金融大數據

銀行定期存款推銷預測

組員：107AB8001 蔡雯惠　107AB8004 陳鴻妮
107AB8002 蕭家希　107AB8005 李予蒨
107AB8003 林芝儀　107AB8406 袁嘉妮

2019.01.02

Content

01 資料集簡介

02 分析目的

03 資料清理過程

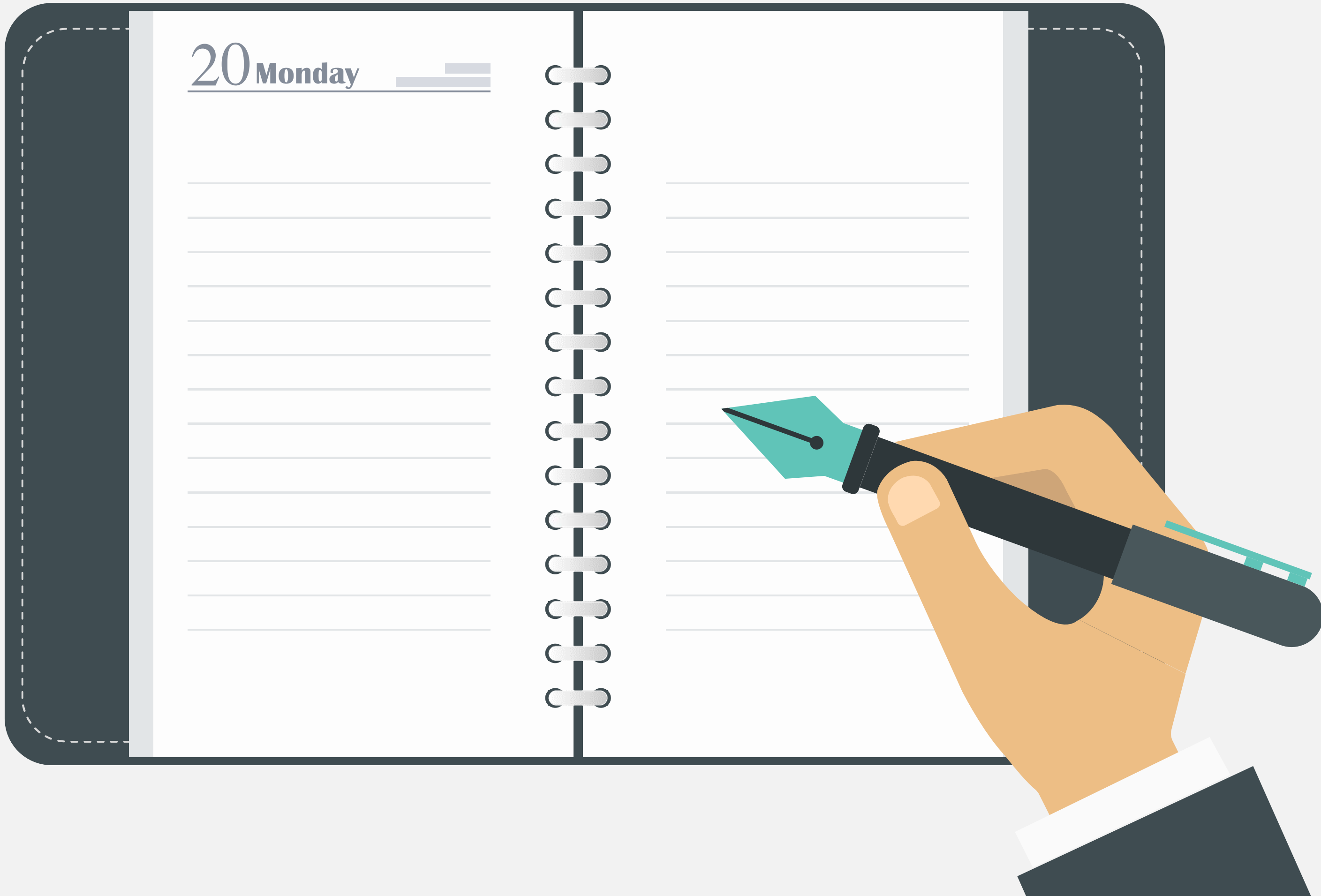
04 模型建構與驗證

05 結論與未來改善方向

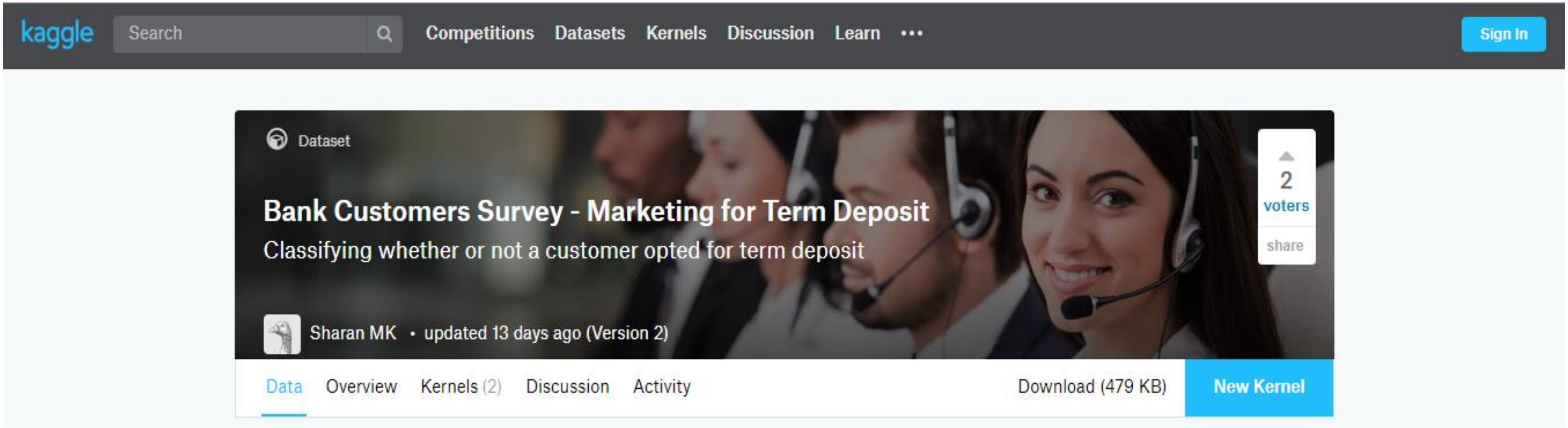
06 組員工作分配

01

資料集簡介



資料集簡介



Bank Customers Survey - Marketing for Term Deposit

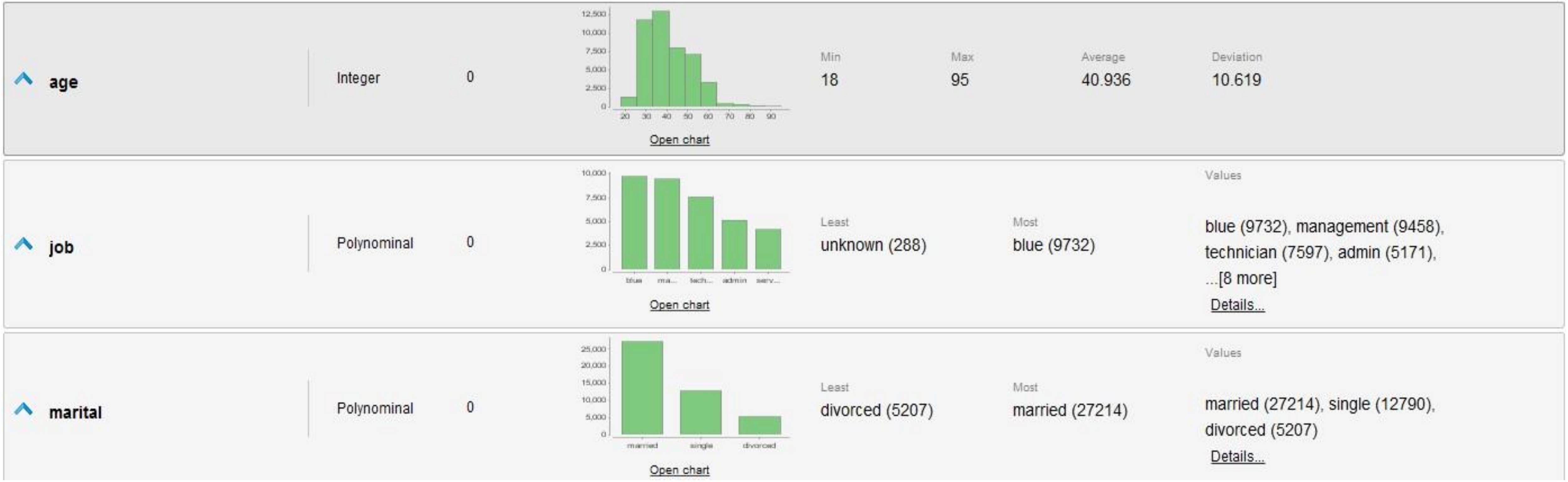
Classifying whether or not a customer opted for term deposit

Sharan MK • updated 13 days ago (Version 2)

資料集名稱：Bank Customers Survey – Marketing for Term Deposit

資料集來源：Kaggle 網站

資料集分佈



age Integer 0

Open chart

Min 18 Max 95 Average 40.936 Deviation 10.619

job Polynominal 0

Open chart

Least unknown (288) Most blue (9732)

Values

blue (9732), management (9458), technician (7597), admin (5171), ...[8 more]

Details...

marital Polynominal 0

Open chart

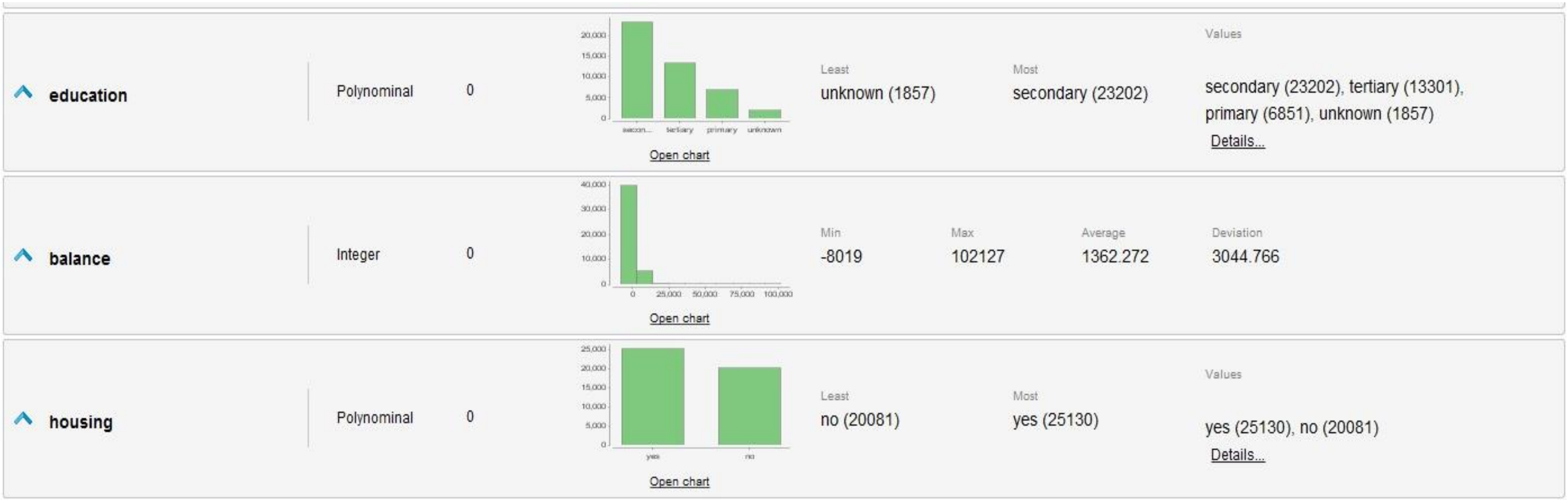
Least divorced (5207) Most married (27214)

Values

married (27214), single (12790), divorced (5207)

Details...

