

FUZZ-IEEE COMPETITION ON EXPLAINABLE ENERGY PREDICTION

<https://ieee-dataport.org/competitions/fuzz-ieee-competition-explainable-energy-prediction>

Evaluation

Our partner E.ON is interested in predicting with the least amount of data as possible a good estimate of the *total year consumption* for reasons of payment adequacy. Accurately predicting a customer's annual consumption allows E.ON to set their direct debit payments up correctly. Making predictions for smart meters for which we have very little historical consumption data is usually more challenging and relevant to avoiding excessive debt or credit at the end of the year. In addition to that, an accurate monthly consumption prediction is also very valuable for Energy Trading teams, who need to be able to predict with some accuracy how much electricity to buy on the energy market. Moreover, E.ON demands not only accurate but self-explaining predictions (i.e., accountable, transparent and easy to understand) with the aim of providing customers with trustworthy products; thus attracting and retaining more customers. This is in agreement with the European General Data Protection Regulation ([GDPR](#)) which states that humans have a right to an explanation of decisions affecting them, no matter who (or what intelligent system) makes such decisions.

Having that in mind, submissions are evaluated as follows:

- **Total Year Consumption:** Your predicted total year consumption will be computed as the sum of all the predictions you've made for each individual month. Then, for each smart meter, we will compute a **Relative Absolute Error** between the predicted total year consumption values and real values.

$$year_{rAE} = \frac{\frac{1}{N} \sum_k |y_k - t_k|}{\frac{1}{N} \sum_k |t_k - \bar{t}|}$$

Where $\bar{t} = \frac{1}{N} \sum_k |t_k|$, N is the total number of smart meters, y_k is the predicted total year consumption of the k -th meter, t_k is the true total year consumption of the k -th meter.

- **Monthly Consumption:** For each smart meter, we will also compute the **Relative Absolute Error** between the predicted monthly consumption values and the real values.

$$month_{rAE} = \frac{1}{N} \sum_k \frac{\frac{1}{12} \sum_i |y_k^i - t_k^i|}{\frac{1}{12} \sum_i |t_k^i - \bar{t}_k|}$$

Where $\bar{t}_k = \frac{1}{12} \sum_i |t_k^i|$, y_k is the predicted monthly consumptions of the k -th meter, $y_k = [y_k^1 \ \dots \ y_k^{12}]$ includes predicted total month consumption for 12 months. t_k is the true monthly consumptions of the k -th meter, $t_k = [t_k^1 \ \dots \ t_k^{12}]$ includes true total month consumption for 12 months.

Both metrics will be considered equally important and aggregated as:

$$total_{r_{AE}} = \frac{1}{2} year_{r_{AE}} + \frac{1}{2} month_{r_{AE}}$$

You can download a Python Implementation of these metrics [here](#).

- **Explanations:**

Only those submissions achieving the given performance threshold will be carefully analysed from the point of view of explainability. A panel of 5 experts will measure goodness and effectiveness of narrative explanations associated to predictions of both total year and monthly consumption, regarding a selection of N cases in the test set. In case there were more than 10 submissions above the threshold, only the top 10 would be evaluated. The jury will pay attention to structural and semantic complexity, regarding accountability, clarity, naturalness, completeness, and length of explanations. Namely, the following criteria are assessed:

- **C1. The explanation of the prediction is correct** (i.e., readable, fluent and grammatically correct) [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C2. From the explanation, I understand how the prediction is made** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C3. From the explanation, I understand the cause-and-effect model underlying the prediction** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C4. The explanation comprises factual and counterfactual pieces of information** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C5. The explanation has sufficient detail** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C6. The explanation is complete** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C7. The explanation is useful to my goals** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C8. The explanation shows me how accurate the prediction is** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C9. The length of the explanation is appropriate** [5* Likert Scale]: 1) Strongly Disagree; 2) somewhat disagree; 3) neutral about it; 4) somewhat agree; 5) Strongly Agree
- **C10. Overall Assessment** [5* Likert Scale]: 1) Very Dissatisfied; 2) Dissatisfied; 3) Neither Satisfied Nor Dissatisfied; 4) Satisfied; 5) Very Satisfied

It is worth noting that those submissions with an Overall Assessment (C10) below 3 will be discarded.

Finalists will be ranked based on an aggregated score of prediction and explanation performance. The jury will be looking at both error rates and explanations as a whole. The **overall performance-explainability trade-off** includes annual and monthly predictions. The Total rAE weights both indicators equally. Scores of C1 – C10 are averaged for all given explanations.

In addition, the jury will do a qualitative review of the description of the methodology applied by each finalist in terms of:

- **Clarity and justification of the proposed methodology.** Does it appropriately use Computational Intelligence techniques if any (not required!)?
- **Robustness and potential generalisation** of the proposed approach, regarding both prediction and automatic narrative explanation generation. Is this a heavily ad-hoc solution?
- **Experimental methodology and parameter tuning.** Have the participants tuned their parameters against the test dataset?

In case of ties, the qualitative feedback of reviewers will be considered.

To give a flavour of the difficulty of the problem and provide some expected thresholds, we provide a very simple and highly explainable solution to the problem. This solution consists of using the average monthly consumption for each customer as prediction for the following 12 months. This provides a year_rAE of 0.3886, month_rAE of 1.6926 and total_rAE of 1.0406. For the sake of comparison, the best solution from the previous competition (with the exact same data) which was not focused on interpretability obtained a year_rAE of 0.2864, month_rAE of 1.0078 and a total_rAE of 0.6471.

The Overall Assessment will be considered together with the aggregated scores and qualitative reviewer feedback to determine the winners.

The Leader Board update will ONLY be provided for the prediction performance, but not for the explainability component, which will be judged at final submission only. Also note that the Leader Board table is updated every 5 minutes *and it will keep the best solution (ever) of each participant. In case you submit a wrongly formatted submission, you will be listed at the top of the Submissions table with an empty row.*

At the end of the competition, the best solutions will be invited to write a joint paper discussing the problem and solutions.

Important:

- (1) Note that use of data other than the one provided is not allowed. If you do so, we will not consider your submission.
- (2) For the sake of reproducibility, to guarantee that the description you provide of your method returns the results you have submitted, you are asked to provide the source code associated to your final submission. We understand that this might involve running multiple software tools and various steps. So, please provide a detailed description and requirements in a README file, and if possible, some sort of simple script (e.g. in R, Python, or Unix) to run it. Note that the code will solely be used by the Technical Committee to guarantee the fairness of the competition. We will not share this with any third party or make any commercial use of it, as this belongs to you.