Machine Learning to explain the price of electricity [1]

1 Context

Every day, a multitude of factors impact on the price of electricity. Local weather variations will affect both electricity generation and demand for instance. Long term phenomena, such as global warming, will also have a significant influence. Geopolitical events, such as the war in Ukraine, may affect in parallel the price of commodities, which are key inputs in electricity generation, knowing that each country relies on a particular energy mix (nuclear, solar, hydro, gas, coal, etc). Moreover, each country may import/export electricity with its neighbors through dynamical markets, like in Europe. These various elements make quite complex the modelisation of electricy price in a given country.

2 Project Goal

The aim is to model the electricity price from weather, energy (commodities) and commercial data for two European countries - France and Germany. Let us stress that the problem here is to explain the electricity price with simultaneous variables and thus this is not only a **prediction** problem.

More precisely, the goal of this challenge is to learn a model that outputs from these explanatory variables a good estimation for the daily price variation of electricity futures) contracts, in France and Germany. These contracts allow you to receive (or to deliver) a given amount of electricity at a specified price by the contract delivered at a specified time in the future (at the contract's maturity). Thus, futures contracts are financial instruments that give you some expected value on the future price of electricity under actual market conditions. In this project, we focus on short-term maturity contracts (24h). Let us stress that electricity future exchange is a dynamic market in Europe.

Regarding the explanatory variables, the participants are provided with daily data for each country which involve weather quantitative measurements (temperature, rain, wind), energetic production (commodity price changes), and electricity use (consumption, exchanges between the two countries, import-export with the rest of Europe).

3 Project Steps

The CRISP [2] method (originally known as CRISP-DM) was originally developed by IBM in the 1960s for data-mining projects. In Data Science, it remains the most widely used methodology. It is composed of 6 steps from the understanding of the business problem to the deployment and production. This method is agile and iterative, i.e. each iteration brings additional business knowledge that allows to better tackle the next iteration. In what follows, we adopt this methodology. Thus, do not forget, at the end of each iteration, to refine and update previous steps if necessary (based on new obtained additional informations).

[1] This project is issued from the Challenge Data proposed by the Data team of ENS Paris.

[2] CRISP-DM : Cross-Industry Standard Process for Data Mining. Cette annexe est issue de https ://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist

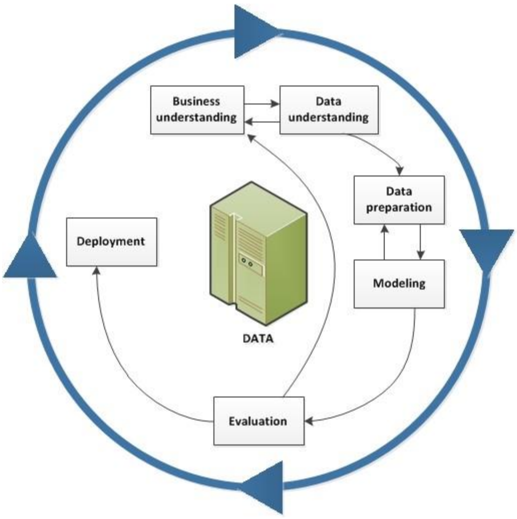

FIGURE 1 – CRISP-DM Phases

3.1 Data Description

We provide three csv file data sets :

- Input data **Data_X** and its output **Data_Y**,
- New Input unlabeled data : **DataNew_X.**

The input data **Data_X** and **DataNew_X** represent the same explanatory variables but over two different time periods.
Input data sets comprise 35 columns :

- ID : Unique row identifier, associated with a day (DAY_ID) and a country (COUNTRY),
- DAY_ID : Day identifier - dates have been anonymized, but all data corresponding to a specific day is consistent,
- COUNTRY : Country identifier - DE = Germany, FR = France,

and then contains daily commodity price variations,

- GAS_RET : European gas,
- COAL_RET : European coal,
- CARBON_RET : Carbon emissions futures,

weather measures (daily, in the country x),

- x_TEMP : Temperature,
- x_RAIN : Rainfall,
- x_WIND : Wind,

energy production measures (daily, in the country x),

- x_GAS : Natural gas,
- x_COAL : Hard coal,
- x_HYDRO : Hydro reservoir,
- x_NUCLEAR : Daily nuclear production,
- x_SOLAR : Photovoltaic,
- x_WINDPOW : Wind power,
- x_LIGNITE : Lignite,

and electricity use metrics (daily, in the country x),

- x_CONSUMPTON : Total electricity consumption,
- x_RESIDUAL_LOAD : Electricity consumption after using all renewable energies,
- x_NET_IMPORT : Imported electricity from Europe,

- x_NET_EXPORT : Exported electricity to Europe,
- DE_FR_EXCHANGE : Total daily electricity exchange between Germany and France,
- FR_DE_EXCHANGE : Total daily electricity exchange between France and Germany.

