Volvo's Discovery challenge at ECML PKDD 2024

Last updated on May 22^{th} 2024

1 About the Competition

Welcome to Volvo's Discovery Challenge at the European Conference on Machine Learning (ECML PKDD 2024). In collaboration with Volvo Group Truck Technologies, Halmstad University challenges participants to predict the risk of failure for an undisclosed component of Volvo trucks. This challenge invites participants to work with a real-world dataset incorporating the measurements from a fleet of more than 10,000 Volvo heavy-duty trucks. The task is to develop a machine learning model to predict risk levels (*Low*, *Medium*, and *High*) for a component of the trucks in the test set. Your solution could enhance the predictive maintenance strategies for trucks, improving reliability and efficiency and reducing the environmental impact and CO_2 emissions. The entire challenge material is conveniently located within a shared Box folder accessible via this link.

2 Prizes

The top three competitors receive a time slot to present their work at the conference. The first place receives **free registration** for the ECML-PKDD 2024. In addition, there is prize money for the top three contenders: the team in the first place will receive **500€**, the second place **300€**, and the third place **200€**. Please note the contestants may be subject to income tax if they win, and participants need to carry any tax effects of receiving the prize.

3 Competition Phases

The competition has two phases:

1. **Development phase** (May 15 to June 15): The competition starts with the development phase. In this phase, you can submit 5 predictions per day. The submitted prediction will be evaluated against 20 percent of the ground truth data. The participants are welcome to submit their prediction results on the Codabench portal and compare their results with other participants.

2. **Final phase** (June 16 to June 30): Once the development phase finishes, the final phase starts. In this phase, you can submit only 3 predictions in total. The submitted prediction will be evaluated against the whole ground truth data. You should choose your best-performing submission among those three submissions in the final phase and mark it to be displayed in the Leaderboard (please refer to Figure 3). Note that your submissions in the development phase will NOT automatically migrate to the final phase, and you need to **make at least one submission in the final phase and mark it to be displayed in the Leaderboard**; otherwise, you will not be included in the final Leaderboard. The ranking in the Leaderboard of the final phase determines the final ranking.

4 How to Participate

All the data required for the competition, as well as an up-to-date version of this description, is available here. Participants are expected to submit their results on the Codabench portal. Registering for a portal account is required to engage in the challenge and share your results. The challenge page can be

located using this link: https://www.codabench.org/competitions/3022/?secret_key=c5bb4004-b280-456e-84f6-3bb42737e8dc. The menus on the challenge page remain locked until the challenge begins. Once the challenge begins, you can upload your predictive outputs, compare your results with those of fellow competitors during the Development phase on the portal's leaderboard, and ultimately submit your final results in the Final phase of the competition. To sign up in the Codabench portal, please use this link. For detailed instructions on how to start with Codabench and submit results, please visit this page.

5 How to Submit your Predictions

The prediction should be a single ZIP file called *"prediction.zip"*. Inside the *"prediction.zip"*, there must be a CSV file *"prediction.csv"* containing a header called "pred" followed by one prediction (Low, Medium, or High) per row (for the total of 33590 rows). The index of each row in the prediction file is aligned with the index of the *"public_X_test.csv"* file. You should then submit your *"prediction.zip"* file in the Codabench portal.

To make the submission process easy to follow, the participants are provided with a Python file named *startkit.py*. This file serves as a starting point, where it first reads the training data, then trains a Decision Tree classifier as a baseline, performs predictions on the test set, and finally creates the *"prediction.zip"*. You can directly submit the resulting *"prediction.zip"* file in the Codabench portal. You can run the *startkit.py* to get your first submission. Figures 2 and 3 in the section 9 (Appendix) provide more information about the submission form and leaderboard in the Codabench portal.

6 Description of the Data

The dataset includes three files: *"train_gen1.csv"*, *"public_X_test.csv"*, and *"variants.csv"*. The files will be added to the shared box folder once the challenge begins.

