Journey to zero

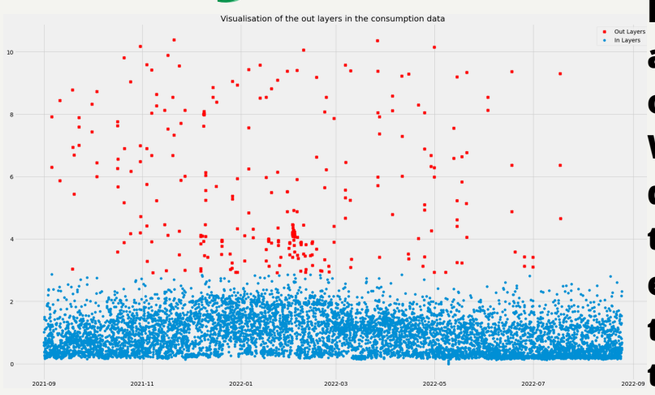
Introduction

Journey to zero is a competition in which competitors must find the best possible model to predict a household's energy consumption for one week using almost one year of weather data.

Cleaning Data

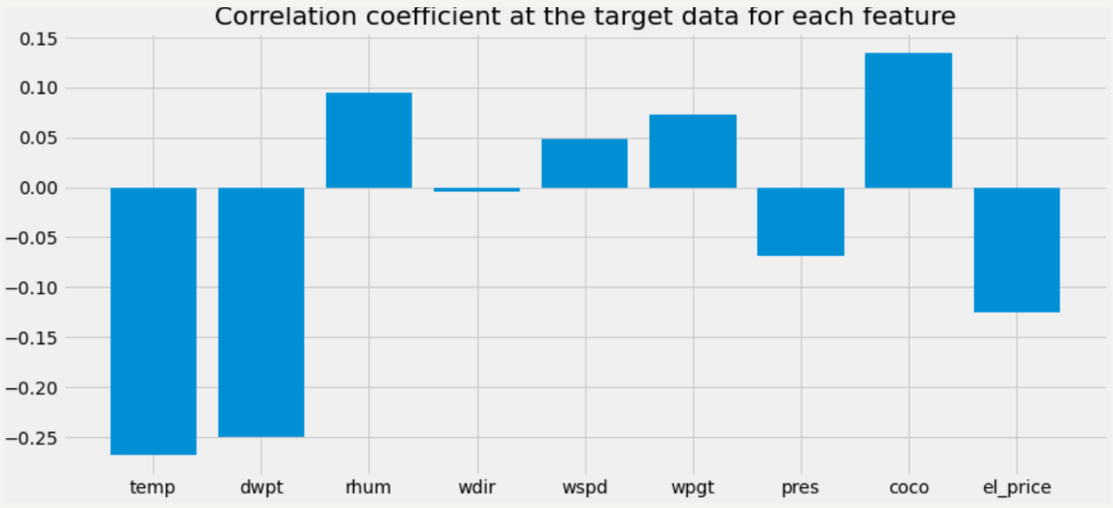
13 columns were given however the columns "snow" (quantity of snow) and "prcp" (quantity of precipitation) had respectively 75% and 99% of NaN values, so we deleted them. The whole month of June in the "coco" column had NaN values. To fill this lack, we implemented in each data the average of the next month. The consumption column also had NaN values, so we replaced them with the previous values.
Also, we have changed the form of the data in the time column by displaying them in Times stamp format.

Outlayers



For data pre-processing we looked at the outlayers that could distort our data and then remove them. We then drew boxplots to determine them and thus isolate these data. You can see an example of the outlayers related to consumption (here in red). In total we found more than 2000 outlayers on the 8592 lines of data.

