機器與深度學習概論

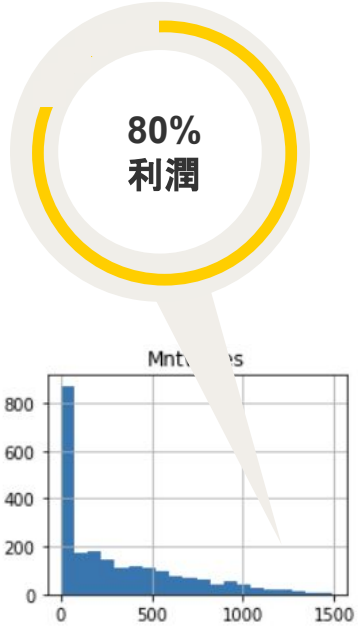
- 預測購買酒水的高消費族群

何子安 E44065020 ｜ 盧欣誼 P37091019 ｜ 王宜安 D54071179

Problem Statement

Motive

- 行銷理論中的「**二八定律**」
- 顧客在酒精上的花費分布圖，確實有極少數的人提供著高消費金額。
- **實際計算後**發現整筆資料的**前35%酒水消費金額佔了整體酒水消費金額的80%**。
- 想更進一步區分**高消費**與**低消費族群**，有利往後針對不同消費族群提供相對應的廣告投放及行銷策略。



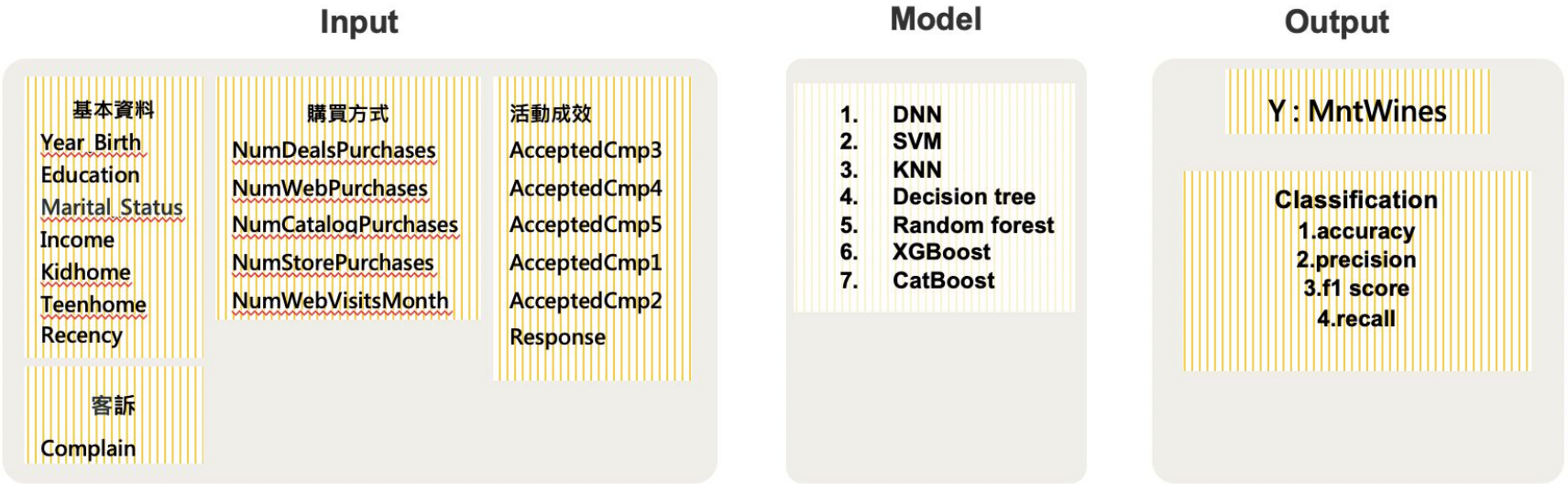
Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

Expectation

- 猜測在酒水的開銷上可能較高的人可能跟一些特徵有關

Supervised Learning

- Classification:
  購買酒水的族群，分成二類(高消費族群、低消費族群)



資料集介紹

基本資料

- ID
- Year_Birth
- Education
- Marital_Status
- Income
- Kidhome
- Teenhome
- Dt_Customer
- Recency

購買方式

- NumDealsPurchases
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

客訴

- Complain

各類產品花費

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds

活動成效

- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- AcceptedCmp1
- AcceptedCmp2
- Response

1

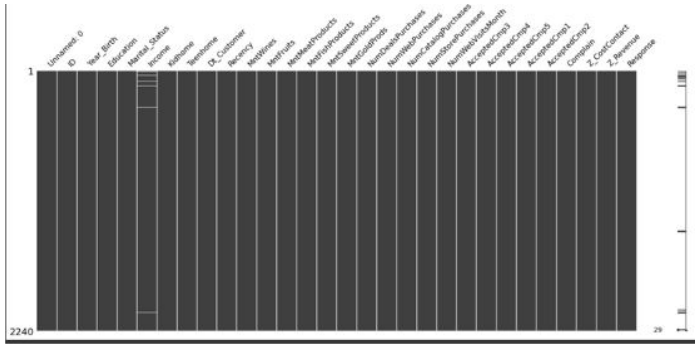
Data Preprocess

缺失值處理 | 新增向度

Data Preprocess

缺失值處理

- 只有 `'income'` 有缺失值
- 以平均值/KNN填補



新增向度

- 從 `'Year_Birth'` 計算 `'Age'`
- 從 `'Dt_Customer'` 計算`'Customer_Years`

	ID	Year_Birth	Dt_Customer	Age	Customer_Years
0	5524	1957	04-09-2012	65	10
1	2174	1954	08-03-2014	68	8
2	4141	1965	21-08-2013	57	9
3	6182	1984	10-02-2014	38	8
4	5324	1981	19-01-2014	41	8

