Task 5
Dimension reduction

By 白宸宇

變數剔除

0	u_userId	751889 non-null	int64
1	u_phonePrice	751889 non-null	int64
2	u_browserLifeCycle	751889 non-null	int64
3	u_browserMode	751889 non-null	int64
4	u_feedLifeCycle	751889 non-null	int64
5	u_refreshTimes	751889 non-null	int64
6	u_newsCatInterests	751889 non-null	object
7	u_newsCatDislike	751889 non-null	object
8	u_newsCatInterestsST	751889 non-null	object
9	u_click_ca2_news	751889 non-null	object
10	i_docId	751889 non-null	object
11	i_s_sourceId	751889 non-null	object
12	i_regionEntity	751889 non-null	int64
13	i_cat	751889 non-null	int64
14	i_entities	729245 non-null	object
15	i_dislikeTimes	751889 non-null	int64
16	i_upTimes	751889 non-null	int64
17	i_dtype	751889 non-null	int64
18	e_ch	751889 non-null	int64
19	e_m	751889 non-null	int64
20	e_po	751889 non-null	int64
21	e_pl	751889 non-null	int64
22	e_rn	751889 non-null	int64
23	e_section	751889 non-null	int64
24	e_et	751889 non-null	int64
25	label	751889 non-null	int64
26	cillabel	751889 non-null	int64
27	pro	751889 non-null	int64

Object類別過於複雜且獨特
因此剔除

e_et為時間戳，CTGAN無法處理

降維

i_regionEntity 文章地域詞的 ID

Unique分佈	處理方式
0佔了大部分的比例(百分比) 剩下所有unique值都很少	選用將佔比低於1%的值重新分類成1

i_regionEntity	
0	0.697171
879	0.027549
1674	0.023043
2859	0.013235
3093	0.010482
744	0.009724
2018	0.007500
2388	0.007432
2658	0.007197
1445	0.005718
1192	0.005668
2934	0.005563
2684	0.005561
1117	0.005301
3139	0.005124
2158	0.004679
1177	0.004557
1728	0.004481
2780	0.004275
1841	0.004223
215	0.004087
214	0.004040
2647	0.003760
2543	0.003717
...	
2516	0.001662
1402	0.001658

i_regionEntity	
0	0.697171
1	0.228520
879	0.027549
1674	0.023043
2859	0.013235
3093	0.010482

降維

i_cat 文章類別的ID	
Unique分佈	處理方式
種類多達207種， 分佈雜亂 i_cat 98 0.093902 168 0.077267 112 0.053685 78 0.043481 219 0.040542 17 0.034875 199 0.032945 171 0.029695 65 0.027543 173 0.024514 109 0.024009 10 0.018544 157 0.017366 21 0.016078 216 0.015899 44 0.014551 108 0.014445 86 0.014347 195 0.013737 218 0.013491 57 0.012051 8 0.011434 27 0.011171 206 0.010823 ... 191 0.004410 128 0.004269 49 0.004201 20 0.004168	大致可以分為3%以上、 1-3%與1%以下 因此選用將佔比低於 1%的值重新分類成1 1-3%分類成3。 i_cat 1 0.312856 3 0.310447 98 0.093902 168 0.077267 112 0.053685 78 0.043481 219 0.040542 17 0.034875 199 0.032945

降維

i_dislikeTimes 文章負面反饋數

Unique分佈	處理方式
Unique值的分佈屬於ordinal scale	因此選用將 0-3編碼為1 (Low) 4-6編碼為2 (Mid) 7-9編碼為3 (High)

	i_dislikeTimes
0	0.467574
9	0.100035
2	0.069137
1	0.067519
4	0.064951
6	0.051006
3	0.050441
7	0.049561
8	0.043125
5	0.036652

	i_dislikeTimes
1	0.654671
3	0.192720
2	0.152609

降維

i_upTimes 文章點讚數	
Unique分佈	處理方式
同i_dislikeTimes Unique值的分佈屬於 ordinal scale i_upTimes 9 0.388137 0 0.246991 8 0.064273 7 0.055285 4 0.049801 6 0.047713 2 0.042996 1 0.040119 3 0.034948 5 0.029737	因此選用將 0-3編碼為1 (Low) 4-6編碼為2 (Mid) 7-9編碼為3 (High) i_upTimes 3 0.507695 1 0.365054 2 0.127251

降維

e_ch 頻道	
Unique分佈	處理方式
19佔了大部分的比例 剩下所有unique值都很少	選用將佔比低於1%的值重新分類成1

e_ch	
19	0.909372
11	0.034457
16	0.016415
3	0.009233
20	0.008557
2	0.005881
1	0.004711
7	0.003361
12	0.002713
13	0.001609
17	0.001087
14	0.000825
6	0.000751
15	0.000448
5	0.000165
8	0.000161
9	0.000128
4	0.000094
18	0.000032

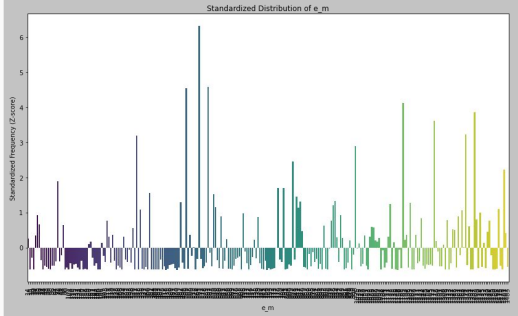
e_ch	
19	0.909372
1	0.039756
11	0.034457
16	0.016415

降維

e_m 事件來源設備機型

Unique分佈

種類很多，但大部分種類的比例皆不高



e_m	
565	0.041916
591	0.031417
506	0.031200
1185	0.028672
1403	0.027102
1277	0.025586
1377	0.023305
327	0.023083
998	0.021296
822	0.018584
1481	0.017183
73	0.015199
778	0.014073
784	0.014011
369	0.013176
619	0.013034
841	0.012579
929	0.011874
847	0.011704
487	0.011580
1205	0.011463
1123	0.011270
928	0.011067
620	0.010780
...	
1087	0.007232
320	0.007146
1347	0.006947
1348	0.006807

處理方式

大致可以分為3%以上、1-3%與1%以下

因此選用將佔比低於1%的值重新分類成1
1-3%分類成3。

e_m	
1	0.507266
3	0.388201
565	0.041916
591	0.031417
506	0.031200

降維

e_po 第幾位	
Unique分佈	處理方式
6、7的比例偏高， 10以後的的比例偏低	由於有順序關係，因此 選用將10以後的類別 轉成10 (10以上)

e_po	
6	0.263898
7	0.228836
9	0.090110
2	0.052552
5	0.046911
10	0.044763
11	0.043278
8	0.027418
4	0.026367
1	0.024595
13	0.019805
12	0.019204
3	0.017515
15	0.016613
14	0.015893
19	0.015304
16	0.014434
18	0.011649
17	0.010983
20	0.009640
21	0.000064
22	0.000056
23	0.000043
26	0.000021
25	0.000021
27	0.000017
24	0.000011

e_po	
6	0.263898
7	0.228836
10	0.221799
9	0.090110
2	0.052552
5	0.046911
8	0.027418
4	0.026367
1	0.024595
3	0.017515

降維

e_pl 拜訪地	
Unique分佈	處理方式

Unique有3089個,
且沒有順序或規律關係

```
e_pl
1728    17327
879     10309
1920     9846
2305     8989
1396     8201
1435     6195
2073     6006
2086     4428
1202     4155
1674     4124
2658     3998
140      3598
2191     3340
1374     3119
2327     3092
214      3037
2399     2905
215      2805
2999     2787
607      2737
1760     2730
2859     2656
2543     2597
3093     2556
...
494      1909
2511     1873
3022     1866
2576     1853
```

因此選用bin的方式,
依照出現的頻率
分成6個bin

```
Bin Ranges:
Bin 1: (-16.326, 2888.667]
Bin 2: (2888.667, 5776.333]
Bin 3: (5776.333, 8664.0]
Bin 4: (8664.0, 11551.667]
Bin 5: (11551.667, 14439.333]
Bin 6: (14439.333, 17327.0]
```

```
e_pl
4    0.210410
3    0.187056
5    0.169097
2    0.153215
6    0.141614
1    0.138608
```

降維

e_rn 第幾刷	
Unique分佈	處理方式
接近70%比例是1 隨後呈現急速遞減, 10之後更是掉到0.5%以下	因此選用將10以後的 類別轉成10 (10以上)

e_rn	
1	0.698808
2	0.093212
3	0.058232
4	0.038554
5	0.025278
6	0.018015
7	0.013023
8	0.009755
9	0.007525
10	0.006099
11	0.004848
12	0.003837
13	0.003153
14	0.002482
15	0.002107
16	0.001837
17	0.001585
18	0.001290
19	0.001099
20	0.000964
21	0.000789
22	0.000708
23	0.000657
24	0.000579
...	
47	0.000068
44	0.000068
50	0.000063
49	0.000061

e_rn	
1	0.698808
2	0.093212
3	0.058232
4	0.038554
10	0.037599
5	0.025278
6	0.018015
7	0.013023
8	0.009755
9	0.007525

降維

pro 文章瀏覽進度

Unique分佈

分佈集中在0、20、40、60、80、100,
其中又以0佔大多數

pro	
0	704278
100	27920
80	5809
60	4641
40	4607
20	4592
95	4
63	3
76	3
53	2
54	2
10	2
90	2
98	2
91	2
2	1
24	1
18	1
9	1
37	1
26	1
93	1
99	1
88	1
...	
96	1
32	1
84	1
34	1

