分析案例

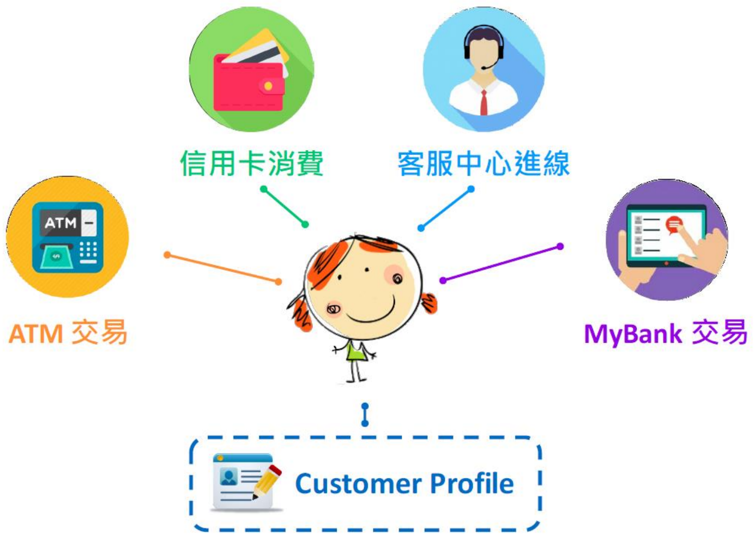
客戶流失模型預測

客戶流失的挑戰

開發新客戶的成本是維繫舊客戶的 5 倍

挽留住 5%顧客，可以降低 18%運營成本

客戶資料構面

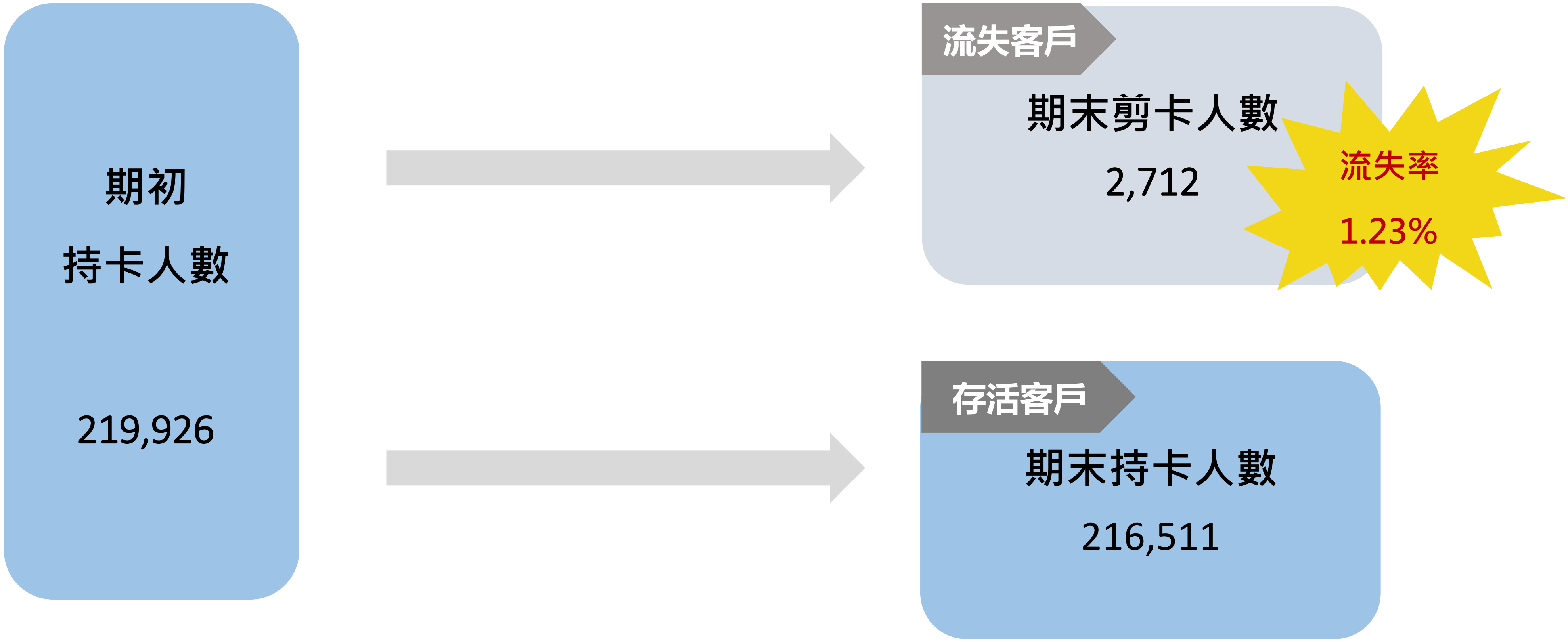


- 含23萬筆客戶, 300萬筆服務紀錄

- 客戶資料含人口特徵，資產餘額，服務使用情況等標籤

- 信用卡交易含卡別，等級，消費類別、與金額

信用卡流失評估



期初

持卡人數

219,926

流失客戶

期末剪卡人數

2,712

流失率

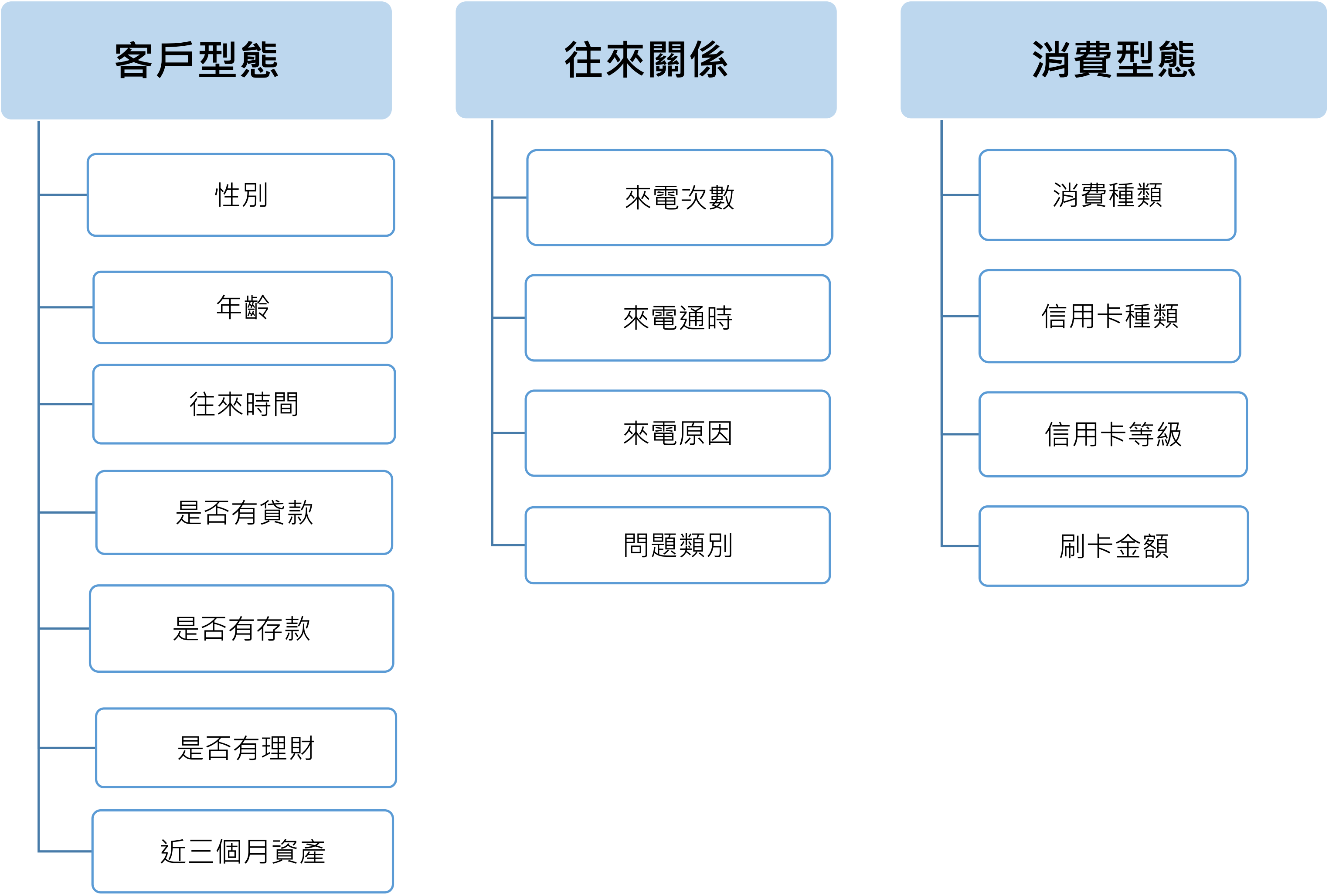
1.23%

存活客戶

期末持卡人數

216,511

問題分析思路

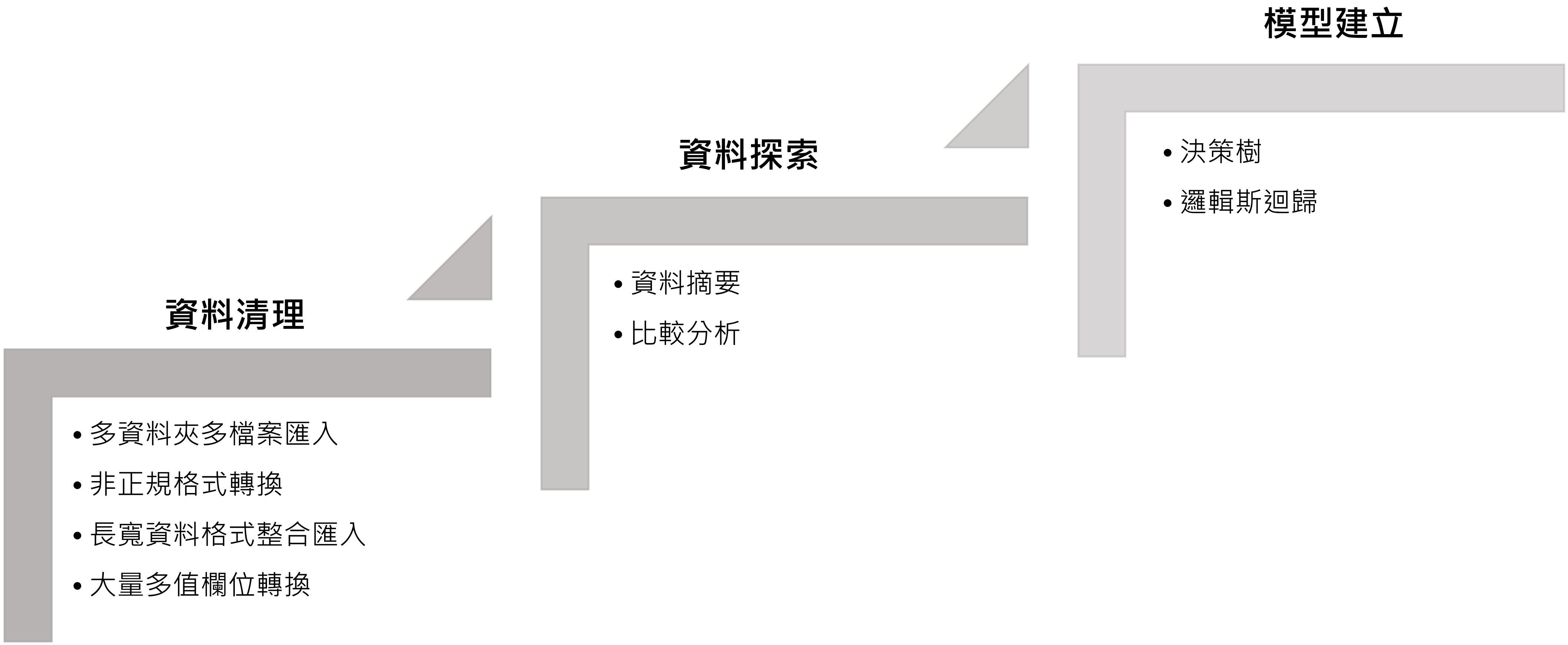


客戶型態
- 性別
- 年齡
- 往來時間
- 是否有貸款
- 是否有存款
- 是否有理財
- 近三個月資產

往來關係
- 來電次數
- 來電通時
- 來電原因
- 問題類別

消費型態
- 消費種類
- 信用卡種類
- 信用卡等級
- 刷卡金額

資料建模流程



資料清理

- 多資料夾多檔案匯入
- 非正規格式轉換
- 長寬資料格式整合匯入
- 大量多值欄位轉換

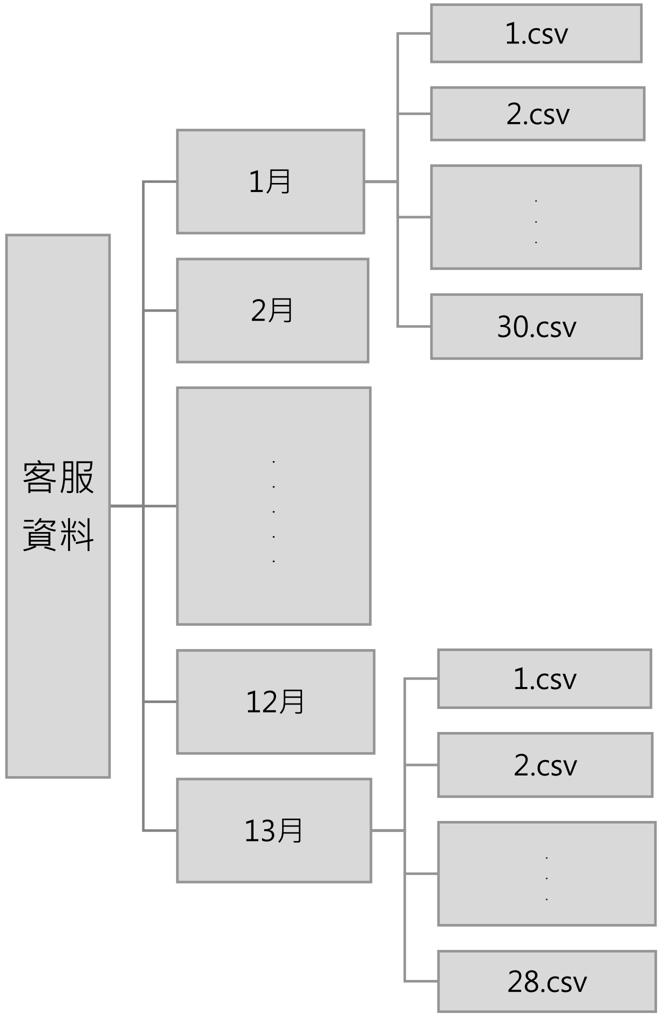
資料探索

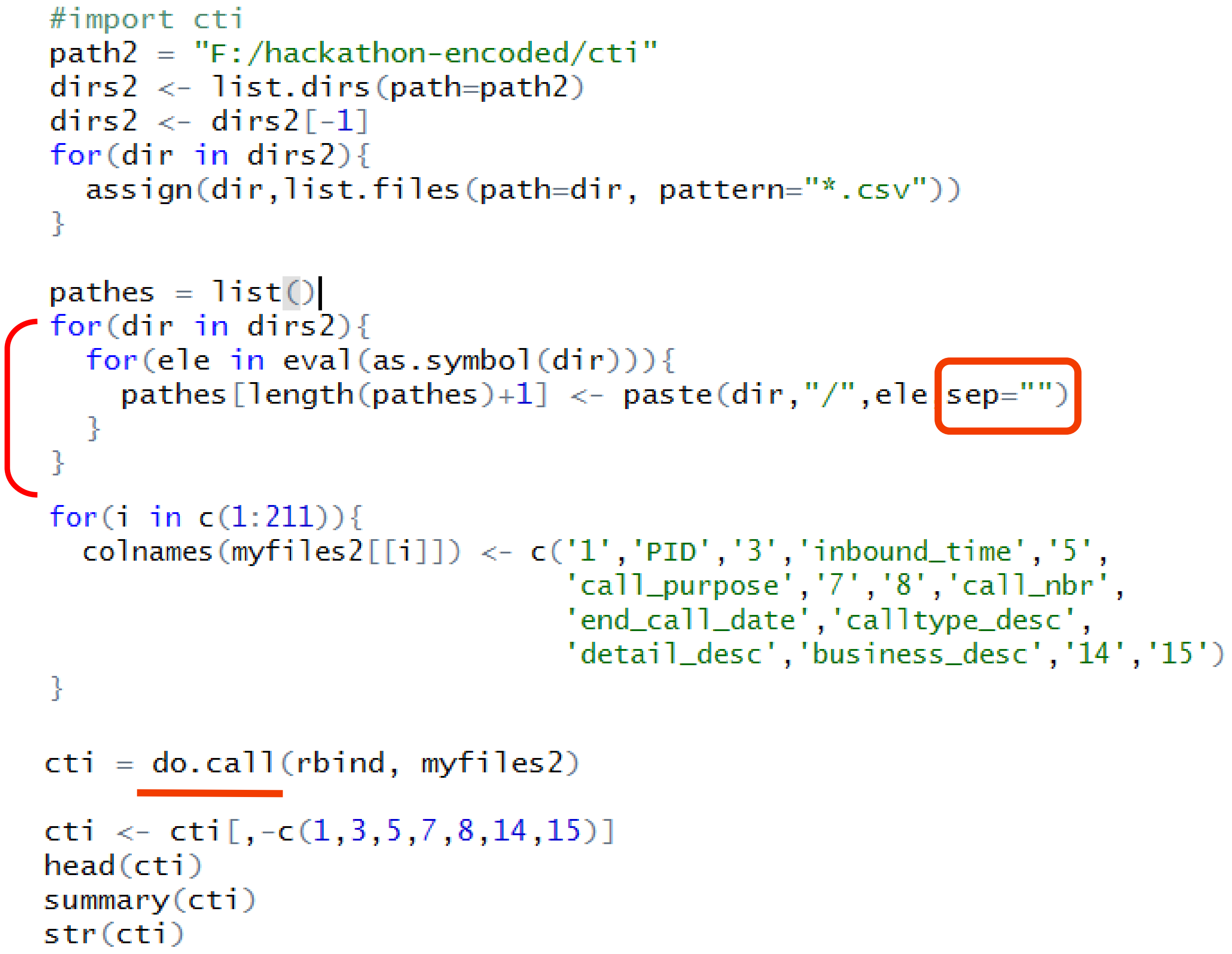
- 資料摘要
- 比較分析

模型建立

- 決策樹
- 邏輯斯迴歸

資料清理1：多資料夾多檔案匯入

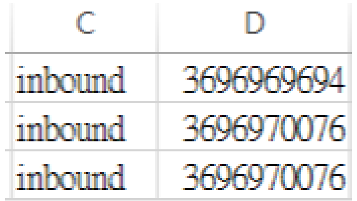




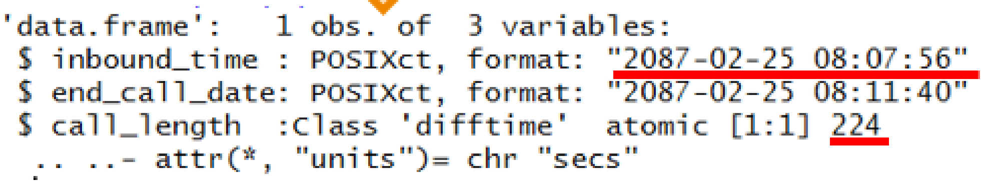