2

Exploratory Data Analysis

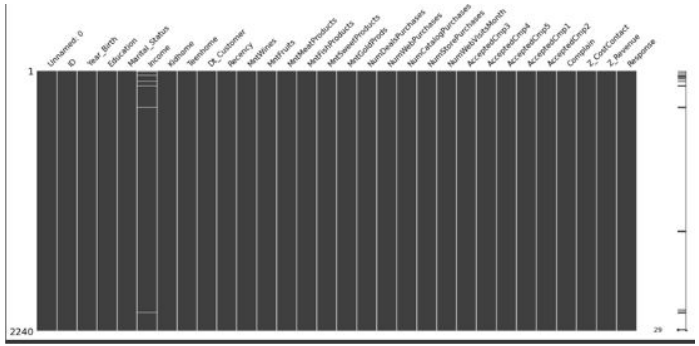
Numerical Variable | histogram |

Correlation Heatmap | factors

Data Preprocess

缺失值處理

- 只有 `'income'` 有缺失值
- 以平均值填補



新增向度

- 從 `'Year_Birth'` 計算 `'Age'`
- 從 `'Dt_Customer'` 計算`'Customer_Years`

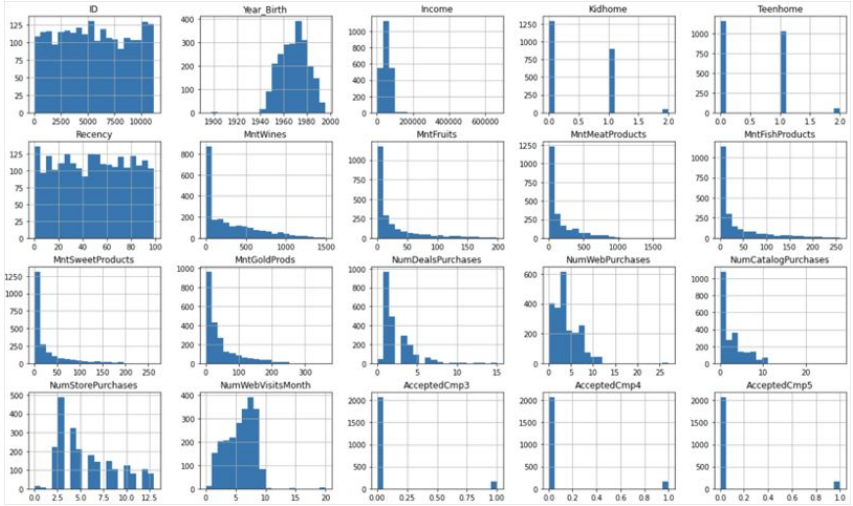
	ID	Year_Birth	Dt_Customer	Age	Customer_Years
0	5524	1957	04-09-2012	65	10
1	2174	1954	08-03-2014	68	8
2	4141	1965	21-08-2013	57	9
3	6182	1984	10-02-2014	38	8
4	5324	1981	19-01-2014	41	8

