Technische Universität München

Business Analytics & Machine Learning

Analytics Cup 2024

Decision Sciences & Systems (DSS)

Department of Informatics

Technical University of Munich

Analytics Cup

Dates and General Information

Dates

- From **December 18th , 2023 at 6:00 pm** until **January 29th , 2024 at 8:00 am (submission deadline)**
- Registration opens December 18th, 6:00 pm at https://analytics-cup.dss.in.tum.de

What is it?

- Students team up in groups of up to 4
  - You are allowed to participate on your own (just make a 1-person team), but we encourage you to find team mates!
- A dataset with a classification task will be published
- Each team solves the task and uploads its solutions
- Be prepared to present explain your solution

What is the benefit?

- Firsthand experience with real-world data and analytics task
- Receive bonus for your grade
  - This project will be graded for each team. If the project grade is better than your final exam grade, the project will be weighted by 33% and the exam by 67%. If not, only the exam will count. -> You can only improve your grade!
  - **Bonus is only valid for first attempted AND passed exam in 2024!**

Analytics Cup Grading

- A bonus grade (1.0 – 5.0) will be given with respect to the achieved *balanced accuracy* on the *private test data* set. If better then exam grade, the bonus will count 33% percent of final grade for the course.
- Your source code will **not** be graded directly, but **must reproduce your submission.** (more below)
- You can make up to 10 submissions. The grade is based only on your team's **best** submission in terms of Balanced Accuracy.
- If you are *registered* for the Endterm Exam (29 February) AND receive a failing grade, your AC bonus will NOT apply to the Retake Exam (tba). (If you are not registered for the February exam, your bonus remains valid for the Retake Exam. If you have already registered for the Endterm exam but no longer plan to attend, you can *deregister* in TUMonline until 22.02.2024.)

Examples:

- Endterm Exam grade (2.0), Analytics Cup grade (2.7)
  AC not considered, because worse than exam grade. → Final grade: **2.0**
- Endterm Exam grade (2.0), Analytics Cup grade (1.0)
  The bonus applies. → *Final grade: 0.67*2.0+0.33*1.0=1.67, rounded down to* ***1.6***
- Endterm Exam: *not registered*, Retake Exam grade (2.0), Analytics Cup grade (1.0)
  Bonus applies to retake exam. → *Final grade:* ***1.6***
- Endterm Exam grade (5.0), Retake Exam: not registered, Analytics Cup grade (1.0)
  Bonus does NOT apply to a failed exam. → Final grade **5.0**.
- Endterm Exam grade (5.0-X), Retake Exam grade (2.0), Analytics Cup grade (1.0)
  Student r*egistered for Endterm exam but then didn't attend (without a medical exception or similar). The bonus no longer applies to Retake exam.* → *Final grade* ***2.0***.

Detailed examples at: https://www.moodle.tum.de/mod/page/view.php?id=2670007

Performance Measure

Performance of your submission will be measured by **balanced accuracy**

$$BAC = \left(\frac{Recall + Specificity}{2}\right)$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Rules of the Analytics Cup

Teams

- Team size: 1 – 4 members.
- **All team members must have joined the team before the first submission!**
  (After your team's first submission, no membership changes are possible.)
- Each student can only be member of one team.
- Looking for a team or additional team mates? AC Moodle Subforum open now.

Submissions

- Maximum number of valid submissions per team: 10. You must submit your predictions AND your code. Your submitted code file must **reproducibly** generate the submitted predictions!
  See detailed instructions provided with the data for details!
  E.g.: run the file in order, do not change data file names, use random seeds, …
- Besides, prepare a presentation, where you explain your approach (~5 minutes)

Reasons for disqualification:

- **Non-reproducible** submissions (submitted predictions **must be reproducible** using the submitted script)
  (If there are reproducibility issues with your script, we will first invite your team to a meeting on January 31 or February 1. If the issues can be sorted out in the meeting and we decide that you have acted in good faith, you will not be disqualified.)
- **Hard-coded** classifications (even if the best ranked submission is not hard-coded!)
- **Copying or cooperation** from/with other groups (disqualification of both teams)
- **Refusal/Failure** to present your solution in the final lecture after winning the cup. (see slide 20)
- Usage of **Automatic Machine Learning** is forbidden.