處理方式

因此選用將較細節的數字併入較大的類別中
0 → 0
1-30 → 20
31-50 → 40
51-70 → 60
71-90→ 80
90-100→ 100

pro	
0	704278
100	27931
80	5818
60	4650
40	4611
20	4601

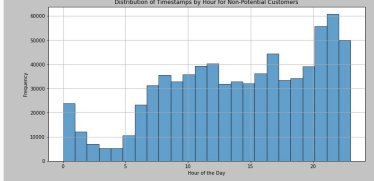
pro	
0	0.936678
100	0.037148
80	0.007738
60	0.006184
40	0.006133
20	0.006119

降維

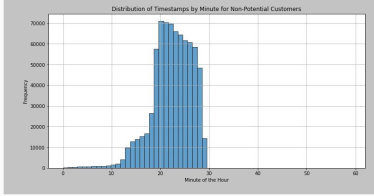
e_et 時間戳	
Unique分佈	處理方式

參照week 3

By Hour



By Minute



將e_et變數依日期及小時分成
e_et_day、e_et_hour,
以便copula與CTGan使用

分鐘因為分佈不均及會導致維度過高
，故不採用

小時近一步分成時段 e_et_timeslot
0-6→1
7-12→2
13-18→3
19-24→4

e_et_day的數字意義為2022/6/XX號

e_et_day	
4	0.160710
5	0.160509
7	0.142048
3	0.140128
8	0.137648
9	0.133799
6	0.125158

e_et_timeslot	
4	0.362963
3	0.289007
2	0.263022
1	0.085009

Run time comparison

```
sample_df = non_potential_customers_3.sample(frac=0.01, random_state=202401801)
```

```
synthesizer = CopulaGANSynthesizer(metadata)
synthesizer.fit(sample_df)

synthetic_data = synthesizer.sample(num_rows=10)
```

✓ 7m 56.2s

```
from sdv.single_table import TVAESynthesizer

synthesizer = TVAESynthesizer(
    metadata, # required
    enforce_min_max_values=True,
    enforce_rounding=False,
    epochs=500
)
synthesizer.fit(sample_df)

synthetic_data = synthesizer.sample(num_rows=10)
```

✓ 3m 6.8s

```
from sdv.single_table import CTGANSynthesizer

synthesizer = CTGANSynthesizer(metadata)
synthesizer.fit(sample_df)

synthetic_data = synthesizer.sample(num_rows=10)
```

✓ 10m 44.4s

CopulaGAN

```
synthetic_data = synthesizer.sample(
    num_rows=1_000_000,
    batch_size=10_000
)
```

✓ 1m 7.8s

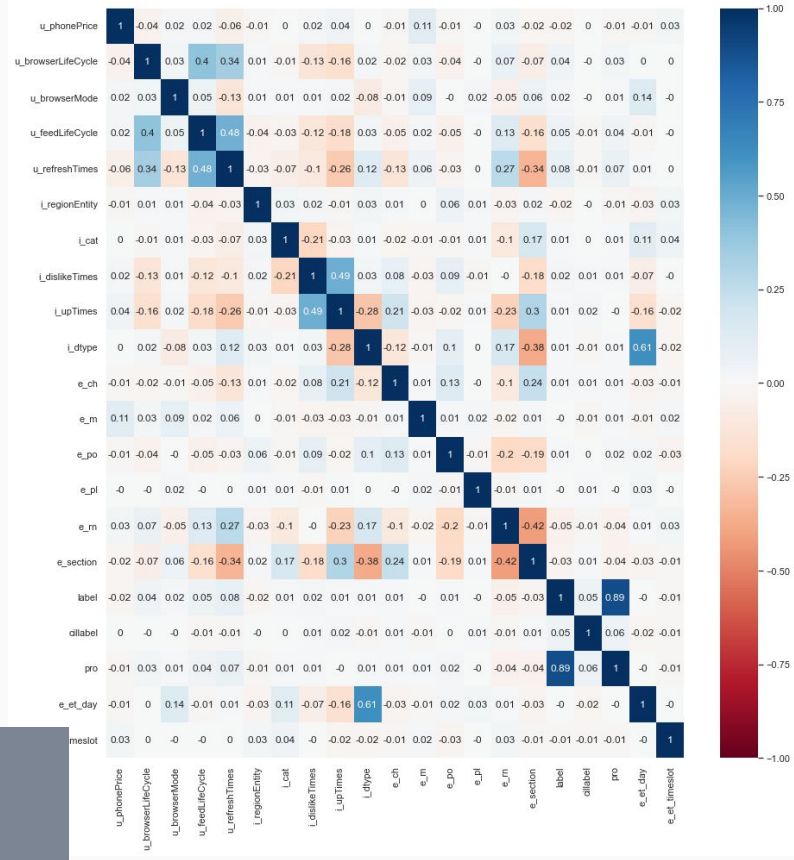
```
Generating report ...

(1/2) Evaluating Data Validity: |██████████| 21/21 [00:00<00:00, 140.06it/s]|
Data Validity Score: 100.0%

(2/2) Evaluating Data Structure: |██████████| 1/1 [00:00<00:00, 754.78it/s]|
Data Structure Score: 100.0%

Overall Score (Average): 100.0%
```

```
Generating report ...

(1/2) Evaluating Column Shapes: |██████████| 21/21 [00:01<00:00, 11.89it/s]|
Column Shapes Score: 90.54%

(2/2) Evaluating Column Pair Trends: |██████████| 210/210 [01:22<00:00,  2.55it/s]|
Column Pair Trends Score: 76.24%

Overall Score (Average): 83.39%
```

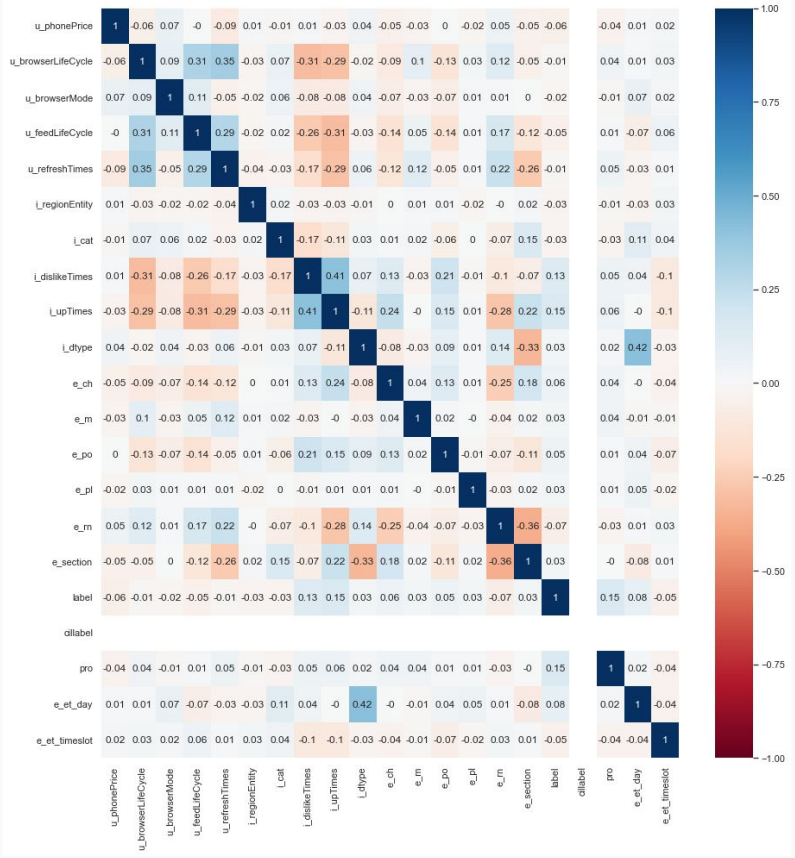
	Column	Metric	Score
0	u_phonePrice	TVComplement	0.819082
1	u_browserLifeCycle	TVComplement	0.922293
2	u_browserMode	TVComplement	0.900006
3	u_feedLifeCycle	TVComplement	0.897912
4	u_refreshTimes	TVComplement	0.888621
5	i_regionEntity	TVComplement	0.830654
6	i_cat	TVComplement	0.827273
7	i_dislikeTimes	TVComplement	0.905793
8	i_upTimes	TVComplement	0.907412
9	i_dtype	TVComplement	0.913285
10	e_ch	TVComplement	0.978110
11	e_m	TVComplement	0.969353
12	e_po	TVComplement	0.892465
13	e_pl	TVComplement	0.772054
14	e_rn	TVComplement	0.965963
15	e_section	TVComplement	0.979750
16	label	KSComplement	0.977896
17	cillabel	KSComplement	0.999601
18	pro	TVComplement	0.935840
19	e_et_day	TVComplement	0.872843
20	e_et_timeslot	TVComplement	0.856670