Data Preprocess

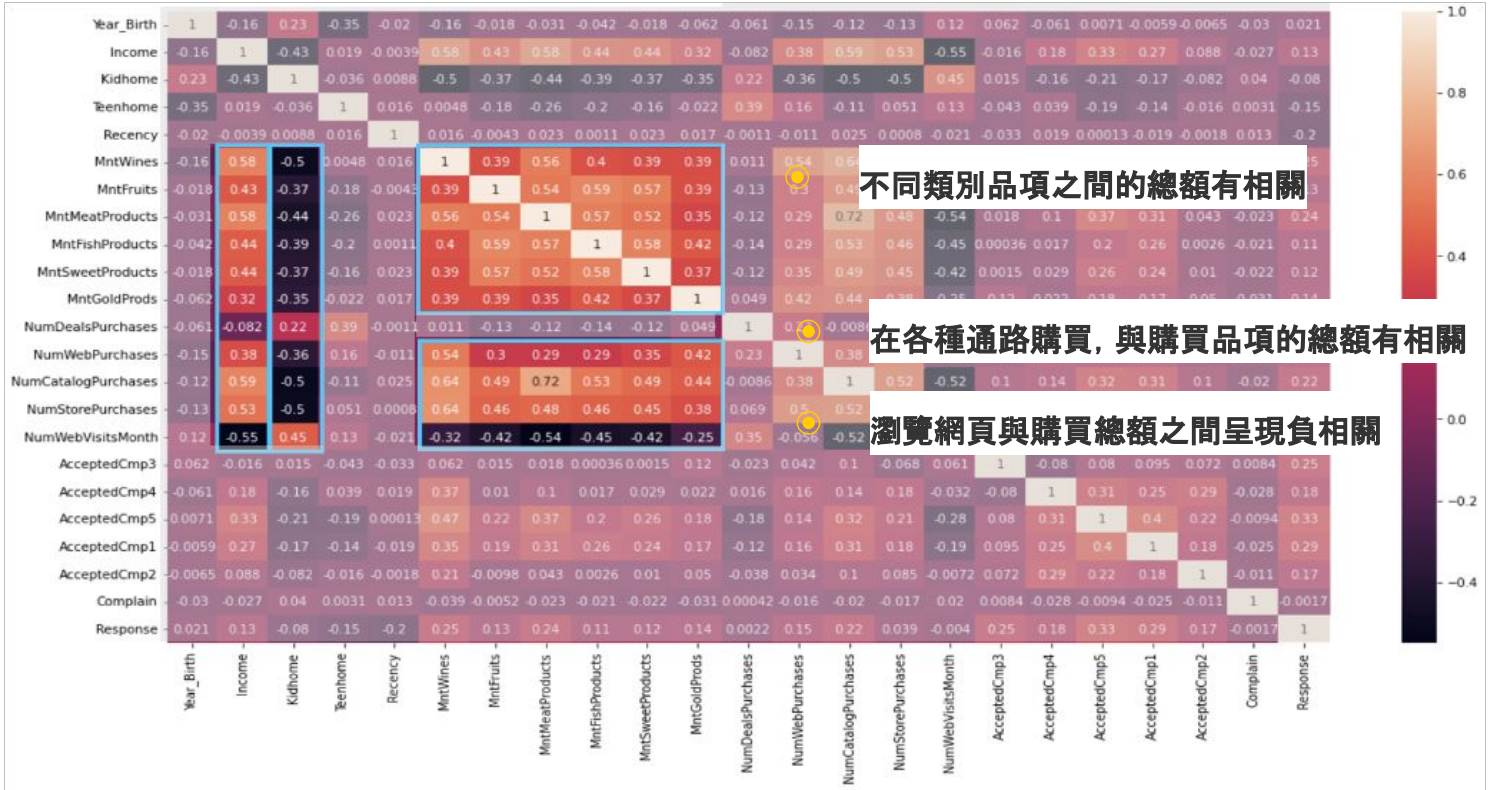
Numerical Variable

	ID	Year_Birth	Income
count	2240.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.251354
std	3246.662198	11.984069	25037.797168
min	0.000000	1893.000000	1730.000000
25%	2828.250000	1959.000000	35538.750000
50%	5458.500000	1970.000000	51741.500000
75%	8427.750000	1977.000000	68289.750000
max	11191.000000	1996.000000	666666.000000

