

2018 國泰大數據競賽



隊伍名稱：一南尬三北

參賽者 隊長：周承翰

隊員：柴宇恆

黃俊凱

范姜紹瑋

I. 資料處理與特徵選取

將訓練資料集與測試資料集（各三個檔案），先將其客戶的所有資料分別使用Pandas套件讀入Python中，並以CUST_ID為Key，將個三個資檔案Merge成兩個dataframe，並觀察兩個dataframe之資料數量與差異。

再做資料的整理，步驟如下：

1. 檢查train/test dataframe各行缺失值數量（如附件一）
2. 先將train的dataframe缺失值過多(>200000筆缺失)的行與test的行做刪除。
3. train的dataframe再將有缺失值的列亦作刪除，test的dataframe則是將有缺失值的列做平均數（mean）的fillna。
4. 最後將兩者行內元素為字元者做文數字轉換或是分類展開，使資料內容皆剩數字型態（int/uint/float）且分別具189/182行。

5. 保留所有未被刪除之資料(train約20萬筆/test 1萬筆)，最後個別輸出為csv檔作儲存（如附件二、附件三）。

II. 模型選擇與驗證成效說明

使用Keras套件的類神經網路(神經元分佈180-30-60-20-7)與上述附件二之輸出(train set)作為訓練模型，將資料切分群訓練神經元，使用matplotlib.pyplot將訓練過程圖形匯出（如附件四），並將切分群的另一份資料(未納入訓練模型)放入模型預測結果，以預測結果與原始資料檢驗此模型之準確度約為86~90%。

為驗證模型之正確，同時使用ScikitLearn套件中的Random Forest演算法(max_depth=35,n_estimators=40)，並用相同的訓練資料集建立模型，並將切分群的另一份資料(未納入訓練模型)放入模型預測結果，以預測結果與原始資料檢驗此模型之準確度約為85~89%（附件五-程式碼執行結果與模型特徵重要性）。

將準備好的test set（附件三之檔案）分別放入類神經網路與隨機森林所建立之模型中做預測，並將預測結果轉為提交之格式，分別輸出為Submit.csv檔。

將兩個submit檔做驗證，得10000筆資料中，約為8929筆資料一致，1071筆不一致，約為89.3%的一致性，以證明此兩模型對於此保單預測資料之正確性。（如附件六）

III. 附件

附件一：檢查缺失值數量之程式碼執行結果

檢查各個column缺失值數量		IS_APP	8
CUST_ID	0	IS_SPECIALMEMBER	28
BEHAVIOR_1	5041	PARENTS_DEAD	0
BEHAVIOR_2	5041	REAL_ESTATE_HAVE	0
BEHAVIOR_3	6494	IS_MAJOR_INCOME	0
STATUS1	211071	BUY_TYPE	0
STATUS2	211071	AGE	0
STATUS3	211071	SEX	0
STATUS4	211071	HEIGHT	16808
EDUCATION	0	WEIGHT	16808
IS_NEWSLETTER	160314	OCCUPATION	0
CHARGE_WAY	64095	CHILD_NUM	0
IS_EMAIL	8	BUY_MONTH	0
IS_PHONE	8	BUY_YEAR	0
INTEREST1	209654	CITY_CODE	0
INTEREST2	209654	BUDGET	0
INTEREST3	209654	MARRIAGE	0
INTEREST4	209654	BUY_TPY1_NUM_CLASS	0
INTEREST5	209654	BUY_TPY2_NUM_CLASS	0
INTEREST6	209654	BUY_TPY3_NUM_CLASS	0
INTEREST7	209654	BUY_TPY4_NUM_CLASS	0
INTEREST8	209654	BUY_TPY5_NUM_CLASS	0
INTEREST9	209654	BUY_TPY6_NUM_CLASS	0
INTEREST10	209654	BUY_TPY7_NUM_CLASS	0