資料集分佈



education Polynominal 0

Least: unknown (1857)

Most: secondary (23202)

Values: secondary (23202), tertiary (13301), primary (6851), unknown (1857)
Details...

balance Integer 0

Min: -8019

Max: 102127

Average: 1362.272

Deviation: 3044.766

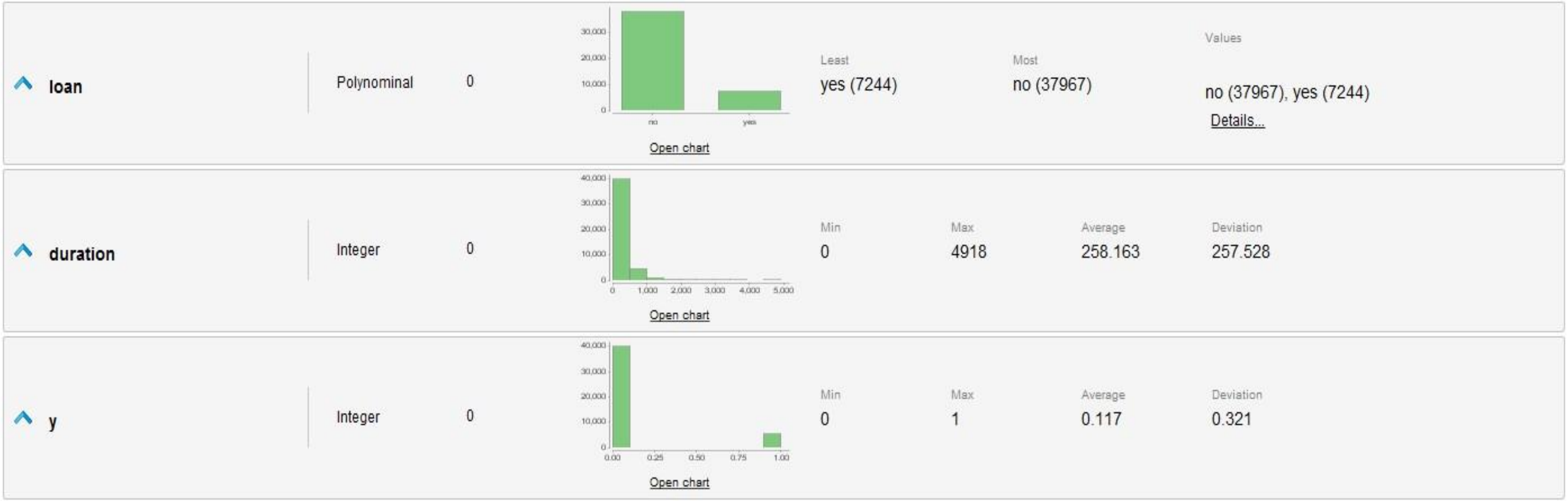
housing Polynominal 0

Least: no (20081)

Most: yes (25130)

Values: yes (25130), no (20081)
Details...

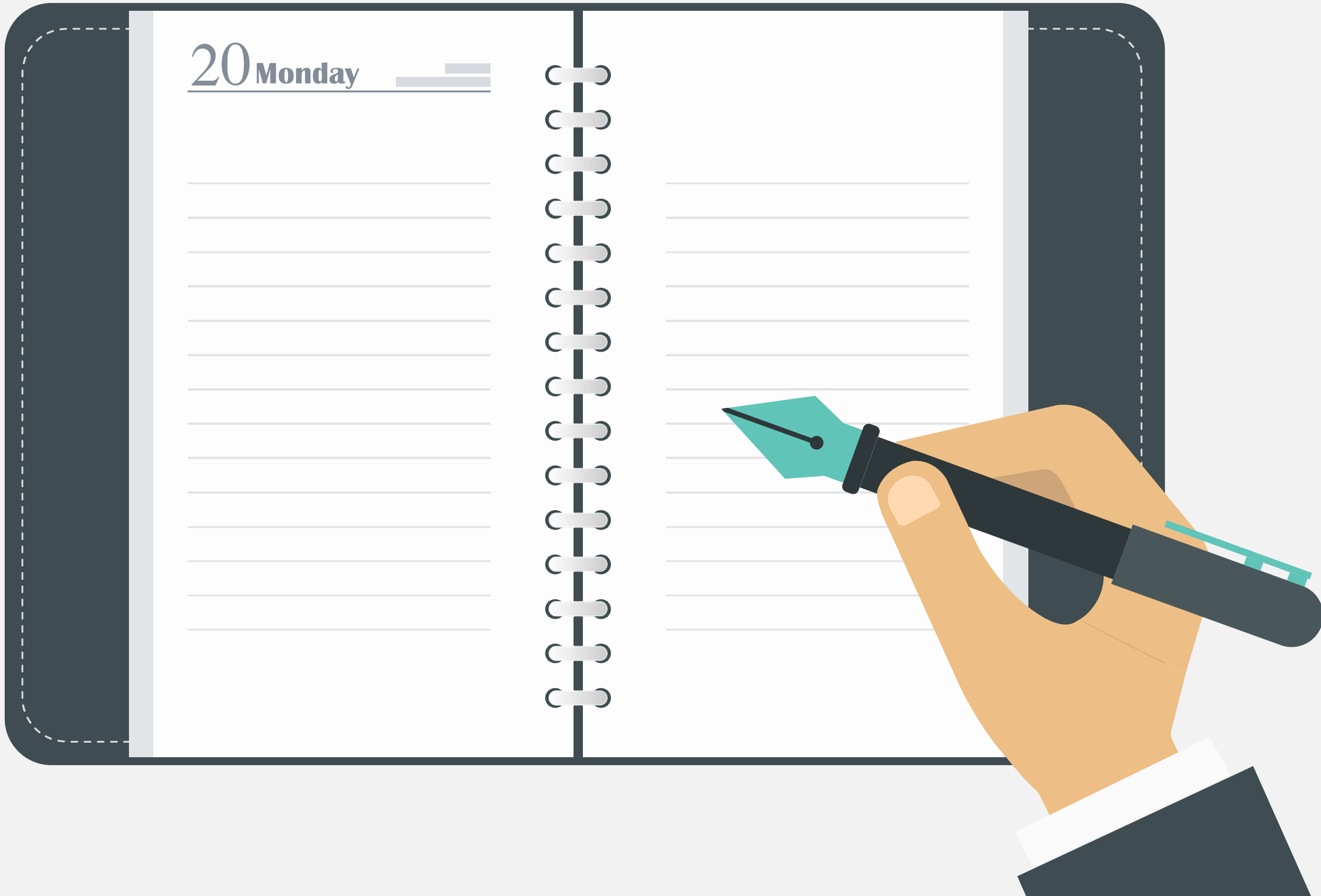
資料集分佈



loan	Polynominal	0	Open chart	Least yes (7244)	Most no (37967)		Values no (37967), yes (7244) Details...
duration	Integer	0	Open chart	Min 0	Max 4918	Average 258.163	Deviation 257.528
y	Integer	0	Open chart	Min 0	Max 1	Average 0.117	Deviation 0.321

02

分析目的



分析目的

分析方法

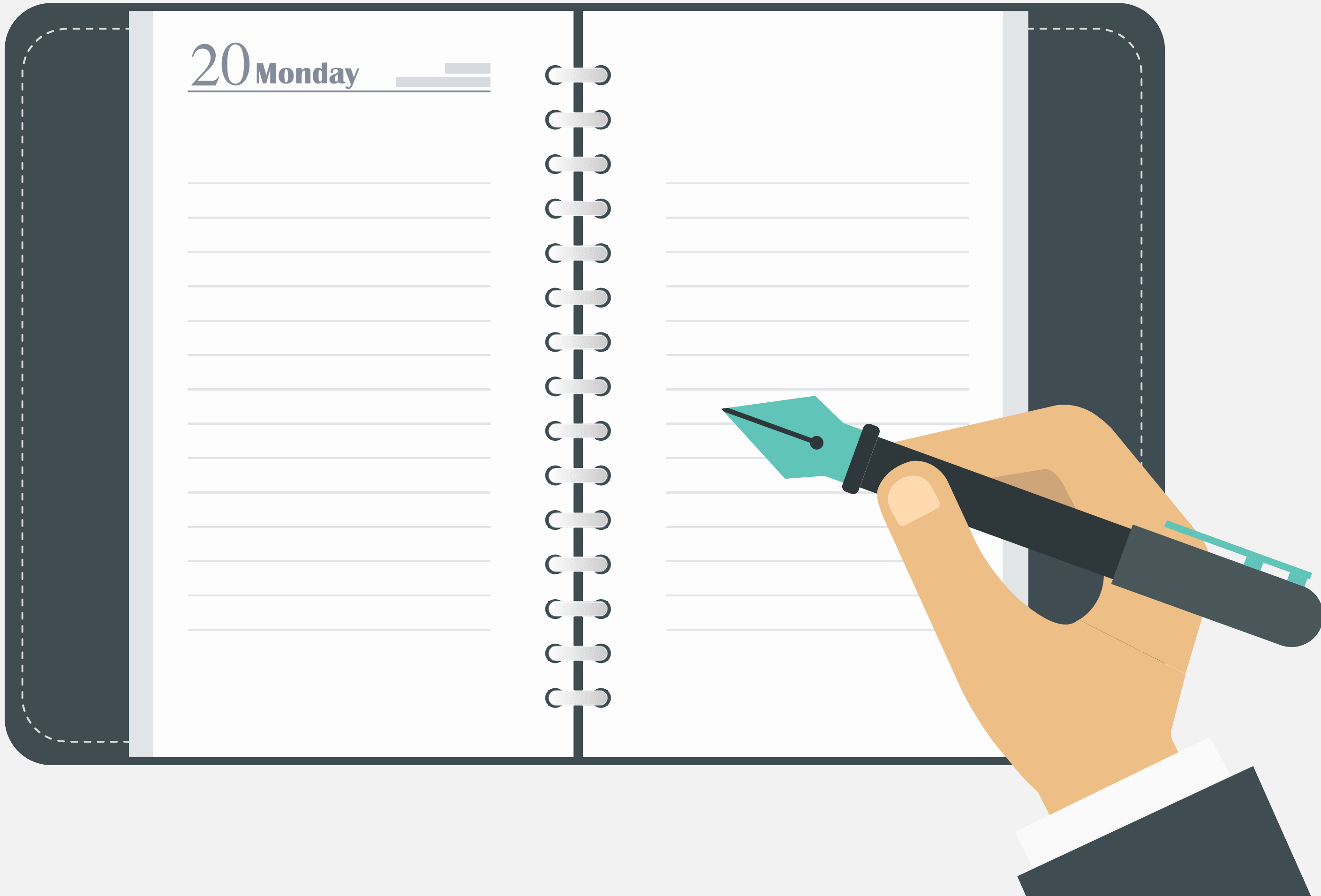
隨機森林(Random Forest)

