南山人壽專題-運用機器學習於客戶接觸點資訊以優化精準行銷

指導業師: 賴昌作 協理　指導老師: 石百達 教授
第四組
陳永進 鄭晴文 孟家瑜 林羿帆

大綱

1. 專題及資料簡介
2. 檔案合併及EDA
3. 資料前處理-缺漏值處理
4. 機器學習模型Y值處理
5. 機器學習模型X值處理
   - feature importance
   - feature tools
6. 特徵選取
7. 處理資料不平衡
8. 模型建立
9. 報表展示

專題簡介

動機	● 運用客戶數位足跡或接觸點上最新的資料 ● 即時並精準預測客戶的人生階段及商品服務需求
痛點	● 已有「動態」客戶接觸資料卻未善加利用 ● 隨著數位時代來臨，運用最新動態資料才能即時預測並結合精準行銷動態調整客群名單，以掌握客戶需求
目的及期許	● 專案主要目的在於提升客戶商機模型預測精準度 ● 解決方案需就模型準確性、穩定性、可解釋性等說明其優劣及相關數據推論根據

資料簡介

- 客戶輪廓檔

  客戶的基本資料

- 客戶接觸點資料: 客戶歷程檔、客戶金流事件檔

  客戶申購保單、申請契變、數位接觸、電話客服、理賠、預期金流的紀錄

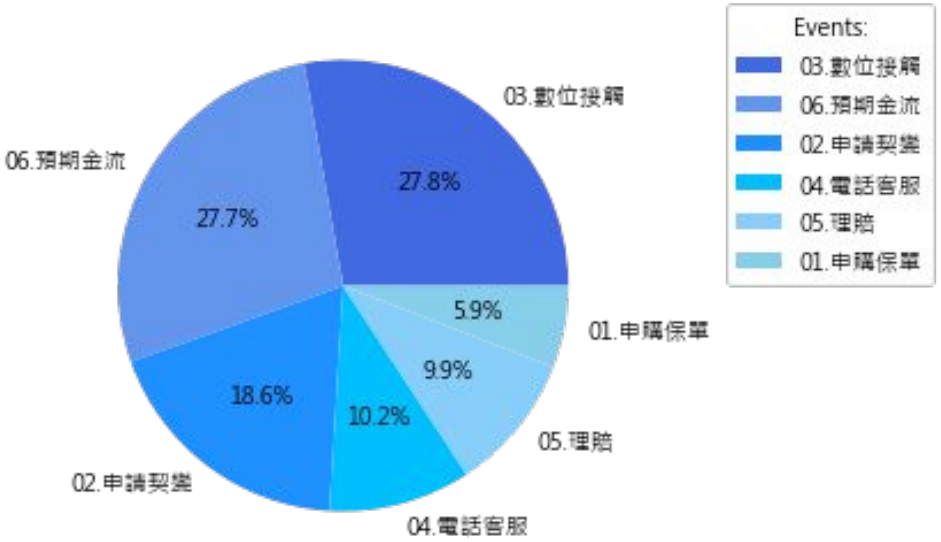
- 客戶再購檔

  客戶再購保單的類型、金額

檔案合併及EDA

- EVENT細項處理完將加入表格
- 合併客戶歷程檔、客戶金流事件檔
- 將客戶歷程檔之EVENT 中是否再購欄位與客戶輪廓檔合併
- 將業務員分群新資料加入

▼ 各個Event所佔的比例



資料前處理-缺漏值處理

<u>刪除缺漏值</u>

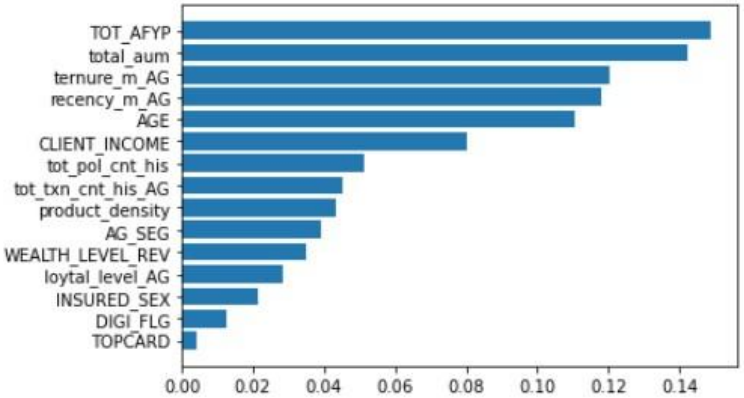
- 刪除輪廓檔性別為空值的客戶資料(共3筆)
- 刪除接觸點檔性別為空值的客戶資料

<u>填充缺漏值</u>

- 輪廓檔CLIENT_INCOME、total_aum（總資產-客戶總繳保費）和TOT_AFYP（客戶年繳保費）欄位

資料前處理-缺漏值處理

- Random Forest’s Feature Importance如下:



- 利用已知客戶的tenure_m_AG(客戶戶齡)、recency_m_AG(最近生效日距今)及Age這三種特徵分組，再將結果補回有缺失的客戶

機器學習模型Y值處理

- **Y值類別 : 0,1**

0	隔年沒有申購保單
1	隔年有申購保單

機器學習模型X值處理(1)

- 行為(X值)區間:2019/1～2019/12
- 預測(Y值)區間:2020/1～2020/12
- 統計日 : 2019/12/31
- 輪廓檔資料: 15種資訊, 產生15個X變數
- 接觸點事件: 申購保單、申請契變、數位接觸、電話客服、理賠、預期金流