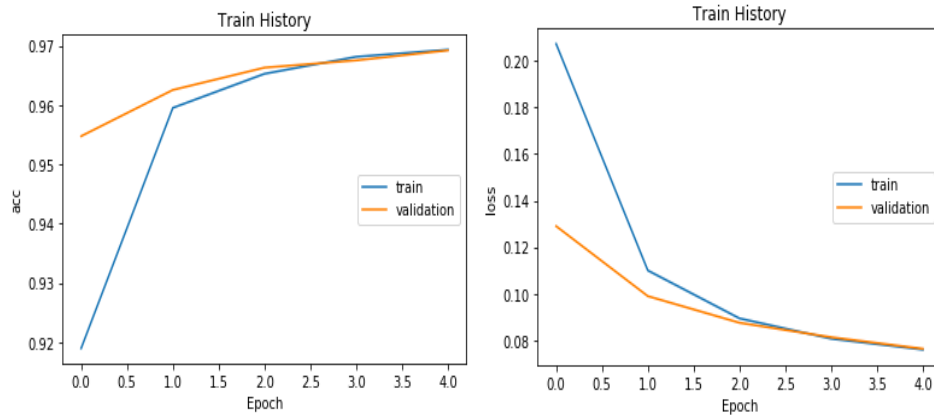
附件二：資料整理過後的train set（約剩20萬筆資料）

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	CUST_ID	BEHAVIOR	BEHAVIOR	BEHAVIOR	EDUCATN	IS_EMAIL	IS_PHONE	IS_APP	IS_SPECIA	PARENTS	REAL_EST	IS_MAJOR	SEX	HEIGHT	WEIGHT	CHILD_N	BUY_MON	BUDGET	MARRIAG	BUY_TPY	BUY_TPY	BUY_TPY	BUY_TP
2	0	1	2	3	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
3	1	31	1	2	1	4	1	1	0	0	1	1	1	0	0	0	0	11	0	5	6	5	5
4	2	49	1	2	3	2	1	0	1	1	1	1	1	0	0	0	0	5	0	5	6	6	6
5	3	55	1	2	3	2	0	1	0	0	1	0	0	0	0	0	3	5	0	5	4	5	5
6	4	62	1	2	3	3	0	1	0	1	1	1	1	0	0	0	0	4	0	0	1	6	6
7	5	84	1	2	3	3	0	1	0	1	1	1	1	0	0	0	10	0	0	1	4	6	6
8	6	99	1	2	3	1	1	0	0	1	1	1	1	0	0	0	10	0	0	1	6	6	6
9	7	107	1	2	3	2	1	0	0	1	1	1	1	0	0	0	2	0	0	1	5	6	6
10	8	112	1	2	3	2	0	1	0	0	1	1	1	0	0	0	3	10	0	0	4	5	5
11	9	115	1	2	1	4	0	1	0	1	0	1	1	1	1	1	0	12	0	0	1	6	6
12	10	116	1	2	3	1	1	0	0	1	1	1	1	0	0	0	11	0	0	1	5	6	6
13	12	140	1	2	2	3	1	1	0	0	1	1	1	0	0	0	2	0	0	5	6	5	5
14	14	153	1	2	2	3	1	1	0	0	1	1	1	0	0	0	1	0	0	1	6	5	5
15	15	160	1	2	1	3	0	1	0	1	1	1	0	0	0	0	3	12	0	0	5	4	6
16	16	161	1	1	2	4	1	1	0	0	1	1	0	0	0	0	7	0	0	1	6	5	5
17	17	168	1	2	3	4	0	1	0	0	1	1	1	0	0	0	3	0	0	5	6	4	4
18	19	204	1	2	3	3	0	1	0	1	0	1	0	0	0	0	3	0	0	1	4	6	6
19	20	218	1	2	3	1	1	0	0	0	1	1	0	0	0	0	2	0	0	5	6	6	6
20	21	245	1	2	1	4	1	1	0	0	1	1	1	1	1	1	0	3	0	0	1	4	6
21	22	247	1	2	3	3	0	1	0	0	1	1	1	1	1	1	2	11	0	0	5	6	6
22	23	266	1	2	1	4	0	1	0	0	1	1	0	0	0	0	12	0	0	1	5	6	6
23	24	269	1	2	3	2	0	1	0	0	0	1	1	0	0	0	7	0	0	5	6	6	6
24	25	289	1	2	3	4	0	1	0	0	1	1	1	0	0	0	11	0	0	5	6	6	6
25	27	308	1	2	3	1	1	1	0	0	0	0	1	0	0	0	6	0	0	1	6	5	5
26	29	314	1	2	3	2	0	1	0	0	1	1	1	0	0	0	10	0	0	5	6	5	5
27	31	327	1	2	3	2	0	1	0	1	1	0	0	0	0	0	3	10	0	5	3	5	5
28	32	334	1	2	2	1	0	1	0	0	0	1	1	1	1	1	7	0	0	1	5	5	5
29	33	357	1	2	1	1	1	0	1	1	1	1	0	0	0	0	2	-0.192511	0	5	6	6	6
30	34	360	1	2	3	2	0	1	0	0	1	1	1	1	1	1	0	6	0	0	1	4	6
31	36	373	1	2	3	1	1	1	0	0	1	0	1	0	1	1	0	9	0	0	1	6	6
32	37	392	1	2	3	4	0	1	0	1	1	1	1	0	1	1	0	10	0	0	1	5	6
33	38	410	1	2	1	1	0	1	0	0	1	1	1	1	1	1	0	0.2572349	0	1	0	4	6
34	39	428	1	2	3	4	0	1	0	0	1	1	1	0	0	0	1	-0.2315	0	0	5	6	6
35	40	445	1	2	3	1	0	1	0	1	0	1	0	1	0	0	0	3	0	0	5	6	6
36	41	451	1	2	3	3	1	1	0	0	1	1	1	1	1	1	0	5	0	0	5	4	6
37	43	492	1	2	3	3	1	1	0	0	0	0	0	1	1	1	0	10	0	0	5	6	6

