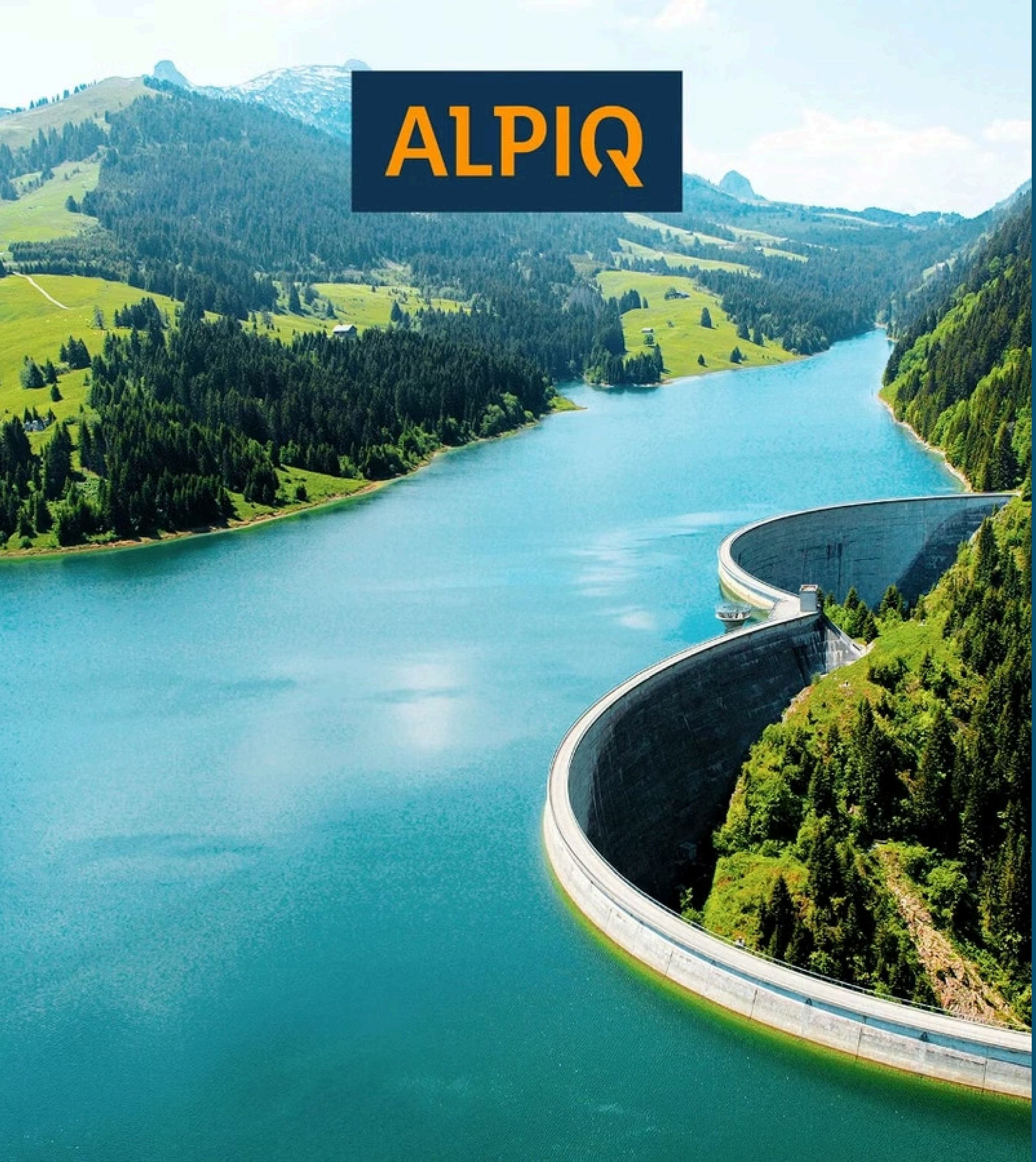


Modelling Tomorrow's Energy Needs

HANGUK ML

ETH Datathon 2025

Alpiq Challenge Presentation, 6th April 2025

ALPIQ

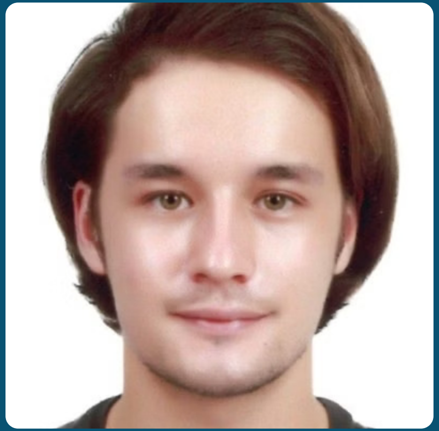
Our Team: HANGUK ML

Nikolai

MS in Quant Finance, ETH & UZH

Shyngys

MS in Data Science, ETH

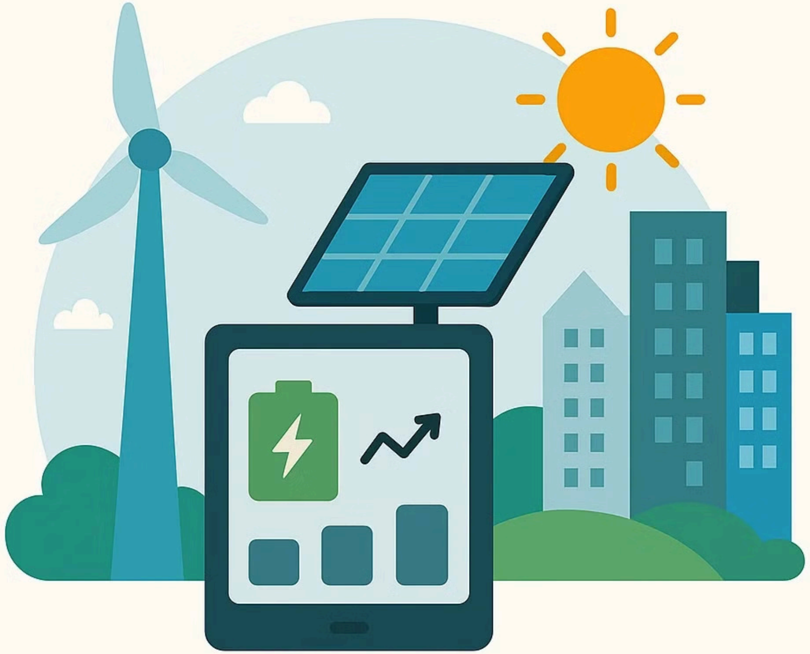


Alexander

MS in Data Science, EPFL

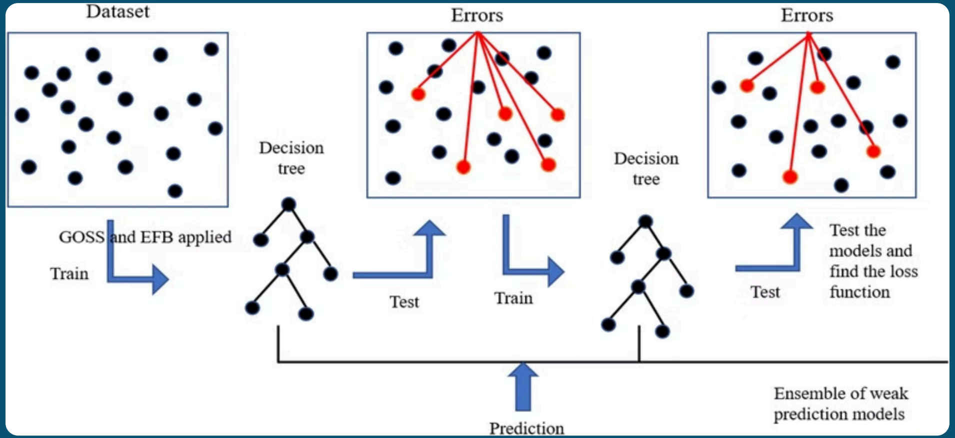
Azamat