```r
#import cti
path2 = "F:/hackathon-encoded/cti"
dirs2 <- list.dirs(path=path2)
dirs2 <- dirs2[-1]
for(dir in dirs2){
  assign(dir,list.files(path=dir, pattern="*.csv"))
}

pathes = list()
for(dir in dirs2){
  for(ele in eval(as.symbol(dir))){
    pathes[length(pathes)+1] <- paste(dir,"/",ele,sep="")
  }
}

for(i in c(1:211)){
  colnames(myfiles2[[i]]) <- c('1','PID','3','inbound_time','5',
                               'call_purpose','7','8','call_nbr',
                               'end_call_date','calltype_desc',
                               'detail_desc','business_desc','14','15')
}

cti = do.call(rbind, myfiles2)

cti <- cti[,-c(1,3,5,7,8,14,15)]
head(cti)
summary(cti)
str(cti)
```

資料清理2：非正規格式轉換

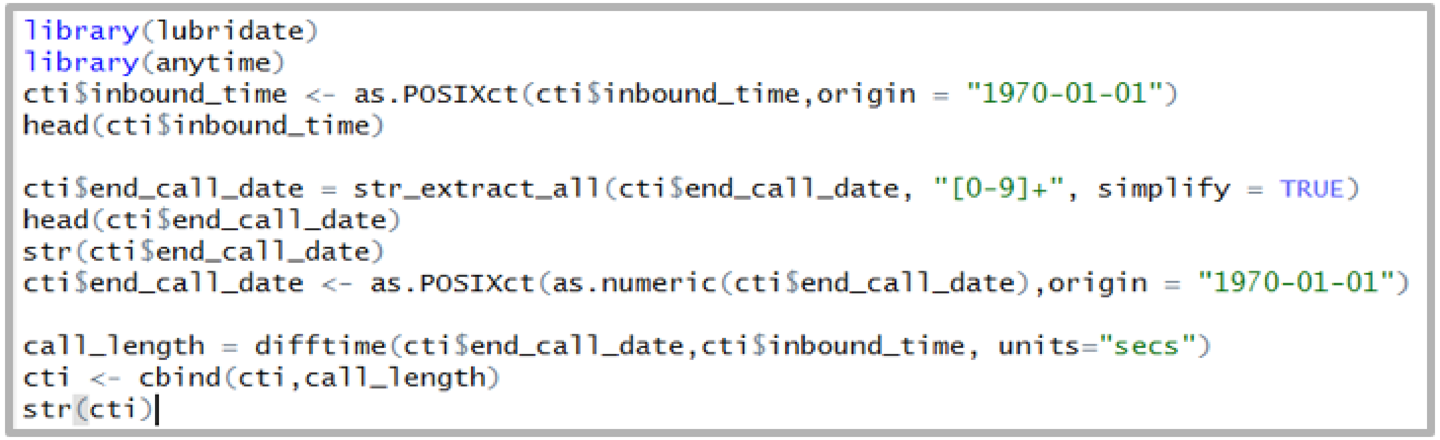
UTC位移後的Timestamp



C	D
inbound	3696969694
inbound	3696970076
