



特徵：

- 交易次數特徵 (3個)
 - 同一張卡(cano)的總交易次數
 - 同一張卡(cano)跟交易日子(locdt)的總交易次數
 - 同一個用戶(bacno)與商店(mchno)的總交易次數
- 交易時間特徵(3個)
 - 上／下一次交易時間(loctm) 的差
 - 同一張卡(cano)跟交易日子(locdt)， 交易時間(loctm)的標準差
- 交易金額特徵(3個)
 - 同一張卡(cano)跟交易日子(locdt)，最大／最少的交易金額(conam)
 - 同一張卡(cano)跟交易日子(locdt)，該交易與交易金額(conam)是零（如有）的交易時間(loctm)差
- 換卡特徵(2個)
 - 同一張卡(cano)，該交易與最後一筆交易的日子(locdt)差
 - 同一個用戶(bacno)，一張卡(cano A)最後的交易日子(locdt)跟接著一張卡(cano B)的第一個交易日子(locdt)的差
- 商店特徵(2個)
 - 同一個用戶(bacno)跟商店(mchno)，最後跟第一筆交易的日子(locdt)差
 - 同一張卡(cano)跟商店(mchno)的第幾筆交易
- 特殊特徵(4個)
 - 同一張卡(cano)，該商店(mchno)在之前有沒有出現出現過盜刷(fraud_ind)
 - 同一張卡(cano)，該商店(mchno)在之前有沒有出現出現過但不是盜刷(fraud_ind)
 - 同一張卡(cano)，該交易金額(conam)在之前有沒有出現出現過盜刷(fraud_ind)
 - 同一張卡(cano)，第一次盜刷(fraud_ind)(如有)與當天交易的日子(locdt)差

模型：

- 五個模型：1個基礎模型 + 4個特殊模型
- 基礎模型： 20個原始特徵 + 13個生成特徵
- 特殊模型：針對一些testing 跟training 都有出現過的用戶(bacno)，建立4個特殊模型，利用盜刷標注(fraud_ind)去獲取用戶(bacno)被盜刷的信息。每個特殊模型由基礎模型的特徵加一個特殊特徵訓練。
- 如果那筆交易符合特殊特徵的條件，就會用該特殊模型去預測，如果都不符合就會用基礎模型去預測。
- 注意：特殊特徵在現實生活中未必有用，因為我們在預測時未必知道該用戶以往的盜刷標注

結論：

- 只用到小量的生成特徵訓練相對簡單的模型，減低過度擬合(overfitting)
- 把一些只有在訓練集出現過的類別型特徵的值替換成NA，減少模型學到一些不能應用在測試集上的規則
- 使用Early Stopping 和 GroupKfold以用戶(bacno)去分割訓練習，減少訓練的迭代次數，避免過度擬合(overfitting)
- 使用10 folds，以一個標準差去排除一些極端的fold，再以平均值估算
- 詳細請參閱：https://github.com/aarontong95/TBrain_Credit_Card