申購保單、申請契變、數位接觸、電話客服、理賠

- 2019年的資料
- 74種事件, 74個X變數

預期金流

- 2019、2020年的資料
- 3種事件, 6個X變數

機器學習模型X值處理(2)-feature importance

```
TOT_AFYP: 0.113
total_aum: 0.112
recency_m_AG: 0.099
ternure_m_AG: 0.097
AGE: 0.090
CLIENT_INCOME: 0.070
tot_pol_cnt_his: 0.046
tot_txn_cnt_his_AG: 0.043
product_density: 0.040
AG_SEG: 0.035
WEALTH_LEVEL_REV: 0.032
loytal_level_AG: 0.024
INSURED_SEX: 0.020
 2.還本金20: 0.011
 3.繳費期滿20: 0.010
repurchase_2019: 0.010
 2.還本金19: 0.009
 3.繳費期滿19: 0.009
DIGI_FLG: 0.007
N1.續期保費改為轉帳/信用卡繳費: 0.007
20.疾病和死亡的外因: 0.006
CP(保戶圓地網頁): 0.006
A1.要被保人聯絡資訊變更(地址/電話/Email): 0.005
 1.滿期金20: 0.005
```

```
99.其他: 0.004
02.收費相關: 0.004
B7.保障內容變更(或復效): 0.004
10.其他: 0.004
A7.FATCA變更: 0.004
TOPCARD: 0.004
01.保單解說: 0.003
 1.滿期金19: 0.003
CLUB(南山聚樂部): 0.003
B1.受益人變更: 0.003
11.消化系統疾病: 0.003
S.終止契約: 0.003
05.指定AG共同服務/更換AG: 0.003
L.保單借款: 0.003
B4.繳法變更: 0.003
APP(保戶圓地APP): 0.003
03.契變/復效: 0.002
02.腫瘤: 0.002
S.投資型保單_贖回/提領: 0.002
GUI.投資型保單相關變更: 0.002
14.泌尿生殖系統疾病: 0.002
10.呼吸系統疾病: 0.002
15.妊娠、分娩和產褥期: 0.002
B3.職業變更: 0.002
A3.續期保費改為自行繳費: 0.002
```

- 使用 RandomForestClassifier().feature_importances_
- 如果某個特徵的feature importance太低，可優先剔除該特徵
- 剔除feature importance = 0.000, 0.001的特徵共46個，剩下49個特徵

