預測購買

2020統計學習與機器學習 Group20
張語樵 廖郁華 華敏學

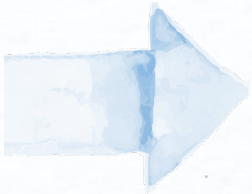
目錄

01 研究動機

預測客戶在下期的購買機率

若企業能夠提前預測訪客是否會購買產品，將會大幅提升企業的經營效率

預測下期的購買機率

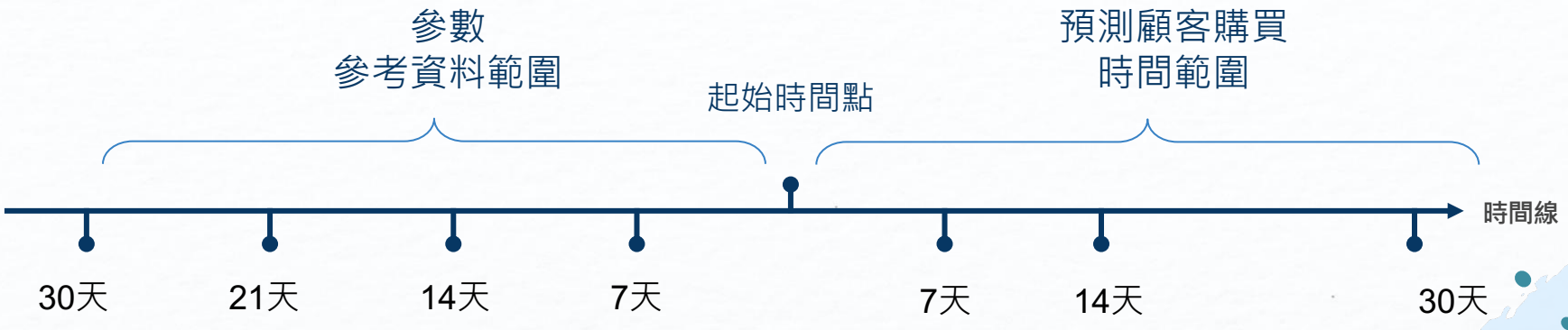


1. 建立營收預估模型
2. 加強簡訊推播
3. 預測現金流量

找出最佳資料範圍及預測範圍

假設

存在一個最佳的行為資料參考範圍以及最佳的預測期間



02 資料集介紹

資料集介紹 – 911APP客戶

業態 - 時尚流行女鞋 (1間商家)

- **Member 會員資料**
  大小:329,871 rows × 10 columns
  時間:～2020/04/30

- **Order 交易資料 (主單資料)**
  OrderSlave (子單資料)
  大小:329,871 rows × 10 columns
  時間:2013/09/30～2020/04/30

- **Behavior 行為資料**
  時間:2018/06/06 ～ 2020/04/30

03 分析流程

分析流程

參數分佈、
參數萃取方法

針對選定的資料集做
多個模型的訓練，
呈現多個模型的比較
結果，選出最佳模型

針對選出的最佳模型，
調整取出資料集的月
份參數，呈現各資料
集的比較結果

選出最佳模型
和最佳資料集，
呈現模型表現

設定參數 – 三大面向

會員資料

- Gender
- Age
- MemberCardLevel

會員特徵

- RFM
- CAI

會員行為

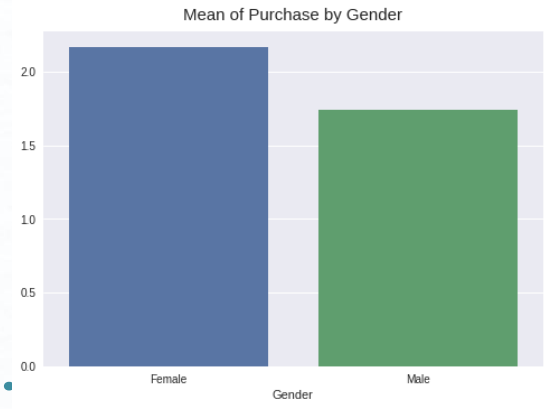
- Mean of Search　平均搜尋字數
- Behavior per Day　日平均行為次數
- Cart Porpotion　加入購物籃比例
- Attention of Product　產品關注度
- Concentration Rate　不同行為的集中度

