Date 2023-03-28

Author Ellevio

Background

To meet society's coming electricity need and changes in energy consumption patterns, the electrical grid is evolving. Data analytics is one of the key solutions Ellevio is focusing on to meet this demand and make cost-effective investments in the future.

One of the areas Ellevio wants to invest in is to proactively avoid exceeding its tariff to the overlying transmission grid. By forecasting when these occasions arise, Ellevios can order more electricity, or purchase flexibility solutions from their customers to avoid exceeding the tariffs.

A crucial aspect of the prognosis is being able to predict the total electricity consumption per hour for the grid. The electricity consumption varies over seasons, weeks and even during the day. One variable that affects the total electricity consumption to large extent is the temperature, as a large amount of the electricity today is used for residential housing heating.

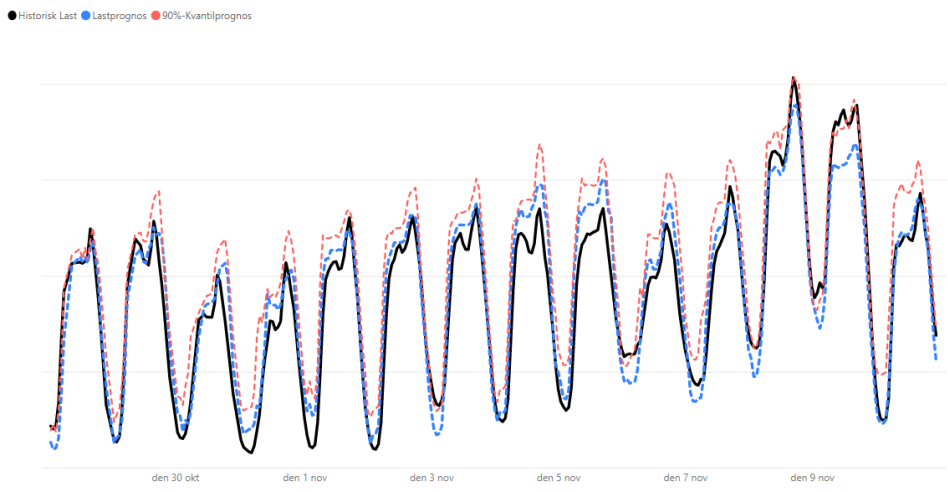

Figure 1: A visual evaluation of the predicted power load of two quantile models. The red line corresponds to a quantile model with alpha=0.9 and the blue line corresponds to a quantile model with alpha=0.5. The black line is the actual power.

Competition

Ellevio needs your help with producing a prognostics model that will be used to predict the load on the grid on the coming days, to optimize the operation and provide all customers with power.

Dataset

You are provided a dataset with data from 2012-01-01 to 2018-03-31. The dataset contains hourly data with the following columns:

timestamp – The time of the row in timezone UTC+1.
temperature_forecast – The forecasted temperature 24 hours ahead of the timestamp.
cal_day_in_week – Indicates day of the week.
holiday_flag – Indicates whether the day is a holiday or not.
holiday_desc – Describes which holiday (in Swedish).
work_day – Indicates whether the day is a working day or not.
cal_week_in_year – Indicates the week of the year.
cal_day_in_year - Indicates the day of the year.
cal_year – Indicates the year.
power – Total average power usage in the area in MW per hour.
hour – Indicates the hour of the day.
electricity_price – The price of electricity in SEK per MWh.
temperature – The observed temperature (from SMHI). The value is not an average. It is the temperature observed exactly at the timestamp.
wind_direction – The observed wind direction at the timestamp (from SMHI).
wind_speed – The observed wind speed at the timestamp (from SMHI).
humidity – The observed relative humidity at the timestamp (from SMHI).
global_irradiance – The observed global irradiance at the timestamp (from SMHI).

You are provided with a temperature forecast 24 hours ahead of time (*temperature_forecast*). You can create your own date features from the date parameters in the dataset.

Important assumption

Forecasts are not provided for *electricity_price*, *wind_direction*, *wind_speed*, *humidity* and *global_irradiance*. Assume that you are able to predict these values 24 hours ahead, i.e. it is OK to shift these values 24 hours ahead and use as a forecasted_value.

Tasks

The contest consists of two parts:

Feature Engineering

a) Select and create features that are best used to predict the power load 24 hours ahead. Describe the method that was used to derive these features.
b) Transform the original dataset into a training and test datasets, based on the features that were selected. The test period should be from 2017-11-01 00:00:00 to 2018-03-31 23:00:00.

Accuracy and Performance

a) Create a model that is trained on your training dataset and validated on your test dataset. It is OK to assume that the training dataset is available up until 1 hour before the forecast is created, i.e. if a forecast is created based on input data for 2017-11-15 14:00:00 it is OK that the model is re-trained/re-tuned with a dataset that includes the data for 2017-11-15 13:00:00.
b) Create predictions for every hour using the test dataset as input. Save the predictions as a comma-separated CSV file with utf-8 encoding (same format as the provided dataset file). The CSV file should contain two columns: the predicted power and corresponding timestamp.

Evaluation

The winner is decided by evaluating your prediction model on three metrics. The goal is to have an as low score as possible.

1. *Feature Engineering. (Pass/Fail)*

   The selected features are clearly explained and it is described why and how they were chosen and/or created.

2. *Accuracy and performance of the model to predict the power load for an entire winter season. (Contest score)*

   The accuracy and performance score will be calculated as

   $$Score = |Mean\ Error| + Absolute\ Mean\ Error + Mean\ Pinball\ Loss\ Function$$

   The submission with the lowest score will win.

3. *Quality of the code written. (Pass/Fail)*

   A data scientist at Ellevio should be able to read and understand the code. A recommendation is to use a code formatter, for example black.

To ensure fair play, the accuracy and performance of the models will also be tested against a secret validation dataset.

How to submit your entry

Please provide the following by attaching it as a zip archive in an email and send it to dadc@ellevio.se from your kth email before the 15th of May. The code files can be submitted either in the zip archive or as a link to your GitHub repo where the code is located. Notebooks or script files are both accepted.

- The CSV file that contains the predictions based on the test dataset.
- The code(s) containing:
    - The feature engineering process
        - This should include how the original dataset is transformed into the training and test datasets.
        - Optional: Provide a short document (1-3 pages) that explains the process. Word/PDF/markdown files are accepted.
    - The creation/training of the model
    - The prediction/test process
    - Alternatively, send your algorithm directly as an attachment

Not sure how to get started?

Go to Kaggle.com for inspiration and tutorials on how to get started with Machine Learning!