分析目的

主要是利用Bank Customers Survey – Marketing for Term Deposit資料集，將挑選出來的屬性透過隨機森林的方法去預測顧客是否被推銷成功。

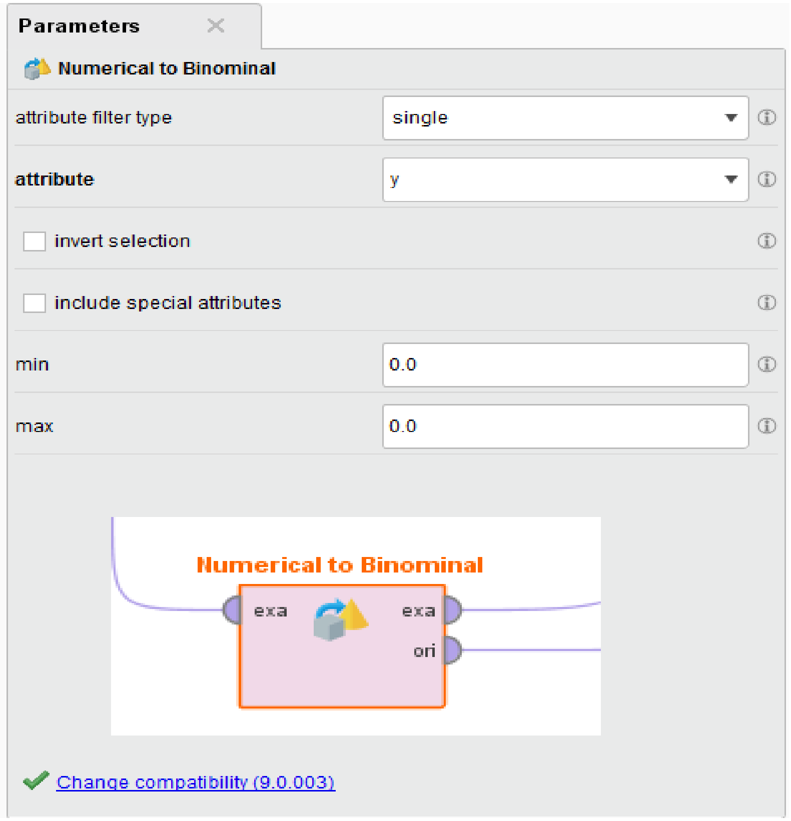
03

資料清理過程



資料清理、轉換

- 將欄位y轉換成類別資料，再將此資料指定為要分析的欄位



Parameters ×

Numerical to Binominal

attribute filter type: single

attribute: y

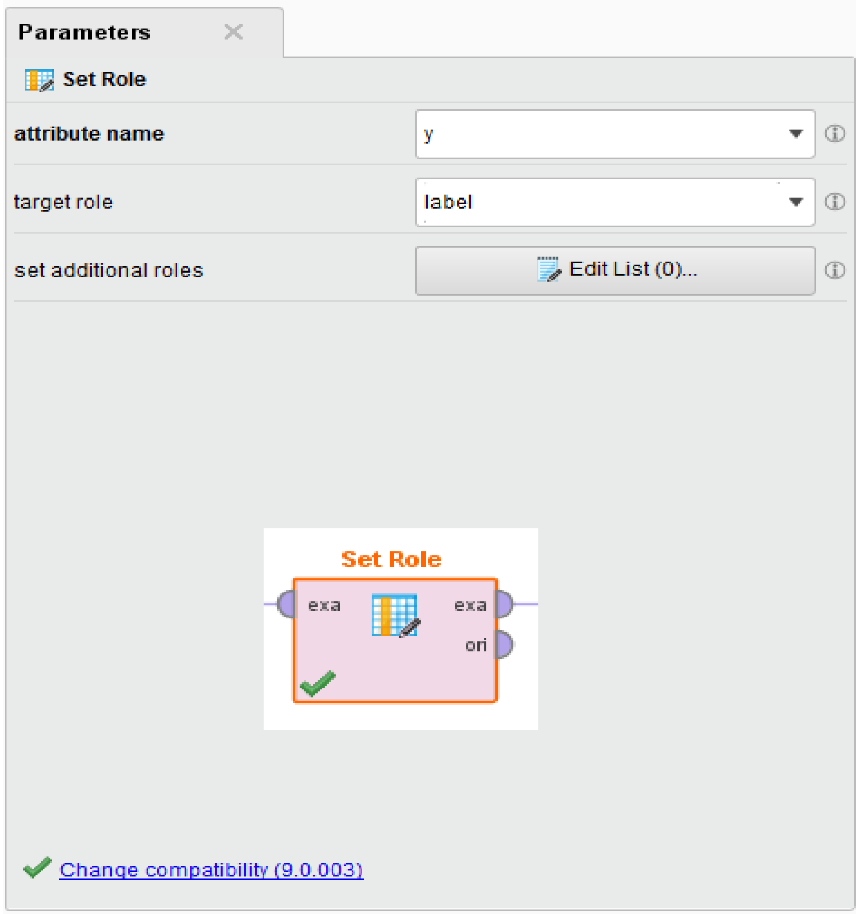
□ invert selection

□ include special attributes

min: 0.0

max: 0.0

✓ Change compatibility (9.0.003)



Parameters ×

Set Role

attribute name: y

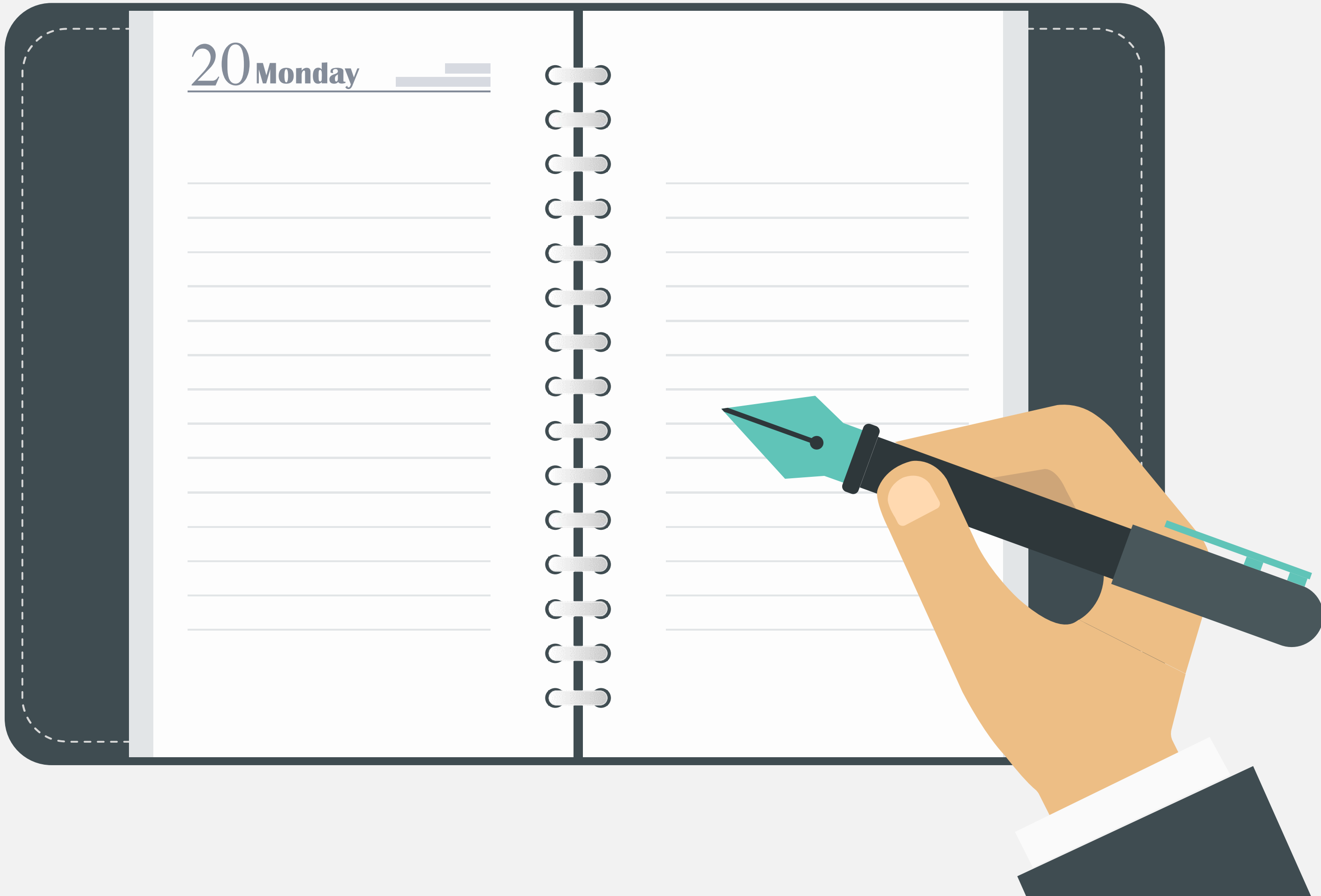
target role: label

set additional roles: Edit List (0)...