CopulaGAN

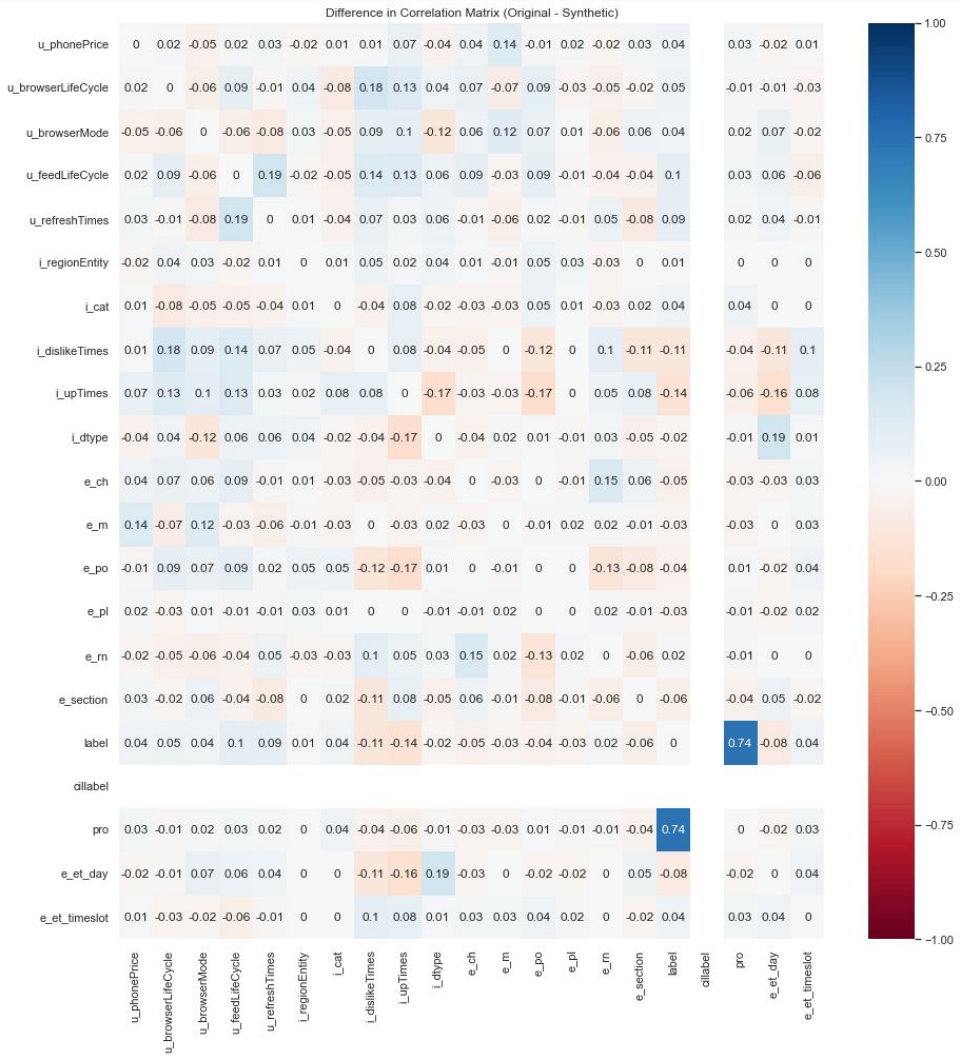
原始資料



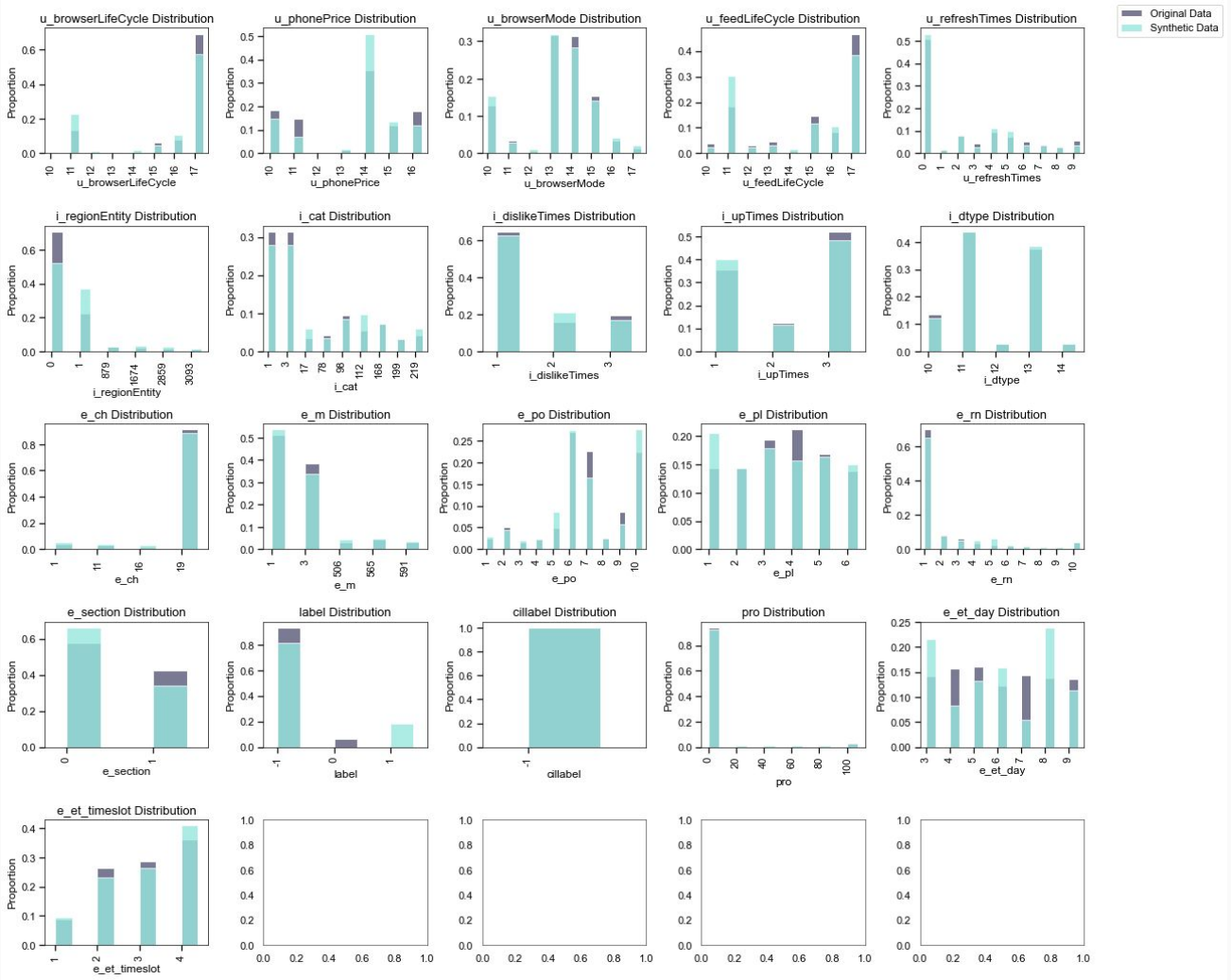
合成資料



相關係數差異熱點圖


Difference in Correlation Matrix (Original - Synthetic)
u_phonePrice 0 0.02 -0.05 0.02 0.03 -0.02 0.01 0.01 0.07 -0.04 0.04 0.14 -0.01 0.02 -0.02 0.03 0.04 0.03 -0.02 0.01
u_browserLifeCycle 0.02 0 -0.06 0.09 -0.01 0.04 -0.08 0.18 0.13 0.04 0.07 -0.07 0.09 -0.03 -0.05 -0.02 0.05 -0.01 -0.01 -0.03
u_browserMode -0.05 -0.06 0 -0.06 -0.08 0.03 -0.05 0.09 0.1 -0.12 0.06 0.12 0.07 0.01 -0.06 0.06 0.04 0.02 0.07 -0.02
u_feedLifeCycle 0.02 0.09 -0.06 0 0.19 -0.02 -0.05 0.14 0.13 0.06 0.09 -0.03 0.09 -0.01 -0.04 -0.04 0.1 0.03 0.06 -0.06
u_refreshTimes 0.03 -0.01 -0.08 0.19 0 0.01 -0.04 0.07 0.03 0.06 -0.01 -0.06 0.02 -0.01 0.05 -0.08 0.09 0.02 0.04 -0.01
i_regionEntity -0.02 0.04 0.03 -0.02 0.01 0 0.01 0.05 0.02 0.04 0.01 -0.01 0.05 0.03 -0.03 0 0.01 0 0 0
i_cat 0.01 -0.08 -0.05 -0.05 -0.04 0.01 0 -0.04 0.08 -0.02 -0.03 -0.03 0.05 0.01 -0.03 0.02 0.04 0.04 0 0
i_dislikeTimes 0.01 0.18 0.09 0.14 0.07 0.05 -0.04 0 0.08 -0.04 -0.05 0 -0.12 0 0.1 -0.11 -0.11 -0.04 -0.11 0.1
i_upTimes 0.07 0.13 0.1 0.13 0.03 0.02 0.08 0.08 0 -0.17 -0.03 -0.03 -0.17 0 0.05 0.08 -0.14 -0.06 -0.16 0.08
i_dtype -0.04 0.04 -0.12 0.06 0.06 0.04 -0.02 -0.04 -0.17 0 -0.04 0.02 0.01 -0.01 0.03 -0.05 -0.02 -0.01 0.19 0.01
e_ch 0.04 0.07 0.06 0.09 -0.01 0.01 -0.03 -0.05 -0.03 -0.04 0 -0.03 0 -0.01 0.15 0.06 -0.05 -0.03 -0.03 0.03
e_m 0.14 -0.07 0.12 -0.03 -0.06 -0.01 -0.03 0 -0.03 0.02 -0.03 0 -0.01 0.02 0.02 -0.01 -0.03 -0.03 0 0.03
e_po -0.01 0.09 0.07 0.09 0.02 0.05 0.05 -0.12 -0.17 0.01 0 -0.01 0 0 -0.13 -0.08 -0.04 0.01 -0.02 0.04
e_pl 0.02 -0.03 0.01 -0.01 -0.01 0.03 0.01 0 0 -0.01 -0.01 0.02 0 0 0.02 -0.01 -0.03 -0.01 -0.02 0.02
e_rn -0.02 -0.05 -0.06 -0.04 0.05 -0.03 -0.03 0.1 0.05 0.03 0.15 0.02 -0.13 0.02 0 -0.06 0.02 -0.01 0 0
e_section 0.03 -0.02 0.06 -0.04 -0.08 0 0.02 -0.11 0.08 -0.05 0.06 -0.01 -0.08 -0.01 -0.06 0 -0.06 -0.04 0.05 -0.02
label 0.04 0.05 0.04 0.1 0.09 0.01 0.04 -0.11 -0.14 -0.02 -0.05 -0.03 -0.04 -0.03 0.02 -0.06 0 0.74 -0.08 0.04
olllabel
pro 0.03 -0.01 0.02 0.03 0.02 0 0.04 -0.04 -0.06 -0.01 -0.03 -0.03 0.01 -0.01 -0.01 -0.04 0.74 0 -0.02 0.03
e_et_day -0.02 -0.01 0.07 0.06 0.04 0 0 -0.11 -0.16 0.19 -0.03 0 -0.02 -0.02 0 0.05 -0.08 -0.02 0 0.04
e_et_timeslot 0.01 -0.03 -0.02 -0.06 -0.01 0 0 0.1 0.08 0.01 0.03 0.03 0.04 0.02 0 -0.02 0.04 0.03 0.04 0
u_phonePrice u_browserLifeCycle u_browserMode u_feedLifeCycle u_refreshTimes i_regionEntity i_cat i_dislikeTimes i_upTimes i_dtype e_ch e_m e_po e_pl e_rn e_section label olllabel pro e_et_day e_et_timeslot
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

