BUSINESS INTELLIGENCE LABORATORY

Lab Practice on Classification - Solution

Task 1: Preprocessing

- 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: Maximaze accuracy

- 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

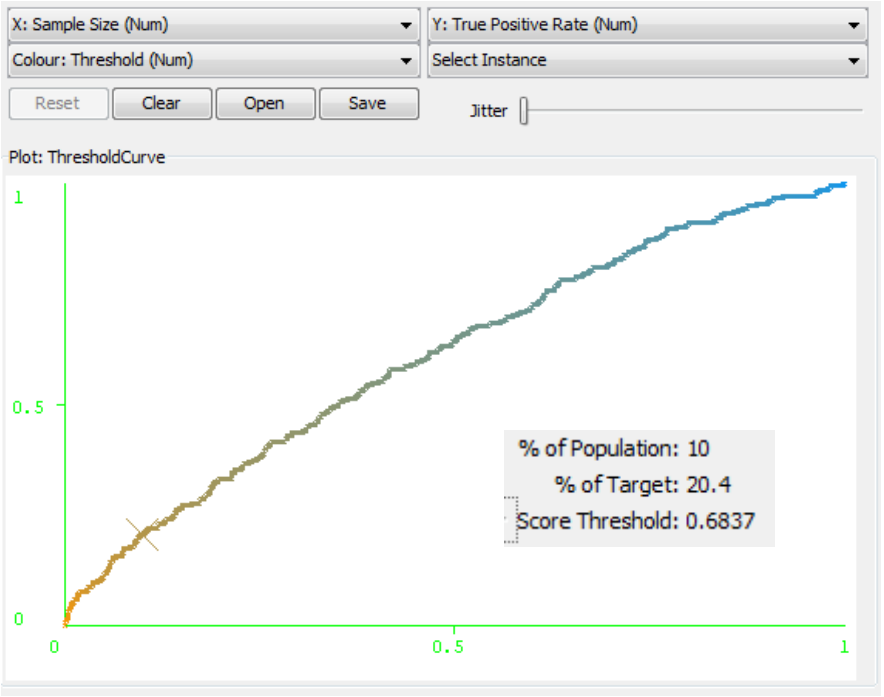
Task 3: Revise objectives

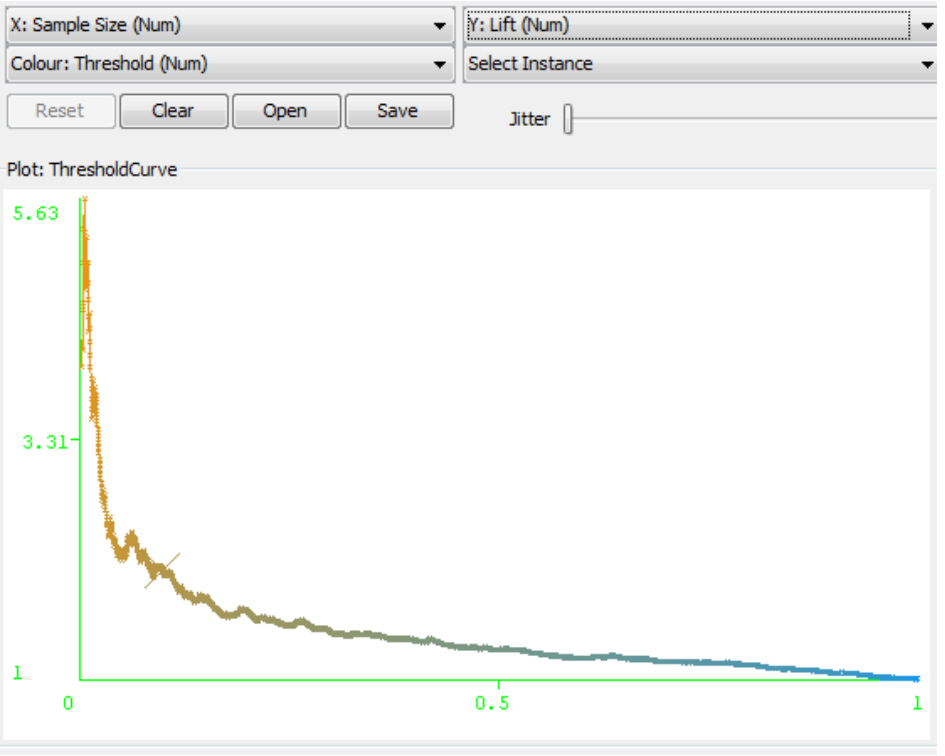
- Majority classifier: € 64,80 per customer
  - Lower bound gain = (2.250 * 72,00) /2500
- Oracle classifier: € 69,18 per customer
  - Upper bound gain = (2.250 * 72,00 + 250 * 43,80) /2500
- Best models with Weka algs
  - € 65,10 CostSensitive (0, 20, 60, 0) + J48
  - € 65,23 CostSensitive (0, 1, 9, 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

