

1. 載入customer_churn.csv，列出資料筆數、屬性數量以及每個欄位的空值個數(5%)

資料屬性(numeric or nominal)

1: CustomerID	2: Churn	3: Tenure	4: PreferredLoginDevice	5: CityTier	6: WarehouseToHome	7: PreferredPaymentMode	8: Gender	9: HourSpendOnApp	10: NumberOfDeviceRegistered
Numeric	Numeric	Numeric	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Numeric
11: PreferredOrderCat	12: SatisfactionScore	13: MaritalStatus	14: NumberOfAddress	15: Complain	16: OrderAmountHikeFromlastYear	17: CouponUsed	18: OrderCount	19: DaySinceLastOrder	20: CashbackAmount
Nominal	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric

資料筆數：3083

Current relation	
Relation: customer_churn	Attributes: 20
Instances: 3083	Sum of weights: 3083

空值個數 (以下欄位是有空值的，如圖)

Selected attribute		
Name: Tenure		Type: Numeric
Missing: 153 (5%)	Distinct: 35	Unique: 3 (0%)
Selected attribute		
Name: WarehouseToHome		Type: Numeric
Missing: 154 (5%)	Distinct: 33	Unique: 1 (0%)
Selected attribute		
Name: HourSpendOnApp		Type: Numeric
Missing: 150 (5%)	Distinct: 6	Unique: 1 (0%)
Selected attribute		
Name: OrderAmountHikeFromlastYear		Type: Numeric
Missing: 131 (4%)	Distinct: 16	Unique: 0 (0%)
Selected attribute		
Name: CouponUsed		Type: Numeric
Missing: 126 (4%)	Distinct: 17	Unique: 3 (0%)
Selected attribute		
Name: OrderCount		Type: Numeric
Missing: 128 (4%)	Distinct: 16	Unique: 0 (0%)
Selected attribute		
Name: DaySinceLastOrder		Type: Numeric
Missing: 166 (5%)	Distinct: 20	Unique: 1 (0%)

2. 請刪除重覆多餘的資料(僅保留一筆)，並列出剩餘的資料筆數(5%)

使用RemoveDuplicates

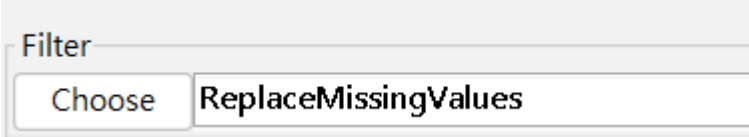
Filter
Choose RemoveDuplicates

資料總數變成3078

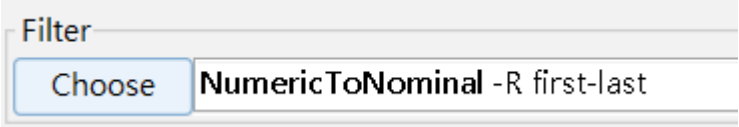
Current relation	
Relation: customer_churn-weka.filters.unsupervised.instance.RemoveDuplicates	Attributes: 20
Instances: 3078	Sum of weights: 3078

3. 資料前處理(5%)

使用ReplaceMissingValues 處理空值

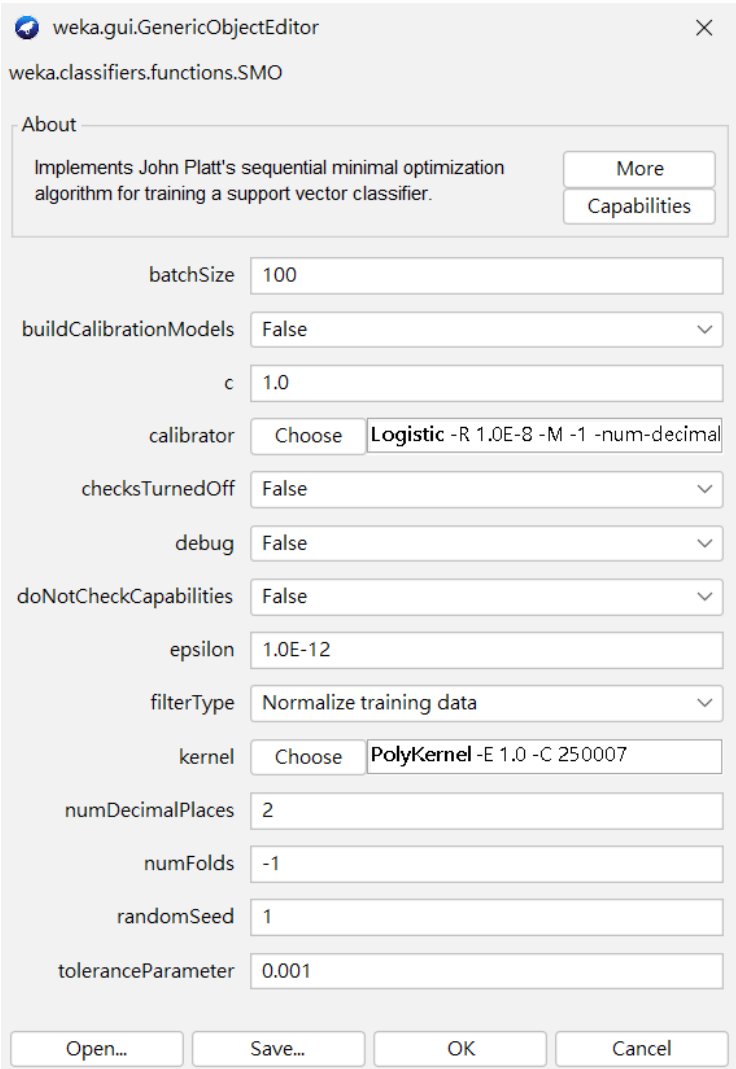


再用NumericToNominal



4. 訓練、測試SVM、Logistic Regression、Decision Tree模型，請以Accuracy評估模型表現(10%)

