META COST-CLASSIFICATION

EXAMPLE OVER GERMAN CREDIT CARD DATASET

German Credit Data

- Available at UCI repository but also in Azure ML Studio

- Contains observations on 30 variables for 1000 past applicants for credit

- Each applicant was rated as "**good customer**" (700 cases) or "**bad customer**" (300 cases).

- Develop a classifier to determine if a new applicant is a good customer or a bad customer

The German Credit Data

- We have **Good Customer** (who will pay) and **Bad Customer** (who will not pay)

- The majority class is **Good Customer** so for a classification is hard to classify a **Bad Customer** because we have a very low number of **Bad Customers**

COST-MATRIX

- **The problem in this case is when we predict a customer as good, but in reality it is bad.**

- So we need to assign a bigger weigth to this kind of missclassification, like in the matrix below:

Classified as ➔	Bad Customer	Good Customer
Bad Customer	0	**5**
Good Customer	1	0

Meta Cost: cost-sensitive

- On the train data apply a classifier getting probability of a class label *P(j|x)*

- Compute expected risk of classifying x with class i:

$$R(i|x) = \sum_j P(j|x)C(i,j)$$

- Re-label the train data with the class i having lower risk

- Learn a model on the cost-sensitive train data