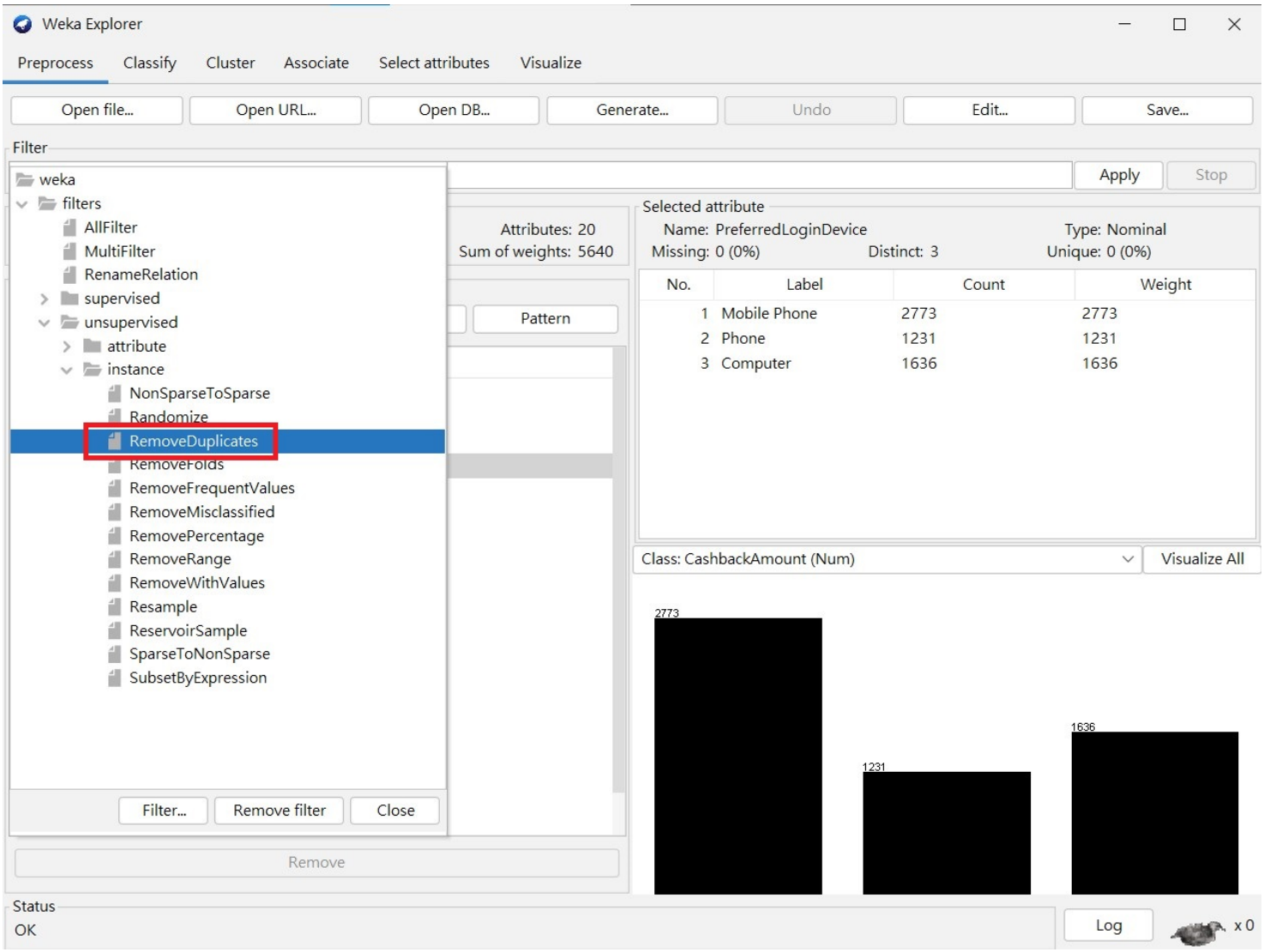
ECT_HW5_108403523

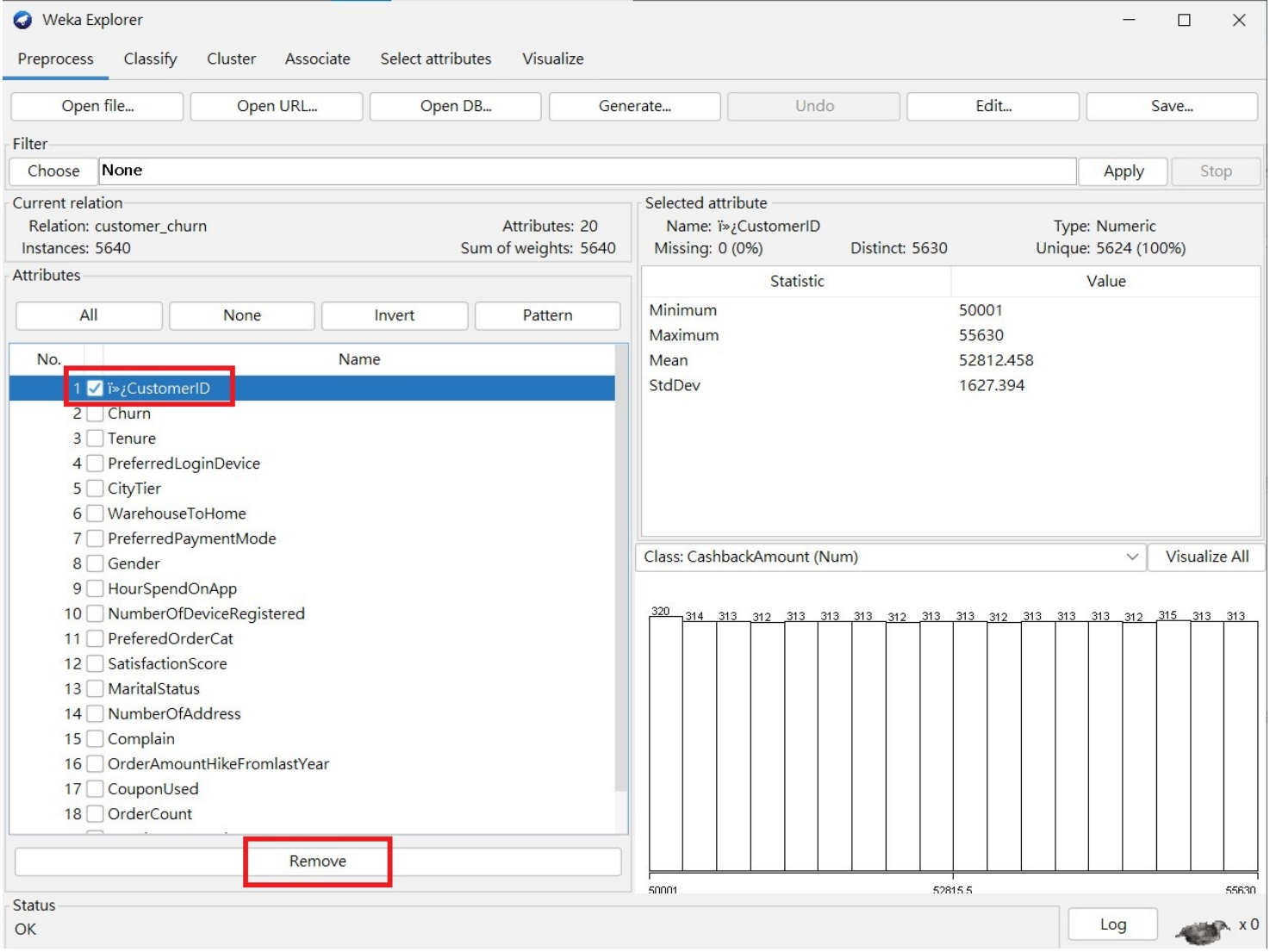
1. 使用 Stratified sampling 從原本的資料集中取60%的資料

(因為 Weka 的 stratified resample 要先做前處理才能執行，所以我把前處理的步驟截圖放在第1題，第3題就不再做前處理，直接以處理後的檔案完成後續題目要求。)

