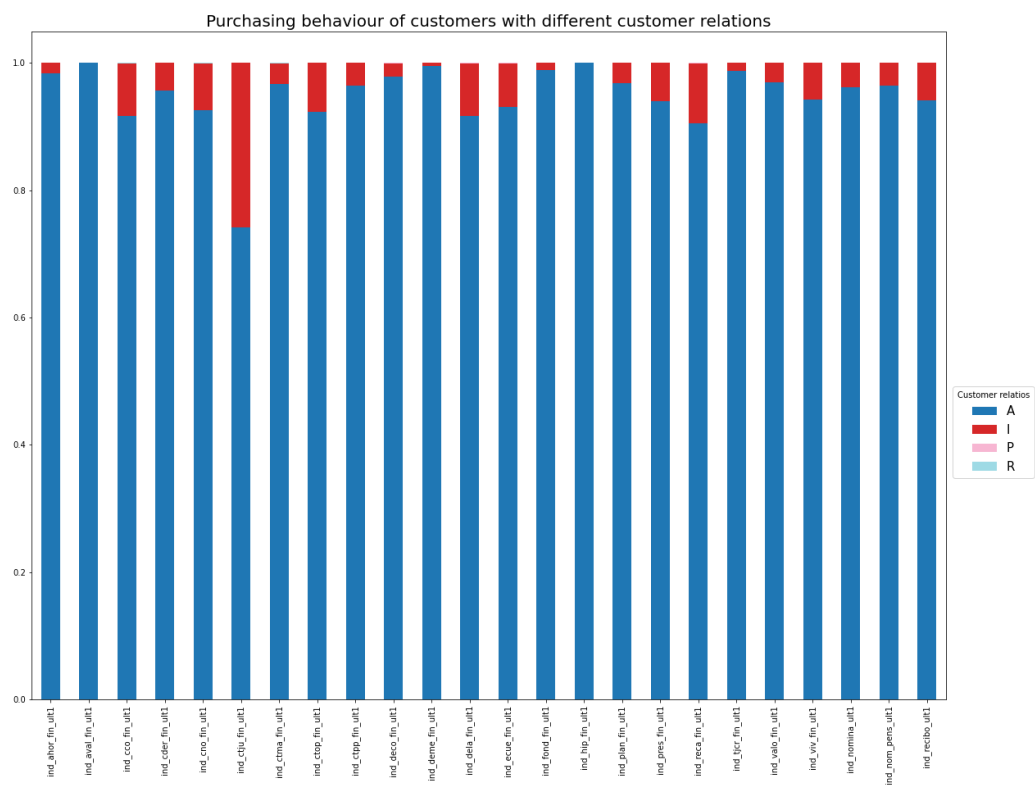


圖9 消費者購買行為與消費與銀行之關係比例圖

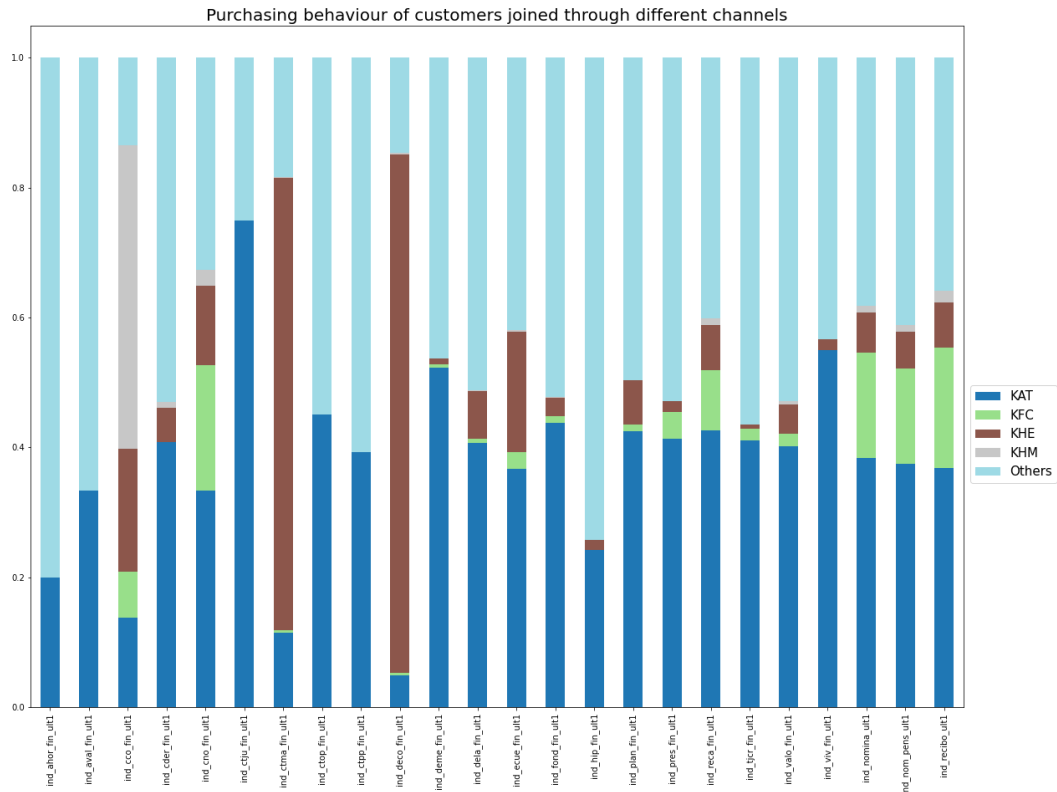


圖10 消費者購買行為與加入途徑的關係比例圖

我們發現canal_entrada (channel used by the customer to join)同樣也是非常好的特徵。通過圖10我們可以發現，在不同商品下，每個途徑加入的比例分佈都相差甚大。例如：通過KAT途徑加入的消費者的消費能力是所有途徑中最高的；通過KHE加入的消費者有偏好購買ind_ctma_fin_ult1 (Más particular Account)和ind_deco_fin_ult1 (Short-term deposits)的現象。

我們也同樣分析過segmento (segmentation: 01 - VIP, 02 - Individuals 03 - college graduated)，並比較每樣商品不同分級消費者的購買比例，如圖11。可以觀察到individuals的購買力最高。而VIP以及college graduated則是在消費數量特別突出。因此這個特徵也同樣非常適合作為訓練用的資料。