機器學習模型X值處理(3)-feature tools

```
feature_matrix,feature_names = ft.dfs(entityset=es,   #使用DFS來自動建立新特徵
                                      target_entity = 'repurchase_rate',
                                      max_depth = 2,
                                      verbose   = 1,
                                      n_jobs = 3)
```

```
Built 674 features
EntitySet scattered to 3 workers in 18 seconds
Elapsed: 00:51 | Progress: 100%|
```

- 輪廓檔利用特徵工程 產出更多相關的feature
- 將cashflow & journey 資料合併，刪除再購欄位，再匯入整理好的再購與否資料
- 搭配feacturetools 去自動生成新特徵，也有搭配人工方式去找出其他特徵。feature tools共生成674個新特徵。
- 利用catboost驗證生成特徵的效果
- 之後再用feacture selection 挑出較好的特徵去建模

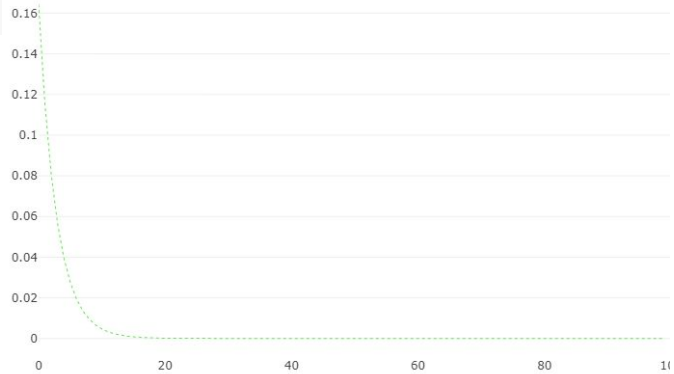
機器學習模型X值處理(4)-feature tools結果驗證

- CatBoost驗證生成特徵的效果

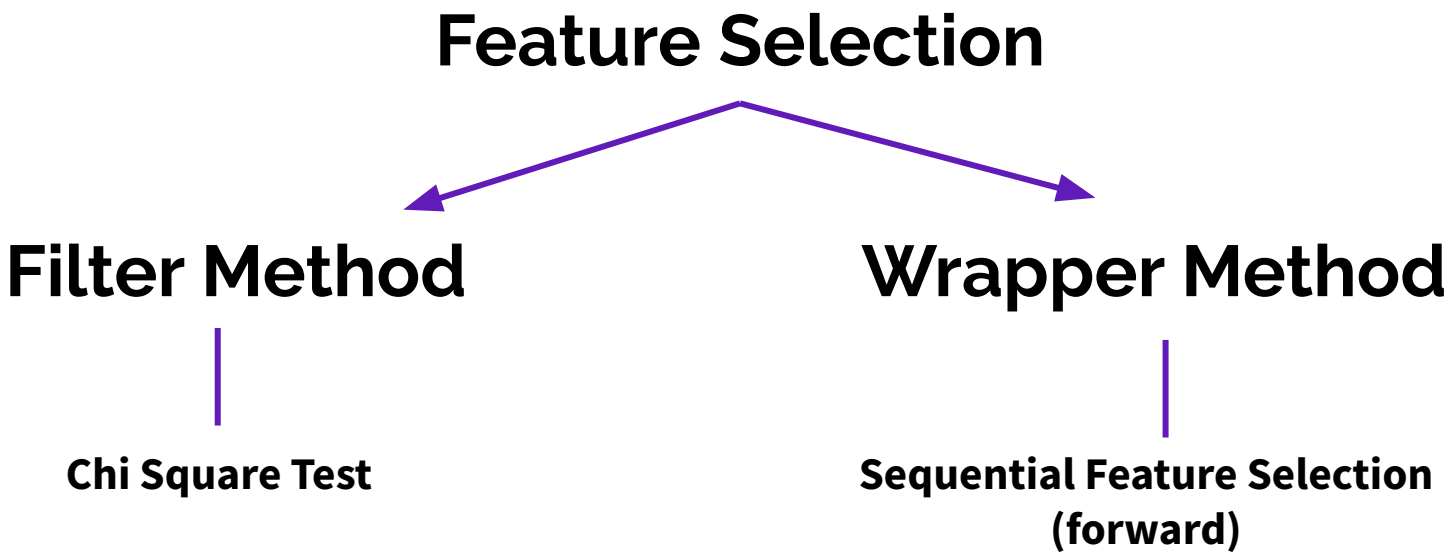
```
model_cat = CatBoostRegressor(iterations=100, learning_rate=0.3,      #評估指標是RMSE（均方根誤差）
                              depth=6, eval_metric='RMSE',loss_function='RMSE',
                              random_seed=7)
# training model
model_cat.fit(X_train, y_train, cat_features=categorical_features,
              use_best_model=True,plot=True)
# validation score
model_cat.score(X_test, y_test)
```

```
                                .
                                .
 96:     learn: 0.0000001        total: 3.49s    remaining: 108ms
 97:     learn: 0.0000001        total: 3.53s    remaining: 72ms
 98:     learn: 0.0000001        total: 3.56s    remaining: 36ms
 99:     learn: 0.0000001        total: 3.6s     remaining: 0us

 0.9999999999965865
```