Output data sets are composed of two columns :

- ID : Unique row identifier - corresponding to the input identifiers,
- TARGET : Daily price variation for futures of 24H electricity baseload.

Note : The input data X_train and X_test represent the same explanatory variables but over two different time periods. The columns ID in X_train et Y_train are identical, and the same holds true for the testing data. 1494 rows are available for the training data sets while 654 observations are used for the test data sets.

3.2 Data Preparation

Depending on the kind of the problem and the objectives to be achieved, the preparation of the data usually involves the following tasks :

- Merging of datasets and/or records
- Selection of a subset of the data
- Calculation of new attributes
- Sort data for modeling
- Remove or replace blanks or missing values
- Split into training and test subsets

In this project, we ask you in particular to :

- Check for missing values in the data.
- Check if the values of the different attributes are comparable

Note. You can always suggest further data preprocessing but you will have to explain why in each case. To understand this project phase, you are invited to consult the following link : Data Preparation with pandas

3.3 Exploratory Data Analysis

This step consists in :

1. Identify the target variable to be predicted. In this project the predicted variable is the daily variation of futures prices (the TARGET column in the datasets Y_train and Y_test).
2. Carry out an Exploratory Data Analysis (EDA) using a variety of charts and statistics by following these steps :
   - Conduct an overview of the variables by examining their type, distribution, range of values, and significance
   - Examine the relationship between the characteristic variables and the target variable using graphical charts such as histograms, box plots, and scatter plots
   - Compute the correlation matrix between variables
   - Interpret the results of the EDA to identify important characteristics that influence the price of electricity and significant relationships between variables

Note. To understand this phase, you are invited to consult the following link : Python for Data Science: Implementing Exploratory Data Analysis (EDA) and K-Means Clustering

3.4 Data Modeling

Many Machine Learning algorithms could be used to train prediction models from the data. We propose to consider and implement the following six regression models :

1. Linear Regression
2. Regularied Linear Regression (RIDGE Regression/ LASSO regression) : see the following link : Régression Ridge, Lasso et nouvel estimateur
3. K-Nearest Neighbors for Regression (K-NN, k-Nearest Neighbors regressor)
4. Decision Tree Regression
5. **Bonus**, Random Forest for regresion (Random Forest regressor)

You should understand both variants of regularized linear regression and describe them succinctly in your final report. Similarly for the bonus method.

3.5 Evaluation

The score function (metric) used is the Spearman's correlation between the participant's output and the actual daily price changes over the testing data set sample.
Models will be evaluated using regression-related metrics such as :

- la corrélation de Spearman,
- Determination coefficient R^2
- Root mean square error (RMSE).

Within this step you have to :

1. Optimize each model : Vary the hyperparameters of each model and retain those that result in the best performance.
2. Comparison of different models : Compare the performance of data prediction algorithms and choose the best performing one by making a ranking.
3. For the best model selected, evaluate the importance of the variables (attributes) that contributed to the best prediction (Evaluez l'importance des variables, How to Calculate Feature Importance With Python, Feature importance, Feature Importance Explained)

4 Programming Language and librairies

1. Programming Language : Python
2. Librairies :
   - Data Manipulation : Pandas
   - Data Visualisation : Matplotlib & Seaborn
   - Data Modeling : Scikit-learn

5 Delivrables

- Jupyter notebooks (you can use Goggle Colab Notebook - Google Colaboratory Notebooks) allowing to review and understand all the work (it must include an explanation of the ML methods not seen in class, the experimental results and their interpretations).
- **Or** A Zip file of python scripts and a concise report describing your project (it must include an explanation of the ML methods not seen in class, the experimental results and their interpretations).
- A 10 minutes presentation.

6 Deadlines

1. Project sent by mail : September 11, 2024 at 11:55.
2. Duration of the project : 10 days

7 Rating/Score

Your project will be graded by your teacher during the last session of the course (Juin 2024) based onyour presentation and the provided work.

1. Report (clarity, interpretations) : 50%
2. Code (Meets needs quality) : 50%