預測洗錢交易

想法

1.預測有無洗錢洗錢 → 比較近期與長期交易紀錄差異

2.多個資料集怎麼合併 →

方法1.將多個資料集以天為單位合併(加總、平均…)，並以label data日期為起始點將近N天合併作為近期資料，N天以前為長期資料。

類別特徵資訊損失大
Public score:0.012

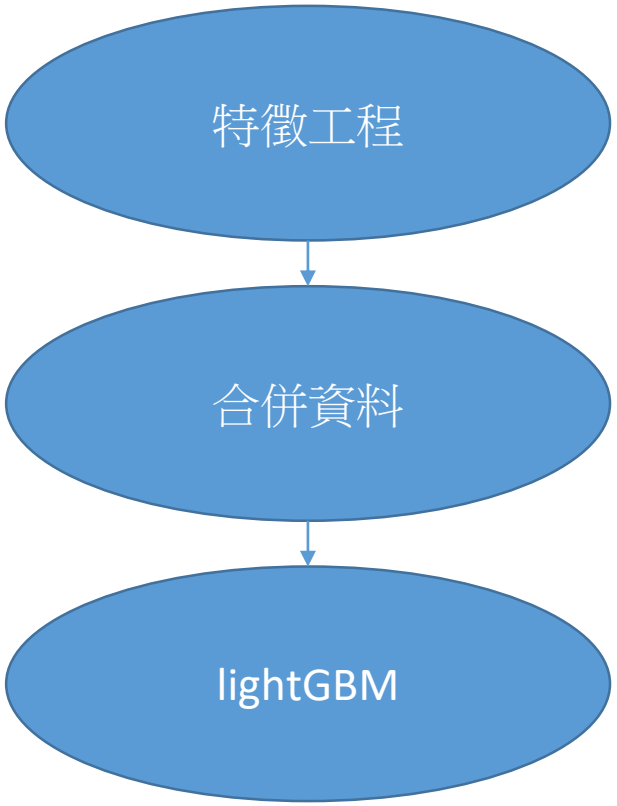
方法2.將多個資料集以天為單位合併(加總、平均…)，並以label data日期為起始點將近N天的資料當作特徵作為近期特徵，N天以前為長期特徵。

類別特徵資料過多，速度慢
Public score:0.011

方法3.將多個資料集以天為單位合併(加總、平均…)，並以label data日期為起始點將近N天的資料以topk形式來選擇k個最常出現的類別特徵當成近期特徵，N天以前的為長期特徵。

