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 week to discuss the current progress

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

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

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

• April 20<sup>th</sup>: Intermediate presentation & discussion
• April 27<sup>th</sup>: Intermediate presentation & discussion
• May 4<sup>th</sup>: Intermediate presentation & discussion
• May 11<sup>th</sup>: Intermediate presentation & discussion
• May 13<sup>th</sup> (Wednesday): Internal submission of DMC solutions to Robert and Heiko
• May 19<sup>th</sup> (Tuesday): DMC Deadline
• May 25<sup>th</sup> (Monday): Submission of final report
• May 26<sup>th</sup> (Tuesday): 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 two teams, but joint meetings
  – You are allowed to use ideas from the other team
    • 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 will 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?
  – Four individual models?
  – A chain of models?
  – Use three individual models for all three coupons, or one for all?
• 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 users in the test set?
  - How is the distribution of products/users/brands, e.g., are there top sellers/users/brands?
  - How many distinct products, brands, categories are there?
  - How are categories distributed?
- Plus: what is the performance of a default model?
Question 2: Which Overall Approach?

- We need to predict four variables
  - coupon1, coupon2, coupon3 redeemed or not (binary)
  - overall basket value

- Possible variants (non-exhaustive)
  - Four individual prediction problems
  - Predict coupon redemption first, use results in basket value prediction
  - Predict basket value first, use result in coupon redemption prediction
Question 2: Which Overall Approach?

- We need to predict four variables
  - coupon1, coupon2, coupon3 redeemed or not (binary)
  - overall basket value

- Three vs. one model for coupon1..3
  - first look: do they differ somehow?
  - e.g., different product categories?

- Variants for coupon redemption (0/1 variable)
  - ordinary classification
  - classification with confidence, use confidence
  - regression, with ex post restriction to [0;1] interval
Question 3: What Features can we Use?

• Let's make some brainstorming:
  – dealing with categories (comma separated list!)
  – features for the user
  – features for the product
  – features for the coupon
  – features for the brand
  – features for the order
Now You Know what to Do!

It's... Time!

To Go to Work!

04/16/15 Heiko Paulheim, Robert Meusel