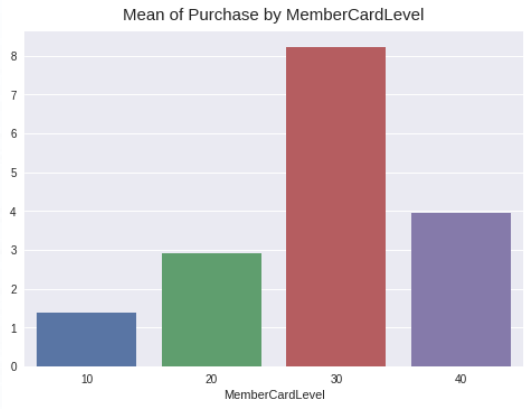
會員資料

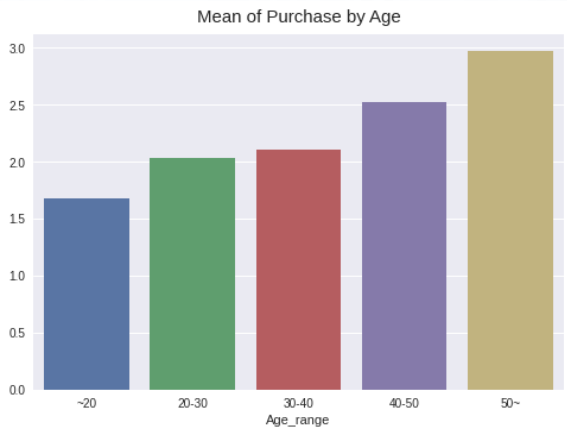
平均購買次數
女性較男性高

會員卡等級為30時
擁有最高的購買次數

購買次數隨著年齡上升





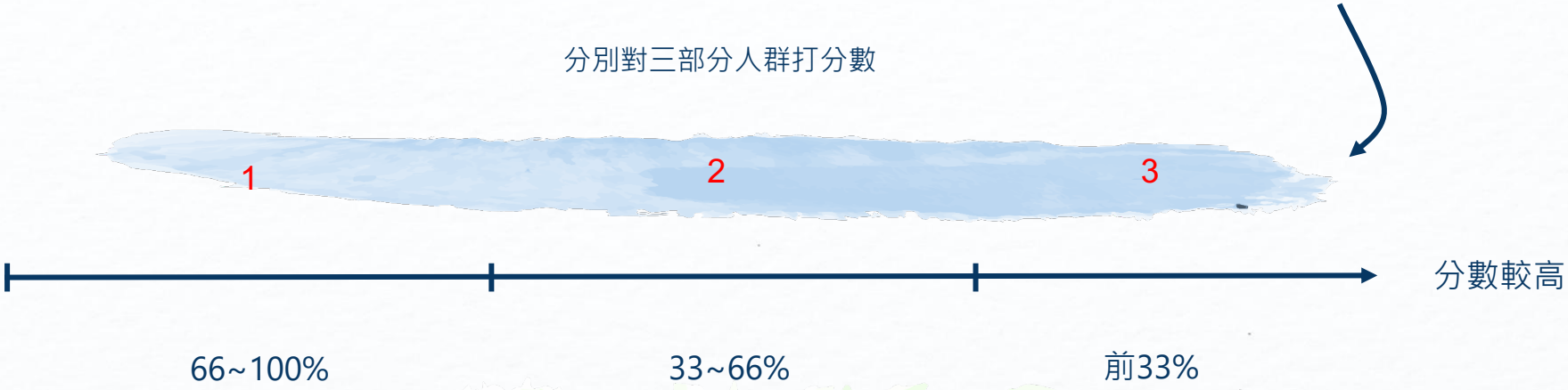


會員特徵 - RFM

RFM分群

Recency : 顧客最後一次購買日
Frequency : 顧客總交易次數
Monetary : 顧客總交易金額

分別對三部分人群打分數



會員特徵 – CAI活躍度

算法

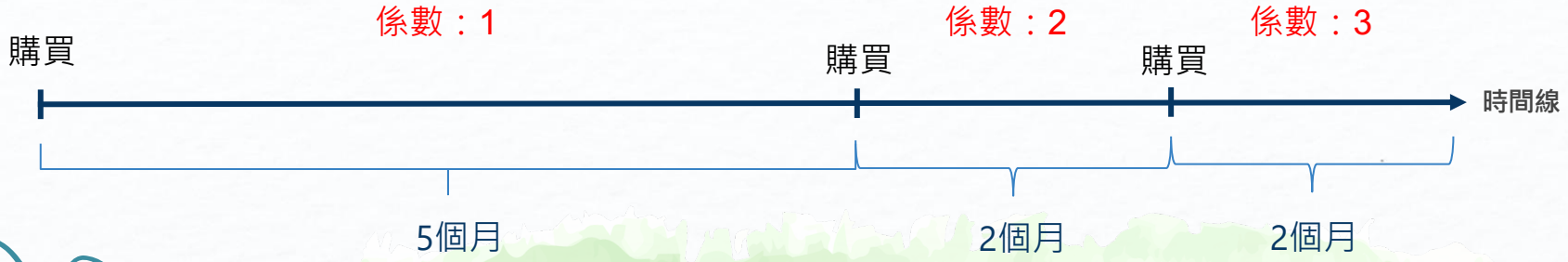
$$CAI = \left\{\frac{\text{平均購買期間} - \text{加權平均購買期間}}{\text{平均購買期間}}\right\} * 100\%$$

概念

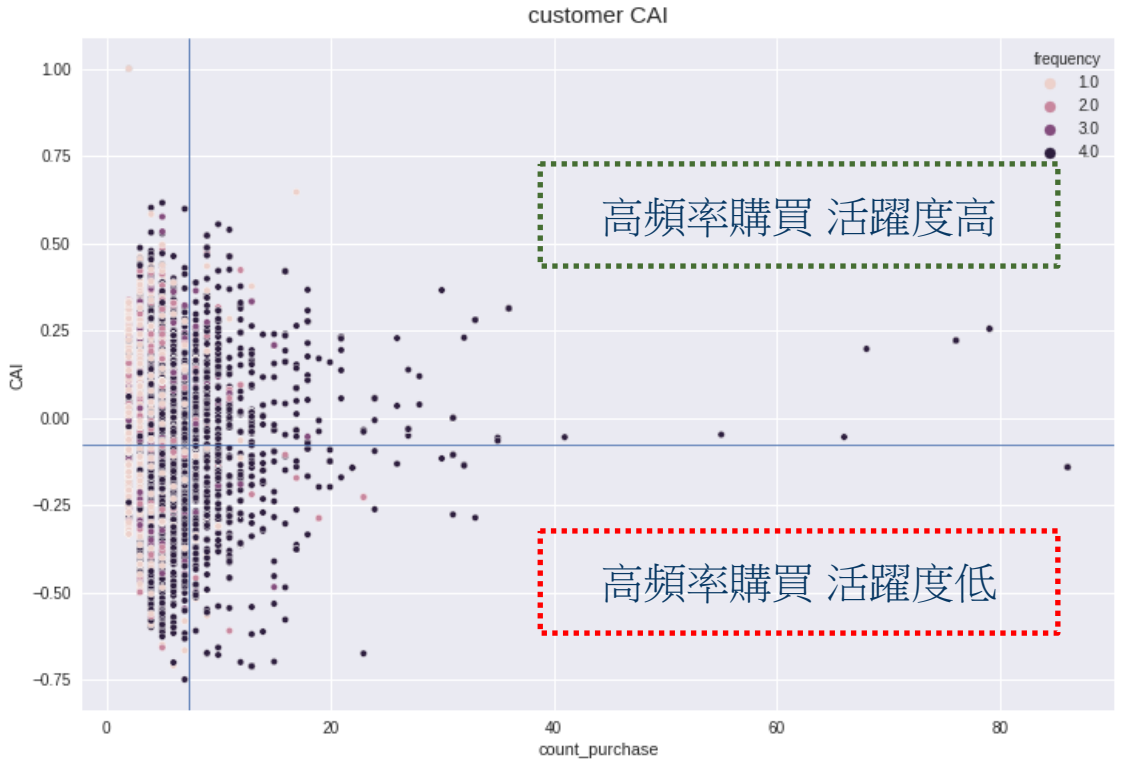
數值 > 0 代表客戶活躍
數值 < 0 代表客戶不活躍

範例

平均購買期間 = 9/3 = 3
加權平均購買期間 = (1*5 + 2*2 + 2*3) / 6 = 2.5
CAI指標 = (3 – 2.5) / 3 = 0.16 * 100 % = 16%