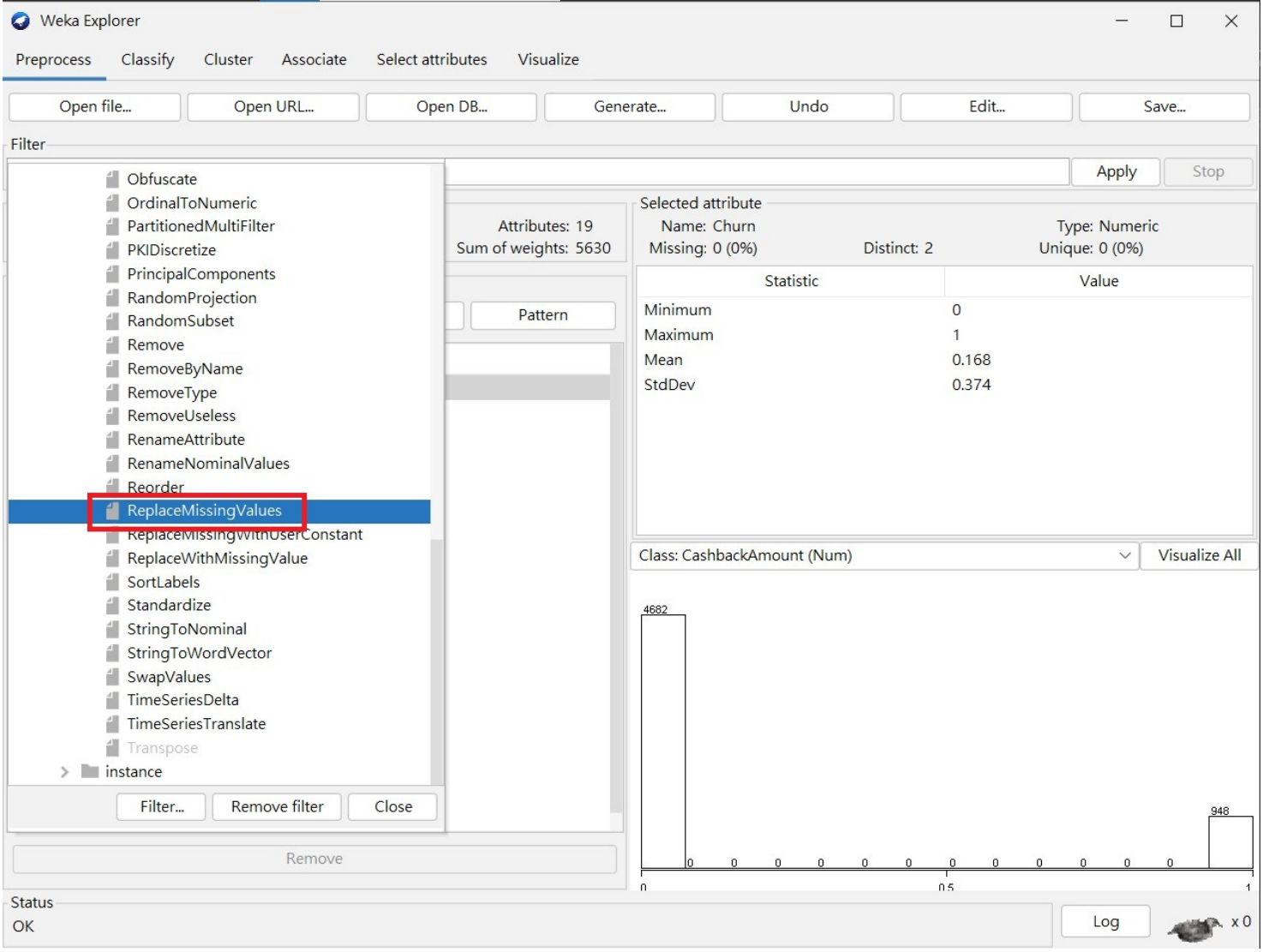
1.1 刪除重複 CustomerID 的資料



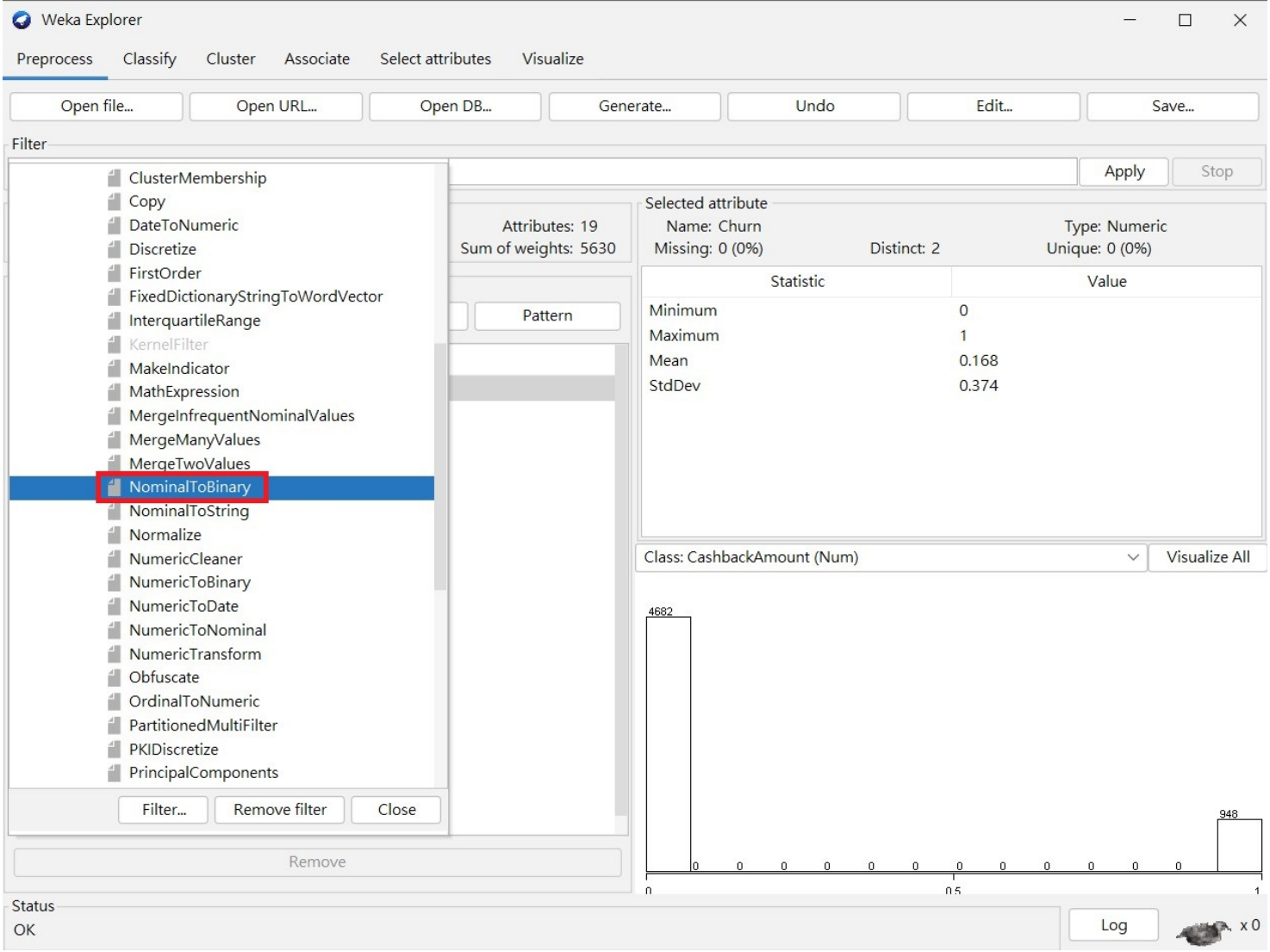
1.2 刪除 CustomerID

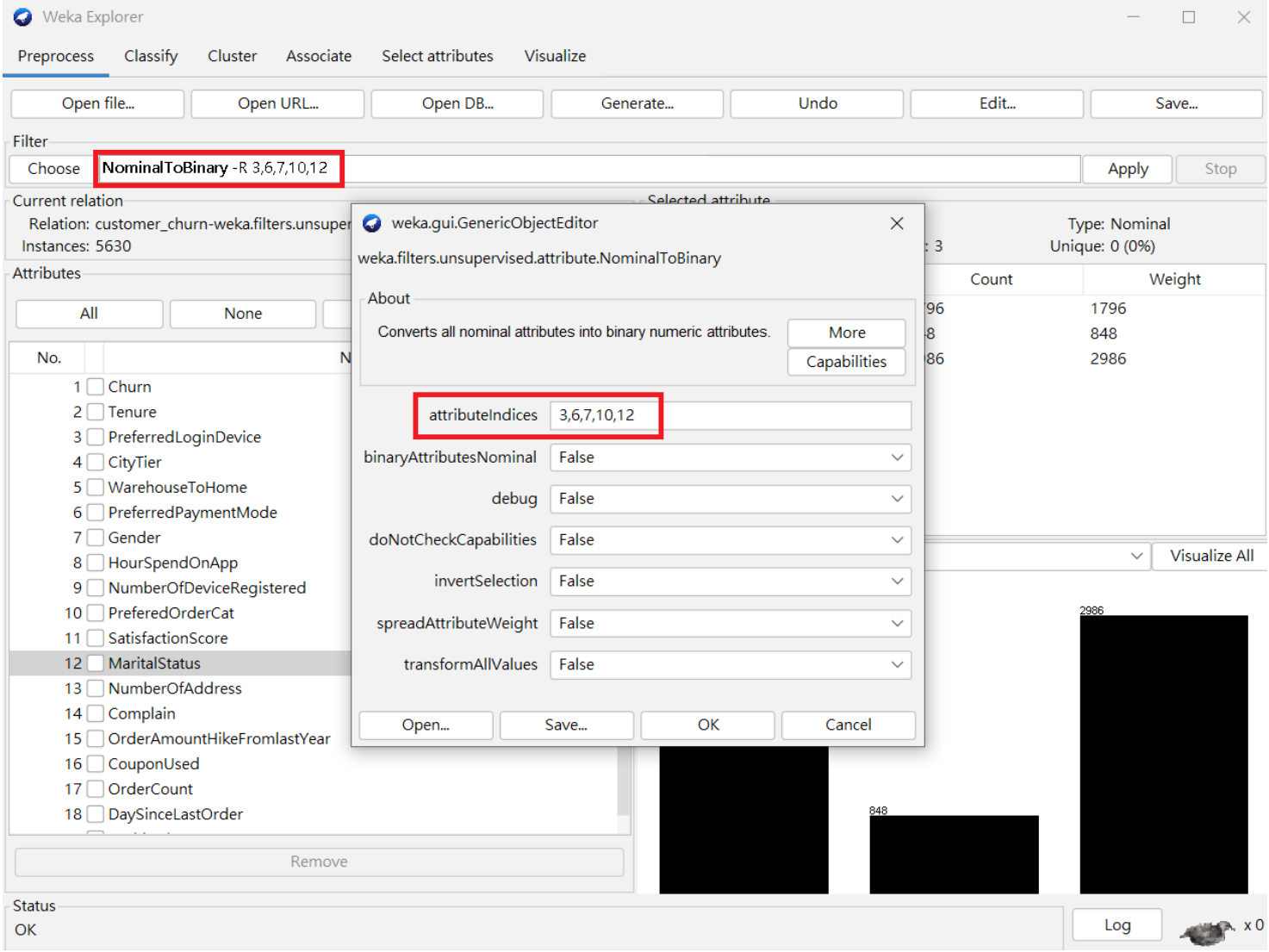

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: customer_churn
Instances: 5640
Attributes: 20
Sum of weights: 5640
Selected attribute
Name: ï»¿CustomerID
Missing: 0 (0%)
Distinct: 5630
Type: Numeric
Unique: 5624 (100%)
Statistic
Value
Minimum
50001
Maximum
55630
Mean
52812.458
StdDev
1627.394
Attributes
All
None
Invert
Pattern
No.
Name
1 ï»¿CustomerID
2 Churn
3 Tenure
4 PreferredLoginDevice
5 CityTier
6 WarehouseToHome
7 PreferredPaymentMode
8 Gender
9 HourSpendOnApp
10 NumberOfDeviceRegistered
11 PreferedOrderCat
12 SatisfactionScore
13 MaritalStatus
14 NumberOfAddress
15 Complain
16 OrderAmountHikeFromlastYear
17 CouponUsed
18 OrderCount
Remove
Class: CashbackAmount (Num)
Visualize All
320 314 313 312 313 313 313 312 313 313 312 313 313 313 312 315 313 313
50001
52815.5
55630
Status
OK
Log
x 0