- **train_gen1.csv** includes more than 157,437 readouts from an undisclosed component of 7,280 Volvo heavy-duty trucks, each identified by an anonymous chassis ID called "ChassisId_encoded". The "train_gen1.csv" file contains 308 columns. The first four columns are 'Timesteps', 'ChassisId_encoded', 'gen', 'risk_level', followed by 304 feature columns. Note that the "train_gen1.csv" file only contains data from "generation1" of the component. These readouts are recorded at consecutive timesteps, starting from timestep 1 and extending until either component failure or the end of data collection. While some components experienced failure during the study period (unhealthy components), many others did not fail (healthy components). The time intervals between these consecutive readouts are undisclosed but can be considered consistent and equally spaced.

  The Training set includes the target variable named "risk_level" that determines three levels of "*Low*", "*Medium*", and "*High*" risk for the trucks – one label per readout, depending on that readout's proximity to the component failure time. As can be seen in Figure 1, the *High* label is assigned to the readouts within 9 timesteps before a failure. The *Medium* and *Low* labels are assigned to the readouts between 9 to 18 timesteps, and those more than 18 timesteps prior to failure, respectively. Additionally, it is important to mention that the training set is limited to observations from a single generation, called "gen1", of the trucks' component. However, the competition's task extends beyond this scope, and the model is expected to provide predictions for two generations of components ("gen1" and "gen2") in the test dataset. The goal is to evaluate how well your model can perform on the test data from gen1 and also how well it can generalize to a new, unseen, generation (gen2).

- **public_X_test.csv** includes 33590 rows and 307 columns where the first three columns are "Timesteps", "ChassisId_encoded", "gen", and the remaining 304 columns are sensor readings, mirroring the features that exist in the training set. Unlike the training set, the test data spans two generations of the case study's component (as indicated in the "gen" column), adding more complexity to the prediction task. Your prediction file should include one prediction per row of "public_X_test" file (i.e., 33590 rows). Moreover, in the test set, the sequence length from each chassis ID is fixed at 10 Timesteps. For the healthy components, a sequence with a length of 10

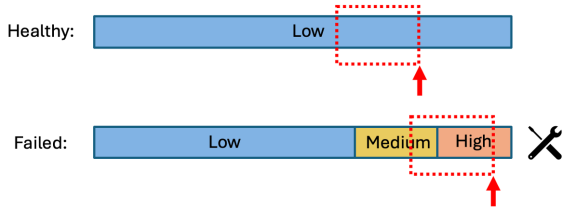

Figure 1: Sequence Extraction from Healthy and Failed components. Note that the length of all sequences is 10 time steps.

is randomly selected from each "ChassisId_encoded". In failure cases, a random timestep from the "High risk" period is selected as the last readout, with the preceding 10 Timesteps initiating the sequence for that particular test chassis ID. The sequence selection process is illustrated in Figure 1.

- **variants.csv** contains specifications for all trucks included in the *"train_gen1.csv"* and *"public_X_test.csv"* files. The file has 10,639 rows and 13 columns. The first column gives "ChassisId_encoded" followed by 12 columns of encoded specifications of the trucks. The specifications include information such as the engine type, cabin type, number of wheels, number of axles, etc.

7 Evaluation

The macro-average f1-score is the main evaluation metric in this competition. To calculate the macro-average F1-score for a classifier with three classes (i.e., Low, Medium, High), one first computes the F1-score for each class individually and then takes the average of these F1-scores. The macro-average F1-score treats all classes equally, regardless of class imbalance.