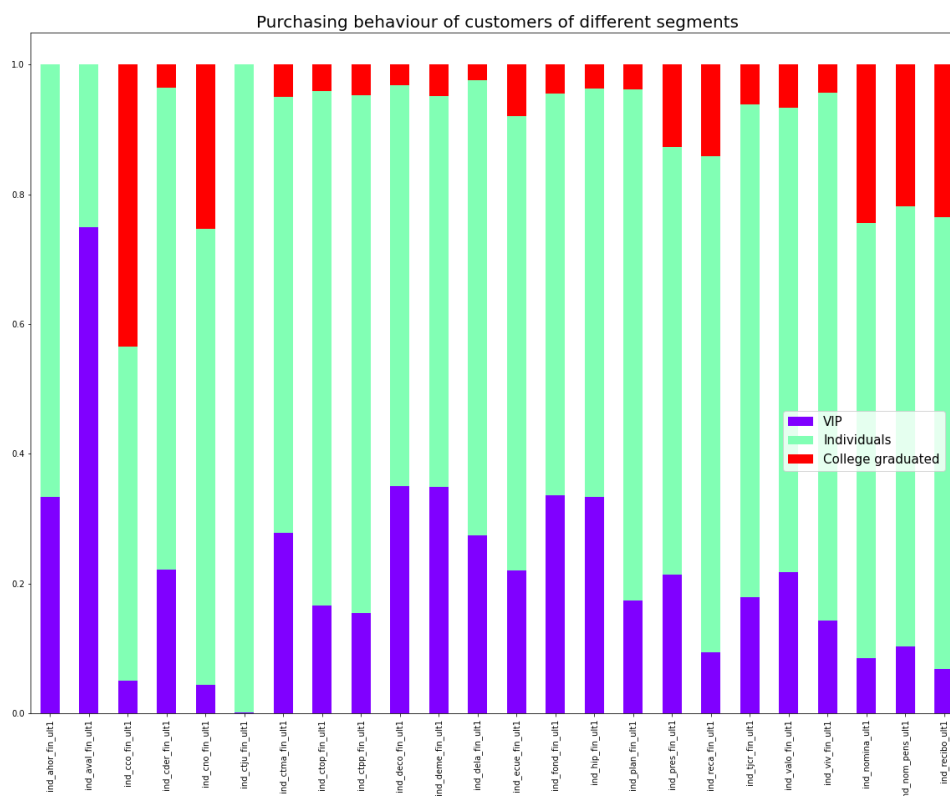


圖11 不同分級的購買行為比較

3.2 模型訓練與結果

我們主要利用 CatBoost 之模型架構，並以 feature selection 挑出來的 7 個最重要資料特徵當作訓練資料集。

CatBoost為分類模型的一種，名稱源於 Category 和 Boost 兩個單詞。由於透過分類和數字特徵組合的各種統計量為類別型的特徵做編碼，因此能夠處理非數值型態的資料，無需對數據特徵進行太複雜預處理就可以將類別轉換為數字。相較於XGBoost 和 LightGBM 這些相似的分類模型，CatBoost的架構更加優化，承襲 Boosting 的優點之外，該演算法也在類別型的特徵上做了一些更公平的特徵工程。

結果 1 - 使用 preprocessing 過後的原始資料進行訓練。

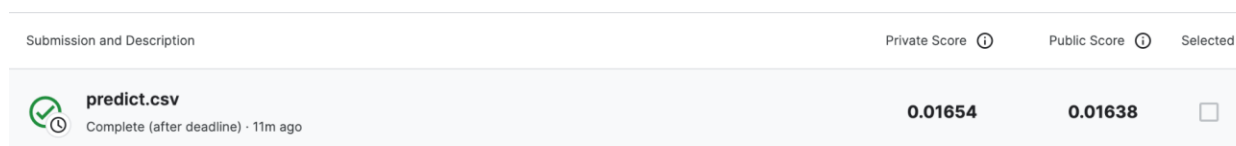



圖12 使用 preprocessing 過後的原始資料測試結果

此資料集會將客戶已購買的商品呈現在往後的月份。由於客戶一般不會再購買前幾個月之前已購買的商品，相同客戶在不同月份的資料若有同樣的商品購買紀錄，則是重複性資料。相對這些重複的紀錄，只考慮客戶購買新商品(前 N 個月前尚未購買的商品)的紀錄也許對於訓練上更有幫助，減少這些重複資料也能降低訓練資料集的大小。

ncodper s	ind_sho r_fin_u ltl	ind_ava l_fin_u ltl	ind_cco r_fin_u ltl	ind_cde r_fin_u ltl	ind_cno r_fin_u ltl	ind_ctj u_fin_u ltl	ind_ctm a_fin_u ltl	ind_cto p_fin_u ltl	ind_ctp p_fin_u ltl	ind_dec o_fin_u ltl	ind_dem e_fin_u ltl	ind_del a_fin_u ltl	ind_ecu e_fin_u ltl	ind_fon d_fin_u ltl	ind_hip r_fin_u ltl	ind_pla n_fin_u ltl	ind_pre s_fin_u ltl	ind_rec a_fin_u ltl	ind_tjc r_fin_u ltl	ind_val o_fin_u ltl	ind_viv r_fin_u ltl	ind_nom ina_ult l	ind_nom pens_u ltl	ind_rec ibo_ult l