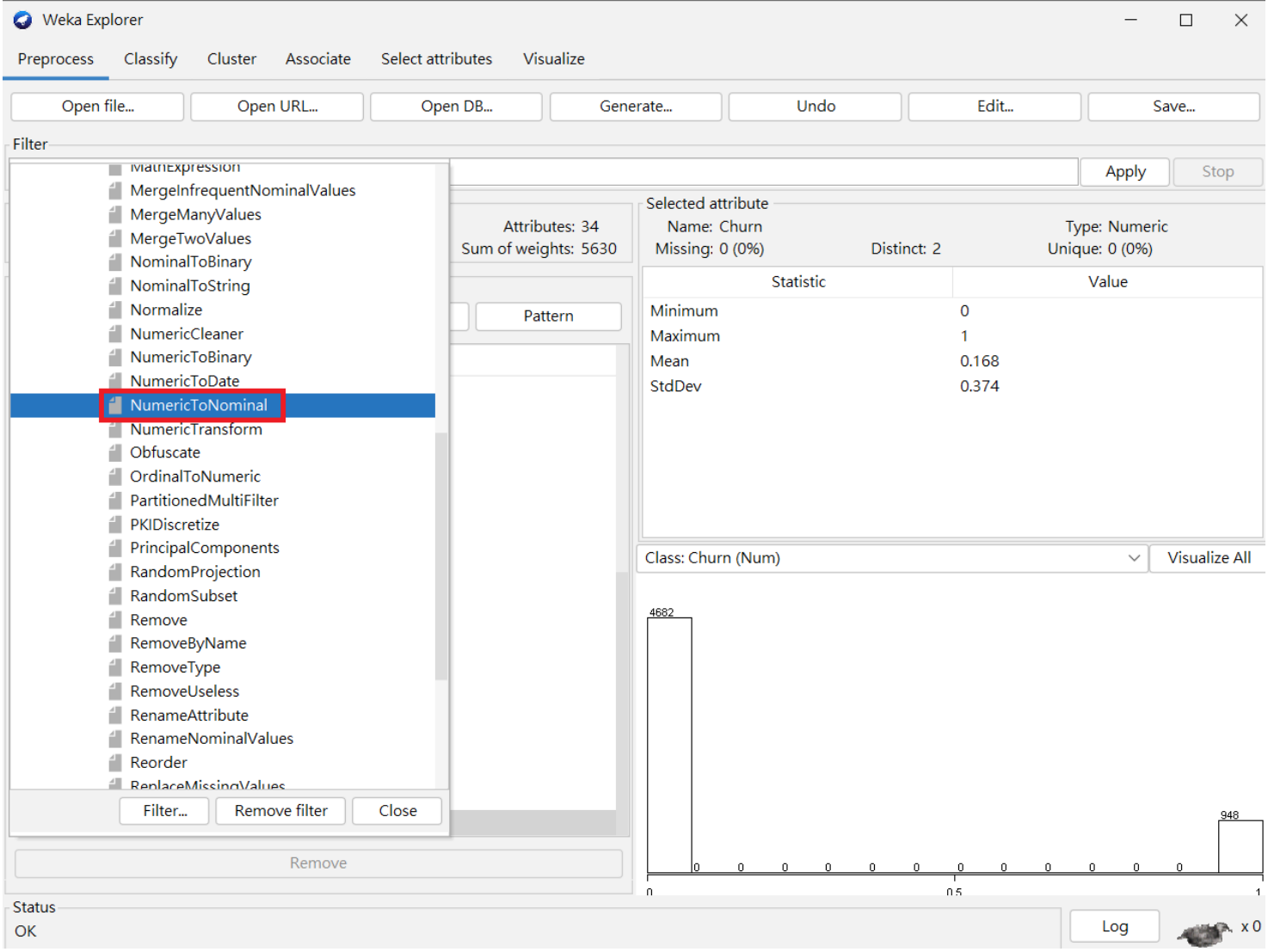
1.3 填補缺失值

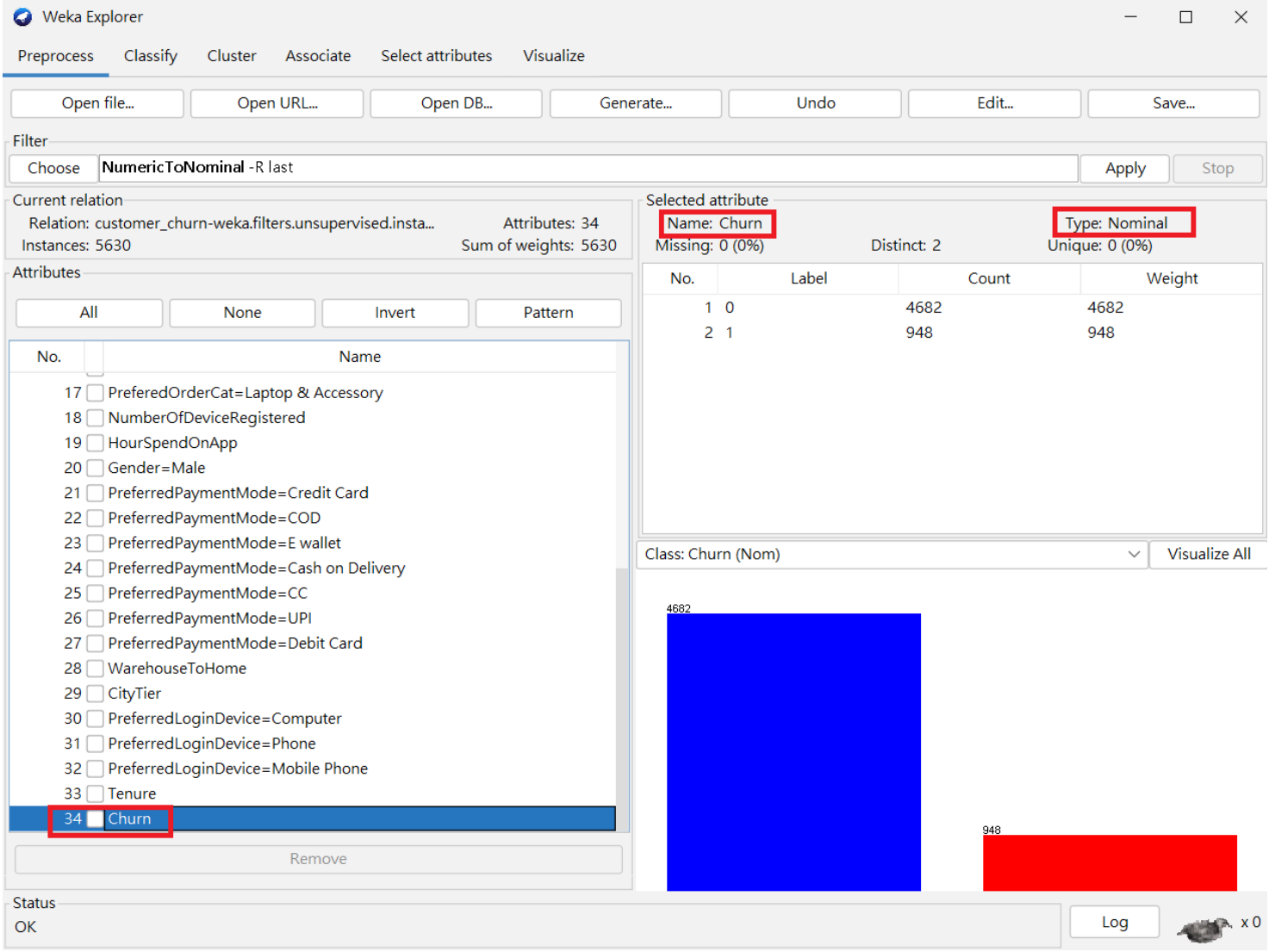


1.4 將資料進行轉換，僅保留 Churn 為唯一的 "Nominal" 欄

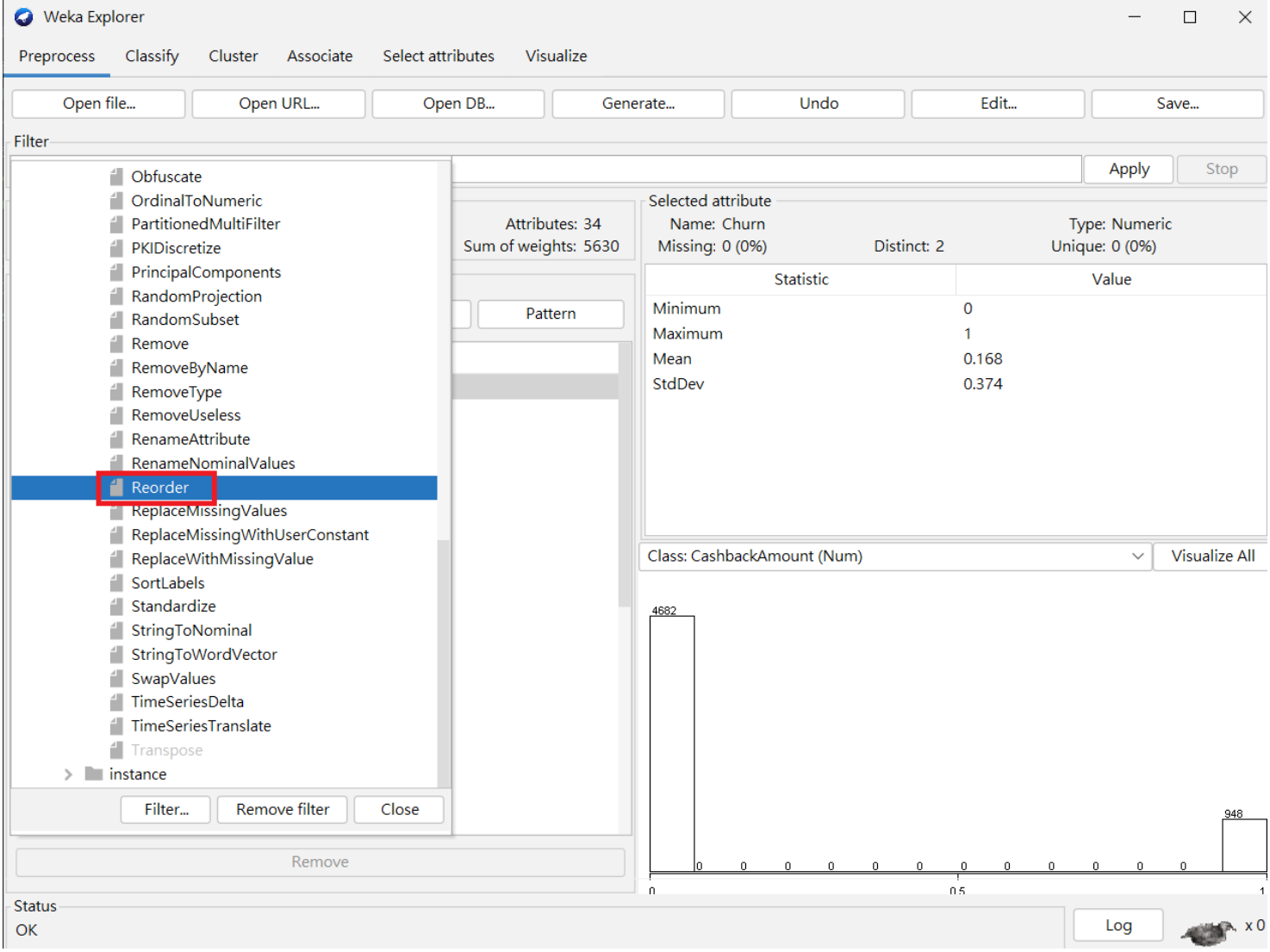

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Apply
Stop
ClusterMembership
Copy
DateToNumeric
Discretize
FirstOrder
FixedDictionaryStringToWordVector
InterquartileRange
KernelFilter
MakeIndicator
MathExpression
MergeInfrequentNominalValues
MergeManyValues
MergeTwoValues
NominalToBinary
NominalToString
Normalize
NumericCleaner
NumericToBinary
NumericToDate
NumericToNominal
NumericTransform
Obfuscate
OrdinalToNumeric
PartitionedMultiFilter
PKIDiscretize
PrincipalComponents
Filter...
Remove filter
Close
Attributes: 19
Sum of weights: 5630
Pattern
Remove
Selected attribute
Name: Churn
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum
0
Maximum
1
Mean
0.168
StdDev
0.374
Class: CashbackAmount (Num)
Visualize All
4682
948
0
0.5
1
Status
OK
Log
x 0

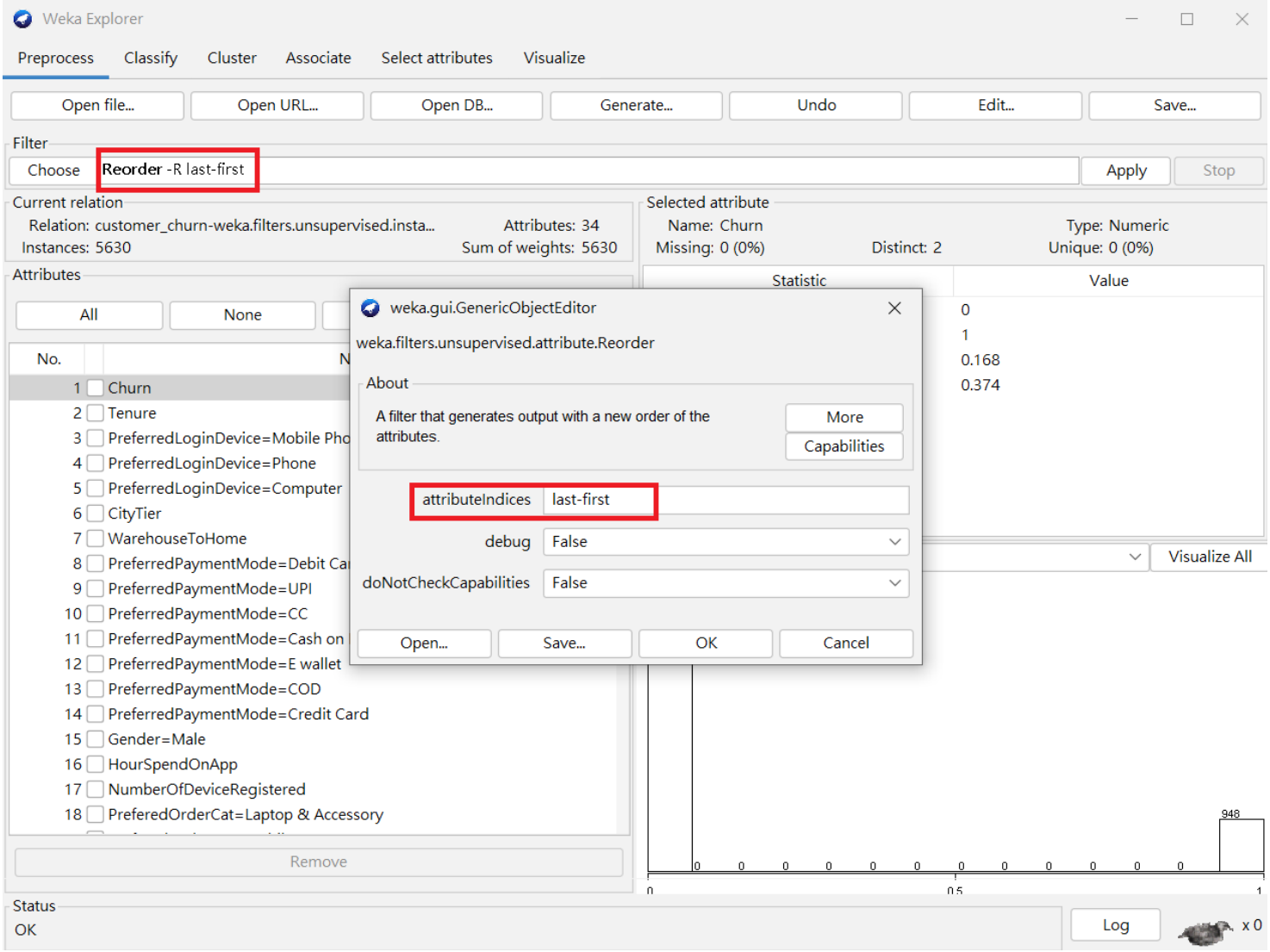

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NominalToBinary -R 3,6,7,10,12
Apply
Stop
Current relation
Relation: customer_churn-weka.filters.unsuper
Instances: 5630
Attributes
All
None
No.
1 Churn
2 Tenure
3 PreferredLoginDevice
4 CityTier
5 WarehouseToHome
6 PreferredPaymentMode
7 Gender
8 HourSpendOnApp
9 NumberOfDeviceRegistered
10 PreferedOrderCat
11 SatisfactionScore
12 MaritalStatus
13 NumberOfAddress
14 Complain
15 OrderAmountHikeFromlastYear
16 CouponUsed
17 OrderCount
18 DaySinceLastOrder
Remove
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NominalToBinary
About
Converts all nominal attributes into binary numeric attributes.
More
Capabilities
attributeIndices
3,6,7,10,12
binaryAttributesNominal
False
debug
False
doNotCheckCapabilities
False
invertSelection
False
spreadAttributeWeight
False
transformAllValues
False
Open...
Save...
OK
Cancel
Type: Nominal
Unique: 0 (0%)
Count
Weight
1796
848
2986
Visualize All
2986
848
Status
OK
Log
x 0

