AmuEval: A User-Friendly Educational Platform for Machine Learning Challenges

Krzysztof Jassem
krzysztof.jassem@amu.edu.pl
Adam Mickiewicz University
Poznań, Poland

Andrzej Gajda
andrzej.m.gajda@gmail.com
Adam Mickiewicz University
Poznań, Poland

Mateusz Tylka
mattyl@st.amu.edu.pl
Adam Mickiewicz University
Poznań, Poland

Ryszard Staruch
ryszard.staruch@amu.edu.pl
Adam Mickiewicz University
Poznań, Poland

Grzegorz Lipiecki
grzlip@st.amu.edu.pl
Adam Mickiewicz University
Poznań, Poland

Szymon Bartanowicz
szybar@amu.edu.pl
Adam Mickiewicz University
Poznań, Poland

Abstract

This article presents a tool designed to teach machine learning (ML) evaluation in a gamification environment. The open-source educational platform, named *AmuEval*, enables teachers and their students to participate in or create their own ML challenges. The need for designing a new tool arose during the course for the students of data analysis, when it turned out that organizing a simple ML challenge is a laborious and time-consuming process. Compared with existing solutions, such as Kaggle, CodaLab, AIcrowd, DrivenData or EvalAI, *AmuEval* offers much easier challenge creating procedure while containing all necessary features. Our primary objective is to provide teachers and their students with a user-friendly solution that simplifies the process of creating a challenge.

Keywords: ML evaluation, gamification, challenge platform

1. INTRODUCTION

Evaluation is a crucial component of the ML workflow in real-world applications. It helps to assess the performance of various models designed for a specific task and to select the most effective one. Proper evaluation is essential to reduce the risk of overfitting and improve robustness. In an industrial environment, evaluation helps to determine whether industry standards are met and facilitates the deployment of a model that is tailored to the business needs. Therefore, it is essential in the ML educational process that the students are made aware of the importance of evaluation as early as possible.

One of the best mechanisms for keeping students engaged is gamification – the use of game design elements in non-game contexts (Deterding, S. 2011). The method can be implemented in a ML course by organizing challenges. These encourage students to benchmark their performance against their peers, while also illustrating the practical applications of the evaluation. The benefits of such an approach include: for teachers, objective verifiability of the results delivered

by students; and for students, immediate feedback on the quality of a solution.

A characteristic feature of a challenge platform is a leaderboard. In (Ortiz-Rojas, M. 2019) the authors discuss the impact of a leaderboard in STEM (science, technology, engineering, mathematics) education. Their experiments show that "studying in a gamified condition leads to a significantly higher learning performance." However, the experiments also indicate that learning in a gamified environment does not affect students' intrinsic motivation or engagement.

One of the ideas to increase students' engagement while acquainting them with evaluation is to encourage them to create their own challenges. However, in most cases it is a laborious and time-consuming process. The main goal of *AmuEval* is to simplify the creation of challenges. This aims to achieve "value co-creation", which consists in students' active participation in creating and enhancing educational experiences and outcomes. Tsourela (2015) examines the impact of value co-creation on higher education, and concludes with the statement that a student "must be transformed into an active collaborative participant of the learning process". That is why in *AmuEval* we focus on simplifying the challenge creation process.

2. RELATED WORK

The concept of introducing competitions in the development of machine learning solutions was popularized worldwide by the Kaggle platform founded in 2010. Other popular platforms today are CodaLab, DrivenData, AIcrowd, and EvalAI. A brief overview of these platforms is presented in the following sections.

2.1 Kaggle

Kaggle[1] is the most advanced platform, offering a wide range of challenges from beginner to expert level and providing a rich set of resources, such as datasets, notebooks, discussion forums and educational content.

To launch a challenge on Kaggle, the following steps must be taken:

(1) Create a competition.
(2) Set a deadline.
(3) Edit rules.
(4) Edit overview tab.
(5) Add dataset description.
(6) Upload dataset.
(7) Select a dataset license.
(8) Select scoring metric.
(9) Set up solution.
(10) Upload sandbox submission.

The Kaggle solution is very flexible and allows one to organize multiple-purpose challenges without any programming. However, the creation procedure is complex, particularly for individuals who are unfamiliar with participating in ML competitions.

2.2 CodaLab

CodaLab (Pavao, A. 2023) is an open-source platform that is particularly favored for academic purposes and research-oriented competitions. It supports both competitions and collaborative projects (worksheets) for experimenting with datasets.

To organize a challenge on the CodaLab platform, you need to prepare a bundle of files that contain the following items:

- A configuration file that outlines the competition's features and provides links to the necessary resources (HTML files, data, programs).
- Descriptive text and instructions to participants (HTML format).
- Data files, i.e. training and reference data for the competition.