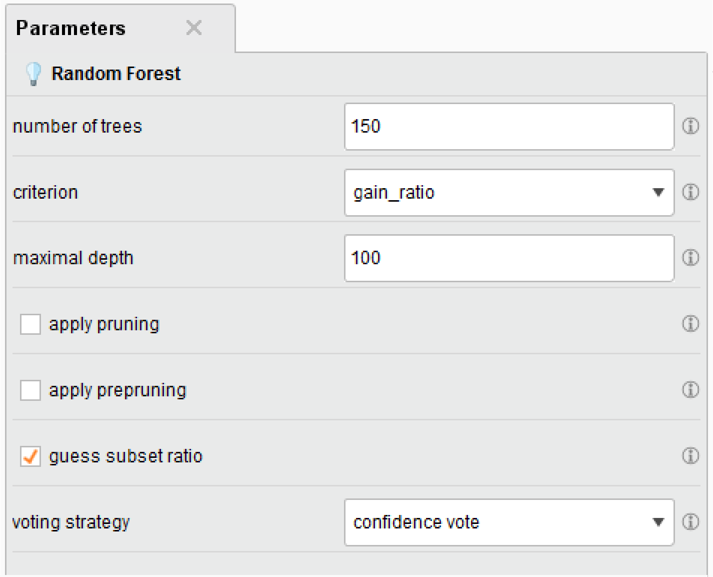
✓ Change compatibility (9.0.003)

04

模型建構與驗證



模型建構、驗證



Parameters

Random Forest

number of trees: 150

criterion: gain_ratio

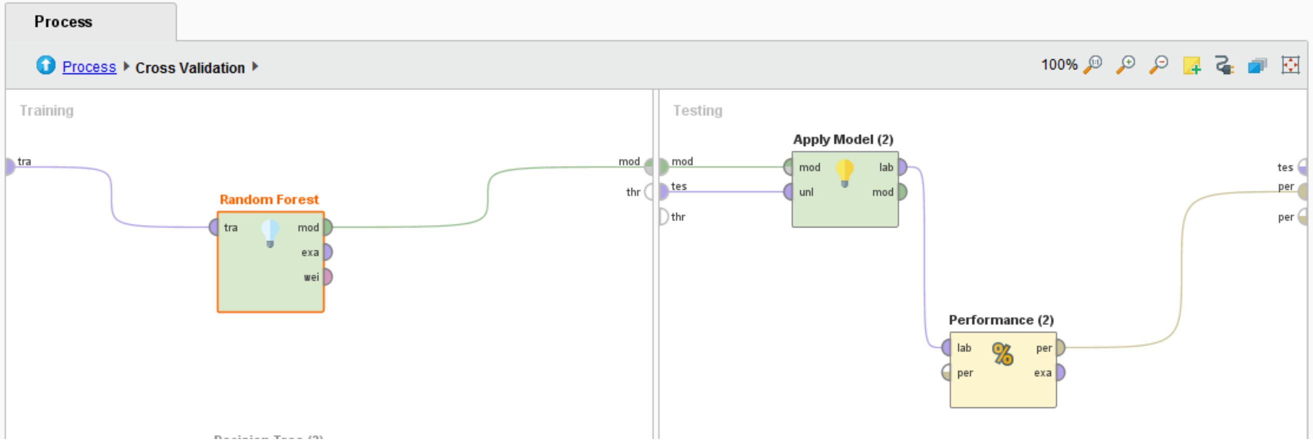
maximal depth: 100

□ apply pruning

□ apply prepruning

☑ guess subset ratio

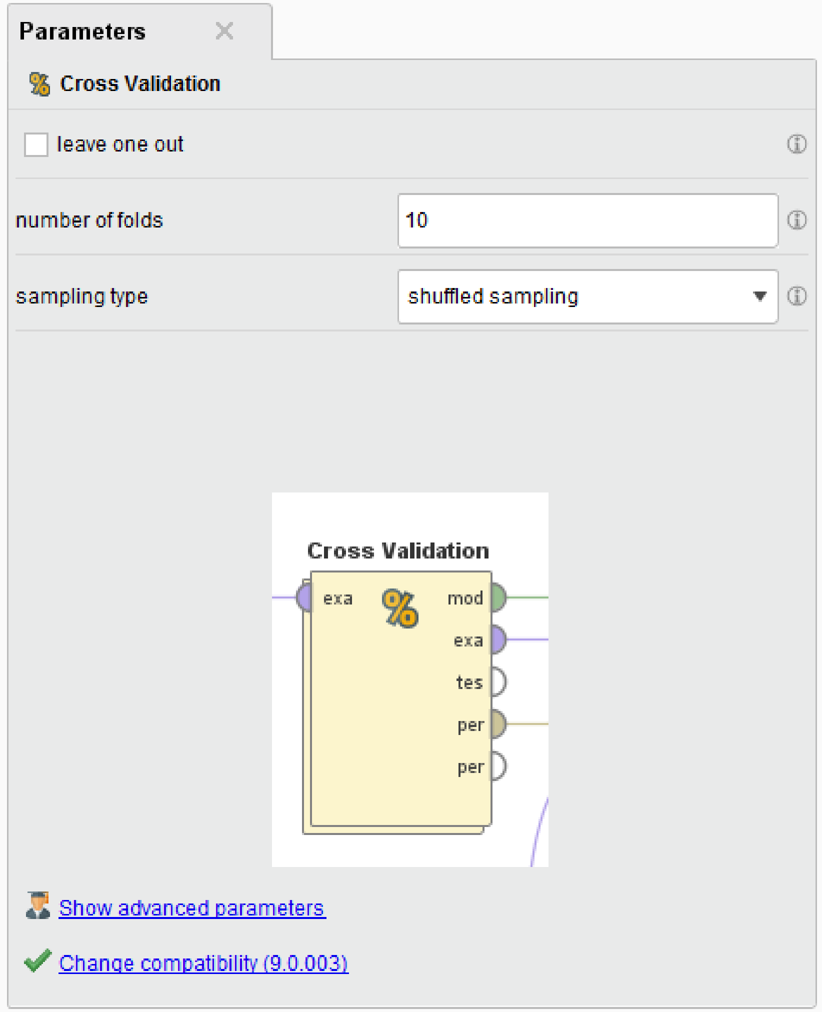
voting strategy: confidence vote



Process

Process ▸ Cross Validation ▸

100%



Parameters

Cross Validation

□ leave one out

number of folds: 10

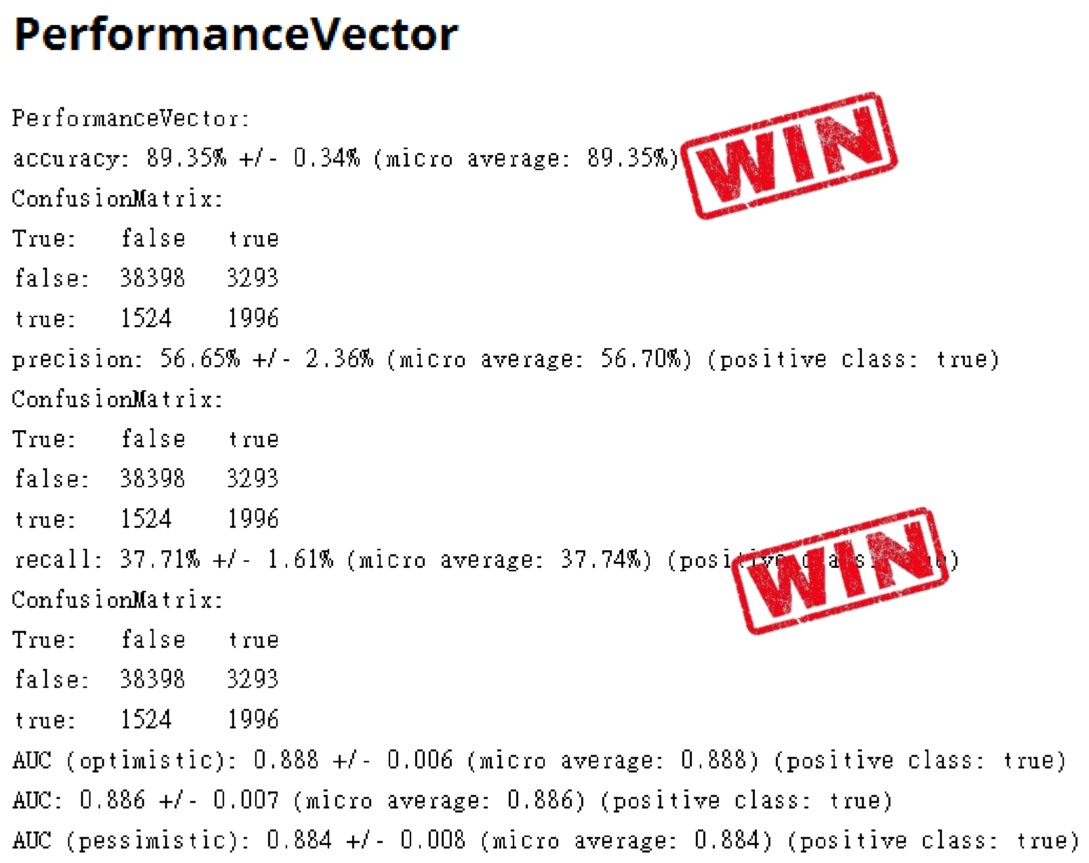
sampling type: shuffled sampling

Show advanced parameters

Change compatibility (9.0.003)