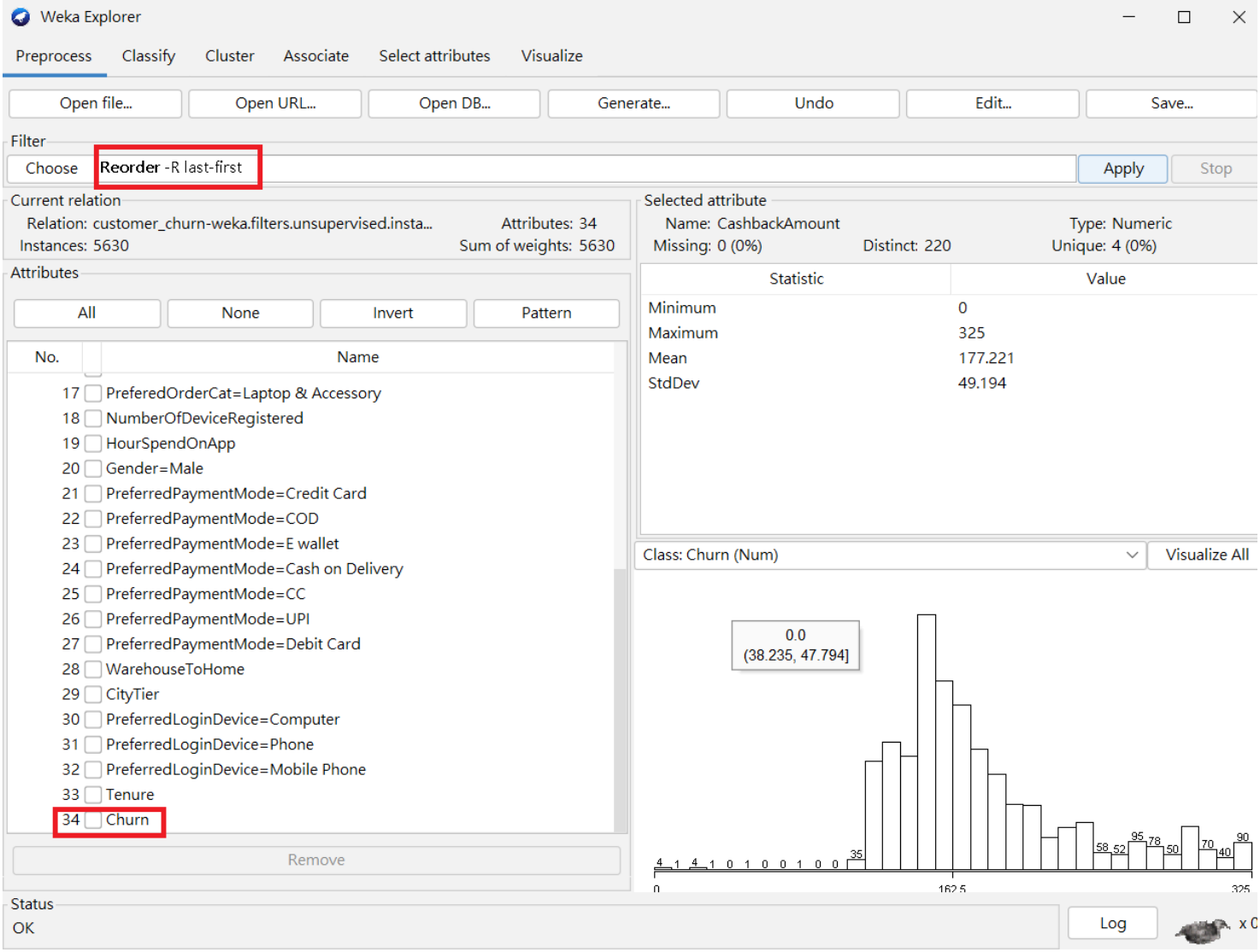

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Apply
Stop
MergeInfrequentNominalValues
MergeManyValues
MergeTwoValues
NominalToBinary
NominalToString
Normalize
NumericCleaner
NumericToBinary
NumericToDate
NumericToNominal
NumericTransform
Obfuscate
OrdinalToNumeric
PartitionedMultiFilter
PKIDiscretize
PrincipalComponents
RandomProjection
RandomSubset
Remove
RemoveByName
RemoveType
RemoveUseless
RenameAttribute
RenameNominalValues
Reorder
Filter...
Remove filter
Close
Attributes: 34
Sum of weights: 5630
Pattern
Remove
Selected attribute
Name: Churn
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 0
Maximum 1
Mean 0.168
StdDev 0.374
Class: Churn (Num)
Visualize All
4682
948
0
0.5
1
Status
OK
Log
x 0



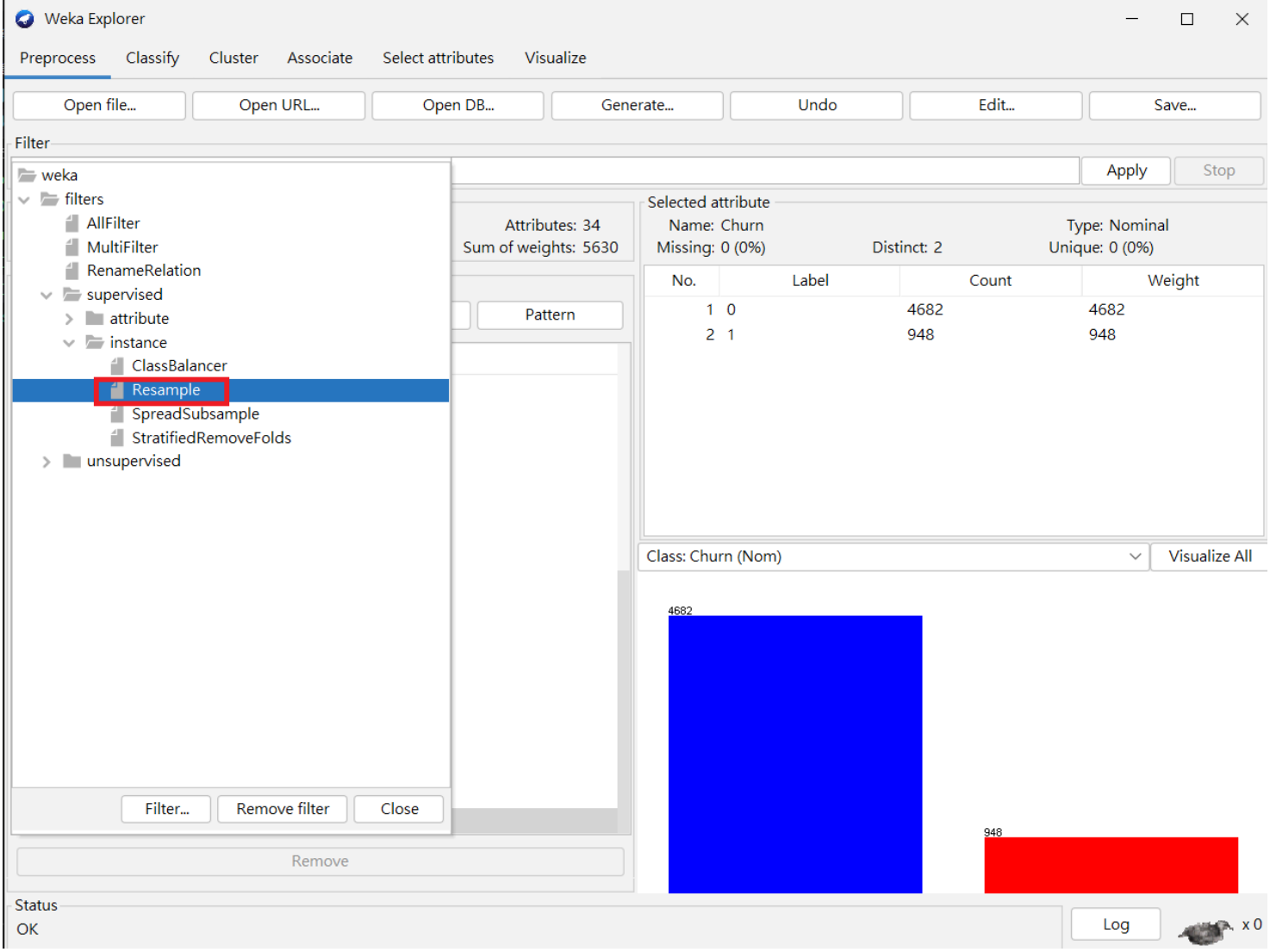
1.5 重新排列屬性(顛倒排列，Churn 排最後)

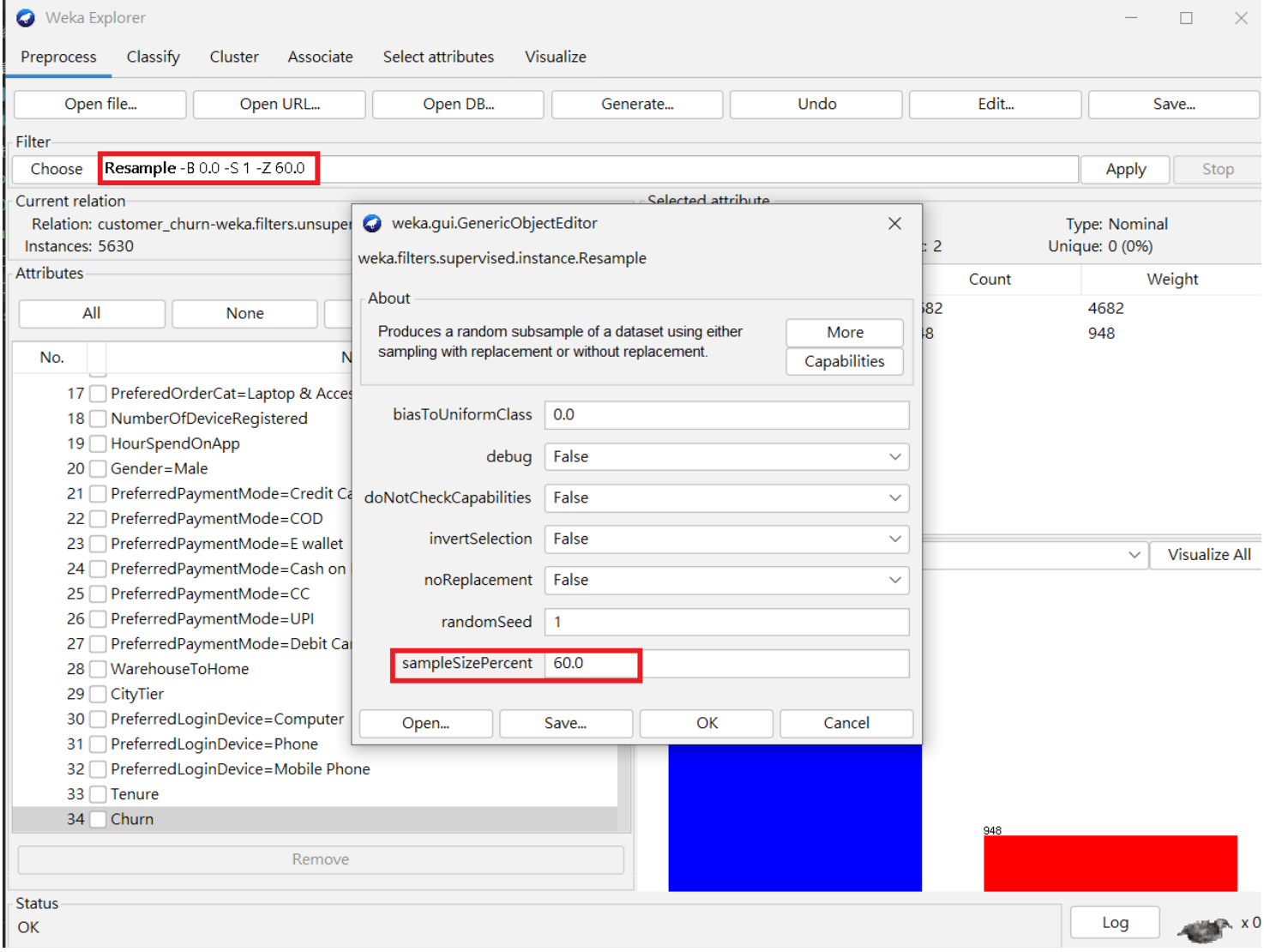

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Apply
Stop
Obfuscate
OrdinalToNumeric
PartitionedMultiFilter
PKIDiscretize
PrincipalComponents
RandomProjection
RandomSubset
Remove
RemoveByName
RemoveType
RemoveUseless
RenameAttribute
RenameNominalValues
Reorder
ReplaceMissingValues
ReplaceMissingWithUserConstant
ReplaceWithMissingValue
SortLabels
Standardize
StringToNominal
StringToWordVector
SwapValues
TimeSeriesDelta
TimeSeriesTranslate
Transpose
instance
Filter...
Remove filter
Close
Attributes: 34
Sum of weights: 5630
Pattern
Remove
Selected attribute
Name: Churn
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum
0
Maximum
1
Mean
0.168
StdDev
0.374
Class: CashbackAmount (Num)
Visualize All
4682
948
0
0.5
1
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Reorder -R last-first
Apply
Stop
Current relation
Relation: customer_churn-weka.filters.unsupervised.insta...
Attributes: 34
Instances: 5630
Sum of weights: 5630
Selected attribute
Name: Churn
Type: Numeric
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
Statistic
Value
0
1
0.168
0.374
Attributes
All
None
No.
1 Churn
2 Tenure
3 PreferredLoginDevice=Mobile Pho
4 PreferredLoginDevice=Phone
5 PreferredLoginDevice=Computer
6 CityTier
7 WarehouseToHome
8 PreferredPaymentMode=Debit Ca
9 PreferredPaymentMode=UPI
10 PreferredPaymentMode=CC
11 PreferredPaymentMode=Cash on
12 PreferredPaymentMode=E wallet
13 PreferredPaymentMode=COD
14 PreferredPaymentMode=Credit Card
15 Gender=Male
16 HourSpendOnApp
17 NumberOfDeviceRegistered
18 PreferedOrderCat=Laptop & Accessory
Remove
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Reorder
About
A filter that generates output with a new order of the attributes.
More
Capabilities
attributeIndices
last-first
debug
False
doNotCheckCapabilities
False
Open...
Save...
OK
Cancel
Visualize All
948
0
0.5
1
Status
OK
Log
x 0