Feature selection



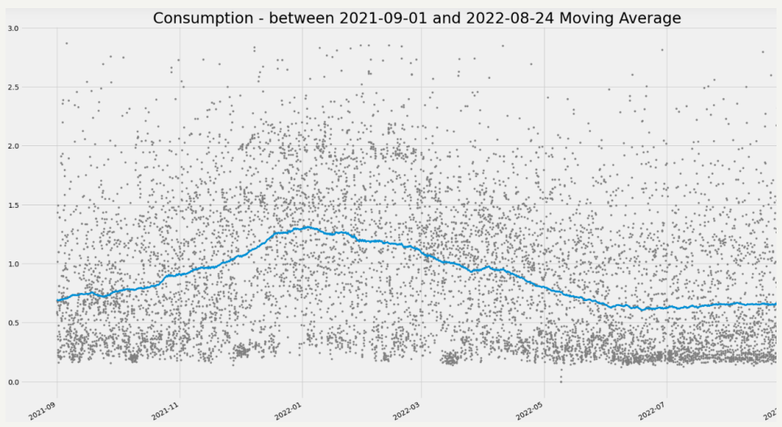
We plotted the correlation coefficients between all the features and consumption to see which ones are the most interesting to compute for predictions. We could conclude with this graph that we could start by dropping the 5 central features, i.e. "rhum", "wdir", "wspd", "wpgt" and "pres" columns.

Training

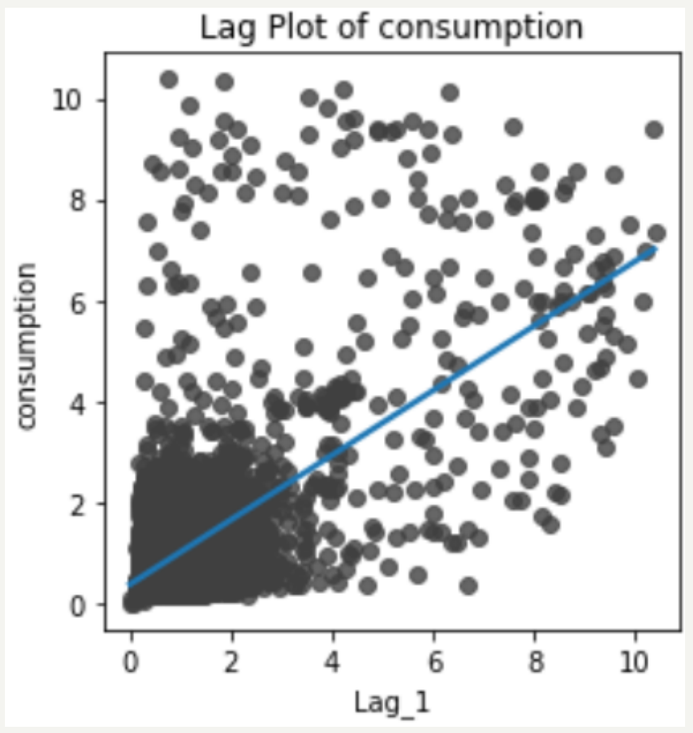
We did a 5-Fold cross validation to better determine the model results.

After several trainings with different data, we realized that the result was better when we did not use all the features. So we used only the temporal data which we separated into columns of hours, days, months, up to the year, the trend data and the two lag data at 7 days before and after.