選擇Random Forest原因

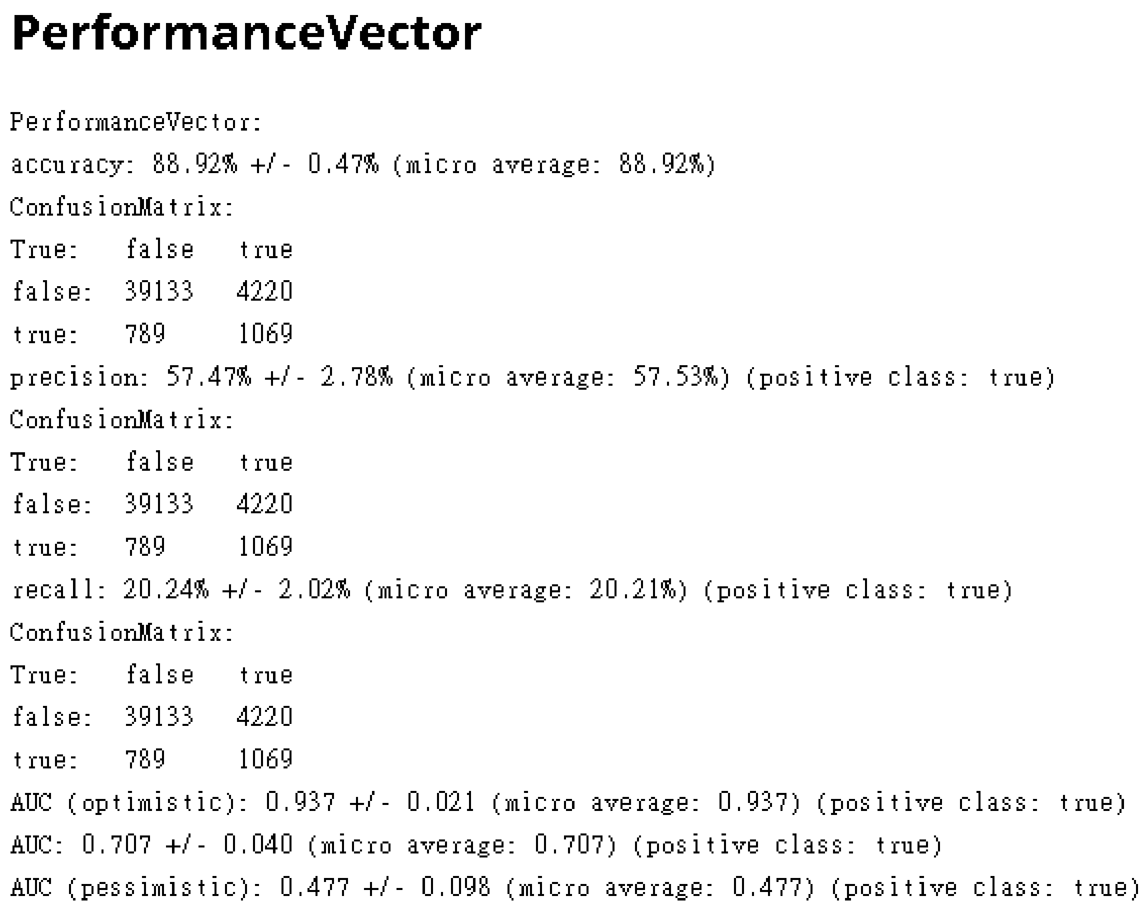
Random Forest



PerformanceVector

```
PerformanceVector:
accuracy: 89.35% +/- 0.34% (micro average: 89.35%)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
precision: 56.65% +/- 2.36% (micro average: 56.70%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
recall: 37.71% +/- 1.61% (micro average: 37.74%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
AUC (optimistic): 0.888 +/- 0.006 (micro average: 0.888) (positive class: true)
AUC: 0.886 +/- 0.007 (micro average: 0.886) (positive class: true)
AUC (pessimistic): 0.884 +/- 0.008 (micro average: 0.884) (positive class: true)
```

Decision Tree

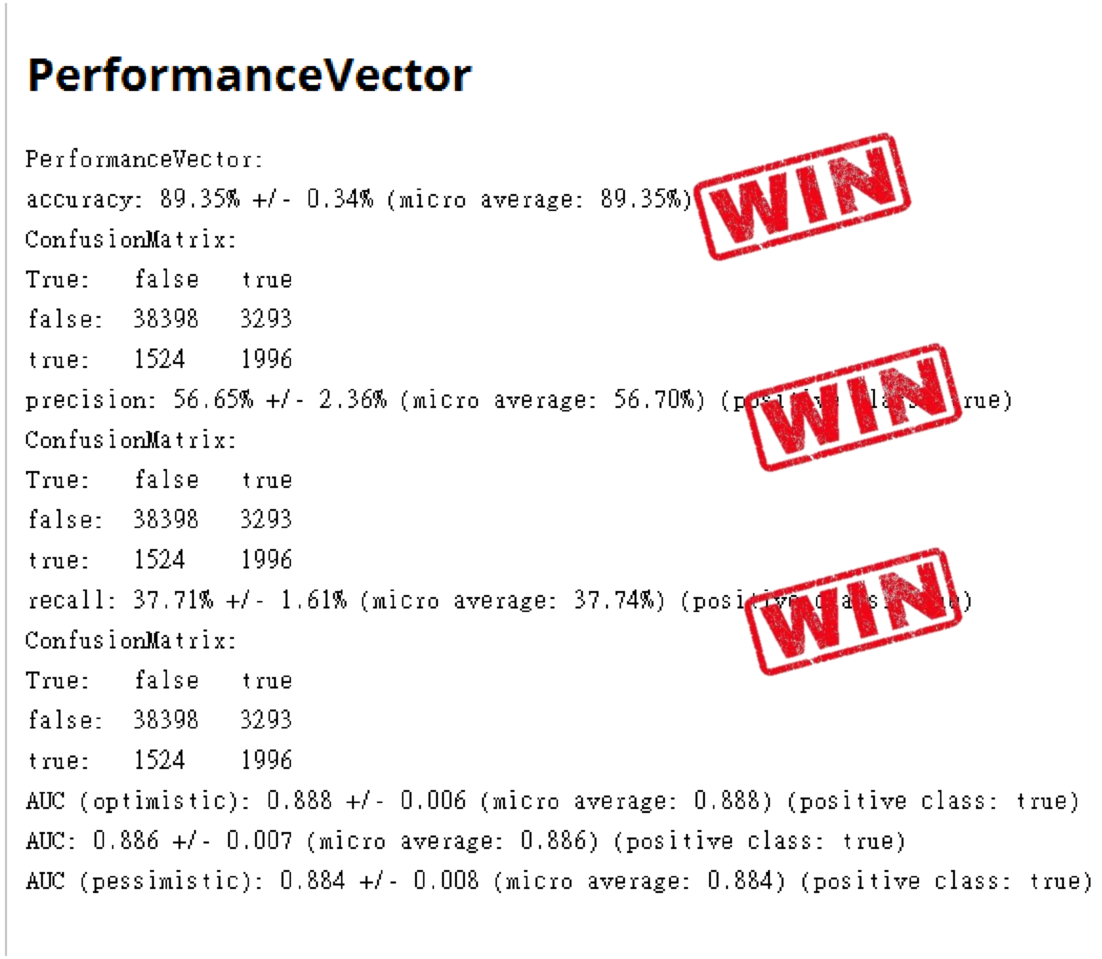


PerformanceVector

```
PerformanceVector:
accuracy: 88.92% +/- 0.47% (micro average: 88.92%)
ConfusionMatrix:
True:   false   true
false:  39133   4220
true:   789     1069
precision: 57.47% +/- 2.78% (micro average: 57.53%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  39133   4220
true:   789     1069
recall: 20.24% +/- 2.02% (micro average: 20.21%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  39133   4220
true:   789     1069
AUC (optimistic): 0.937 +/- 0.021 (micro average: 0.937) (positive class: true)
AUC: 0.707 +/- 0.040 (micro average: 0.707) (positive class: true)
AUC (pessimistic): 0.477 +/- 0.098 (micro average: 0.477) (positive class: true)
```

選擇Random Forest原因

Random Forest



PerformanceVector

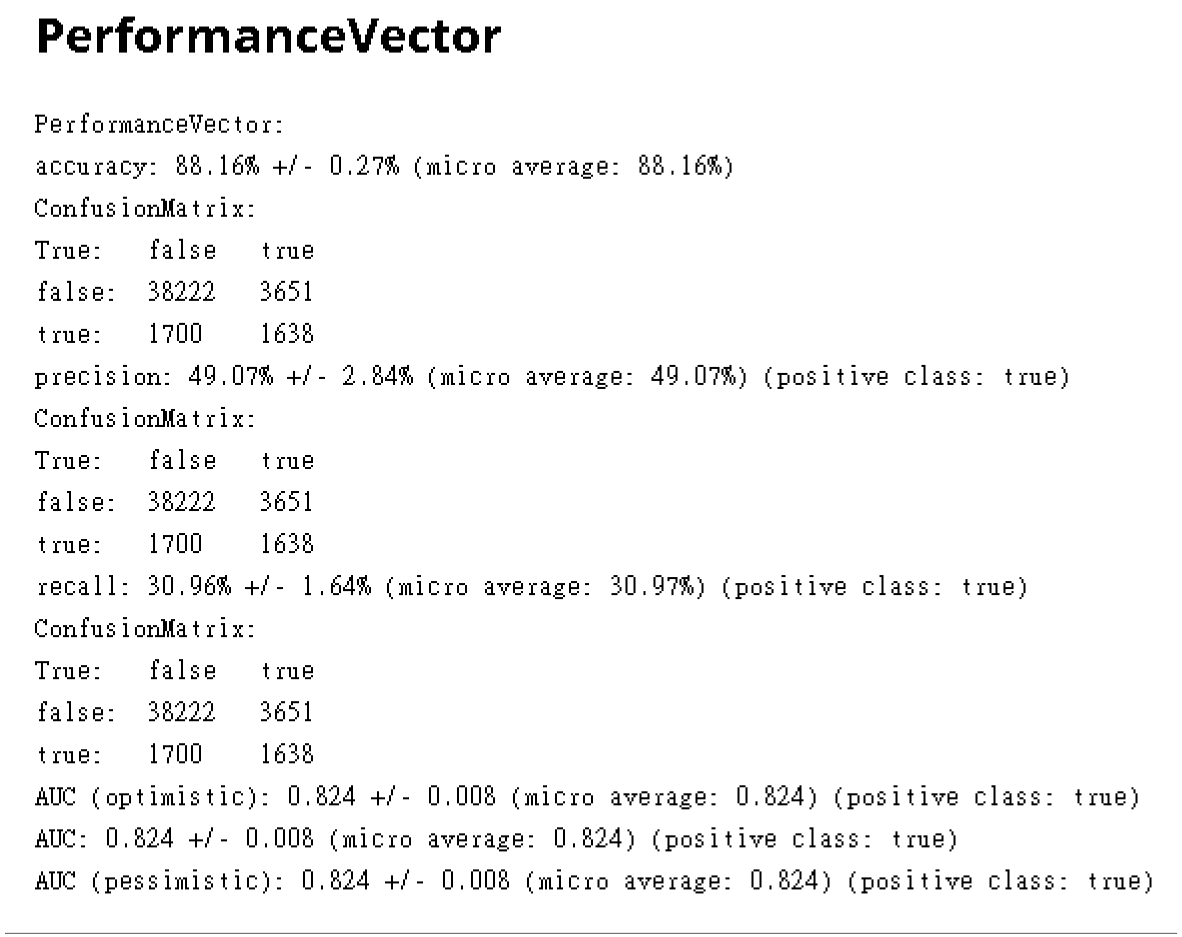
```
PerformanceVector:
accuracy: 89.35% +/- 0.34% (micro average: 89.35%)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
precision: 56.65% +/- 2.36% (micro average: 56.70%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
recall: 37.71% +/- 1.61% (micro average: 37.74%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
AUC (optimistic): 0.888 +/- 0.006 (micro average: 0.888) (positive class: true)
AUC: 0.886 +/- 0.007 (micro average: 0.886) (positive class: true)
AUC (pessimistic): 0.884 +/- 0.008 (micro average: 0.884) (positive class: true)
```