會員特徵 – CAI活躍度


customer CAI
frequency
1.0
2.0
3.0
4.0
高頻率購買 活躍度高
高頻率購買 活躍度低
CAI
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
20
40
60
80
count_purchase

會員行為 – 自訂參數

Behavior per Day

每日行為次數越高
代表消費者可能想要購買商品

公式：總行為次數 / 造訪天數

Mean of Search

搜尋字數越長
代表消費者清楚自己要購買什麼產品

公式：總行為次數 / 造訪天數

Attention of Product

若瀏覽商品越集中
代表消費者專注在某項產品上

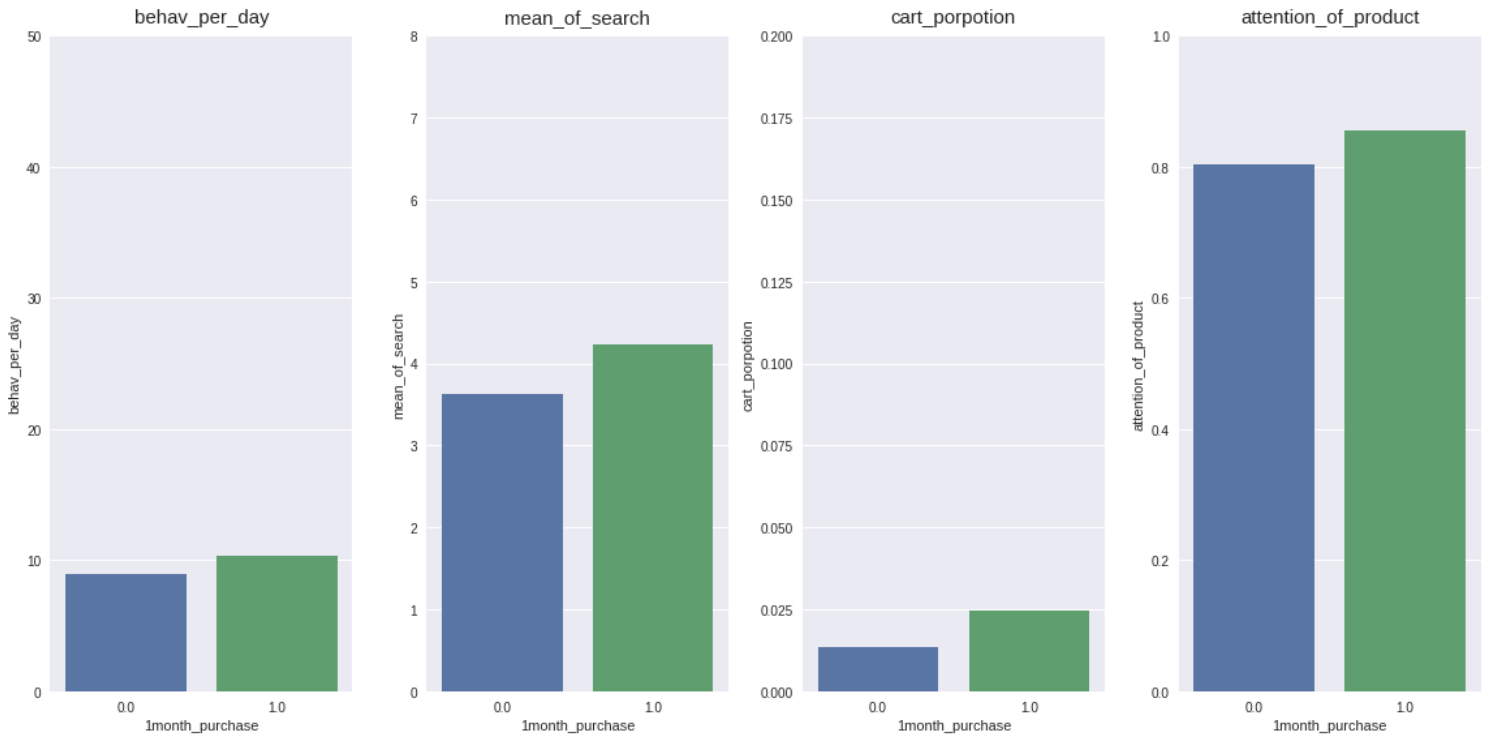
公式： 瀏覽商品次數 / 瀏覽不同商品數

Cart Proportion

放入購物籃的比例越高
代表消費者購買的機率越高

公式：加入購物籃次數 / 總行為次數

會員行為－在購買上有差異



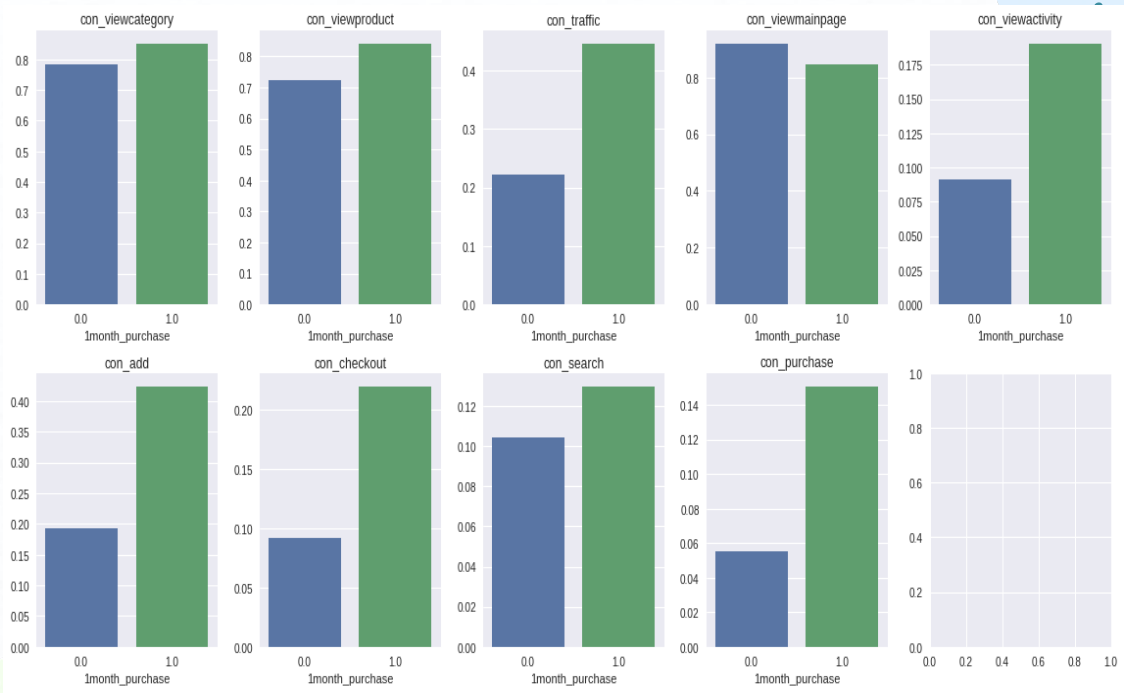
會員行為 – 行為集中度

Con_behavior

集中度越高
代表客戶在此資料期間的行為較活躍

公式：

$$\frac{\text{資料月份平均行為次數}}{\text{3倍資料期間平均行為次數}}$$



訓練資料

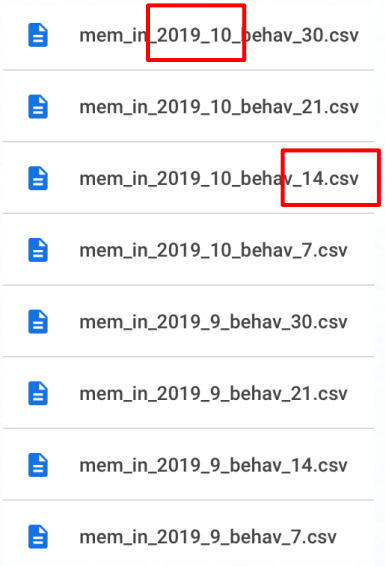
資料起迄時間

會員資料：- 2020/04/22
行為資料：2019/1/1 – 2019/12/31

生成資料

取2019總共12個月份
設定每個月一號為起始點，