特徴選取


Feature Selection
Filter Method
Wrapper Method
Chi Square Test
Sequential Feature Selection
(forward)

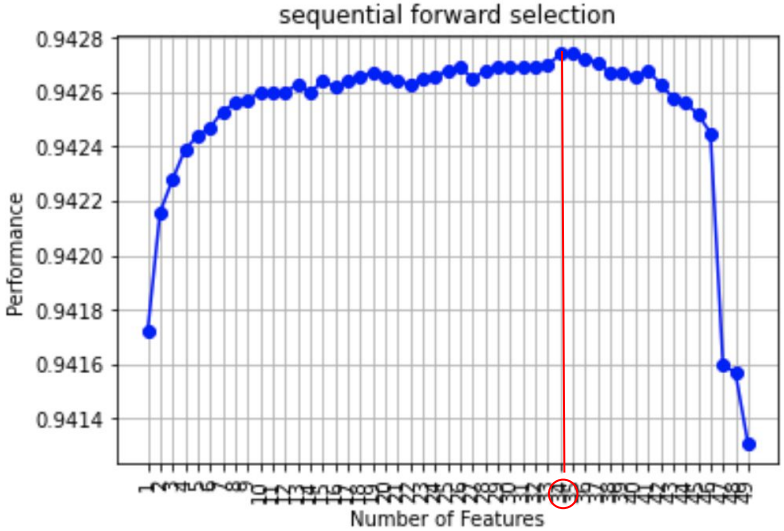
特徵選取-Chi square test

- sklearn　-SelectKBest-
- 採卡方檢驗，透過計算特徵統計值來選出重要且相關的特徵給機器學習
- 圖為49取35 因此方法並非最佳選取結果 後來採用sequential forward selection 方法

```
['AGE', 'WEALTH_LEVEL_REV', 'loytal_level_AG', 'CLIENT_INCOME', 'total_aum', 'TOPCARD', 'DIGI
_FLG', 'ternure_m_AG', 'recency_m_AG', 'product_density', 'tot_pol_cnt_his', 'tot_txn_cnt_his
_AG', 'TOT_AFYP', 'repurchase_2019', 'N1.續期保費改為轉帳/信用卡繳費', 'A1.要被保人聯絡資訊變更(地址/
電話/Email)', 'CP(保戶園地網頁)', 'B1.受益人變更', 'APP(保戶園地APP)', 'S.終止契約', 'A7.FATCA變更',
'B7.保障內容變更(或復效)', 'L.保單借款', '20.疾病和死亡的外因', '99.其他', 'A3.續期保費改為自行繳費',
'S.投資型保單_贖回/提領', 'CLUB(南山聚樂部)', 'B4.繳法變更', 'GUI.投資型保單相關變更', '1.滿期金20',
'3.繳費期滿19', '3.繳費期滿20', '2.還本金19', '2.還本金20']
```