inbound	3696970076

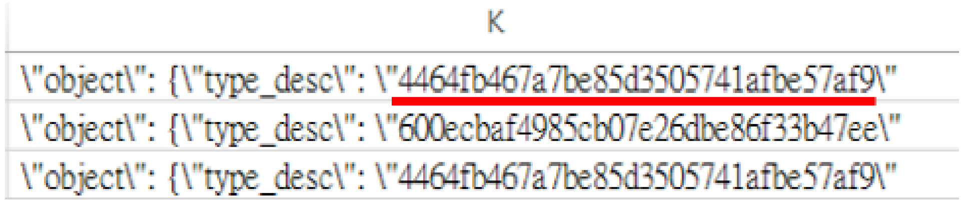


```
'data.frame':   1 obs. of  3 variables:
 $ inbound_time : POSIXct, format: "2087-02-25 08:07:56"
 $ end_call_date: POSIXct, format: "2087-02-25 08:11:40"
 $ call_length  :Class 'difftime'  atomic [1:1] 224
  .. ..- attr(*, "units")= chr "secs"
```

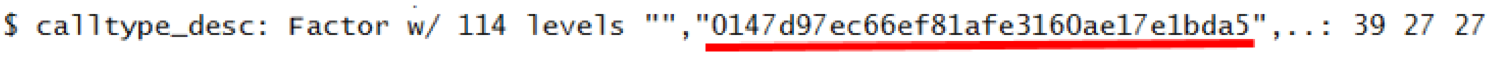


```
library(lubridate)
library(anytime)
cti$inbound_time <- as.POSIXct(cti$inbound_time,origin = "1970-01-01")
head(cti$inbound_time)

cti$end_call_date = str_extract_all(cti$end_call_date, "[0-9]+", simplify = TRUE)
head(cti$end_call_date)
str(cti$end_call_date)
cti$end_call_date <- as.POSIXct(as.numeric(cti$end_call_date),origin = "1970-01-01")

call_length = difftime(cti$end_call_date,cti$inbound_time, units="secs")
cti <- cbind(cti,call_length)
str(cti)
```