Explainability of Results

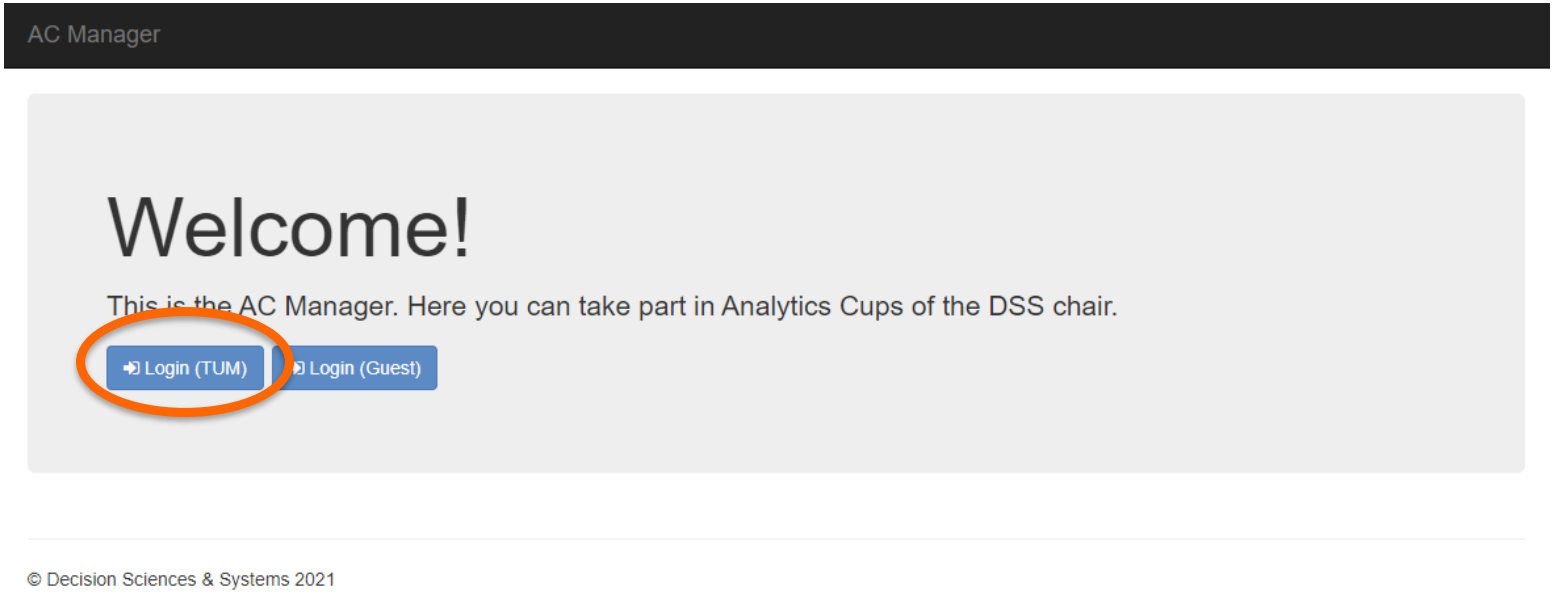
- The development of a black-box model that generates predictions is in many situations insufficient.
  - Businesses try to predict which customers are likely to buy a product. Understanding the reasons behind this could reveal potential for marketing campaigns.
  - It is quite common that doctors utilize predictive models to identify the disease of a patient. However, the data input of those systems is limited, and thus, doctors need to understand which factors influence the prediction to determine the optimal treatment for a patient.

- For this reason, you should be able to explain your prediction.
  - We will sample several teams and ask them to present their prediction and the corresponding explanation either in the final lecture or in a separate meeting.
  - You can find an example regarding the explainability in homework 7.

Build Team in AC Manager

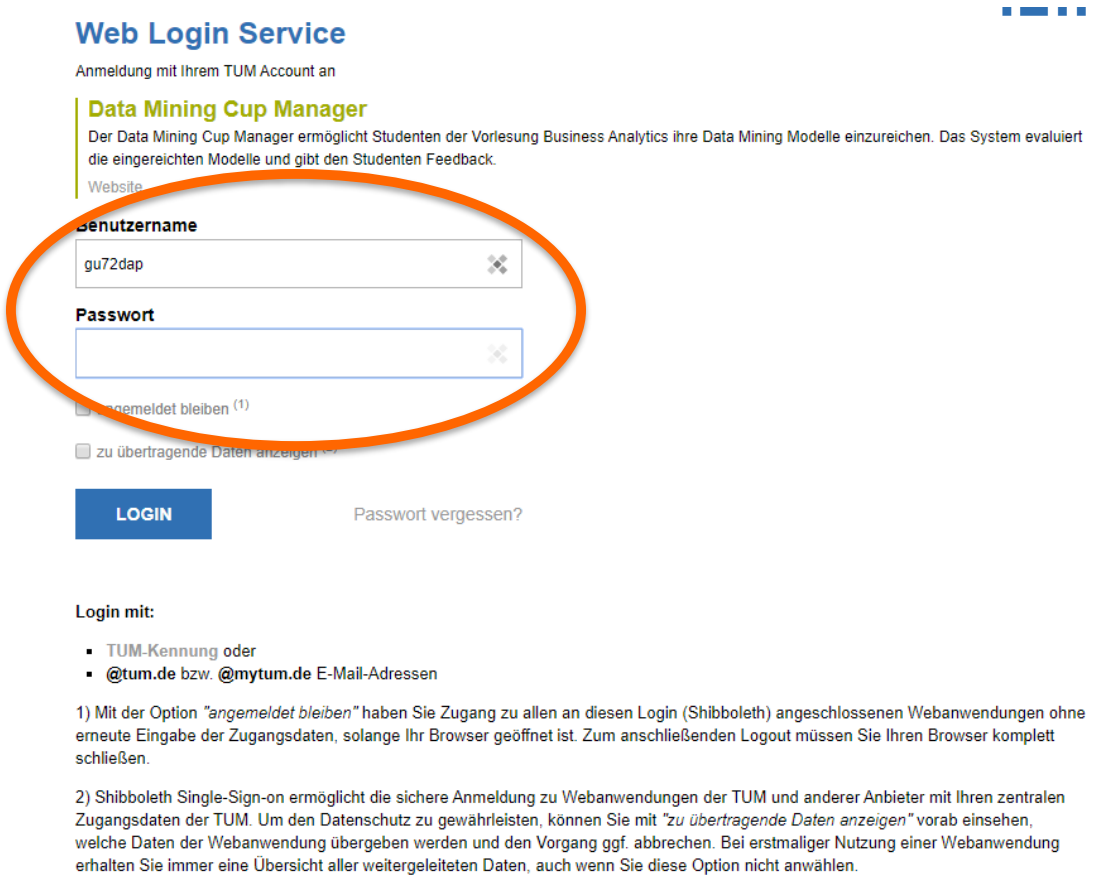
Login with your TUM login data (“TUM Kennung”)

