

BUSINESS INTELLIGENCE LABORATORY

Lab Practice on Classification - Solution

Task 1: Preprocessing

2

- Split into training and test sets
 - ▣ Stratified sampling
- Preprocessing tasks
 - ▣ Discretization of continuous attributes
 - ▣ Replace missing values
 - ▣ Attribute selection
 - ▣ Oversampling
 - ▣ **Before or after the split?**
 - Typically after the split:
 - e.g., discretization must be done by looking at the values in the training set only

Task 2: Maximize accuracy

3

- Quality measure: accuracy
 - ▣ Majority classifier: 90%
 - ▣ Unbalanced class values
 - ▣ Predictive attributes not really discriminatory
 - Oversampling does not improve
 - Discretization does not improve
 - Parameters of Decision tree based methods do not improve
 - Bayes models do not improve
 - Meta-classifiers do not improve

Task 3: Revise objectives

4

- Majority classifier: € 64,80 per customer
 - Lower bound gain = $(2.250 * 72,00) / 2500$
- Oracle classifier: € 69,18 per customer
 - Upper bound gain = $(2.250 * 72,00 + 250 * 43,80) / 2500$
- Best models with Weka algs
 - € 65,10 CostSensitive (0, 20, 60, 0) + J48
 - € 65,23 CostSensitive (0, 1, 9, 0) + SimpleLogistic
 - € 65,26 no SCMW* att + CostSensitive (0, 1, 9, 0) + SimpleLogistic
 - € 65,41 CostSensitive (0, 1, 9, 0) + Jrip (10 folds)

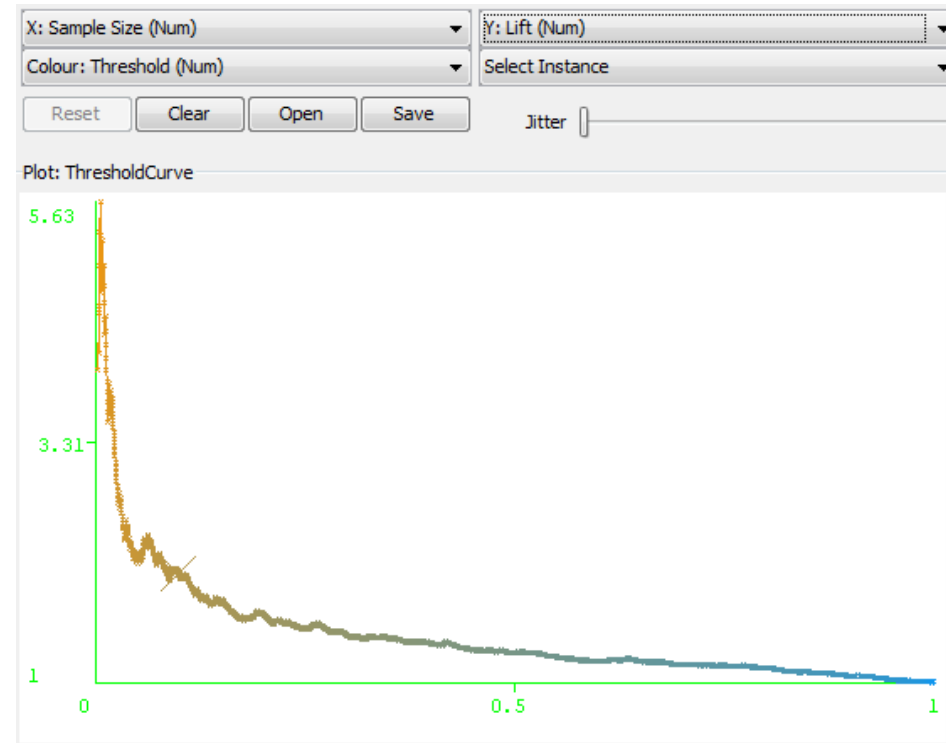
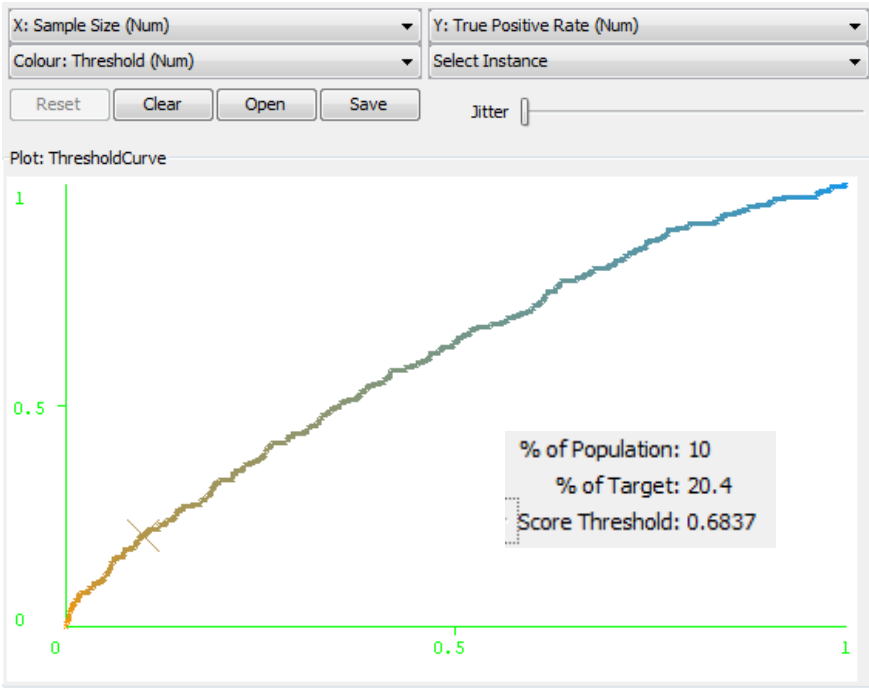
Task 5: Lift Chart