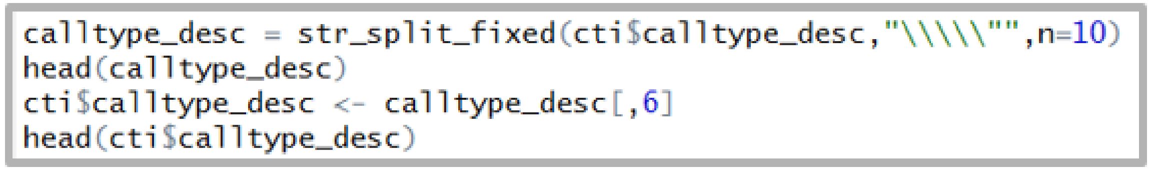
參雜Json格式



K
\"object\": {\"type_desc\": \"4464fb467a7be85d3505741afbe57af9\"
\"object\": {\"type_desc\": \"600ecbaf4985cb07e26dbe86f33b47ee\"
\"object\": {\"type_desc\": \"4464fb467a7be85d3505741afbe57af9\"

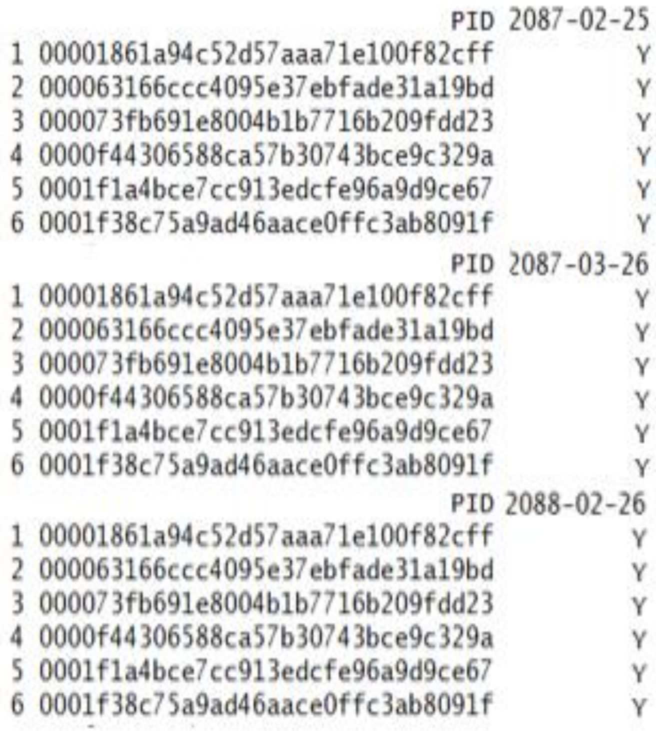


```
$ calltype_desc: Factor w/ 114 levels "","0147d97ec66ef81afe3160ae17e1bda5",..: 39 27 27
```

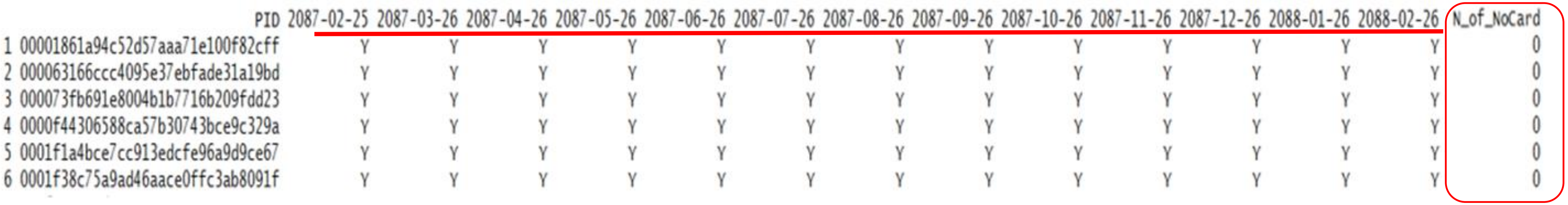


```
calltype_desc = str_split_fixed(cti$calltype_desc,"\\\\\"",n=10)
head(calltype_desc)
cti$calltype_desc <- calltype_desc[,6]
head(cti$calltype_desc)
```

資料清理3：長寬資料格式整合匯入

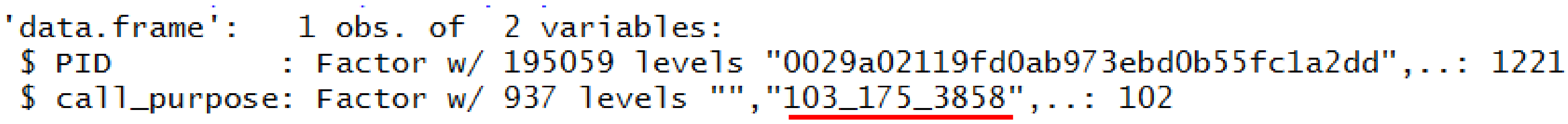


```
                               PID 2087-02-25
1 00001861a94c52d57aaa71e100f82cff          Y
2 000063166ccc4095e37ebfade31a19bd          Y
3 000073fb691e8004b1b7716b209fdd23          Y
4 0000f44306588ca57b30743bce9c329a          Y
5 0001f1a4bce7cc913edcfe96a9d9ce67          Y
6 0001f38c75a9ad46aace0ffc3ab8091f          Y
                               PID 2087-03-26
1 00001861a94c52d57aaa71e100f82cff          Y
2 000063166ccc4095e37ebfade31a19bd          Y
3 000073fb691e8004b1b7716b209fdd23          Y
4 0000f44306588ca57b30743bce9c329a          Y
5 0001f1a4bce7cc913edcfe96a9d9ce67          Y
6 0001f38c75a9ad46aace0ffc3ab8091f          Y
                               PID 2088-02-26
1 00001861a94c52d57aaa71e100f82cff          Y
2 000063166ccc4095e37ebfade31a19bd          Y
3 000073fb691e8004b1b7716b209fdd23          Y
4 0000f44306588ca57b30743bce9c329a          Y
5 0001f1a4bce7cc913edcfe96a9d9ce67          Y
6 0001f38c75a9ad46aace0ffc3ab8091f          Y
```

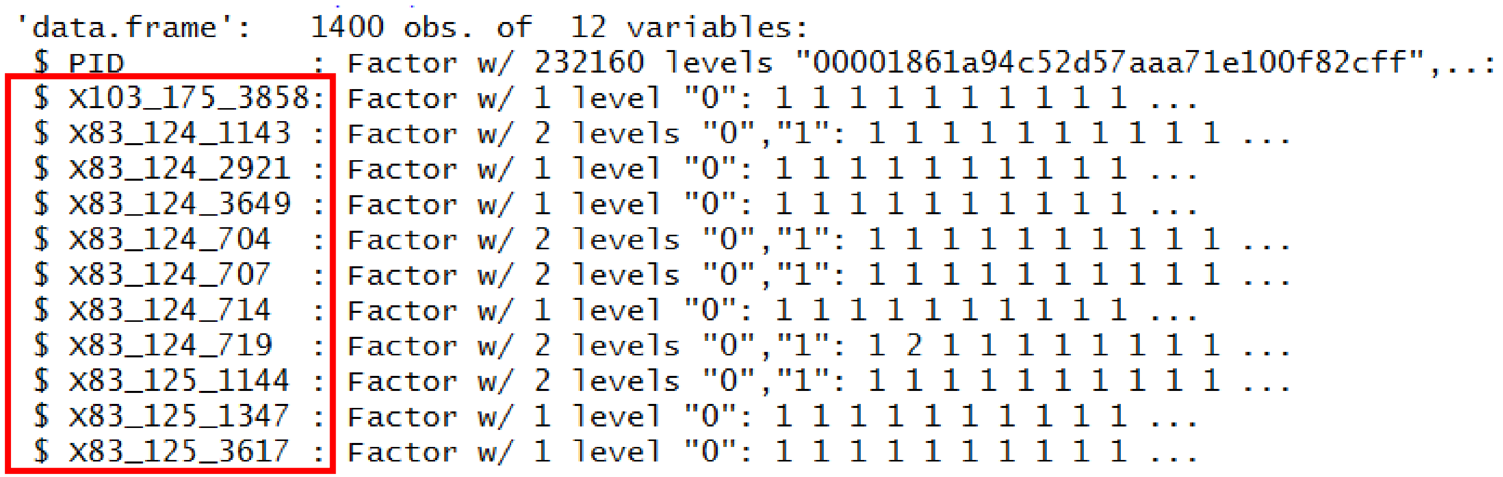


```
                               PID 2087-02-25 2087-03-26 2087-04-26 2087-05-26 2087-06-26 2087-07-26 2087-08-26 2087-09-26 2087-10-26 2087-11-26 2087-12-26 2088-01-26 2088-02-26 N_of_NoCard
1 00001861a94c52d57aaa71e100f82cff          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
2 000063166ccc4095e37ebfade31a19bd          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
3 000073fb691e8004b1b7716b209fdd23          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
4 0000f44306588ca57b30743bce9c329a          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
5 0001f1a4bce7cc913edcfe96a9d9ce67          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
6 0001f38c75a9ad46aace0ffc3ab8091f          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y          Y           0
```