- 特徵工程使用此方法以674 取25

特徵選取-Sequential Forward Selection



▲ 表現最好時為34個特徵

特徵選取-34個特徵

tot_pol_cnt_his	B4.繳法變更	2.還本金19
loytal_level_AG	A3.續期保費改為自行繳費	A7.FATCA變更
ternure_m_AG	S.終止契約	1.滿期金19
product_density	INSURED_SEX	B1.受益人變更
05.指定AG共同服務/更換AG	DIGI_FLG	GUI.投資型保單相關變更
CLUB(南山聚樂部)	B3.職業變更	WEALTH_LEVEL_REV
1.滿期金20	03.契變/復效	20.疾病和死亡的外因
CP(保戶園地網頁)	14.泌尿生殖系統疾病	TOPCARD
repurchase_2019	APP(保戶園地APP)	3.繳費期滿19
AGE	A1.要被保人聯絡資訊變更(地址/電話/Email)	11.消化系統疾病
AG_SEG	01.保單解說	
02.腫瘤	S.投資型保單_贖回/提領	

特徵選取與解釋-前10名特徵(1)

特徵	說明	解釋
tot_pol_cnt_his	客戶曾持有之保單數	忠誠度和客戶曾持有之保單數有關，忠誠度高的客戶較有可能再購保單
loytal_level_AG	忠誠度	
ternure_m_AG	客戶戶齡	戶齡較久的客戶，可能更有再購保單的需求
product_density	產品密度	購買公司多種保單的客戶，更有需求再購保單
05.指定AG共同服務/更換AG	指定業務員共同服務/更換業務員	與業務員的互動可能產生購買保單的商機

特徵選取與解釋-前10名特徵(2)

特徵	說明	解釋
CLUB(南山聚樂部)	客戶瀏覽南山聚樂部網頁的次數	客戶可能考慮購買保單，才會登入聚樂部網站
1.滿期金20	2020年的滿期金	今年可獲得滿期金的客戶較可能再購保單
CP(保戶園地網頁)	客戶瀏覽保戶園地網頁的次數	和CLUB的解釋類似
repurchase_2019	2019年的購買記錄	去年曾購買健康暨意外險(AH)保單的客戶可能今年續買保單
AGE	年齡	壯年的客戶較有經濟能力再購保單

處理資料不平衡-SMOTE

- 利用oversampling的方式，解決資料不平衡問題
- 產生相似合成樣本，隨機增大少數的樣本數量

	原本	使用SMOTE後
2020無申購	75347	75347
2020有申購	4650	**75347**

模型建立及表現評估

- 建立四種模型

1. Random Forest
2. SVM
3. XGBoost → Logistic Regression
4. NN

- 使用套件sklearn.{model}.predict_proba，預測每位客戶的再購機率
- 將資料依照此機率由高至低排序，計算出模型的捕捉率

Random Forest

report

	precision	recall	f1-score	support
0	0.96	0.78	0.86	18810
1	0.11	0.43	0.18	1190
accuracy			0.76	20000
macro avg	0.53	0.61	0.52	20000
weighted avg	0.91	0.76	0.82	20000

Random Forest_Training Data

前20%客戶約可補捉51%再購客戶

Score	客戶數			累計			補捉率
	客戶數	再購客戶數	再購率	客戶數	再購客戶數	再購率	
1%	1,000	350	35.0%	1,000	350	35.0%	6.0%
5%	4,000	670	16.8%	5,000	1,020	20.4%	17.5%
10%	5,000	655	13.1%	10,000	1,675	16.8%	28.7%
20%	9,999	1,277	12.8%	19,999	2,952	14.8%	50.5%
30%	9,999	778	7.8%	29,998	3,730	12.4%	63.9%
40%	9,999	662	6.6%	39,997	4,392	11.0%	75.2%
50%	9,999	515	5.2%	49,996	4,907	9.8%	84.0%
60%	9,996	355	3.6%	59,992	5,262	8.8%	90.1%
70%	10,001	241	2.4%	69,993	5,503	7.9%	94.2%
80%	9,999	170	1.7%	79,992	5,673	7.1%	97.1%
90%	10,000	120	1.2%	89,992	5,793	6.4%	99.2%
100%	10,005	47	0.5%	99,997	5,840	5.8%	100.0%
全部	99,997	5,840	5.8%	99,997	5,840	5.8%	