往前抓取7、14、21、30天之資料
總共**48份會員資料集**

2019/10/1為起始點



往前取14天
2019/9/16 – 2019/10/1
的客戶行為資料

訓練資料

會員資料

uid	Age	Gender	IsAppInstalled	IsEnableEmail	IsEnablePushNotification	IsEnableShortMessage	MemberCardLevel
mGhnbm0GNCLpyil%2BF7pEIbh%2BBwdwENsx8O0TzX5DdL...	28.000000	Female	True	True	True	True	10

會員特性

CAI	recency	frequency	monetary
0.027523	1.0	4.0	1.0

預測購買

會員在一週、兩週、一個月後是否購買

1week_purchase	2week_purchase	1month_purchase
0.0	0.0	0.0

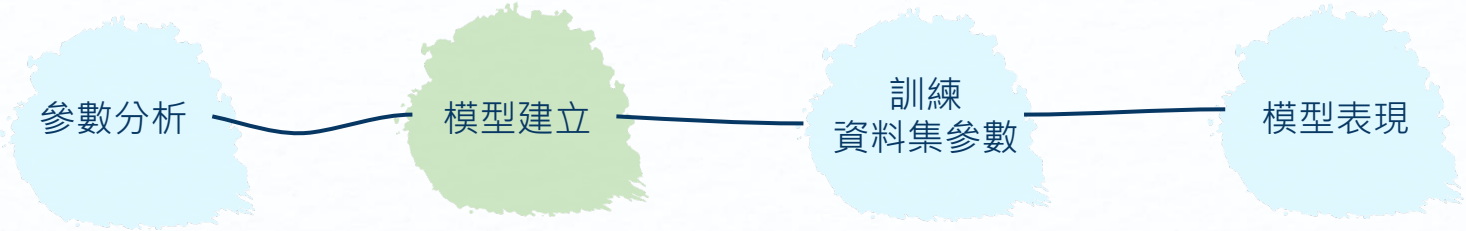
行為資料

behav_per_day	mean_of_search	cart_porpotion	attention_of_product
3.000000	0.0	0.000000	1.000000

集中度計算

con_viewcategory	con_viewproduct	con_traffic	con_viewmainpage	con_viewactivity	con_add	con_checkout	con_search	con_purchase
2.318182	2.25	0.0	2.142857	0.0	0.0	0.0	1.5	0.0