histogram



Correlation Heatmap



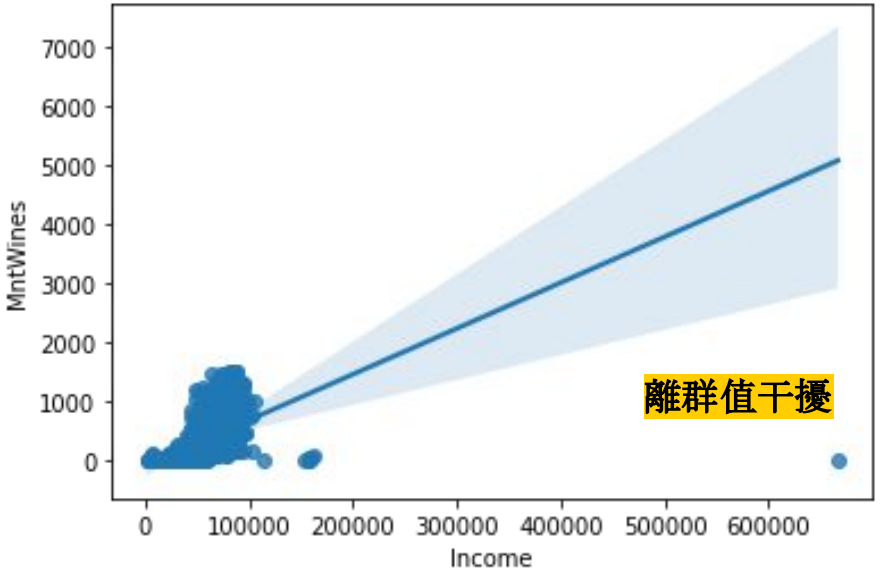
不同類別品項之間的總額有相關

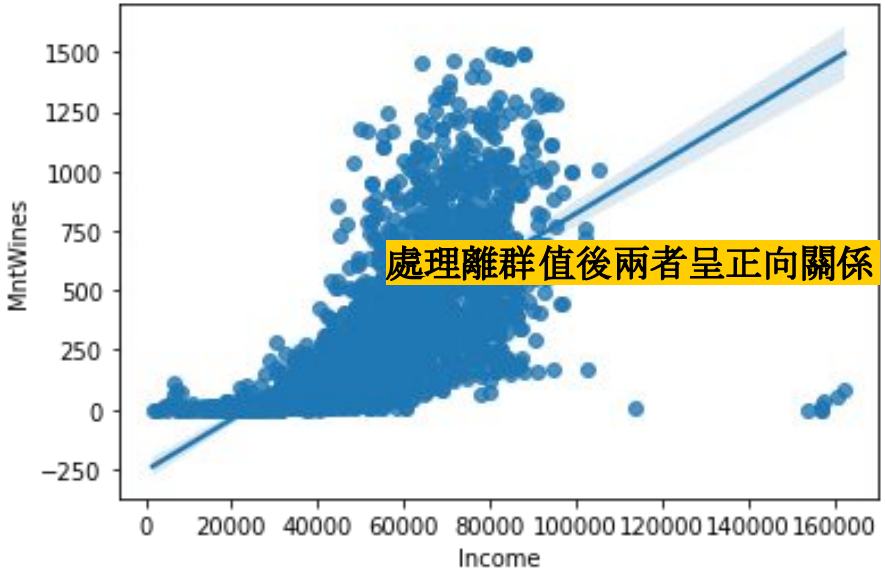
在各種通路購買，與購買品項的總額有相關

瀏覽網頁與購買總額之間呈現負相關

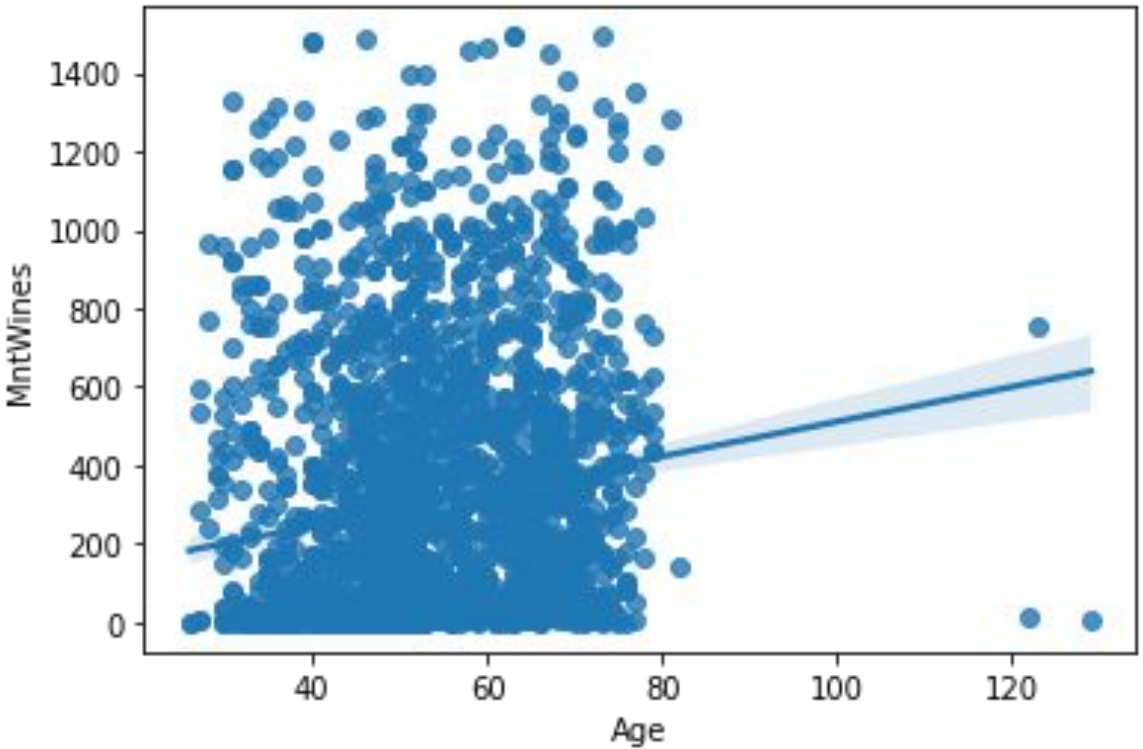
- **較高收入的人：**
  - 購賣不同品項↑
  - 瀏覽網站↓
  - 購買行為↑
- **有兒童的家庭：**
  - 購買不同品項↓
  - 購買行為↓
  - 瀏覽網站行為↑