Random Forest_Testing Data

前20%客戶約可補捉41%再購客戶

Score	客戶數			累計			補捉率
	客戶數	再購客戶數	再購率	客戶數	再購客戶數	再購率	
1%	1,000	234	23.4%	1,000	234	23.4%	3.9%
5%	4,000	538	13.5%	5,000	772	15.4%	12.9%
10%	5,000	577	11.5%	10,000	1,349	13.5%	22.5%
20%	9,999	1,094	10.9%	19,999	2,443	12.2%	40.7%
30%	9,999	753	7.5%	29,998	3,196	10.7%	53.3%
40%	9,996	710	7.1%	39,994	3,906	9.8%	65.1%
50%	10,002	550	5.5%	49,996	4,456	8.9%	74.3%
60%	9,999	479	4.8%	59,995	4,935	8.2%	82.3%
70%	9,999	337	3.4%	69,994	5,272	7.5%	87.9%
80%	9,999	328	3.3%	79,993	5,600	7.0%	93.3%
90%	9,999	257	2.6%	89,992	5,857	6.5%	97.6%
100%	9,999	142	1.4%	99,991	5,999	6.0%	100.0%
全部	99,991	5,999	6.0%	99,991	5,999	6.0%	

SVM

report

	precision	recall	f1-score	support
0	0.95	0.83	0.89	18810
1	0.10	0.31	0.15	1190
accuracy			0.80	20000
macro avg	0.53	0.57	0.52	20000
weighted avg	0.90	0.80	0.84	20000

SVM_Testing Data

前20%客戶約可補捉33%再購客戶

Score	客戶數			累計			補捉率
	客戶數	再購客戶數	再購率	客戶數	再購客戶數	再購率	
1%	999	110	11.0%	999	110	11.0%	1.8%
5%	4,000	438	11.0%	4,999	548	11.0%	9.1%
10%	5,000	546	10.9%	9,999	1,094	10.9%	18.2%
20%	10,000	888	8.9%	19,999	1,982	9.9%	33.0%
30%	10,000	748	7.5%	29,999	2,730	9.1%	45.5%
40%	10,000	669	6.7%	39,999	3,399	8.5%	56.7%
50%	10,000	538	5.4%	49,999	3,937	7.9%	65.6%
60%	10,000	538	5.4%	59,999	4,475	7.5%	74.6%
70%	10,000	480	4.8%	69,999	4,955	7.1%	82.6%
80%	10,000	367	3.7%	79,999	5,322	6.7%	88.7%
90%	10,000	333	3.3%	89,999	5,655	6.3%	94.3%
100%	9,991	344	3.4%	99,990	5,999	6.0%	100.0%
全部	99,991	5,999	6.0%	99,991	5,999	6.0%	