https://analytics-cup.dss.in.tum.de



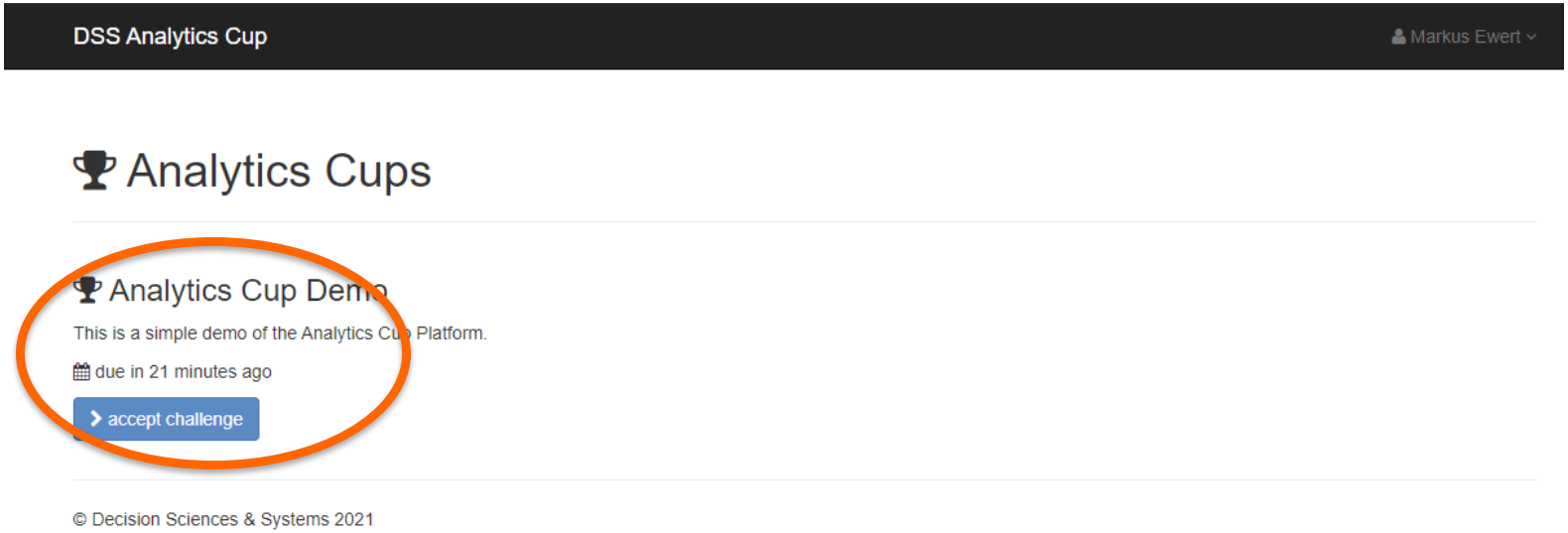
Build Team in DMC Manager

Login via “Shibboleth” with your TUM login data (“TUM Kennung”)



