玉山人工智慧公開挑戰賽
2021冬季賽
信用卡消費類別推薦

隊伍名稱：把拉拉嘟嘟

Private leaderboard：0.725021 (6th)

蔡岳霖，王維綱，林芊，黃宇瑍，
王書偉，魏旻柔，藍雲瀚，李宇堂

Outline

- 資料前處理
- 特徵工程
- 模型與超參數
- 類別推薦方法
- 集成學習

資料前處理

- 此問題本質為序列推薦(sequential recommendation)任務，資料的時序關係是關鍵特徵，因此本隊以移動窗格方法整理顧客的消費紀錄，最終取sliding window = 18

  前18個月份的特徵預測下個月的交易狀況

- 選用特徵：16類產品的**消費金額**、**次數**、**線下國內**、**線上國內**、**線下海外**、**線上海外金額佔比**、與移動窗格中**金額占筆最高的卡片類別**。

- 每筆資料的特徵數量： 16(類別數) x 7(選用特徵) x 18 (窗格大小) = 2,016

遺漏值補0，並刪除完全無交易行為的使用者

特徵工程

- 除前處理完的資料，我們額外增加以下特徵

masts, educd, trdtp, naty, poscd, cuorg, slam, gender_code, age, primary_card

- 顧客特徵：
  - 以每位顧客在資料的最後一筆有效資料為主 (+10)
  - 另加入最初18個月平均消費金額比例最高的卡號 (+1)

- 其餘特徵工程：
  - 所有商品18個月的消費金額/次數之平均值/標準差 (+4)
  - 所有商品近3個月的消費金額/次數之平均值/標準差 (+4)
  - 上述兩者平均數的比值 (+2)
  - 每類商品18個月的消費金額/次數之平均值/標準差 (+64)
  - 每類商品線上下國內外18個月的消費金額比例平均值 (+64)

模型與超參數

- 以上述資料為輸入特徵，建立16個分類模型與16個迴歸模型分別預測16類商品下個月的消費行為

- 分類：下個月此類別商品是否有交易紀錄(True/False)
- 迴歸：下個月此類別商品之交易金額

- 選用模型：Light GBM classifier / regressor
  - n_estimators:1000
  - subsample:0.8
  - colsample_bytree:0.8

類別推薦方法

- 在測試資料上，以每類商品的消費機率(prob)與消費金額(price)進行以下計算

$$Expected\ value = prob^{1.5} * price$$

$$\overset{500,000}{\begin{bmatrix} 0.03^{1.5} & \cdots & 0.14^{1.5} \\ \vdots & \ddots & \vdots \\ 0.62^{1.5} & \cdots & 0.09^{1.5} \end{bmatrix}} * \overset{16}{\begin{bmatrix} 0 & \cdots & 2713.4 \\ \vdots & \ddots & \vdots \\ 5812.57 & \cdots & 93.51 \end{bmatrix}}$$

- 再依照各類商品之期望值進行排序後取前三高之商品類別依序作為推薦結果。

集成學習

- 在ensemble作法上，我們取不同特徵製作了三份原始資料，並分別依此訓練模型，將模型之預測機率與預測消費金額分別做平均後再以相同方法做類別推薦。
  - Ver. A： 消費金額、次數、線下國內、線上國內、線下海外、線上海外金額佔比、與當月金額占筆最高的卡片類別
  - Ver. B：消費金額、次數、線下國內、線上國內、線下海外、線上海外消費次數、與當月金額占筆最高的卡片類別
  - Ver. C：消費金額、次數、線下國內、線上國內、線下海外、線上海外消費次數、與當月消費次數最高的卡片類別

本次比賽能勝出之關鍵因素

- 硬體資源足以支援資料維度
  - 168 GB ram
  - 最終做法中未使用GPU做運算

- 使用lightGBM演算法，能夠快速進行線下測試與實驗各種特徵
  - LGB : ~5mins / 16 models
  - XGB : ~4hrs / 16 models

- 分類與迴歸模型計算之期望值做推薦依據有效提升NDCG分數

Experiment result

作法	提升分數(offline validation)
Prob * Price	+0.000
Prob only	-0.0045
Price only	-0.0085

單以price或prob皆得到較低的NDCG分數

Experiment result

作法	提升分數(private leaderboard)
Ensemble (3 models)	+0.0003
刪除窗格內沒有交易紀錄之使用者	+0.0012
加入每月每類商品最常使用之卡號	+0.0004
加入國內外與線上下之特徵	+0.0004
Prob^1.5 * Price	+0.0005
加入顧客特徵	+0.0010

base model: 0.720 (LGB with default hyperparams , 576 features(16x2x18))

其餘分數提升主要為特徵工程與模型超參數

Codes: https://reurl.cc/Mbb3y4

Thanks for listening

Q & A

補充資料

Validation mode

data set	start of window	end of window	y label
train	dt 1	dt 17	dt 18
	dt 2	dt 18	dt 19
	dt 3	dt 19	dt 20
	dt 4	dt 20	dt 21
	dt 5	dt 21	dt 22
	dt 6	dt 22	dt 23
valid	dt 7	dt 23	dt 24

線下驗證時，取window size = 17並以dt24作為驗證NDCG之資料

Testing mode

data set	start of window	end of window	y label
train	dt 1	dt 18	dt 19
	dt 2	dt 19	dt 20
	dt 3	dt 20	dt 21
	dt 4	dt 21	dt 22
	dt 5	dt 22	dt 23
	dt 6	dt 23	dt 24
test	dt 7	dt 24	-

實際提交時，重新以window size = 18做訓練