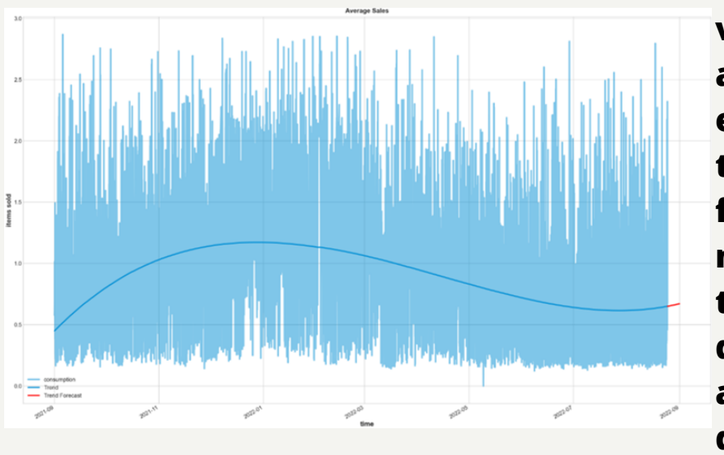
Time series



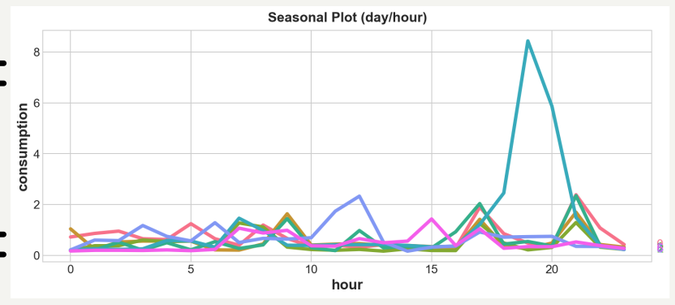
We then proceeded to a TimeSeries approximation of the data. To start we have to visualize the trends of the consumption. We can see on the rolling average plot a peak of consumption during the winter.



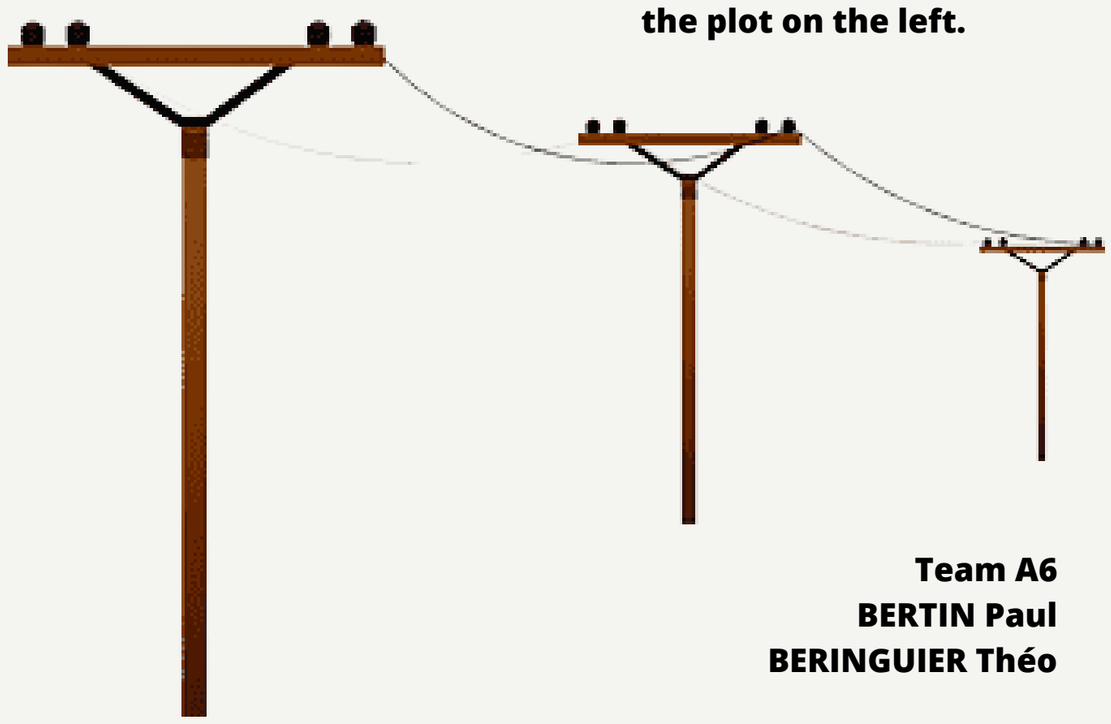
We can see on this graph of lag an increasing line. This shows that these are important and can improve our data. So, with this statement, we performed feature ingeneering by adding a lag column which is a shift of data in time. In our case, we did a lag of one week which is a total of 168 values. Moreover, if we loop the year by assuming that each year is similar, then the week after the week to predict is the first week of our training data. We can therefore make a lag on with the weeks before and after the week of the test set.



We have calculated the consumption trend (blue curve) which we have then visualized on this graph. It allows us to visualize the evolution of the data and at the same time the trend forecast (curve in red) of the missing week, that is to say the next 168 values. These data are useful since they allow us to increase the data by doing some feature engineering and improve the forecast result.



To improve our model, we also wanted to check if there was a seasonality of consumption and more precisely recurrent consumption over several days. However, neither of these two approaches was conclusive as we can see on the plot on the left.



Team A6
BERTIN Paul
BERINGUIER Théo