MS in Quant Finance, ETH & UZH



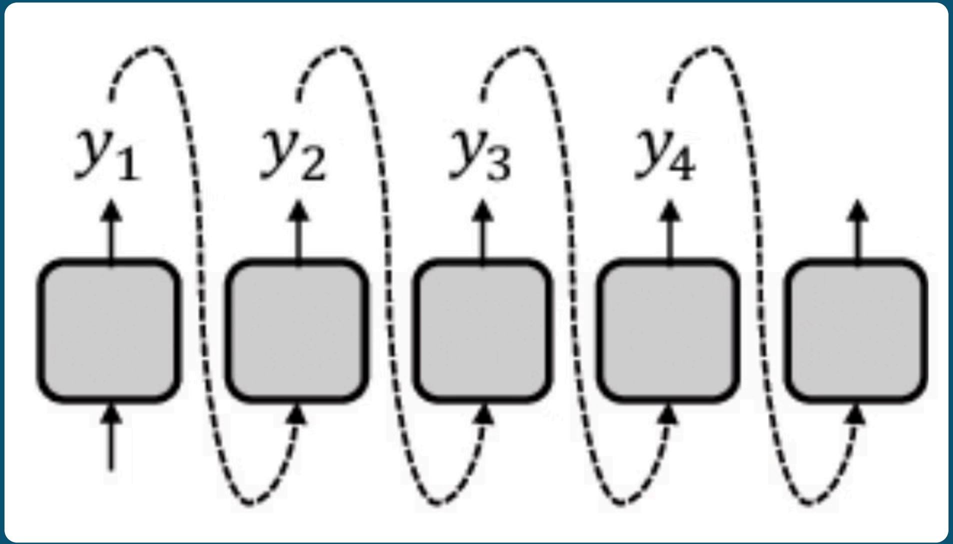
Problem Description

- Energy supply must match demand in real-time, deviations lead to financial penalties.
- Given a heterogeneous, imbalanced historical dataset for Spain and Italy with missing values
- Our goal is to develop an hourly energy consumption forecasting model for one month horizon at both:
    - Individual consumer level
    - Aggregated portfolio level



