Xavier Leprintemps LEKEUKA FOUELEFACK

Clara Rita NASSIF INF442 Programming Project

A feasibility study of predicting household power consumption based on meteorological data.

Some background information

Before diving into the problem, we thought it would be best to keep in mind that:

- There are four seasons in France.
  - Winter (December - February)
  - Spring (March - May)
  - Summer (June - August)
  - Autumn (September - November)
- Holiday seasons and vacations must be considered when it comes to power consumption: whether a house is occupied or not during vacations directly impacts the use of household equipment in ways that don't necessarily match the power consumption's everyday routine of said house.
- No particularly exceptional weather conditions were reported between 2007 and 2010.

Data pre-processing

Pre-processing aims at cleaning the data set by removing variables with the least data, mostly noticeable by an important proportion of "not a number" (nan) indications in the data. In addition, its goal is to detect outliers and remove them or wisely modify their values so that they don't add bias to our data. Since we're working on a time scale, which is continuous, we choose to include outliers in the interpolation rather than delete them, to maintain the same evolution tendency between the original and the pre-processed dataset and thus, have the most exact predictor possible.

To lighten our datasets in terms of number of variables, we also studied correlations between variables (Fig.1): in each set of perfectly correlated variables, we kept one variable. This was only the case for global_intensity and global_active_power: we kept global_active_power because intuitively, it could represent a power consumption label later.

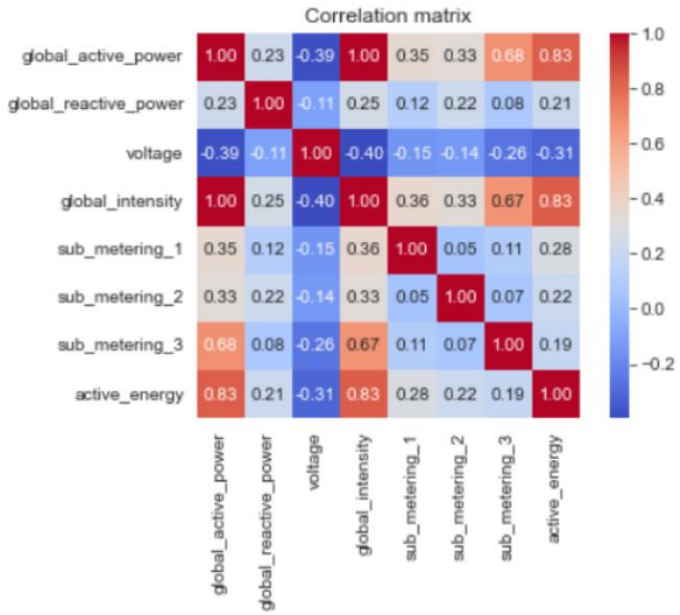

Fig.1: Correlation matrix between different features of the power consumption dataset

Depending on the content, pre-processing turned out to be different for each dataset.

Weather dataset

- Some columns were identified as having between 60% and 100% of data registered as "nan". We started by removing these columns.
- Then, we calculated Z-scores for each cell to detect outliers: every cell with a Z-score outside of the interval [-3;3] was considered an outlier.
- Finally, the remaining "nan" were replaced by backward fill or forward fill, and we proceeded to a linear interpolation using Python's default interpolation method.

Power consumption dataset

- Some lines were identified as having no data but "nan". Thus, we temporarily removed the corresponding lines.
- Then, we calculated Z-scores for each cell to detect outliers: every cell with a Z-score outside of the interval [-3;3] was considered an outlier. Every outlier is replaced by "nan".
- Afterwards, we re-added the lines deleted in the beginning.
- Finally, we operated a linear interpolation using Python's default interpolation method.

We had to bear in mind that both the weather dataset and the power consumption dataset had to be merged at some point to detect correlations between the house's power consumption and the weather conditions. We choose to transform every three-hour interval into minutes by duplicating the line to be able to switch from an hourly time scale to a daily time scale easily. We formulated the hypothesis that, on a three-hour interval, the meteorological conditions are roughly the same, allowing us to duplicate the corresponding vector for every minute of this interval.