SVM (SMO)結果



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **SMO** -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation Folds

☒ Percentage split %

(Nom) Churn

Result list (right-click for options)

02:24:27 - functions.SMO

Classifier output

```

+ 0.3677 + (normalized) CashbackAmount*323
+ 0.2701 + (normalized) CashbackAmount*324
- 1.0331

Number of kernel evaluations: 31694050 (68.08% cached)

Time taken to build model: 166.85 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.07 seconds

=== Summary ===

Correctly Classified Instances      900      85.9559 %
Incorrectly Classified Instances    147      14.0401 %
Kappa statistic                    0.661
Mean absolute error                 0.1404
Root mean squared error             0.3747
Relative absolute error             33.0802 %
Root relative squared error         81.9446 %
Total Number of Instances          1047

=== Detailed Accuracy By Class ===


      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.906    0.251    0.895    0.906    0.901    0.661    0.828    0.877    0
      0.749    0.094    0.772    0.749    0.760    0.661    0.828    0.653    1
Weighted Avg.    0.860    0.204    0.859    0.860    0.859    0.661    0.828    0.811

=== Confusion Matrix ===

  a  b  <-- classified as
667 69 | a = 0
 78 233 | b = 1

```

Status OK

 x 0

Logistic Regression

weka.gui.GenericObjectEditor

weka.classifiers.functions.Logistic

About

Class for building and using a multinomial logistic regression model with a ridge estimator.

batchSize

debug

doNotCheckCapabilities

doNotStandardizeAttributes

maxIts

numDecimalPlaces

ridge

useConjugateGradientDescent

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseLogistic -R 1.0E-8 -M -1 -num-decimal-places 4

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation

☒ Percentage split

Set...

Folds10

%66

More options...

(Nom) Churn

Start

Stop

Result list (right-click for options)

02:24:27 - functions.SMO

02:34:05 - functions.Logistic

Classifier output

CashbackAmount=3193.133022127705087826

CashbackAmount=3201.4449368097422742812

CashbackAmount=3218.78622631971232825

CashbackAmount=3225.876361933509605824

CashbackAmount=32310.0304

CashbackAmount=32413.1064

Time taken to build model: 14.87 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances89285.1958 %

Incorrectly Classified Instances15514.8042 %

Kappa statistic0.6523

Mean absolute error0.1776

Root mean squared error0.3523

Relative absolute error41.8538 %

Root relative squared error77.0376 %

Total Number of Instances1047

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.880	0.215	0.906	0.880	0.893	0.653	0.891	0.947	0
	0.785	0.120	0.735	0.785	0.759	0.653	0.891	0.710	1
Weighted Avg.	0.852	0.187	0.855	0.852	0.853	0.653	0.891	0.876	

=== Confusion Matrix ===

a b <-- classified as

648 88 | a = 0

67 244 | b = 1

StatusOK

Log

Decision Tree

weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

More

Capabilities

batchSize100

binarySplitsFalse

collapseTreeTrue

confidenceFactor0.25

debugFalse

doNotCheckCapabilitiesFalse

doNotMakeSplitPointActualValueFalse

minNumObj2

numDecimalPlaces2

numFolds3

reducedErrorPruningFalse

saveInstanceDataFalse

seed1

subtreeRaisingTrue

unprunedFalse

useLaplaceFalse

useMDLcorrectionTrue

Open...

Save...

OK

Cancel

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

ChooseJ48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation

☒ Percentage split

Set...

Folds10

%66

More options...

(Nom) Churn

Start

Stop

Result list (right-click for options)

02:24:27 - functions.SMO

02:34:05 - functions.Logistic

02:36:11 - trees.J48

Classifier output

Tenure = 60: 0 (1.0)

Tenure = 61: 0 (1.0)

Number of Leaves : 790

Size of the tree : 839

Time taken to build model: 0.07 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances88484.4317 %

Incorrectly Classified Instances16315.5683 %

Kappa statistic0.6276

Mean absolute error0.2261

Root mean squared error0.3458

Relative absolute error53.2779 %

Root relative squared error75.6245 %

Total Number of Instances1047

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.889	0.260	0.890	0.889	0.889	0.628	0.871	0.927	0
	0.740	0.111	0.737	0.740	0.738	0.628	0.871	0.724	1
Weighted Avg.	0.844	0.216	0.844	0.844	0.844	0.628	0.871	0.867	

=== Confusion Matrix ===

a b <-- classified as

654 82 | a = 0

81 230 | b = 1

Status

OK

Log

x 0