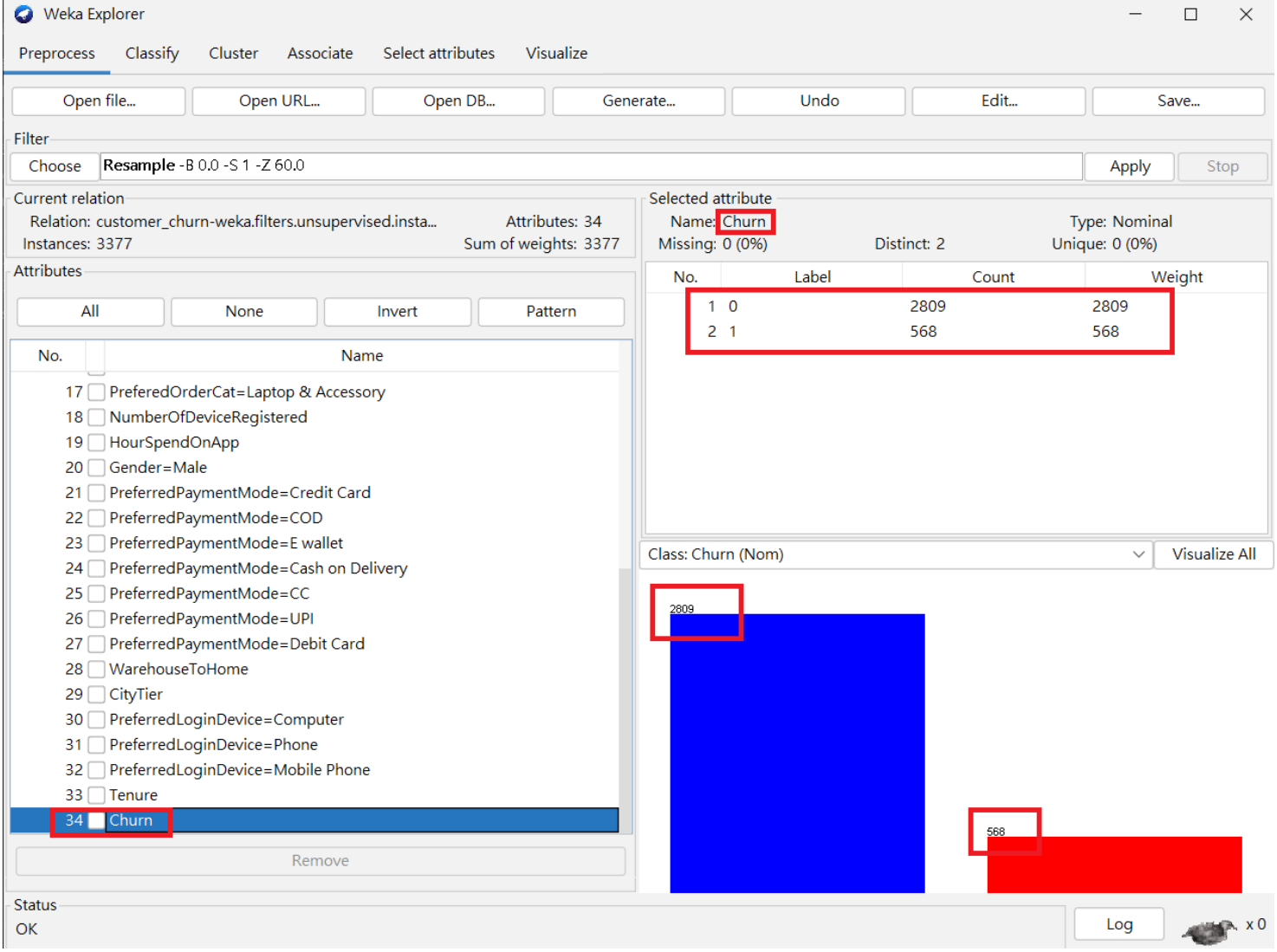
1.6 做 60% stratified resample


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Apply
Stop
weka
filters
AllFilter
MultiFilter
RenameRelation
supervised
attribute
instance
ClassBalancer
Resample
SpreadSubsample
StratifiedRemoveFolds
unsupervised
Filter...
Remove filter
Close
Attributes: 34
Sum of weights: 5630
Pattern
Remove
Selected attribute
Name: Churn
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 0 4682 4682
2 1 948 948
Class: Churn (Nom)
Visualize All
4682
948
Status
OK
Log
x 0