[1]http://www.kaggle.com/competitions

- Program files, i.e. files for the competition that include a mandatory scoring program.

The CodaLab platform, which is publicly available and deployed at the Université Paris-Saclay, operates under the Apache 2.0 license. This availability enables the platform to be redeployed in any educational institution. The platform supports three user roles – administrator, organizer, and participant – thereby facilitating the organization of competitions within closed educational teams. In contrast to Kaggle, the process of organizing a challenge in CodaLab requires some programming knowledge.

2.3 DrivenData

DrivenData[2] focuses primarily on addressing social issues, frequently collaborating with non-profit organizations and government agencies to address challenges in public health, environmental conservation, and education. Platform competitions typically involve predictive modeling and optimization problems.

Running an ML challenge in DrivenData requires contacting the site administrators. As the platform is designed for professional competitions with monetary prizes, it is not feasible for students to create an ML challenge there in a short period of time.

2.4 AIcrowd

AIcrowd[3] serves a wide range of artificial intelligence (AI) challenges, including video game AI, robotics, and computer vision. It has strong ties to academic and research institutions and often hosts competitions associated with major conferences.

Organizing a challenge on AIcrowd requires filling out a special form with detailed information about the challenge. Unfortunately, it is not possible to create a custom-made challenge just by submitting the required information.

2.5 EvalAI

EvalAI (Yadav, D. 2019) is another open-source platform for creating and participating in ML competitions.

The platform offers key features such as custom evaluation protocols and human-in-the-loop evaluation. The process of creating a ML challenge in EvalAI involves integration with a GitHub repository. The following steps must be completed in order to create the challenge:

(1) Clone the starter repository.
(2) Generate GitHub personal access token.
(3) Add the access token to the repository secrets.
(4) Create the 'challenge' branch from the 'master' branch.
(5) Configure the config file.
(6) Change the YAML file, HTML templates and evaluation script (instructions are provided in the documentation).
(7) Commit and push the changes.

The challenge creation process requires a certain level of proficiency in managing GitHub repositories and modifying HTML, YAML, and JSON files. This may result in students devoting a substantial amount of time to preparing their challenges, rather than focusing on solutions and exploring different models.

2.6 Why *AmuEval*?

AmuEval provides an easy procedure for challenge creation. It enables students to prepare and publish a new challenge without spending much time on acquainting with the challenge creating

[2]https://www.drivendata.org/competitions
[3]https://www.aicrowd.com/challenges

procedure. The user-friendly interface also encourages students to take part in challenges organized by their peers.

3. EDUCATIONAL EXPERIENCE WITH MACHINE LEARNING CHALLENGES PLATFORM

The use of competitions in the education of machine learning has been introduced at our university since 2016. Then, a dedicated platform was created to host challenges for educational and research purposes, primarily in the field of natural language processing. The motivation for the creation of a new solution at that time was the following: Unlike Kaggle, the platform supported academic activities by providing a comprehensive evaluation library that enables participants to evaluate and optimize their solutions locally before submitting them. The platform's advantage over CodaLab was that it allowed users to track changes in submissions through a version control system. The platform is fully open-source – the code is available for download and compilation, as well as a running instance of the system. Since its creation, the platform has been widely used for educational purposes and in various conference tasks.[4]

Extensive use of the tool has revealed its drawbacks. The system is written in Haskell, which makes it difficult to maintain, as that programming language has been losing popularity. Moreover, the platform does not support the use of an external metric – any new metric must be Haskell-coded into the evaluation library. This has led to organizational problems in the preparation stage for shared tasks.

In the fall semester of 2023, the platform was introduced to MSc students of data analysis (in previous years, it was used only by advanced students of computer science). This new type of users had only basic programming skills and a solid understanding of statistics and machine learning methods. The primary objective of the course was to enable the students to apply their knowledge of machine learning to real-world scenarios. Initially, six teams were presented with a classical challenge involving the prediction of apartment prices, with an award given to the top three performers on the leaderboard. Later, each team was expected to design and deploy their own functional computer system that leverages ML methods. Furthermore, each team was required to create a challenge related to their project on the platform.

The course experience led to the following conclusions:

- Introducing a challenge platform at the beginning of the course proved successful – all groups were well aware of the necessity of evaluation in ML.
- Allowing the students to create ML systems according to their interests at a later stage was also an educational success – the students were creative and engaged.
- Using the platform to organize a challenge proved to be the most time-consuming part of the assignment. There was no cross-solving of challenges between groups.

The key takeaway from the course was the need to redesign the challenging platform so that it was as student-friendly as possible.