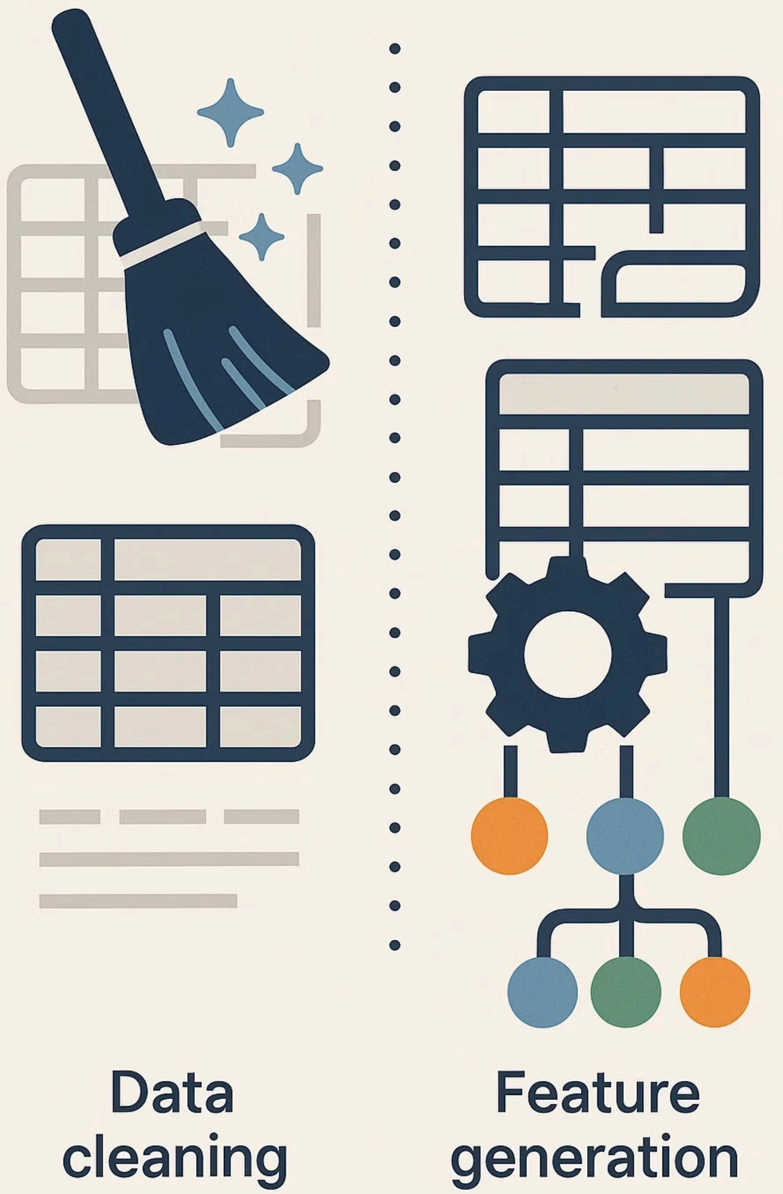
Light GBM

- Light Gradient-Boosting Machine that constructs an ensemble of decision trees
- Can automatically handle missing features => no need for imputation
- Faster to train compared with many other gradient boosting methods, such as XGBoost
- As a result, we can train 1 LightGBM per client! => solve the Spain v.s. Italy class imbalance and learn individual behaviors of each client better



AutoRegression

- Starting with the last available consumption timestamp, we predict the consumption in the next hour