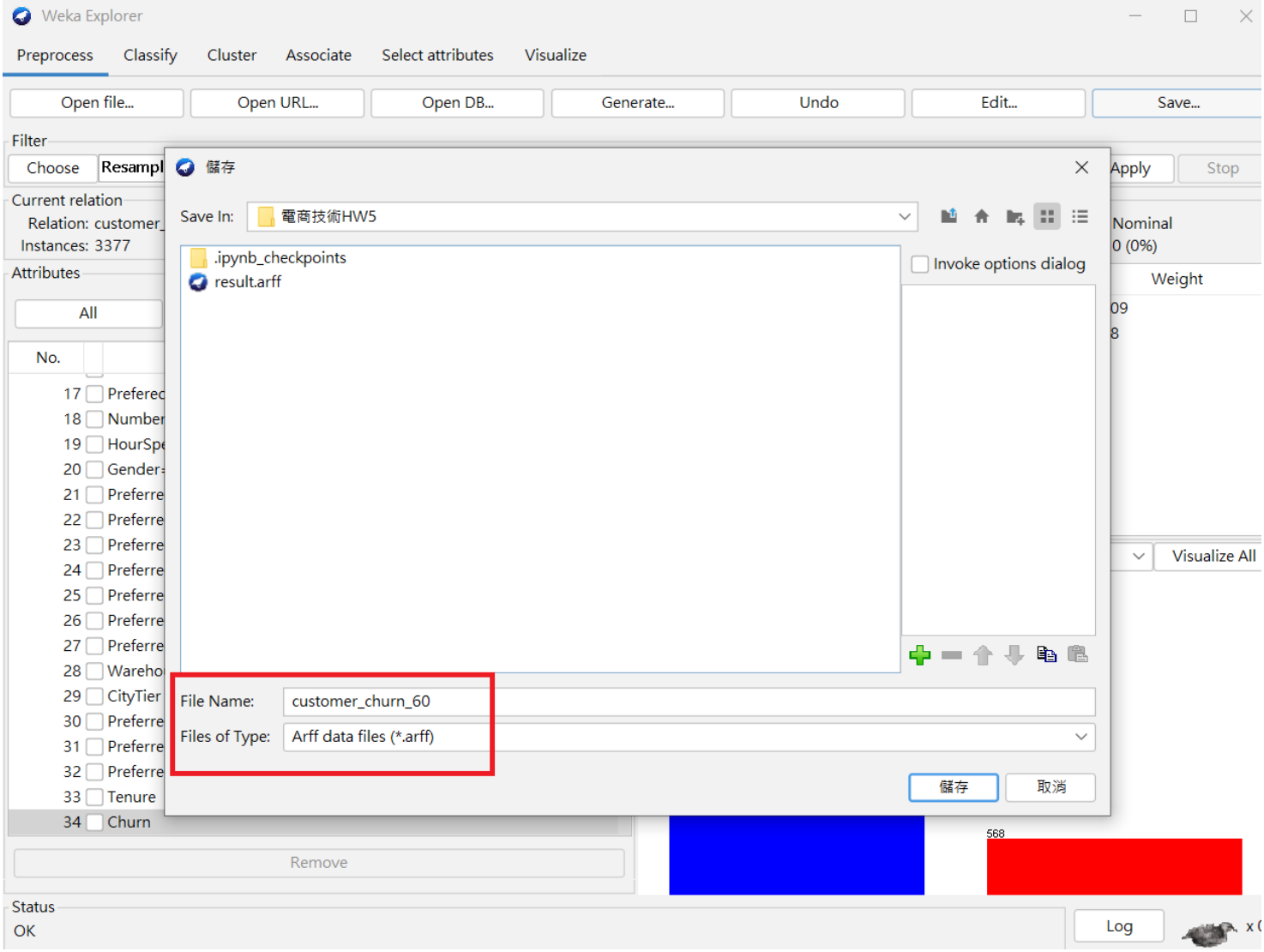


2. 顯示取樣後各類別的資料數量

2.1 紅框內為分層抽樣後各類別資料數量



2.2 存成新的檔案做後續第3、4題

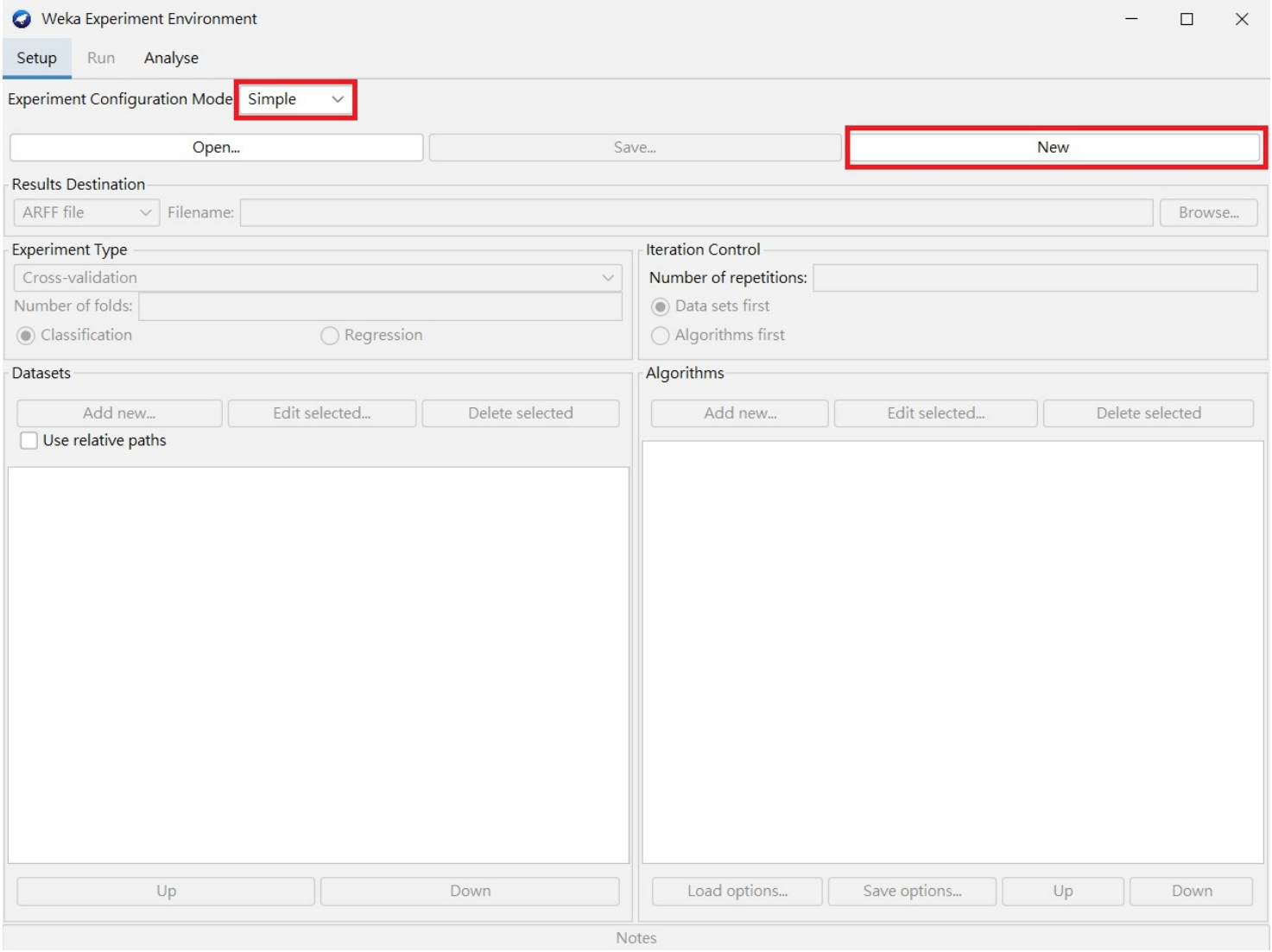


---

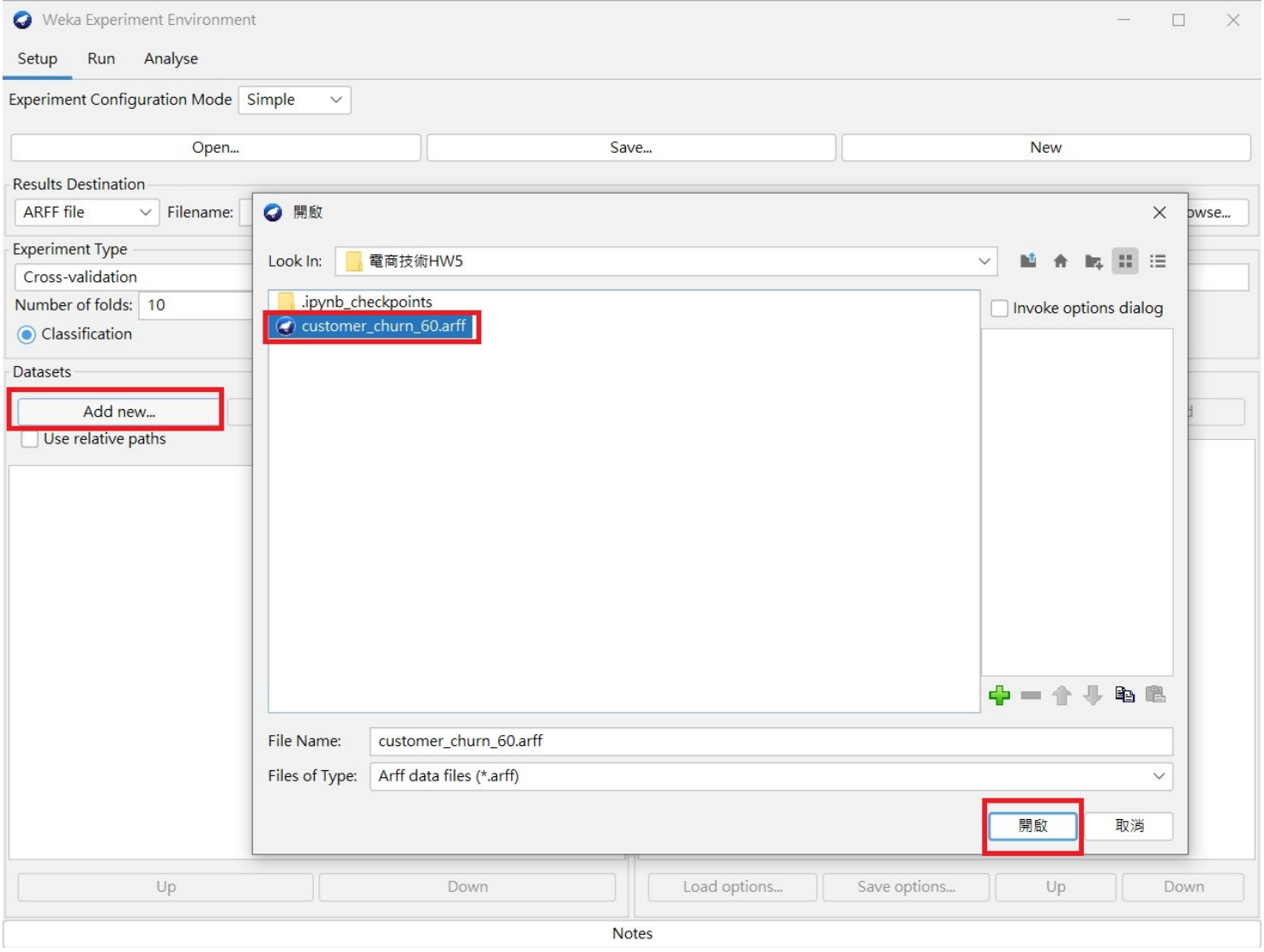
3. 資料前處理，並以 repeated 10 folds cross-validation (重複 10 次) Paired t-test 比 較 Logistic Regression 及 SVM模型

(在第1、2題已完成資料前處理，並將分層抽樣結果另存新檔)

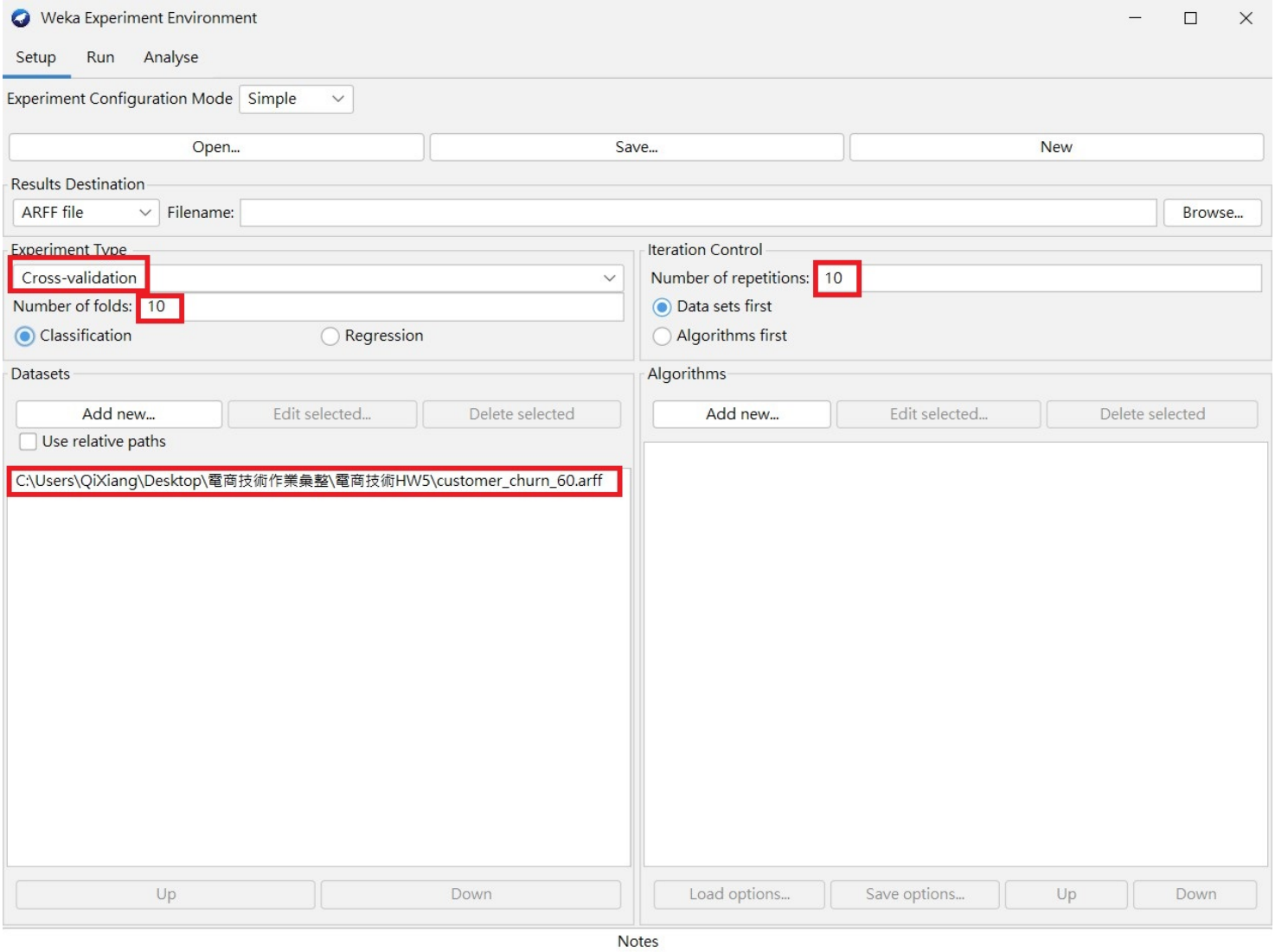
3.1 New Experiment


Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode
Simple
Open...
Save...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
Classification
Regression
Iteration Control
Number of repetitions:
Data sets first
Algorithms first
Datasets
Add new...
Edit selected...
Delete selected
Use relative paths
Up
Down
Algorithms
Add new...
Edit selected...
Delete selected
Load options...
Save options...
Up
Down
Notes