原始與合成資料分佈比較



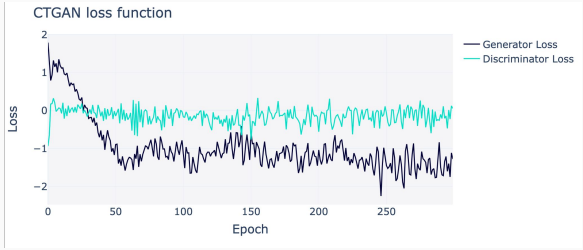
CTGAN

```
synthetic_data = synthesizer.sample(
    num_rows=1_000_000,
    batch_size=10_000
)
```

✓ 1m 33.5s

```
Generating report ...

(1/2) Evaluating Data Validity: |██████████| 21/21 [00:00<00:00, 115.07it/s]|
Data Validity Score: 100.0%

(2/2) Evaluating Data Structure: |██████████| 1/1 [00:00<00:00, 251.38it/s]|
Data Structure Score: 100.0%

Overall Score (Average): 100.0%
```

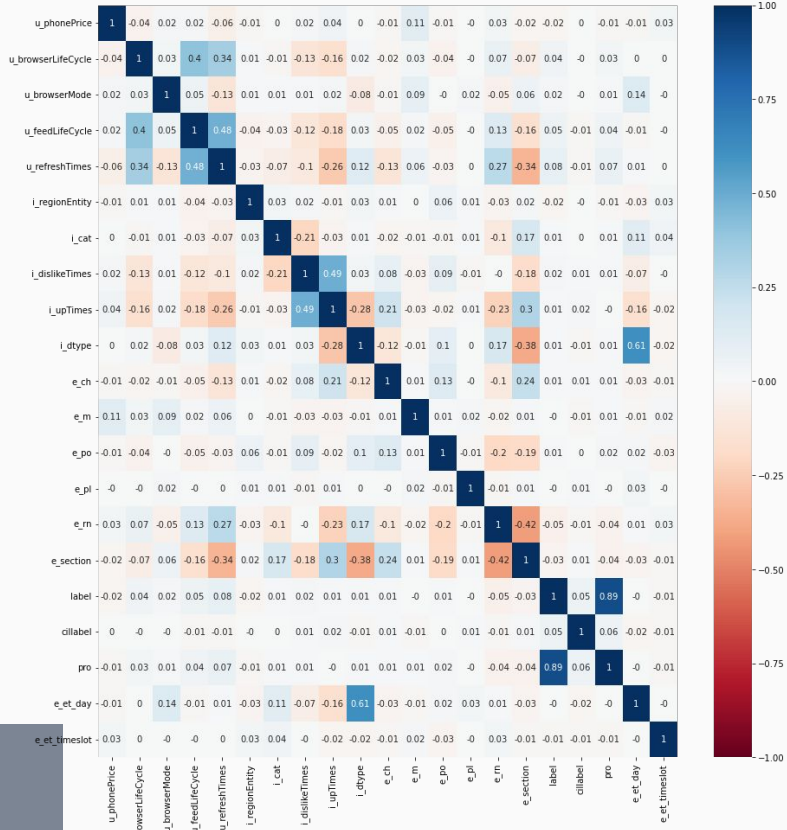
```
Generating report ...

(1/2) Evaluating Column Shapes: |██████████| 21/21 [00:01<00:00, 12.14it/s]|
Column Shapes Score: 90.57%

(2/2) Evaluating Column Pair Trends: |██████████| 210/210 [01:26<00:00,  2.44it/s]|
Column Pair Trends Score: 76.67%

Overall Score (Average): 83.62%
```



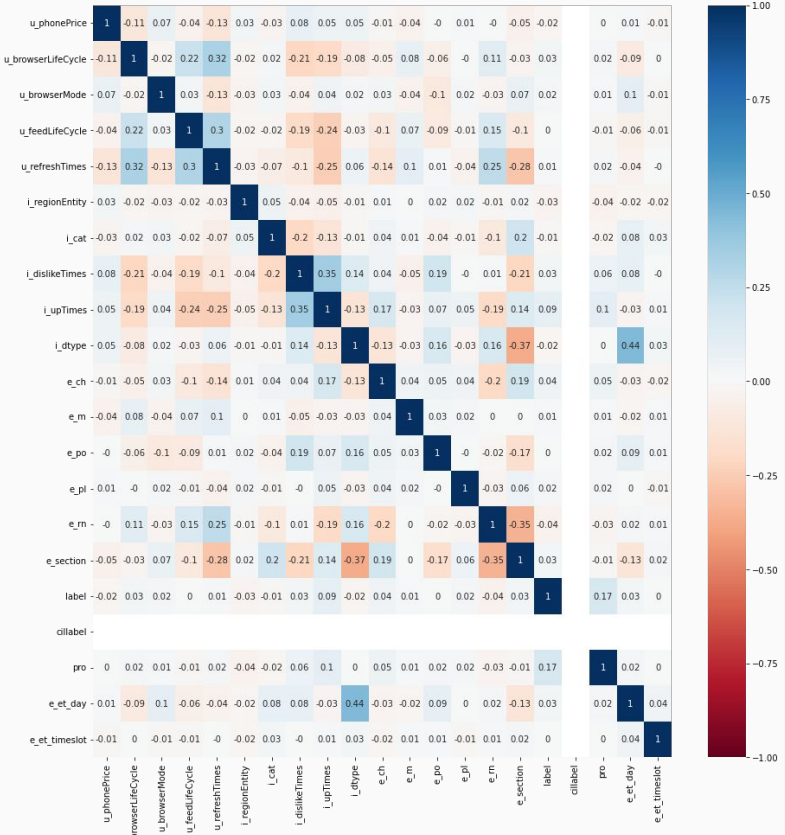
	Column	Metric	Score
0	u_phonePrice	TVComplement	0.877514
1	u_browserLifeCycle	TVComplement	0.936233
2	u_browserMode	TVComplement	0.916469
3	u_feedLifeCycle	TVComplement	0.896563
4	u_refreshTimes	TVComplement	0.895003
5	i_regionEntity	TVComplement	0.932775
6	i_cat	TVComplement	0.868479
7	i_dislikeTimes	TVComplement	0.931143
8	i_upTimes	TVComplement	0.873122
9	i_dtype	TVComplement	0.950579
10	e_ch	TVComplement	0.957059
11	e_m	TVComplement	0.904771
12	e_po	TVComplement	0.830528
13	e_pl	TVComplement	0.809742
14	e_rn	TVComplement	0.908893
15	e_section	TVComplement	0.904689
16	label	KSComplement	0.998817
17	cillabel	KSComplement	0.999601
18	pro	TVComplement	0.974483
19	e_et_day	TVComplement	0.771026
20	e_et_timeslot	TVComplement	0.882441