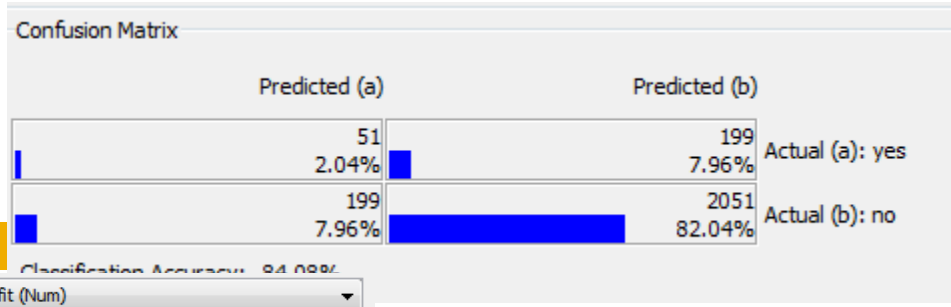
- Test set: 2500 cases
- 250 offers = 10% of the test set
- Random classifier: recall 10%
- CostSensitive (0,1,9,0)+ SimpleLogistic:
  - Recall ~ 20.4%
  - Lift ~ 2.04
    - More than twice the cancelers are reached wrt random offers

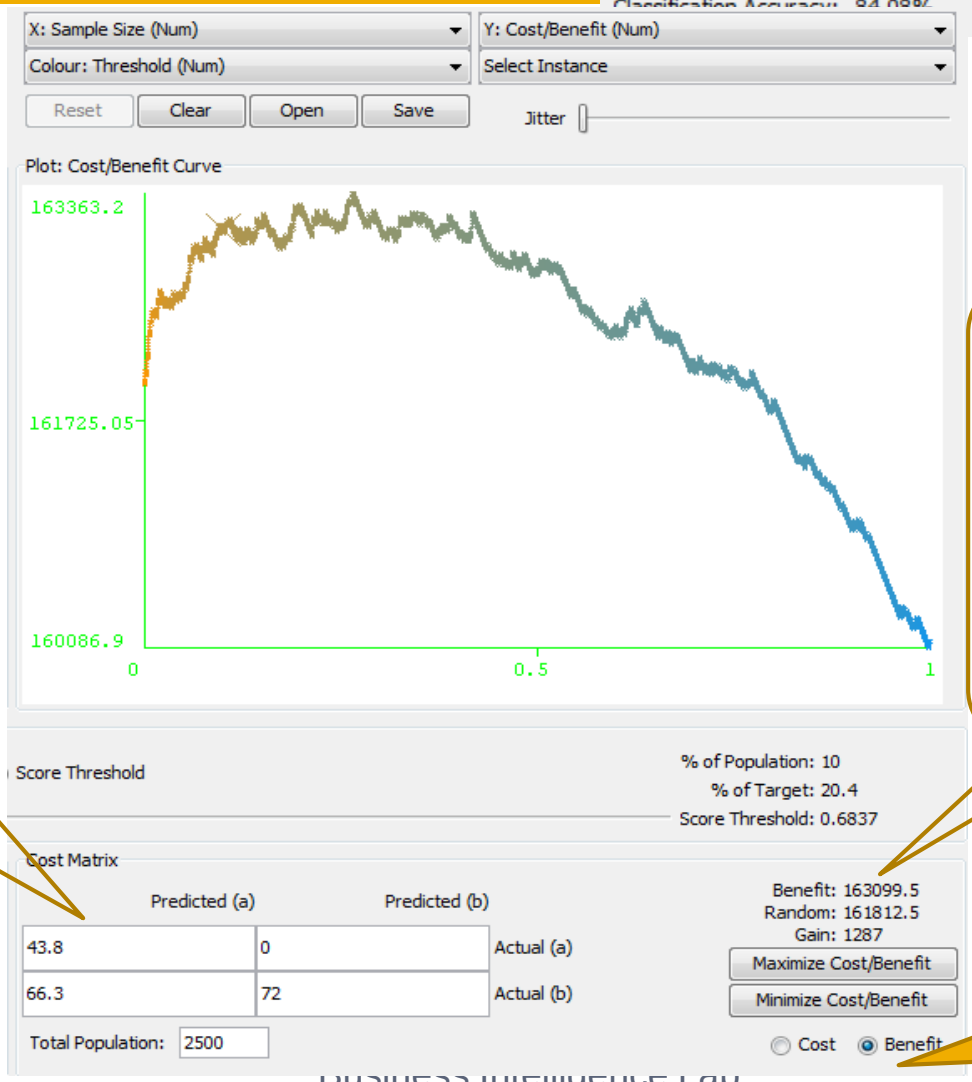
Lift Chart


X: Sample Size (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: ThresholdCurve
1
0.5
0
0
0.5
1
% of Population: 10
% of Target: 20.4
Score Threshold: 0.6837