Seasonality detection

Applying the elbow method on the power consumption dataset showed that the optimal number of clusters for the k-means algorithm is 2, at the elbow of the plot representing the variation of the minimum variance (function of the clusters' centroids and the standard deviation) in terms of the number of clusters (Fig.2).

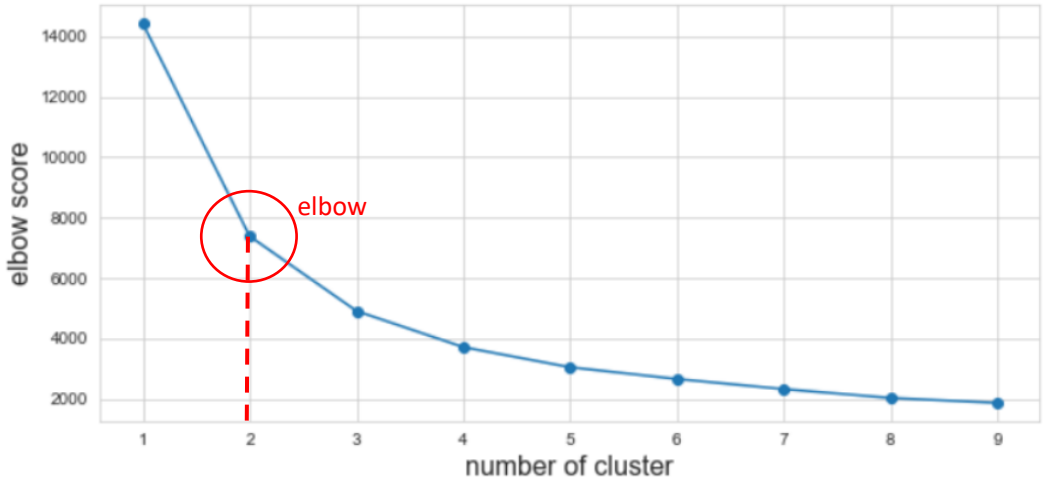

Fig.2: Elbow method

We confirmed this result by plotting, for each year, the average silhouette, to avoid the bias of 2 being the minimal number of clusters for which an elbow can be observe: the seasonal power consumption tendencies are similar from a year to another, as the silhouette method gave an optimal number of clusters of 2 for years 2008 to 2010 and 3 for 2007, which is coherent with the result of the elbow method.

Therefore, we can identify two definitive periods of the year where same or very similar power consumption patterns are found. Intuitively, we can associate two seasons out of four to each of them to justify this result: the first period of the year would be colder seasons (winter and fall), whereas the second would correspond to hotter seasons (spring and summer) (Fig.3).