可能影響酒水花費的因素-收入

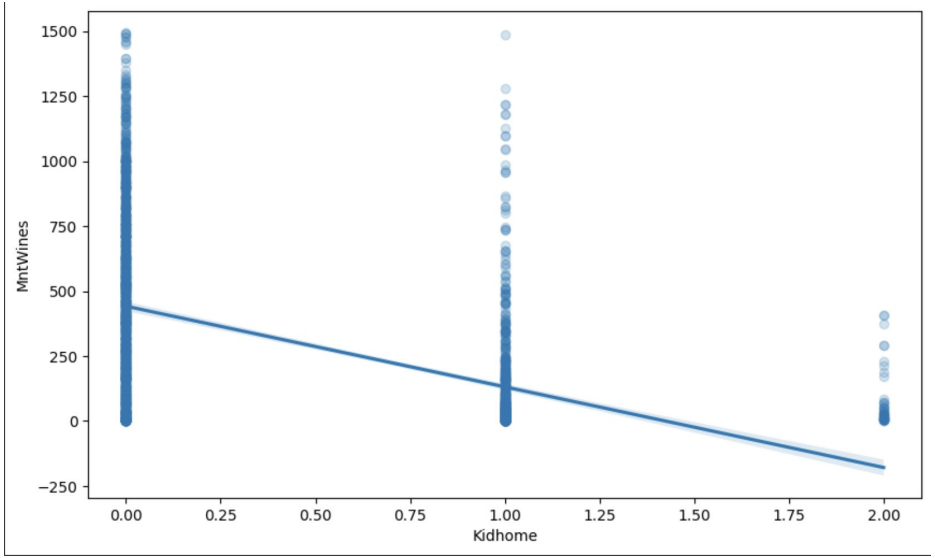


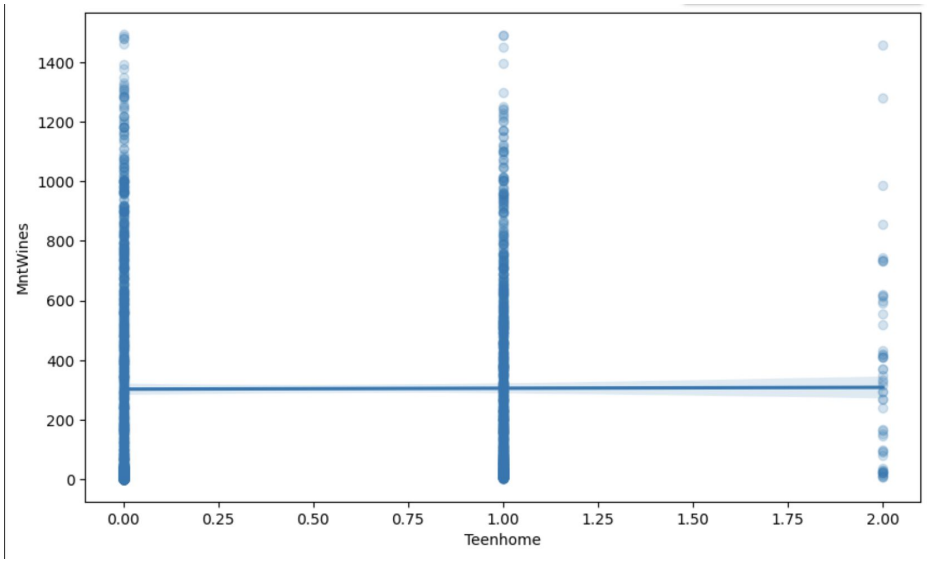


可能影響酒水花費的因素- 年齡



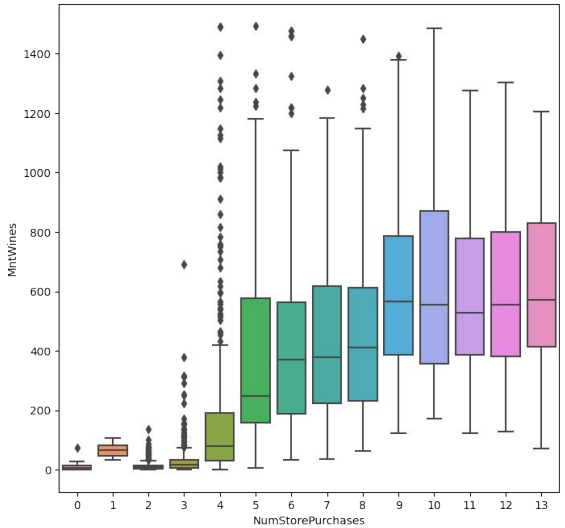
可能影響酒水花費的因素- 家庭成員

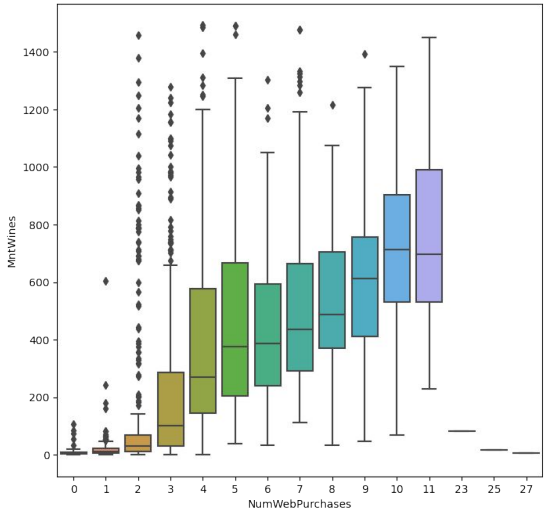


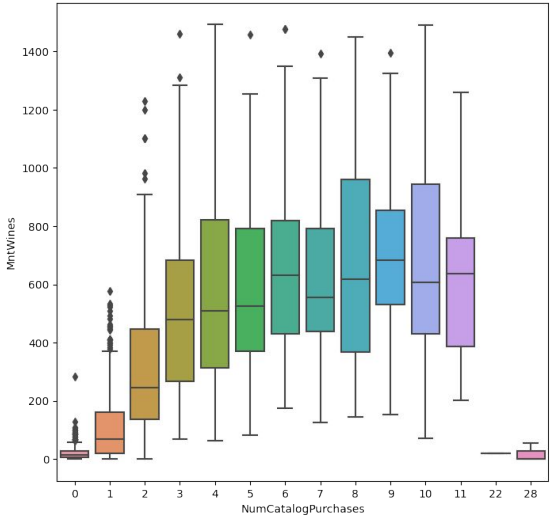


- 家中的兒童越多，酒水的開銷有較少的趨勢
- 家中青少年的多寡，與酒水的開銷無明顯趨勢

可能影響酒水花費的因素- 實體店購買

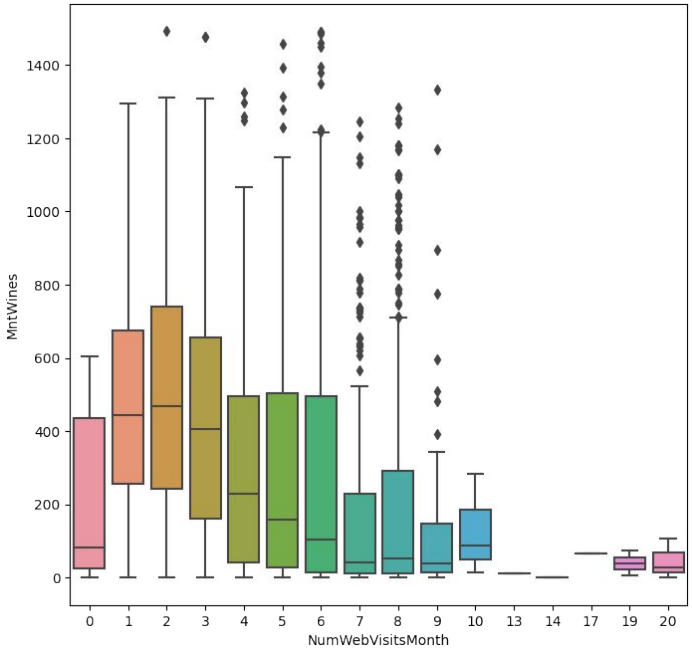






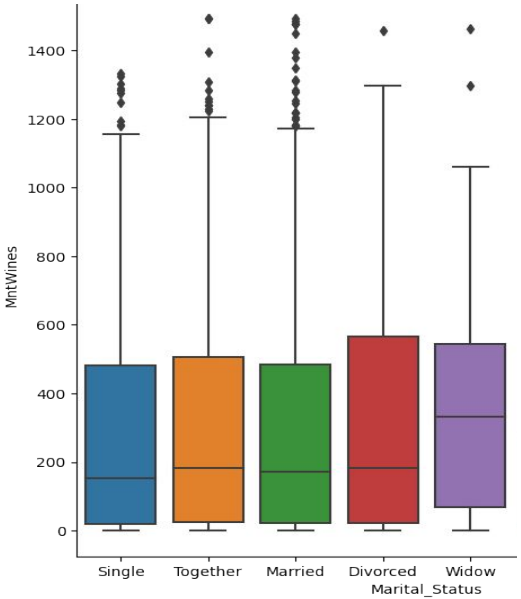
	在商店購買次數	在網站購買次數	在型錄購買次數
酒水的花費>200元	>五次	>四次	>二次
酒水的花費>400元	>九次	>八次	>六次