5

- Test set: 2500 cases
- 250 offers = 10% of the test set
- Random classifier: recall 10%
- CostSensitive (0,1,9,0)+ SimpleLogistic:
 - ▣ Recall ~ 20.4%
 - ▣ Lift ~ 2.04
 - More than twice the cancelers are reached wrt random offers

Lift Chart

6



Benefit Chart

7

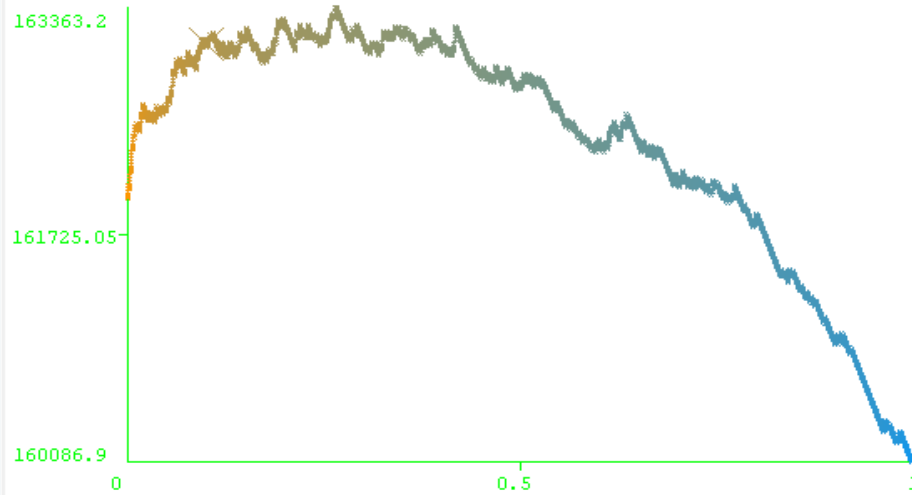
Confusion Matrix

	Predicted (a)	Predicted (b)	
Actual (a): yes	51 2.04%	199 7.96%	
Actual (b): no	199 7.96%	2051 82.04%	

Classification Accuracy: 84.08%

X: Sample Size (Num) Y: Cost/Benefit (Num)
 Colour: Threshold (Num) Select Instance
 Reset Clear Open Save Jitter

Plot: Cost/Benefit Curve



Score Threshold

% of Population: 10
 % of Target: 20.4
 Score Threshold: 0.6837

Cost Matrix

	Predicted (a)	Predicted (b)	
43.8	0	Actual (a)	
66.3	72	Actual (b)	