WIN

WIN

WIN

Naïve Bayes

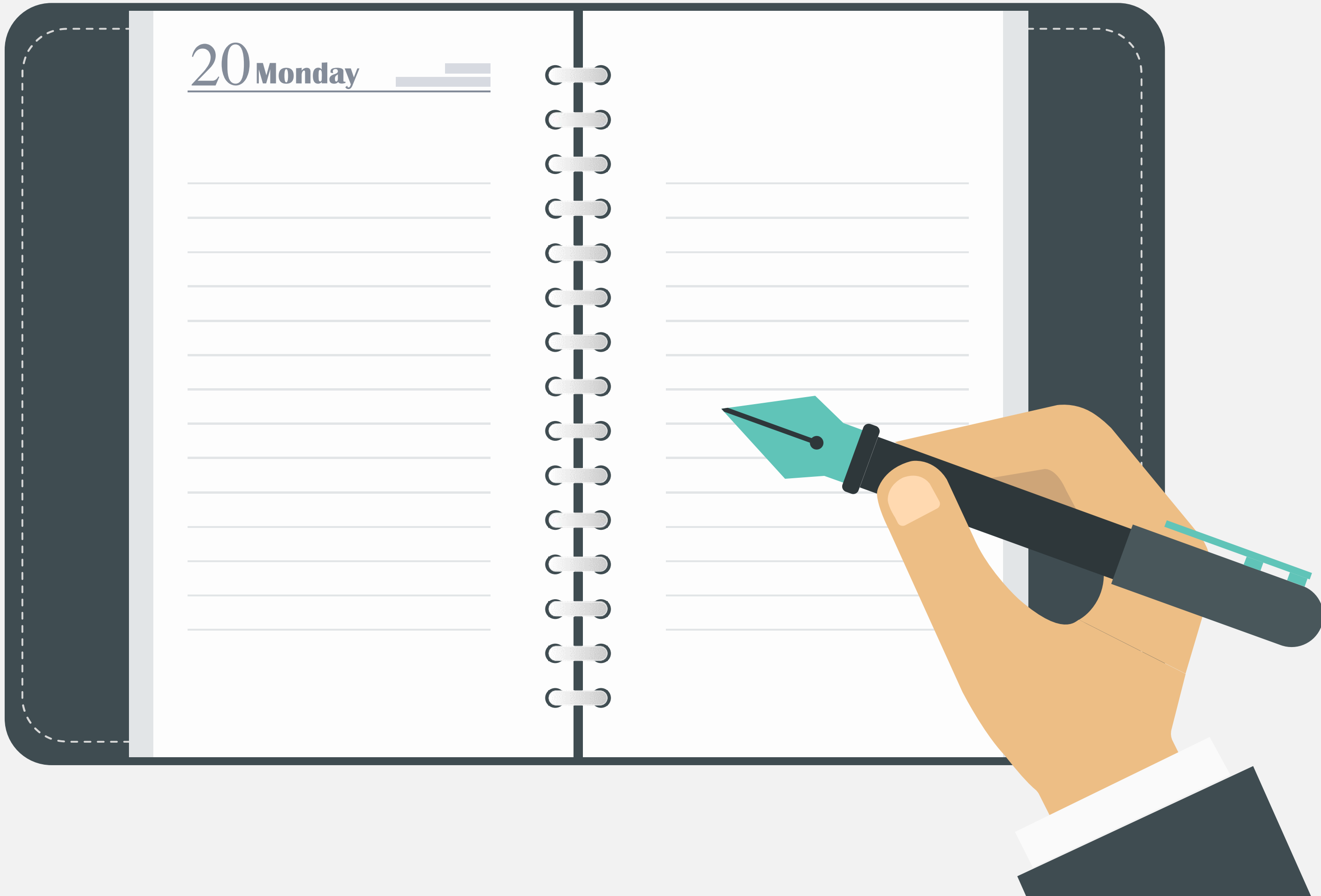


PerformanceVector

```
PerformanceVector:
accuracy: 88.16% +/- 0.27% (micro average: 88.16%)
ConfusionMatrix:
True:   false   true
false:  38222   3651
true:   1700    1638
precision: 49.07% +/- 2.84% (micro average: 49.07%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38222   3651
true:   1700    1638
recall: 30.96% +/- 1.64% (micro average: 30.97%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38222   3651
true:   1700    1638
AUC (optimistic): 0.824 +/- 0.008 (micro average: 0.824) (positive class: true)
AUC: 0.824 +/- 0.008 (micro average: 0.824) (positive class: true)
AUC (pessimistic): 0.824 +/- 0.008 (micro average: 0.824) (positive class: true)
```

05

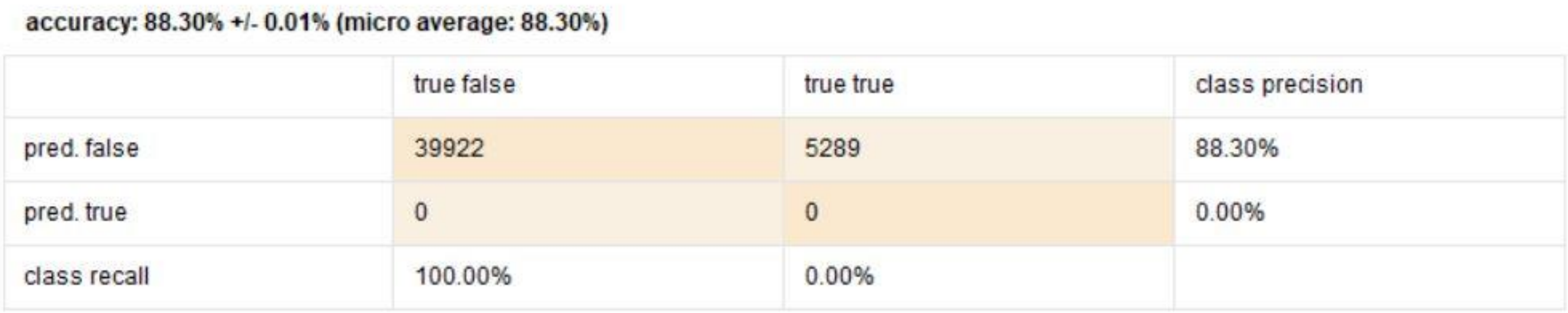
結論與未來改善方向



結論與改善方向

- Y值，false的資料筆數太多，true的筆數太少，不管怎麼分，他的accuracy都會很高，因此我們就去看precision跟recall，發現很低。

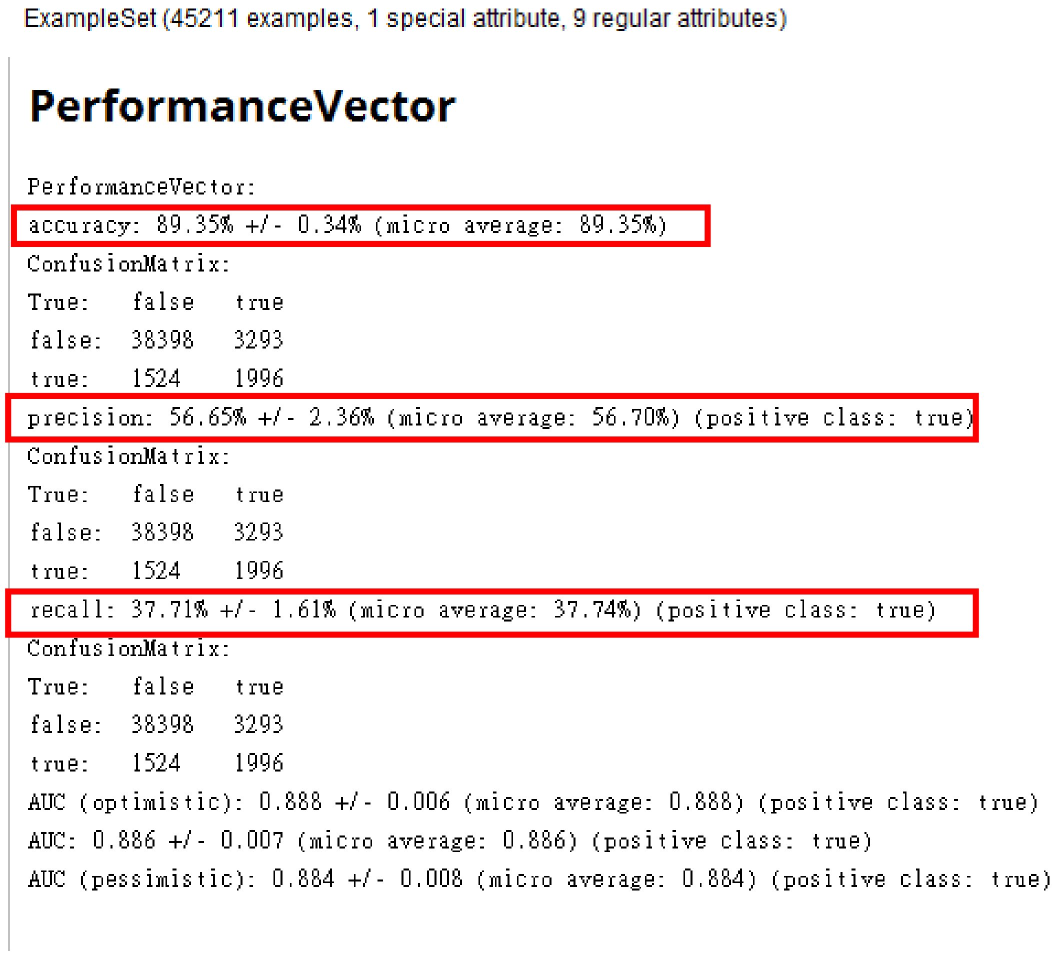
- 解方: (1)增加Y=true的筆數 (2) 減少Y=false的筆數



accuracy: 88.30% +/- 0.01% (micro average: 88.30%)

	true false	true true	class precision
pred. false	39922	5289	88.30%
pred. true	0	0	0.00%
class recall	100.00%	0.00%	

