

Junction 2025 Hackathon – Fortum Challenge Participant Brief

INTRODUCTION

The Fortum challenge at Junction 2025 invites you to tackle a real-world energy forecasting problem. Fortum, a leading clean energy company, has provided an anonymized dataset of electricity consumption for several customer groups (in a fun unit we're calling "FortumWattHours"). Your task is to build models that **predict future power usage** – both in the short term (the next 48 hours, hour by hour) and the long term (the next 12 months, month by month). By leveraging historical consumption patterns, corresponding day-ahead electricity prices and any open dataset that you can find, you'll develop forecasts that can help optimize energy production and grid management. This challenge is a chance to apply your data science skills to a sustainability-focused problem and demonstrate how much better your predictions are compared to a simple baseline. In the sections that follow, you'll find details about the dataset, the specific forecasting tasks, how your work will be evaluated, and submission guidelines. Good luck, and enjoy exploring the future of energy with Fortum!

DATASET & FORECASTING TASKS

Fortum has provided an anonymized dataset and a real-world scenario that reflects how we **trade and hedge electricity**. When forecasting consumption, think of Fortum's role as an electricity seller: we must buy the right amount of power for our customers on the hourly market and secure longer-term contracts for future months. Accurate forecasts help us avoid imbalances and manage price risks. Here's what you'll be working with:

- **groups sheet – Customer Group Definitions:** This sheet lists unique numeric IDs for each customer group (112 groups in total, with non-sequential ID numbers). Every group represents a specific segment of Fortum's customer base, defined by a combination of attributes:
 - **Macro Region:** One of Finland's top-level administrative regions (Regional State Administrative Agency areas, e.g., Southern Finland, Eastern Finland, etc.).
 - **County/Region:** An official region within that macro area (Finland has 19 regions, like Uusimaa, Pirkanmaa, etc.).
 - **Municipality:** A major city or area within the county. If this field equals the county name, the group includes all customers of that type across the county. If it's a specific city name (e.g., Helsinki, Espoo, Vantaa), the group is limited to customers in that city. If it's listed as "{County}_Others," it covers all other municipalities in that county outside the major city.
 - **Segment:** The customer sector – either "private" (household consumers) or "enterprise" (small business customers).
 - **Product Type:** The type of electricity contract or pricing the customers have. For example, one common type is a **spot price** contract (where usage is billed at fluctuating market prices). Other groups might be on fixed-rate plans or other pricing models.

- **Consumption Bucket:** A broad classification of the customers' annual electricity usage: "low," "medium," or "high" consumption. (The thresholds for these categories differ between private and enterprise customers, reflecting that businesses generally use more energy than households.)

Each group ID in the dataset corresponds to a unique combination of the above attributes. *For instance, a group could represent something like "Southern Finland – Uusimaa – Espoo – Enterprise – Spot Price – Medium Consumption."* All groups are sufficiently large (100+ customers each) to ensure no single customer's behavior dominates the data or can be identified.

- **training_consumption sheet:** Historical electricity consumption per group at an hourly interval. This spans several years up to end of September 2024. The consumption values are given in a fictional unit called **"FortumWattHour (FWH)"** – essentially a transformed kWh used to mask actual volumes. That means all trends and patterns are intact, but you don't need to worry about the actual magnitude (just work with FWH as you would with any energy unit). The data is continuous hourly records (no large gaps), capturing daily cycles, weekly patterns, and seasonal variations for each group.
- **training_prices sheet:** Historical day-ahead electricity prices for Finland (EUR per MWh) for each hour corresponding to the consumption data timeframe. Essentially, for every hour in our training period, you have what the Nordpool energy market day-ahead price was. We've included price data up to 24 hours beyond the last consumption timestamp (through October 1, 2024) to simulate that you know the next day's prices. These prices can be used as an input feature for forecasting.

