Capstone project proposal

Domain Background

For this final project, the goal is to work with the data from the RTE open data platform [1] there is information about the electricity production and consumption in France at the national and regional scale. For this project, I will focus my works on the national consumption data. This data is basically the average power consumed during the last 30 minutes at the scale of the French territories without the overseas territorial departments [2]. In the figure 1, there is an illustration of this data.

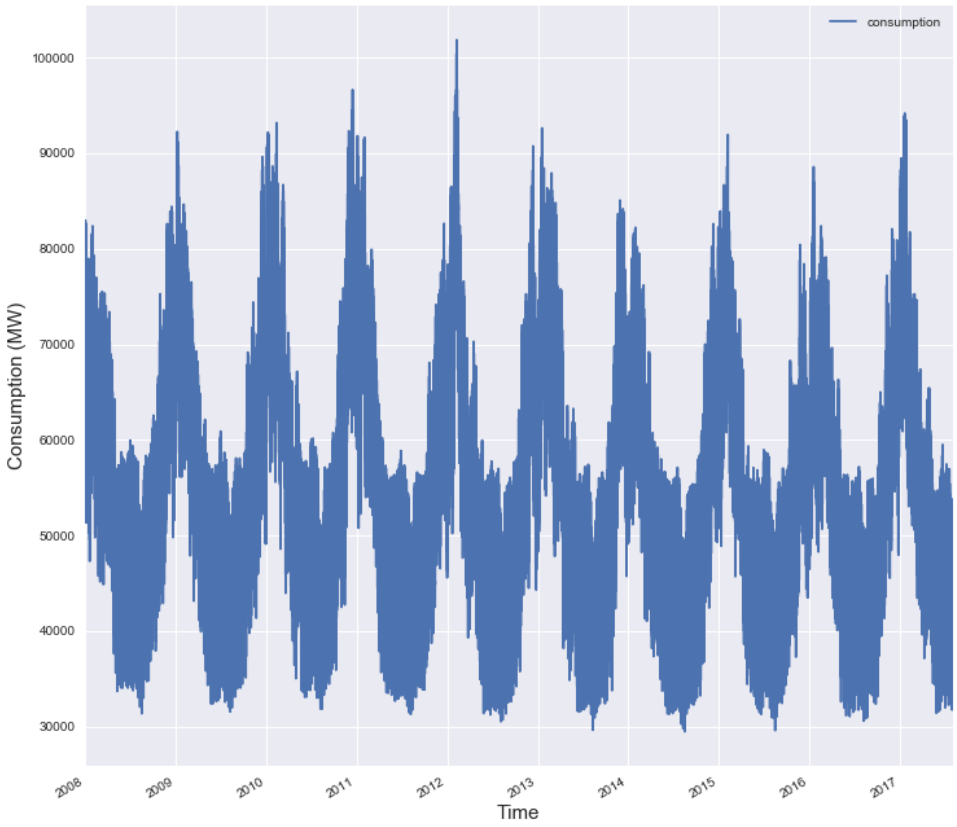

Figure 2: National consumption at 30min scale

My idea in this project is to create an algorithm that can forecast the electricity consumption at the national scale. The condition of France for their electrical consumption is that electricity is used for everything (heating and other usages as appliances), the electrical heating system are very popular in France by the fact that there is a big usage of the nuclear power plants to produce electricity so this energy is “cheap”. My personal interest to make this project, is that I am currently working for a utility company (EDF the biggest utility company in France) and I am a data scientist so it could definitely be a good project to support my future applications for my next job in this domain or in the data science in general.

To support my analytic, I will add two more datasets:

- The weather data from 73 airports in France, with different time scales, from weather underground [3]

- The data about the cities in France from IGN, like the number of habitants and the area of the cities in France [4]

There is no proper scientific paper on this topic (it's quite a sensitive topic for utility company), but there is paper where the researchers try to forecast the evolution of a timeseries like this model to predict the solar forecasting based on ANN [7] and this model to predict the wind speed at an hourly scale [8] that is using an ARMA model.

You can find an extract of the data, in the data folder.

Problem Statement

The inputs for this problem will be:

- The electrical consumption with 30 min time scale
- The selected weather stations with the parameters describe below
- The information associated on the cities around the selected weather stations

For these analytics, I will segment my approach in two parts:

- An analysis to try to predict the amount energy consumed at the national scale per day
- A 30-min forecast to try to predict in "real time" the consumption

In the first case, there is no time series involved but for the second case it will be a time series forecasting. I think that we can use the results of the first analysis in the second analysis to make the time series forecast.

In term of technics, like I said before I will mostly use supervised learning technics like polynomial regression or neural network. To assess the model, I will use the metrics like the r^2. To test different models, we can use the grid search and the Kfold to find the best and most efficient models to used.