資訊損失較小，速度快，所以選擇這方法
Public score:0.018

方法流程



特徵工程

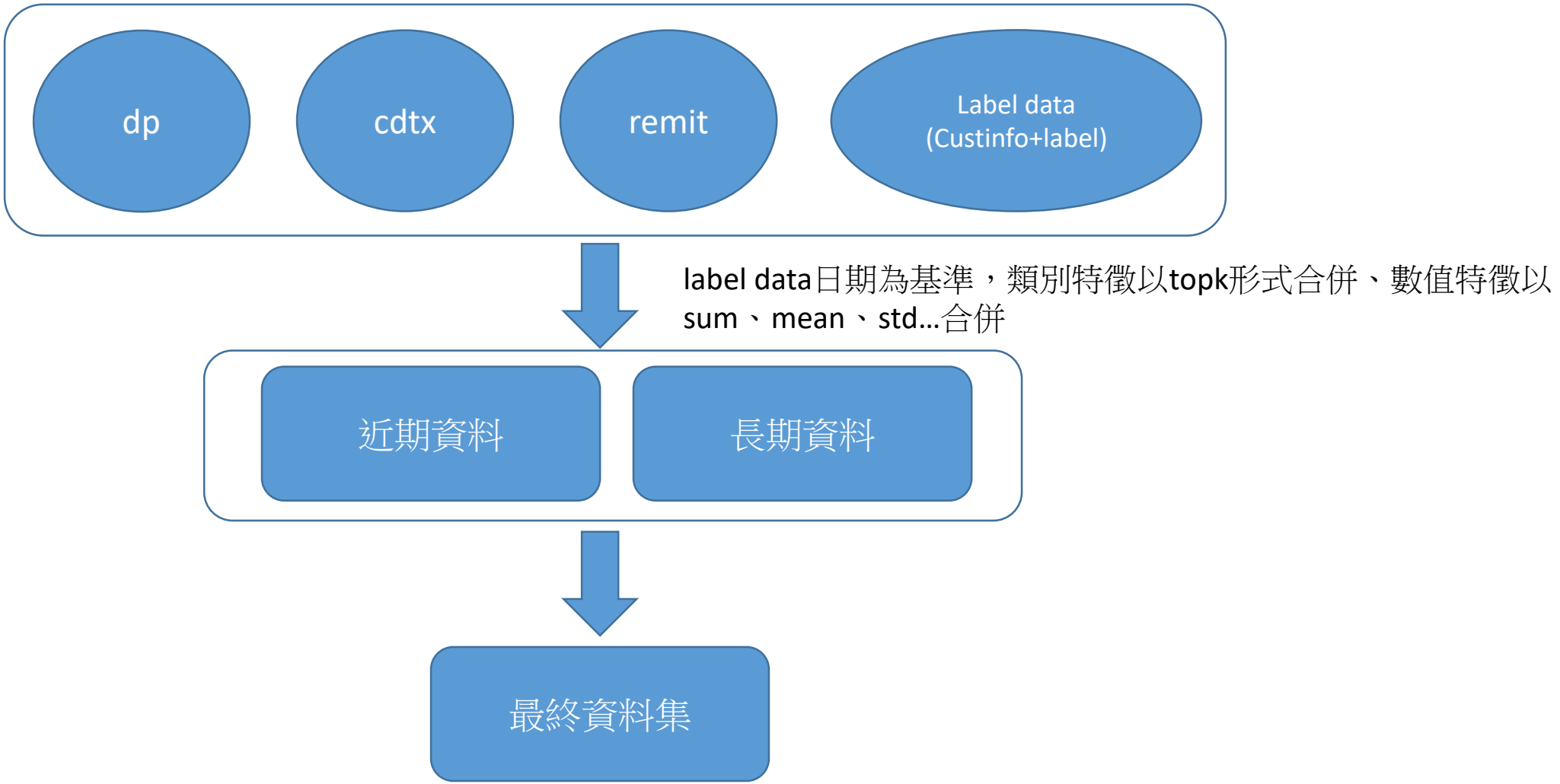
將借貸資料以借貸別分成兩個資料集，其他資料集不變。

- 數值特徵:以sum、mean、std、min、max合併成一天資料
- 類別特徵:1.將各個類別特徵組合，例如借貸資料裡面有交易代碼與分行代碼兩個資料，透過轉成字串後相接變成新的組合特徵，範例如下，2.匯率轉成類別特徵。

交易代碼:4,分行代碼:167 ⟶ 組合特徵:4167

- 特徵選擇使用mlxtend的SequentialFeatureSelector

合併資料


dp
cdtx
remit
Label data
(Custinfo+label)
label data日期為基準，類別特徵以topk形式合併、數值特徵以sum、mean、std…合併
近期資料
長期資料
最終資料集

lightgbm

- 將資料集切成5份，以交叉驗證的方法來預測測試資料集的label。

改進

- 類別特徵:

新增類別眾數差特徵:比較近期資料與長期資料差異，有差異為1

新類別數:比較近期類別特徵比長期類別特徵新出現多少個特徵

推敲不同時間段:星期、平日、假日、連假。

特徵選擇:

mlxtend的SequentialFeatureSelector可以搭配Multi-Fidelity Search(halvingGridSearchCV)這種trick來加快選擇速度，也就是使用少量訓練資料來選擇特徵，再漸進的加大資料來選特徵。