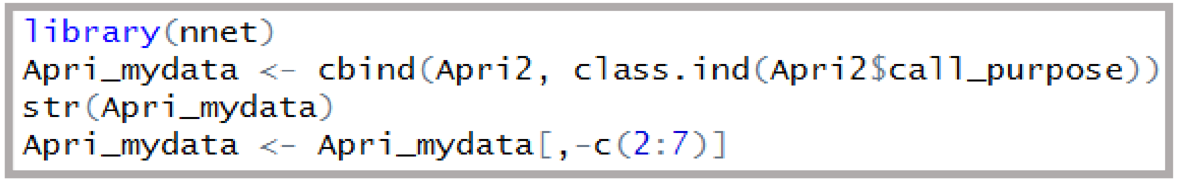
資料清理4：大量多值欄位轉換虛擬變數



```
'data.frame':   1 obs. of  2 variables:
 $ PID         : Factor w/ 195059 levels "0029a02119fd0ab973ebd0b55fc1a2dd",..: 1221
 $ call_purpose: Factor w/ 937 levels "","103_175_3858",..: 102
```

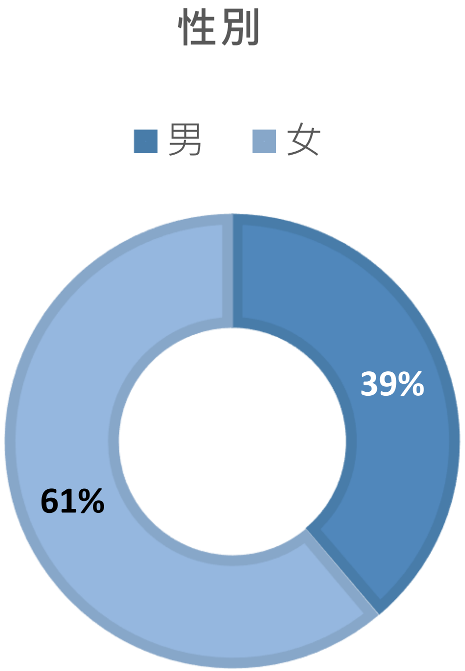


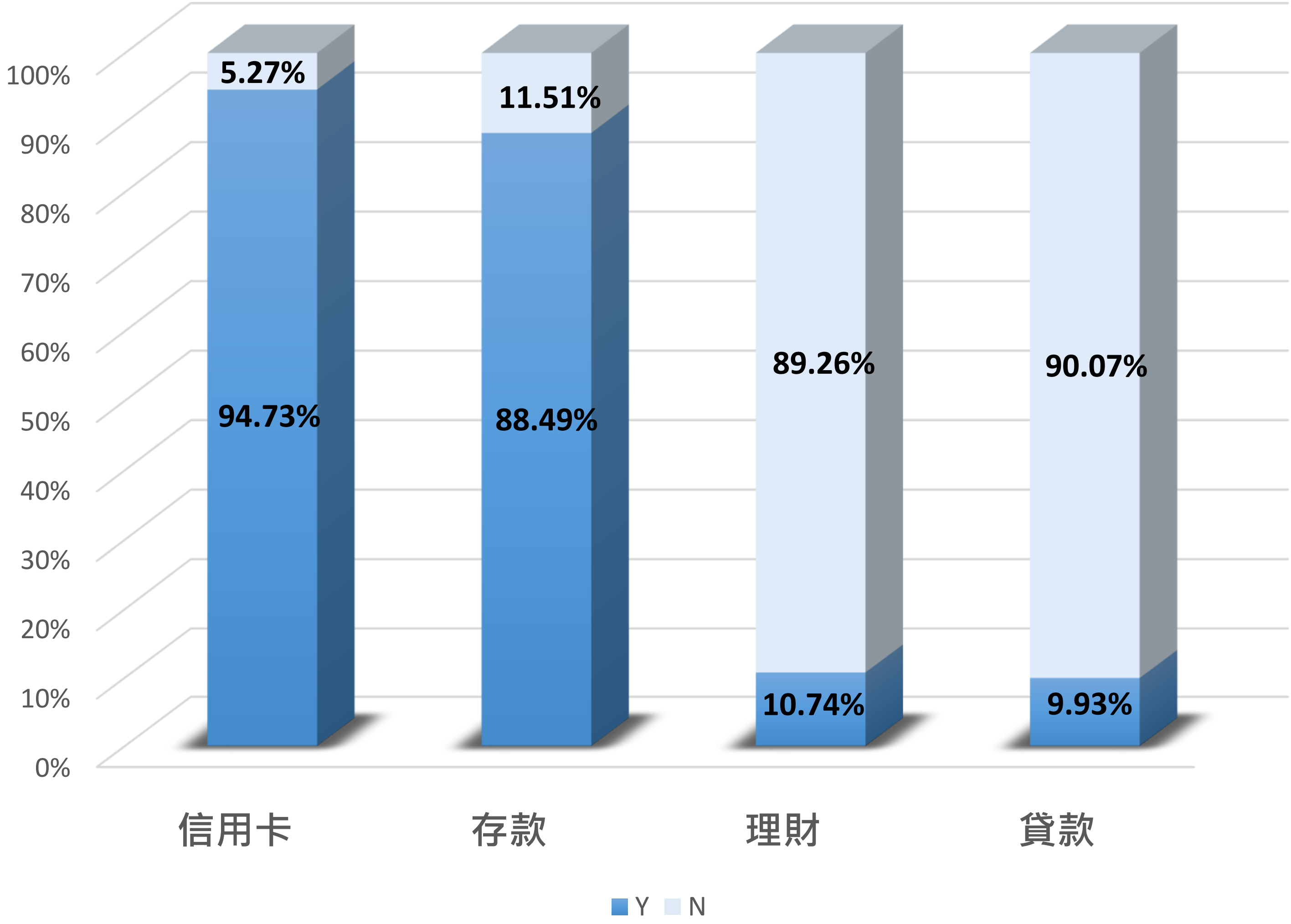
```
'data.frame':   1400 obs. of  12 variables:
 $ PID          : Factor w/ 232160 levels "00001861a94c52d57aaa71e100f82cff",..:
 $ X103_175_3858: Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_1143 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_2921 : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_3649 : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_704  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_707  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_714  : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_124_719  : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ X83_125_1144 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_125_1347 : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ X83_125_3617 : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
```



```
library(nnet)
Apri_mydata <- cbind(Apri2, class.ind(Apri2$call_purpose))
str(Apri_mydata)
Apri_mydata <- Apri_mydata[,-c(2:7)]
```