CTGAN

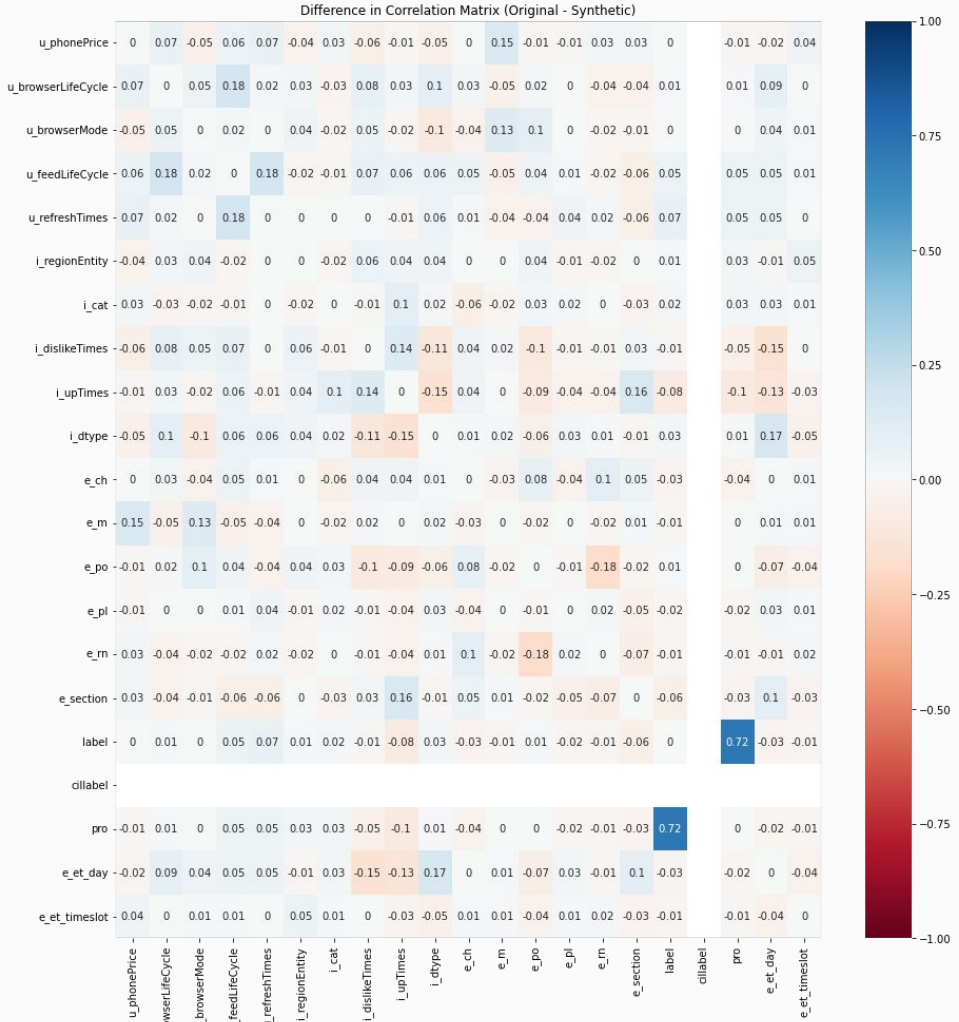
原始資料



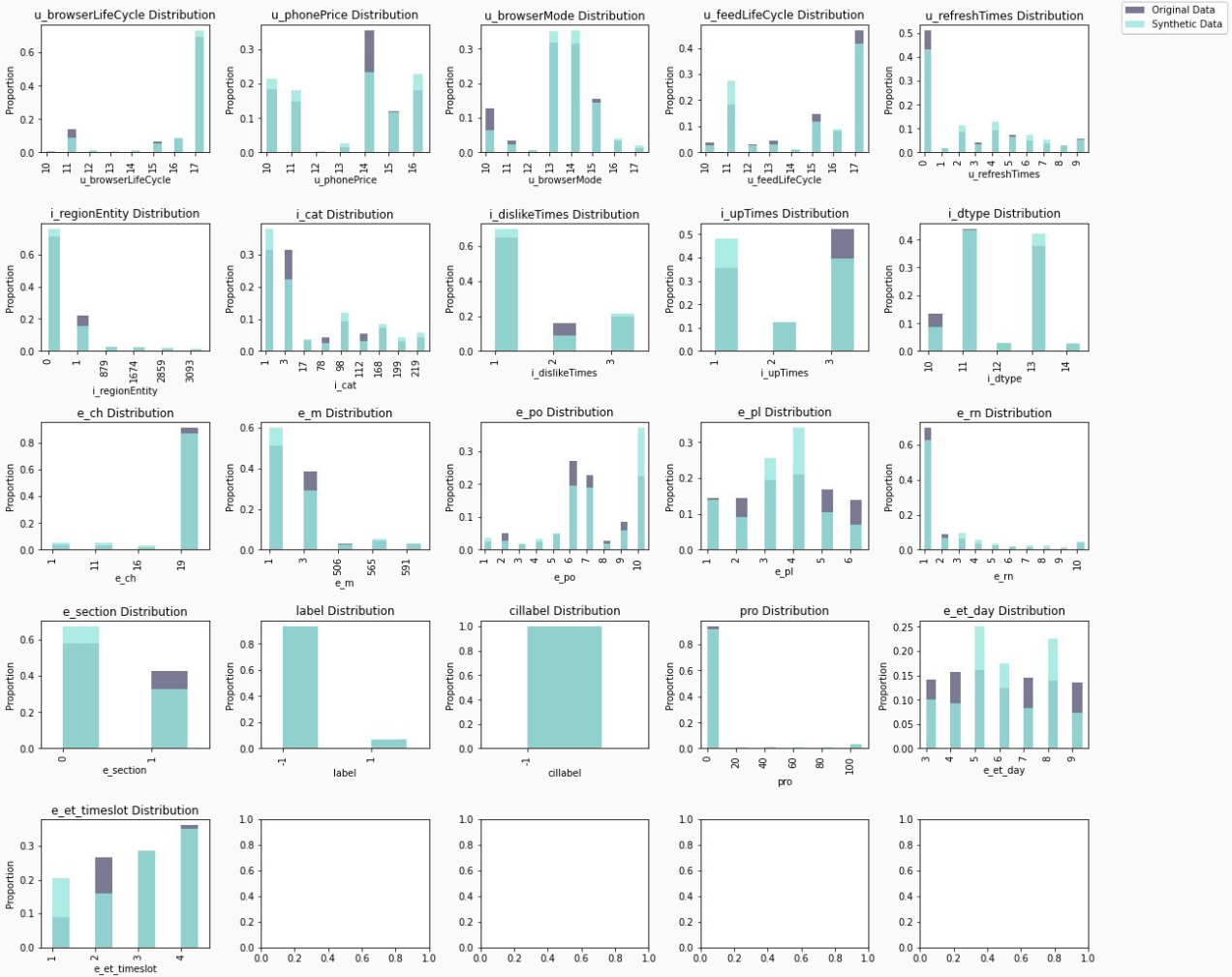
合成資料



相關係數差異熱點圖


Difference in Correlation Matrix (Original - Synthetic)
u_phonePrice
u_browserLifeCycle
u_browserMode
u_feedLifeCycle
u_refreshTimes
i_regionEntity
i_cat
i_dislikeTimes
i_upTimes
i_dtype
e_ch
e_m
e_po
e_pl
e_rn
e_section
label
cillabel
pro
e_et_day
e_et_timeslot
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

原始與合成資料分佈比較



TVAE

```
synthetic_data = synthesizer.sample(
    num_rows=1_000_000,
    batch_size=10_000
)
```

✓ 48.9s

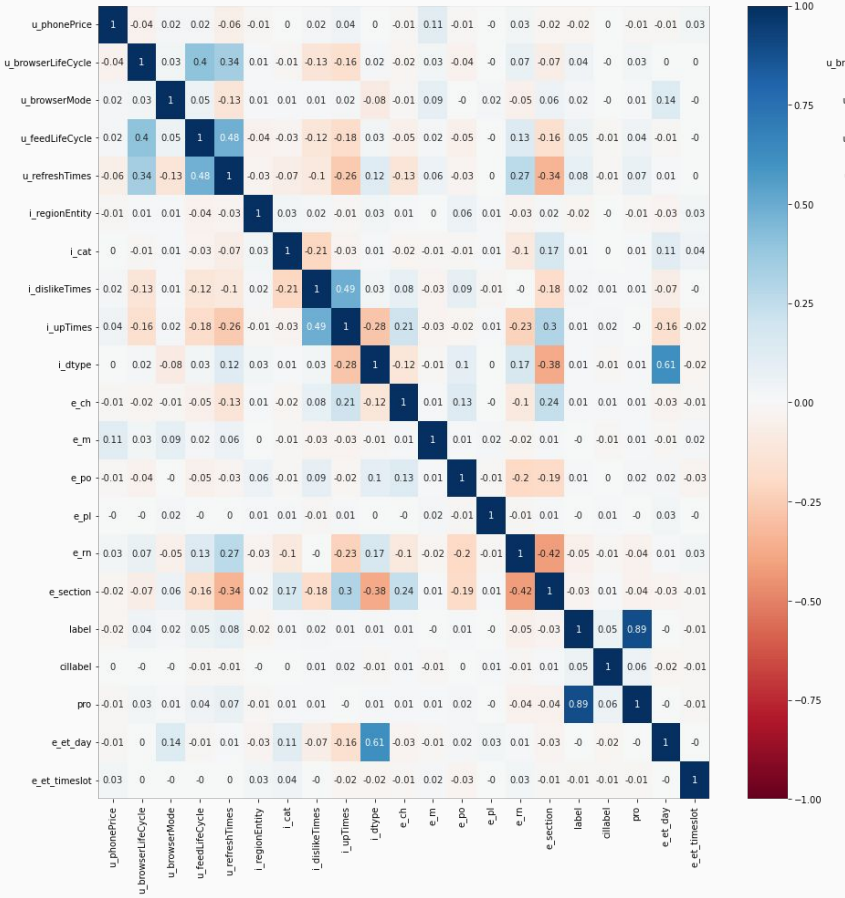
```
Generating report ...

(1/2) Evaluating Data Validity: |          | 21/21 [00:00<00:00, 163.01it/s]|
Data Validity Score: 100.0%

(2/2) Evaluating Data Structure: |          | 1/1 [00:00<00:00, 725.78it/s]|
Data Structure Score: 100.0%

Overall Score (Average): 100.0%
```

```
Generating report ...

(1/2) Evaluating Column Shapes: |          | 21/21 [00:01<00:00, 12.53it/s]|
Column Shapes Score: 86.96%

(2/2) Evaluating Column Pair Trends: |          | 210/210 [01:22<00:00,  2.53it/s]|
Column Pair Trends Score: 69.59%

Overall Score (Average): 78.28%
```

