Data Mining II
Data Mining Cup – Let's Go!

Heiko Paulheim, Robert Meusel
Requirements

• Final exam
  – 50 % written exam
  – 50 % project work

• Project work
  – work on DMC tasks
  – we meet every Monday to discuss the current progress

• Presentations
  – Three intermediate presentations
    • open questions, problems, current results (numbers in 10-fold CV)
  – one final presentation
  – everybody has to present once during those four presentations

• Final report
  – 10 pages per team
  – solutions, results, lessons learned
DMC Timeline

• Today: First look at the task, organization
  – Build teams
  – Understand the task

• April 25\textsuperscript{th}: Intermediate presentation & discussion
• May 2\textsuperscript{nd}: Intermediate presentation & discussion
• May 9\textsuperscript{th}: Intermediate presentation & discussion
• May 12\textsuperscript{th} (Thursday): pre-submission of results to Robert and Heiko
• May 18\textsuperscript{th} (Wednesday): DMC Deadline
• May 19\textsuperscript{th} (Thursday): submission of final report
• May 23\textsuperscript{th}: Final presentation
Project Grading

• Projects will be graded based on
  – Innovation of ideas created and pursued
  – Intermediate and final presentations
  – Quality of the final report

• We will have 3-4 teams, but joint meetings
  – You are allowed to use ideas from the other teams
    • but you have to mark them in the final report
  – And you send us your slides of each intermediate presentation
    • so that we can track the origin of ideas
Individual Grading

• In each team, there may be smaller sub teams working on different tasks
  – In each presentation, you have at least one slide per sub team / task
  – With names!

• Peer grading
  – At the end of the project, you will give grades to your team mates
  – Your grades will be kept secretly
  – We only use them to confirm (and, if necessary, adjust) our assessment
Let's Get Started with the Task

• You have looked at the data
• ...and read the task

• Question 1: what does the data look like?
  – we need a detailed profile until next week

• Question 2: what will the overall approach look like?
  – Classification vs. regression
  – Sampling?
  – Models per product group?

• Question 3: which features do we use?
  – we start off with a brainstorming today
  – then, prepare some preliminary study until next week
Question 1: What does the Data Look Like?

- Detailed questions include, but are not limited to
  - Are there unseen products in the test set?
  - Are there unseen customers in the test set?
  - How is the distribution of products/customers, e.g., are there top selling products and/or larger customers?
  - How many distinct products etc. are there?
  - How are product groups distributed?
  - Which vouchers are there?
  - Distribution and impact of missing values?

- Plus: what is the performance of a default model?
Question 2: Which Overall Approach?

- Sampling
  - The data is probably too large to handle in experiments
  - We need to create meaningful samples
  - Try to avoid tearing apart orders, customers, products…

- Important: make sure all teams use the same samples!

- We need to predict a numerical value
  - Idea 1: regression
  - Idea 2: classification 0/1 (only a small fraction (0.2%) is >1!)
Question 3: What Features can we Use?

• Let's make some brainstorming:
  – dealing with sizes
  – features for the customer
  – features for the product group
  – features for the order
  – features for the voucher?
Now You Know what to Do!

It's... Time!

To Go to Work!