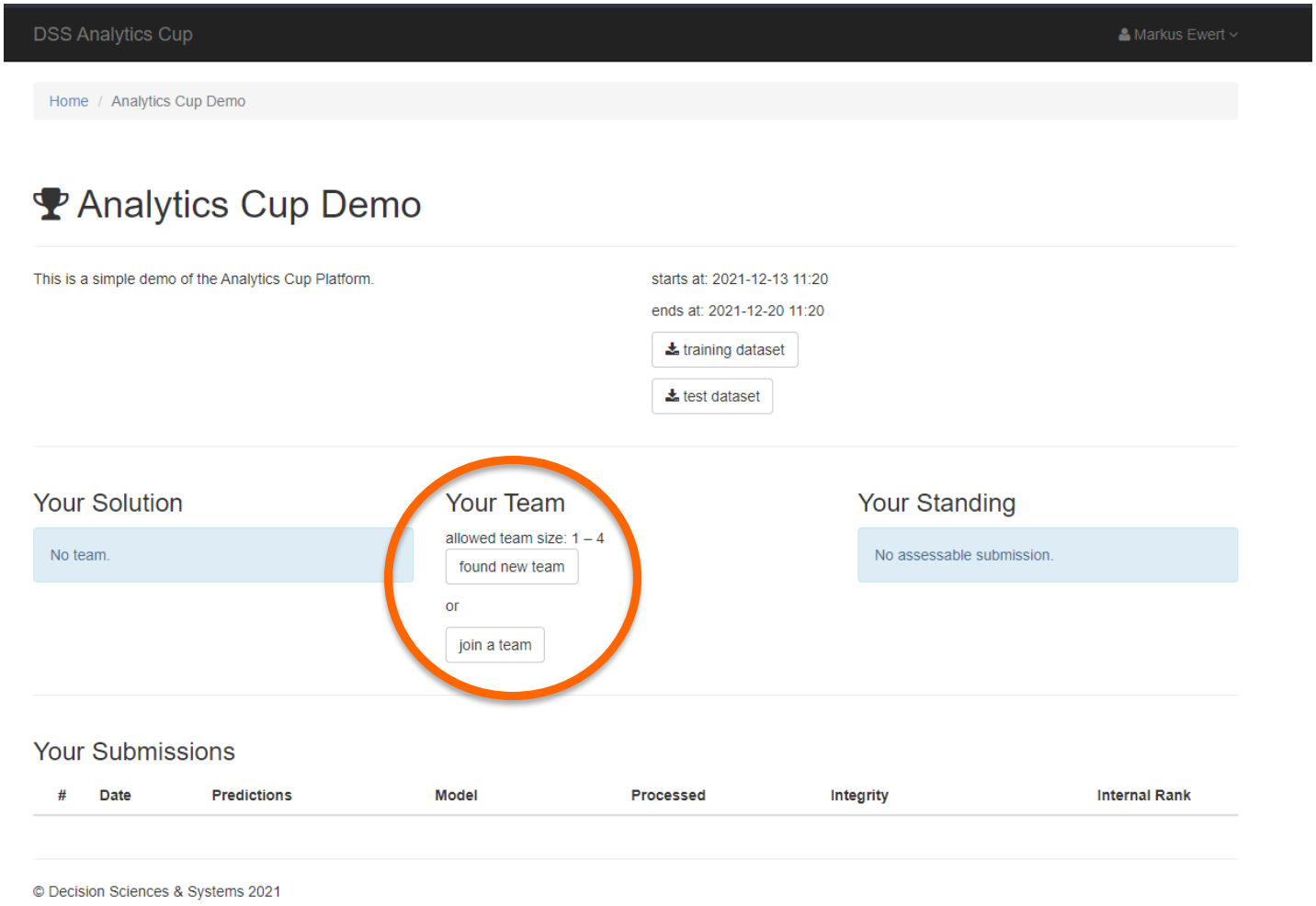
Build Team in DMC Manager

Choose the AC instance Analytics Cup 2022 in the AC Manager



Build Team in DMC Manager

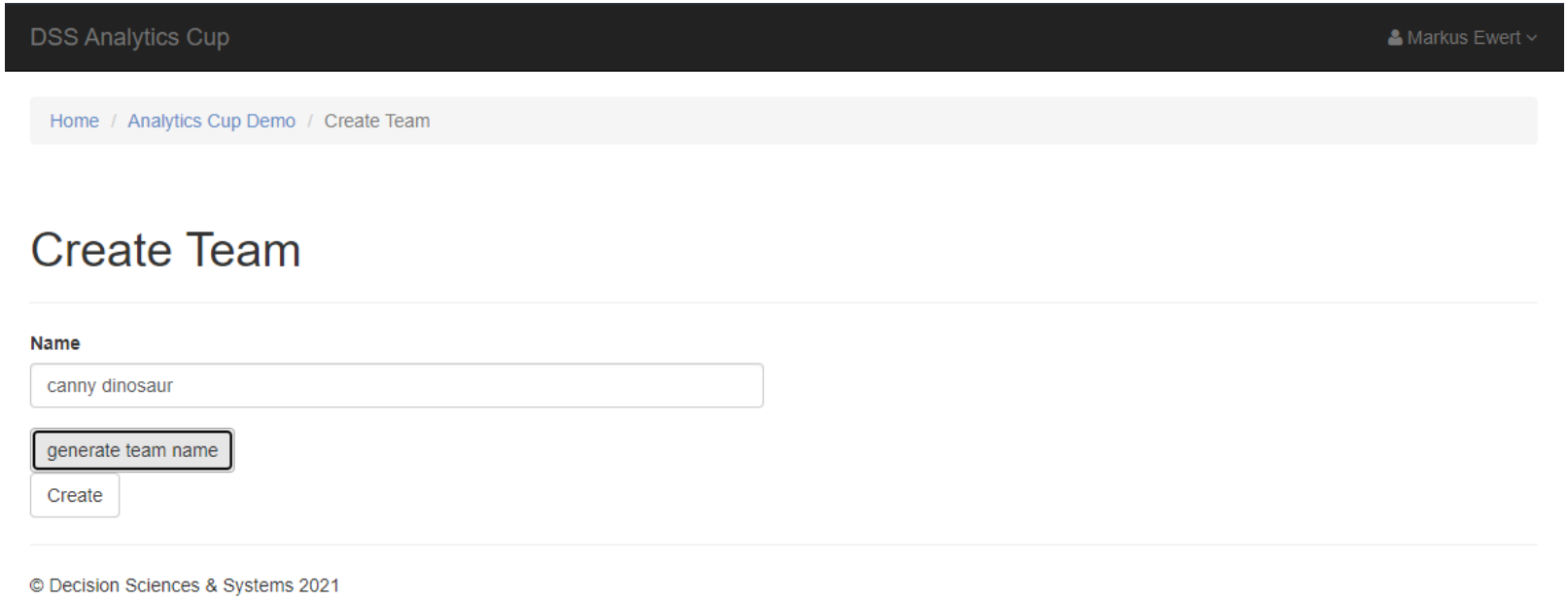
Found new team or join an existing team



Build Team in DMC Manager

Creating a new team