Datasets and Inputs

In the following figure, there is an illustration in the heatmap of the evolution of the consumption in function of the month of the year and the hour of the day.

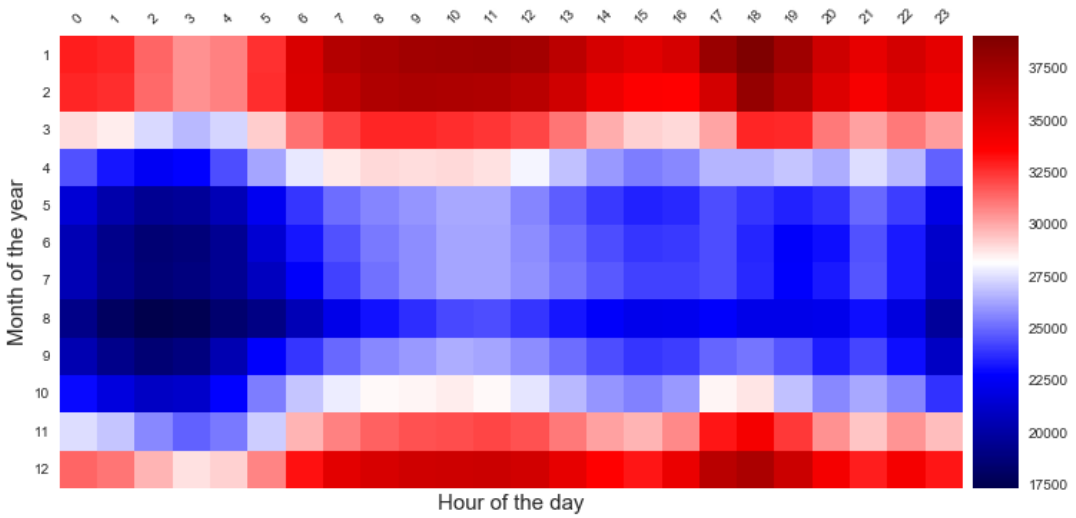

Figure 2: Heatmap of the average daily consumption in function of the month and the hour

It can be notice that the consumption is linked with the hour of the day with the morning peak (08:00 to 12:00) and the evening peak around 18:00. There is a seasonal effect of the electrical consumption with a higher consumption during the winter (by the usage of the heating system).

But in this dataset, basically the data can have two statuses:

- Definitive: The data are verified by RTE
- Intermediaire: the data are not verified completely by RTE (after January 2016)

I will have to focus my training/test period on the data that are before 2016 (with the definitive status).For the timeseries point of view that represents 140254 points to make the model, for the daily forecast it's 2922 days.

The seasonal effect could be assessed in this analytic with the data from the 73 weather stations of weather underground.

Figure 3: Map of the weather stations

The weather stations are on the different regions of France and the distribution on the territory is quite good, for each station we have the following parameters:

- Outdoor temperature
- Outdoor humidity
- Precipitation
- Gust speed
- Wind speed
- Wind direction

I don't know yet what will be all the right features to use for the weather data but the outdoor temperature and the wind speed are a good start. But I think that it could exist a good combination of weather features that could help me to make a proper forecast algorithm.

The IGN data is presented for each city in France by number of inhabitants and metropolitan area (city size). This data will be used to give a weight to the weather data on the national forecast because if there is less people in this area, there will be less consumption. So, I will have to find the right index to make the combination with the weather data.

For this model, it will be important to use as an input the time of the day, day of the year and day of the week ,these three parameters have a clear impact on the electrical consumption.

To train the models, like I said previously I will use the consumption until 2016, I will split my training /testing set and validation set basically with 90%/10% ratio. I will use on the training test set the Kfold approach on the randomised dataset.

In the following figure, there is a summary of the elements that will be implied in the conception of this model.