資料探索：資料摘要

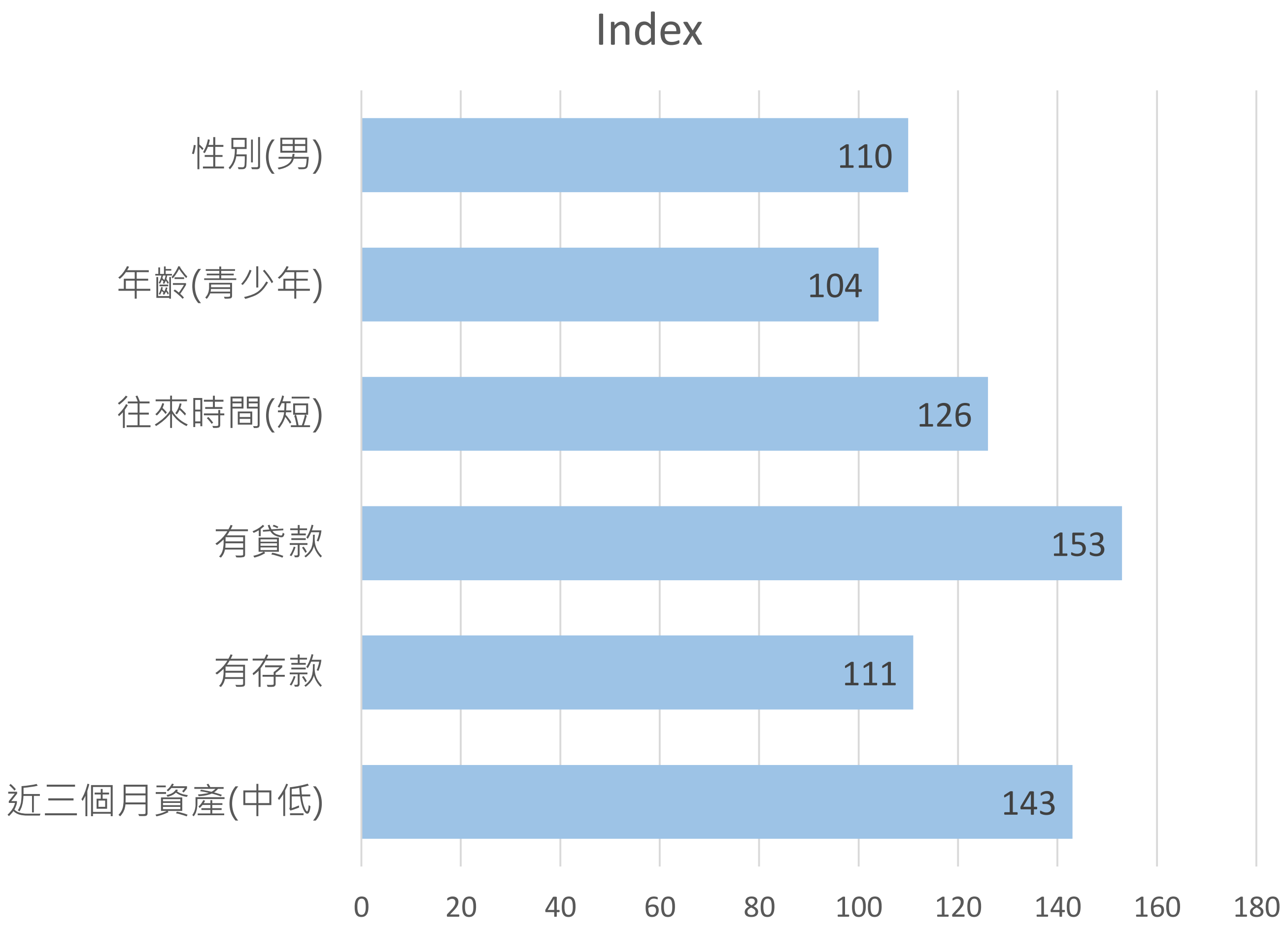

性別
男
女
39%
61%


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
5.27%
94.73%
11.51%
88.49%
89.26%
10.74%
90.07%
9.93%
信用卡
存款
理財
貸款
Y
N