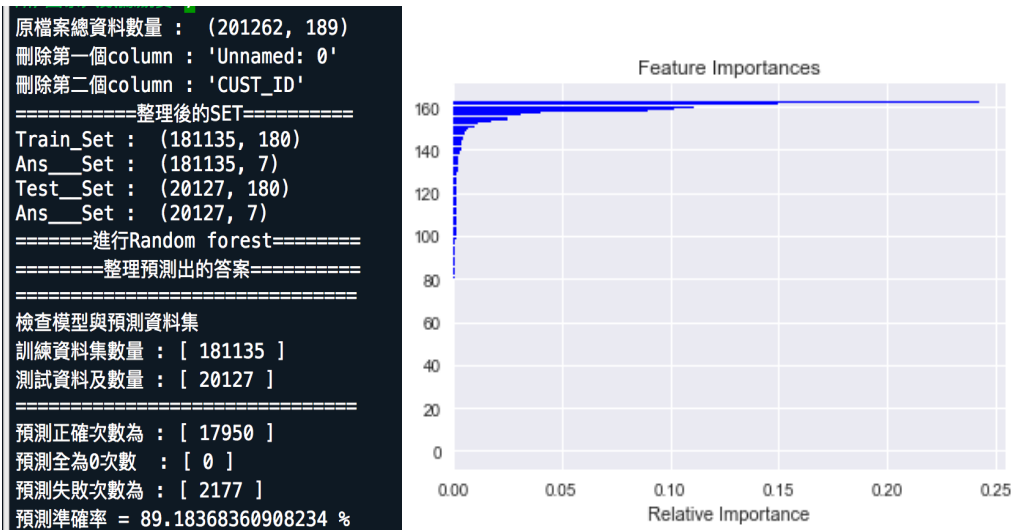
附件三：資料整理過後的test set（一萬筆資料）

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	CUST_ID	BEHAVIOR	BEHAVIOR	BEHAVIOR	EDUCATN	IS_EMAIL	IS_PHONE	IS_APP	IS_SPECIA	PARENTS	REAL_EST	IS_MAJOR	SEX	HEIGHT	WEIGHT	CHILD_N	BUY_MON	BUDGET	MARRIAG	BUY_TPY	BUY_TPY	BUY_TPY	BUY_TP
2	0	233335	1	2	3	3	1	1	0	0	1	0	1	0	0	0	4	0	0	0	6	5	5
3	1	1097571	1	2	3	2	1	0	0	0	1	1	1	1	1	0	0	6	0	0	1	6	6
4	2	2795034	1	2	3	3	1	1	0	0	0	0	1	1	1	1	0.8811298	0	10	0	0	5	4
5	3	48210	1	2	3	2	1	1	0	0	1	1	1	1	1	0	0	4	0	0	5	4	6
6	4	1582776	1	2	3	4	0	1	0	0	1	1	0	0	0	0	0	2	0	0	1	5	6
7	5	1121183	1	2	3	4	0	1	0	0	1	1	0	0	0	0	1	10	0	0	1	5	4
8	6	1510150	1	2	1	2	1	1	0	0	0	1	0	0	0	0	1	12	0	0	5	4	6
9	7	2177914	1	2	3	4	0	1	0	0	1	1	1	0	0	1	1	10	0	0	5	4	5
10	8	2423517	1	2	3	1	1	1	0	0	1	1	1	0	0	1	0.8811298	0	11	-0.23272	1	5	6
11	9	460100	1	2	3	1	1	1	0	0	1	1	1	0	0	0	0	6	0	0	5	6	6
12	10	1812989	1	2	3	1	0	1	0	1	0	0	0	1	0	0	2	3	0	0	5	6	6
13	11	2864129	1	2	2	3	1	0	1	0	1	1	1	0	1	1	0	7	-0.23824	0	5	6	6
14	12	419210	1	2	3	3	1	0	0	0	0	0	0	0	0	1	0	10	0	0	1	6	5
15	13	342686	1	2	3	1	1	1	0	0	0	1	0	1	0	1	0	11	0	0	1	6	6
16	14	2063578	1	2	3	3	0	1	1	1	1	1	1	1	1	1	0	6	0	0	5	6	6
17	15	1301099	1	2	3	1	1	1	0	0	1	1	1	1	1	1	0	5	0	0	1	6	6
18	16	1860256	1	2	3	4	0	1	0	1	1	1	1	1	1	1	0	8	0	0	5	6	5