- Next, we use the output of our model as the input to the model for the next timestamp prediction, etc



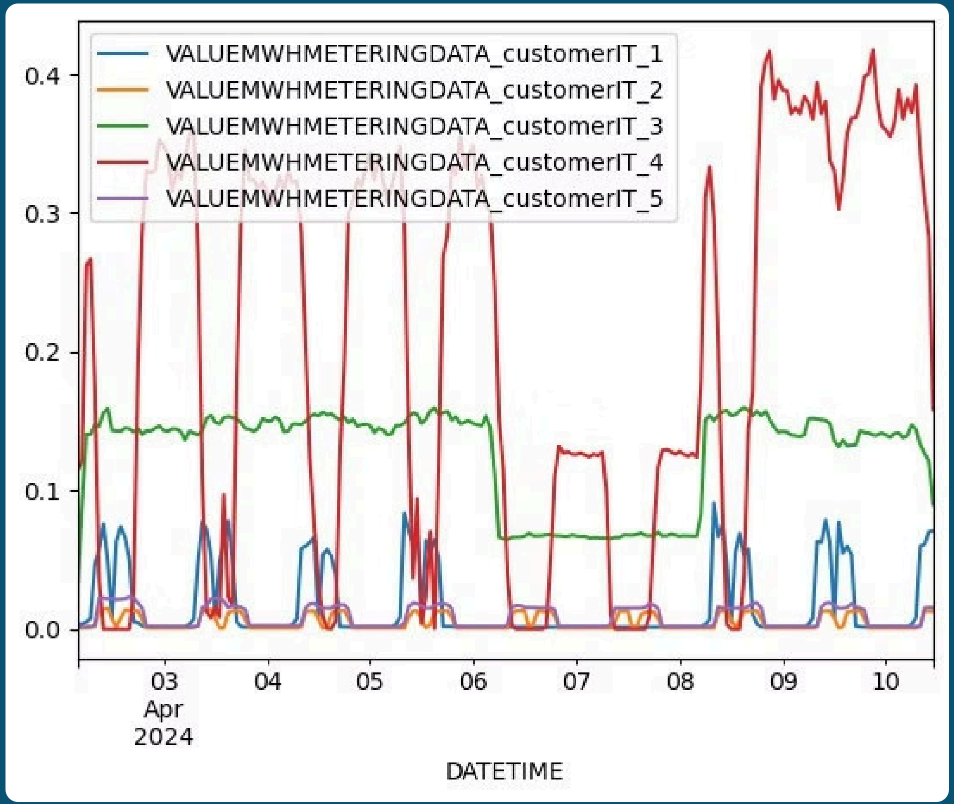
ALPIQ

Data Cleaning & Features Generation

- We used the original input features (regional holidays, temperature, photovoltaic production, and initial rollout)
- To enrich the input features, we supplied:
  - Temporal attributes
  - Statistical summaries over a rolling window of consumption measurements
  - Lag variables (consumption several hours ago)
  - Fourier transforms over the rolling window consumption

