

## Data Insights

- The data is heavily imbalanced
- No Duplicates
- Some channel sales and origin up channels have a few number of observations. We will treat them as missing ones
- margin\_net\_pow\_ele and margin\_gross\_pow\_ele are almost identical except for two observations!
- Viewing the scatter plot of net\_margin" and "margin\_net\_pow\_ele", they did not sound to have a direct correlation. This is sounds very odd as the margin of net\_pow\_ele should be somehow composited inside the net margin
- The number of products implies customers that have more than 5 services or products are less likely to churn from others and also antique customers.

## Data Augmentation

- In some months some SME are more actually to renew. However it wouldd be helpful to cluster months into yearly periods divided as "Peak", "Off Peak", "Mid Peak" To be consistent with the idea that consumbtion and price does change with the period of the year.
- Some channel sales and origin up channels have a few number of observations. We will treat them as missing ones.
- Data below are heavily skewed, however, their log transformation shows a normal distribution
  - cons\_12m
  - gas\_cons\_12m
  - margin gross pow ele
  - net margin
  - Current paid consumption
  - maximum power
  - forecast\_cons\_year
  - forecast\_meter\_rent\_12m
  - forecast\_price\_energy\_off\_peak
  - Forecast\_price\_pow\_off\_peak

## Columns to drop

- Forecast\_discount\_energy only 3% of this feature has values and the rest of observations are zeros which is very unhelpful
- Forecast\_price\_energy\_peak has 50% of its values set to zero which is also very unhelpful

## Data Suggestions to be provided

- Provide regions of companies to understand if SMEs in a certain location and more likely to churn or what.
- Provide clients feedbacks if applicable.