3.2 導入資料

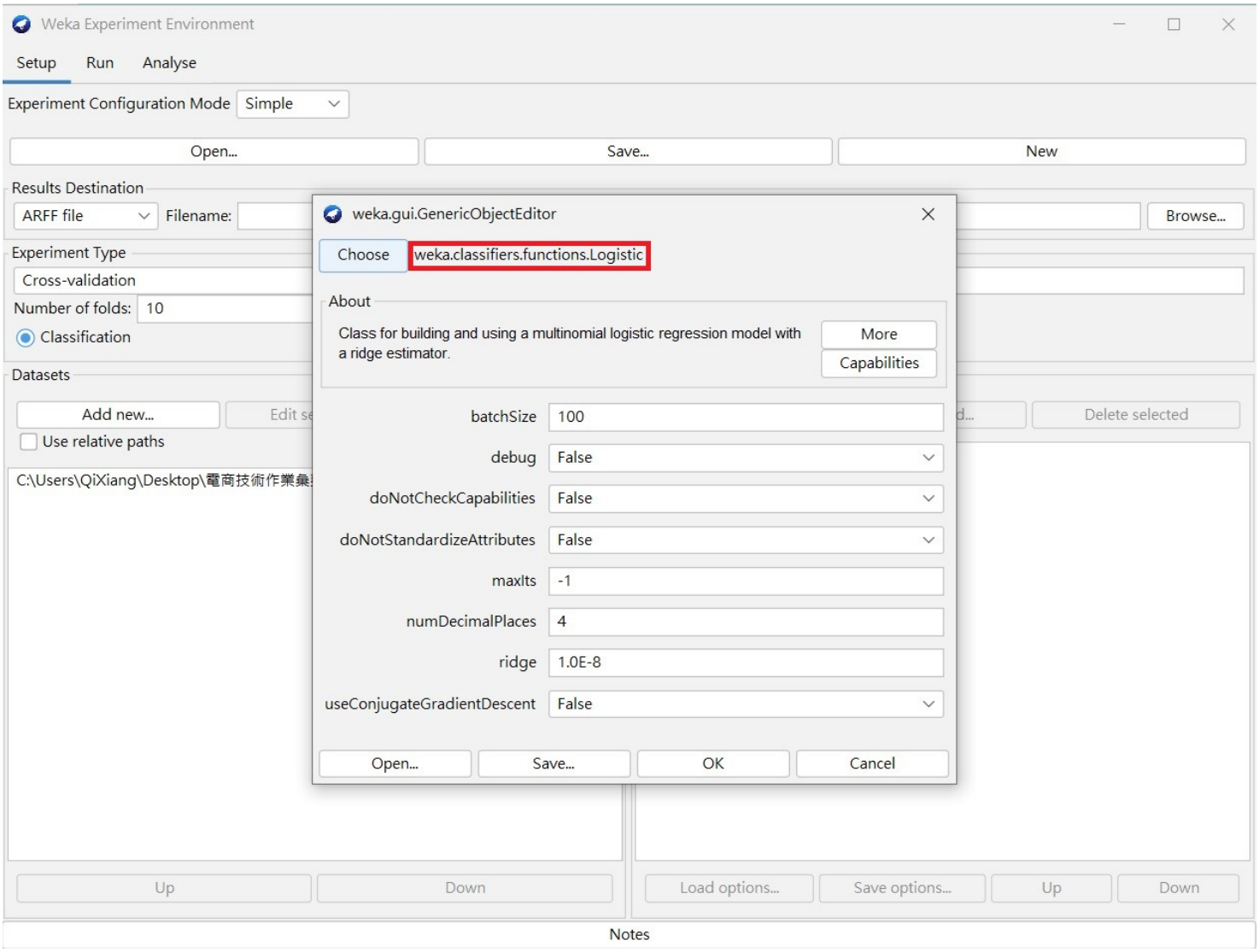


3.3 設定 10 run 10 folds cross validation

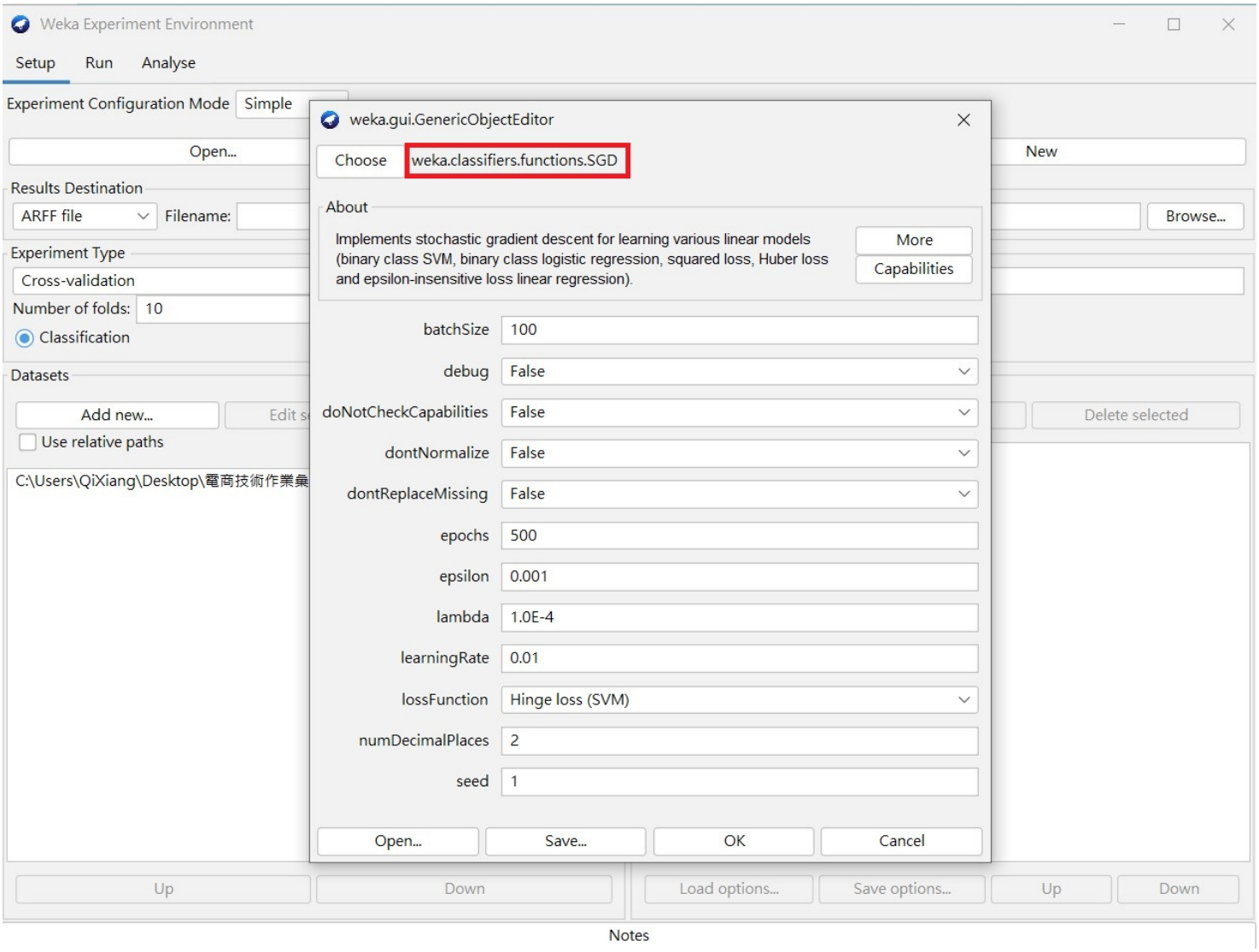


3.4 設定要跑的兩個模型

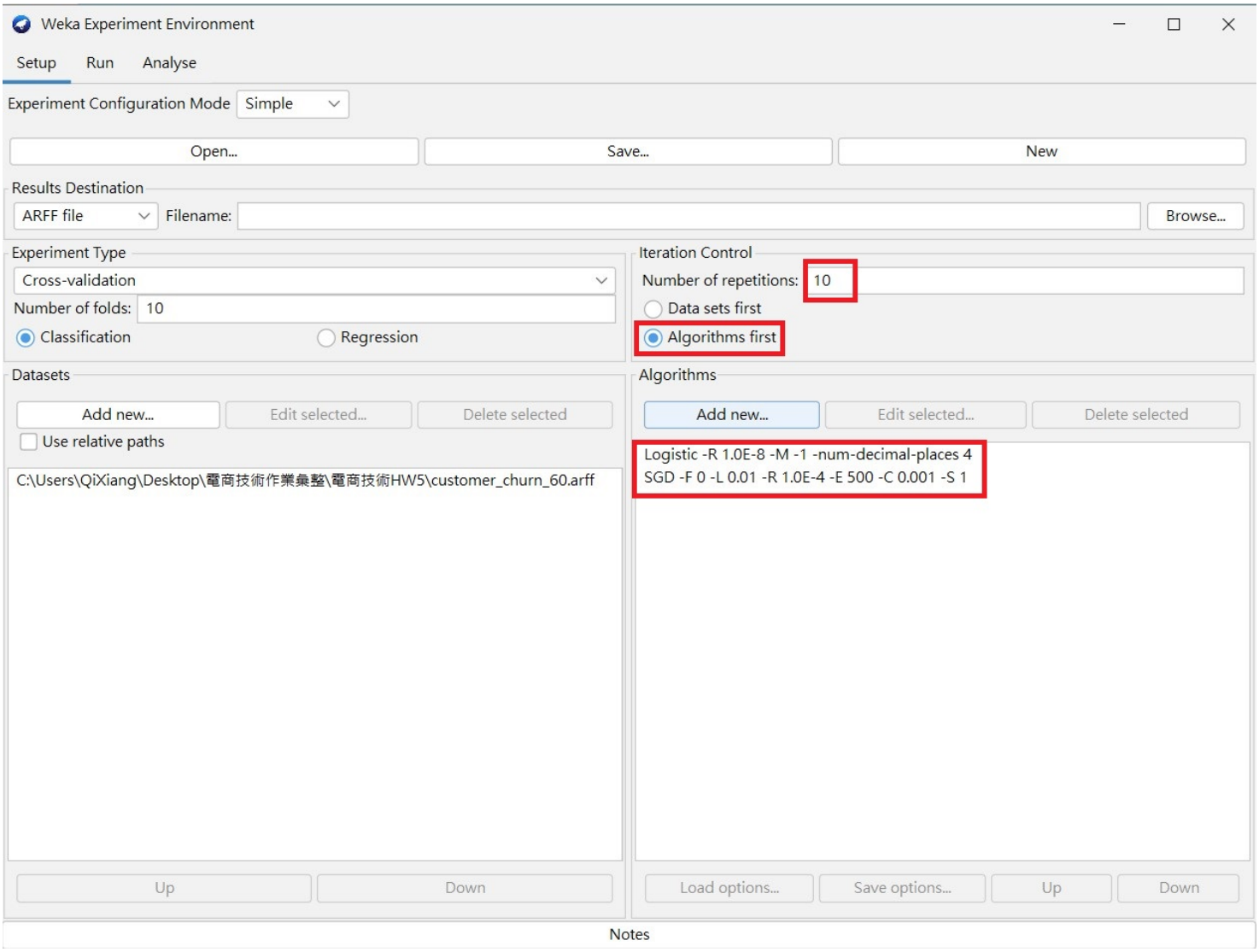
(Logistic Regression)



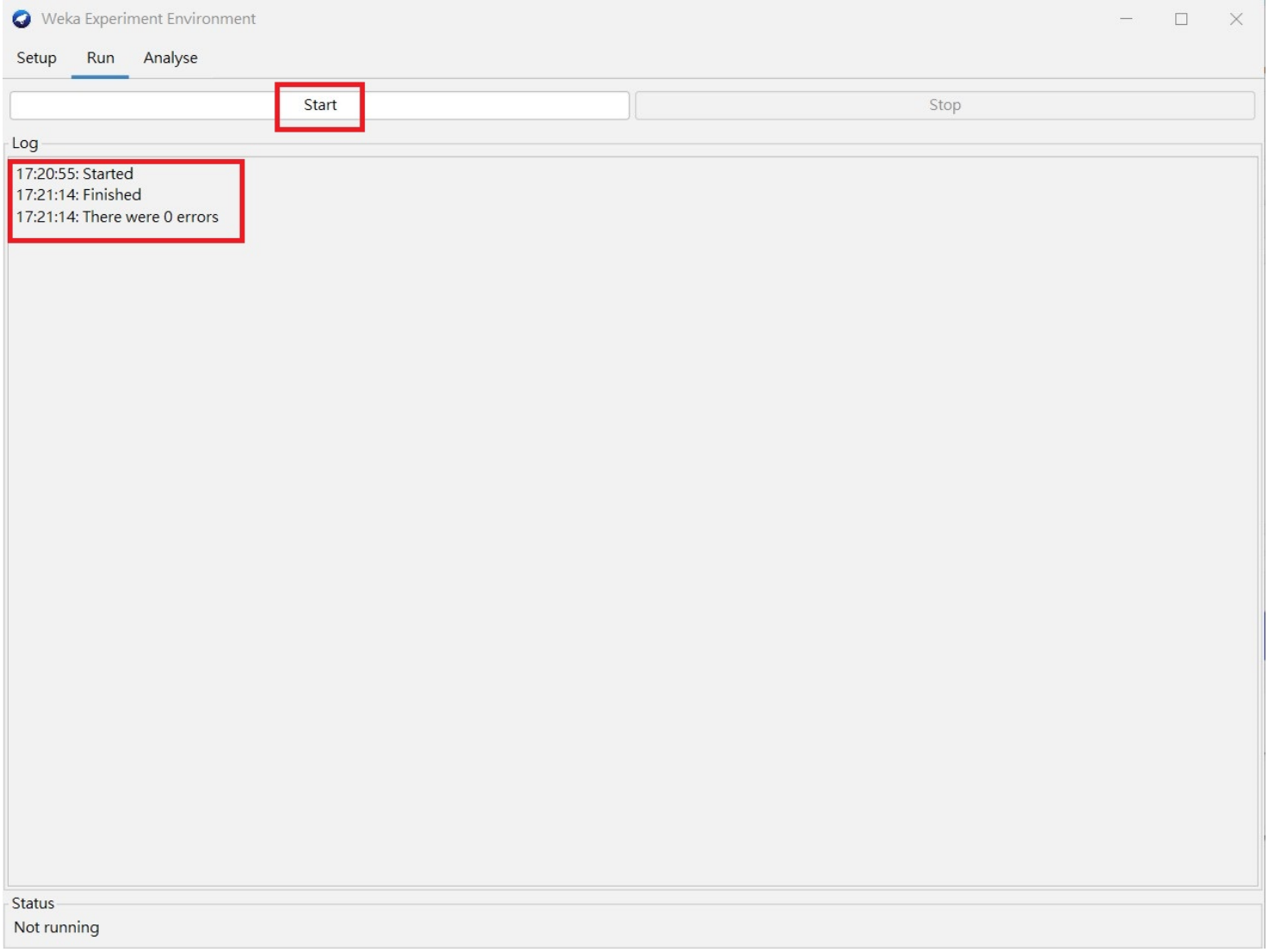
(SVM)


Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode
Simple
Open...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
10
Classification
Datasets
Add new...
Delete selected
Use relative paths
C:\Users\QiXiang\Desktop\電商技術作業集
Up
Down
Load options...
Save options...
Up
Down
Notes
weka.gui.GenericObjectEditor
Choose
weka.classifiers.functions.SGD
About
Implements stochastic gradient descent for learning various linear models (binary class SVM, binary class logistic regression, squared loss, Huber loss and epsilon-insensitive loss linear regression).
More
Capabilities
batchSize
100
debug
False
doNotCheckCapabilities
False
dontNormalize
False
dontReplaceMissing
False
epochs
500
epsilon
0.001
lambda
1.0E-4
learningRate
0.01
lossFunction
Hinge loss (SVM)
numDecimalPlaces
2
seed
1
Open...
Save...
OK
Cancel