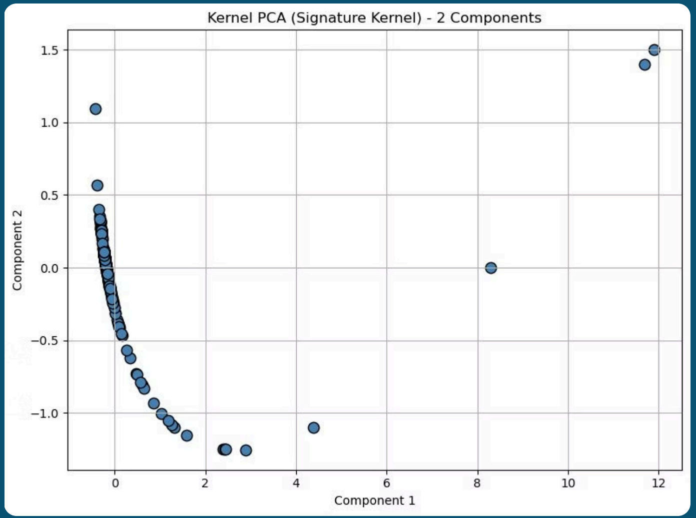
Customer-specific modeling approach

Due to low similarity across customers, clustering was not used.



Consumption pattern for 5 Italian customers

Figure shows a high degree of heterogeneity

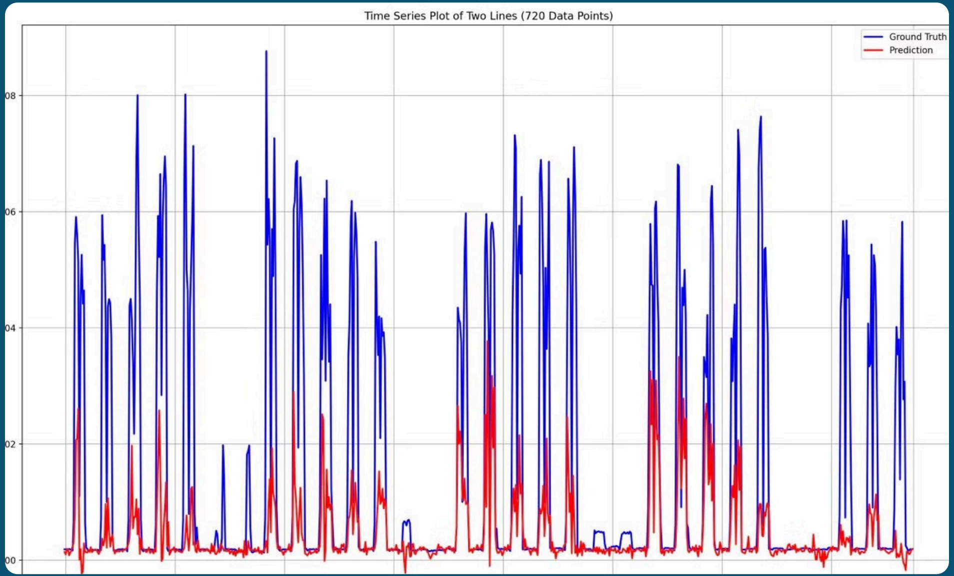


Kernel PCA with the Signature Kernel

Figure shows the lack of clear separation

Model predictions

- As an example, consider the prediction of the trained Light GBM for a particular customer in July
    - The model learned to produce lower consumption on the weekend
    - The output closely follows the trends in the ground truth
- Limitations:
    - There is still a significant mismatch between predictions and ground truths!



Model prediction visualization

ALPIQ

Model Evaluation

We implemented a comprehensive time-series cross-validation pipeline for model evaluation.

- Uses TimeSeriesSplit to create training and testing splits that respect the temporal order of the data.
- Evaluates the model across multiple folds for each customer.
- Computes the mean and standard deviation of the following metrics across all cross-validation folds:
    - Absolute error (per customer)
    - Portfolio-level error (aggregated across all customers)
    - A combined score using a weighted penalty scheme.

Results

- The model accurately captured the overall trend of actual consumption.
- The model effectively learned key temporal dynamics, including seasonality and holiday effects.
- The cross-validation showed the model is consistently robust across customers and folds.

Next steps

- As an additional step to increase the model complexity, we can:
  - Increase the number of iterations the Light GBM is trained for
  - Decrease the learning rate
  - Increase the depth of the trees and the number of leaves
  - Perform further feature engineering, removing some redundancy
  - Include Daylight Saving time adjustment