Forecasting Tasks: Using this data, your challenge is to predict future consumption for each group on two horizons:

1. **48-Hour Forecast (Short-Term, Hourly):** Predict the electricity consumption (in FWH) for each of the next 48 hours **after** the training period. Specifically, forecast from October 1, 2024, 00:00 up to October 2, 2024, 23:00 for each group. For the first 24 hours of this period (Oct 1), you have the actual day-ahead prices provided – just like Fortum would know today the prices for tomorrow. For the following 24 hours (Oct 2), no price info is given (since in a real scenario those prices aren't known yet). Your models should handle this by either not relying on price for that part or perhaps forecasting consumption based on typical price patterns without actual values. The goal here is to mimic real daily operations: **predict the next two days** of demand so Fortum can buy the right amount of energy in the day-ahead market.
2. **12-Month Forecast (Long-Term, Monthly):** Predict the total consumption for each group for each of the next 12 months, from October 2024 through September 2025. Provide one forecasted value (in FWH) for each group for each month (the sum of consumption over that month). These long-term forecasts will help Fortum in **hedging and planning** – e.g., deciding how much energy to contract or hedge for the upcoming year. You have no future price or weather data provided for this horizon, so you'll rely on historical trends, seasonality, and any other patterns in the consumption (and perhaps how it correlated with past price movements) to make these projections.

Use of External Data: You are encouraged to get creative and enrich your models with **publicly available external data sources**! For instance, historical or forecasted weather data, public holidays calendar, daylight hours, economic indicators, or news events could all be relevant to electricity usage. **Two important rules** when using external data: (a) it must be truly open/public data that any team can access (no private datasets), and (b) you **cannot use data from the actual future** that wouldn't be known in real-time. In other words, don't incorporate any information that includes October 2024–September 2025 actual realized data. For example, using historical weather up to 2024 is fine, or even weather forecasts issued before Oct 2024, but not actual observed temperatures in 2025. Similarly, you shouldn't use actual market prices beyond what's provided (since those would only be known after the fact). The spirit is to simulate making forecasts as if we are at the end of September 2024, with all prior data (and any relevant external

info up to that point) at your disposal, but nothing from the future. Using additional data smartly can improve your model's accuracy – e.g., incorporating weather trends might explain seasonal consumption changes better, or economic indicators might help capture growth or decline in usage.

By understanding the data and the business context (trading and hedging), you're set to build solutions that are not only technically sound but also aligned with real energy decision-making. In the next sections, we'll cover how to format your results and how we'll evaluate your forecasts against a baseline.

SUBMISSION FORMAT & DELIVERABLES

To participate in the Fortum challenge, please prepare the following deliverables and submit them via the Junction platform by **Sunday, Nov 16, 2025 at 10:00 EET**:

- **Forecast Files (CSV format):** You must produce two CSV files with your predictions. Both files should be encoded in **UTF-8** and use a semicolon (;) as the delimiter (following the European CSV format). Decimal points should be represented with a comma (.). Ensure every required value is present – no missing timestamps or empty cells. The format of each file is as follows:
 - **48-Hour Forecast File (Hourly Predictions):** This file should contain your hourly consumption forecasts for each group for the period Oct 1, 2024 00:00 to Oct 2, 2024 23:00. Each **row** represents one hourly timestamp in ISO 8601 format (**UTC** time with a "Z" suffix, e.g., 2024-10-01T00:00:00.000Z), and each **column** (after the first) corresponds to a customer group ID. The first column header must be measured_at, and each subsequent column header must be the exact group ID (numeric, in the same order as in the training data. Each data row will have the timestamp in the first column and 112 forecast values (one for each group) in the following columns. Make sure to include **all 112 groups** in each row. An example file is provided (with dummy values).

Note: The values in the example use commas as decimal separators (e.g., 2,664540354 for 2.664540354) and are separated by semicolons. Your file should follow this format exactly. No group columns should be omitted, even if a forecast might logically be zero. Include a row for every hour in the 48-hour range.

- **12-Month Forecast File (Monthly Predictions):** This file should contain your monthly total consumption forecasts for each group from October 2024 through September 2025. Each **row** represents one month, using the timestamp of the first day of that month at 00:00 (ISO format). For example, use 2024-10-01T00:00:00.000Z for October 2024, 2024-11-01T00:00:00.000Z for November 2024, and so on up to 2025-09-01T00:00:00.000Z for September 2025. As with the hourly file, the first column header is measured_at, followed by a column for each group ID. Each data row will have the month timestamp and 112 forecast values (one per group) representing the total consumption (in FWH) for that group in that month.

Again, follow the exact format shown in the example: semicolon separators and comma for decimal point. There should be 12 rows of data (one for each month Oct 2024–Sep 2025). Include all group columns in each row. If a group's forecast for a month is zero, you should still explicitly write 0 (or 0,0 with the comma decimal format) rather than leaving it blank.