X: Sample Size (Num)
Y: Lift (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: ThresholdCurve
5.63
3.31
1
0
0.5
1

Benefit Chart



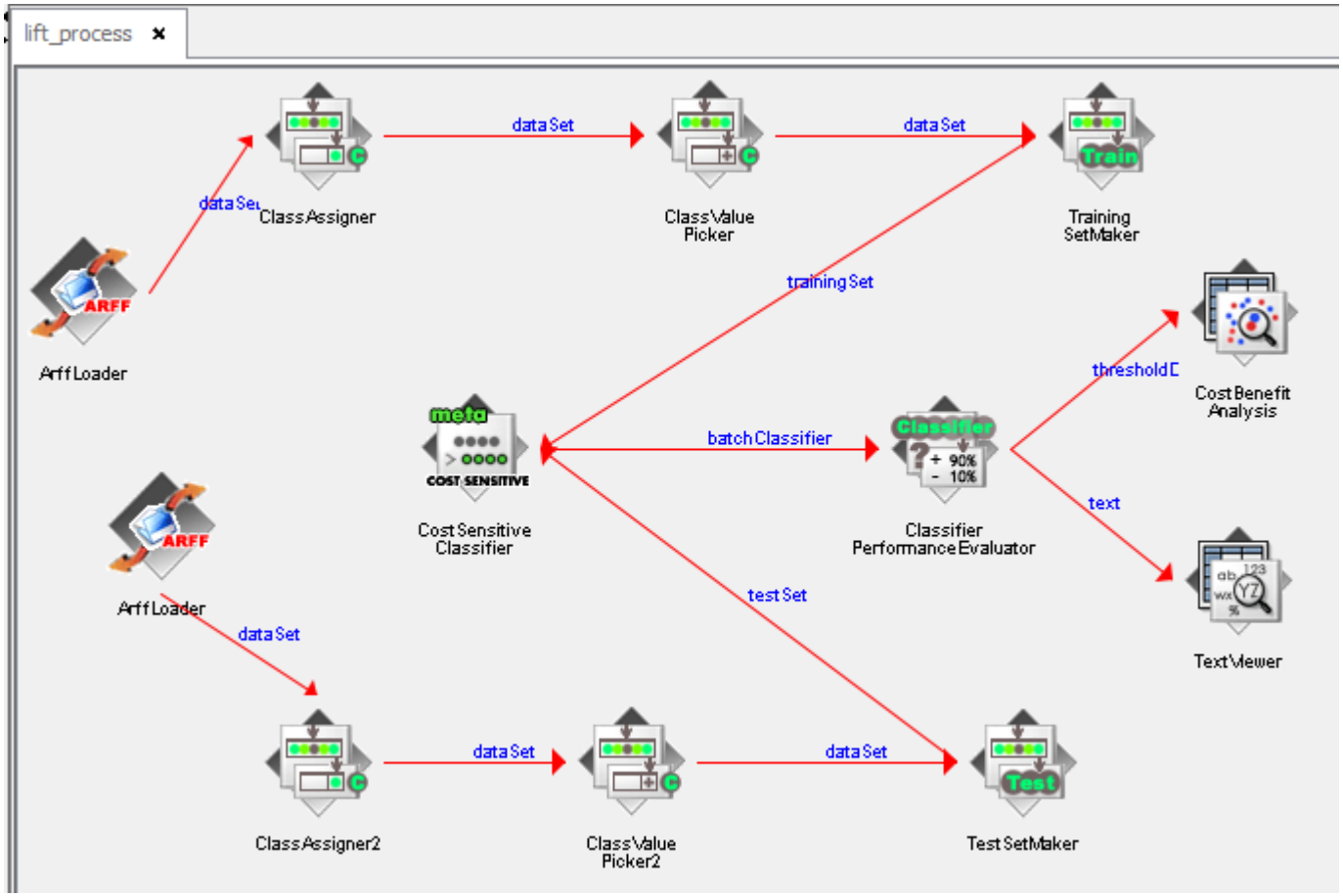


Benefit matrix (**rows/columns are swapped**!)

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

Benefit chart

Knowledge Flow Process



Knowledge Flow Process

- 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!