分析流程



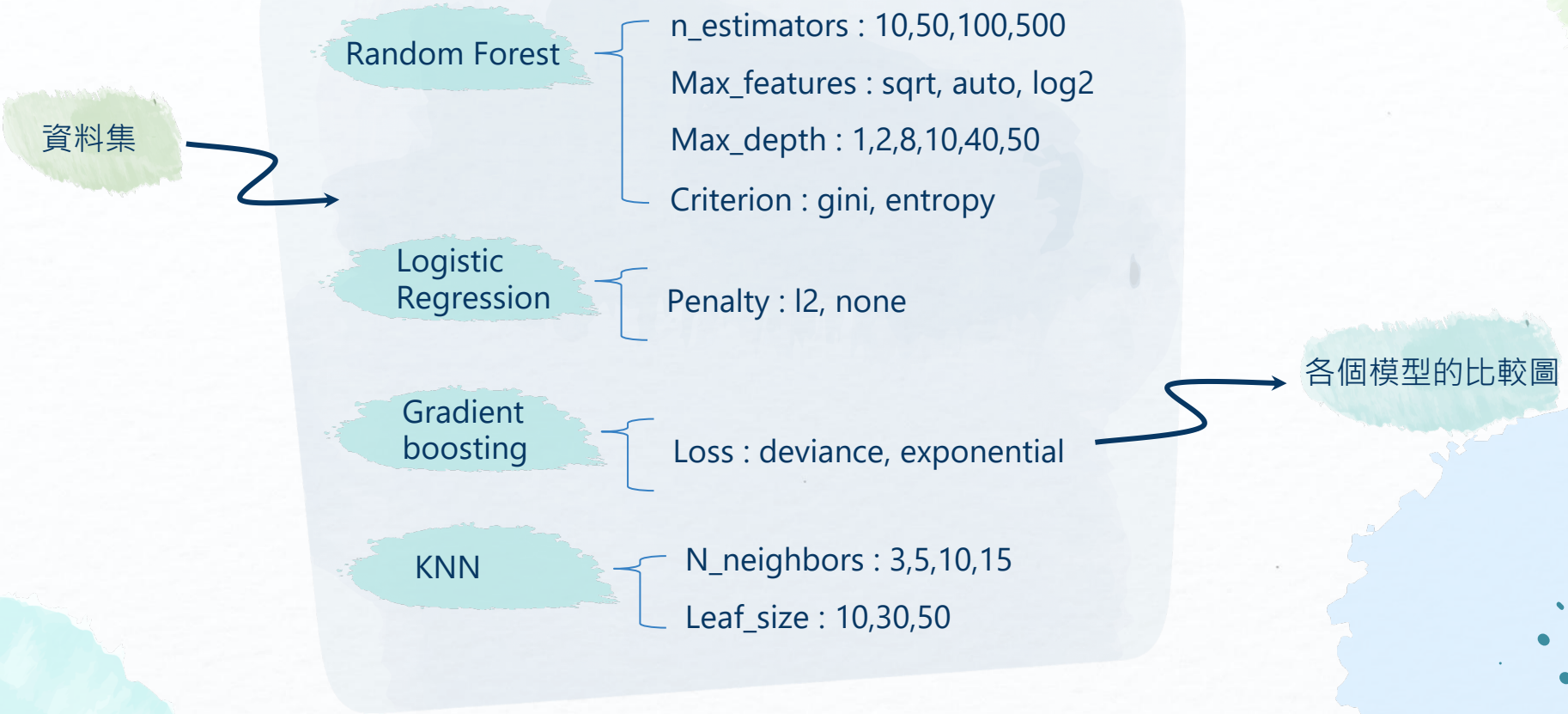
參數分佈、
參數萃取方法

針對選定的資料集做多個模型的訓練，呈現多個模型的比較結果，選出最佳模型

針對選出的最佳模型，調整取出資料集的月份參數，呈現各資料集的比較結果

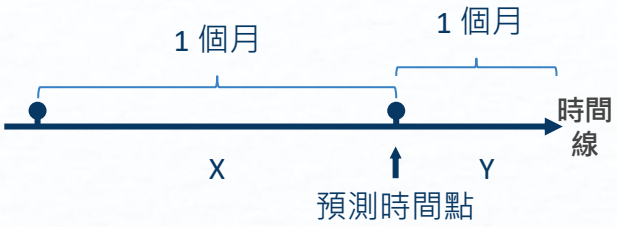
選出最佳模型和最佳資料集，呈現模型表現

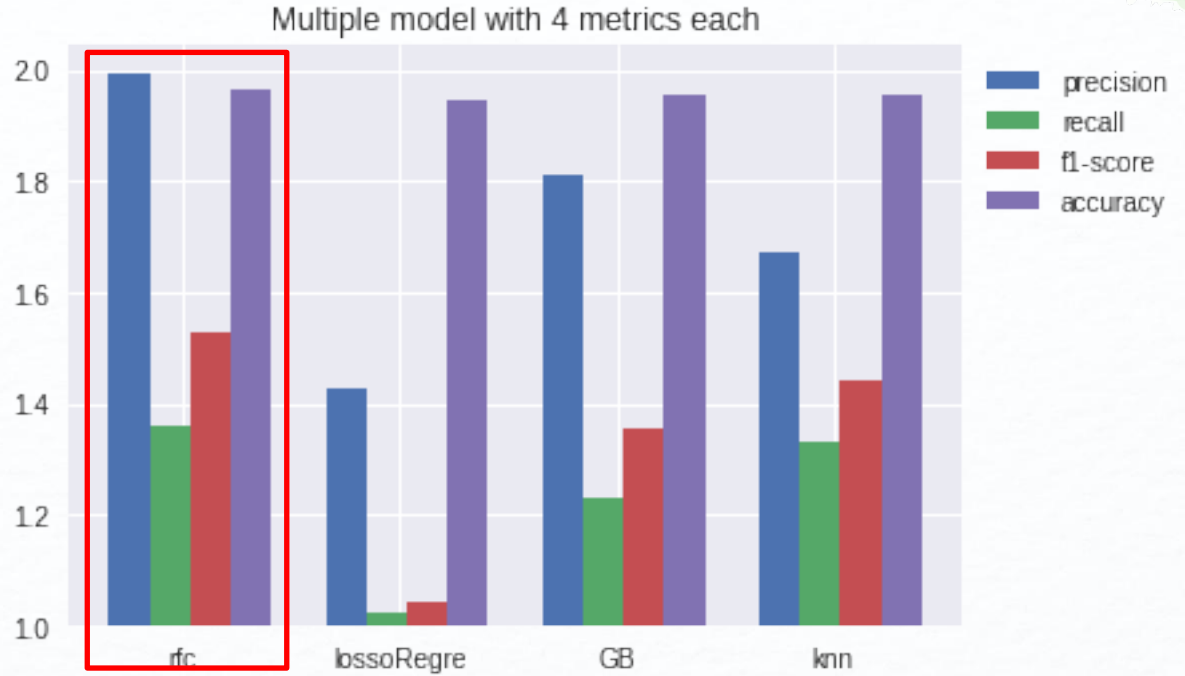
模型建置


資料集
Random Forest
n_estimators : 10,50,100,500
Max_features : sqrt, auto, log2
Max_depth : 1,2,8,10,40,50
Criterion : gini, entropy
Logistic Regression
Penalty : l2, none
Gradient boosting
Loss : deviance, exponential
KNN
N_neighbors : 3,5,10,15
Leaf_size : 10,30,50
各個模型的比較圖