可能影響酒水花費的因素- 瀏覽行為

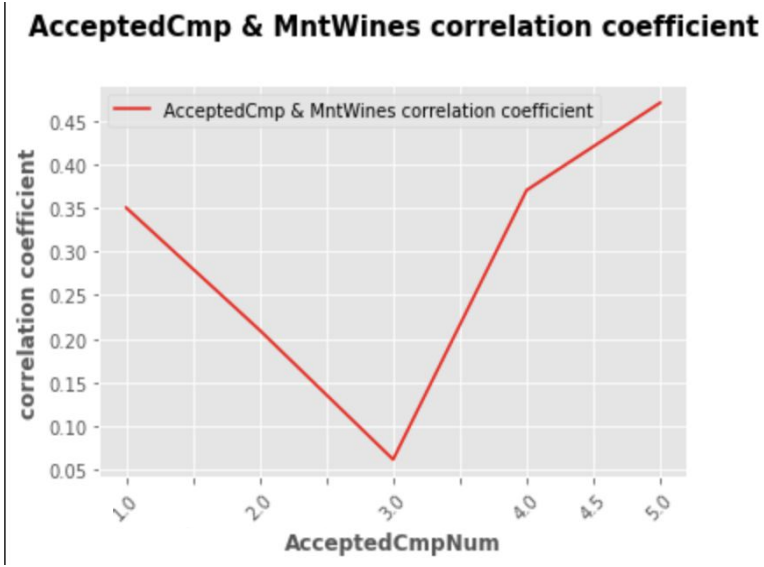


- 瀏覽網頁次數1~3次的顧客
  - **在酒水的花費高於瀏覽次數>4次的顧客**
  - **瀏覽次數越多反倒減少花費**

可能影響酒水花費的因素- 活動促銷



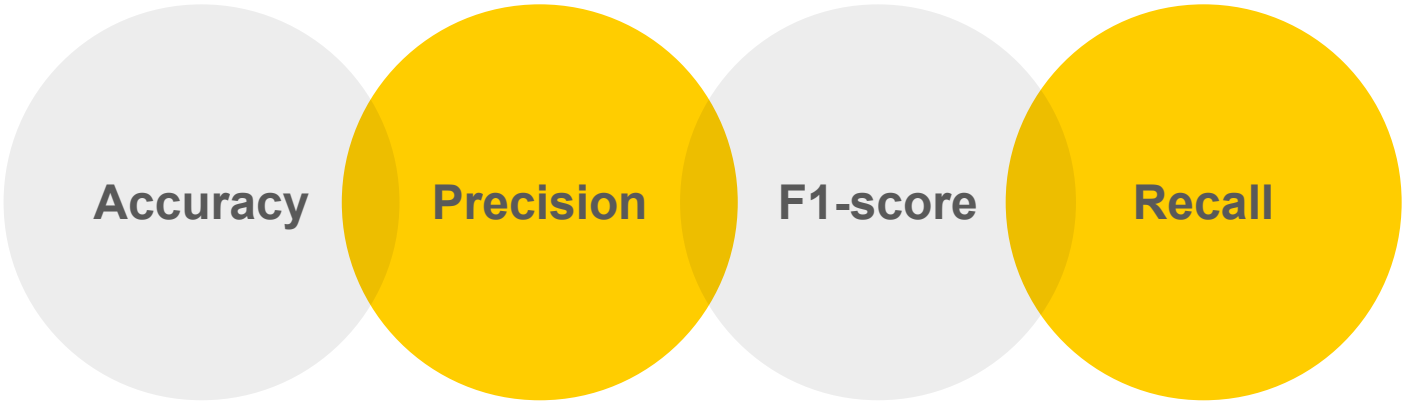
- 不同**婚姻狀況**在酒水的開銷分佈差不多



- **酒水上的開銷與促銷次數相關係數**在第一次促銷後開始逐次遞減
- 直到第四、五次促銷時，酒精消費力與促銷次數相關性再次提升

Metrics

Classification



Model & Insight

DNN / KNN / DecisionTree / RandomForest / XGBoost / CatBoost

DNN

DNN baseline

- 8 hidden layers

```
on training data:
    loss : 0.5790244340896606
    accuracy : 0.6540312767028809
    precision : 0.5
    recall : 1.0
    f1_score : [0.790833   0.51408136]

on test data:
    loss : 0.5478824377059937
    accuracy : 0.6371841430664062
    precision : 0.5
    recall : 1.0
    f1_score : [0.77839035 0.5324503 ]
```

DNN final

- 3 hidden layers

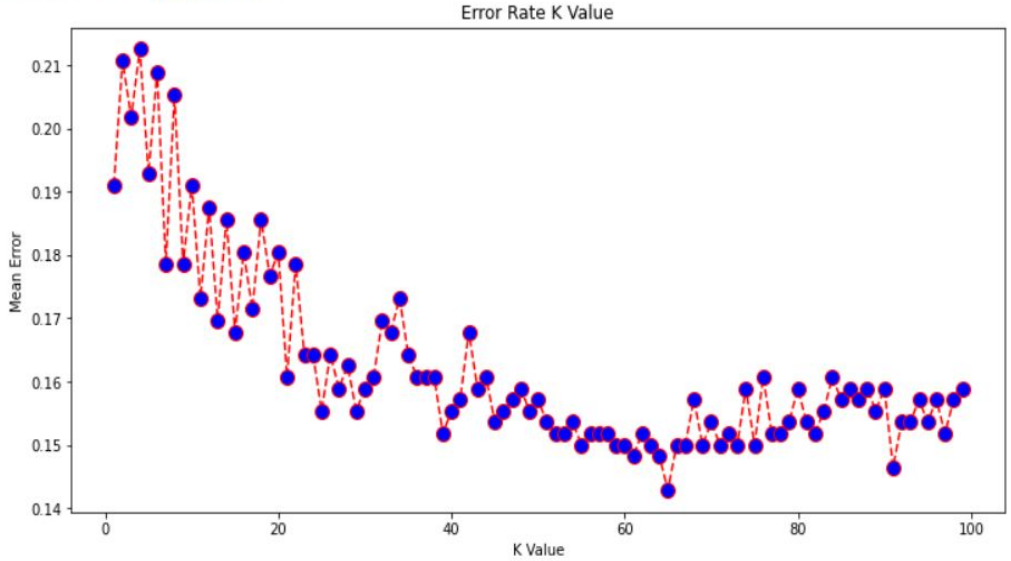
```
on training data:
    loss : 0.007712917868047953
    accuracy : 0.9988088011741638
    precision : 0.9993961453437805
    recall_2 : 0.9857057929039001
    f1_score : [0.99315387 0.99125874]

on test data:
    loss : 1.050164818763733
    accuracy : 0.8785714507102966
    precision : 0.8963531851768494
    recall_2 : 0.8339285850524902
    f1_score : [0.8908555  0.81885856]
```