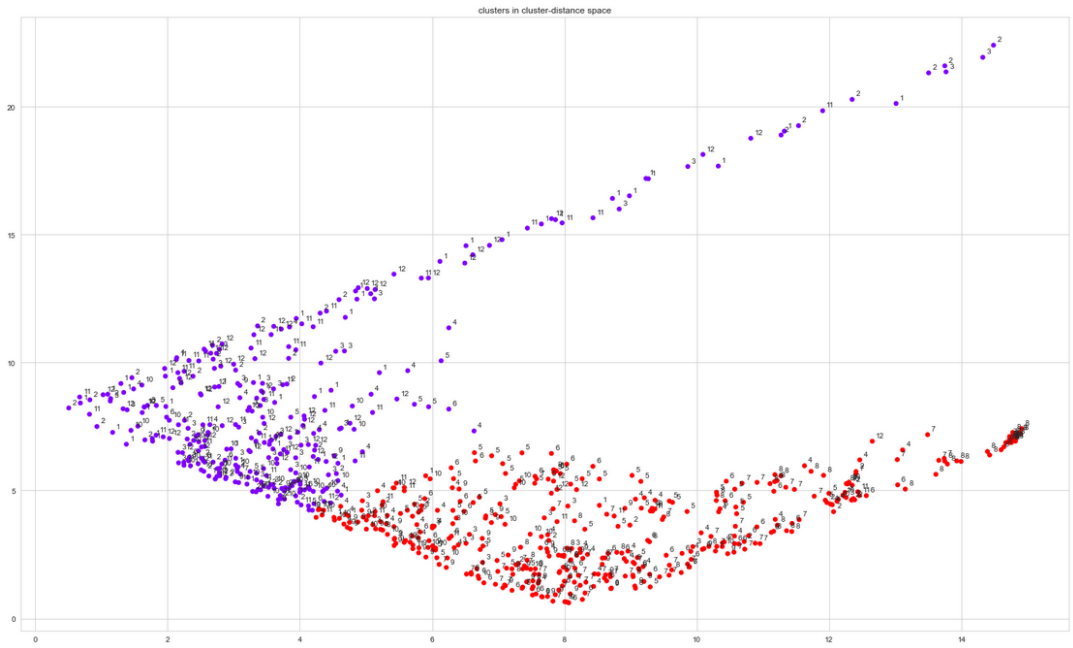

Fig.3: Representation of the clusters – each point has a label between 1 and 12 corresponding to the month it belongs to

In addition, we visualized the label attributed to each data point (i.e., each day) during a year: it showed us that the label 1 was attributed to days at the beginning and at the end of the year, which corresponds to winter and fall respectively. Likewise, most days in the middle of the year were labeled 0, and this time of year corresponds indeed to spring and summer.

Distribution of hourly consumption

In theory, plotting the variation of global_active_power in terms of time (in hours) gives an idea of the distribution of hourly consumption. During hotter seasons, one would expect that the power consumption plot will draw a peak in the middle of the day, where it is most sunny and hot and air conditioning is most needed. Similarly, during colder seasons, one can expect to see a peak at nighttime, when it is colder and the heat is most needed, and in the morning, to get a supply of hot water for the day.

In practice however, plotting the variation of global_active_power in terms of time (in hours) for some days selected at random during each year, turned out to give irregular functions, with punctual peaks at hours that would vary from one day to another. We used the Kolmogorov-Smirnov test to determine whether the shape of hourly power consumption resembles a Gaussian. It turns out that plotting the variation of power consumption in terms of time, even in minutes during a single hour of the day, isn't a Gaussian: the distribution changes a lot between two hours of the same day and between the same hour taken on distinct days.

Therefore, we tried to plot a histogram for each day (at least some of them), counting for how many hours each value of the power consumption was reached. This last plot turned out to be a gaussian.

Calculating the p-value for the power consumption for each season lead to the following results:

Season	**p-value**
Winter	$\approx 10^{-29}$
Spring	$\approx 10^{-29}$
Summer	$\approx 10^{-24}$
Autumn	$\approx 10^{-34}$

Given the fact that the reference value for the p-value is 0.05 to consider the distribution to be at least close to a Gaussian, we can conclude that the distribution isn't a Gaussian in any season. However, distributions are similar between seasons, especially winter and spring.

Weather influence

When we computed the correlation matrix with the features of the merged datasets, we noticed that, surprisingly, correlations were only visible between features of the power consumption dataset separately, and features of the weather dataset separately. In other words, there is roughly no correlation (or a low correlation) between two features each taken from a different dataset: only humidity, snow height and temperature are a little correlated with global_active_power, with a maximum (in absolute value) correlation coefficient of -0.35. This could be explained by the fact that a lot of measured parameters in terms of power consumption (sub_metering_1 and sub_metering_2), include household equipment that is used indifferently in cold or hot weather, such as a fridge or a dishwasher.

We selected the features that we felt were linked to global_active_power (our label), i.e. those that cause the highest correlation coefficients in the matrix: ht_neige (snow height); t (temperature); td (*point de rosée*); u (humidity). The idea was to determine a minimal absolute value v of the correlation matrix coefficient such that for each coefficient a of the global_active_power line, if $|a| \geq v$, then we consider the variables to be correlated. This minimal value must be easy to reach -maybe less than 0.1- since there are very little correlations between the weather and power consumption datasets. With a linear regression, we obtained a Mean Squared Error of 0.69.

We wanted to try an approach with SVM, but we didn't know how to proceed since the labels are continuous. We also realized that with this many data, a neural network could be more helpful and accurate, but we didn't have time to implement it.