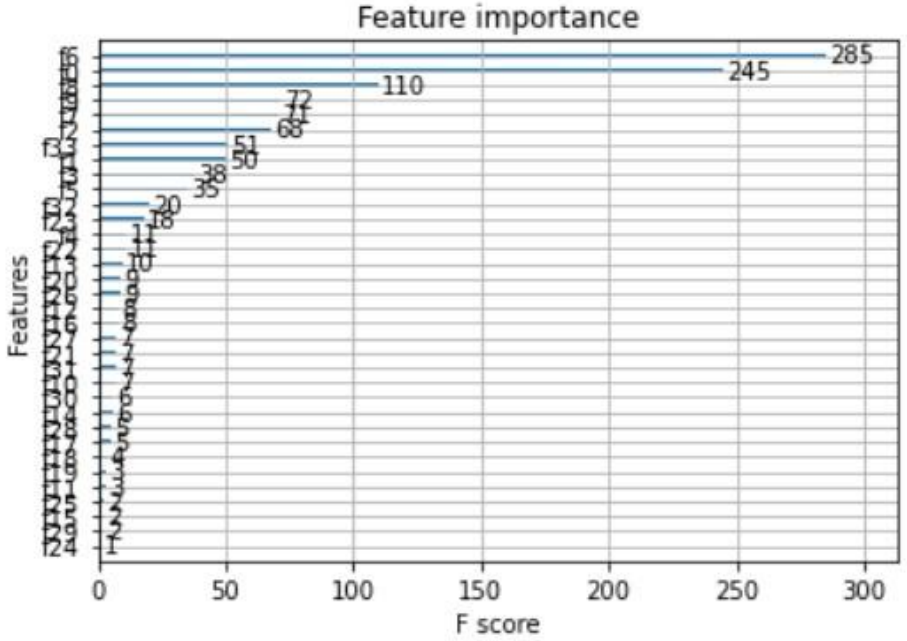
XGBoost

訓練集: 0.9819485160435041
測試集: -0.8084519944076627

- 經過smote後，模型似乎overfitting,

→重調learning rate、tree numbers也無顯著改善,
→利用Dropout、Regularization 也沒有改成功。

→改變模型:
使用Logistic Regression



特徵重要程度: [0.01499156 0.02953278 0.08169486 0.45473295 0.03501862 0.0259969
0.00954186 0.00789679 0.00980184 0.04156391 0.12062828 0.00127256
0.01001489 0.01998973 0.00393776 0.00073318 0.00656763 0.00773742
0.00086032 0.00715727 0.00269499 0.00451456 0.01492047 0.02015316
0.00073593 0.00247908 0.0040194 0.00101682 0.00083141 0.00371458
0.01204671 0.00108012 0.00798907 0.03413267]

Logistic Regression

report

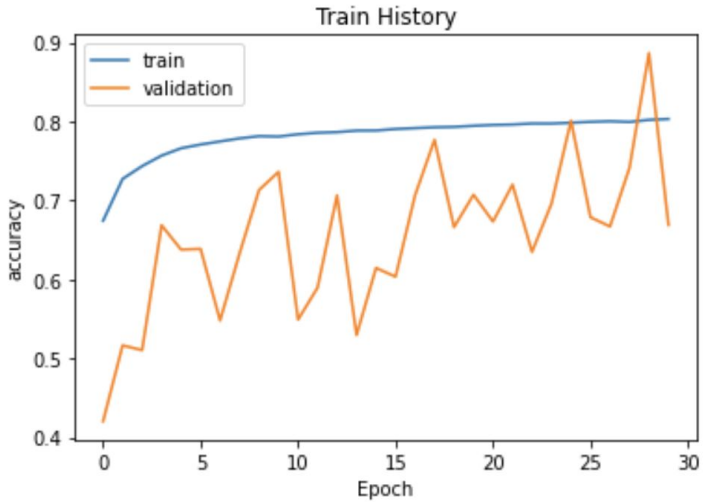
	precision	recall	f1-score	support
0	0.95	0.78	0.86	18810
1	0.10	0.39	0.16	1190
accuracy			0.75	20000
macro avg	0.53	0.58	0.51	20000
weighted avg	0.90	0.75	0.81	20000