KNN Classifier

```
error = []
for i in range(1, 100):
    knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(x_train, y_train)
    pred_i = knn.predict(x_test)
    error.append(np.mean(pred_i != y_test))
```

找出最適合的k值

```
Text(0, 0.5, 'Mean Error')
```

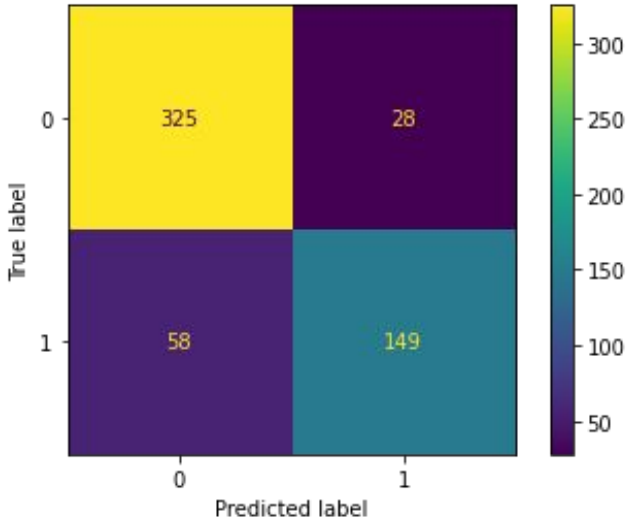


Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

KNN Classifier

找出最佳k值的預測結果

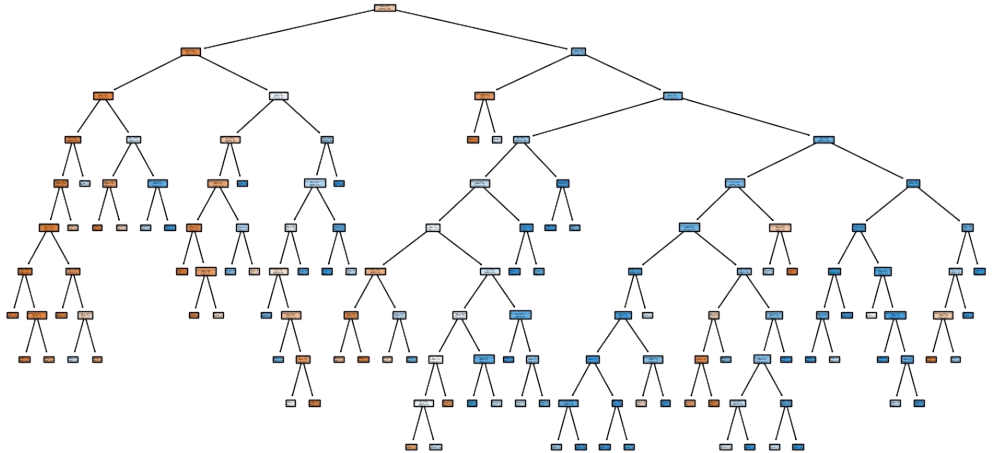
```
KNN
on training data:
    accuracy: 0.8695652173913043
    precision: 0.8509803921568627
    recall: 0.7521663778162911
    f1-score: 0.7985280588776448

on test data:
    accuracy: 0.8464285714285714
    precision: 0.8418079096045198
    recall: 0.7198067632850241
    f1-score: 0.7760416666666666
```



Decision Tree

參數設定

```
DecisionTreeClassifier(max_depth=10, min_samples_leaf=5, min_samples_split=10,
                       random_state=48)
```



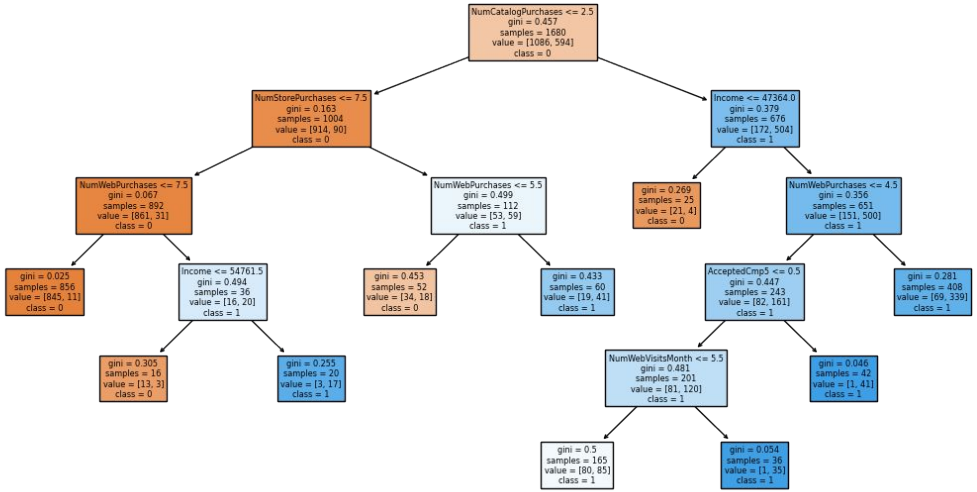
```
Income:        0.0822
Kidhome:       0.0
Teenhome:      0.0004
Recency:       0.0231
NumDealsPurchases: 0.0054
NumWebPurchases: 0.069
NumCatalogPurchases: 0.59
NumStorePurchases: 0.0994
NumWebVisitsMonth: 0.0437
AcceptedCmp3: 0.0
AcceptedCmp4: 0.004
AcceptedCmp5: 0.018
AcceptedCmp1: 0.0103
AcceptedCmp2: 0.0
Complain:      0.0
Response:      0.0005
Age:           0.022
Customer_Years: 0.008
Education_2n Cycle: 0.0081
Education_Basic: 0.0
Education_Graduation: 0.0093
Education_Master: 0.0
Education_PhD: 0.0
Marital_Status_Absurd: 0.0
Marital_Status_Alone: 0.0
Marital_Status_Divorced: 0.0
Marital_Status_Married: 0.0031
Marital_Status_Single: 0.0
Marital_Status_Together: 0.0036
Marital_Status_Widow: 0.0
Marital_Status_YOLO: 0.0
```

Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

Decision Tree Model

參數調整:增加 max leaf nodes = 10

```
DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10, min_samples_leaf=5,
                       min_samples_split=10, random_state=48)
```




```
Income:        0.0534
Kidhome:       0.0
Teenhome:      0.0
Recency:       0.0
NumDealsPurchases: 0.0
NumWebPurchases: 0.0737
NumCatalogPurchases: 0.7256
NumStorePurchases: 0.1006
NumWebVisitsMonth: 0.0258
AcceptedCmp3: 0.0
AcceptedCmp4: 0.0
AcceptedCmp5: 0.0209
AcceptedCmp1: 0.0
AcceptedCmp2: 0.0
Complain:      0.0
Response:      0.0
Age:           0.0
Customer_Years: 0.0
Education_2n Cycle: 0.0
Education_Basic: 0.0
Education_Graduation: 0.0
Education_Master: 0.0
Education_PhD: 0.0
Marital_Status_Absurd: 0.0
Marital_Status_Alone: 0.0
Marital_Status_Divorced: 0.0
Marital_Status_Married: 0.0
Marital_Status_Single: 0.0
Marital_Status_Together: 0.0
Marital_Status_Widow: 0.0
Marital_Status_YOLO: 0.0
```

Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

Decision Tree Model

結果評估

- baseline

```
average train acc: 0.9232142857142858
average train precision: 0.9105901039789639
average train f1score: 0.9180168800792594
average train recall: 0.9291673642502372

average test acc: 0.875
average test precision: 0.8577329808327825
average test f1score: 0.8629581462991709
average test recall: 0.8695590327169275

array([[328,  42],
       [ 28, 162]])
```

- 限制最大葉節點

```
average train acc: 0.8755952380952381
average train precision: 0.8627016126125997
average train f1score: 0.8697807241203467
average train recall: 0.8900468776159385

average test acc: 0.8767857142857143
average test precision: 0.8592358878702491
average test f1score: 0.8681907958697054
average test recall: 0.8862731152204837

array([[317,  53],
       [ 16, 174]])
```

Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

Experimental Results

初步結果(Baseline)

- without standardization
- without impute income
- without one-hot (drop)
- include income's outlier

model	acc	precision	f_score	recall
DNN	0.6372	0.5	0.6554	1
SVM	0.852	0.725	0.7797	0.8435
KNN	0.8252	0.7477	0.7711	0.7960
Decision tree	0.8791	0.7659	0.8521	0.9602
RandomForest	0.9025	0.8657	0.8657	0.8657
XGBoost	0.8953	0.8357	0.8599	0.8856
CatBoost	**0.9043**	**0.8627**	**0.8691**	**0.8756**

Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

調整後結果(Final)

- with standardization
- imputed income with KNN
- one-hot categorical features
- not included income's outlier

model	acc	precision	f_score	recall
DNN	0.6372 **0.8786**	0.5 **0.8964**	0.6554 **0.8549**	1 0.8340
SVM	0.852 **0.8804**	0.725 **0.8182**	0.7797 **0.8431**	0.8435 **0.8696**
KNN	0.8252 **0.8464**	0.7477 **0.8418**	0.7711 **0.7760**	0.7960 0.7198
DecisionTree	0.87910.8446	0.76590.7564	0.85210.8027	0.96020.8551
RandomForest	0.9025 **0.9160**	0.8657 0.8478	0.8657 **0.8924**	0.8657 **0.9420**
XGBoost	0.8953 0.8946	0.8357 0.8162	0.8599 **0.9227**	0.8856 0.8662
CatBoost	0.9043 **0.9161**	0.8627 0.8419	0.8691 **0.8934**	0.8756 **0.9517**

Data Sources https://www.kaggle.com/imakash3011/customer-personality-analysis

Discussions

預測及EDA後發現, 原本認為會顯著影響結果的特徵, 跟預期的不一樣

Discussions

	預期狀況	實際結果
顧客屬性	收入較高	● 符合預期 ● 在EDA階段或模型的特徵重要性皆顯示為關鍵特徵
	家中沒有兒童及青少年	● 在EDA階段發現：家中的兒童↑，酒水的開銷↓ ● 然而此特徵在Decision tree model中，沒有扮演特別關鍵的特徵
	顧客年齡	● 在EDA階段發現：年齡與酒水開銷沒有正向趨勢 ● 此特徵在Decision tree model中重要性佔0.022
購買方式	多在店面購買	● 在EDA階段發現：不同銷售通路的次數會與銷售總額有關連，但因數據不足無法推論多在何種通路購買
促銷活動成效	第一次看到廣告後便會購買	● 在EDA階段發現：第四、五次投放廣告時，酒精消費力與廣告投放次數相關性再次提升，甚至大於第一次投放 ● 「第五次投放廣告接受度」在Decision tree model中重要性佔0.021

Thanks!

Any questions ?