This led to the *AmuEval* platform project. The development team consisted of six members:

- One of the two teachers who had supervised the classes in fall 2023, having an interest in preserving the key functionalities of the previous platform while minimizing the time spent on organizing challenges.
- Two computer science graduates who had encountered all of the technical pitfalls associated with the challenging platform during their studies, driven by

a motivation to spare future generations from similar experiences.
- An MSc computer science student with no prior experience with challenging platforms, who recognized the need for a system that would benefit himself and his colleagues.
- A UX designer with no prior experience in ML, seeking to learn the concept of evaluation by testing *AmuEval*.
- A backend developer with expertise in both Haskell and Python; his task was to translate Haskell code into Python, and his main motivation was to create reliable code that would minimize future problems.

4. *AMUEVAL* WORKFLOW

4.1 Organizing a Challenge

Organizing a challenge in *AmuEval* is a process that requires only three simple steps:

Step 1: Share the dataset link. Prepare the data in your preferred repository in two folders:

- train
- test

The *train* folder should contain two files: *in.tsv* for the input data to the model and *expected.tsv* with the values that the model should predict.

The *test* folder should contain only the *in.tsv* file; the *expected.tsv* file will be hidden from the participants.

If the data cannot be represented as a tsv file (e.g. images), then only the IDs of the observations must be provided in the *in.tsv* file. For example, if a student wants to create an image classification challenge, then the images can be placed in a different folder, and *in.tsv* will contain the filenames as IDs.

The only action taken on the *AmuEval* platform is to share the repository link.

Step 2: Upload solutions. Upload the *expected.tsv* file that contains the expected values for the test set.

Step 3: Choose your metrics. Select metrics appropriate for your task from the list.

Automatic setup. The setup for the challenge is created automatically:

- The deadline is set by default for six months after publication.
- The name is identical to the name of your repository.
- The challenge description is automatically generated.
- The validation process (see 4.1, below) is triggered.

The challenge is now ready for publication; you can publish it or choose to customize it before publishing (see Figure 1).

Customizing the challenge. You can customize the challenge by:

- changing its deadline,
- changing its name,
- modifying its description and/or adapting the metrics (see 4.1, below).

Adapting the metrics. The organizer can select evaluation metrics for the challenge from a list. Within each metric, the user may set the metric parameters relevant to the task. An interactive window helps the organizer set the appropriate parameters for each metric. Figure 2 shows the panel responsible for the configuration of the accuracy metric. The "Sklearn metrics URL" non-editable window contains the link to the *scikit-learn* documentation on the chosen metric. The editable windows that appear below are used to set the parameter values. For the accuracy metric, two edit windows become visible: for the *Normalize*

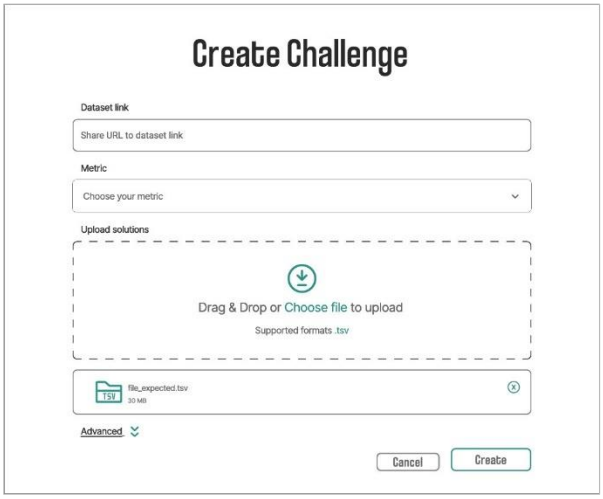

Figure 1: Three-step challenge creation in *AmuEval*

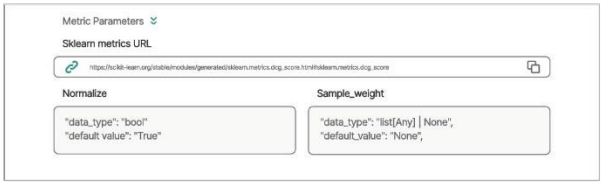

Figure 2: Metric configuration panel

and *Sample-weight* parameters, each of them displaying a prompt for a user. In the *Normalize* window, the user is supposed to skip typing or input the "False" value (then, the accuracy will be represented by the number, not the fraction, of correctly classified samples). In the *Sample-weight* window, the user can skip typing (then all samples are treated equally) or assign non-trivial weights to samples in a list.

Validation. The system validates the data submitted by calculating the values of the selected metric on *in.tsv* and *expected.tsv* from the test folder. If validation is successful, the challenge is published.

4.2 Participating in a challenge

To participate in a challenge, a user must complete the following steps:

Cloning the public repository. Participants download the data from the challenge repository to work on their solutions.