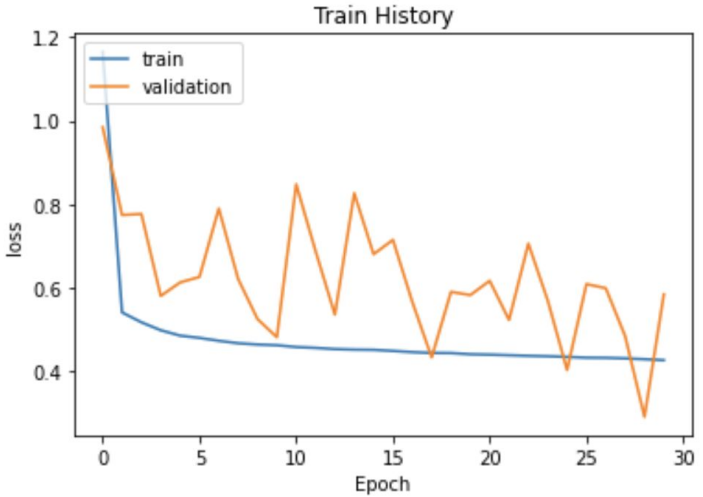
Logistic Regression_Testing Data

前20%客戶約可補捉19.8%再購客戶

Score	客戶數			累計			補捉率
	客戶數	再購客戶數	再購率	客戶數	再購客戶數	再購率	
1%	999	55	5.0%	999	55	5.5%	0.9%
5%	4,000	234	5.9%	4,999	289	5.8%	4.8%
10%	5,000	290	5.8%	9,999	579	5.8%	9.7%
20%	10,000	607	6.1%	19,999	1,186	5.9%	19.8%
30%	10,000	642	6.4%	29,999	1,828	6.1%	30.5%
40%	10,000	563	5.6%	39,999	2,391	6.0%	39.9%
50%	10,000	573	5.7%	49,999	2,964	5.9%	49.4%
60%	10,000	604	6.0%	59,999	3,568	5.9%	59.5%
70%	10,000	727	7.3%	69,999	4,295	6.1%	71.6%
80%	10,000	549	5.5%	79,999	4,844	6.1%	80.7%
90%	10,000	634	6.3%	89,999	5,478	6.1%	91.3%
100%	9,992	521	5.2%	99,991	5,999	6.0%	100.0%
全部	99,991	5,999	6.0%	99,991	5,999	6.0%	

NN






```
training set accuracy= 0.8697826266288757
testing set accuracy= 0.8605999946594238
```

NN

report

	precision	recall	f1-score	support
0	0.95	0.90	0.92	18810
1	0.13	0.24	0.17	1190
accuracy			0.86	20000
macro avg	0.54	0.57	0.55	20000
weighted avg	0.90	0.86	0.88	20000

NN_Testing Data

前20%客戶約可補捉37.5%再購客戶

Score	客戶數			累計			補捉率
	客戶數	再購客戶數	再購率	客戶數	再購客戶數	再購率	
1%	999	227	22.7%	999	227	22.7%	3.8%
5%	4,000	617	15.4%	4,999	844	16.9%	14.1%
10%	5,000	551	11.0%	9,999	1,395	14.0%	23.3%
20%	9,999	854	8.5%	19,998	2,249	11.2%	37.5%
30%	9,999	664	6.6%	29,997	2,913	9.7%	48.6%
40%	9,999	533	5.3%	39,996	3,446	8.6%	57.4%
50%	10,000	494	4.9%	49,996	3,940	7.9%	65.7%
60%	9,999	459	4.6%	59,995	4,399	7.3%	73.3%
70%	9,999	456	4.6%	69,994	4,855	6.9%	80.9%
80%	9,999	401	4.0%	79,993	5,256	6.6%	87.6%
90%	9,999	353	3.5%	89,992	5,609	6.2%	93.5%
100%	9,999	390	3.9%	99,991	5,999	6.0%	100.0%
全部	99,991	5,999	6.0%	99,991	5,999	6.0%	

預測再購客戶名單聯集

- 從四個模型的csv檔案中找出預測再購機率最高20%的客戶ID，合併到同一欄
- 用資料>移除重複項的功能，得到再購客戶名單的聯集
- 原本總計有79,996位客戶，取聯集後剩下35,207位客戶

分工表

組員	分工
陳永進	檔案合併、資料前處理、模型X值處理、負責Random Forest模型、簡報製作
鄭晴文	檔案合併、 資料前處理、模型Y值處理、處理資料不平衡、負責SVM模型、簡報製作
孟家瑜	特徵工程、特徵選取、負責NN模型
林羿帆	特徵工程、特徵選取、負責XGBoost、Logistic Regression模型

Thank you for your listening.