19	17	998480	1	2	3	4	0	1	0	0	1	1	0	1	0	0	0	10	0	0	5	6	5
20	18	180333	1	2	3	1	0	1	0	1	0	1	0	1	0	0	2	5	0	0	5	6	5
21	19	430400	1	2	3	2	1	1	0	0	1	1	1	1	1	1	0	6	0	0	5	6	6
22	20	1414603	1	2	1	1	1	1	0	0	1	1	1	1	1	1	0	1	-0.220561	0	1	4	6
23	21	2701567	1	2	3	3	1	1	0	0	0	1	1	1	0	0	0	10	0	0	1	4	5
24	22	680379	1	2	3	2	1	0	0	0	0	1	1	1	1	1	0	0	0	0	5	6	6
25	23	2970723	1	2	2	2	0	1	1	1	1	1	1	1	1	1	1	1.0288013	1.0039763	2	3	0	4
26	24	2730714	1	2	2	3	1	1	0	0	0	0	1	1	1	1	0	6	0	0	1	6	5
27	25	1999163	1	2	3	3	1	1	0	1	0	1	1	1	1	1	0	9	0	0	1	4	6
28	26	2225550	1	2	1	2	1	1	0	0	0	0	1	0	1	0	3	9	0	0	5	6	5
29	27	1128520	1	2	3	1	0	1	0	0	1	1	1	1	1	1	0	5	0	0	5	6	6
30	28	201408	1	2	2	2	1	1	0	0	0	1	1	0	1	0	4	-0.21782	0	1	1	6	6
31	29	425512	1	2	3	4	0	1	0	0	1	1	1	1	1	1	0	10	0	0	5	6	6
32	30	1950099	1	2	1	1	1	1	0	0	0	1	1	1	0	0	3	2	0	0	5	6	5
33	31	1129991	1	2	3	2	1	0	0	0	1	1	1	1	1	1	0	1	0	0	1	4	6
34	32	1515646	1	2	3	1	1	1	0	0	0	0	1	1	1	1	0	6	0	0	1	6	5
35	33	818984	1	2	3	2	1	1	0	0	0	0	1	1	1	1	0	3	3	0	0	5	6
36	34	1096002	1	2	3	3	1	1	0	0	0	0	1	1	1	1	0	7	-0.23840	0	1	4	6
37	35	1129151	1	2	3	1	1	1	0	0	1	1	1	1	1	1	0	10	0	0	5	6	5

附件四：類神經網路訓練神經元



附件五：隨機森林驗證之程式碼執行結果與模型特徵重要性



附件六：以兩模型分別預測之結果核對

```
Submit_NN: (10000, 2)
Submit_RF (10000, 2)
兩個模型分別預測之相同的答案數量 : 8929
其相同的百分比為 : 89.29 %
```