結論與改善方向-(1)增加Y=true的筆數



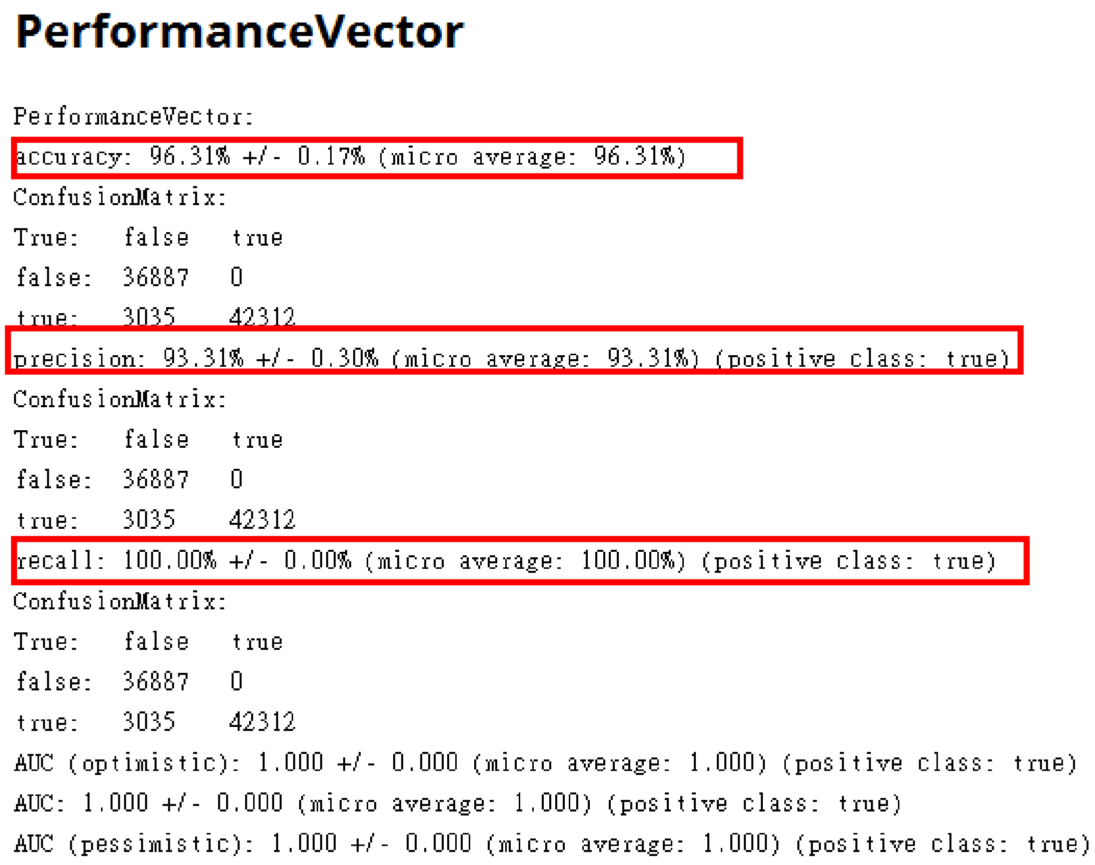
ExampleSet (45211 examples, 1 special attribute, 9 regular attributes)

PerformanceVector

```
PerformanceVector:
accuracy: 89.35% +/- 0.34% (micro average: 89.35%)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
precision: 56.65% +/- 2.36% (micro average: 56.70%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
recall: 37.71% +/- 1.61% (micro average: 37.74%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
AUC (optimistic): 0.888 +/- 0.006 (micro average: 0.888) (positive class: true)
AUC: 0.886 +/- 0.007 (micro average: 0.886) (positive class: true)
AUC (pessimistic): 0.884 +/- 0.008 (micro average: 0.884) (positive class: true)
```

增加前

八萬多筆，true跟false各一半



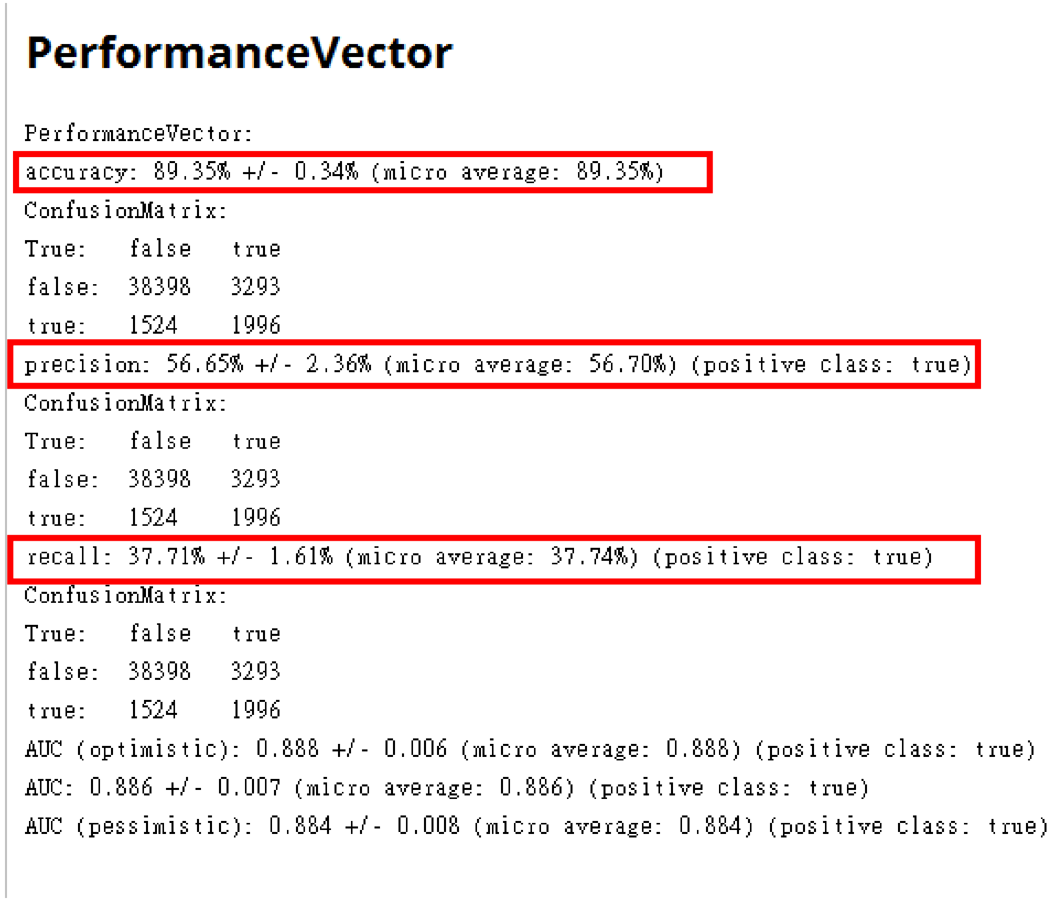
PerformanceVector

```
PerformanceVector:
accuracy: 96.31% +/- 0.17% (micro average: 96.31%)
ConfusionMatrix:
True:   false   true
false:  36887   0
true:   3035    42312
precision: 93.31% +/- 0.30% (micro average: 93.31%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  36887   0
true:   3035    42312
recall: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  36887   0
true:   3035    42312
AUC (optimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: true)
AUC: 1.000 +/- 0.000 (micro average: 1.000) (positive class: true)
AUC (pessimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: true)
```

增加後

結論與改善方向-(2)減少Y=false的筆數

ExampleSet (45211 examples, 1 special attribute, 9 regular attributes)

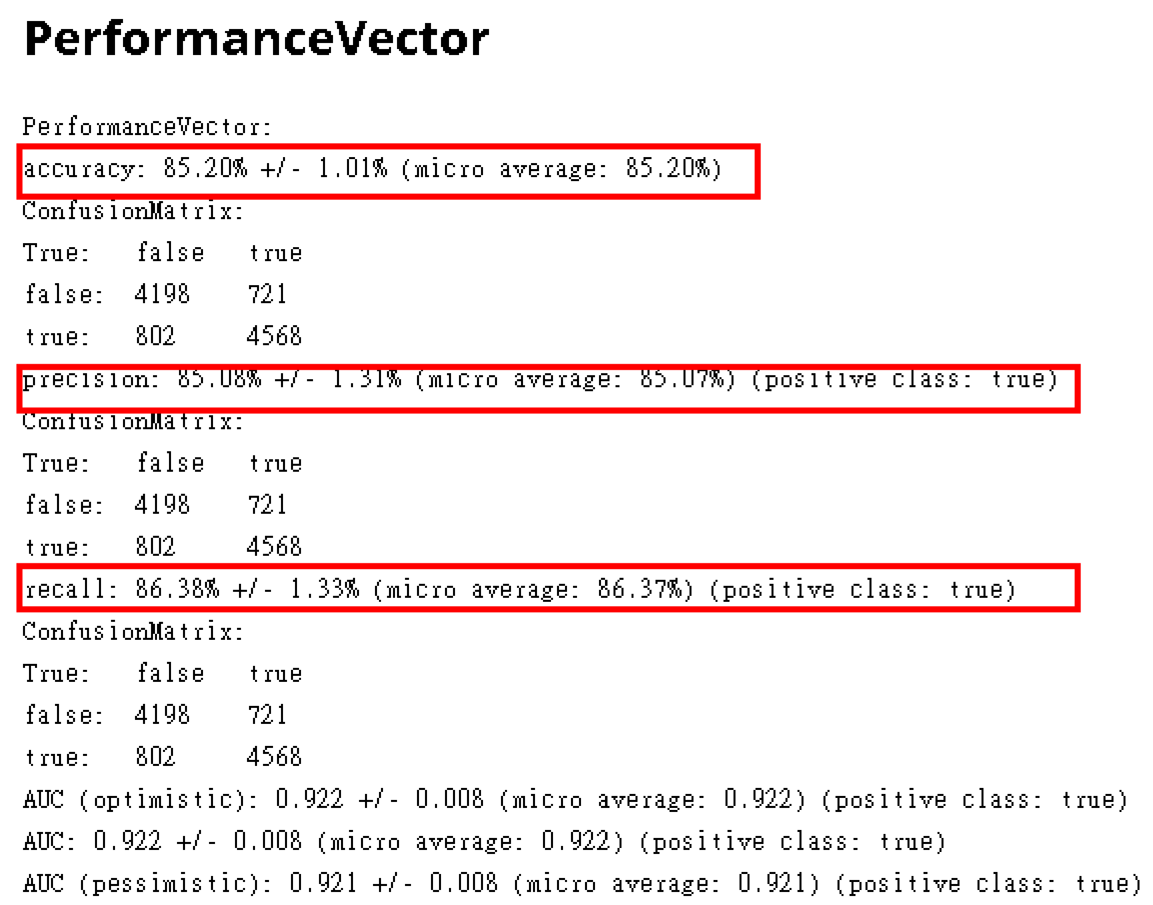


PerformanceVector

```
PerformanceVector:
accuracy: 89.35% +/- 0.34% (micro average: 89.35%)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
precision: 56.65% +/- 2.36% (micro average: 56.70%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
recall: 37.71% +/- 1.61% (micro average: 37.74%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  38398   3293
true:   1524    1996
AUC (optimistic): 0.888 +/- 0.006 (micro average: 0.888) (positive class: true)
AUC: 0.886 +/- 0.007 (micro average: 0.886) (positive class: true)
AUC (pessimistic): 0.884 +/- 0.008 (micro average: 0.884) (positive class: true)
```

減少前

ExampleSet (10289 examples, 1 special attribute, 8 regular attributes) True跟false各半