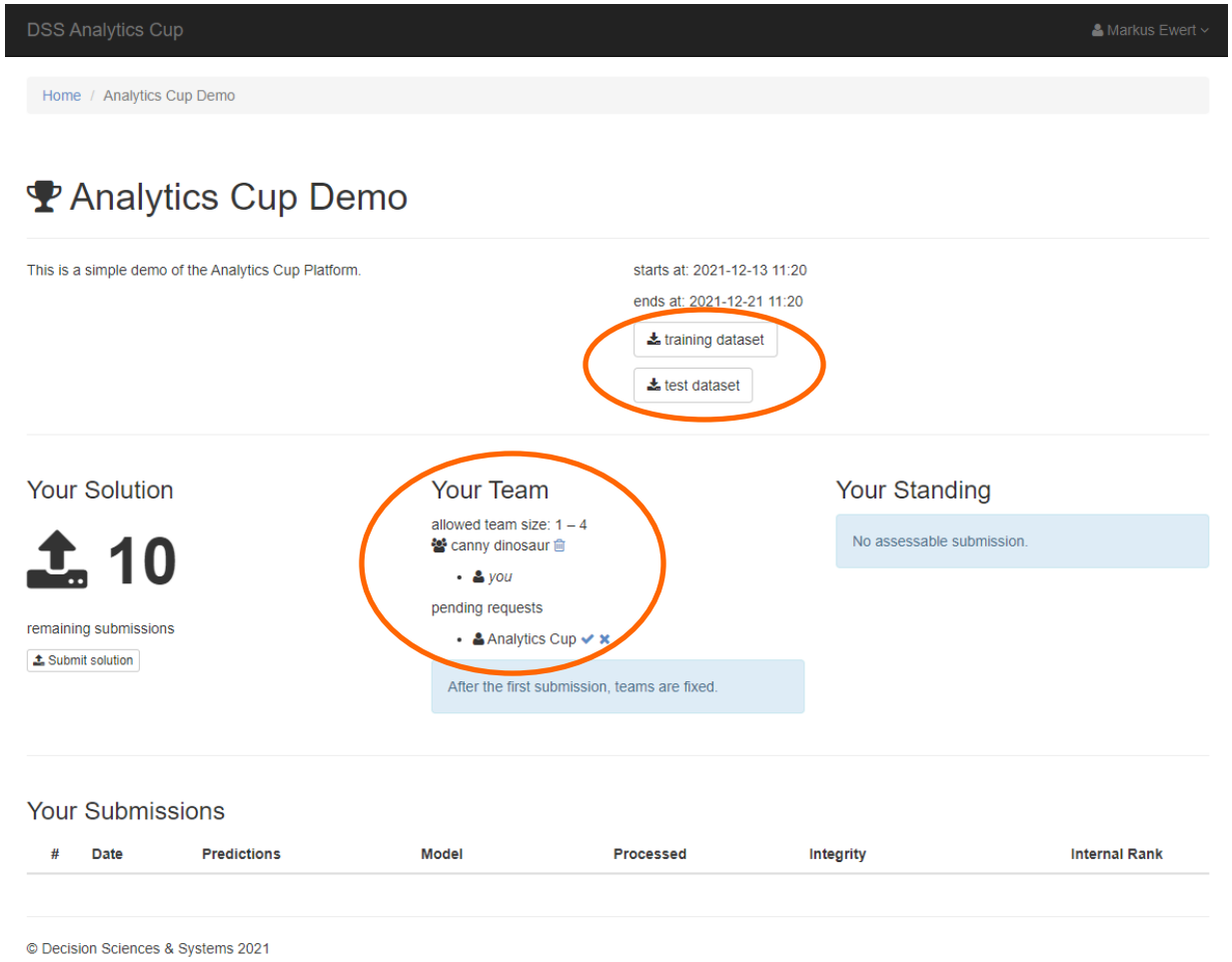
- Team size: 1-4 members



Load the Data Set

Download the training and test datasets from the AC Manager

Note: The full public test set is included in the zip accessed via the training data button. The 'test dataset' button download only contains instance IDs of the test set.



