一、資料處理

1. AGE：類別代號 a:q 依序轉為整數 1:17
2. SEX：類別代號 a:b 依序轉為 0,1
3. HEIGHT、WEIGHT：屬於連續型特徵，所以使用 R 程式套件 mice 的 cart 補遺失值
4. OCCUPATION：以開頭的英文字母作為分類依據，同一字母 OCCUPATION 視為同一職業，每一欄職業別代號為 1:從事此職業，0:不從事此職業，所以有限制式，因此共有 23 種職業，22 種特徵
5. BUY_MONTH：轉為類別型特徵，與 OCCUPATION 處理方式相同，每一月份別代號為 1:此月份，0:非此月份，有限制式，因此共有 12 個月份，11 種特徵
6. BUY_YEAR：因為代號一樣為同一年，無其他資訊，所以不採用
7. CITY_CODE：與 OCCUPATION 處理方式相同，每一城市代號為 1:在此城市，0:不在此城市，有限制式，因此共有 23 個城市，22 種特徵
8. MARRIAGE：與 OCCUPATION 處理方式相同，每一婚姻型態代號為 1:此婚姻型，0:不為此婚姻型，有限制式，因此共有 6 種婚姻型態，5 種特徵
9. STATUS1、STATUS2、STATUS3：皆補遺失值成為類別 b，再轉代號 0:a,代號 1:b
10. STATUS4：補遺失值成為類別 a，再轉代號 0:a,代號 1:b
11. EDUCATION：與 OCCUPATION 處理方式相同，每一教育型態代號為 1:此教育型，0:不為此教育型，有限制式，因此共有 4 種教育水準，3 種特徵
12. IS_NEWSLETTER 、CHARGE_WAY、IS_EMAIL、IS_PHONE、IS_APP、IS_SPECIALMEMBER：其遺失值新增為一 label，與 OCCUPATION 處理方式相同，每一型別代號(a,b)為 1:此型別，0:非此型別，所以共有 3 種型別，且有限制式，因此有 2 個特徵
13. INTEREST1，INTEREST2，...，INTEREST10：其遺失值新增為一 label，且與 OCCUPATION 處理方法相同，每一種興趣代號(a,b)為 1:此型別，0:非此型別，所以 10 種興趣總共有 30 種型別，且有限制式，因此有 20 個特徵
14. PARENTS_DEAD、REAL_ESTATE_HAVE、IS_MAJOR_INCOME：與 OCCUPATION 處理方法相同，每一型別代號(A,B)為 1:此型別，0:非此型別，所以共有 2 種型別，且有限制式，因此有 1 個特徵

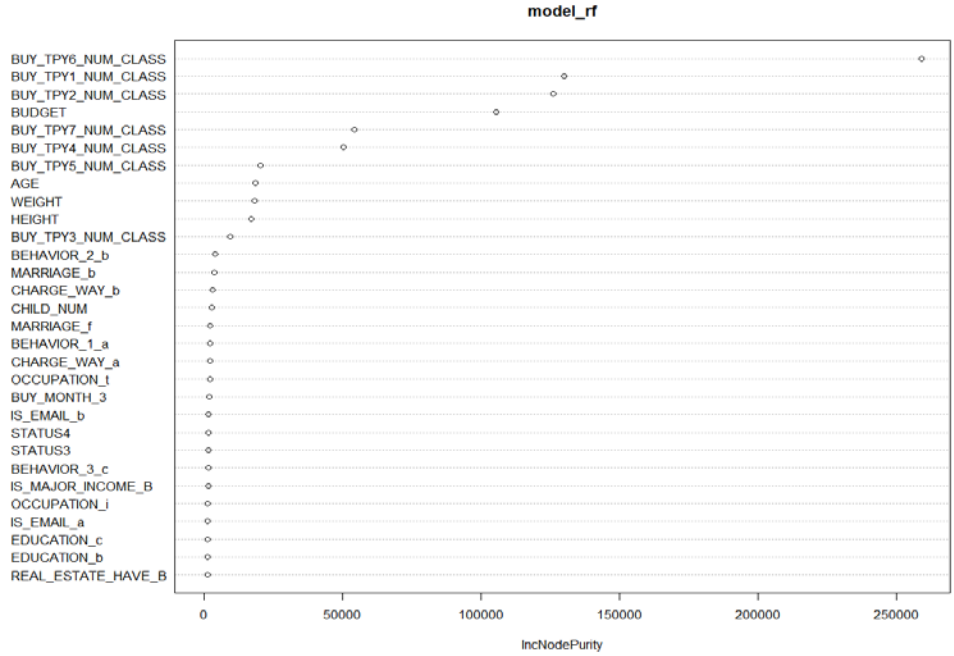
15. BUY_TPY1_NUM_CLASS，...，BUY_TPY7_NUM_CLASS：先將特徵 A:G 依序轉成整數 1:7 後，扣掉 7 取絕對值處理
16. BUY_TPYE: 發現到 BUY_TPY1_NUM_CLASS，...，BUY_TPY7_NUM_CLASS 與 BUY_TPYE 有對應關係，分別為 1->a, 2->e, 3->b, 4->c, 5->f, 6->d, 7->g

二、特徵選取

整理後的特徵個數一共有 124 個。

有經過特徵篩選或降維的有 pca 和 lasso 兩種方法；經 pca 計算後保留 pc1 到 pc42(約 95%解釋量)，共 42 個線性轉換後的特徵；lasso 則總共選取了 51 個特徵；其餘模型則用全部 124 個特徵。

隨機森林所篩選的重要特徵排序:



透過隨機森林演算法所得到的特徵重要性如上圖，可以看出過去曾購買商品件數與預算對我們預測客戶購買的商品類型有非常大的貢獻。

三、模型選擇與驗證模型成效

我們一共嘗試了五種模型，分別是

1. 先對特徵值做主成分分析，再利用 logistic 預測
2. LASSO
3. 隨機森林
4. Support vector machine
5. XGBOOST

接著利用 grid search 找出每個 model 的最佳參數，各 model 所得到的 5-fold CV error 如下表:

Model	5-folds cv_error
PCA	0.1787177
LASSO	0.1718649
Randomforest	0.06520048
SVM	0.1178611
XGBOOST	0.049976

由於 Logistic 與 LASSO 的 CV error 過高所以我們不採用，最後就選擇隨機森林、SVM 以及 XGBOOST 三種模型做整合，得出最終預測模型。

最終的模型採用 Stacking 的方法來做整合，將以上三個模型配適的預測值再進一步採用 xgboost 作預測，藉此達到三個臭皮匠勝過一個諸葛亮的最大效益。