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: Maximize 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 = \((2.250 \times 72,00) / 2500\)

- **Oracle classifier:** € 69,18 per customer
  - Upper bound gain = \((2.250 \times 72,00 + 250 \times 43,80) / 2500\)

- **Best models with Azure algos**
  - € 65,29 CostSensitive (0,9,1,0) Bayes + Two-Class Decision Forest (100 trees)
  - € 65,26 CostSensitive (0,1,9,0) Bayes + Two-Class Decision Jungle (100 trees)
  - € 65,04 CostSensitive (0,1,9,0) Bayes + Two-Class Boosted Decision
  - € 64,86 CostSensitive (0,1,9,0) Bayes + Two-Class Logistic Regression

- Numbers can differ for different train-test split datasets
Task 5: Lift Chart - AZURE

- Test set: 2500 cases
- 250 offers = 10% of the test set
- Random classifier: recall 10%
- CostSensitive (0,9,1,0)+ Two-Class Boosted Decision:
  - Recall ~ 19.5%
  - Lift ~ 1.95
    - Almost twice the cancelers are reached wrt random offers
Task 3: Revise objectives

- **Majority classifier: €64,80 per customer**
  - Lower bound gain = \((2.250 \times 72,00) / 2500\)

- **Oracle classifier: €69,18 per customer**
  - Upper bound gain = \((2.250 \times 72,00 + 250 \times 43,80) / 2500\)

- **Best models with Weka algs**
  - €65,10 CostSensitive \((0, 20, 60, 0)\) + J48
  - €65,23 CostSensitive \((0, 1, 9, 0)\) + SimpleLogistic
  - €65,26 no SCMWh* 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
Gain & Lift Chart

Gain Chart

Lift Chart
Benefits/Cost Chart

Profit of with top 10% of customers: 163288.5
Lift Chart - WEKA
Benefit Chart

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

Benefit matrix (rows/columns are swapped!)
Knowledge Flow Process - WEKA
Knowledge Flow Process - WEKA

- 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!