The macro-average f1-score is calculated for predictions of gen1 and gen2 separately. The final score that determines the winner is the average of the macro f1-score for gen1 and gen2. The startkit.py file contains an implementation of the evaluation metric used in the competition and evaluates the predictions of the baseline model against a mocked-up ground truth. Once you submit your predictions to Codabench, the following explains how you can check the detailed evaluation results for your submission:

1. Click "My Submissions"
2. Select your submission from the list at the bottom of the page.
3. Click "LOGS"
4. Click "Scoring Logs" (see Figure 4)

8 Terms and Conditions

All participants of the Volvo Challenge ECML PKDD 2024 will gain access to Volvo's dataset. The dataset distributed during the challenge is referred as the "Volvo GTT" dataset and belongs to Volvo. By participating in the challenge, Volvo grants you and you accept to receive a personal, non-exclusive, non-transferable, non-sublicensable, royalty free license to use the "Volvo GTT" dataset solely for the purpose of participating in the challenge. The license shall last only during the time of the Volvo Challenge ECML PKDD 2024. Apart from the abovementioned license, Volvo reserves all rights in the "Volvo GTT" dataset.

By enrolling in the competition, all participants grant Halmstad University (as the organizer) and Volvo Group a license to the solutions proposed in the contest. Given the prize levels, the contestants may be subject to income tax if they win, and participants need to carry any tax effects of receiving the prize.

Contact Information For Further Questions: mahmoud.rahat@hh.se, peyman.mashhadi@hh.se

Organizers: mahmoud.rahat@hh.se, peyman.mashhadi@hh.se, shamik.choudhury@consultant.volvo.com, leo.petrin@volvo.com, thorsteinn.rognvaldsson@hh.se

The organizers wish all participants a thrilling and successful competition ahead!

Appendix Do not forget to check Section 9 (Appendix) on the next page.

