BUSINESS INTELLIGENCE LABORATORY

Practice on a Classification Problem
Dataset

- **ee_dataset.arff**
- A dataset of 7,500 customers of a German electric power company
- Some customers intend to cancel their subscription (attribute *canceler*)
- A special promotion consisting of a discount on the price of electricity must be planned to prevent cancelers to abandon.
Task 1: Preprocessing

- Split the dataset into training and test
- Investigate the meaning of attributes from the provided documentation
- Study the distribution of data and the relevance of attributes
- If needed, create derived attributes
Task 2: Maximaze accuracy

- Extract a classification model that predicts whether a customer is a canceler, so that its accuracy is maximized.
Task 2: Classification methods

- Classification model
  - Decision tree, NaiveBayes, Metaclassification
- Parameters of classification algorithm
  - tree pruning, confidence, stop earlier, ...
- Input dataset:
  - Preprocessing on attributes (selection, derived, …)
  - Preprocessing on instances (missing values, oversampling)
- …
Task 3: Revise objectives

- Is it really the accuracy that one intends to maximize?
  - Maximize the following gain function:

<table>
<thead>
<tr>
<th></th>
<th>No offer sent</th>
<th>Offer sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Canceler</td>
<td>72,00 Euro</td>
<td>66,30 Euro</td>
</tr>
<tr>
<td>Canceler</td>
<td>0 Euro</td>
<td>43,80 Euro</td>
</tr>
</tbody>
</table>
Task 4: Descriptive use

Does your classifier provide a description of the profiles of canceler customers?
Task 5: Lift Chart

- Assume to have limited amount of resources, so that at most 250 offers can be sent out. How many cancelers does your classifier can reach?
Task 6: Validation set

- Answer to Task 3 and Task 5 using as test set a totally new set of data \((\text{ee\_validation.arff})\). How do the performances of your classifier change?