Total Population: 2500

Benefit: 163099.5
 Random: 161812.5
 Gain: 1287

Maximize Cost/Benefit

Minimize Cost/Benefit

Cost Benefit

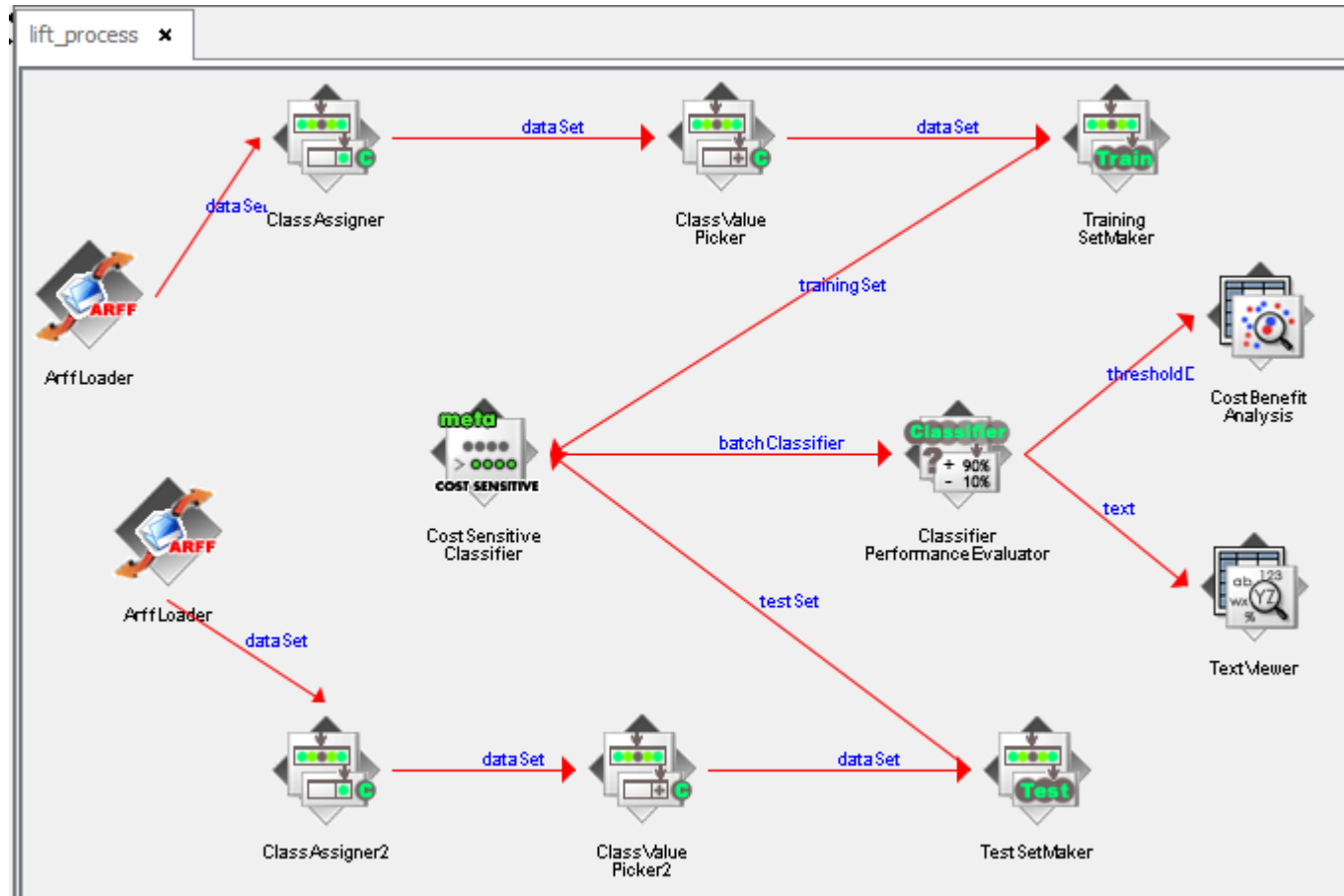
Benefit at 10% is
 $43.8 * 51 + 66.3 * 199 + 0 * 199 + 72 * 2051 = 163099.5$

Benefit matrix (rows/columns are swapped!)

Benefit chart

Knowledge Flow Process

8



Knowledge Flow Process

9

□ Important!

- By setting canceler=yes in the ClassValuePicker steps, classes are swapped in the confusion matrix
- Hence, they must be swapped in the cost matrix too!