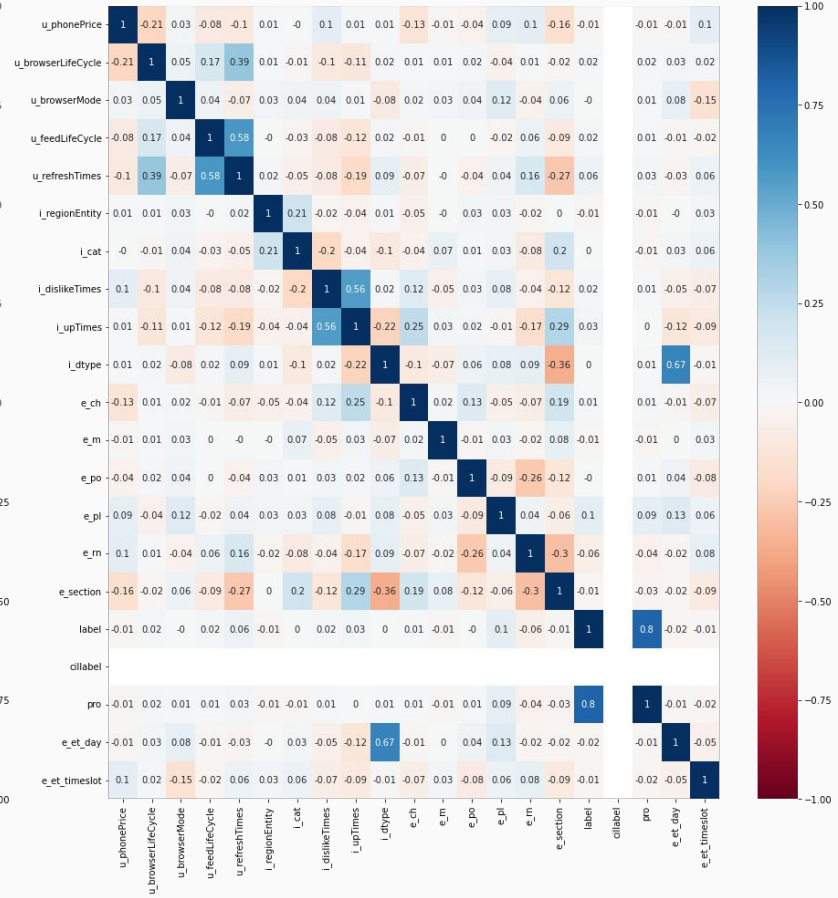
	Column	Metric	Score
0	u_phonePrice	TVComplement	0.800330
1	u_browserLifeCycle	TVComplement	0.960472
2	u_browserMode	TVComplement	0.907558
3	u_feedLifeCycle	TVComplement	0.910495
4	u_refreshTimes	TVComplement	0.843157
5	i_regionEntity	TVComplement	0.941102
6	i_cat	TVComplement	0.892115
7	i_dislikeTimes	TVComplement	0.963560
8	i_upTimes	TVComplement	0.965840
9	i_dtype	TVComplement	0.960160
10	e_ch	TVComplement	0.960357
11	e_m	TVComplement	0.788545
12	e_po	TVComplement	0.741827
13	e_pl	TVComplement	0.592093
14	e_rn	TVComplement	0.845202
15	e_section	TVComplement	0.974164
16	label	KSComplement	0.996419
17	cillabel	KSComplement	0.999601
18	pro	TVComplement	0.977179
19	e_et_day	TVComplement	0.705024
20	e_et_timeslot	TVComplement	0.537401