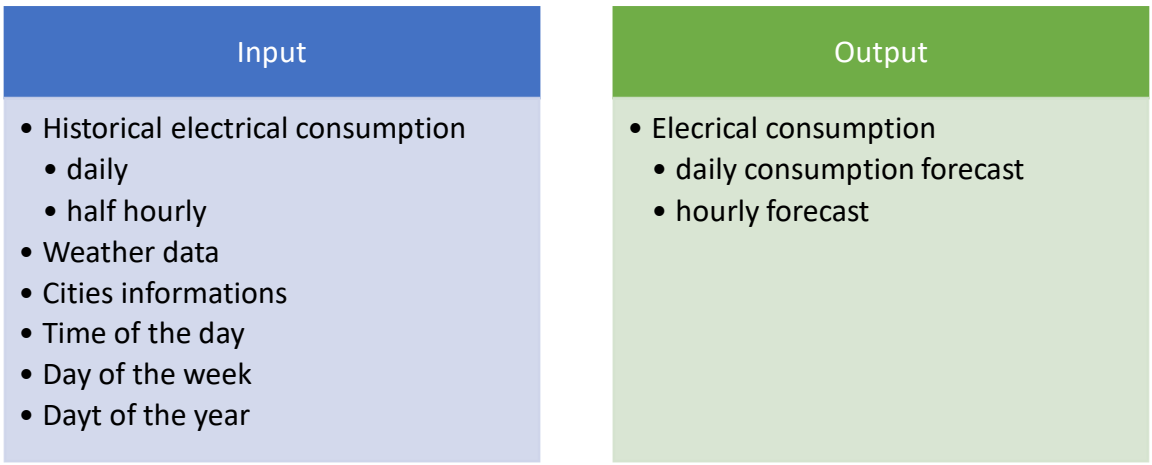

Figure 4: State of the input and outputs for this project

Solution statement & Benchmark model

There are some papers that are illustrating the evolution of the energy consumption of a household in function of the average outdoor temperature [5]. If we apply the same approach at a national scale with the daily consumption data in function of the average outdoor temperature of a specific place in this case Clermont-Ferrand (in the centre of France).

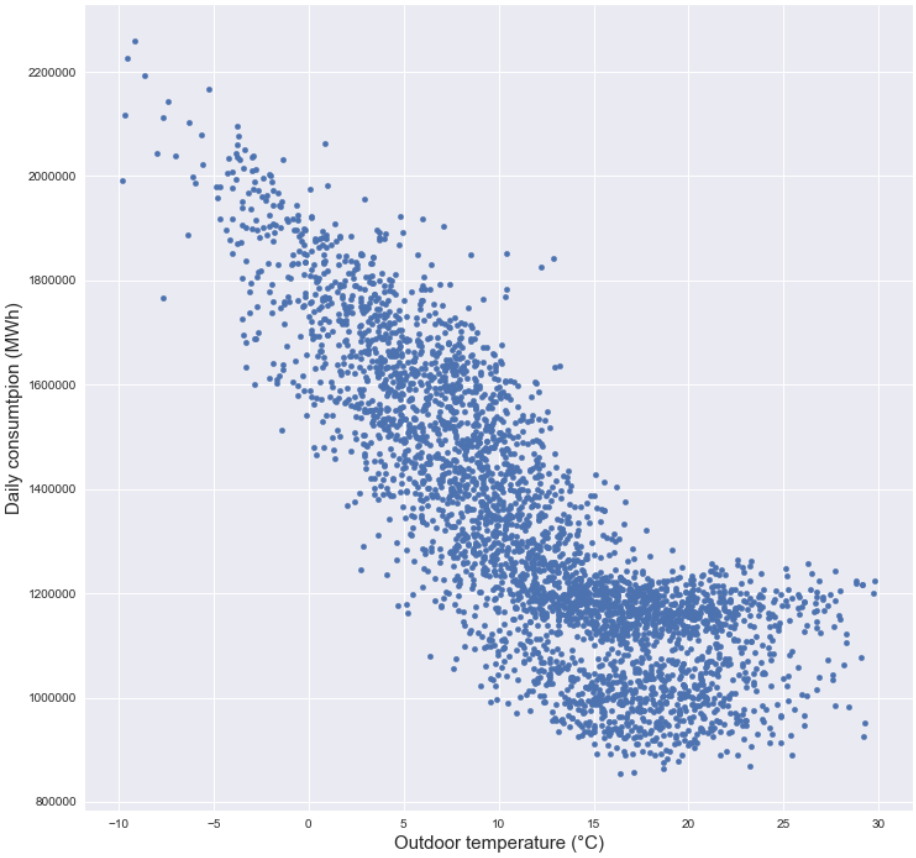

Figure 3: Daily consumption in function of the average outdoor temperature

It can be noticed that there is a relation (linear regression by piece) between the consumption and the outdoor temperature as we expected by the works on the subject [5]. This first model could be a benchmark model for the daily forecast of the national consumption.

For the case of the timeseries forecast, use a similar approach than in [7] or [8] with a neural network based on the outdoor temperature could be a good start as a benchmark model. To improve the neural network, adding new layers, new activation functions and new index to replace the outdoor temperature could be a way to find the right model.

For example, in the UK, there is an index that is used to forecast the gas consumption, this index is call CWV for Composite Weather Variable that is function of the outdoor temperature, the wind speed and a seasonal factor. The relation between the gas consumption and the CWV is quite similar that the one with the outdoor temperature but the dispersion of the linear model is less important [6].

The technics that I will used will be the following:

- Polynomial regression
- Neural network
- K-nearest neighbour
- Random forest

In both case (daily analysis and timeseries analysis) it will be important to create new index based on the weather data and the population in the regions of the weather stations to be the inputs with the training data for the different models.

Project Design & Evaluation Metrics

For this project, I will use the same approach for the daily model and the hourly model. In the following figure, there is an illustration of the process to complete the project.