9 Appendix

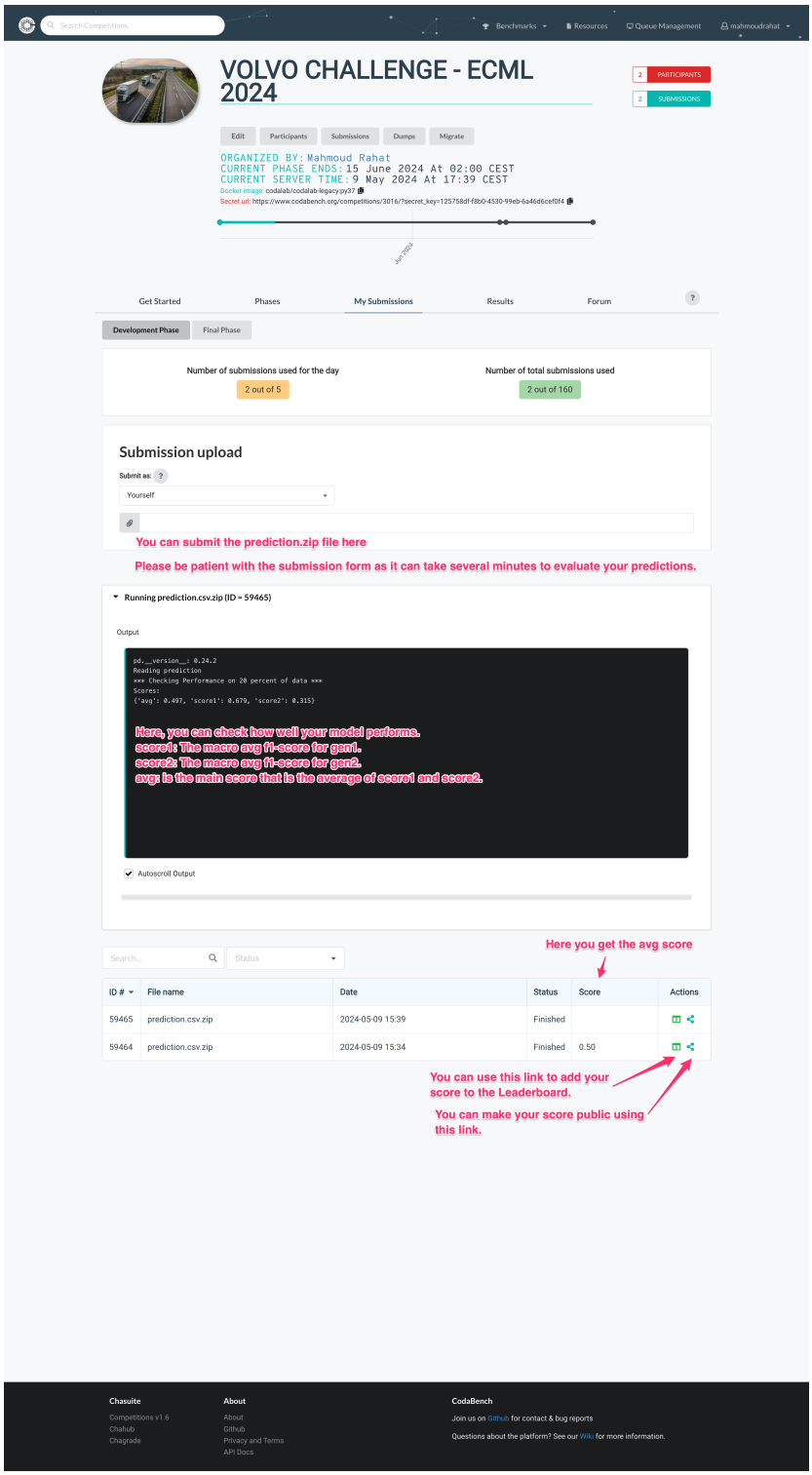

Figure 2: The submission form

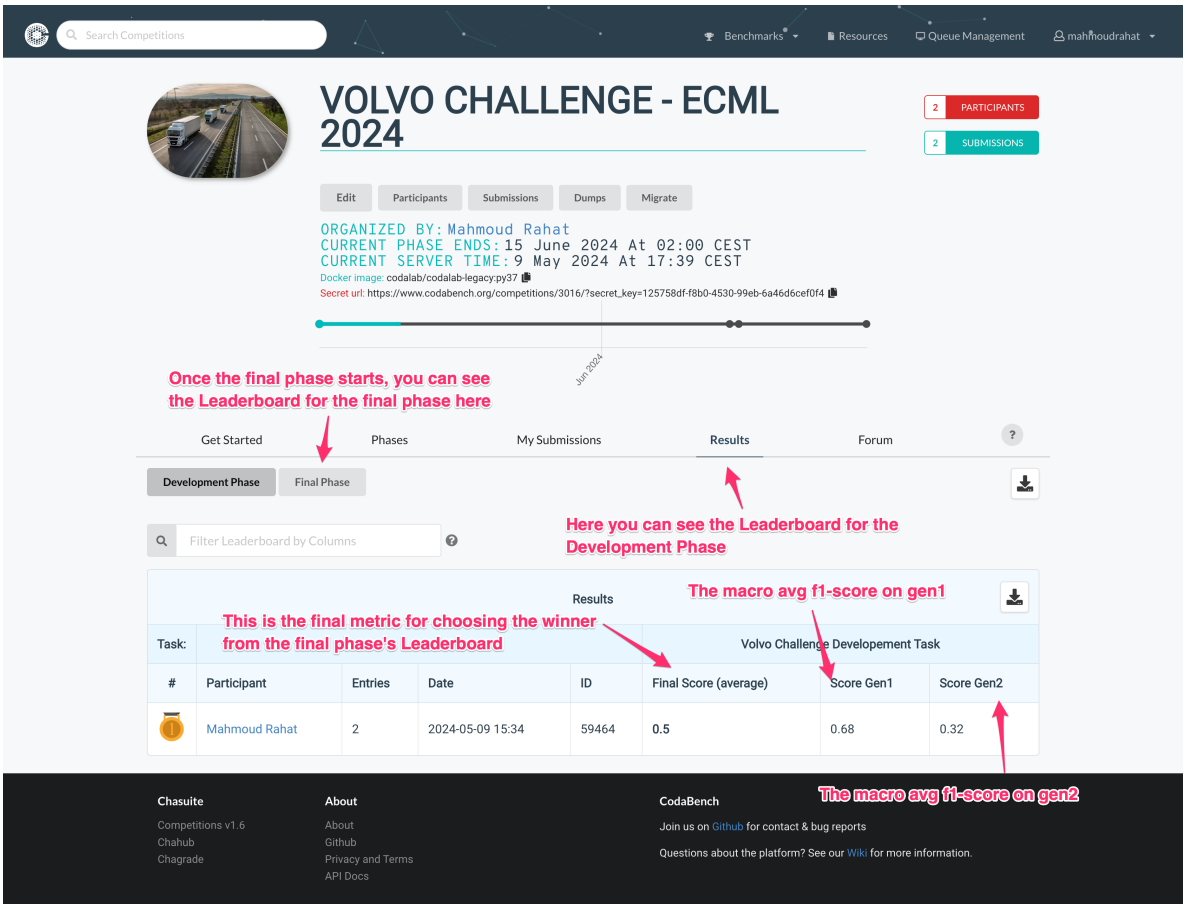

Figure 3: The Leaderboard

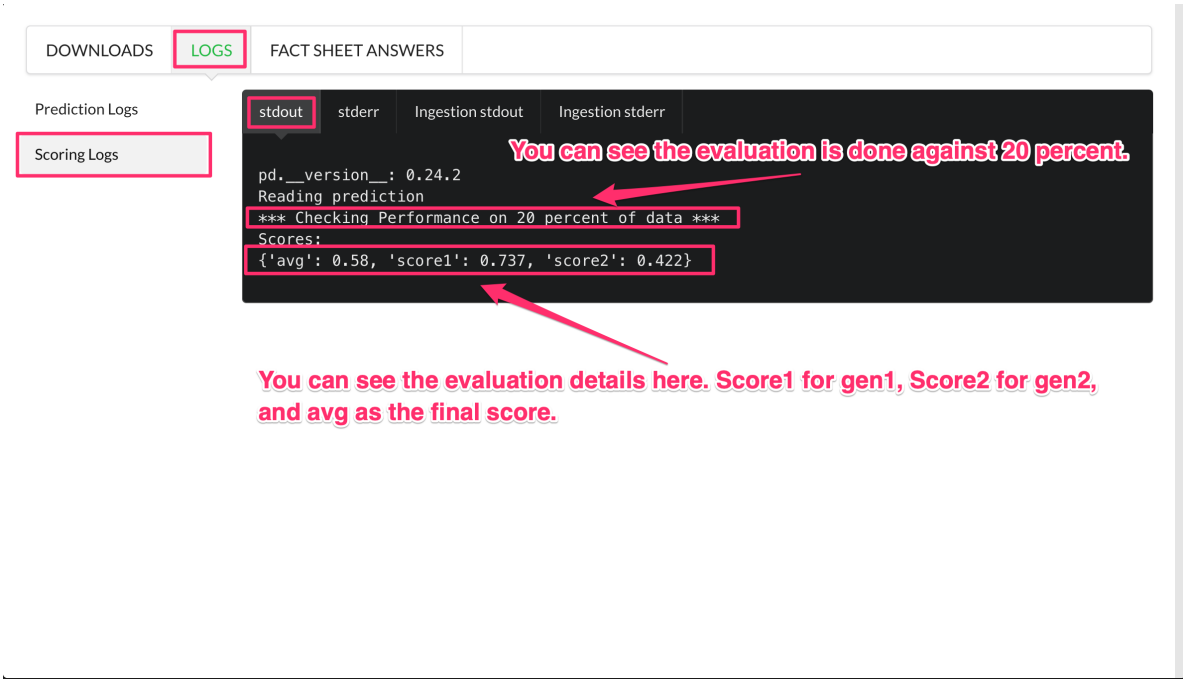

Figure 4: The evaluation form (above) is accessible from your Codabench account through: "My Submissions">select a submission from the list>"LOGS">"Scoring Logs"