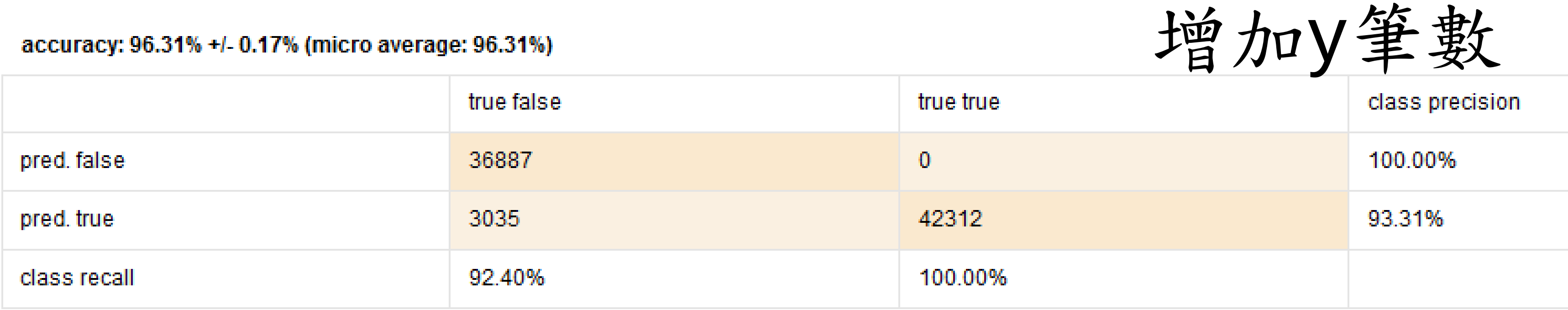
PerformanceVector

```
PerformanceVector:
accuracy: 85.20% +/- 1.01% (micro average: 85.20%)
ConfusionMatrix:
True:   false   true
false:  4198    721
true:   802     4568
precision: 85.08% +/- 1.31% (micro average: 85.07%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  4198    721
true:   802     4568
recall: 86.38% +/- 1.33% (micro average: 86.37%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  4198    721
true:   802     4568
AUC (optimistic): 0.922 +/- 0.008 (micro average: 0.922) (positive class: true)
AUC: 0.922 +/- 0.008 (micro average: 0.922) (positive class: true)
AUC (pessimistic): 0.921 +/- 0.008 (micro average: 0.921) (positive class: true)
```

減少後

結論與改善方向-混淆矩陣

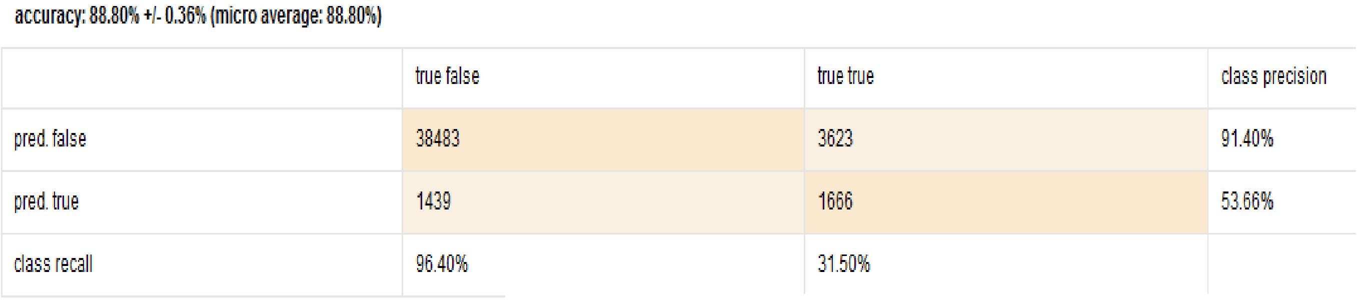
增加y筆數



accuracy: 96.31% +/- 0.17% (micro average: 96.31%)

	true false	true true	class precision
pred. false	36887	0	100.00%
pred. true	3035	42312	93.31%
class recall	92.40%	100.00%	

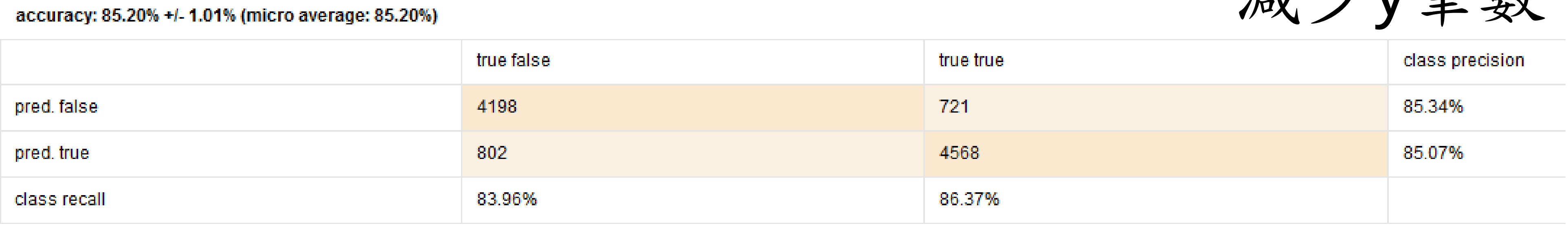
原始結果



accuracy: 88.80% +/- 0.36% (micro average: 88.80%)

	true false	true true	class precision
pred. false	38483	3623	91.40%
pred. true	1439	1666	53.66%
class recall	96.40%	31.50%	

減少y筆數

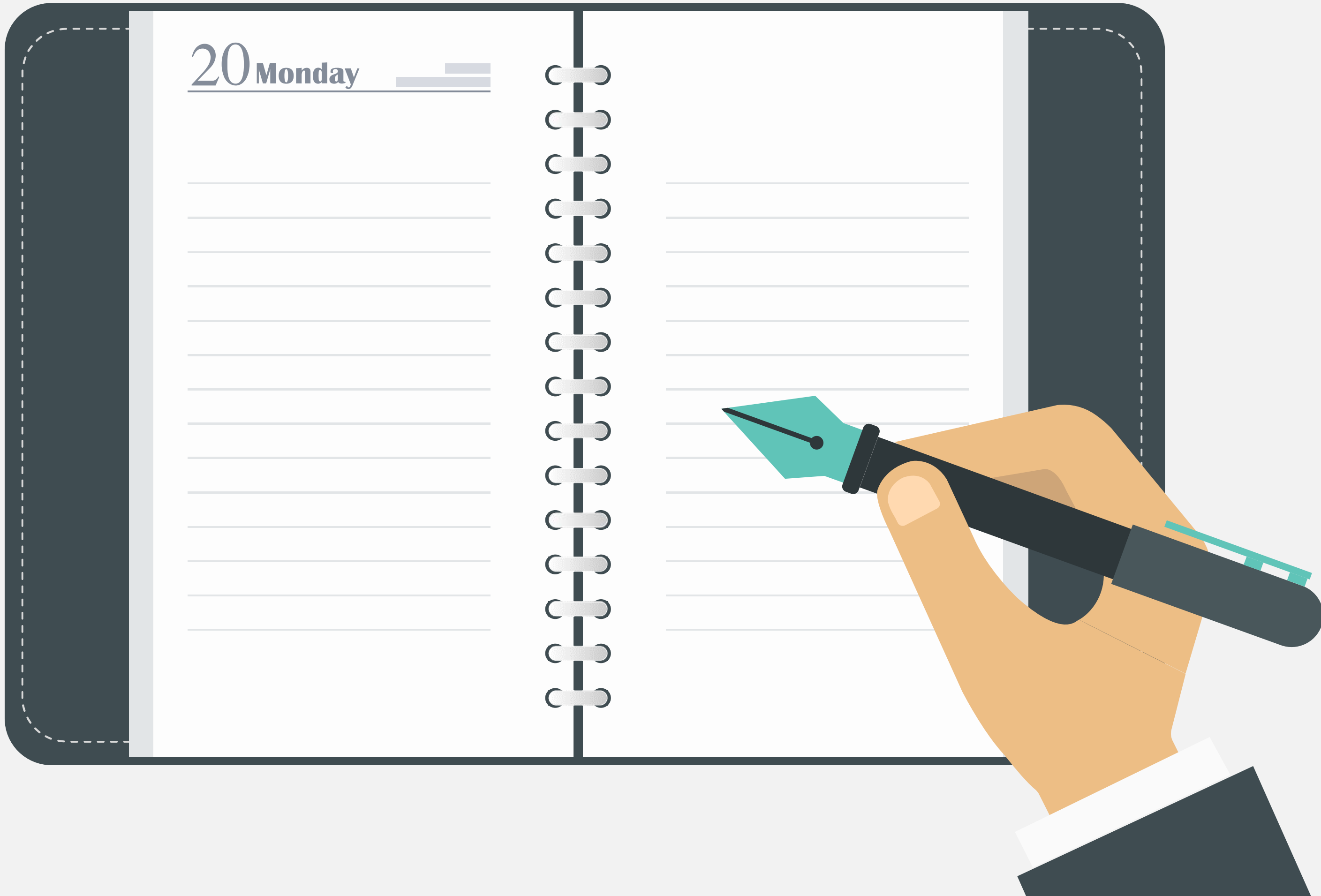


accuracy: 85.20% +/- 1.01% (micro average: 85.20%)

	true false	true true	class precision
pred. false	4198	721	85.34%
pred. true	802	4568	85.07%
class recall	83.96%	86.37%	

06

組員工作分配



組員工作分配

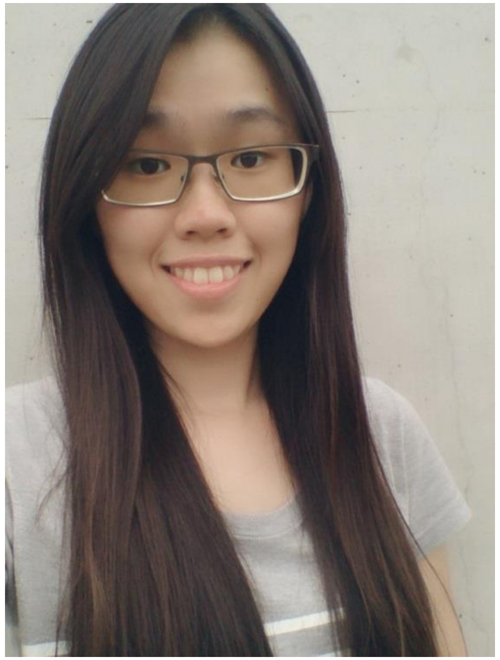
107AB8001　蔡雯惠
負責資料集分析、測試

107AB8002　蕭家希
負責資料集分析、口頭報告

107AB8003 林芝儀
負責資料集分析、測試

組員工作分配

107AB8004　陳鴻妮
負責口頭報告



107AB8005 李予蒨
負責資料統整為PPT

107AB8406 袁嘉妮
負責資料統整為PPT

Thanks for listening.