Build your model and export the predictions

when your team is ready to make a first submission…

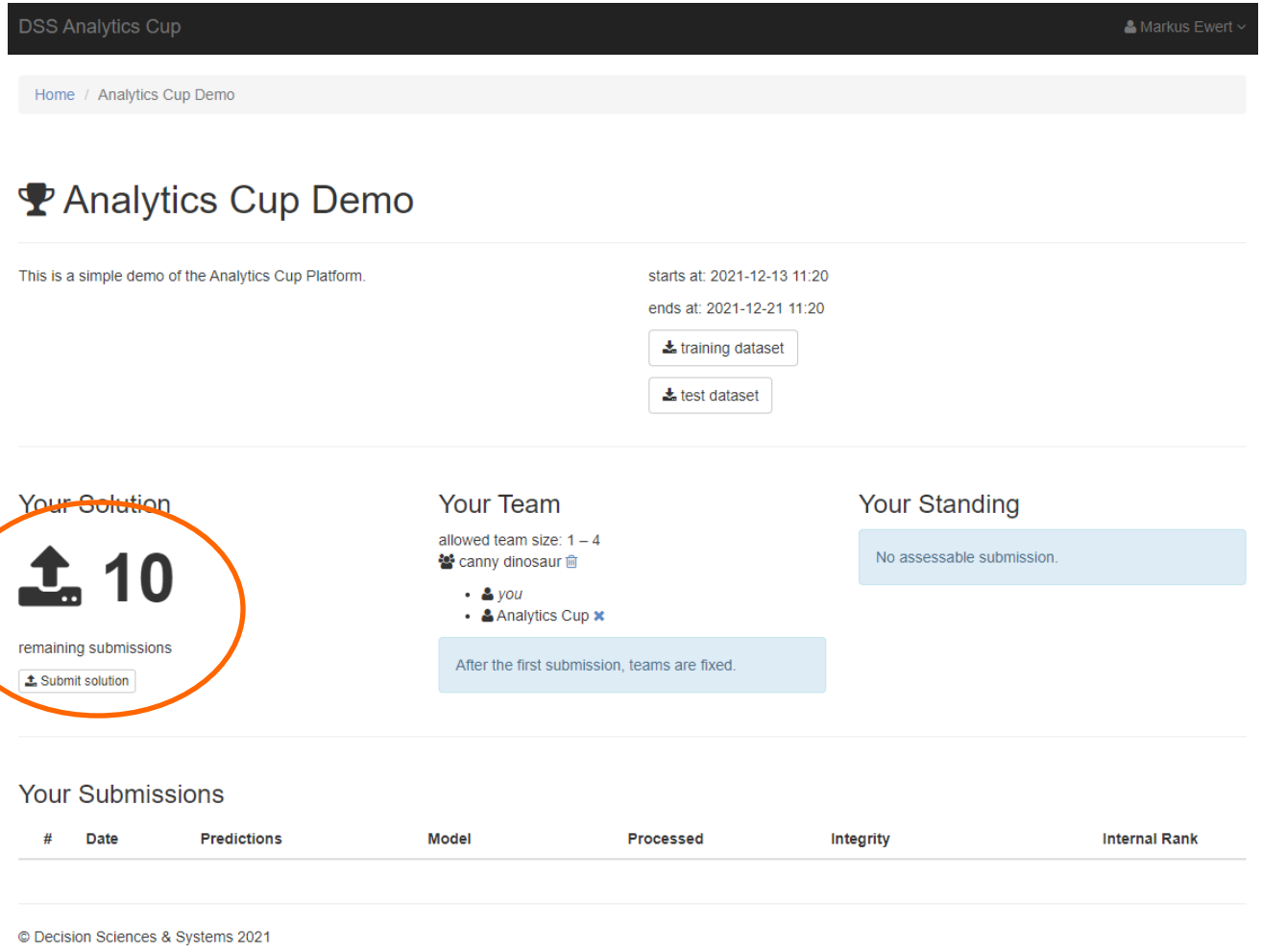
- Export predictions into csv-file
  - Format: id, prediction
  - Predictions must be 0 or 1 (not 0.5, not 'Yes', not 'FALSE')
  - Must contain all instances of the original test dataset

```
➢ pd.to_csv(submission,
  'submissions/predictions_group_name_number.csv', index = False)
```