TVAE

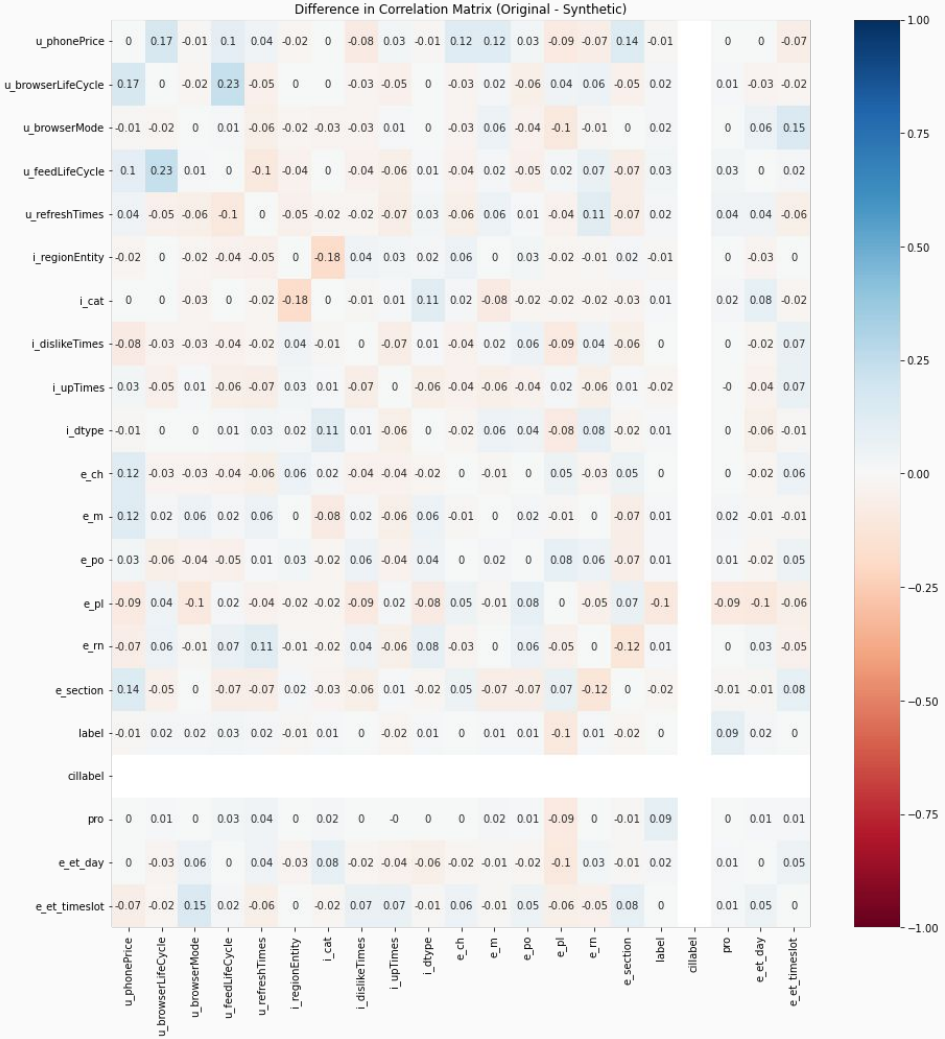
原始資料



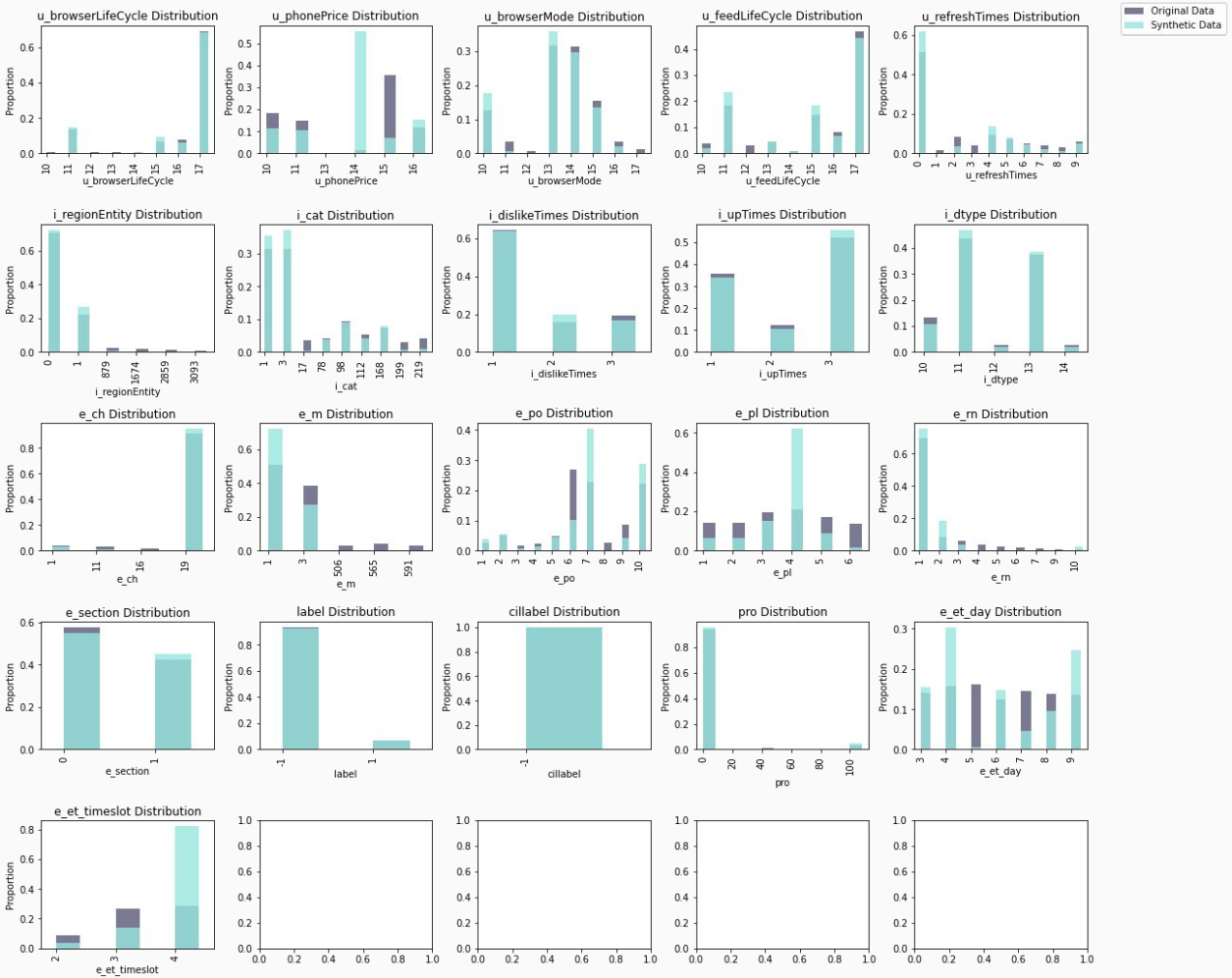
合成資料



相關係數差異熱點圖


Difference in Correlation Matrix (Original - Synthetic)
u_phonePrice 0 0.17 -0.01 0.1 0.04 -0.02 0 -0.08 0.03 -0.01 0.12 0.12 0.03 -0.09 -0.07 0.14 -0.01 0 0 -0.07
u_browserLifeCycle 0.17 0 -0.02 0.23 -0.05 0 0 -0.03 -0.05 0 -0.03 0.02 -0.06 0.04 0.06 -0.05 0.02 0.01 -0.03 -0.02
u_browserMode -0.01 -0.02 0 0.01 -0.06 -0.02 -0.03 -0.03 0.01 0 -0.03 0.06 -0.04 -0.1 -0.01 0 0.02 0 0.06 0.15
u_feedLifeCycle 0.1 0.23 0.01 0 -0.1 -0.04 0 -0.04 -0.06 0.01 -0.04 0.02 -0.05 0.02 0.07 -0.07 0.03 0.03 0 0.02
u_refreshTimes 0.04 -0.05 -0.06 -0.1 0 -0.05 -0.02 -0.02 -0.07 0.03 -0.06 0.06 0.01 -0.04 0.11 -0.07 0.02 0.04 0.04 -0.06
i_regionEntity -0.02 0 -0.02 -0.04 -0.05 0 -0.18 0.04 0.03 0.02 0.06 0 0.03 -0.02 -0.01 0.02 -0.01 0 -0.03 0
i_cat 0 0 -0.03 0 -0.02 -0.18 0 -0.01 0.01 0.11 0.02 -0.08 -0.02 -0.02 -0.02 -0.03 0.01 0.02 0.08 -0.02
i_dislikeTimes -0.08 -0.03 -0.03 -0.04 -0.02 0.04 -0.01 0 -0.07 0.01 -0.04 0.02 0.06 -0.09 0.04 -0.06 0 0 -0.02 0.07
i_upTimes 0.03 -0.05 0.01 -0.06 -0.07 0.03 0.01 -0.07 0 -0.06 -0.04 -0.06 -0.04 0.02 -0.06 0.01 -0.02 -0 -0.04 0.07
i_dtype -0.01 0 0 0.01 0.03 0.02 0.11 0.01 -0.06 0 -0.02 0.06 0.04 -0.08 0.08 -0.02 0.01 0 -0.06 -0.01
e_ch 0.12 -0.03 -0.03 -0.04 -0.06 0.06 0.02 -0.04 -0.04 -0.02 0 -0.01 0 0.05 -0.03 0.05 0 0 -0.02 0.06
e_m 0.12 0.02 0.06 0.02 0.06 0 -0.08 0.02 -0.06 0.06 -0.01 0 0.02 -0.01 0 -0.07 0.01 0.02 -0.01 -0.01
e_po 0.03 -0.06 -0.04 -0.05 0.01 0.03 -0.02 0.06 -0.04 0.04 0 0.02 0 0.08 0.06 -0.07 0.01 0.01 -0.02 0.05
e_pl -0.09 0.04 -0.1 0.02 -0.04 -0.02 -0.02 -0.09 0.02 -0.08 0.05 -0.01 0.08 0 -0.05 0.07 -0.1 -0.09 -0.1 -0.06
e_rn -0.07 0.06 -0.01 0.07 0.11 -0.01 -0.02 0.04 -0.06 0.08 -0.03 0 0.06 -0.05 0 -0.12 0.01 0 0.03 -0.05
e_section 0.14 -0.05 0 -0.07 -0.07 0.02 -0.03 -0.06 0.01 -0.02 0.05 -0.07 -0.07 0.07 -0.12 0 -0.02 -0.01 -0.01 0.08
label -0.01 0.02 0.02 0.03 0.02 -0.01 0.01 0 -0.02 0.01 0 0.01 0.01 -0.1 0.01 -0.02 0 0.09 0.02 0
cillabel
pro 0 0.01 0 0.03 0.04 0 0.02 0 -0 0 0 0.02 0.01 -0.09 0 -0.01 0.09 0 0.01 0.01
e_et_day 0 -0.03 0.06 0 0.04 -0.03 0.08 -0.02 -0.04 -0.06 -0.02 -0.01 -0.02 -0.1 0.03 -0.01 0.02 0.01 0 0.05
e_et_timeslot -0.07 -0.02 0.15 0.02 -0.06 0 -0.02 0.07 0.07 -0.01 0.06 -0.01 0.05 -0.06 -0.05 0.08 0 0.01 0.05 0
u_phonePrice u_browserLifeCycle u_browserMode u_feedLifeCycle u_refreshTimes i_regionEntity i_cat i_dislikeTimes i_upTimes i_dtype e_ch e_m e_po e_pl e_rn e_section label cillabel pro e_et_day e_et_timeslot
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

原始與合成資料分佈比較



ADS.csv

降維

residence 居住地	
Unique分佈	處理方式
共35種，佔比到後面變低	因此選用將佔比低於3%的值編類成1(其他)

residence	
33	0.134142
20	0.118027
16	0.074987
21	0.054394
18	0.050050
30	0.043205
17	0.042630
41	0.040565
46	0.040234
26	0.038632
45	0.033841
13	0.033240
27	0.032639
29	0.031112
19	0.028424
32	0.023546
34	0.023184
25	0.021257
11	0.018619
24	0.017635
39	0.014271
23	0.012576
42	0.010496
14	0.010376
...	
36	0.002310
40	0.000788
43	0.000113
35	0.000080

residence	
1	0.232302
33	0.134142
20	0.118027
16	0.074987
21	0.054394
18	0.050050
30	0.043205
17	0.042630
41	0.040565
46	0.040234
26	0.038632
45	0.033841
13	0.033240
27	0.032639
29	0.031112

降維

city 居住城市 ID	
Unique分佈	處理方式
種類多達341種， 分佈雜亂	大致可以分為3%以上、 1-3%與1%以下 因此選用將佔比低於 1%的值重新分類成1 1-3%分類成3。

```
city
319    0.137387
113    0.043205
162    0.032639
328    0.024811
372    0.024288
354    0.018112
431    0.017635
135    0.016086
170    0.015369
297    0.014887
429    0.014271
310    0.014020
179    0.013749
343    0.012828
220    0.012570
207    0.011922
424    0.010732
191    0.010220
282    0.010100
212    0.009840
371    0.009612
380    0.009245
120    0.007087
199    0.006969
...
203    0.004068
342    0.004019
283    0.004011
262    0.003948
```

```
city
1      0.545169
3      0.241600
319    0.137387
113    0.043205
162    0.032639
```