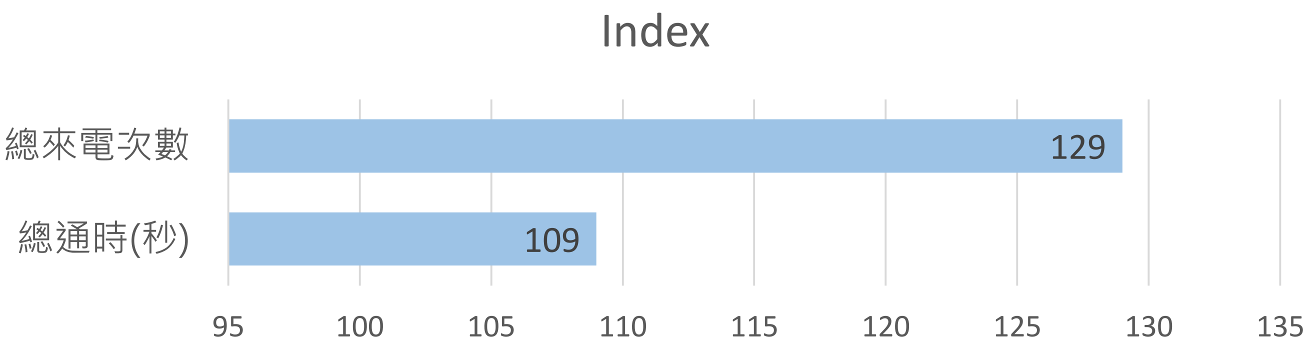
資料探索：客戶型態傾向性比較

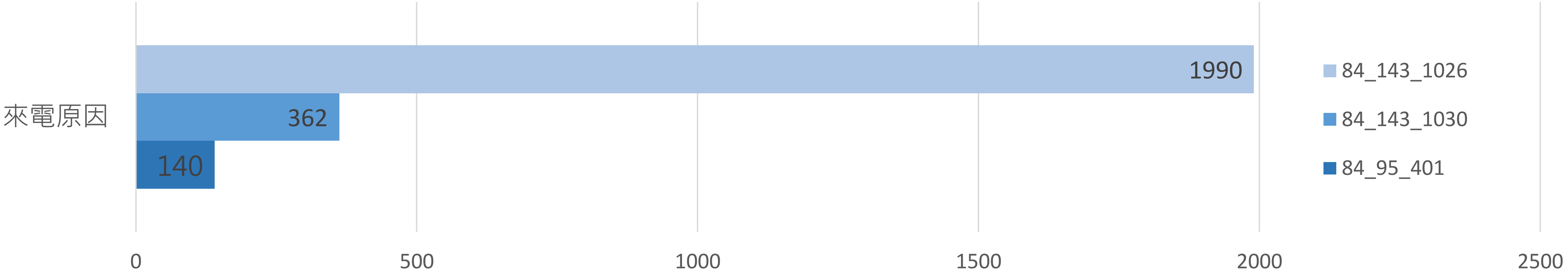
- Index 代表流失傾向越高

	流失客戶	存活客戶	index
有貸款	15.7%	10.3%	153
無貸款	84.3%	89.7%	94
total	2,712	216,511	

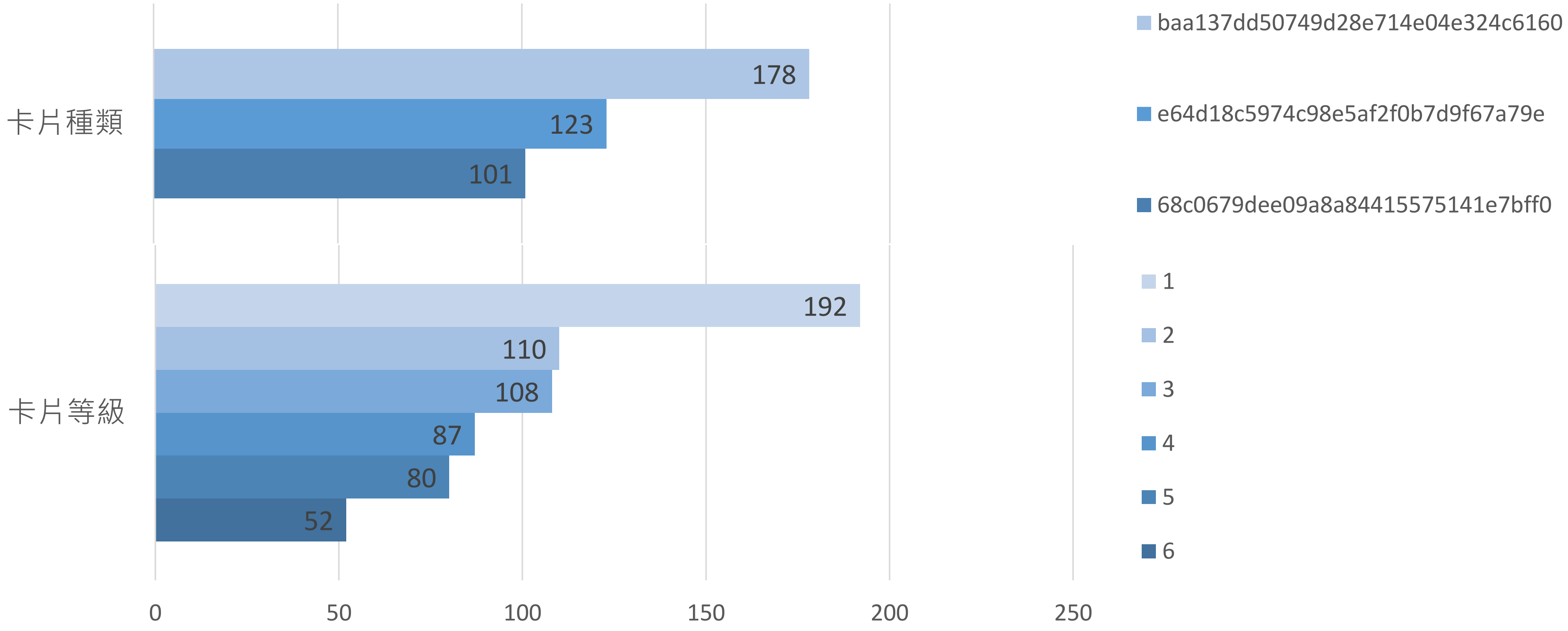


資料探索：往來關係傾向性比較



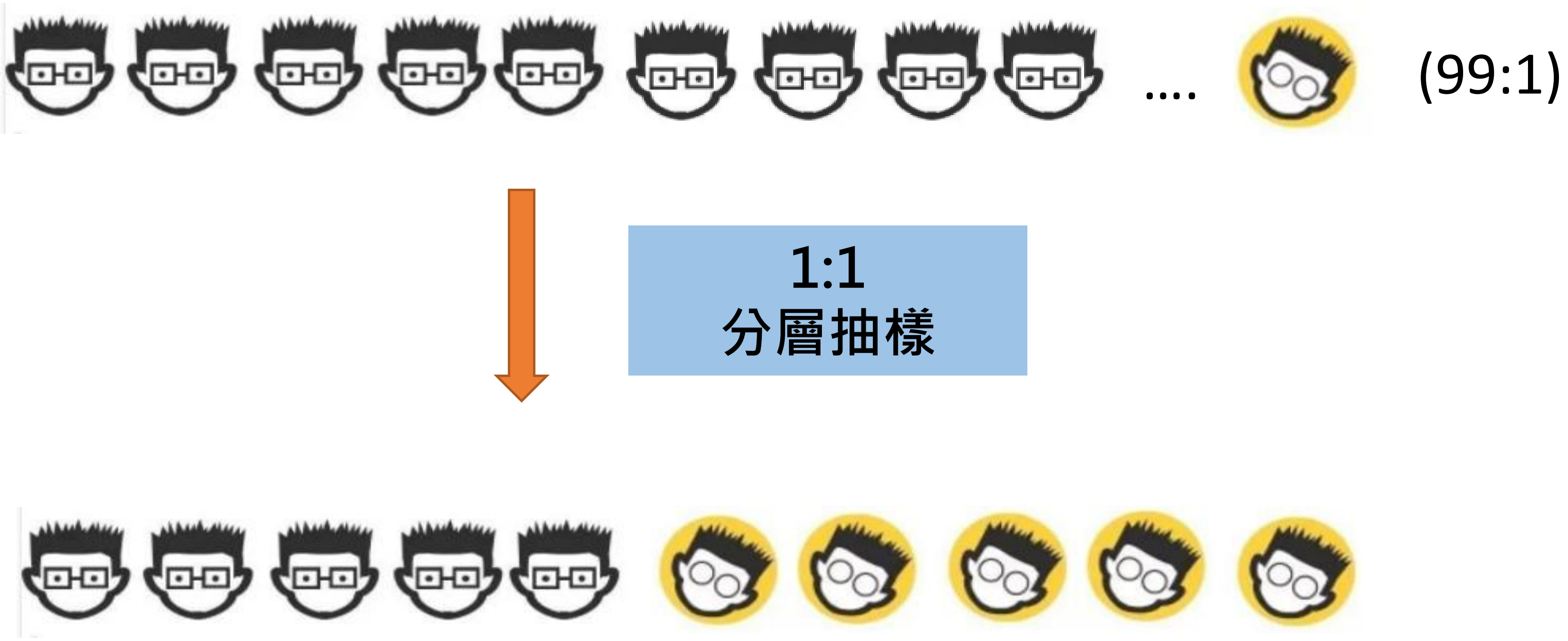


資料探索：消費型態傾向性比較



模型建立：樣本權重調整

以流失客戶樣本數為基準，從存活客戶中隨機抽出等筆資料（即調整成1:1的比例）



....

(99:1)

1:1
分層抽樣

模型建立：分類模型

決策樹

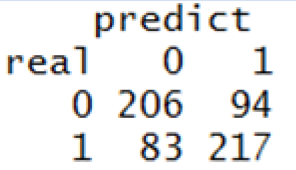
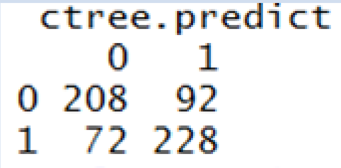
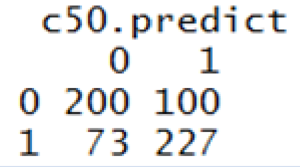
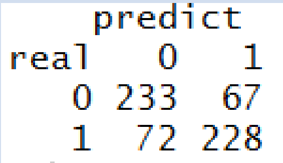
- 結果直觀視覺化
- 過程需修剪樹分支較為繁瑣
- 可能過度配適

邏輯斯迴歸

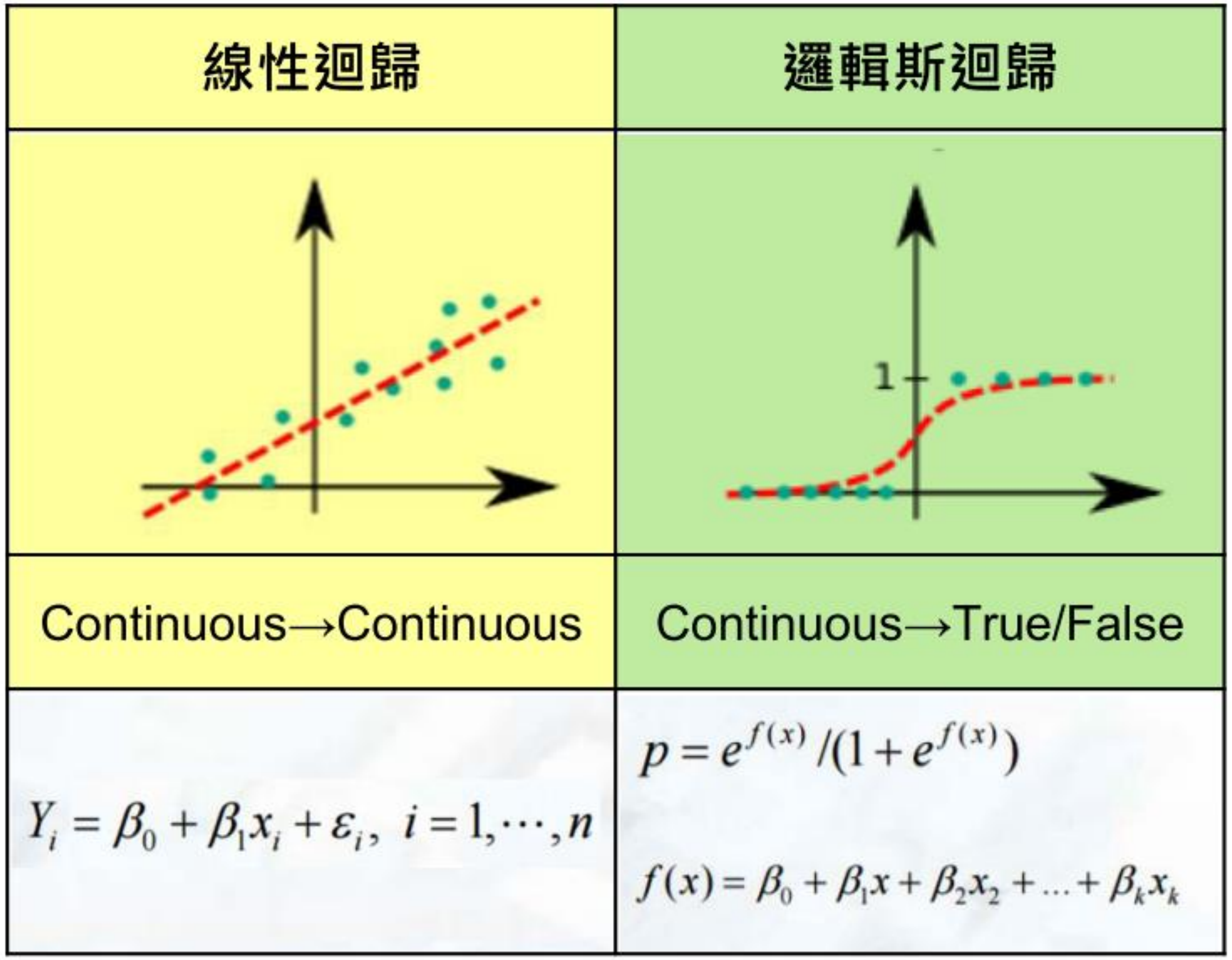
- 可作統計顯著性檢定
- 可模擬參數實質效果
- 可能過度配適(變數數量過多時)