模型結果

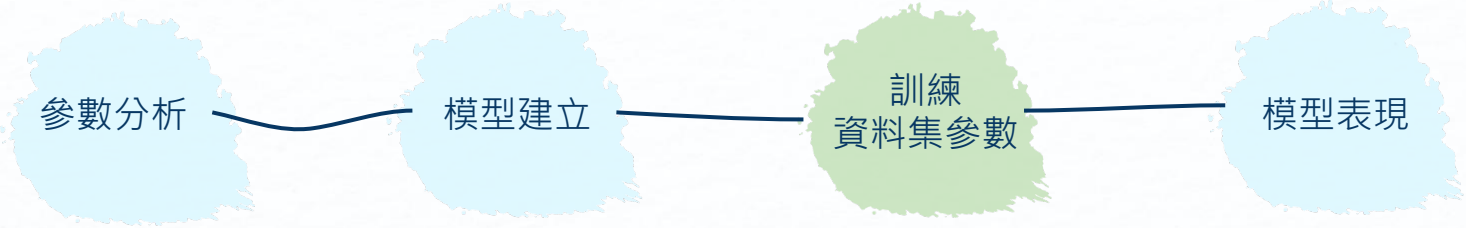
輸入資料





選此模型來調整
資料集日期參數

分析流程



參數分佈、
參數萃取方法

針對選定的資料集做多個模型的訓練，呈現多個模型的比較結果，選出最佳模型

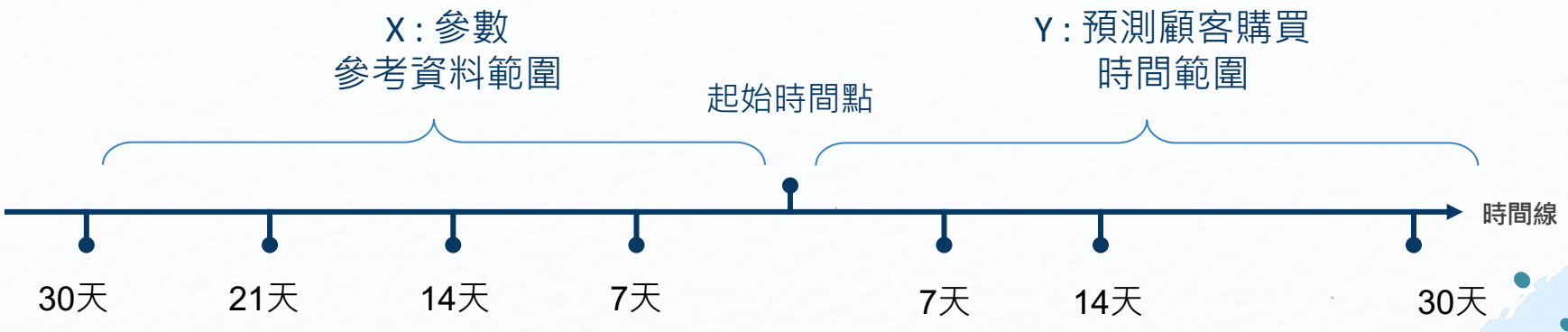
針對選出的最佳模型，調整取出資料集的月份參數，呈現各資料集的比較結果

選出最佳模型和最佳資料集，呈現模型表現

調整資料集參數

Remind 目標

找出最佳的行為資料區間和預測區間



調整資料集參數

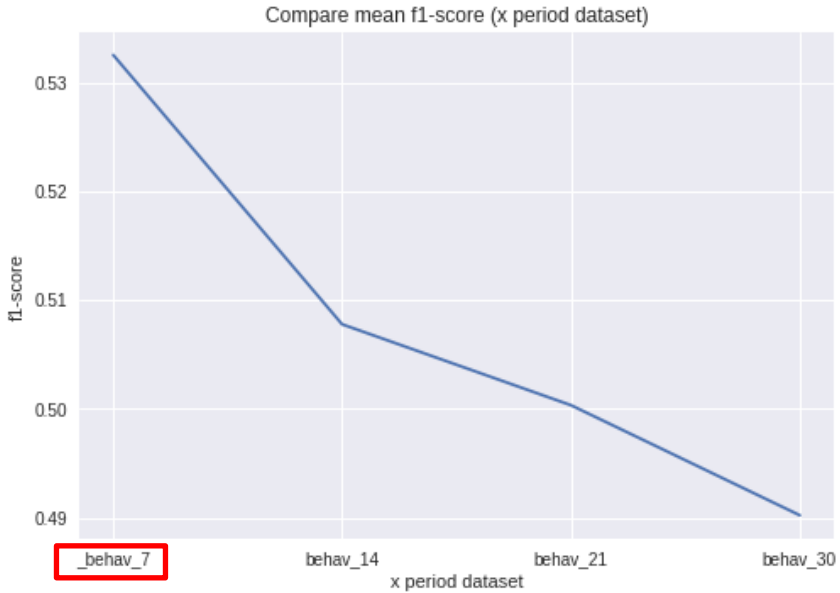
	x period	y period	dataset1	dataset2	dataset3	dataset4	dataset5	dataset6	dataset7	dataset8	dataset9	dataset10	dataset11
0	behav_7	1week	0.483221	0.552632	0.561111	0.472393	0.343195	0.412500	0.500000	0.538153	0.540925	0.543147	0.630000
1	behav_7	2week	0.405694	0.548544	0.523894	0.559350	0.342105	0.433898	0.647191	0.563981	0.507246	0.646154	0.616046
2	behav_7	1month	0.535922	0.533762	0.514286	0.430556	0.426426	0.626230	0.618182	0.581921	0.632558	0.636150	0.665302
3	behav_14	1week	0.472050	0.520000	0.509284	0.370588	0.397727	0.364780	0.513514	0.546075	0.568047	0.531250	0.555831
4	behav_14	2week	0.390093	0.557957	0.522835	0.488688	0.373626	0.441989	0.636559	0.544402	0.575758	0.629393	0.565161
5	behav_14	1month	0.497638	0.512758	0.451351	0.390438	0.404348	0.535565	0.545794	0.544843	0.593407	0.598383	0.606178
6	behav_21	1week	0.453039	0.503704	0.493438	0.379404	0.447368	0.357143	0.454545	0.531250	0.589812	0.479508	0.569507
7	behav_21	2week	0.403270	0.525328	0.496753	0.463768	0.407216	0.422886	0.639098	0.550098	0.573503	0.622289	0.557870
8	behav_21	1month	0.478873	0.549451	0.413333	0.419463	0.393939	0.530026	0.489985	0.547798	0.592718	0.571181	0.603503
9	behav_30	1week	0.400000	0.525180	0.490654	0.426606	0.408163	0.367568	0.467532	0.528662	0.548718	0.494624	0.537118
10	behav_30	2week	0.385027	0.527273	0.492958	0.492537	0.366071	0.416476	0.613430	0.541586	0.533113	0.630712	0.537415
11	behav_30	1month	0.460705	0.522876	0.421281	0.365891	0.396040	0.494062	0.509537	0.540773	0.566692	0.578606	0.589490

