東吳巨資課程
發票資料集說明

行為即時追蹤

消費特徵洞察

麻布數據科技 | Jay Chen

資料集說明 (品類)

訓練資料

- 資料時間：2021/01/01~2021/01/14
- 資料數量：92,306 筆
- 資料集有多重標籤 (Multi Label)
- 可能包含人為分類錯誤

資料分布

- 訓練集: 217 類
- 測試集: 217 類

測試資料

- 資料時間：2022/01/01~2022/01/14
- 資料數量： 108,851 筆
- 資料集有多重標籤 (Multi Label)

評分標準

- Macro F1

資料集說明 (品牌)

訓練資料

- 資料時間：2021/02/01~2021/02/14
- 資料數量：70,046 筆
- 每筆明細只會有一個品牌
- 可能包含人為分類錯誤

資料分布

- 訓練集: 4,093 品牌
- 測試集: 5,389 品牌

測試資料

- 資料時間：2022/02/01~2022/02/14
- 資料數量：97,158 筆
- 會有訓練集未包含的品牌
- 每筆明細只會有一個品牌

評分標準

- 現有標籤：Macro F1 * 1
- 未知標籤：Jaccard(y_true, y_pred) * 1.2
- 得分：現有標籤 + 未知標籤

資料集欄位說明

品類資料集	品牌資料集	欄位名稱	備註
name	name	明細名稱	
occurred_at	occurred_at	消費日期	
iv_price	iv_price	發票總金額	
product		明細類別	測試資料集的該欄位為空值，請各團隊將測試資料預測結果提供給業師，業師會將判定結果提供各團隊
	brand	明細品牌	
unit	unit	明細數量	需注意：商家所上傳的明細原始資料，可能發生「明細數量 x 明細單價 不等於 明細總價」
unit_price	unit_price	明細單價	
total_price	total_price	明細總價	
channel	channel	通路	營業秘密之故，通路名稱以 Hash 後的結果提供

Jaccard 程式碼

- Jaccard 原理，請參考 Kaggle 說明 (點我)
- 由於資料屬性議題，請各團隊使用以下調整後的 Jaccard 程式碼，進行品牌標記的評分

```
def jaccard(str1, str2):
    a = set(str1.lower().replace(' ',''))
    b = set(str2.lower().replace(' ',''))
    c = a.intersection(b)
    return float(len(c)) / (len(a) + len(b) - len(c))
```

評分標準

1.品類分類：Macro F1

2.品牌標記：

- 已知品牌標籤：Macro F1 * 1
- 未知品牌標籤：Jaccard(y_true,y_pred) * 1.2

3.分數比重：品類分類 50%，已知品牌標記 50%，滿分為 100 分

4.額外加分：未知品牌標籤

5.如果有使用外部資料，則需要提供資料來源