模型建立：決策樹效果比較

★

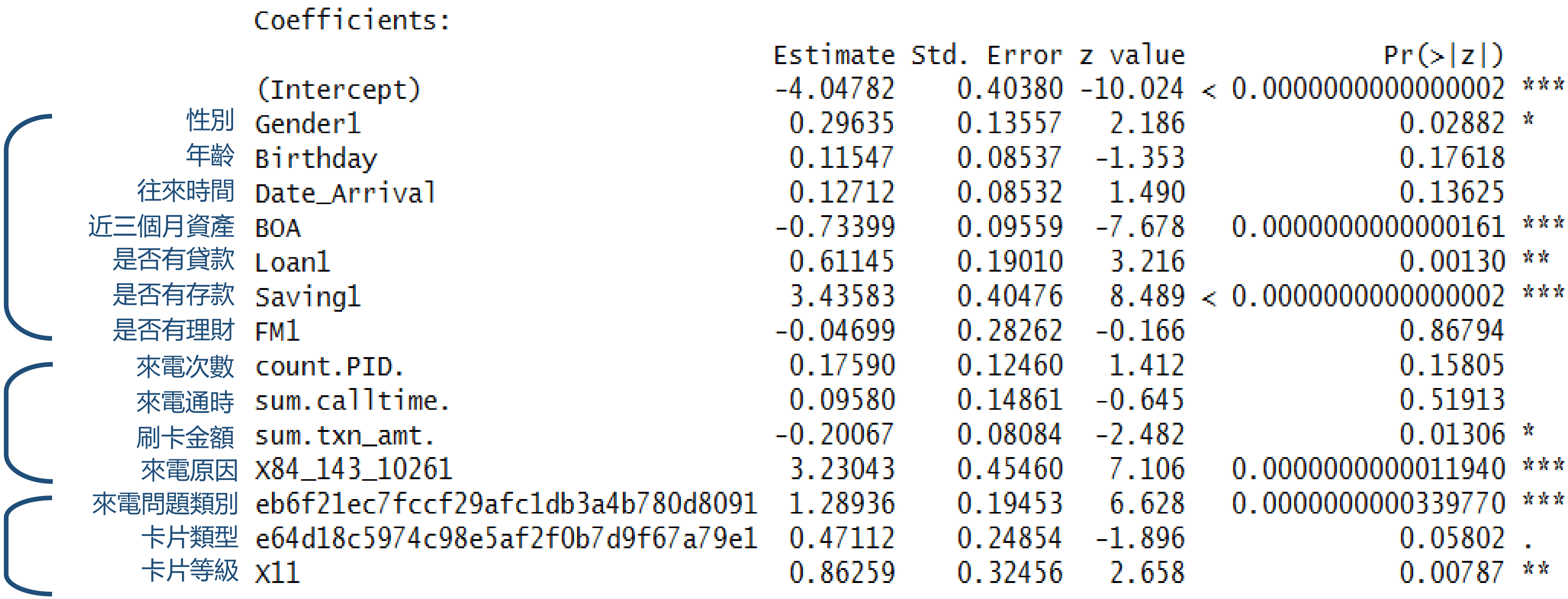
	Rpart	Ctree	C50	Random Forests
	 predict real 0 1 0 206 94 1 83 217	 ctree.predict 0 1 0 208 92 1 72 228	 c50.predict 0 1 0 200 100 1 73 227	 predict real 0 1 0 233 67 1 72 228
準確率 Accuracy	70.5%	72.7%	71.2%	76.83%
精確率 Precision	69.8%	71.3%	69.4%	77.29%
召回率 Recall	72.3%	76.0%	75.7%	76.39%

模型建立：邏輯斯迴歸



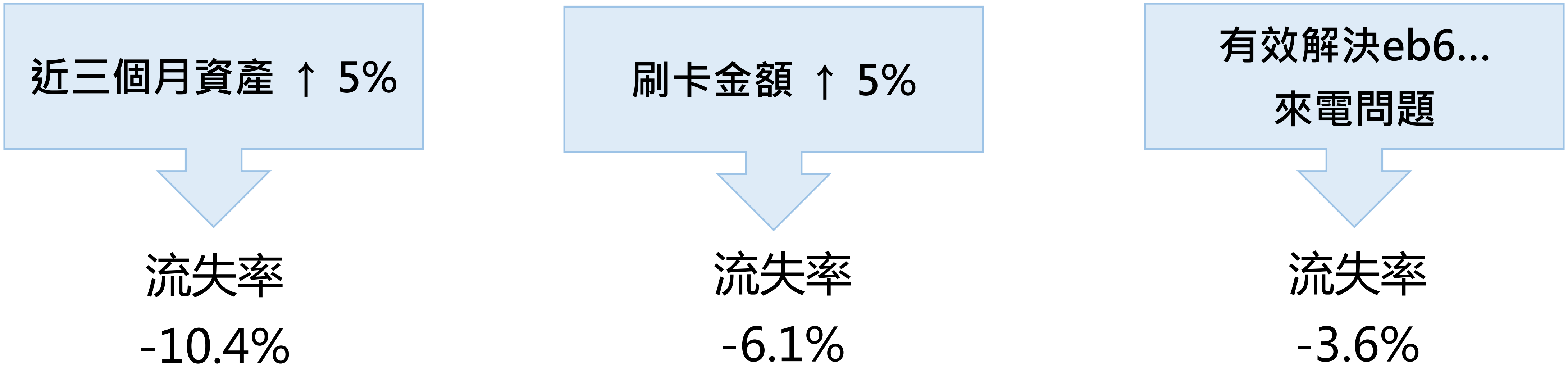
線性迴歸	邏輯斯迴歸
Continuous→Continuous	Continuous→True/False
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,\cdots,n$	$p = e^{f(x)}/(1+e^{f(x)})$ $f(x) = \beta_0 + \beta_1 x + \beta_2 x_2 + \ldots + \beta_k x_k$

模型建立：邏輯斯迴歸顯著性比較



```
                  Coefficients:
                                                        Estimate Std. Error z value           Pr(>|z|)
                  (Intercept)                           -4.04782    0.40380 -10.024 < 0.0000000000000002 ***
          性別    Gender1                                0.29635    0.13557   2.186              0.02882 *
          年齡    Birthday                               0.11547    0.08537  -1.353              0.17618
      往來時間    Date_Arrival                           0.12712    0.08532   1.490              0.13625
  近三個月資產    BOA                                   -0.73399    0.09559  -7.678   0.0000000000000161 ***
    是否有貸款    Loan1                                  0.61145    0.19010   3.216              0.00130 **
    是否有存款    Saving1                                3.43583    0.40476   8.489 < 0.0000000000000002 ***
    是否有理財    FM1                                   -0.04699    0.28262  -0.166              0.86794
      來電次數    count.PID.                             0.17590    0.12460   1.412              0.15805
      來電通時    sum.calltime.                          0.09580    0.14861  -0.645              0.51913
      刷卡金額    sum.txn_amt.                          -0.20067    0.08084  -2.482              0.01306 *
      來電原因    X84_143_10261                          3.23043    0.45460   7.106   0.0000000000011940 ***
  來電問題類別    eb6f21ec7fccf29afc1db3a4b780d8091      1.28936    0.19453   6.628   0.0000000000339770 ***
      卡片類型    e64d18c5974c98e5af2f0b7d9f67a79e1      0.47112    0.24854  -1.896              0.05802 .
      卡片等級    X11                                    0.86259    0.32456   2.658              0.00787 **
```

模型建立：參數效果模擬



近三個月資產 ↑ 5%

流失率

-10.4%

刷卡金額 ↑ 5%

流失率

-6.1%

有效解決eb6...
來電問題

流失率

-3.6%

共降低20.1%客戶流失機率