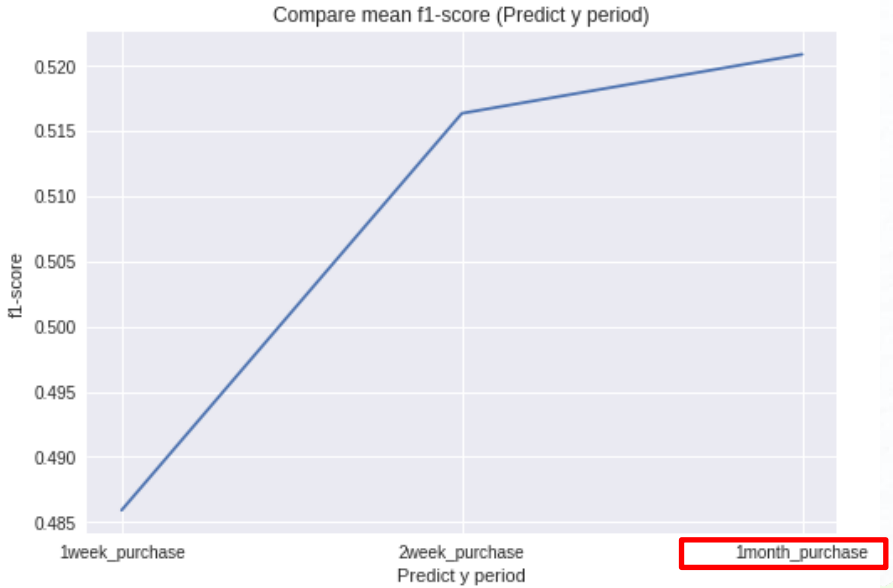
加總

調整資料集參數

	x period	y period	dataset1	dataset2	dataset3	dataset4	dataset5	dataset6	dataset7	dataset8	dataset9	dataset10	dataset11
0	behav_7	1week	0.483221	0.552632	0.561111	0.472393	0.343195	0.412500	0.500000	0.538153	0.540925	0.543147	0.630000
1	behav_7	2week	0.405694	0.548544	0.523894	0.559350	0.342105	0.433898	0.647191	0.563981	0.507246	0.646154	0.616046
2	behav_7	1month	0.535922	0.533762	0.514286	0.430556	0.426426	0.626230	0.618182	0.581921	0.632558	0.636150	0.665302
3	behav_14	1week	0.472050	0.520000	0.509284	0.370588	0.397727	0.364780	0.513514	0.546075	0.568047	0.531250	0.555831
4	behav_14	2week	0.390093	0.557957	0.522835	0.488688	0.373626	0.441989	0.636559	0.544402	0.575758	0.629393	0.565161
5	behav_14	1month	0.497638	0.512758	0.451351	0.390438	0.404348	0.535565	0.545794	0.544843	0.593407	0.598383	0.606178
6	behav_21	1week	0.453039	0.503704	0.493438	0.379404	0.447368	0.357143	0.454545	0.531250	0.589812	0.479508	0.569507
7	behav_21	2week	0.403270	0.525328	0.496753	0.463768	0.407216	0.422886	0.639098	0.550098	0.573503	0.622289	0.557870
8	behav_21	1month	0.478873	0.549451	0.413333	0.419463	0.393939	0.530026	0.489985	0.547798	0.592718	0.571181	0.603503
9	behav_30	1week	0.400000	0.525180	0.490654	0.426606	0.408163	0.367568	0.467532	0.528662	0.548718	0.494624	0.537118
10	behav_30	2week	0.385027	0.527273	0.492958	0.492537	0.366071	0.416476	0.613430	0.541586	0.533113	0.630712	0.537415
11	behav_30	1month	0.460705	0.522876	0.421281	0.365891	0.396040	0.494062	0.509537	0.540773	0.566692	0.578606	0.589490

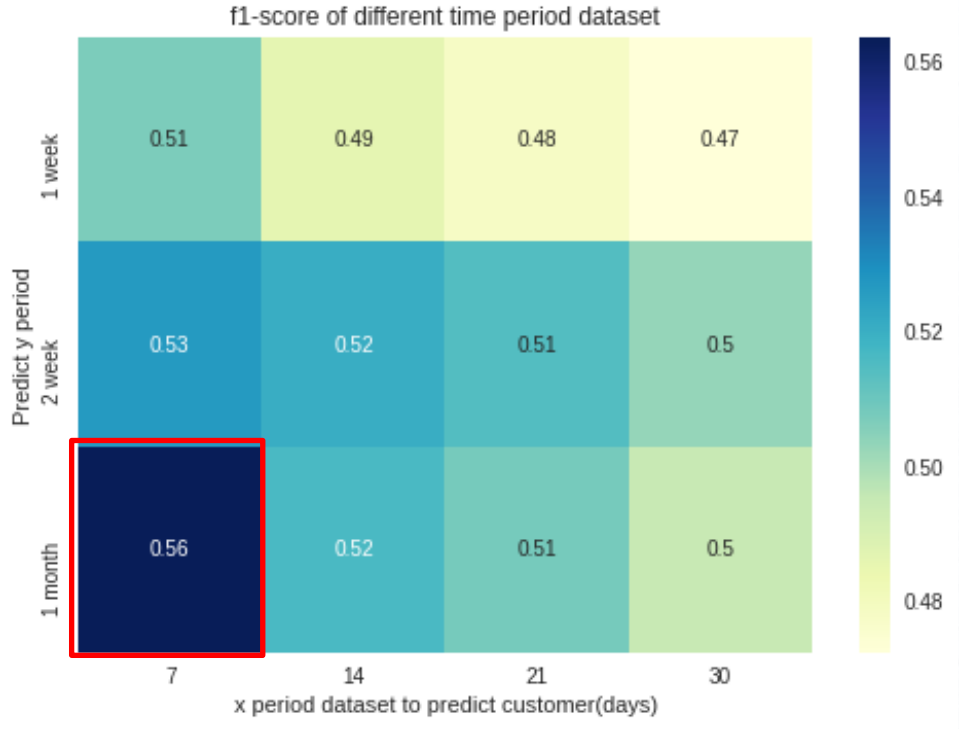
調整資料集參數


Compare mean f1-score (x period dataset)
f1-score
0.53
0.52
0.51
0.50
0.49
_behav_7
behav_14
behav_21
behav_30
x period dataset

