

## 從觀察資料對比賽做了哪些假設? 消費面切入

非盜刷金額分布

- 經常性消費居多 -消費金額集中在NTD3000內
- ●視卡號為最小顆粒度
- 即使一人多卡,也會因活動模式不同而產生不同的消費模式
- 一分析玉山信用卡商品類別
- -利用歷史消費偏好(mcc,國外消費,網購消費)作為卡片類別的判斷依據,依此對每張卡號進行客戶貼標







Diff

## 基於信用卡特性做了哪些特徵?抓出異常差異



盜刷金額分布

特定通路任務有加碼,幹嘛去刷其他通路?

-統計每張卡號在此筆消費以前,過去消費紀錄之特店代號眾數類別



每張卡的消費模式會隨意變動?

-計算本筆消費金額與過去平均消費金額之差異



消費分數差異!

-統計常用消費組合並依Baseline Model 之特徵重要度給予權重

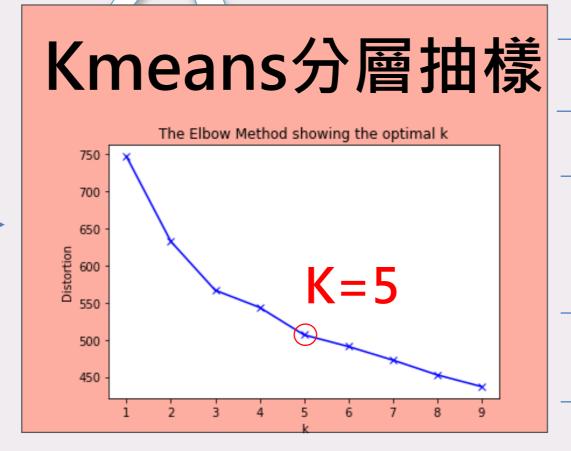
How

## 如何處理不平衡資料?對無盜刷紀錄資料做Undersampling

有盜刷 紀錄 (141625) 無盜刷 紀錄 (1389938)

模型應不受 消費次數影響

删除交易 筆數 > 19 筆卡號之 消費等 最早前 最早前 30%資料 Stage 消費金額相關變數 Gain很大



保留各消費金額級 距的特徵多樣性

Group 3
Group 4
Group 5

Group 1
Group 2

無盜刷紀錄

全部保留

Random 80%
Random 80%

(1006783) 少了 38萬等

有盜刷

紀錄

(141625)

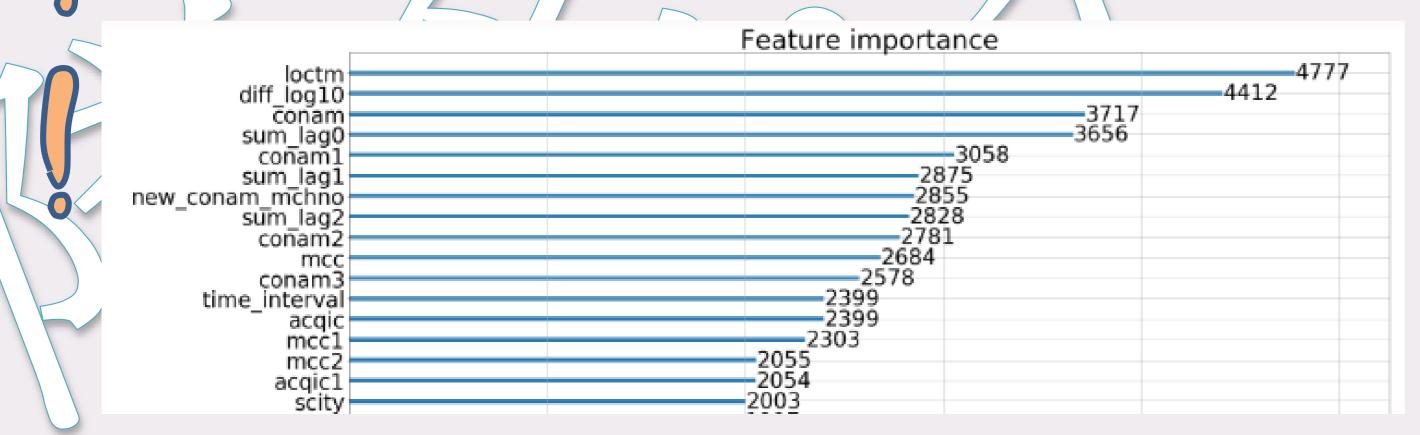
Model

## 編碼,stacking & RIPPER演算法觀察分裂規則

Target Encoding&一對一逐數轉換對類別變數進行編碼

StratifiedKFold + GridSearch 決定XGBoost最適參數

Stacking XGBoost & LightGBM



RIPPER Train Data Set

XGBoost Train Step

Model

Test Data Set

Performance
Performance
better

RIPPER tranform

Rule Set

Create Variable

F1 Private Score = 0.668