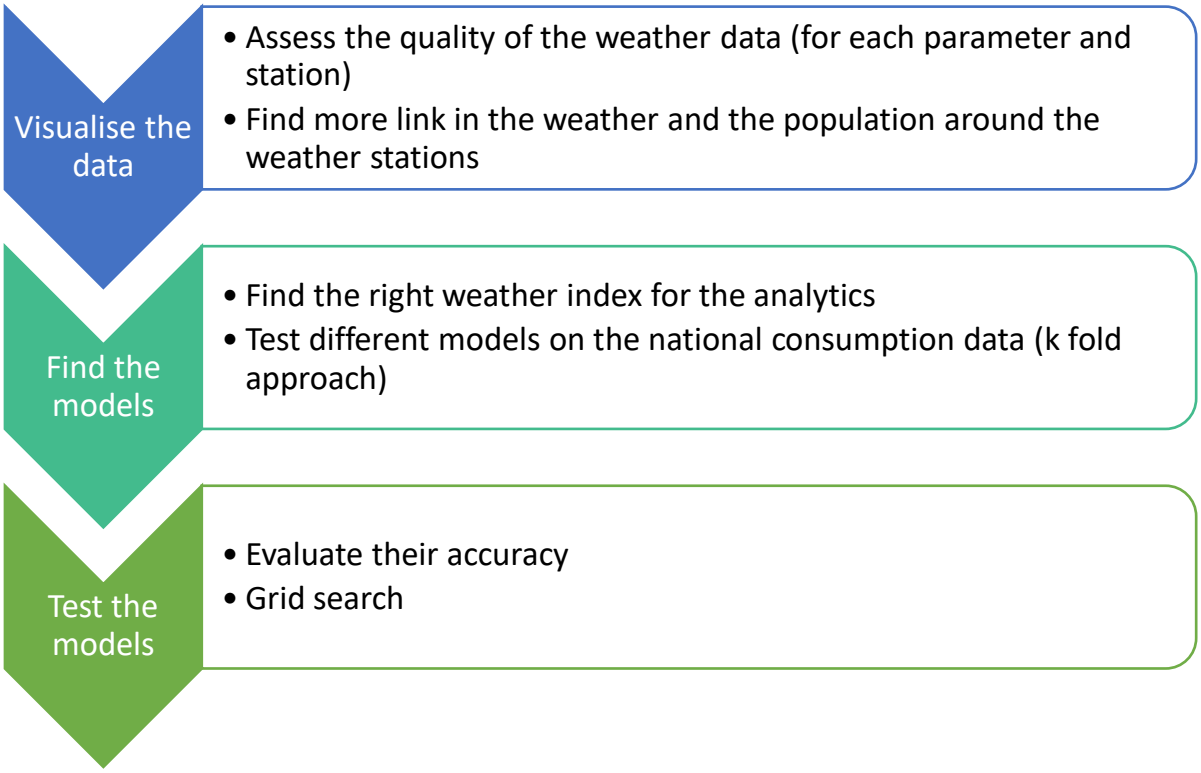

Figure 4: Process for the analysis

For the visualisation part, I will have to find the right correlation (and maybe combination) between the weather parameters and the electrical consumption at a daily and half hourly scale. This part can also help me to find the right weather stations, the one with the best data quality. When I will do the split between training test and validation set I will have to randomize the selection of the data.

For the second step, I will have to create the different models based on the training dataset. The k fold approach will be the best way to find the model .

For the testing part, I will have to use the validation set (10% of the initial dataset) with the models and if I am referring to the scikit learn documentation [9] I will use the r^2 score to assess the efficiency of the model on the validation set.

To complete the analysis, it could be interesting to work on the size of the different trainin/test set to see the impact on the models.

References

[1] RTE Open data platform, https://rte-opendata.opendatasoft.com/pages/accueil/

[2] RTE Open data platform, Electricity consumption load graph (january 2008 - july 2017), https://rte-opendata.opendatasoft.com/explore/dataset/cdc_conso/?disjunctive.qualite

[3] Weather underground, https://www.wunderground.com/

[4] IGN, http://professionnels.ign.fr/geofla

[5] Jonathan Chambers, http://www.ibpsa.org/proceedings/BS2015/p2854.pdf

[6] National Grid, Gas demand forecasting methodology, file:///C:/Users/daign/Downloads/Gas%20Demand%20Forecasting%20%20Methodology%20(1).pdf, page 20-27

[7] Marquez R, Proposed Metric for evaluation of solar forecasting models, http://solarenergyengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1662255

[8] Stefsos A,A comparison of various forecasting techniques applied to mean hourly wind speed time series.

[9] Scikit learn, Model evaluation: quantifying the quality of predictions , http://scikit-learn.org/stable/modules/model_evaluation.html