Preparing the solution using the train set. The challenge organizer does not provide a validation set. Instead, participants are tasked with splitting the training set as an educational exercise. It is essential to note that submissions based on overfitted models are unlikely to yield satisfactory results on the test set. The results will also verify whether the proper evaluation was conducted.

Pushing the results to the platform. The results (that is, the expected values for the test files) should be saved to the *out.tsv* file.

Posting the evaluation results. Once the solution has been sent to *AmuEval*, the result of its evaluation appears on the leaderboard.

4.3 Viewing the leaderboard

The leaderboard contains the best submission for each participant. Other submissions can be viewed in the "Submission" section.

5. EXAMPLES

To verify the usefulness of the platform for educational purposes, we designed two simple challenges: based on tabular data and based on an image dataset.

The first case mirrors the Kaggle challenge of predicting whether a passenger on the Titanic survived the ship's sinking. To design the challenge, we needed to prepare a repository with three files: *in.tsv* and *expected.tsv* in the *train* folder and *in.tsv* in the test folder. We then selected the accuracy metric from the list

and uploaded *expected.tsv* for the test set. The experiment confirmed that a fully functional challenge can be created in three short steps.

The second challenge involves recognizing the type of nut from its picture. Here, the repository contains an additional folder *NutsImages* with 900 images. Figure 3 shows the first five lines of *in.tsv* and *expected.tsv*; combining them we can infer that *image.627.jpg* displays a nut of type 0.

	in.tsv	expected.tsv	
1	image_627.jpg	1	0
2	image_297.jpg	2	3
3	image_590.jpg	3	3
4	image_242.jpg	4	0
5	image_465.jpg	5	0

Figure 3: in.tsv and expected.tsv files for the train set in the NutsClassification challenge

In order to submit the solutions, a participant uploads an *out.tsv* file with the predictions for the test file. We simulated three submissions: a perfect solution (identical to *expected.tsv*), a one-class solution (all predictions equal to 0), and a nearly perfect solution modeled with a convolutional neural network.

We also asked a researcher who had no prior experience with creating machine learning challenges to create his own challenge. He was able to successfully upload the challenge to *AmuEval* platform without any questioning.

6. FUTURE WORK

In the next academic semester (starting October 2024) we plan to make a user experience report and an advanced assessment of the *AmuEval* platform during the university courses.

Moreover, the usability of the system for humanist-oriented tasks will be tested during an EU-funded project that has started in June 2024. Under the project, it is planned to organize the following tasks on the *AmuEval* platform:

- predicting the date of origin for a historical text from its content,
- generating the contemporary spelling of a historical text,
- finding contemporary synonyms of ancient words and proper names,
- linking names of characters appearing in novels, interpreting hand-made notes collected from consecutive translators.

7. LICENSE

The platform is available under the Apache 2.0 license[4]. The service is available on the amueval.pl website. New instances of *AmuEval* can be deployed for educational or business purposes.

8. CONCLUSIONS

Challenges play an important role in the development of research solutions for ML. They can also be very helpful in education, engaging students by providing them with the entertainment of competition. However, most platforms require some experience in programming and competence in evaluation to organize a new challenge. We propose a new open-source platform, *AmuEval*, which reduces the organizational effort of creating a challenge to three simple steps. The goal is

[4]https://github.com/amu-cai

to encourage ML beginners to organize their own challenges and engage their peers to participate. We believe that this gamification-oriented approach will result in students recognizing the importance of evaluation at a very early stage of their ML education.

REFERENCES

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (Tampere, Finland) *(MindTrek '11).* Association for Computing Machinery, New York, NY, USA, 9–15. https://doi.org/10.1145/2181037.2181040

Ortiz-Rojas, M., Chiluiza, K., & Valcke, M. (2019). Gamification through leaderboards: an empirical study in engineering education. *COMPUTER APPLICATIONS IN ENGINEERING EDUCATION* 27, 4 (2019), 777–788. http://doi.org/10.1002/cae.12116

Pavao, A., Guyon, I., Letournel, A.-C., Tran, D.-T., Baro, X., Escalante, H. J., Escalera, S., Thomas, T., & Xu, Z. (2023). CodaLab Competitions: An Open-Source Platform to Organize Scientific Challenges. *Journal of Machine Learning Research* 24, 198 (2023), 1–6. http://jmlr.org/papers/v24/211436.html

Tsourela, M., Tarabanis, K., Fragidis, G., & Paschaloudis, D. (2015). Value co-creation in education: scope methods and insights. *International Journal of Advance Research and Innovative Ideas in Education* 1 (2015), 160–171. https://api.semanticscholar.org/CorpusID:113949975

Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Singh, S., Lee, S., & Batra D. (2019). EvalAI: Towards Better Evaluation Systems for AI Agents. *CoRR* abs/1902.03570 (2019). arXiv:1902.03570 http://arxiv.org/abs/1902.0357