降維

series_dev 設備系列	
Unique分佈	處理方式
大部分集中在其中幾種 series_dev 16 0.19664708 30 0.16609083 27 0.11873454 31 0.11294001 34 0.10284740 11 0.09493419 32 0.06850066 17 0.03506669 21 0.02975252 23 0.02079521 19 0.01519741 24 0.00932823 15 0.00514076 26 0.00458406 20 0.00433104 12 0.00360184 33 0.00329072 28 0.00325867 14 0.00291616 36 0.00188157 29 0.00007009 13 0.00004013 37 0.00002254 25 0.00001238 18 0.00001003 35 0.00000508 22 0.00000013	因此選用將佔比低於3%的值編類成1(其他) series_dev 16 0.19664708 30 0.16609083 27 0.11873454 31 0.11294001 1 0.10423858 34 0.10284740 11 0.09493419 32 0.06850066 17 0.03506669

降維

emui_dev emui 版本號碼	
Unique分佈	處理方式
大部分集中在其中幾種	因此選用將佔比低於3%的值編類成1(其他)

emui_dev	
20	0.178605
21	0.163529
11	0.147316
35	0.085591
28	0.083945
13	0.078010
19	0.050669
29	0.044397
30	0.032693
32	0.032087
23	0.024011
36	0.019959
16	0.013326
12	0.011847
18	0.007753
31	0.007116
17	0.006504
14	0.004551
37	0.004430
33	0.002570
25	0.000748
34	0.000247
22	0.000044
27	0.000032
24	0.000017
15	0.000007
26	0.000000

emui_dev	
20	0.178605
21	0.163529
11	0.147316
1	0.103160
35	0.085591
28	0.083945
13	0.078010
19	0.050669
29	0.044397
30	0.032693
32	0.032087

降維

device_name 使用者手機機型	
Unique分佈	處理方式
沒有特別高的類別	大致可以分為3%以上、1-3%與1%以下 因此選用將佔比低於 1%的值重新分類成1(低頻率) 1-3%分類成2(中頻率) 3%以上分類成3(高頻率)

```
device_name
346   0.038468
278   0.037629
319   0.032407
151   0.030309
210   0.027572
109   0.024295
123   0.022768
272   0.022099
229   0.020565
334   0.017779
351   0.017182
344   0.016720
248   0.016643
299   0.014758
137   0.014227
228   0.014168
168   0.013490
310   0.013299
202   0.013079
127   0.012922
170   0.012562
164   0.012479
183   0.012458
240   0.012243
...
292   0.007338
139   0.006749
117   0.006664
141   0.006634
```

```
device_name
1   0.486698
2   0.374489
3   0.138813
```

降維

device_size 使用者手機尺寸

Unique分佈	處理方式
大部分集中在少數類別	因此選用將佔比低於3%的值編類成1(其他)

device_size	
2117	0.189847
2032	0.091451
2401	0.085506
1186	0.047918
2541	0.030736
1656	0.028077
1555	0.024482
2438	0.024261
1505	0.021293
2231	0.019120
2397	0.017948
1815	0.016866
1199	0.016407
2177	0.016297
2482	0.012065
2103	0.011821
2353	0.011487
1821	0.011062
2469	0.010786
1171	0.009812
2485	0.009805
1413	0.009237
1966	0.009169
1269	0.008860
...	
1524	0.002975
2383	0.002923
2579	0.002757
1135	0.002701

device_size	
1	0.554542
2117	0.189847
2032	0.091451
2401	0.085506
1186	0.047918
2541	0.030736

降維

task_id　廣告任務ID

Unique分佈	處理方式
大部分集中在少數類別	因此選用依次數做分桶 10000以下為1 10001-20000為2 20001-50000為3 50001-100000為4 100001以上為5

```
task_id
22100    154812
14584    126367
34382    122094
34975     92381
31941     77470
31996     69691
34848     68921
14404     67865
10670     65328
10699     49888
27573     47745
10653     46242
20692     46090
35039     45008
33083     44617
29211     44374
17020     40879
30157     39634
10769     39408
21493     39063
22097     37265
32082     35061
21567     35051
34041     34803
...
28290     20596
21104     20533
24106     20373
35422     20182
```

```
task_id
1    4304423
3    1273220
2    1252945
4     441656
5     403273
```

降維

adv_id 廣告任務對應的素材 **ID**

Unique分佈	處理方式
大部分集中在少數類別	因此選用依次數做分桶 10000以下為1 10001-20000為2 20001-50000為3 50001-100000為4 100001以上為5

adv_id	
18060	154812
17683	126367
11752	122094
15769	92381
21695	77470
10724	69691
20207	68921
20934	67865
11794	65328
22477	49888
10957	47745
13957	46242
19451	46090
16335	45008
18274	44617
12555	44374
17973	40879
10974	39634
23247	39408
15213	39063
21323	37265
21961	35061
10585	35051
13608	34803
...	
23039	20373
11123	19363
22701	19242
17413	19107

adv_id	
1	4474937
3	1205078
2	1150573
4	441656
5	403273

降維

adv_prim_id 廣告任務對應的廣告主 ID

Unique分佈	處理方式
大部分集中在少數類別	因此選用依次數做分桶 10000以下為1 10001-20000為2 20001-50000為3 50001-100000為4 100001以上為5

adv_prim_id	
1036	684617
1236	241924
2066	237504
1220	233549
1486	205201
1005	137486
1913	132252
1363	123505
1187	121133
1314	116444
1041	102221
1562	100364
1518	91187
1097	90625
1535	90249
1611	83190
1669	82610
1403	78831
1619	74789
1482	73797
1106	72461
1690	72222
1223	71628
1731	67865
...	
1244	34570
1637	34426
1542	33770
1114	33227

adv_prim_id	
5	2436200
3	1846431
4	1468934
2	1036819
1	887133

降維

slot_id 廣告位ID	
Unique分佈	處理方式
大部分集中在少數類別	因此選用依次數做分桶 100000以下為1 100001-300000為2 300001-500000為3 500001-1000000為4 1000001以上為5

slot_id	
16	1794442
17	555059
54	438012
38	388710
50	384803
59	356804
22	348515
13	286164
35	283984
12	244190
30	231938
28	224140
26	162163
21	149335
34	140681
40	136728
14	120833
69	110775
53	105232
67	102812
47	91774
63	84757
31	76628
65	73122
...	
70	115
39	57
48	51
51	9

slot_id	
2	2298975
3	1916844
5	1794442
1	1110197
4	555059

降維

spread_app_id 投放廣告對應的 App ID

Unique分佈	處理方式
大部分集中在少數類別	因此選用依次數做分桶 100000以下為1 100001-300000為2 300001-500000為3 500001-1000000為4

spread_app_id	
213	897391
162	880552
312	709396
190	492803
152	368226
344	364280
114	343475
246	280992
309	237625
280	233549
240	230462
257	214807
168	214629
301	161329
350	159333
317	139638
372	122008
306	97809
321	87724
174	82487
298	75174
181	73365
250	65966
199	64471
...	
206	9521
243	7707
256	7451
343	7092

spread_app_id	
4	2487339
2	1994372
1	1625022
3	1568784

降維

hispace_app_tags 廣告任務對應的 App標籤

Unique分佈	處理方式
大部分集中在少數類別	因此選用將佔比低於5%的值編類成1(其他)

```
hispace_app_tags
47   0.250307
43   0.179963
49   0.142641
18   0.117104
20   0.094315
39   0.048251
12   0.032159
26   0.022676
19   0.021043
27   0.012743
46   0.011741
16   0.010747
48   0.007217
41   0.006372
14   0.006137
30   0.005996
51   0.004597
15   0.004467
31   0.004109
44   0.003855
23   0.003036
50   0.001867
33   0.001277
37   0.001038
...
22   0.000026
36   0.000020
13   0.000016
11   0.000013
```

```
hispace_app_tags
47   0.250307
1    0.215671
43   0.179963
49   0.142641
18   0.117104
20   0.094315
```

降維

app_second_class 廣告任務對應的 App 的二級分類

Unique分佈	處理方式
大部分集中在少數類別	因此選用將佔比低於5%的值編類成1(其他)

app_second_class	
14	0.230159
18	0.209799
23	0.169997
17	0.149588
13	0.092149
15	0.057069
29	0.038156
20	0.020505
28	0.007386
25	0.006326
16	0.004740
22	0.004461
27	0.003867
21	0.002486
19	0.001261
30	0.000635
11	0.000529
24	0.000404
12	0.000253
26	0.000228

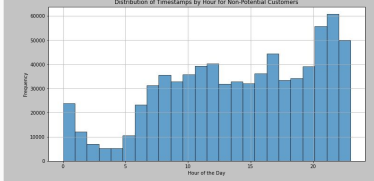
app_second_class	
14	0.230159
18	0.209799
23	0.169997
17	0.149588
13	0.092149
1	0.091238
15	0.057069

降維

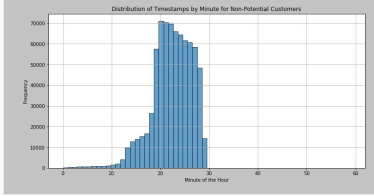
pt_d 時間戳	
Unique分佈	處理方式

參照week 3

By Hour



By Minute



將e_et變數依日期及小時分成
e_et_day、e_et_hour,
以便copula與CTGan使用

分鐘因為分佈不均及會導致維度過高
，故不採用

小時近一步分成時段 e_et_timeslot
0-6→1
7-12→2
13-18→3
19-24→4

e_et_day的數字意義為2022/6/XX號

e_et_day	
4	0.160710
5	0.160509
7	0.142048
3	0.140128
8	0.137648
9	0.133799
6	0.125158

e_et_timeslot	
4	0.362963
3	0.289007
2	0.263022
1	0.085009

Thanks for watching