(各跑10次)

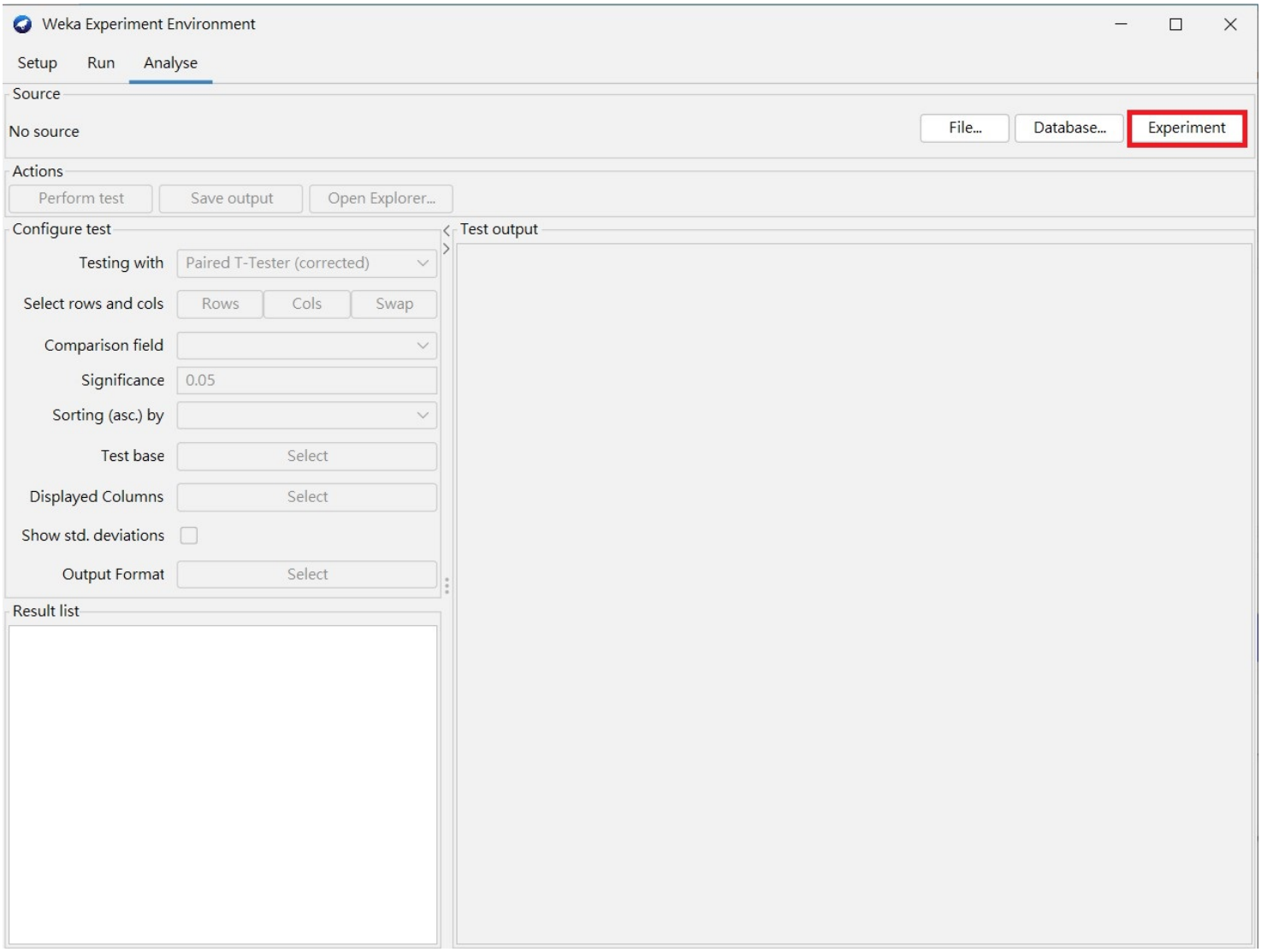


3.5 開始跑 Experiment

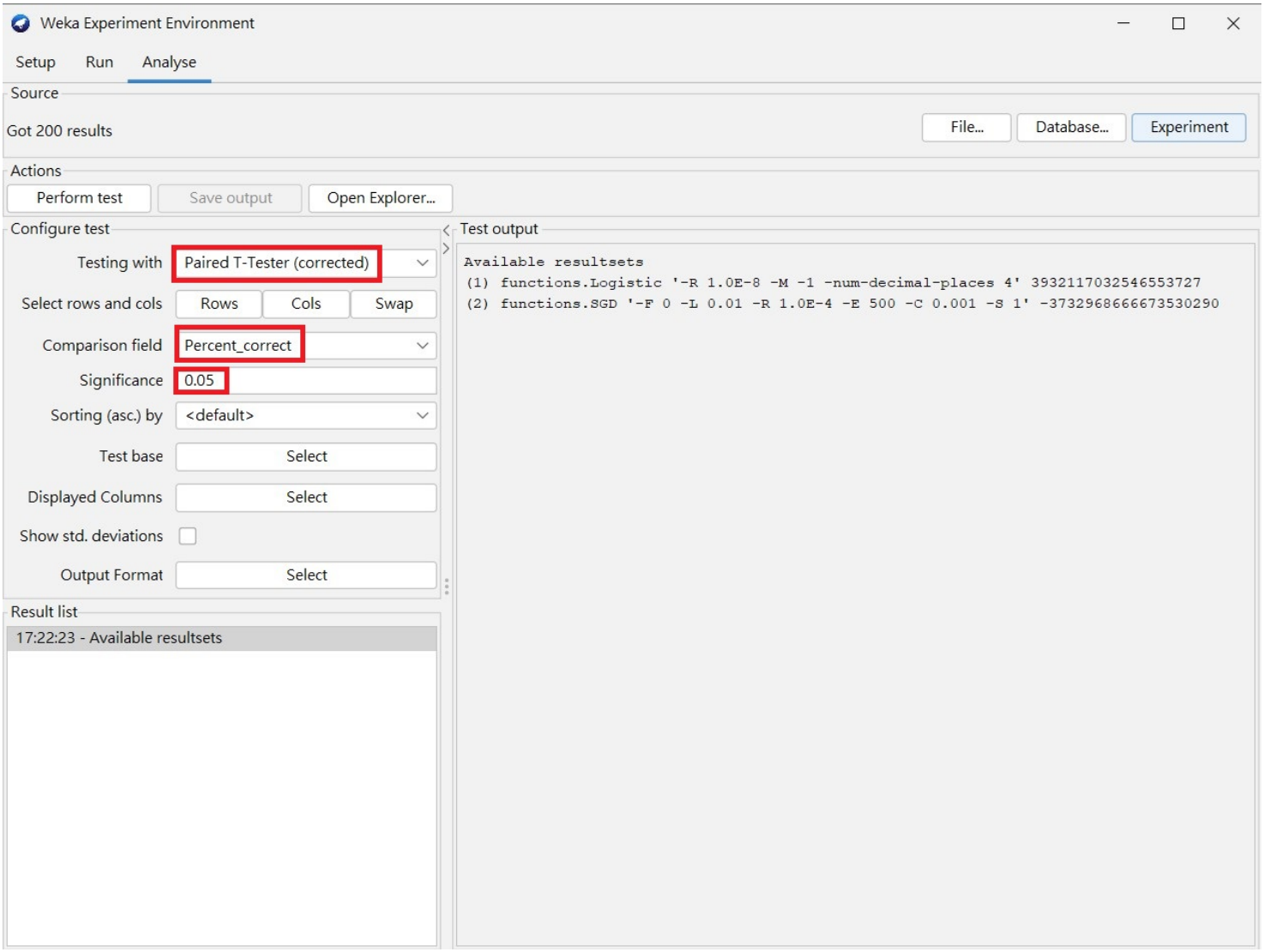

Weka Experiment Environment
Setup
Run
Analyse
Start
Stop
Log
17:20:55: Started
17:21:14: Finished
17:21:14: There were 0 errors
Status
Not running

3.6 分析

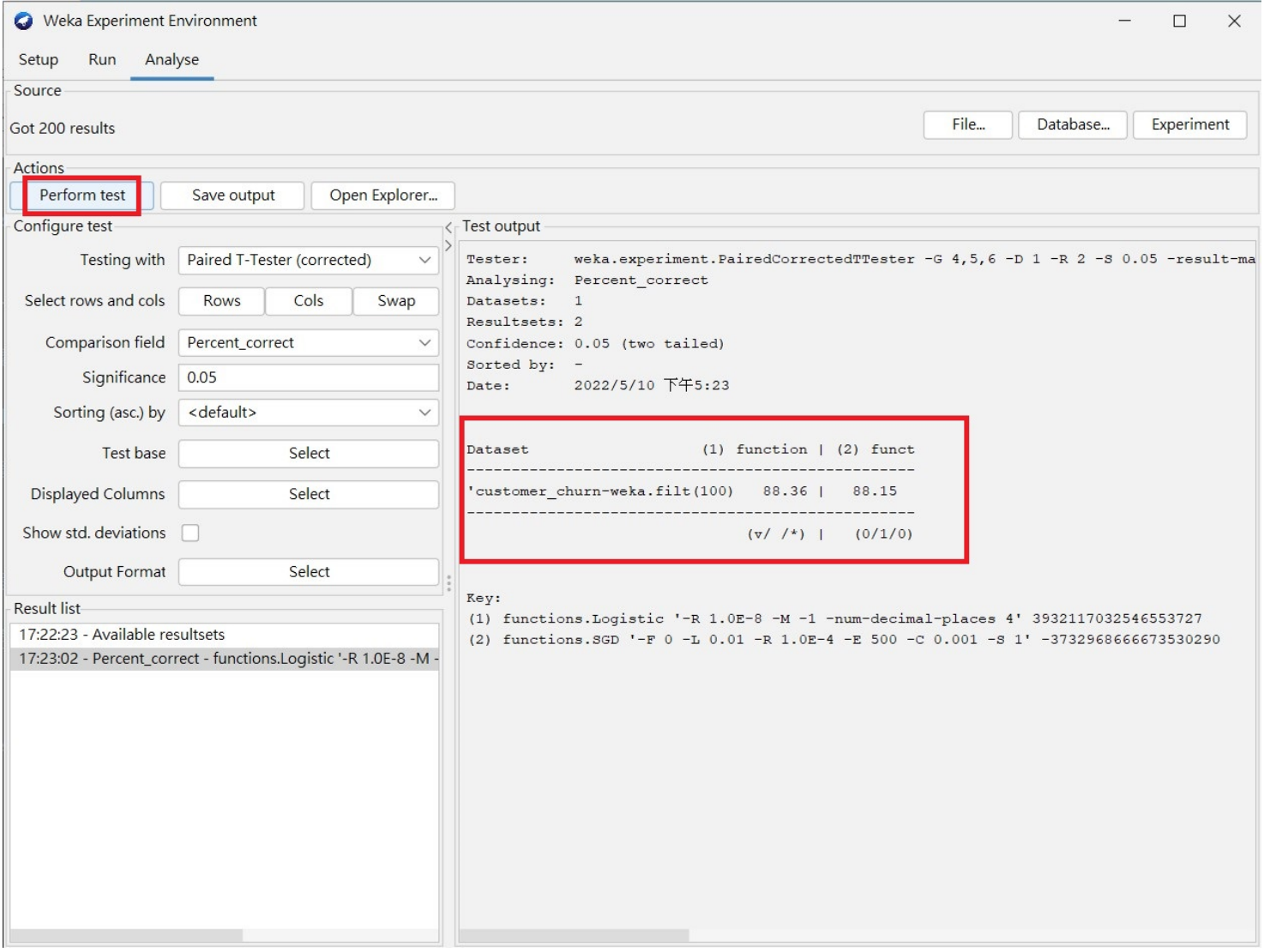
(點擊Experiment)



(做Paired T-test、參數調整)



3.7 輸出結果



4. 根據weka 的輸出說明結論

4.1 最終結果

Test output

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-ma
Analysing:  Percent_correct
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       2022/5/10 下午5:23

Dataset                    (1) function | (2) funct
----------------------------------------------------
'customer_churn-weka.filt(100)   88.36 |   88.15
----------------------------------------------------
                                (v/ /*) |   (0/1/0)

Key:
(1) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
(2) functions.SGD '-F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' -3732968666673530290
```

做完兩個模型的 Paired T-test 後，可以發現 SVM 和 Logistic Regression 在預測成功比例上的表現並沒有什麼差異，所以只要任選其一即可。