Journey to zero - Predict electricity consumption

Beringuier Théo / Bertin Paul

Task 2

Electricity prices have skyrocketed and consumers around the world are looking for the options to reduce electricity costs and environmental footprint. Enefit, one of the largest energy companies in Baltic countries wants to help its customers to reduce their energy consumption.

Both electricity cost and the environmental footprint could be drastically reduced by forecasting the consumption of the household and optimizing its energy usage (controlling smart energy devices in such a way that minimizes the cost and environmental footprint of the consumption).

The goal of this competition is to create an energy consumption prediction model for a single household. With the data of a previous year, we will train a model to predict which will be the next consumptions throughout the current year and make the client aware of it. This will enable every client to reduce its energy consumption.

The evaluation of the work is done with the mean absolute error (MAE) which is a measure of errors between predicted and real values. The mean absolute error is calculated as the sum of absolute errors divided by the sample size. We must submit the work under the following format, a table of two columns, in the first one, the time and the other the energy consumption.

We are two people working on this project, Theo and Paul.
The resources available are the train data and test data that are two csv files that show energy consumption every hour for one year and other meteorological events.

The training data could be bad quality and we will need to work a longer time on it.

Time series: It is a sequence of successive time equally spaced.

Forecasting: It is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends.

Log: It is a reminder of the previous record.

The goal of the project is to train a model that would achieve the lowest MAE score and be the highest ranked in the kaggle competition. We will start with basic models such as RandomForest or Regressor and then we will also try the XGBoost model.

To have interesting results, it is necessary that the data be as qualitative as possible. It is also necessary to have as much data as possible so that we can train the models and have a better accuracy and in the same time, the lowest MAE.

Task 3

Every data collected by Enefit for 1 year is not used to train our model. Indeed, only 11 over 13 data are used which are the columns of time, temperature, dew point, humidity, wind direction, wind speed, peak wind gust, pressure, weather conditions, electricity price and consumption. The two columns we removed are the precipitation and snow columns. Recordings are made every two hours. This gives us a large amount of data over a short period of time (1 year for 8592 records).

We deleted those data because there is a very large amount of missing values (99% missing values for the characteristic 'snow' and 75% for 'prcp'). They have a low impact on the energy consumptions or even no link. We determined which of these columns were not necessary by calculating the correlation coefficients between consumption and the columns in question. We also determined them with some graphical representations and verified the correlation between the variations of values and the peaks.

To upload the trained file the time column must be put in a pandas Timestamp type as asked in the sample submission. To do this we had to modify the 'time' data line by line to separate the string and remove the '+03:00' part. Passing time data in Pandas Timestamp format then allows for easier TimeSeries work and, also the use of specialist libraries. All data are spaced one hour apart between August 31, 2021, and August 24, 2022.

The shape of the required data is a table of two columns presenting the time and the energy consumption. Time data should be given in the following form:

Year-Month-Day
Hours:Minutes:Seconds

We noticed that one month's worth of values was missing from the weather condition column. To overcome this problem, we averaged the values of the previous month and entered the data in the empty slots.
We will also look for outliers that might appear in the columns we keep to train the model. So, we will delete those outliers to replace them with the mean of the values around. To find the outliers, we will search the min and max values of each columns, calculate the mean and calculate the standard deviation to find which values are out of the bounds.
We have also done a lot of data engineering. For example, by adding lag columns or trends to the power of 1 and 2 to capture the trend of the series.

Task 4

We have prepared the data. The aim is to identify the characteristics correlated with energy consumption. To do this, we have drawn several curves that give us information on this subject. The different variance of the characteristics is also an important indicator of the importance of a column. As we have already said, some decisions have already been made concerning certain characteristics. For example, the characteristics 'snow' and 'prpc' have been completely removed because they have too many missing values and are not very correlated with energy consumption. In addition, some consumption values were missing. We replaced them with the previous values because the recordings are every two hours and the line by line readings are quite close. Finally, we did a TimeSeries job trying to capture the trend of the consumption data over time as well as the seasonality.