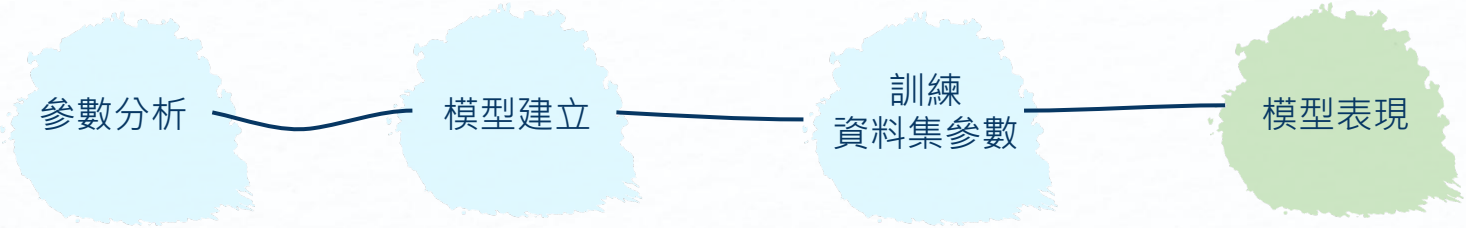

Compare mean f1-score (Predict y period)
f1-score
0.520
0.515
0.510
0.505
0.500
0.495
0.490
0.485
1week_purchase
2week_purchase
1month_purchase
Predict y period

調整資料集參數

選此資料集參數
當作訓練資料



分析流程



參數分佈、
參數萃取方法

針對選定的資料集做
多個模型的訓練，
呈現多個模型的比較
結果，選出最佳模型

針對選出的最佳模型，
調整取出資料集的月
份參數，呈現各資料
集的比較結果

選出最佳模型
和最佳資料集，
呈現模型表現

模型表現

before

	precision	recall	f1-score	support
1	0.892704	0.341544	0.494062	609.000000
0	0.961461	0.997507	0.979152	10029.000000
accuracy	0.959955	0.959955	0.959955	0.959955
macro avg	0.927082	0.669525	0.736607	10638.000000
weighted avg	0.957525	0.959955	0.951382	10638.000000

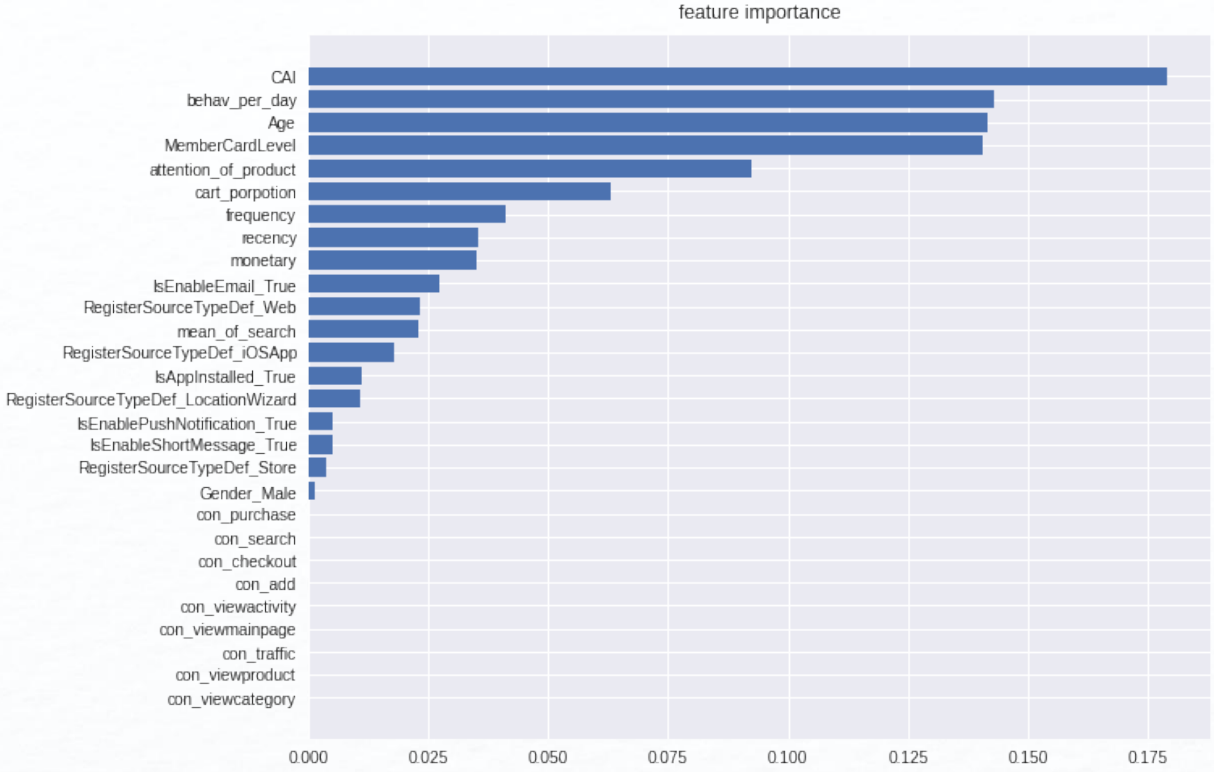
資料集：30 天行為資料 -> 預測 1 個月使否購買
模型：Random forest

after

	precision	recall	f1-score	support
1	0.890756	0.417671	0.563754	424.000000
0	0.946057	0.994440	0.969635	4238.818182
accuracy	0.943514	0.943514	0.943514	0.943514
macro avg	0.918406	0.706056	0.766694	4662.818182
weighted avg	0.940897	0.943514	0.934080	4662.818182

資料集：7 天行為資料 -> 預測 1 個月使否購買
模型：Random forest

參數重要性



資料集：7 天行為資料 + 1 個月回購顧客

模型：Random forest

06

結論

結論

特徵分析

表現較好的特徵：
- CAI
- Age
- Behavior per day

表現差的特徵：
- 行為集中度表現差

資料集及預測時間區段

最佳預測區間
行為資料：前七天
預測時段：一個月內

模型

Random Forest
Precision： 89%
Recall： 42%
F1 score： 56%