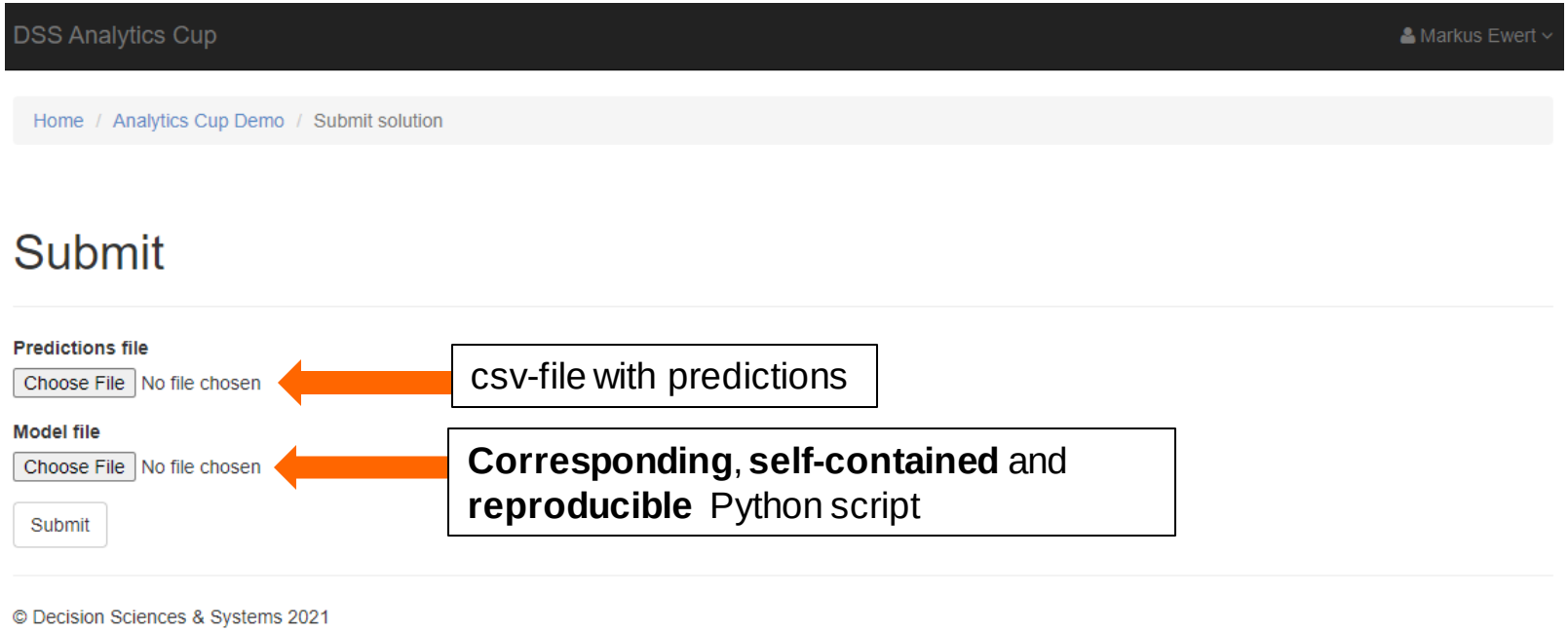
predictions_group_name_number.csv

```
id,prediction
130200,1
394720,0
87847,1
228637,1
189299,0
262991,1
…
```

Upload the Predictions and the Corresponding Python Script on the AC Manager Platform



Upload the Predictions and the Corresponding Python Script to the AC Manager



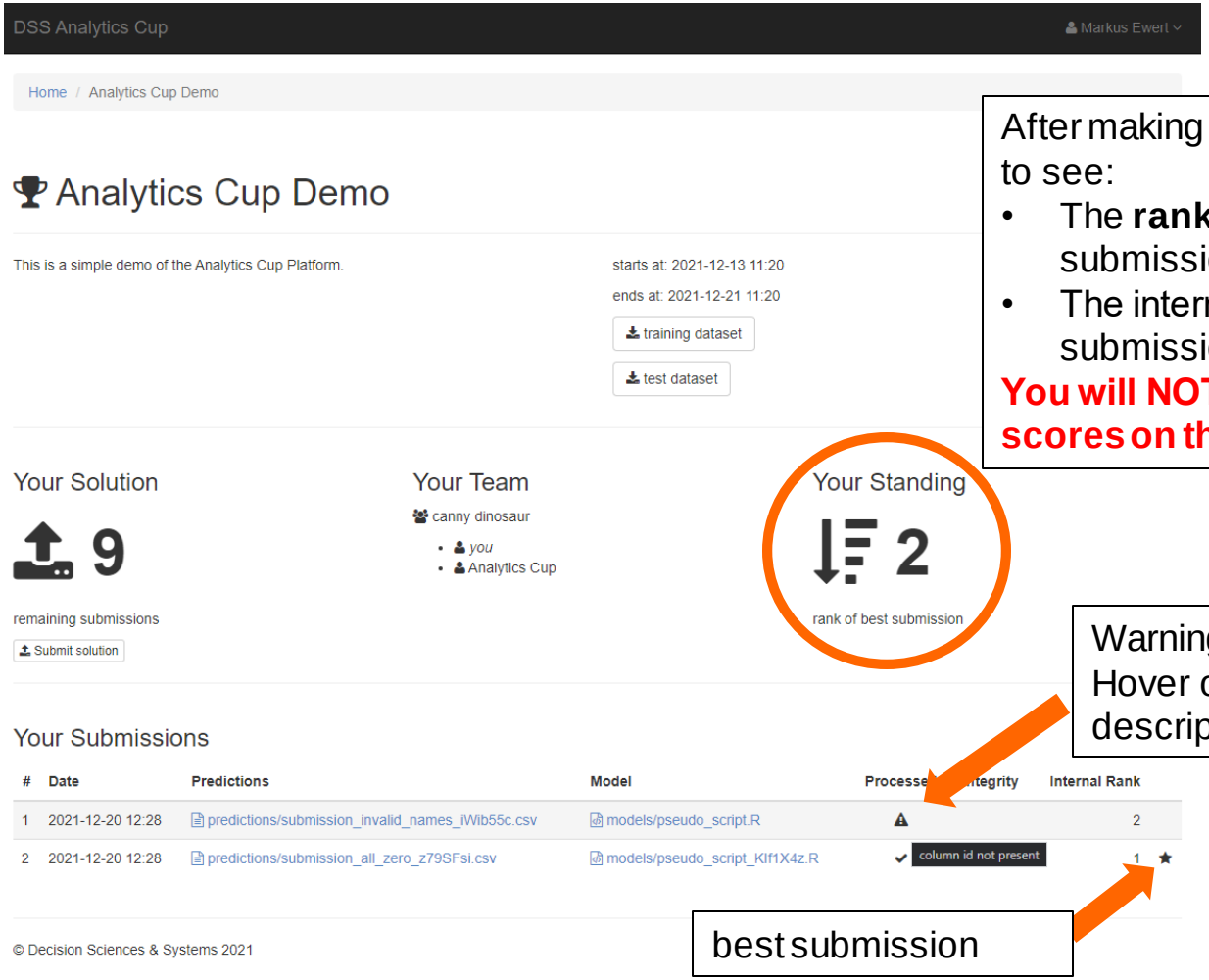
Upload the Predictions and the Corresponding Python Script to the AC Manager