- Please use the provided template files as a reference to ensure your submission is formatted correctly. Submissions with incorrect format (missing columns, wrong headers, etc.) may not be evaluated, so double-check against the examples.

How Your Forecasts Will Be Scored: We will measure how much your predictions improve upon a naive baseline forecast. The baseline assumes consumption stays the same as typical patterns (for hourly forecasts, it uses the **same hour one week earlier**; for monthly forecasts, the **same month one year earlier**). We calculate a **Forecast Value Added (FVA)** percentage to quantify your improvement. In simple terms:

$$\text{FVA\%} = 100 \times (\text{Error_baseline} - \text{Error_your_model}) / \text{Error_baseline}$$

Here “Error” refers to the Mean Absolute Percentage Error (MAPE). A positive FVA% means your model’s error is lower than the baseline’s (good!), while 0% means you’re on par with the baseline. We will compute one FVA% for your 48-hour forecast and another for your 12-month forecast, then take the **average of the two** (giving equal weight to short-term and long-term accuracy) to determine your final score. Keep this in mind: the goal is to beat the baseline on both horizons for the best overall result.

- **GitHub Repository:** Host all your project materials in a public (or publicly accessible) **GitHub repository**. This repository should contain:
 - Your two CSV forecast files (hourly and monthly).
 - Your code, scripts, and any notebooks used to develop your solution.
 - **Methodology Documentation** – see below.
 - A brief **README** that describes how to run your code or understand the project structure (for completeness).

Once your repository is ready, submit the **GitHub repo link** through the Junction hackathon platform. Make sure the repo is public or that you have granted access to the judges so we can review your work. The deadline for submitting the repo link (and all deliverables within) is **10:00 EET on Sunday, Nov 16, 2025** (the end of the hackathon).

- **Methodology Document:** Include a concise report (in PDF or Markdown format) in your GitHub repo that clearly explains your approach. This document should be **self-contained** and cover:
 - Modeling Techniques – Describe the models or algorithms applied for both the 48-hour and 12-month forecasts. Explain your rationale for selecting these approaches.
 - Feature Selection & External Data – Detail your data preprocessing steps and the features used in your models. Indicate if you incorporated any external data sources and discuss their impact.
 - Model Training & Validation – Summarize your model training and tuning process. Outline your validation strategy and how you ensured the robustness of your models.
 - Business Understanding – Show how your forecasting approach aligns with Fortum’s operational context and objectives.
 - Results Summary – Provide a brief assessment of your model’s performance or comparison to baseline methods, if available.

The methodology document should be written clearly enough that judges can understand your solution without needing to ask you questions. Remember, there will be **no live Q&A** during judging, so this document (and your code) is our window into your thinking.

- **Demo Video:** As part of your Junction submission, you’ll also upload a short demo video. In this video, introduce your team’s solution – walk through your approach, highlight any novel features of your work, and summarize your results. Think of it as a quick presentation for the judges. You can show slides, your system in action, or any visualizations of your forecasts. Aim for a length and format as specified by Junction. We will watch this to complement your written documentation. Ensure the video link or file is provided via the Junction submission platform and is accessible.

By the submission deadline, make sure:

- Your GitHub repo is finalized and the link submitted.
- Both CSV forecast files are in the repo and adhere to the required format and naming.
- Your methodology document (and any **clearly marked** supplementary analysis) is in the repo.
- Your demo video is uploaded or linked as required.

What We Expect from Participants: We expect you to develop your forecasts using sound data science practices. This means using historical data (and any allowed external data) to train predictive models, rather than hard-coding or guesswork. Show creativity in improving upon the baseline – whether through advanced modeling, clever features, or insightful use of external information. Also, maintain a focus on the business goal: accurate forecasts that provide value beyond the naive approach. In your documentation and video, convey not just *what* you did, but *why* it's effective for this problem. We're excited to see solutions that are both technically strong and aligned with Fortum's energy forecasting context.

Good luck with your submission! We look forward to reviewing your forecasts and insights. Make sure to double-check your files and documentation for completeness before the deadline. Once submitted, our judges will evaluate your work based on accuracy (as described in the evaluation section) and the clarity and soundness of your approach.