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1027006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

圖13 客戶在不同月份無購買紀錄的變化



CatBoostClassifier.csv

Complete (after deadline) · 1d ago

0.03059

0.0302

☐


圖14 嘗試上述的方法，在原始資料集所得到的測試結果

Submission and Description

Private Score ⓘ

Public Score ⓘ

Selected



custom_cb.csv

Complete (after deadline) · 3h ago

0.03066

0.03032

☐

圖15 嘗試上述的方法，在我們 feature selection 後資料集所得到的測試結果

由此可見，對原始資料集使用上述 data cleaning 以及 feature selection 的處理當作訓練資料，有助於提升 CatBoost 模型在 Kaggle 上的測試表現。

4. 結論

本篇報告著重在Santander銀行產品資料集的分析上。我們將產品分為數個類別，透過觀察資料集各項特徵的性質與關聯性，來挑選出對預測模型的訓練最有幫助的特徵，如年齡, 性別等。我們亦透過圖表，來更有說服力的表達這些特徵的分布情況。

除了系統性的觀察與分析資料集本身，我們也利用ANOVA進行特徵選擇，挑選出7個最重要的特徵，最後發現由演算法得出的結果和觀察的結果十分相近，證明了兩種方法相輔相成的可靠性。

除了使用原始資料，我們還利用結合特徵選擇和資料清理這些技巧所篩選出的資料集進行訓練，最後在Kaggle得到0.03的分數，在競爭激烈的排行榜上可以到第150名左右，證明透過完善前處理過的資料，對於預測模型的訓練還是很有幫助的。