Submissions & Possible Errors

- Maximum number of submission: 10 (valid submissions)
  - Best submission counts
  - Only formally valid submissions count towards your total, uploads that are rejected by the system due to formatting errors do not count as one of your ten submissions.

- **Most common formatting errors**
  - Incorrect column names
  - Unknown IDs (if not in Test Data)
  - Missing IDs (if in Test Data but not in Predictions)
  - Wrong file format
  - Wrong encoding (separators, extra line breaks, quotation marks,…)
  - …

Upload the Predictions and the Corresponding Python Script to the AC Manager



After making submissions you will be able to see:

- The **rank** of your teams best submission compared to other teams.
- The internal ranking of your team's submissions.

You will NOT see your submissions' scores on the test set.

Warning indicates invalid upload. Hover over triangle for error description

best submission

Your Task: Developing a Recommender Model for Recipes generated by a Large Language Model.

- The rise of LLMs enable the development of new business models.
  - Develop a classification model to determine whether customers likes recipes generated by LLMs.
- Dataset: Over 140'000 reviews of 75'000 recipes.
- 25 features across four tables
- Goal:
  - Predict whether a customer likes a recipe.
  - Explain your prediction
- Detailed explanation enclosed with the training data.

Getting Started

1. Read the instructions and materials **carefully,** especially the pdf included with the training data. Understand the task at hand and how you will be evaluated.
   Ask questions in the AC forum on Moodle or in the Q&A Webinar on December 22 if something is unclear.
2. Build a first simple model that produces a valid prediction file.
   - You can use the demo script from Homework 7 for "inspiration". *Submit early to make sure you got the format right, you will still have 9 more submissions to improve. Do not submit before all group members have joined your team!*
   - Leverage the provided template script.
1. Improve your model and iterate
   - Evaluation strategy on your labeled training data
   - Data cleaning, feature-selection, -engineering, preprocessing
   - Try different algorithms, tune your models. Read tutorials and documentation for help and ideas!
   - Ask questions in the Moodle AC subforum or book a support session! (Mid January)

If you feel overwhelmed, get started simple. Complicated solutions are not automatically better than simple ones.
You do not need every trick in the book to be successful!

Support During the Cup

- For questions about the rules or the data set / challenge that are not covered in the pdf instructions, please use the specific *Analytics Cup Forum* on Moodle (rather than the general forum for the course).
- If you cannot ask your question publicly because it would reveal your team's method, you may also send us an email.
- We will conduct a general Q&A session on Friday 22nd, December 2023 at 10.00 am via Zoom.

1 on 1 Support Sessions

- In the weeks January 17-23, we will offer individual 1-on-1 support sessions with your team (held via video-call or in person), where you can ask us your individual questions. You will be able to book these sessions via Moodle.
  Details will be announced in early January.
- (There will also be regular tutorials in these weeks, those will **not** be affected.)

After The Cup

- **Submission Deadline:** Monday, January 29, 8 am

- **Grading of Submissions** by the BA team.
  - We may ask you for individual meetings in case there are problems or irregularities with your submissions (e.g. reproducibility). Failing to attend these meetings will lead to disqualification of your team.

- **Final Lecture** (February 5th):
  - Awards Ceremony and Recap
  - 2-4 of the best teams will be asked to give a 5 minute-presentation about their solution. If you're asked, you're expected to present.